A Generalised Method for Empirical Game Theoretic Analysis

Karl Tuyls
DeepMind
London, UK
karltuyls@google.com

Julien Perolat
DeepMind
London, UK
perolat@google.com

Marc Lanctot
DeepMind
Edmonton, Canada
lanctot@google.com

Joel Z Leibo
DeepMind
London, UK
jzl@google.com

Thore Graepel
DeepMind
London, UK
thore@google.com

ABSTRACT

This paper provides theoretical bounds for empirical game theoretical analysis of complex multi-agent interactions. We provide insights in the empirical meta game showing that a Nash equilibrium of the meta-game is an approximate Nash equilibrium of the true underlying game. We investigate and show how many data samples are required to obtain a close enough approximation of the underlying game. Additionally, we extend the meta-game analysis methodology to asymmetric games. The state-of-the-art has only considered empirical games in which agents have access to the same strategy sets and the payoff structure is symmetric, implying that agents are interchangeable. Finally, we carry out an empirical illustration of the generalised method in several domains, illustrating the theory and evolutionary dynamics of several versions of the *AlphaGo* algorithm (symmetric), the dynamics of the Colonel Blotto game played by human players on Facebook (symmetric), and an example of a meta-game in Leduc Poker (asymmetric), generated by the PSRO multi-agent learning algorithm.

KEYWORDS

Empirical Games; Asymmetric Games; Replicator Dynamics

1 INTRODUCTION

Using game theory to examine multi-agent interactions in complex systems is a non-trivial task. Works by Walsh et al. [14, 15] and Wellman et al. [16], have shown the great potential of using heuristic strategies and empirical game theory to examine such interactions at a higher meta-level, instead of trying to capture the decision-making processes at the level of the atomic actions involved. Doing this turns the interaction in a smaller normal form game, or meta-game, with the higher-level strategies now being the primitive actions of the game, making the complex multi-agent interaction amenable to game theoretic analysis.

Others have built on this empirical game theoretic methodology and applied these ideas to no limit Texas hold'em Poker and various types of double auctions for example, see [4, 7–9, 12], showing that a game theoretic analysis at the level of meta-strategies yields novel insights into the type and form of interactions in complex systems. A major limitation of this empirical game theoretic approach is that it comes without theoretical guarantees on the approximation of the true underlying game by an estimated game based on sampled data, and that it is unclear how many data samples are required to achieve a good approximation. Additionally, the method remains limited to symmetric situations, in which the agents or players have access to the same set of strategies, and are interchangeable. One approach is to ignore asymmetry (types of players), and average over many samples of types resulting in a single expected payoff to each player in each entry of the meta-game payoff table. Many real-world situations though are asymmetric in nature and involve various roles for the agents that participate in the interactions. For instance, buyers and sellers in auctions, or games such as Scotland Yard, but also different roles in e.g. robotic soccer (defender vs striker) and even natural language (hearer vs speaker).

In this paper we tackle these problems. We prove that a Nash equilibrium of the estimated game is a 2ϵ-Nash equilibrium of the real underlying game, showing that we can closely approximate the real Nash equilibrium as long as we have enough data samples from which to build the meta-game payoff table. Furthermore, we also examine how much data samples are required to confidently approximate the underlying game. We also show how to generalise the heuristic payoff or meta-game method introduced by Walsh *et al.* to two-population asymmetric games.

Finally, we illustrate the generalised method in several domains. We carry out an experimental illustration on the *AlphaGo* algorithm [10], Colonel Blotto [5] and an asymmetric Leduc poker game. In the *AlphaGo* experiments we show how a symmetric meta-game analysis can provide insights into the evolutionary dynamics and strengths of various versions of the *AlphaGo* algorithm while it was being developed, and how intransitive behaviour can occur by introducing a non-related strategy. In the Colonel Blotto game we illustrate how the methodology can provide insights into how humans play this game, constructing several symmetric meta-games from data collected on Facebook. Finally, we illustrate the method in Leduc poker, by examining an asymmetric meta-game, generated by a recently introduced multiagent reinforcement learning algorithm, policy-space response oracles (PSRO) [6]. For this analysis

we rely on some theoretical results that connect an asymmetric normal form game to its symmetric counterparts [13].

2 PRELIMINARIES

In this section, we introduce the necessary background to describe our game theoretic meta-game analysis of the repeated interaction between p players.

2.1 Normal Form Games:

In a p-player Normal Form Game (NFG), players are involved in a single round strategic interaction. Each player i chooses a strategy π^i from a set of k strategy $S^i = \{\pi_1^i, \ldots, \pi_k^i\}$ and receives a payoff $r^i(\pi^1, \ldots, \pi^p) : S^1 \times \cdots \times S^p \to \mathbb{R}$. For the sake of simplicity, we will write $\boldsymbol{\pi}$ the joint strategy $(\pi^1, \ldots, \pi^p)$ and $\boldsymbol{r}(\boldsymbol{\pi})$ the joint reward $(r^1(\boldsymbol{\pi}), \ldots, r^p(\boldsymbol{\pi}))$. Then a p-player NFG is a tuple $G = (S^1, \ldots, S^p, r^1, \ldots, r^p)$. Each player interacts in this game by following a strategy profile x^i which is a probability distribution over S^i.

A symmetric NFG captures interactions where players can be interchanged. The first condition is therefore that the strategy sets are the same for all players, (*i.e.* $\forall i, j \ S_i = S_j$ and will be written S). In a symmetric NFG, if a permutation is applied to the joint strategy $\boldsymbol{\pi}$, the joint payoff is permuted accordingly. Formally, a game G is symmetric if for all permutations of p elements σ we have $\boldsymbol{r}(\boldsymbol{\pi}_\sigma) = \boldsymbol{r}_\sigma(\boldsymbol{\pi})$ (where $\boldsymbol{\pi}_\sigma = (\pi^{\sigma(1)}, \ldots, \pi^{\sigma(p)})$ and $\boldsymbol{r}_\sigma(\boldsymbol{\pi}) = (r^{\sigma(1)}(\boldsymbol{\pi}), \ldots, r^{\sigma(p)}(\boldsymbol{\pi}))$). So for a game to be symmetric there are two conditions, the players need to have access to the same strategy set and the payoff structure needs to be symmetric, such that players are interchangeable. If one of these two conditions is violated the game is asymmetric.

In the asymmetric case our analysis will focus on the two-player case (two roles) and thus we introduce specific notations for the sake of simplicity. In a two-player normal-form game, each player's payoff can be seen as a $k \times k$ matrix. We will write $A = (a_{l,l'})_{1 \leq l, l' \leq k}$ for the payoff matrix of player one (*i.e.* $a_{l,l'} = r^1(\pi_l^1, \pi_{l'}^2)$) and $B = (b_{l,l'})_{1 \leq l, l' \leq k}$ for the payoff matrix of player two (*i.e.* $b_{l,l'} = r^2(\pi_l^1, \pi_{l'}^2)$). In this two-player game, the column vector x is the strategy of player one and y the one of player two. In the end, a two player NFG is defined by the following tuple $G = (S^1, S^2, A, B)$.

2.2 Nash Equilibrium

In a two-player game, a pair of strategies (x, y) is a Nash equilibrium of the game (A, B) if no player has an incentive to switch from their current strategy. In other words, (x, y) is a Nash equilibrium if $x^\top A y = \max A y$ and $x^\top B y = \max x^\top B$.

Evolutionary game theory often consider a single strategy x that plays against itself. In this situation, the game is said to have a single population. In a single population game, x is a Nash equilibrium if $x^\top A x = \max A x$.

2.3 Replicator Dynamics

The replicator dynamics equation describes how a strategy profile evolves in the midst of others. This evolution is described according to a first order dynamical system. In a two-player NFG (A, B, S^1, S^2), the replicator equations are defined as:

$$\dot{x}_l = x_l \left((Ay)_l - x^\top A y\right) \qquad \dot{y}_{l'} = y_{l'} \left((x^\top B)_{l'} - x^\top B y\right) \quad (1)$$

The dynamics defined by these two coupled differential equations changes the strategy profile to increase the probability of the strategies that have the best return or are the *fittest*.

In the case of a symmetric two-player game ($A = B^\top$), the replicator equations assume that both players play the same strategy profile (*i.e.* player one and two play according to x) and the dynamics is defined as follows:

$$\dot{x}_l = x_l \left((Ax)_l - x^\top A x\right) \quad (2)$$

2.4 Meta Games

A meta game is a simplified model of a complex interaction. In order to analyze complex games like e.g. poker, we do not need to consider all possible strategies but a set of relevant meta-strategies that are often played [9]. These meta strategies (or styles of play), over atomic actions, are commonly played by players such as for instance "passive/aggressive" or "tight/loose" in poker. A p-type meta game is now a p-player repeated NFG where players play a limited number of meta strategies. Following our poker example, the strategy set of the meta game will now be defined as the set $S = \{\text{"aggressive"}, \text{"tight"}, \text{"passive"}\}$ and the reward function as the outcome of a game between p-players using different profiles.

3 METHOD

There are now two possibilities, either the meta-game is symmetric, or it is asymmetric. We will start with the simpler symmetric case, which has been studied in empirical game theory, then we continue with asymmetric games, in which we consider two populations, or roles.

3.1 Symmetric Meta Games

We consider a set of agents or players A with $|A| = n$ that can choose a strategy from a set S with $|S| = k$ and can participate in one or more p-type meta-games with $p \leq n$. If the game is symmetric then the formulation of meta strategies has the advantage that the payoff for a strategy does not depend on which player has chosen that strategy and consequently the payoff for that strategy only depends on the composition of strategies it is facing in the game and not on who is playing the strategy. This symmetry has been the main focus of the use of empirical game theory analysis [7, 9, 14, 16].

If we were to construct a classical payoff table for $\mathbf{r}$ we would require k^p entries in the table (which becomes large very quickly). Since all players can choose from the same strategy set and all players receive the same payoff for being in the same situation, we can simplify our payoff table.

Let N be a matrix, where each row N_i contains a discrete distribution of p players over k strategies. The matrix yields $\binom{p+k-1}{p}$ rows. Each distribution over strategies can be simulated (or derived from data), returning a vector of expected rewards $u(N_i)$. Let U be a matrix which captures the rewards corresponding to the rows in N, i.e., $U_i = u(N_i)$. We refer to a meta payoff table as $M = (N, U)$. So each row yields a *discrete profile* $(n_{\pi_1}, \ldots, n_{\pi_k})$ indicating exactly how many players play each strategy, with $\sum_j n_{\pi_j} = p$. A strategy profile $\mathbf{x}$ then equals $(\frac{n_{\pi_1}}{p}, \ldots, \frac{n_{\pi_k}}{p})$.

Suppose we have a meta-game with 3 meta-strategies ($|S| = 3$) and 6 players ($|A| = 6$) that interact in a 6-type, this leads to a meta game payoff table of 28 entries (which is a good reduction from 3^6 cells. An important advantage of this type of table is that it easily extends to many agents, as opposed to the classical payoff matrix.

Table 1 provides an example for three strategies π_1, π_2 and π_3. The left-hand side expresses the discrete profiles and corresponds to matrix N, while the right-hand side gives the payoffs for playing any of the strategies given the discrete profile and corresponds to matrix U.

$$P = \begin{pmatrix} N_{i1} & N_{i2} & N_{i3} & U_{i1} & U_{i2} & U_{i3} \\ 6 & 0 & 0 & 0 & 0 & 0 \\ & \cdots & & & \cdots & \\ 4 & 0 & 2 & -0.5 & 0 & 1 \\ & \cdots & & & \cdots & \\ 0 & 0 & 6 & 0 & 0 & 0 \end{pmatrix}$$

Table 1: An example of a meta game payoff table

In order to analyse the evolutionary dynamics of high-level meta-strategies, we also need to estimate the expected payoff of such strategies relative to each other. In evolutionary game theoretic terms, this is the relative fitness of the various strategies, dependent on the current frequencies of those strategies in the population.

In order to approximate the payoff for an arbitrary mix of strategies in an infinite population of replicators distributed over the species according to $\mathbf{x}$, p individuals are drawn randomly from the infinite distribution. The probability for selecting a specific row N_i can be computed from $\mathbf{x}$ and N_i as

$$P(N_i|\mathbf{x}) = \binom{p}{N_{i1}, N_{i2}, \ldots, N_{ik}} \prod_{j=1}^{k} x_j^{N_{ij}}.$$

The expected payoff of strategy π^i, $r^i(\mathbf{x})$, is then computed as the weighted combination of the payoffs given in all rows:

$$r^i(\mathbf{x}) = \frac{\sum_j P(N_j|\mathbf{x})U_{ji}}{1 - (1 - x_i)^p}.$$

This expected payoff function can be used in Equation 2 to compute the evolutionary population change according to the replicator dynamics by replacing $(Ax)_i$ by $r^i(\mathbf{x})$. Note that we need to re-normalize (denominator) by ignoring rows that do not contribute to the payoff of a strategy because it is not present in the distribution N_j in the HPT.

3.2 Asymmetric Meta Games

One can now wonder how the previously introduced method extends to asymmetric games, which has not been considered in the literature. An example of an asymmetric game is the famous battle of the sexes game illustrated in Table 2. In this game both players do have the same strategy sets, i.e., go to the opera or go to the movies, however, the corresponding payoffs for each are different, expressing the differences in preferences that both players have.

	O	M
O	3, 2	0, 0
M	0, 0	2, 3

Table 2: Battle of the Sexes game: strategies O and M correspond to going to the Opera and going to the Movies respectively.

	C_1	C_2	C_3
R_1	r_{11}, c_{11}	r_{12}, c_{12}	r_{13}, c_{13}
R_2	r_{21}, c_{21}	r_{22}, c_{22}	r_{23}, c_{23}
R_3	r_{31}, c_{31}	r_{32}, c_{32}	r_{33}, c_{33}

Table 3: General 3x3 normal form game.

If we aim to carry out a similar evolutionary analysis as in the symmetric case, restricting ourselves to two populations or roles, we will need two meta game payoff tables, one for each player over its own strategy set. We will also need to use the asymmetric version of the replicator dynamics as shown in Equation 1. Additionally, in order to compute the right payoffs for every situation we will have to interpret a discrete strategy profile in the meta-table slightly

different. Suppose we have a 2-type meta game, with three strategies in each player's strategy set. We introduce a generalisation of our meta-table for both players by means of an example shown in Table 4, which corresponds to the general NFG shown in Table 3.

$$P = \begin{pmatrix} N_{i1,j1} & N_{i2,j2} & N_{i3,j3} & U_{i1,j1} & U_{i2,j2} & U_{i3,j3} \\ (1,1) & 0 & 0 & (r_{11},c_{11}) & 0 & 0 \\ & \cdots & & & \cdots & \\ (1,0) & (0,1) & 0 & (r_{12},0) & (0,c_{12}) & 0 \\ (0,1) & (1,0) & 0 & (0,c_{21}) & (r_{21},0) & 0 \\ & \cdots & & & \cdots & \\ 0 & 0 & (1,1) & 0 & 0 & (r_{33},c_{33}) \end{pmatrix}$$

Table 4: An example of an asymmetric meta game payoff table

Let's have a look at the first entry in Table 4, i.e., $[(1,1),0,0]$. This entry means that both agents (i and j) are playing their first strategy, expressed by $N_{i1,j1}$, meaning the number of agents N_{i1} playing strategy π_i^1 in the first population equals 1 and that the number of agents N_{j1} playing strategy π_j^2 in the second population equals 1 as well. The corresponding payoff for each player $U_{i1,j1}$ equals (r_{11},c_{11}). Now lets have a look at the discrete profiles: $[(1,0),(0,1),0]$ and $[(0,1),(1,0),0]$. The first one means that the first player is playing its first strategy while the second player is playing their second strategy. The corresponding payoffs are r_{12} for the first player and c_{12} for the second player. The profile $[(0,1),(1,0),0]$ shows the reverted situation in which the second player plays his first strategy and the first player plays his second strategy, yielding payoffs r_{21} and c_{21} for the first player and second player respectively. In order to turn the table into a similar format as for the symmetric case, we can now introduce p meta-tables, one for each player. More precisely, we get Tables 5 and 6 for players 1 and 2 respectively.

$$P = \begin{pmatrix} N_{i1,j1} & N_{i2,j2} & N_{i3,j3} & U_{i1,j1} & U_{i2,j2} & U_{i3,j3} \\ 2 & 0 & 0 & r_{11} & 0 & 0 \\ & \cdots & & & \cdots & \\ 1 & 1 & 0 & r_{12} & r_{21} & 0 \\ & \cdots & & & \cdots & \\ 0 & 0 & 2 & 0 & 0 & r_{33} \end{pmatrix}$$

Table 5: A decomposed asymmetric meta payoff table for Player 1.

$$Q = \begin{pmatrix} N_{i1,j1} & N_{i2,j2} & N_{i3,j3} & U_{i1,j1} & U_{i2,j2} & U_{i3,j3} \\ 2 & 0 & 0 & c_{11} & 0 & 0 \\ & \cdots & & & \cdots & \\ 1 & 1 & 0 & c_{12} & c_{21} & 0 \\ & \cdots & & & \cdots & \\ 0 & 0 & 2 & 0 & 0 & c_{33} \end{pmatrix}$$

Table 6: A decomposed asymmetric meta payoff table for Player 2.

One needs to take care in correctly interpreting these tables. Let's have a look at row $[1,1,0]$ for instance. This should now be interpreted in two ways: one, the first player plays his first strategy while the other player plays his second strategy and he receives a payoff of r_{12}, two, the first player plays his second strategy while the other player plays his first strategy and receives a payoff of r_{21}. The expected payoff $r^i(\mathbf{x})$ can now be estimated in the same way as explained for the symmetric case as we will be relying on symmetric replicator dynamics by decoupling asymmetric games in their *symmetric counterparts* (explained in the next section).

3.3 Linking symmetric and asymmetric games

Here we summarize the most important results on the link between an asymmetric game and its symmetric counterpart games. For a full treatment and discussion of these results see [13]. In a nutshell, this work proves that if x, y is a Nash equilibrium of the bimatrix

game (A, B) (where x and y have the same support[1]), then y is a Nash equilibrium of the single population, or symmetric, game A and x is a Nash equilibrium of the single population, or symmetric, game $B^\top$. Both symmetric games are called the *counterpart games* of the asymmetric game (A, B). The reverse is also true: If y is a Nash equilibrium of the single population game A and x is a Nash equilibrium of the single population game $B^\top$ (and if x and y have the same support), then x, y is a Nash equilibrium of the game (A, B). In our empirical analysis, we use this property to analyze an asymmetric games (A, B) by looking at the counterpart single population games A and $B^\top$.

4 THEORETICAL INSIGHTS

As illustrated in the previous section the procedure for empirical meta-game analysis consists of two parts. Firstly, one needs to construct an empirical meta-game utility function for each player. This step can be performed using logs of interactions between players, or by playing the game sufficiently enough. Secondly, one expects that analyzing the empirical game will give insights in the true underlying game itself (i.e. the game from which we sample).This section provides insights in the following: how much data is enough to generate a good approximation of the true underlying game? Is uniform sampling over actions or strategies the right method?

4.1 Main Lemma

Sometimes players receive a stochastic reward $R^i(\pi^1, \ldots, \pi^p)$ for a given joint action $\boldsymbol{\pi}$. The underlying game we study is $r^i(\pi^1, \ldots, \pi^p) = E\left[R^i(\pi^1, \ldots, \pi^p)\right]$ and for the sake of simplicity the joint action of every player but player i will be written $\boldsymbol{\pi}^{-i}$. In the two following definitions, we introduce the concept of Nash equilibrium and ϵ-Nash equilibrium in p-player games (as we only introduced it in the 2-player case):

Definition : A joint strategy $\boldsymbol{x} = (x^1, \ldots, x^p) = (x^{-i}, \boldsymbol{x}^i)$ is a Nash equilibrium if for all i:

$$E_{\boldsymbol{\pi} \sim \boldsymbol{x}}\left[r^i(\boldsymbol{\pi})\right] = \max_{\pi^i} E_{\boldsymbol{\pi}^{-i} \sim \boldsymbol{x}^{-i}}\left[r^i(\pi^i, \boldsymbol{\pi}^{-i})\right]$$

Definition : A joint strategy $\boldsymbol{x} = (x^1, \ldots, x^p) = (x^{-i}, \boldsymbol{x}^i)$ is an ϵ-Nash equilibrium if for all i:

$$\max_{\pi^i} E_{\boldsymbol{\pi}^{-i} \sim \boldsymbol{x}^{-i}}\left[r^i(\pi^i, \boldsymbol{\pi}^{-i})\right] - E_{\boldsymbol{\pi} \sim \boldsymbol{x}}\left[r^i(\boldsymbol{\pi})\right] \le \epsilon$$

When running an analysis on a meta game, we do not have access to the average reward function $r^i(\pi^1, \ldots, \pi^p)$ but to an empirical estimate $\hat{r}^i(\pi^1, \ldots, \pi^p)$. The following lemma shows that a Nash equilibrium for the empirical game $\hat{r}^i(\pi^1, \ldots, \pi^p)$ is an 2ϵ-Nash equilibrium for the game $r^i(\pi^1, \ldots, \pi^p)$ where $\epsilon = \sup_{\boldsymbol{\pi}, i} |\hat{r}^i(\boldsymbol{\pi}) - r^i(\boldsymbol{\pi})|$.

Lemma: If $\boldsymbol{x}$ is a Nash equilibrium for $\hat{r}^i(\pi^1, \ldots, \pi^p)$, then it is an 2ϵ-Nash equilibrium for the game $r^i(\pi^1, \ldots, \pi^p)$ where $\epsilon = \sup_{\boldsymbol{\pi}, i} |r^i(\boldsymbol{\pi}) - \hat{r}^i(\boldsymbol{\pi})|$.

Proof:
First we have the following relation:

$$E_{\boldsymbol{\pi} \sim \boldsymbol{x}}\left[r^i(\boldsymbol{\pi})\right] = E_{\boldsymbol{\pi} \sim \boldsymbol{x}}\left[\hat{r}^i(\boldsymbol{\pi})\right] + E_{\boldsymbol{\pi} \sim \boldsymbol{x}}\left[r^i(\boldsymbol{\pi}) - \hat{r}^i(\boldsymbol{\pi})\right]$$

---

[1] x and y have the same support if $I_x = I_y$ where $I_x = \{i \mid x_i > 0\}$ and $I_y = \{i \mid y_i > 0\}$

Then:

$$E_{\boldsymbol{\pi}^{-i} \sim \boldsymbol{x}^{-i}}\left[r^i(\pi^i, \boldsymbol{\pi}^{-i})\right] = E_{\boldsymbol{\pi}^{-i} \sim \boldsymbol{x}^{-i}}\left[\hat{r}^i(\pi^i, \boldsymbol{\pi}^{-i})\right]$$
$$+ E_{\boldsymbol{\pi}^{-i} \sim \boldsymbol{x}^{-i}}\left[r^i(\pi^i, \boldsymbol{\pi}^{-i}) - \hat{r}^i(\pi^i, \boldsymbol{\pi}^{-i})\right]$$
$$\max_{\pi^i} E_{\boldsymbol{\pi}^{-i} \sim \boldsymbol{x}^{-i}}\left[r^i(\pi^i, \boldsymbol{\pi}^{-i})\right] \le \max_{\pi^i} E_{\boldsymbol{\pi}^{-i} \sim \boldsymbol{x}^{-i}}\left[\hat{r}^i(\pi^i, \boldsymbol{\pi}^{-i})\right]$$
$$+ \max_{\pi^i} E_{\boldsymbol{\pi}^{-i} \sim \boldsymbol{x}^{-i}}\left[r^i(\pi^i, \boldsymbol{\pi}^{-i}) - \hat{r}^i(\pi^i, \boldsymbol{\pi}^{-i})\right]$$

Finally,

$$\max_{\pi^i} E_{\boldsymbol{\pi}^{-i} \sim \boldsymbol{x}^{-i}}\left[r^i(\pi^i, \boldsymbol{\pi}^{-i})\right] - E_{\boldsymbol{\pi} \sim \boldsymbol{x}}\left[r^i(\boldsymbol{\pi})\right]$$
$$\le \underbrace{\max_{\pi^i} E_{\boldsymbol{\pi}^{-i} \sim \boldsymbol{x}^{-i}}\left[\hat{r}^i(\pi^i, \boldsymbol{\pi}^{-i})\right] - E_{\boldsymbol{\pi} \sim \boldsymbol{x}}\left[\hat{r}^i(\boldsymbol{\pi})\right]}_{=0 \text{ since } \boldsymbol{x} \text{ is a Nash equilibrium for } \hat{r}^i}$$
$$+ \underbrace{\max_{\pi^i} E_{\boldsymbol{\pi}^{-i} \sim \boldsymbol{x}^{-i}}\left[r^i(\pi^i, \boldsymbol{\pi}^{-i}) - \hat{r}^i(\pi^i, \boldsymbol{\pi}^{-i})\right]}_{\le \epsilon}$$
$$\underbrace{- E_{\boldsymbol{\pi} \sim \boldsymbol{x}}\left[r^i(\boldsymbol{\pi}) - \hat{r}^i(\boldsymbol{\pi})\right]}_{\le \epsilon}$$

$\square$

This lemma shows that if one can control the difference between $|r^i(\boldsymbol{\pi}) - \hat{r}^i(\boldsymbol{\pi})|$ uniformly over players and actions, then an equilibrium for the empirical game $\hat{r}^i(\pi^1, \ldots, \pi^p)$ is almost an equilibrium for the game defined by the average reward function $r^i(\pi^1, \ldots, \pi^p)$.

4.2 Finite Samples Analysis

This section details some concentration results. In practice, we often have access to a batch of observations of the underlying game. We will run our analysis on an empirical estimate of the game denoted by $\hat{r}^i(\boldsymbol{\pi})$. The question then will be either with which confidence can we say that a Nash equilibrium for $\hat{r}$ is a 2ϵ-Nash equilibrium, or for a fixed confidence, for which ϵ can we say that a Nash equilibrium for $\hat{r}$ is a 2ϵ-Nash equilibrium for r. In the case we have access to game play, the question is how many samples n do we need to assess that a Nash equilibrium for $\hat{r}$ is a 2ϵ-Nash equilibrium for r for a fixed confidence and a fixed ϵ. For the sake of simplicity, we will assume that all payoff are bounded in $[0, 1]$.

4.2.1 The batch scenario. Here we assume that we are given $n(i, \boldsymbol{\pi})$ independent samples to compute the empirical average $\hat{r}^i(\boldsymbol{\pi})$. Then, by applying Hoeffding's inequality we can prove the following result:

$$P\left(\sup_{\boldsymbol{\pi}, i} |r^i(\boldsymbol{\pi}) - \hat{r}^i(\boldsymbol{\pi})| < \epsilon\right) \ge \prod_{i \in \{1, \ldots, p\}} \prod_{\boldsymbol{\pi}} \left(1 - 2e^{\left(-2\epsilon^2 n(i, \boldsymbol{\pi})\right)}\right)$$

4.2.2 uniform sampling. In this section we assume that we have a budget of n samples per joint actions $\boldsymbol{\pi}$ and per player i. In that case we have the following bound:

$$P\left(\sup_{\boldsymbol{\pi}, i} |r^i(\boldsymbol{\pi}) - \hat{r}^i(\boldsymbol{\pi})| < \epsilon\right) \ge \left(1 - 2e^{\left(-2\epsilon^2 n\right)}\right)^{|S^1| \times \cdots \times |S^p| \times p}$$

Then, If we want $\sup_{\boldsymbol{\pi}, i} |r^i(\boldsymbol{\pi}) - \hat{r}^i(\boldsymbol{\pi})| < \epsilon$ with a probability of at least $1 - \delta$ we need at least $n = -\dfrac{\ln\left(1 - (1-\delta)^{\frac{1}{|S^1| \times \cdots \times |S^p| \times p}}\right)}{2\epsilon^2}$

5 EXPERIMENTS

This section presents experiments that illustrate the meta-game approach and its feasibility for examining strengths and weaknesses of higher-level strategies in various domains, including *AlphaGo*,

Colonel Blotto, and the meta-game generated by PSRO. Note that we restrict the meta-games to three strategies, as we can nicely visualise this in a phase plot, and these still provide useful information about the dynamics in the full strategy spaces.

5.1 AlphaGo

The data set under study consists of 7 *AlphaGo* variations and a a number of different Go strategies such as Crazystone and Zen (previously the state-of-the-art). α stands for the algorithm and the indexes r, v, p for the use of respectively *rollouts*, *value nets* and *policy nets* (e.g. α_{rvp} uses all 3). For a detailed description of these strategies see [10]. The meta-game under study here concerns a 2-type NFG with $|S| = 9$. We will look at various 2-faces of the larger simplex. Table 9 in [10] summarises all wins and losses between these various strategies (meeting several times), from which we can compute meta-game payoff tables.

5.1.1 Experiment 1: strong strategies. This first experiment examines three of the strongest *AlphaGo* strategies in the data-set, i.e., $\alpha_{rvp}, \alpha_{vp}, \alpha_{rp}$. As a first step we created a meta-game payoff table involving these three strategies, by looking at their pairwise interactions in the data set (summarised in Table 9 of [10]). This set contains data for all strategies on how they interacted with the other 8 strategies, listing the win rates that strategies achieved against one another (playing either as white or black) over several games. The meta-game payoff table derived for these three strategies is described in Table 7.

$$\begin{pmatrix} \alpha_{rvp} & \alpha_{vp} & \alpha_{rp} & U_{i1} & U_{i2} & U_{i3} \\ \hline 2 & 0 & 0 & 0.5 & 0 & 0 \\ 1 & 0 & 1 & 0.95 & 0 & 0.05 \\ 0 & 2 & 0 & 0 & 0.5 & 0 \\ 1 & 1 & 0 & 0.99 & 0.01 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0.5 \\ 0 & 1 & 1 & 0 & 0.39 & 0.61 \end{pmatrix}$$

Table 7: Meta-game payoff table generated from Table 9 in [10] for strategies $\alpha_{rvp}, \alpha_{vp}, \alpha_{rp}$

In Figure 1 we have plotted the directional field of the meta-game payoff table using the replicator dynamics for a number of strategy profiles **x** in the simplex strategy space. From each of these points in strategy space an arrow indicates the direction of flow, or change, of the population composition over the three strategies. Figure 2 shows a corresponding trajectory plot. From these plots one can easily observe that strategy α_{rvp} is a strong attractor and consumes the entire strategy space over the three strategies. This restpoint is also a Nash equilibrium. This result is in line with what we would expect from the knowledge we have of the strengths of these various learned policies. Still, the arrows indicate how the strategy landscape flows into this attractor and therefore provides useful information as we will discuss later.

5.1.2 Experiment 2: evolution and transitivity of strengths. We start by investigating the 2-face simplex involving strategies α_{rp}, α_{vp} and α_{rv}, for which we created a meta-game payoff table similarly as in the previous experiment (not shown). The evolutionary dynamics of this 2-face can be observed in Figure 4a. Clearly strategy α_{rp} is a strong attractor and dominates the two other strategies. We now replace this attractor by strategy α_{rvp} and plot its evolutionary dynamics in Figure 4b. What can be observed from both trajectory plots in Figure 4 is that the curvature is less pronounced

in plot 4b than it is in plot 4a. The reason for this is that the difference in strength between α_{rv} and α_{vp} is less obvious in the presence of an even stronger attractor than α_{rp}. This means that α_{rvp} is now pulling much stronger on both α_{rv} and α_{vp} and consequently the flow goes more directly to α_{rvp}. So even when a strategy space is dominated by one strategy, the curvature (or curl) is a promising measure for the strength of a meta-strategy.

What is worthwhile to observe from the *AlphaGo* dataset, and illustrated as a series in Figures 3 and 4, is that there is clearly an incremental increase in the strength of the *AlphaGo* algorithm going from version α_r to α_{rvp}, building on previous strengths, without any intransitive behaviour occurring, when only considering a strategy space formed by the *AlphaGo* versions.

Finally, as discussed in Section 4, we can now examine how good of an approximation an estimated game is. In the *AlphaGo* domain we only do this analysis for the games displayed in Figures 4a and 4b, as it is similar for the other experiments. We know that α_{rp} is a Nash equilibrium of the estimated game analyzed in Figure 4a (meta Table not shown). The outcome of α_{rp} against α_{rv} was estimated with $n_{\alpha_{rp}, \alpha_{rv}} = 63$ games (for the other pair of strategies we have $n_{\alpha_{vp}, \alpha_{rp}} = 65$ and $n_{\alpha_{vp}, \alpha_{rv}} = 133$). Because of the symmetry of the problem, the bound in section 4.2.1 is reduced to:

$$P\left(\sup_{\pi, i} |r^i(\pi) - \hat{r}^i(\pi)| < \epsilon\right) \geq \left(1 - 2e^{\left(-2\epsilon^2 n_{\alpha_{rp}, \alpha_{rv}}\right)}\right) \times \left(1 - 2e^{\left(-2\epsilon^2 n_{\alpha_{vp}, \alpha_{rp}}\right)}\right)$$
$$\times \left(1 - 2e^{\left(-2\epsilon^2 n_{\alpha_{vp}, \alpha_{rv}}\right)}\right)$$

Therefore, we can conclude that the strategy α_{rp} is an 2ϵ-Nash equilibrium (with $\epsilon = 0.15$) for the real game with probability at least 0.78. The same calculation would also give a confidence of 0.85 for the RD studied in Figure 4b for an $\epsilon = 0.15$ (as the number of samples are $(n_{\alpha_{rv}, \alpha_{vp}}, n_{\alpha_{vp}, \alpha_{rvp}}, n_{\alpha_{rp}, \alpha_{rv}}) = (65, 106, 91)$).

5.1.3 Experiment 3: cyclic behaviour. A final experiment investigates what happens if we add a *pre-AlphaGo* state-of-the-art algorithm to the strategy space. We have observed that even though α_{rvp} remains the strongest strategy, dominating all other *AlphaGo* versions and previous state-of-the-art algorithms, cyclic behaviour can occur, something that cannot be measured or seen from Elo ratings.[2] More precisely, we constructed a meta-game payoff table for strategies α_v, α_p and *Zen* (one of the previous commercial state-of-the-art algorithms). In Figure 5 we have plotted the evolutionary dynamics for this meta-game, and as can be observed there is a mixed equilibrium in strategy space, around which the dynamics cycle, indicating that *Zen* is capable of introducing in-transitivity, as α_v dominates α_p, α_p dominates *Zen* and *Zen* dominates α_v.

5.2 Colonel Blotto

Colonel Blotto is a resource allocation game originally introduced by Borel [2]. Two players interact, each allocating m troops over n locations. They do this separately without communication, after which both distributions are compared to determine the winner. When a player has more troops in a specific location, it wins that location. The player winning the most locations wins the game. This game has many game theoretic intricacies, for an analysis see [5]. Kohli et al. have run Colonel Blotto on Facebook (project

---

[2]An Elo rating or score is a measure to express the relative strength of a player, or strategy. It was named after Arpad Elo and originally introduced to rate chess players. For an introduction see e.g. [3]
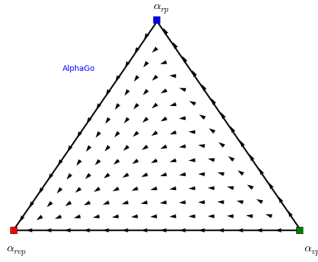
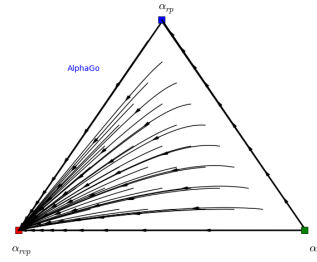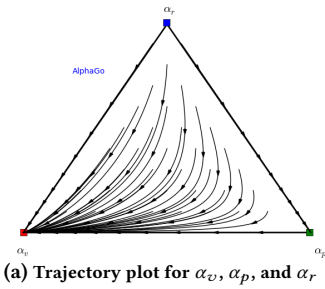**Figure 1: Directional field plot for the 2-face consisting of strategies** α_{rvp}, α_{vp}, α_{rp}

Figure 2: Trajectory plot for the 2-face consisting of strategies α_{rvp}, α_{vp}, α_{rp}

(a) Trajectory plot for α_v, α_p, **and** α_r
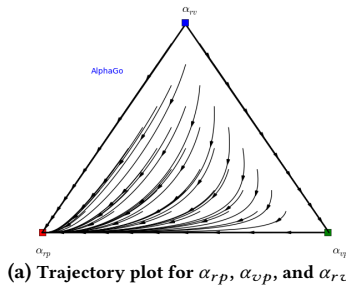
(a) Trajectory plot for α_{rp}, α_{vp}, **and** α_{rv}
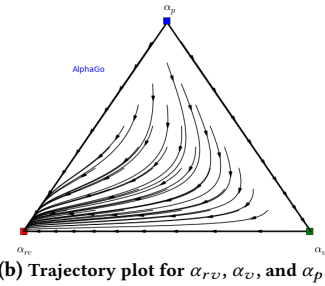
(b) Trajectory plot for α_{rv}, α_v, **and** α_p

Figure 3

(b) Trajectory plot for α_{rvp}, α_{vp}, **and** α_{rv}

Figure 4

(a)
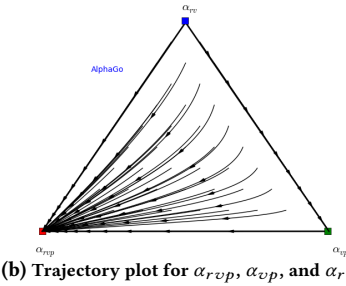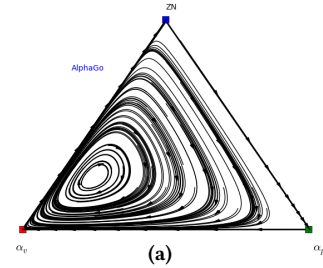
Figure 5: Intransitive behaviour for α_v, α_p, **and** Zen.

Strongest strategies		
Strategy	Frequency	Win rate
[36, 35, 24, 3, 2]	1	.74
[37, 37, 21, 3, 2]	17	.73
[35, 35, 26, 2, 2]	1	.73
[35, 34, 25, 3, 3]	3	.70
[35, 35, 24, 3, 3]	13	.70

Table 8: 5 of the strongest strategies played on Facebook.

Waterloo), collecting data describing how humans play this game, with each player having $m = 100$ troops and considering $n = 5$ battlefields. The number of strategies in the game is vast: a game with m troops and n locations has $\binom{m+n-1}{n-1}$ strategies.

Based on Kohli et al. we carry out a meta game analysis of the *strongest strategies* and the *most frequently played strategies* on Facebook. We have a look at several 3-strategy simplexes, which can be considered as 2-faces of the entire strategy space.

An instance of a strategy in the game of Blotto will be denoted as follows: $[t_1, t_2, t_3, t_4, t_5]$ with $\sum_i t_i = 100$. All permutations σ_i in this division of troops belong to the same strategy. We assume that permutations are chosen uniformly by a player. Note that in this game there is no need to carry out the theoretical analysis of the approximation of the meta-game, as we are are not examining heuristics or strategies over Blotto strategies, but rather these strategies themselves, for which the payoff against any other strategy will always be the same (by computation). Nevertheless, carrying out a meta-game analysis reveals interesting information.

5.2.1 Experiment 1: Top performing strategies. In this first experiment we examine the dynamics of the simplex consisting of the three best scoring strategies from the study of [5]: [36, 35, 24, 3, 2], [37, 37, 21, 3, 2], and [35, 35, 26, 2, 2], see Table 8. In a first step we compute a meta-game payoff table for these three strategies. The interactions are pairwise, and the expected payoff can be easily computed, assuming a uniform distribution for different permutations of a strategy. This normalised payoff is shown in Table 9.

Using table 9 we can compute evolutionary dynamics using the standard replicator equation. The resulting trajectory plot can be observed in Figure 6a. The first thing we see is that we have one strong attractor, i.e, strategy $s_2 = [37, 37, 21, 3, 2]$ and there is transitive behaviour, meaning that [36, 35, 24, 3, 2] dominates [35, 35, 26, 2, 2], [37, 37, 21, 3, 2] dominates [36, 35, 24, 3, 2], and [37, 37, 21, 3, 2] dominates [35, 35, 26, 2, 2]. Although [37, 37, 21, 3, 2] is the strongest

$$\begin{pmatrix} s_1 & s_2 & s_3 & U_{i1} & U_{i2} & U_{i3} \\ 2 & 0 & 0 & 0.5 & 0 & 0 \\ 1 & 0 & 1 & 0.66 & 0 & 0.34 \\ 0 & 2 & 0 & 0 & 0.5 & 0 \\ 1 & 1 & 0 & 0.33 & 0.67 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0.5 \\ 0 & 1 & 1 & 0 & 0.75 & 0.25 \end{pmatrix}$$

Table 9: Meta-game payoff table generated for strategies s_1 = [36, 35, 24, 3, 2], s_2 = [37, 37, 21, 3, 2], **and** s_3 = [35, 35, 26, 2, 2].

Most played strategies	
Strategy	Frequency
[34, 33, 33, 0, 0]	271
[20, 20, 20, 20, 20]	235
[33, 1, 33, 0, 33]	127
[1, 32, 33, 1, 33]	97
[35, 30, 35, 0, 0]	68
[0, 100, 0, 0, 0]	67
[10, 10, 35, 35, 10]	58
[25, 25, 25, 25, 0]	50

Table 10: The 8 most frequently played strategies on Facebook.
strategy in this 3-strategy meta-game, the win rates (computed over all played strategies in project Waterloo) indicate that strategy [36, 35, 24, 3, 2] was more successful on Facebook. The differences are minimal, and on average it is better to choose [37, 37, 21, 3, 2], which was also the most frequently chosen strategy from the set of strong strategies, see Table 8. We show a similar plot for the evolutionary dynamics of strategies [35, 34, 25, 3, 3], [37, 37, 21, 3, 2], and [35, 35, 24, 3, 3] in Figure 6b, which are three of the most frequently played strong strategies from Table 8.

5.2.2 Experiment 2: most frequently played strategies. We compared the evolutionary dynamics of the eight most frequently played strategies and present here a selection of some of the results. The meta-game under study in this domain concerns a 2-type repeated NFG G with $|S| = 8$. We will look at various 2-faces of the 8-simplex. The top eight most frequently played strategies are shown in Table 10. First we investigate the strategies [20, 20, 20, 20, 20], [1, 32, 33, 1, 33], and [10, 10, 35, 35, 10] from our strategy set. In Table 11 we show the resulting meta-game payoff table of this 2-face simplex. Using this table we can again compute the replicator dynamics and investigate the trajectory plots in Figure 7a. We observe that the dynamics cycle around a mixed Nash equilibrium (every interior rest point is a Nash equilibrium). This intransitive behaviour makes sense by looking at the pairwise interactions between strategies and the corresponding payoffs they receive from Table 9. The expected payoff for [20, 20, 20, 20, 20], when playing against [1, 32, 33, 1, 33] will be lower than the expected payoff for [1, 32, 33, 1, 33]. Similarly, [1, 32, 33, 1, 33] will be dominated by [10, 10, 35, 35, 10] when they meet, and to make the cycle complete, [10, 10, 35, 35, 10] will receive a lower expected payoff against [20, 20, 20, 20, 20]. As such, the behaviour will cycle around a the Nash equilibrium.

$$\begin{pmatrix} s_1 & s_2 & s_3 & U_{i1} & U_{i2} & U_{i3} \\ 2 & 0 & 0 & 0.5 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0.5 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0.5 \\ 0 & 1 & 1 & 0 & 0.1 & 0.9 \end{pmatrix}$$

Table 11: Meta-game payoff table generated for strategies s_1 = [20, 20, 20, 20, 20], s_2 = [1, 32, 33, 1, 33], **and** s_3 = [10, 10, 35, 35, 10].

An interesting question is where human players are situated in this cyclic behaviour landscape. In Figure 7b we show the same

trajectory plot but added a red marker to indicate the strategy profile based on the frequencies of these 3 strategies played by human players. This is derived from Table 10 and the profile vector is (0.6, 0.25, 0.15). If we assume that the human agents optimise their behaviour in a *survival of the fittest* style they will cycle along the red trajectory. In Figure 7c we illustrate similar intransitive behaviour for three other frequently played strategies.

5.3 PSRO-generated Meta-Game

We now turn our attention to an asymmetric game. Policy Space Response Oracles (PSRO) is a multiagent reinforcement learning process that reduces the strategy space of large extensive-form games via iterative best response computation. PSRO can be seen as a generalized form of fictitious play that produces approximate best responses, with arbitrary distributions over generated responses computed by meta-strategy solvers. One application of PSRO was applied to a commonly-used benchmark problem known as Leduc poker [11], except with a fixed action space and penalties for taking illegal moves. Therefore PSRO learned to play from scratch, without knowing which moves were legal. Leduc poker has a deck of 6 cards (jack, queen, king in two suits). Each player receives an initial private card, can bet a fixed amount of 2 chips in the first round, 4 chips in the second round, with a maximum of two raises in each round. A public card is revealed before the second round starts.

In Table 12 we present such an asymmetric 3×3 2-player game generated by the first few epochs of PSRO learning to play Leduc Poker. In the game illustrated here, each player has three strategies that, for ease of the exposition, we call $\{A, B, C\}$ for player 1, and $\{D, E, F\}$ for player 2. Each one of these strategies represents an approximate best response to a distribution over previous opponent strategies. In Table 13 we show the two symmetric counterpart games (see section 3.3) of the empirical game produced by PSRO.

	D	E	F
A	−2.26, 0.02	−2.06, −1.72	−1.65, −1.43
B	−4.77, −0.13	−4.02, −3.54	−5.96, −2.30
C	−2.71, −1.77	−2.52, −2.94	−6.10, 1.06

Table 12: Asymmetric PSRO meta game applied to Leduc poker.

	A	B	C			D	E	F
A	−2.26	−2.06	−1.65		D	0.02	−1.72	−1.43
B	−4.77	−4.02	−5.96		E	−0.13	−3.54	−2.30
C	−2.71	−2.52	−6.10		F	−1.77	−2.94	1.06

Table 13: Left - first counterpart game of the PSRO empirical game. Right - second counterpart game of the PSRO empirical game.

Again we can now analyse the equilbrium landscape of this game, but now using the asymmetric meta-game payoff table and the decomposition result introduced in section 3.3. Since the PSRO meta game is asymmetric we need two populations for the asymmetric replicator equations. Analysing and plotting the evolutionary asymmetric replicator dynamics now quickly becomes very tedious as we deal with two simplices, one for each player. More precisely, if we consider a strategy profile for one player in its corresponding simplex, and that player is adjusting its strategy, this will immediately cause the second simplex to change, and vice versa. Consequently, it is not straightforward anymore to analyse the dynamics.

In order to facilitate the process of analysing the dynamics we can apply the counterpart theorems to remedy the problem. In Figures 8 and 9 we show the evolutionary dynamics of the counterpart games. As can be observed in Figure 8 the first counterpart game has only one equilibrium, i.e., a pure Nash equilibrium in
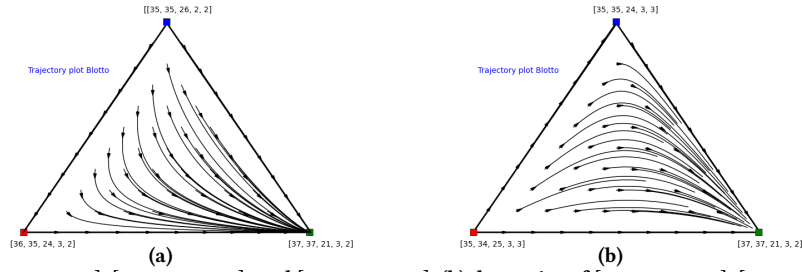
Figure 6: (a) dynamics of [36, 35, 24, 3, 2], [37, 37, 21, 3, 2], **and** [35, 35, 26, 2, 2]. **(b) dynamics of** [35, 34, 25, 3, 3], [37, 37, 21, 3, 2], **and** [35, 35, 24, 3, 3].
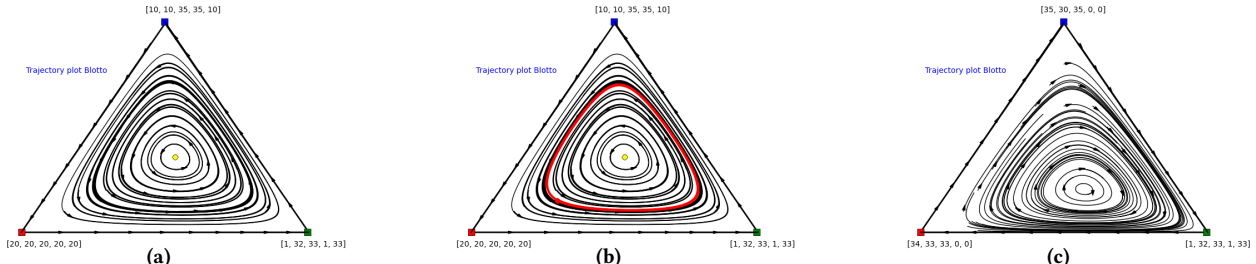

Figure 7: Dynamics of 3 2-faces of the 8-simplex: (a) Nash eq. (b) Human play (c) Another example of intransitive behaviour

which both players play strategy A, which absorbs the entire strategy space. Looking at Figure 9 we see the situation is a bit more complex in the second counterpart game, here we observe three equilibria: one pure at strategy D, one pure at strategy F, and one unstable mixed equilibrium at the 1-face formed by strategies D and F. All these equilibria are Nash in the respective counterpart games[3]. By applying the theory of section 3.3 we now know that we only maintain the combination $((1, 0, 0), (1, 0, 0))$ as a pure Nash equilibrium of the asymmetric PSRO empirical game, since these strategies have the same support as a Nash equilibrium in the counterpart games. The other equilibria in the second counterpart game can be discarded as candidates for Nash equilibria in the PSRO empirical game since they do not appear as equilibria for player 1.
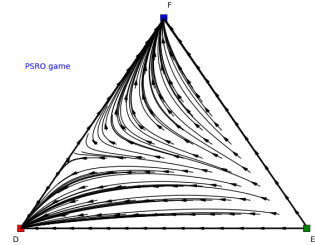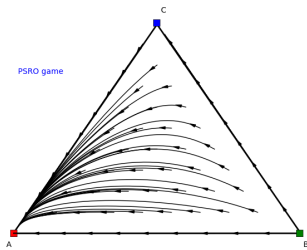

Figure 8: Trajectory plot of the first CP game.

Finally, each joint action of the game was estimated with 100 samples. As the outcome of the game is bounded in the interval $[-13, 13]$ we can only guarantee that the Nash equilibrium of the meta game we studied is a 2ϵ-Nash equilibrium of the unknown underlying game. It turns out that with $n = 100$ and $\epsilon = 0.05$, the confidence can only be guaranteed to be above 10^{-8}. To guarantee a confidence of at least 0.95 for the same value of $\epsilon = 0.05$, we would need at least $n = 886 \times 10^3$ samples.

[3]Banach solver (http://banach.lse.ac.uk/) is used to check Nash equilibria [1]

Figure 9: Trajectory plot of the 2nd CP game.

6 CONCLUSION

In this paper we have generalised the heuristic payoff table method introduced by Walsh et al. [14] to two-population asymmetric games. We call such games *meta-games* as they consider complex strategies instead of atomic actions as found in normal-form games. As such they are well suited to investigate real-world multi-agent interactions, as they summarize behaviour in terms of high-level strategies rather than primitive actions. We have shown that a Nash equilibrium of the meta-game is a 2ϵ Nash equilibrium of the true underlying game, providing theoretical bounds on how much data samples are required to build a reliable meta payoff table. As such our method allows for an equilibrium analysis with a certain confidence that this game is a good approximation of the underlying real game. Finally, we have carried out an empirical illustration of this method in three complex domains, i.e., *AlphaGo*, Colonel Blotto and PSRO, showing the feasibility and strengths of the approach.

ACKNOWLEDGMENTS

REFERENCES

[1] D. Avis, G. Rosenberg, R. Savani, and B. von Stengel. 2010. Enumeration of Nash Equilibria for Two-Player Games. *Economic Theory* 42 (2010), 9–37.

[2] E. Borel. 1953. La théorie du jeu les équations intégrales à noyau symétrique. Comptes Rendus de l'Académie 173, 1304–1308 (1921); English translation by Savage, L.: The theory of play and integral equations with skew symmetric kernels. *Econometrica* 21 (1953), 97–100.

[3] Rémi Coulom. 2008. Whole-History Rating: A Bayesian Rating System for Players of Time-Varying Strength. In *Computers and Games, 6th International Conference, CG 2008, Beijing, China, September 29 - October 1, 2008. Proceedings*. 113–124.

[4] Michael Kaisers, Karl Tuyls, Frank Thuijsman, and Simon Parsons. 2008. Auction Analysis by Normal Form Game Approximation. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, Sydney, NSW, Australia, December 9-12, 2008*. 447–450.

[5] Pushmeet Kohli, Michael Kearns, Yoram Bachrach, Ralf Herbrich, David Stillwell, and Thore Graepel. 2012. Colonel Blotto on Facebook: the effect of social relations on strategic interaction. In *Web Science 2012, WebSci '12, Evanston, IL, USA - June 22 - 24, 2012*. 141–150.

[6] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Perolat, David Silver, and Thore Graepel. 2017. A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). 4190–4203.

[7] Steve Phelps, Kai Cai, Peter McBurney, Jinzhong Niu, Simon Parsons, and Elizabeth Sklar. 2007. Auctions, Evolution, and Multi-agent Learning. In *Adaptive Agents and Multi-Agent Systems III. Adaptation and Multi-Agent Learning, 5th, 6th, and 7th European Symposium, ALAMAS 2005-2007 on Adaptive and Learning Agents and Multi-Agent Systems, Revised Selected Papers*. 188–210.

[8] Steve Phelps, Simon Parsons, and Peter McBurney. 2004. An Evolutionary Game-Theoretic Comparison of Two Double-Auction Market Designs. In *Agent-Mediated Electronic Commerce VI, Theories for and Engineering of Distributed Mechanisms and Systems, AAMAS 2004 Workshop, AMEC 2004, New York, NY, USA, July 19, 2004, Revised Selected Papers*. 101–114.

[9] Marc Ponsen, Karl Tuyls, Michael Kaisers, and Jan Ramon. 2009. An evolutionary game-theoretic analysis of poker strategies. *Entertainment Computing* 1, 1 (2009), 39–45.

[10] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 7587 (2016), 484–489.

[11] Finnegan Southey, Michael Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and Chris Rayner. 2005. Bayes' bluff: Opponent modelling in poker. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence (UAI-05)*.

[12] Karl Tuyls and Simon Parsons. 2007. What evolutionary game theory tells us about multiagent learning. *Artif. Intell.* 171, 7 (2007), 406–416.

[13] Karl Tuyls, Julien Perolat, Marc Lanctot, Rahul Savani, Joel Leibo, Toby Ord, Thore Graepel, and Shane Legg. 2018. Symmetric Decomposition of Asymmetric Games. *Scientific Reports* 8, 1 (2018), 1015.

[14] W. E. Walsh, R. Das, G. Tesauro, and J.O. Kephart. 2002. Analyzing complex strategic interactions in multi-agent games. In *AAAI-02 Workshop on Game Theoretic and Decision Theoretic Agents, 2002*.

[15] W. E. Walsh, D. C. Parkes, and R. Das. 2003. Choosing samples to compute heuristic-strategy Nash equilibrium. In *Proceedings of the Fifth Workshop on Agent-Mediated Electronic Commerce*.

[16] Michael P. Wellman. 2006. Methods for Empirical Game-Theoretic Analysis. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA*. 1552–1556.