Game Theory Project

Mateusz Pyla

Github Repo

Incremental Learning, Game Theory, and Applications 2022

Master IASD, Université Paris Dauphine-PSL

.

ABSTRACT

We consider a typical non-collaborative game theory setup; n (ε-rational) players want to learn strategies for a repeated (finite) stage game. The players use the standard statistical tools in order to validate their beliefs. We assume no prior knowledge, and we do not know the opponents' behaviour nor playoffs.
Dean Foster and H. Peyton Young design learning strategies that come close to subgame equilibrium play. They prove the convergence in probability and show that under certain conditions these hypothesis testing strategies are ε-best responses to their beliefs.[1]

Keywords: Game Theory, Learning, Hypothesis test, Nash Equilibrium, Repeated game, Robustness

1. INTRODUCTION

1.1 Notation

The paper itself is demanding so in order to enhance the clarity of its message we decided to build a small glossary.

Glossary

G (finite) Stage game.

G^∞ Infinitely repeated stage game.

t index for time.

T time limit for finite game.

n Number of players.

i Player i.

ε Given a non-negative parameter, please refer to ε-good predictors..

σ Degree of smoothing - we talk about σ_i-optimal response.

τ Test Tolerance.

s Amount of data collected.

λ Degree of conservation.

M Memory (we say a model has at most memory M. Hypothesis has memory M if it attributes memory M strategies and responses of memory M.).

X Action space, i.e. $X = \prod X_i$.

Δ The set of probability measures over the set of the actions..

Δ^M The set of probability measures over the set of the actions accounting the memory M.

ω^t Vector of ω_i^t, the actions taken in period t.

$\bar{\omega}^t$ Sequence of actions taken in periods from 1 to t.

$\bar{\omega}$ Sequence of actions in the whole history.

$\Omega(\bar{\omega}^t)$ Set of all continuations of the initial history $\bar{\omega}^t$.

$\overrightarrow{a}$ Vector of responses; n-tuple of strategies.

$\mathcal{A}_i$ Response space.

$f_i(\cdot|\bar{\omega}^t)$ Density on the response space.

$\mathcal{A}$ Space of strategies.

$\overrightarrow{u}$ Vector of utility functions, $u_i : X \to \mathcal{R}$.

u_i Utility function for player i scaled between 0 and 1..

u_i^t Utility (function) from time period t for player i.

$u(\cdot\cdot)$ applied to (x_i, ϕ_i^t) - Expected utility for i, discounted to time t, from playing x_i in each period from t on, give the model ϕ_i.

ϕ_i Model. ϕ_i^* is a model fixed point with $\overrightarrow{\phi} = P(A^{\overrightarrow{\sigma}}(\overrightarrow{\phi}))$..

Φ_i Space of all models having memory at most M, $\Phi_i = \prod_{j \neq i} \Delta_j^M$.

Φ Space of vectors of models, $\Phi = \prod_i \Phi_i$.

P mapping all players current responses to the correct models, function taking a vector of responses and returning the vector of correct models $\prod_{j \neq i}(a_j)$.

A_i^σ Continuous mapping from i's model space to i's response space - smoothed best response function.

$A^{\overrightarrow{\sigma}}$ Mapping all players believe to their responses, function taking a vector of models (at time t) and returning the vector of distributions generating the process $A_i^{\sigma_i}(\phi_i)$.

H_0 Null hypothesis.

H_1 Alternative hypothesis.

ν_0 Null distribution.

ν True distribution.

α_{i,s_i} Probability of making type-II error. Similarly, for β_{i,s_i}.

$\beta_{i,s_i,\tau}$ Least upper bound on making a type-II error when the true distribution ν is more than τ away from the null ν_0. Similarly, for $\alpha_{i,s_i,\tau}$ within τ-ball..

R Rejection set.

ρ Discount factor.

v_{a_i,ϕ_i} Probability measure over infinite histories induced by the reposnse a_i and the model ϕ_i.

$p_i^t(x_{-i}|\cdot)$ Conditional probability distribution specified by model ϕ_i for every initial history $\bar{\omega}^{t-1}$.

$h_i(\sigma_i))$ minimum probability that i plays.

1.2 Nature of the game

Let us suppose that n players want to play a repetitive finite-action game, which we denote as stage game G. In general we do not want to bound ourselves by the finite horizon T hence we consider G^{∞}. Each player i wants to maximize their own profit hence we deal with non-cooperative game theory. Each game is in normal form meaning that the actions are performed simultaneously. For each period t, each player chooses their own action ω^t, observe the opponents' actions and record their own payoff.

Although we assumed that the game is of perfect information, this is not a strong constraint. On contrary, we **make no assumption** on prior (common) knowledge, opponents' strategies nor payoffs (including the distribution). Hence we aim at finding robust (subsection 1.4) algorithm for our problem. We call player to be ε-good predictor if they do the predictions such that the mean square error is almost surely bounded by ε as we go with the time to infinity, i.e.

$$\lim_{T\to\infty} \sup \frac{1}{T} \sum_{t=1}^{T} (\phi_i^t - B_i(\overrightarrow{a}^t))^2 \leq \varepsilon$$

Assuming the environment as stated above, our ultimate goal is to model the strategic interactions between a set of players. In particular, we will focus on learning how to make predictions for almost rational players.

1.3 Learning

The question of learning within game theory is not precisely defined. This may be caused by the fact that we have numerous of settings, and depending on the requirements one may find various learning procedures compelling. In the simplest terms, for our case we are interested in finding an algorithm that learns how to map the history of previous games to predictions for the next games. One may point out that this is a very challenging task as the world to be learnt constantly changes under our actions. We experience feedback loop from the model and the objects to be modeled.

1.3.1 Motivation

The framework of multiple agents interacting with each other has found many applications, including economics, business, biology. With the advance of widely understood artificial intelligence, the game theory also plays a role in designing the framework, especially due to its compatibility with reinforcement learning[2] and generative methods.[3][4][5]

1.4 Previous approaches

If players are rational, good predictors, and learn deterministically, there are many games for which neither beliefs nor actions converge to a Nash equilibrium (An impossibility theorem under deterministic learning[6]) The authors of the papers are well-known experts in the field. It is highly likely that they were familiar with all the state of the art methods. They highlighted the following work on learning the strategies:

1. Trivial algorithm (simple search the space of mixed strategies) dynamically converging to Nash equilibrium of the stage game is not adequate from the global picture's perspective. There are two main reasons:
   - The process is not decentralized meaning that the learning is not performed for the individual
   - Finding such Nash equilibrium would require the relaxation of no prior assumption of the payoff knowledge.
2. For some games, using fictitious play leads to Nash equilibrium.[7][8]
   - For many situations the convergence is not useful for predictions.
   - There are still some examples of games in which the algorithm diverges.[9]
3. Knowing the appropriate Bayesian Nash equilibrium prior makes all limit points of the posterior to be also Bayesian Nash equilibrium.[10] Authors of our paper claim that the problem of solving the equilibrium is not solved, but rather rephrased.

4. Assigning positive probabilities in players' prior belief to all events with actual positive probability leads to guarantee of reaching ε-equilibrium ε close enough with probability one.[11] The authors of the main paper claim there is no robust procedure to determine such priors on other players' strategies which makes best-response strategies continuous. Instead, they propose another constraint stated in subsection 1.2.

5. Predicting the previous, already realized events, i.e. conditional regret minimization. Given the history of games one can question whether playing one action would be superior over another. There are numerous methods although they only guarantee convergence to the set of correlated equilibria.[12][13]

1.5 Hypothesis testing

For the repeated games, player's hypothesis consists of the strategies of their opponents. If we denote Δ to be the set of probability distributions on X, finite action space. Thus we can consider each strategy as a point in the Euclidean space of finite dimension. We will impose the simplification that the strategies will rely on last M periods, meaning that the strategy of memory M will be a point in Δ^M. Similarly, we will exhibit the hypotheses of memory M which live in the strategies space $\mathcal{A}$.

The role of hypothesis test within our framework is following:

1. Players engaged in the repeated game do the predictions for the next stage game.
2. They observe the realization, i.e. the actions performed by all players.
3. We introduce the rule: at the end of each period t, agents not currently carrying out a hypothesis test, initialises a new with probability $1/s_i$. After successful launch, after s_i steps they test the empirical distribution under a null hypothesis H_0.
    - If it is unlikely that the observations could have happened assuming the null hypothesis is true then they reject it and propose a new one H_1 (from the uniform distribution).
    - Otherwise the null hypothesis cannot be rejected; they stay with it.

We will need 2 properties of the hypothesis tester:

- Flexible (related to diffusion defined below) if for all $\bar{\omega}^t$, $f_i(\cdot|\bar{\omega}^t)$ is bounded away from zero.
- Conservative (with parameter λ) in the sense with probability at least $1-\lambda_i$ we propose new hypothesis λ_i-close to the old hypothesis.

Null hypothesis induces null distribution $\nu_0 = (A_i^{\sigma_i}(\phi_i^{t0}), \phi_i^{t0})$ We denote true distribution (to be learnt) as ν.

In general, our statisticians-players will commit two types or errors

I Rejecting the H_0 when it is correct. We will study the probability of making it (denoting by α) at each time; for instance $\alpha_{i,s_i} = \sup_{\omega,t,\nu_0} \alpha_{i,s_i}(\nu_i^{t_0}, \bar{\omega}^{t-1})$ we read as least upper bound on making type-I error given the history and our current model and response.

II Accepting the H_0 when it is not correct. We will further put the constraints related to the tolerance: $\beta_{i,s_i,\tau} = \sup_{\omega,t,\nu_0} \sup_{\nu:|\nu-\nu_0|>\tau}(\nu, \nu_0, \bar{\omega}^{t-1})$ would be read as the probability of making type-II error given τ.

1.5.1 Smoothed best response functions

We denote $A_i^{\sigma_i}$ to be σ_i-smoothed best response if:

- We say that the response is almost rational if its expected payoff is within σ_i of the payoff from an optimal strategy.
- We say that the response is diffuse if we assign positive probability to each possible action in every time period.
- We also impose A_i^{σ} to be continuous function from Φ_i to $\mathcal{A}_i$ and continuous in payoffs u_i.

Such family $\{A_i^{\sigma_i}\}_{\sigma_i}$ we will name family of smoothed best response functions.

2. COMPREHENSION

2.1 Theorems

We firstly present our ultimate goal:

> THEOREM 1 (FOSTER & YOUNG, 1998-2003). *Let G be a finite, normal-form, n-person game and let $\varepsilon \geq 0$. If the players are almost rational, use sufficiently powerful hypothesis tests with comparable amount of data, and are flexible in their adoption of new hypothesis, then at least $1-\varepsilon$ of the time:*
>
> *I Their repeated-game strategies are ε-close to subgame perfect equilibrium,*
>
> *II All players are ε-good predictors.*

In order to reach the stated results we will use the intermediate results that we will use as a milestones to prove the main theorem.

LEMMA 2.1 (POWERFUL FAMILY OF TESTS). *We define the family of hypothesis tests to be powerful if there are positive tolerance functions k_i and r_i such that for every tolerance $\tau > 0$ there is at least one test in the family such that*

$$\alpha_{i,s_i} \leq k_i(\tau)e^{-r_i(\tau)s_i} \qquad \textit{(Upper bound on type-I error)}$$
$$\beta i, s_i, \tau \leq k_i(\tau)e^{-r_i(\tau)s_i} \qquad \textit{(Upper bound on type-II error)}$$

LEMMA 2.2 (UPPER BOUND ON TYPE-I ERROR). *Assume that null hypothesis imposes the conditional probability of observing any vector of actions at any time is at least ξ. Then:*

$$\forall \tau > 0 \quad \alpha_{i,s_i,\tau} \leq (1+\frac{\tau}{\xi})^{s_i}\alpha_{i,s_i}$$

COROLLARY 2.2.1 (UPPER BOUND OF TYPE-I AND TYPE-II ERRORS).

$$\beta_{i,s_i,\tau} \leq k_i(\tau)e^{-r_i(\tau)s_i/2}$$

There exists $c_i(\tau) \leq \tau$ such that

$$\alpha_{i,s_i,c_i(\tau)} \leq k_i(\tau)e^{-r_i(\tau)s_i/2}$$

LEMMA 2.3 (FAIRLY GOOD MODEL VECTOR). *We call model vector $\overrightarrow{\phi}$ to be fairly good if it is good for all responsive players. We denote it to be good if it is good for all players, meaning $|\phi_i - P_i(A^{\overrightarrow{\sigma}}(\overrightarrow{\phi}))| \leq \tau$ holds for player i (otherwise it is bad for i). Finally, we call it bad if it is not fairly good.*

The model vector $\overrightarrow{\phi}^t$ is fairly good at least $1-\varepsilon$ of the time.

THEOREM 2.4 (GUARANTEES FOR RESPONSES). *We assume that for the conditions described in the main theorem (1), the participants use M memory hypotheses for which they employ suitable hypothesis tests with comparable amount of data.*

Moreover we assume that the engaged players have σ_i-smoothed best response functions.

Let $\varepsilon > 0$. There exist functions $\sigma(\varepsilon)$, $\tau(\varepsilon,\sigma)$ and $s(\varepsilon,\sigma,\tau)$ bounding the corresponding parameters.

Then at least $1-\varepsilon$ of the time t

I $\lvert a^t - A^{\overrightarrow{\sigma}}(P(\overrightarrow{a}^t))\rvert \leq \varepsilon/2$	*the responses are close to being a fixed point*
II $\lvert U_i^t(a_i^t, P_i(\overrightarrow{a}^t)) - max_{a_i'} U_i^t(a', P_i(\overrightarrow{a}^t))\rvert \leq \varepsilon$ *for all players i*	*the responses are ε-optimal*
III $\lvert \phi_i^t - P_i(A^{\sigma_i}(\overrightarrow{\phi}^t))\rvert \leq \varepsilon$ *for all players i*	*the models are within ε of being correct*

COROLLARY 2.4.1 (GUARANTEE FOR REPEATED GAME). *The second statement of the Theorem 2.4 is equivalent to claiming that the repeated-game strategies are ε-close to the equilibria of the repeated game $G^\infty(\overrightarrow{u}, X)$ for at least $1-\varepsilon$ of the time.*

The third statement imposes that all players are ε-good predictors.

2.1.1 Convergence

In the Theorem 2.4 we obtain the strategies ε-close the equilibrium providing the adequate choice of the parameters σ τ and s. However in most of the optimization algorithms, we adjust the parameters accordingly to the number of iterations.

Assume now that after every period, with probability (ε) we halve the relevant parameters. Then we decrease the probability of shrinking the parameters further, i.e. $p(\varepsilon/2) < p(\varepsilon)$ etc.

After k successful iterations in which we shrink the parameters, we are left with $\varepsilon_k = \varepsilon/2^k$ causing the proximity of ε_k to the set of subgame perfect equilibria; hence arbitrary close.

This learning process guarantees convergence in probability.

2.1.2 Intuition

Although the statements above are aligned in the correct order, it is essential to gain the intuition behind the formulas. By putting up the claims listed in subsection 2.1 together, we build up the proof to finally reach the main theorem results.

We have two big ingredients: fixed point model and hypothesis test tools. We aim at designing hypothesis testing rule that solves the associated fix-point problem.

Firstly, fixed points are equilibria. We are certain about their existences since we have two continuous maps (P, $A^{\vec{\sigma}}$) between compact convex spaces to each other. Those two spaces are namely, the space of vector of models $\Phi \equiv \prod_i \Phi_i$, and the space of tuples of set of responses for players, i.e. strategies $\mathcal{A} \equiv \prod_i \mathcal{A}_i$. Theorem 2.3 yields that with high probability, the players will reject the hypothesis when at the bad model vectors. Once we reach the fixed point, the probability of escaping it is negligible. Both goes to zeros exponentially fast as s_i increases.

There are four parameters controlling the system.

- the degree of smoothing σ_i,
- the tolerance τ_i,
- the amount of data collected s_i,
- the degree of conservatism λ_i.

Let ε be a positive number. Remark 2.4.1 implies that all the players are ε-good predictors and their strategies are ε-close to the equilibrium at least $1 - \varepsilon$ time providing σ_i is sufficiently small, τ_i is sufficiently small and s_i is sufficiently big. Furthermore, if we choose the suitable λ_i then we can guarantee that we spend all the time near the equilibrium.

2.1.3 Individual perspective

The process is decentralized so we look at the algorithm from each player's perspective. If the current model vector ϕ_i is bad (Theorem 2.3) for first player then likely he will reject the hypothesis and update their beliefs. Then the other players will reject their own model hypothesises and move towards to a equilibrium. And finally the first player will correct it for themselves on the following test.

2.1.4 Beliefs

Towards the end of the paper, the authors pointed out that the proposed system is rather simplification of solving the strategies for repeated games. Everything is stated in a discrete manner: using models and responses instead of believes which we define as the conditional probabilities of the opponents' future actions and their own future model changes.

A player's model only change when they are reject it at the end of test phase, and only then the conditional probability distribution changes. If we are conservative enough (governed by λ_i) we can guarantee to bound the difference between the expected payoffs discounted by ρ of the belief and the current model $\varepsilon/2$. If we take response $\varepsilon/2$ close to the expected discount of the model we have that all the time, for every player's response is ε-close to their belief.

2.2 Proofs

For the full versions of proofs we refer to the original paper. Here, we aim at mathematically convincing the reader about the results but do not intent to present all rigorous justifications.

2.2.1 Powerful Family of Tests

Proof. we take a simple test with the property $\alpha \doteq \alpha_{i,s_0} < 0.5$ and $\beta \doteq \beta_{i,s_0,\tau} < 0.5$. We do the construction of such family of tests. The idea is to divide our accumulated data of size s into disjoint buckets of size s_0 ignoring the last elements if they do not fit or make the number of subset even ($s' \doteq s - \delta$). We reject the null hypothesis iff it is rejected on the majority of buckets. Let $\kappa \doteq \lfloor \frac{s'}{s_0} \rfloor$ The probability of type-I is:

$$\alpha_{i,s} \leq \sum_{j=\lceil \kappa/2 \rceil}^{\kappa} \binom{\kappa}{j} \alpha^j (1-\alpha)^{\kappa-j} \leq \alpha \sum_{j=\lceil \kappa/2 \rceil}^{\kappa} \binom{\kappa}{j} (\alpha(1-\alpha))^{\lfloor \kappa/2 \rfloor} \leq 2\alpha(4\alpha(1-\alpha))^{\lfloor \kappa/2 \rfloor} \sum_{j=\lceil \kappa/2 \rceil}^{\kappa} \binom{\kappa}{j} 0.5^{\kappa} \leq \alpha(4\alpha(1-\alpha))^{\lceil \kappa/2 \rceil}$$

The inequalities are deducted due to $\kappa/2$ approximation and the fact that $\alpha < (1-\alpha)$ and $\sum_{j=0}^{\kappa} \binom{\kappa}{j} 0.5^{\kappa} = 1$.

Because $\alpha(1-\alpha) \leq 1/4$ we have $\alpha_{i,s} \leq \alpha[4\alpha(1-\alpha)]^{-2} \cdot e^{-(\ln(4\alpha(1-\alpha)/2 \cdot s_0))s}$

Similarly, one can repeat the same thinking for $\beta_{i,s,\tau}$.

By taking $k = \max\{\alpha[4\alpha(1-\alpha)]^{-2}, \beta[4\beta(1-\beta)]^{-2}\}$ and $r = \min\{\ln[4\alpha(1-\alpha)]/2s_0, \ln[4\beta(1-\beta)]/2s_0\}$ we obtain both required inequalities. □

2.2.2 Upper Bound of Type-I (and Type-II) Errors

Proof. Assume that the conditioned data $\bar{\omega}^{s_i}$ is generated by test in the first s_i periods. Our null hypothesis induces $p(\bar{\omega}^{s_i})$ and let $q(\bar{\omega}^{s_i})$ be any other distribution within τ of p hence $\forall t \leq s_i \forall \omega^t |p(\omega^t|\bar{\omega}^{t-1}) - q(\omega^t|\bar{\omega}^{t-1})| \leq \tau$.

Assuming, $\forall t \leq s_i \quad p(\omega^t|\bar{\omega}^{t-1}) \geq \xi$ we have $q(\omega^t|\bar{\omega}^{t-1}) \leq p(\omega^t|\bar{\omega}^{t-1}) + \tau \leq p(\omega^t|\bar{\omega}^{t-1})$ and $\frac{q(\omega^t|\bar{\omega}^{t-1})}{p(\omega^t|\bar{\omega}^{t-1})} \leq (1 + \frac{\tau}{\xi})$

Let R be the rejection set, i.e. $\{\bar{\omega}^{s_i}\}$ causing H_0 to be rejected.

$q(R) = \sum_{\bar{\omega}^{t-1} \in R} \prod_t q(\omega^t|\bar{\omega}^{t-1}) \leq \sum_{\bar{\omega}^{t-1} \in R} \prod_t p(\omega^t|\bar{\omega}^{t-1}) \cdot (1 + \frac{\tau}{\xi}) = (1 + \frac{\tau}{\xi})^{s_i} p(R)$.

By taking $p(R) \leq \alpha_{i,s_i}$ we reach the desired inequality. □

2.2.3 Corollary on the errors

Proof. Let $h_i(\sigma_i))$ be the minimum probability that i plays every action in all times. We assumed the best response function to be diffuse in subsubsection 1.5.1 hence this value is positive.
Therefore $c_i(\tau) = [\prod_j h_j(\sigma_j)] r_i(\tau)/2$ does the job and we take $c(\tau) = \min_i c_i(\tau)$ □

2.2.4 Fairly Good Model

Proof. Building up from previous result, there exists $c(\tau) \leq \tau$ making the exponentially small probability of type-I error providing ν_o we are within $c(\tau)$ of the truth ν and similarly with probability exponentially close to 1 of rejecting when we are away from at least τ from ν.

Now, let us suppose $|\phi_i - \phi_i^*| < \gamma$ and $\gamma < c(\tau)/2$.

Since $A_i^{\sigma_i}$ is uniformly continuous then $|\phi_i - \phi_i^{'}| \leq \gamma$ implies $|A_i^{\sigma_i}(\phi_i) - A_i^{\sigma_i}(\phi_i^{'})| \leq c(\tau)/(2n)$

Moreover $|P_i(A_i^{\sigma_i}(\phi_i)) - P_i(A_i^{\sigma_i}(\phi_i^*))| \leq \frac{(n-1)c(\tau)}{2n}$

Combining we obtain that $|\phi_i - \phi_i^*| \leq \gamma$ implies $|\phi_i - P_i(A_i^{\sigma_i}(\overrightarrow{\phi}))| \leq c(\tau)$.

Let us denote $s_* \equiv \min_i s_i$ and $s^* = max_i s_i$.

We define state to be great when for all players i, i's model is within γ of ϕ_i^* and no player is currently in a test phase. Assume that we have bad vector $\overrightarrow{\phi}^t$. The following steps may end up with a great state.

1. Without loss of generality, let the first player to have bad model and they starts a new test phase once all other tests are carried out. No other player perform a test during this phase. After first player rejects their hypothesis and obtain new model within γ to $\phi_1^{'}$ such that $A_1^{\sigma_1}(\phi_1^{'}) = a^{'}$ is τ away from $(\phi_j^*)_i$
2. One by one, other players conduct the tests respecting the time order and receive a new model γ-close to ϕ^* of their part.
3. First player again starts a test phase and reject the their test moving to the model within γ of ϕ_1^*.
4. No player starts a new test until the time $t + (n+2)s^*$

Let us compute the probability of the events described above.

There are $n+1$ test phases ended up with rejection, and the timings have to align well. Hence this is at least $[(1/s^*)(1-1/s^*)^{(n-1)s^*}]^{n+1} \geq [(1/s^*)(1/4)^{(n-1)s^*}]^{n+1}$. No other test can occur at this time which happens with probability $(1-1/1s_*)^{n(n+2)s^*} \geq (1/4)^{n(n+2)s^*/s_*}$.

Since the players uses comparable amount of data (subsubsection 2.2.5) $s^*/s_* \leq s_*^{\xi}$ for $\xi \in (0,1)$.

Let $\eta \equiv \alpha s_*^{(n+1)\xi} e^{-\beta s_*^{\xi}}$ be lower bound of the probability for the bad model vector $\overrightarrow{\phi}^t$ being in a great state at $t + (n+2)s^*$.

Recall that the probability that i reject a test is $\alpha_{i,s_i,\tau_0} \leq k_0 e^{-r_0 s_i} \leq k_0 e^{-r_0 s_*}$. During periods $t+1, ..., t+T$ there are at most nT/s_* possible tests hence the probability is bounded by $(nT/s_*)k_0 e^{r_0 s_*} < Te^{-4rs_*} = e^{-rs_*}$ by letting $T = e^{3rs_*}$

Now define E to be the event that in εT of the states in a period of length T are bad. We can divide that in two subevents: E', the ones were all the bad states are from time 0 till k, and the others (E"). Hence $P(E) = P(E') + P(E'')$.

For E' the probability upper bound is $(1-\eta)^{k-1} < e^{-\eta(k-1)}$ using our earlier bounds on variable we can bound that variable from above with e^{-rs_*}, which we can make arbitrary small when s_* is large enough. Hence for instance $P(E') < \varepsilon/2$.

If E" happens the process does not stay in the great states for T periods thus $P(E'') \leq e^{-rs_*}$, and again if we take s_* to be large enough we can make this sufficient small, i.e. $P(E'') < \varepsilon/2$ yielding $P(E) < \varepsilon$. □

2.2.5 Guarantees for Responses

Proof. Recall that we say that the players are using comparable amount of data if for any two players, i.e. $\forall \gamma \in (1,2) s_i \leq s_j^{\gamma}$.

We assume that all players are responsive, for non-responsive case we refer to section 5. We need to prove all three points, one by one finishing with "OK!" message.

Let us take the fixed point $\overrightarrow{a^*}$ of the composition of mapping $A^{\sigma} \circ P$

Let $\delta < \varepsilon/2n$. We take $\tau \leq \frac{\delta}{2(n+1)}$ and a model vector $\overrightarrow{\phi}^t$.

Since $n > 1$ we have $\tau < \delta$, $\delta < \varepsilon$ and finally $\tau < \varepsilon$.

Due to the previous result, we have that $\overrightarrow{\phi}^t$ is fairly good at least $1-\varepsilon$ of the time. Thus $|\phi_i - P_i(A^{\overrightarrow{\sigma}}(\overrightarrow{\phi}))| \leq \tau$.

Combining last two inequalities we reach $|\phi_i - P_i(A^{\overrightarrow{\sigma}}(\overrightarrow{\phi}))| \leq \varepsilon$. OK!

We repeat the same trick: $|\phi_i - P_i(A^{\overrightarrow{\sigma}}(\overrightarrow{\phi}))| \leq \delta$. Due to continuity having $|\phi_i - \phi_i^{'}| \leq \delta$ implies that $|A_i^{\sigma_i}(\phi_i) - A_i^{\sigma_i}(\phi_i^{'})| \leq \varepsilon/2n$ which means $|a_i - A_i^{\sigma}(P_i(a_i))| \leq \varepsilon/2n$ and finally $|\overrightarrow{a} - A^{\sigma}(P(a))| \leq /2$. OK!

Let a_i^* be a best response to $P_i(\overrightarrow{a}^t)$. We use results of previous line: for every player i we have $|a_i^t - A_i^{\sigma}(P(\overrightarrow{a}))| \leq \varepsilon/2n$ and therefore the payoff difference between them is bounded by $\varepsilon/2$ because the utility u_i is suitably scaled. Therefore for u_i^t we have $|U_i^t(a_i^t, P_I(\overrightarrow{a}^t)) - U_i^t(a_i^*, P_i(\overrightarrow{a}^t)| \leq \varepsilon/2 + \varepsilon/2 = \varepsilon$. OK! □

2.2.6 Main Theorem

Proof. Let $\varepsilon > 0$. Denote $\varepsilon_2 = \varepsilon/2$. Due to almost rational play, there exists α such that all players carry about the prediction more than α. Let $\varepsilon^{'} = \min\{\varepsilon_2, \alpha\}$. Due to previous result, applying the last point of remark 2.4.1 we have that all players are $\varepsilon^{'}$-good predictor and since $\varepsilon^{'} \leq \varepsilon$ also ε-good predictors. OK!

For the first statement, we take $\overrightarrow{a}^t$. We denote $A^{\overrightarrow{0}}(\overrightarrow{\phi})$ to be the best response containing $\overrightarrow{a}^t$ iff $\forall_i a_i$ is a best response to ϕ_i. Let $\mathcal{S}$ be the set of fixed points of $A^{\overrightarrow{0}} \circ P$. Every $\overrightarrow{a} \in \mathcal{S}$ generates a subgame perfect equilibrium of the repeated game.

By previous result we have $|a_i^t - A_i^{\sigma_i}(P(\overrightarrow{a}^t))| \leq \varepsilon_2$ for all players at least $1 - \varepsilon_2$ of the time.

As $\sigma_i \to 0, A^{\overrightarrow{\sigma}} \to A^{\overrightarrow{0}}$. By choosing σ_i appropriately small we can guarantee that $\overrightarrow{a}^t$ is within $2n\varepsilon$ of $\mathcal{S}$ at least $1 - \varepsilon_2$ of the time.

We finish the proof with claiming that for $1/\varepsilon$ rounds following t no player will change their model providing $s_* > 2n/\varepsilon_2^2$.

We consider time periods between t and $t + T$. By choosing T suitably large, the number of tests is upper-bounded by $T\varepsilon^2$ since $n + T/s_* \leq n + T\varepsilon^2/2$. Thus at least $1 - \varepsilon_2$ of the time, no player changes their strategies. Let us notice that the constraint is rather demanding in terms of time, however we do care in fact about the infinite unfold.

To sum up, the strategies will be close to $\mathcal{S}$ at least $1 - \varepsilon_2$ of the time and for at least $1 - \varepsilon_2$ the time will be ε_2-steady. Hence for at least $1 - 2\varepsilon_2 = 1 - \varepsilon$ of the time the players will be close to being a subgame perfect equilibrium. □

3. IMPLEMENTATION

We used the Python's library called Gambit.[14] It offers rather easy interface for manipulation of games.

Unscrupulous diner's dilemma is a n-person prisoner's dilemma variant. All players are choosing their meal at the restaurant having agreed on the bill split. For all players, difference of cost between the meals is bigger than the difference of happiness caused by improving the position from menu.

The paper is pretty theoretical and we found it difficult to quickly put the results to the practice. For now, we struggle with the hypothesis representation as we did some progress in the problem and best response modelling.

4. CONCLUSION

Once again, the biggest achievement of the paper is to propose a robust way of learning the players' behaviour strategies. For each player we only assume that the player opponents' actions are public. The obtained results are satisfactory and they are based on well-grounded apparatus mathematical. The only limitation is that the behaviors do not necessarily converge to an ε-equilibrium, or even to the set of ε-equilibria; they are close to equilibrium a large fraction of the time.

4.1 Learning Outcomes

The project was a fantastic opportunity to learn about the challenges that researchers in the field of game theory are facing. The paper is relatively old, nevertheless it has not yet had a natural successor to my best knowledge. This reflects the scale of the difficulty of the problem.

The paper was admittedly quiet tough from the mathematical perspective therefore we are satisfied with the further mathematical horizon expansion and with idea of proofs present in the field.

We had also a good occasion to get familiar with gambit, python's library for game theory.

Overall, the field of game theory serves an elegant way of modelling the interactions between the players and we are extremely delighted to learn the basics. The knowledge can be easily transferred to another problems and there are many links with other areas of science to be explored. Game theory itself is a relatively young subfield of mathematics leaving many doors still open for researchers. Aligning with my current interest, there are numerous options to combine the game theory framework and artificial intelligence. For instance, one can consider its role in the generative setting or reinforcement learning.

4.2 Further Work

Given the time constraints, we were not able to conduct all experiments we would like to do. Given more time in the future I would like to additionally:

- Finish the Unscrupulous diner's dilemma implementation.
- Increase the number of experiments and their complexity using gambit library.
- Work on more practical guarantees.
- Do study ablation for extending this approach to stochastic games.

REFERENCES

[1] Foster, D. P. and Young, H. P., "Learning, hypothesis testing, and nash equilibrium," *Games and Economic Behavior* **45**(1), 73–96 (2003).

[2] Nowé, A., Vrancx, P., and Hauwere, Y.-M. D., "Game theory and multi-agent reinforcement learning," in [*Reinforcement Learning*], 441–470, Springer (2012).

[3] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., "Generative adversarial nets," *Advances in neural information processing systems* **27** (2014).

[4] Zhou, Y., Kantarcioglu, M., and Xi, B., "A survey of game theoretic approach for adversarial machine learning," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **9**(3), e1259 (2019).

[5] Yasodharan, S. and Loiseau, P., "Nonzero-sum adversarial hypothesis testing games," *Advances in Neural Information Processing Systems* **32** (2019).

[6] Foster, D. P. and Young, H. P., "Learning with hazy beliefs," in [*Game Theory, Experience, Rationality*], 325–335, Springer (1998).

[7] Milgrom, P. and Roberts, J., "Adaptive and sophisticated learning in normal form games," *Games and economic Behavior* **3**(1), 82–100 (1991).

[8] Monderer, D. and Shapley, L. S., "Potential games," *Games and economic behavior* **14**(1), 124–143 (1996).

[9] Foster, D. P. and Young, H. P., "On the nonconvergence of fictitious play in coordination games," *Games and Economic Behavior* **25**(1), 79–96 (1998).

[10] Jordan, J. S., "Bayesian learning in repeated games," *Games and Economic Behavior* **9**(1), 8–20 (1995).

[11] Kalai, E. and Lehrer, E., "Rational learning leads to nash equilibrium," *Econometrica: Journal of the Econometric Society* , 1019–1045 (1993).

[12] Hart, S. and Mas-Colell, A., "A simple adaptive procedure leading to correlated equilibrium," *Econometrica* **68**(5), 1127–1150 (2000).

[13] Foster, D. P. and Vohra, R. V., "Calibrated learning and correlated equilibrium," *Games and Economic Behavior* **21**(1-2), 40 (1997).

[14] McKelvey, Richard D., M. A. M. and Turocy, T. L., "Gambit: Software tools for game theory." http://www.gambit-project.org.

5. APPENDIX

For the sake of clarity of the proof we decided to slightly change the enumerations of theorems and lemmas and modify the notation to be more intuitive.

5.1 For all players prediction matters

For the clarity of the statements we decided to assume that for every player i we do the prediction modelling that matters by at least ε. Otherwise, we say that the prediction does not matter for player i if their accuracy in prediction does not affect their payoff.

The original formulation of the last sentence of Theorem 1 would be "All players for whom prediction matters are ε-good predictors.", whereas for instance last point of Theorem 2.4 would be: "[...] for every player i for whom prediction matters by at least ε.

5.2 Responsive players

In subsubsection 2.2.5 we quickly replaced fairly good (good for all responsive players) with good (good for all) assuming for all players the predictions matter are hence they are all responsive players.

Let $d_i > 0$ be the diameter (controlled by $\overrightarrow{u}$) of the image $A_i^{\sigma_i}$ in $\mathcal{A}_i$, the space of i's responses. If d_i then player is responsive.

1. Case 1. All players are unresponsive. Then
2. Case 2. At least one player is responsive. Basically the proof follows as stated with some technicalities that some players are unresponsive meaning for instance $|a_i - A_i^{\sigma_i}(P_i(a_i))| \leq \delta \leq \varepsilon/2n$

Other Variables from the glossary: $\bar{\omega}$ $\Omega(\bar{\omega}^t)$ $u(\cdot\cdot)$ $p_i^t(x_{-i}|\cdot)$ v_{a_i,ϕ_i}