Limits and Limitations of No-Regret Learning in Games

Barnabé Monnot
Singapore University of Technology and Design
8, Somapah Road
Singapore 487372
monnot_barnabe@mymail.sutd.edu.sg

Georgios Piliouras
Singapore University of Technology and Design
8, Somapah Road
Singapore 487372
georgios.piliouras@gmail.com

ABSTRACT

We study the limit behavior and performance of no-regret dynamics in general game theoretic settings. We design protocols that achieve both good regret and equilibration guarantees in general games. In terms of arbitrary no-regret dynamics we establish a strong equivalence between them and coarse correlated equilibria.

We examine structured game settings where stronger properties can be established for no-regret dynamics and coarse correlated equilibria. In congestion games, as we decrease the size of agents, coarse correlated equilibria become closely concentrated around the unique equilibrium flow of the nonatomic game. Moreover, we compare best/worst case no-regret learning behavior to best/worst case Nash in small games. We study these ratios both analytically and experimentally. These ratios are small for 2×2 games, become unbounded for slightly larger games, and exhibit strong anti-correlation.

Keywords

Game Theory; No-Regret; Equilibria; Price of anarchy

1. INTRODUCTION

Understanding the outcome of self-interested adaptive play is a fundamental question of game theory. At the same time understanding systems that arise from coupling numerous intelligent agents together is central to numerous other disciplines such as distributed optimization, artificial intelligence and robotics.

To take the example of a multi-agent system, the problem of routing a large number of agents with repeated interactions offers agents the opportunity to learn from these interactions. In particular, we investigate how much can be gained from the process of learning, synthesized in a ratio which we call the *value of learning*. The reverse is also considered: what agents risk by following no-regret learning procedures, the *price of learning*.

The onset of such inquiries typically focuses on equilibria and their properties. Since games may have multiple Nash equilibria, two approaches have been developed: one focusing on worst case guarantees, known as price of anarchy [17], and one focusing on best case equilibria known as price of stability [1]. Defined as the ratio between the social cost at the worst Nash equilibrium and the optimum, price of anarchy captures the worst possible loss in efficiency due to equilibration. On the other hand, price of stability compares the social cost of the optimal Nash against the optimum.

Both approaches depend on the assumption that the agents converge to an equilibrium in the first case. This is a strong assumption and it is typically weakened to merely asking that the agents' adaptive behavior meets some performance benchmark, such as low regret [22]. An online optimization algorithm is said to exhibit vanishing regret when its time average performance is roughly at least as large as the best fixed action with hindsight.

Price of anarchy bounds for Nash equilibria for several classes of games are known to extend automatically to this larger class of learning behavior [20]. This implies that those worst case games are equally bad both for worst case Nash equilibria as well as for worst case learning behavior. Nevertheless, this does not mean that for individual games there cannot be significant gaps between the worst case performance of no-regret dynamics and Nash equilibria. The existence and size of such gaps for typical games are not well understood. Contrasting best case equilibria versus best case learning seems to be completely unexplored despite being a rather natural way to quantify the benefits of improving the design of our current learning mechanisms.

Our results. We study the limits and limitations of no-regret learning dynamics in games. We start by designing a protocol such that in isolation each such algorithm exhibits vanishing regret against any opponent while at the same time converging to Nash equilibria in self-play in any normal form game. This result establishes that no regret guarantees do not pose in principle a fundamental obstacle to system equilibration.

We establish a strong equivalence between the time average behavior of no-regret dynamics in games and coarse correlated equilibria (CCE) which is a relaxation to the notion of correlated equilibria. Specifically, given any infinite history of play in a game we can define for any time T a probability distribution over strategy outcomes that samples one of the T first outcomes uniformly at random. It is textbook knowledge that for all normal form games and no-regret dynamics the distance of this time average distribution from the set of CCE converges to zero as T grows [22]. We complement this result by establishing an inclusion in the reverse direction as well. Given any CCE we can construct a sequence of no-regret dynamics whose time average distribution converges pointwise to it as T grows. Hence in any normal form game, understanding best/worst case no-regret dynamics reduces to understanding best/worst case CCE.

In the second part of the paper we exploit this reduction to argue properties about best/worst case no regret learning dynamics in different classes of games. We provide a shorter, more intuitive argument that extends to the case of many but small agents by exploiting the connection to coarse correlated equilibria. Specifically, we show that for all atomic congestion games as we increase the number of agents/decrease the amount of flow they control, any coarse correlated equilibrium concentrates most of its probability mass on states where all but a tiny fraction of agents have a small incentive to deviate to another strategy. The uniqueness of the cost of equilibrium flow at the limit implies that for these games there is no distinction between good/bad Nash/learning behavior.

The picture gets completely reversed when we focus on games with few agents. We define Price of Learning (PoL) as the ratio between the worst case no-regret learning behavior and the worst case Nash whereas the Value of Learning (VoL) compares best case learning behavior to best case Nash. For the class of 2×2 (cost minimization) games PoL is at most two and this bound is tight, whereas VoL is at least $3/2$ and we conjecture that this bound is tight as well. Both PoL and VoL become unbounded for slightly larger games (e.g., 2×3).

We conclude the paper with experimentation where we compute PoL, VoL for randomly generated games. When plotted against each other, (PoL,VoL), reveal a strong anti-correlation effect. High price of learning is suggestive of low value of learning and vice versa. Understanding the topology of the Pareto curve on the space of (PoL, VoL) could quantify the tradeoffs between the risk and benefits of learning.

2. RELATED WORK

No-regret dynamics in games are central to the field of game theory and multi-agent learning [21]. Our protocols improve upon prior work that established convergence only in 2×2 games [7]. These dynamics are not efficient. Complexity results strongly indicate that no such dynamics exist for general games [11, 14]. Instead, this is a characterization result studying the tension between achieving no-regret guarantees and equilibration.

The algorithm method presented here for convergence to the one-shot NE while maintaining the no-regret property is similar in spirit to ones found in papers such as [18] in its tit-for-tat strategies in repeated games. However, this paper [18] is concerned with the set of NE obtained with the Folk Theorem conditions, larger in general than the set of one-shot NE. [9] defines a learning algorithm that converges in polynomial time to a NE of the one-shot game for 2 players. We extend this result to the case of N players while adding the requirement of no-regret to the strategies. [10] introduces an online algorithm that all players follow, leading them to convergence to Nash Equilibrium of the one-shot game. It also has the same concept of increasing periods of time after which the algorithm "forgets" and restarts. Indeed, a probabilistic bound of the same type as Hoeffding (in that case, Chebichev) is used to tune the length of these periods. In our case though, the learning part happens over the first three stages, while the last one is simply an implementation of the equilibrium.

The "weak" convergence of time-average no-regret dynamics to the set of CCE [22] has been useful in terms of extending price of anarchy [20] guarantees from NE to no-regret learning, which is usually referred to as the price of total anarchy [6]. Our equivalence result reduces the search for both best/worst case no-regret dynamics to a search over CCE which define a convex polytope in the space of distributions over strategy outcomes. In [12], similar results are proven for calibrated forecasting rules in almost every game. Our results extend easily to no-internal-regret algorithms and correlated equilibria (CE). [13] shows that through the definition of Φ-regret we can have a general definition that encompasses both no-internal and no-external regret.

In nonatomic congestion games regret-minimizing algorithms lead to histories of play where on most days the realized flow is an (ϵ, δ) approximate equilibrium [5] . In atomic congestion games general no-regret dynamics do not converge to NE. If we focus on specific no-regret dynamics such as multiplicative weights updates equilibration can be guaranteed [16]. Our results establish a hybrid of the two results. In atomic congestion games as the size of individual agents decreases, the set of coarse correlated equilibria focuses most of its probability mass on states where all but a tiny fraction of agents have a small incentive to deviate to another strategy. At the limit where the size of each agent becomes infinitesimal small, coarse correlated equilibria becomes arbitrarily focused on the unique nonatomic Nash flow.

In the case of utility games, [2] looks at two different social welfare ratios: the value of mediation defined as the ratio between the best CE and the best NE and the value of enforcement, which compares the worst CE to the worst NE. The value of mediation is shown to be a small constant for 2×2 games while the value of enforcement is unbounded, and they both are unbounded for larger games. Our results for cost (negative utility) games follow more closely the setting of [8] where once again the cost of worst CE is compared to the cost of the worst NE.

[4] shows it is NP-hard to compute a CCE with welfare strictly better than the lowest-welfare CCE. As a result our experimentation focuses on small instances but nevertheless reveals an interesting tension between the risks and benefits of learning.

3. PRELIMINARIES

Let I be the set of players of the game Γ. Each player $i \in I$ has a finite *strategy set* S_i and a cost function $c_i : S_i \times S_{-i} \longrightarrow [0, 1]$, where $S_{-i} = \prod_{j \neq i} S_j$. A player $i \in I$ may choose his strategy from his set of *mixed strategies* $\Delta(S_i)$, i.e the set of probability distributions on S_i. We extend the cost function's domain to the mixed strategies naturally, following the linearity of expectation.

Definition 1. A *Nash equilibrium* (NE) is a vector of distributions $(p_i^*)_{i \in I} \in \prod_{i \in I} \Delta(S_i)$ such that $\forall i \in I, \forall p_i \in \Delta(S_i)$

$$c_i(p_i^*, p_{-i}^*) \leq c_i(p_i, p_{-i}^*)$$

An *ϵ-Nash equilibrium* for $\epsilon > 0$ is one such that

$$c_i(p_i^*, p_{-i}^*) \leq c_i(p_i, p_{-i}^*) + \epsilon$$

We give the definition of a correlated equilibrium, from [3].

Definition 2. A *correlated equilibrium* (CE) is a distribution π over the set of action profiles $S = \prod_i S_i$ such that for all player i and strategies $s_i, s_i' \in S_i, s_i \neq s_i'$,

$$\sum_{s_{-i} \in S_{-i}} c_i(s_i, s_{-i}) \pi(s_i, s_{-i}) \leq \sum_{s_{-i} \in S_{-i}} c_i(s_i', s_{-i}) \pi(s_i, s_{-i})$$

We will also make use of coarse correlated equilibrium ([22]).

Definition 3. A *coarse correlated equilibrium* (CCE) is a distribution π over the set of action profiles $S = \prod_i S_i$ such that for all player i and strategy $s_i \in S_i$,

$$\sum_{s \in S} c_i(s) \pi(s) \leq \sum_{s_{-i} \in S_{-i}} c_i(s_i, s_{-i}) \pi_i(s_{-i})$$

where $\pi_i(s_{-i}) = \sum_{s_i \in S_i} \pi(s_i, s_{-i})$ is the marginal distribution of π with respect to i.

Definition 4. An online sequential problem consists of a feasible set $F \in R^m$, and an infinite sequence of cost functions $\{c^1, c^2 ..., \}$, where $c^t : R^m \to R$.

Given an algorithm A and an online sequential problem $(F, \{c^1, c^2, \dots\})$, if $\{x^1, x^2, \dots\}$ are the vectors selected by A, then the cost of A until time T is $\sum_{t=1}^{T} c^t(x^t)$. Regret compares the performance of an algorithm with the best static action in hindsight:

Definition 5. The regret of algorithm A at time T is defined as $R(T) = \sum_{t=1}^{T} c^t(x^t) - \min_{x \in F} \sum_{t=1}^{T} c^t(x)$.

An algorithm is said to have no regret or that it is Hannan consistent [22], if for every online sequential problem, its regret at time T is $o(T)$. For the context of game theory, which is our focus here, the following definition of no-regret learning dynamics suffices.

Definition 6. The regret of agent i at time T is defined as $R(T) = \sum_{t=1}^{T} c_i(s^t) - \min_{s'_i \in S_i} \sum_{t=1}^{T} c_i(s'_i, s^t_{-i})$.

We will also make use of the following inequality from [15].

THEOREM 1. *Suppose $(X_k)_{k=1}^n$ are independent random variables taking values in the interval $[0,1]$. Let Y denote the empirical mean $Y = \frac{1}{n}\sum_{k=1}^{n} X_k$. Then for $t > 0$*

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq t) \leq 2\exp\big(-2nt^2\big)$$

4. NO-REGRET DYNAMICS CONVERGING TO NASH EQUILIBRIUM IN SELF-PLAY

THEOREM 2. *In a finite game with N players, for any $\epsilon > 0$, there exist learning dynamics that satisfy simultaneously the following two properties: i) against arbitrary opponents their average regret is at most ϵ, ii) in self-play they converge pointwise to a ϵ-Nash equilibrium with probability 1.*

PROOF. We divide the play in four stages. In the first stage, players explore their strategy space sequentially and learn the costs obtained from every action profile. In the second stage, they communicate by cheap talk their costs. In the third stage, they compute the desired ϵ-Nash equilibrium that is to be reached, for $\epsilon > 0$. In the fourth stage, players are expected to use their equilibrium strategies and they monitor other players in case these deviate from equilibrium play.

The players are expected to follow a communication procedure and implement a no-regret strategy in the case of another player's deviation. Since the first three stages have finite length (though very long: exponential in the size of the cost matrix [14]), the no-regret property follows. The restriction on convergence to an ϵ-NE, instead of a mixed NE (so $\epsilon = 0$) arises from the fact that even games with rational costs can possess equilibria that are irrational [19].

Settlement on a particular NE can be decided by a fixed rule before play, such as lexicographically in the players' actions or the NE that has the lowest social cost.

In the fourth stage, players have settled on an equilibrium and will implement it. To fulfill the requirement of pointwise convergence, it is not enough for the players to stick to a deterministic sequence of plays. We want them to pick randomly a move from their equilibrium distribution of actions. During this process, there can happen that the generated sequence of play of an opponent does not closely match his equilibrium distribution. In that case, the players need to decide whether the opponent has been truthful but "unlucky" or deliberately malicious.

We achieve this by dividing the fourth stage in blocks of increasing length. Let $n \in \mathbb{N}$ denote the block number, we set block n to have a length of $l(n) = n^2$ turns. On these blocks, the players will make use of statistical tests to verify that all other opponents are truthful. We want to find a test such that a truthful but possibly unlucky player will fail almost surely a finite number of these tests, while a malicious player will almost surely fail an infinite number of these.

We first look at the case where we have N players with only two strategies, 0 and 1. We can then identify the equilibrium distribution of a player i, to the probability p_i^* that he chooses action 1.

Suppose the play is at the n-th block and player i chooses to implement the mixed strategy p_i. Let $(X_k^i)_{k=1,\dots,l(n)}$ denote the sequence of strategies chosen by player i, such that $X_k^i \sim \mathcal{B}(p_i)$ and all are independent. Let Y_n^i be the empirical frequency of strategy 1 during block n.

$$Y_n^i = \frac{1}{l(n)} \sum_{j=1}^{l(n)} X_j^i$$

If the player is truthful and implements the prescribed NE, then we have $p_i = p_i^*$ and we expect the empirical frequency of strategy 1 Y_n^i to be close to p_i^*. Otherwise, a malicious player will choose $p_i \neq p_i^*$.

Let A_n^i denote the event $A_n^i = \{|Y_n^i - p_i^*| \geq t_n\}$. In other words, we are trying to determine how far the empirical frequency of strategy 1 is from the expected equilibrium distribution. If the event A_n^i is realised, then the test is failed: the empirical distribution of play is too far from the expected NE distribution. The idea is to make block after block the statistical test more discriminating, i.e get a decreasing sequence $(t_n)_n$ such that a truthful player will only see a finite number of events A_n^i happen, while a malicious one will face an infinite number of failures.

We claim that picking $t = n^{-\alpha}$ with $0 < \alpha < 1$ is enough. Indeed by Hoeffding's inequality we have that

$$\mathbb{P}(A_n^i) \quad \leq \quad 2\exp\big(-2n^2t^2\big)$$

if the player is truthful (remember that block n has length $l(n) = n^2$).

Extending the proof to the case where a player i has finite strategy set S_i is not hard. Let $(p_s^i)_{s \in S}$ be the distribution that the i-th player decides to implement, while $(p_s^{i,*})_{s \in S}$ is the NE distribution for player i. Let $X_k^{i,s}$ follow a multinomial distribution of parameters $(p_s^i)_{s \in S}$. Then $Y_n^{i,s}$ is the empirical frequency of strategy s during block n for player i. We define events

$$A_n^{i,s} = \{|Y_n^{i,s} - p_s^{i,*}| \geq t_n\}.$$

Then we define our test A_n^i to be $\cup_{s \in S_i} A_n^{i,s}$. Using Hoeffding's inequality again we obtain:

$$\begin{aligned} \mathbb{P}(A_n^i) &= \mathbb{P}(\cup_{s \in S_i} A_n^{i,s}) \\ &\leq \sum_{s \in S_i} \mathbb{P}(A_n^{i,s}) \leq |S_i| \times 2\exp(-2n^2t^2) \end{aligned}$$

Thus $\sum \mathbb{P}(A_n^i) < +\infty$ for $0 < \alpha < 1$, so by Borel-Cantelli we know that the A_n^i will only ever happen a finite number of times if the player is truthful, i.e if $\mathbb{E}[Y_n^{i,s}] = p_s^{i,*}$.

To satisfy the no-regret property, we do the following: if one of the opponents failed the statistical test described earlier, then all players will implement a no-regret strategy for a time $n^{2+\delta}$ to compensate for that. We call this block of size $n^{2+\delta}$ a *compensating block*.

If a finite number of tests fails, then the whole sequence satisfies the ϵ-regret property, since players are arbitrarily close to the ϵ-Nash equilibrium. When one of the tests fails, say, at block n, the maximum regret accumulated is of size n^2. The following compensating block guarantees that overall regret has grown by a value bounded by $n^{1-\delta}$, so sublinearly.

We also guarantee that the expected turn number that ends the last of the truthful player's potential failed block is not infinity. Indeed let B_n be the event that the last failed block is the n-th one. Then

$$\begin{aligned}\mathbb{P}(B_n) &= \mathbb{P}(A_n) \times \mathbb{P}(A^c_{n+1}) \dots \\ &\leq 2\exp(-2n^2t^2) \times 1 \dots \\ &\leq 2\exp(-2n^2t^2)\end{aligned}$$

We use A^c to denote the complement of event A. The first equality holds by independence of the blocks, the second inequality is true from Hoeffding's and the fact that a probability is less or equal to 1. We then define L to be the index of the turn that ends the last compensating block of a truthful player. L is a random variable on the integers. We have

$$\mathbb{E}[L] \leq \sum_n \Big(\sum_{k=1}^{n} (k^2 + k^{2+\delta}) \Big) \times 2\exp(-2n^2t^2) < +\infty$$

We bound $\mathbb{E}[L]$ by assuming a truthful player got every test wrong up to the latest failed one. Then the last turn L occurs at index $\sum_n (n^2 + n^{2+\delta})$. We multiply this by the bound on $\mathbb{P}(B_n)$ and use the property of the exponential to conclude that $\mathbb{E}[L]$ is bounded. □

5. EQUIVALENCE BETWEEN COARSE CORRELATED EQUILIBRIA AND NO-REGRET DYNAMICS

The long-run average outcome of no-regret learning converges to the set of coarse correlated equilibria [22]. Here, we argue the reverse direction.

THEOREM 3. *Given any coarse correlated equilibrium C of a normal form game with a finite number of players n and finite number of strategies, there exist a set of n-no regret processes such that their interplay converges to the coarse correlated equilibrium C.*

PROOF. Suppose that we are given a coarse correlated equilibrium C of a n-player game*. There exists a natural number K, such that all probabilities are multiples of $1/K$. We can create a sequence of outcomes S of length K, such that the probability distribution that chooses each such outcome with probability $1/K$ is identical to the given coarse correlated equilibrium C. The high level idea is to have the agents play this sequence in a sequential, cyclical manner and punish any observed deviation from it by employing any chosen no regret algorithm (e.g., Regret Matching).

Let's denote the j-th element of this sequence as $< x_1^j, x_2^j, ..., x_N^j >$, where $0 \leq j \leq K-1$. Each element of this sequence will act as a recommendation vector for the no regret algorithm. Given the sequence above we are ready to define for each of the N players a no regret algorithm, such that their interplay converges to the given coarse correlated equilibrium C.

The algorithm for the i-th player is as follows: at time zero she plays the i-th coordinate of the first element in S. As long as the other players' responses up to any point in time t are in unison with S, that is for every $t' < t$ and $j \neq i$ the strategy implemented by player j at time t' was $x_j^{t' \mod K}$ then the i-player will follow the recommendation of the S sequence playing $x_i^{t \mod K}$. However, as soon as the player recognizes any sort of deviation from S by another player then the player will just disregard any following recommendations coming from S and will merely follow from that point on a no regret algorithm of her liking.

It is straightforward to check that in self-play this protocol converges to the given coarse correlated equilibrium C. We need to also prove that all of these algorithms are no-regret algorithms. When analyzing the accumulated regret of any of the algorithms above we split their behavior into two distinct segments. The first segment corresponds to the time periods before any deviation is recorded from the recommendation provided by C. For this segment, the definition of coarse correlated equilibrium implies that each agent experiences bounded total regret (only corresponding to the last partial sequence of length at most K). Once a first deviation is witnessed by the player in question, she turns to her no-regret algorithm of choice and the no regret property then follows from this algorithm. As a result, each algorithm exhibits vanishing (average) regret in the long run. □

6. COLLAPSING EQUILIBRIUM CLASSES

6.1 Congestion games with small agents

We have a finite ground set of elements E. There exist a constant number k of types of agents and each agent of type i has an associated set of allowable strategies/paths S_i. S is the set of possible strategy outcomes. Let N_i be the set of agents of type i. We assume that each agent of type i controls a flow of size $1/|N_i|$, which he assigns to one of his available paths S_i. This can also be interpreted as a probability distribution over the set of strategies S_i. Each element e has a nondecreasing cost functions of bounded slope $c_e : \mathbb{R} \rightarrow \mathbb{R}$ which dictates its latency given its load. The load of an edge e is $\ell_e(s) = \sum_i \frac{k_i}{|N_i|}$, where k_i the number of agents of type i which have edge e in their path in the current strategy outcome. The cost of any agent of type i for choosing strategy $s_i \in S_i$ is $c_{s_i}(s) = \sum_{e \in s_i} c_e(\ell_e(s))$. In many cases, we abuse notation and write ℓ_e, c_{s_i} instead of $\ell_e(s), c_{s_i}(s)$ when the strategy outcome is implied. The social cost, *i.e.*, the sum of agents' costs, is equal to $C(s) = \sum_e c_e(\ell_e)\ell_e$. Finally, it is useful to keep track of the flows going trough a path s_i or an edge e when focusing on agents of a single type i. We denote these quantities as $\ell^i_{s_i}(s)$ and $\ell^i_e(s) = \sum_{s_i \ni e} \ell^i_{s_i}(s)$. For any strategy outcome s and any type i, $\sum_{s_i \in S_i} \ell^i_{s_i}(s) = 1$ defining a distribution over S_i.

We normalize the cost functions uniformly so that the cost of any path as well as the increase to the cost of any path due to the deviation by a single agent are both upper and lower bounded by absolute positive constants. To simplify the number of relevant parameters we treat the number of resources, paths as a constant.

THEOREM 4. *In congestion games with cost functions of bounded slope, as long as the flow that each agent controls is at most ϵ, any coarse correlated equilibrium applies at least $1 - O(\epsilon^{1/4})$ probability to set of outcomes where at most $O(\epsilon^{1/8})$ fraction of agents have more than $O(\epsilon^{1/8})$ incentive to deviate.*

PROOF. Let π be a coarse correlated equilibrium of the game and let $\pi(s)$ the probability that it assigns to strategy outcome s. By definition of CCE, the expected cost of any agent cannot decrease if he deviates to another strategy. We consider two possible deviations for each agent of type i. Deviation A has the agent deviating to a strategy that has minimal expected cost according to π (amongst his available strategies). Deviation B has the agent deviating to the mixed strategy that corresponds to expected flow of all the agents of type i in π. If each agent controlled infinitesimal flow

*We will assume that all involved probabilities are rational. Since the set of coarse correlated equilibria is a convex polytope defined $A\mathbf{x} \leq \mathbf{b}$ where all entries of A, $\mathbf{b}$ are rational every correlated equilibrium involves rational probabilities or can be approximated with arbitrarily high accuracy by using rational probabilities.

then his cost would be equal to

$$\min_{s_i \in S_i} \mathbf{E}_{\mathbf{s}\sim\pi}[\sum_{e \in S_i} c_e(\ell_e(s))]$$

and

$$\sum_{s_i \in S_i} \mathbf{E}_{\mathbf{s}\sim\pi}[\ell^i_{s_i}(s)]\mathbf{E}_{\mathbf{s}\sim\pi}[\sum_{e \in S_i} c_e(\ell_e(s))]$$

when deviating to A and B respectively.

Furthermore, his expected cost at π would be less or equal to his cost when deviating to A, which would again be less or equal to his cost when deviating to B. Due to the normalization of the cost functions and the small flow $\leq \epsilon$ that each agent controls this ordering is preserved modulo $O(\epsilon)$ terms. This ordering and size of error terms is preserved when computing the (expected) social costs according to π, the sum of the deviation costs when each agent deviates according to A and the sum of all deviation costs when they deviate according to B. I.e.

$$\mathbf{E}_{\mathbf{s}\sim\pi}[C(s)] \leq \sum_i \min_{s_i \in S_i} \mathbf{E}_{\mathbf{s}\sim\pi}[\sum_{e \in S_i} c_e(\ell_e(s))] + O(\epsilon)$$
$$\leq \sum_i \sum_{s_i \in S_i} \mathbf{E}_{\mathbf{s}\sim\pi}[\ell^i_{s_i}(s)]\mathbf{E}_{\mathbf{s}\sim\pi}[\sum_{e \in S_i} c_e(\ell_e(s))] + O(\epsilon)$$

By applying Chebyshev's sum inequality we can derive that for each edge e

$$\mathbf{E}_{\mathbf{s}\sim\pi}[\ell_e(s)]\mathbf{E}_{\mathbf{s}\sim\pi}[c_e(\ell_e(s))] \leq \mathbf{E}_{\mathbf{s}\sim\pi}[\ell_e(s)c_e(\ell_e(s))]$$

Taking summation over all edges, we produce the inverse of our first inequality, since $\ell_e(s) = \sum_i \sum_{s_i \ni e} \ell^i_{s_i}(s)$, implying that all related terms are equal to each other up to errors of $O(\epsilon)$.

By linearity of expectation we have that

$$\mathbf{E}_{\mathbf{s}\sim\pi}\Big[\Big(\ell_e(s) - \mathbf{E}_{\mathbf{s}\sim\pi}[\ell_e(s)]\Big)c_e(\mathbf{E}_{\mathbf{s}\sim\pi}[\ell_e(s)])\Big] = 0.$$

Combining everything together we derive that

$$\mathbf{E}_{\mathbf{s}\sim\pi}\Big[\sum_e \Big(\ell_e(s) - \mathbf{E}_{\mathbf{s}\sim\pi}[\ell_e(s)]\Big)$$
$$\cdot \quad \Big(c_e(\ell_e(s)) - c_e(\mathbf{E}_{\mathbf{s}\sim\pi}[\ell_e(s)])\Big)\Big] = O(\epsilon).$$

Since costs $c_e(x)$ are nondecreasing, the function whose expectation we are computing is always nonnegative. In fact, since we have assumed that the slope of the cost functions is upper, lower bounded by some fixed constants we have that

$$\mathbf{E}_{\mathbf{s}\sim\pi}\sum_e \Big(\ell_e(s) - \mathbf{E}_{\mathbf{s}\sim\pi}[\ell_e(s)]\Big)^2 = O(\epsilon).$$

By applying Cauchy-Schwarz inequality, we derive that

$$\mathbf{E}_{\mathbf{s}\sim\pi}\sum_e |\ell_e(s) - \mathbf{E}_{\mathbf{s}\sim\pi}[\ell_e(s)]| = O(\sqrt{\epsilon})$$

The coarse correlated equilibrium π is closely concentrated around its "expected" flow $E_{s\sim\pi}[\ell_e(s)]$. For simplicity we denote this continuous flow y. The set of strategy outcomes $S' \subset S$ with $\sum_e |\ell_e(s') - \ell_e(y)| > \epsilon^{1/4}$ must receive (in π) cumulative probability mass less than $O(\epsilon^{1/4})$. If we consider the rest strategy outcomes, which we denote as "good", then we have that in each "good" outcome both the social cost (i.e. the sum of the costs of all agents) as well as the cost of the optimal path are always within $O(\epsilon^{1/4})$ of the respective social cost and cost of the optimal path under flow y. Finally, by combining our main inequality with the fact that $\mathbf{E}_{\mathbf{s}\sim\pi}\sum_e |\ell_e(s) - \mathbf{E}_{\mathbf{s}\sim\pi}[\ell_e(s)]| = O(\sqrt{\epsilon})$ we have that the social cost under flow y are within $O(\sqrt{\epsilon})$ of the cost of the optimal path under y.[†] Hence, all of the "good" outcomes have social cost within $O(\epsilon^{1/4})$ of the cost of their own optimal path. So, at most $O(\epsilon^{1/8})$ agents in each "good" outcome can decrease their cost by more than $O(\epsilon^{1/8})$ by deviating to another path. □

6.2 CE = CCE for N agents 2 strategy games

PROPOSITION 1. *For games where all players have only two strategies, the set of coarse correlated equilibria is the same as the set of correlated equilibria.*

PROOF. Let i be one of the players, suppose his two strategies are A and D, where we pick D to be the deviating one. Then the requirement for correlated equilibrium states that

$$\sum_{s_{-i} \in S_{-i}} u_i(s_{-i}, D)\pi(s_{-i}, A) \geq \sum_{s_{-i} \in S_{-i}} u_i(s_{-i}, A)\pi(s_{-i}, A)$$

while the corresponding one for coarse correlated equilibrium is

$$\textstyle\sum_{s_{-i} \in S_{-i}} u_i(s_{-i}, D)(\pi(s_{-i}, A) + \pi(s_{-i}, D)) \geq$$
$$\textstyle\sum_{s_{-i} \in S_{-i}} (u_i(s_{-i}, D)\pi(s_{-i}, D) + u_i(s_{-i}, A)\pi(s_{-i}, A))$$

which is equivalent after removing the $\sum_{s_{-i} \in S_{-i}} u_i(s_{-i}, D)\pi(s_{-i}, D)$ term on both sides. □

7. SOCIAL WELFARE GAPS FOR DIFFERENT EQUILIBRIUM CONCEPTS

We define a measure to compare equilibria obtained under no-regret algorithms to Nash equilibria: *the value of learning*. This measure quantifies by how much the players are able to decrease their costs when relaxing the equilibrium requirements from Nash to CCE.

Definition 7. Define the value of learning in cost games *VoL* as the ratio of the social cost of the best Nash equilibrium to that of the best coarse correlated equilibrium.

$$\text{VoL}(\Gamma) = \frac{\text{best NE}}{\text{best CCE}}$$

[†] Since we have

$$\mathbf{E}_{\mathbf{s}\sim\pi}\sum_e |\ell_e(s) - \mathbf{E}_{\mathbf{s}\sim\pi}[\ell_e(s)]| = O(\sqrt{\epsilon})$$

the terms

$$\sum_i \min_{s_i \in S_i} \mathbf{E}_{\mathbf{s}\sim\pi}[\sum_{e \in S_i} c_e(\ell_e(s))]$$

and

$$\sum_i \min_{s_i \in S_i} \sum_{e \in S_i} c_e(\mathbf{E}_{\mathbf{s}\sim\pi}[\ell_e(s)])$$

as well as the pair of

$$\sum_i \sum_{s_i \in S_i} \mathbf{E}_{\mathbf{s}\sim\pi}[\ell^i_{s_i}(s)]\mathbf{E}_{\mathbf{s}\sim\pi}[\sum_{e \in S_i} c_e(\ell_e(s))]$$

with the term

$$\sum_i \sum_{s_i \in S_i} \mathbf{E}_{\mathbf{s}\sim\pi}[\ell^i_{s_i}(s)] \sum_{e \in S_i} c_e(\mathbf{E}_{\mathbf{s}\sim\pi}[\ell_e(s)])$$

are within $O(\sqrt{\epsilon})$ of each other, but the first and last term are within $O(\epsilon)$ of each other, implying that all terms are within $O(\sqrt{\epsilon})$.

Since the set of NE is included in the set of CCE, then the best NE in terms of social cost will always be greater than the best CCE. Thus we take the ratio so that the value of learning is always greater than or equal to 1, a convention also found in other papers related to the price of anarchy [2, 8].

Conversely, we define *the price of learning* as the ratio of the worst CCE to the worst NE.

Definition 8. Define the price of learning *PoL* in a cost game Γ as the ratio of the social cost of the worst coarse correlated equilibrium to that of the worst Nash equilibrium.

$$\text{PoL}(\Gamma) = \frac{\text{worst CCE}}{\text{worst NE}}$$

This approach is not too dissimilar to the one adopted in [8], which defines the ration of the worst CE to the worst NE as the price of mediation. With the help of proposition 1, we can extend their result to learning algorithms that possess the no-regret property.

7.1 2x2 games

Denote by $\Gamma_{2\times 2}$ the class of 2×2 games. We are interested in the best-case scenario: how high the ratio of the value of learning can get for all 2×2 games.

Definition 9. Denote by $\text{VoL}(\Gamma_{2\times 2}) = \sup_{\Gamma \in \Gamma_{2\times 2}} \text{VoL}(\Gamma)$ the value of learning for the class of 2×2 games.

PROPOSITION 2. *VoL*$(\Gamma_{2\times 2}) \geq \frac{3}{2}$

PROOF. Consider the following cost game for $x > 1$

$$\begin{array}{c} \\ T \\ B \end{array} \begin{array}{cc} L & R \\ \left(\begin{array}{cc} 0, x-1 & x, x \\ 1, 1 & x-1, 0 \end{array}\right) \end{array}$$

The game admits three NE: (T, L), (B, R) and $((0.5, 0.5), (0.5, 0.5))$. The first two have social cost equal to $x-1$ while the latter's is $x/2$. Hence for $x > 2$, the social cost of the best NE is $x - 1$.

The correlated equilibrium that minimizes social cost assigns probability $1/3$ to every action profile except for (T, R). Its social cost is $2x/3$. Hence, in this game, VoL $= \frac{3(x-1)}{2x}$. Taking $x \longrightarrow +\infty$, we derive $\text{VoL}(\Gamma_{2\times 2}) \geq \frac{3}{2}$. □

We conjecture that this $\frac{3}{2}$ bound is tight, i.e, there is no 2×2 game Γ such that $\text{VoL}(\Gamma) > 3/2$.

To support this claim, we have run numerical simulations on games generated from a random uniform distribution. An interesting result is the predominance of games for which the ratios are 1, i.e mediation does not better the social welfare/cost. We then observe higher ratios at a lower rate, hence our histograms look like those of a power law (figure 1). The obtained ratios come close to the 3/2 threshold, without going further (only a few ratios approaching 1.4 were observed over 10^7 simulations).

PROPOSITION 3. *PoL*$(\Gamma_{2\times 2}) = 2$

PROOF. By proposition 1, the social cost of the worst CE is equal to the social cost of the worst CCE, since the set of CE is the same as the set of CCE. Then by [8], we have that $PoL(\Gamma_{2\times 2}) = 2$. □

In figure 2 we present a 2D histogram of the joint distribution of the VoL and PoL. 10^6 games were generated and for each we compute both values. The size of the dot is representative of how many games possess particular values for the VoL and the PoL.

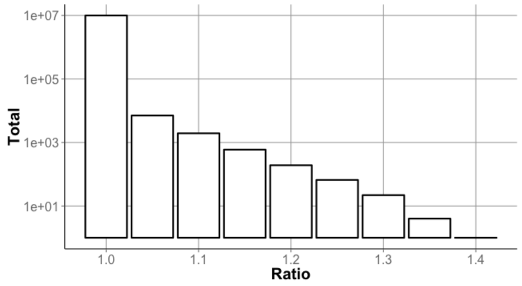

Figure 1: Histogram of values of learning obtained over 10^7 simulations for 2×2 games. A $\log_{10}$ scale is used for the y-axis.

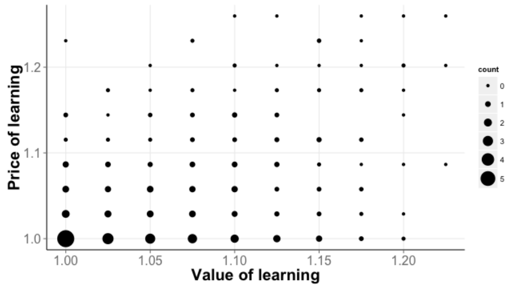

Figure 2: 2D histogram of VoL and PoL over 10^6 simulations for 2×2 games. The count legend is to be interpreted as a power of ten (so count of 5 is 10^5)

7.2 Larger games

Next, we examine larger games, i.e., games with more than 2 players and/or more than 2 strategies per player. Let Γ_{m_1,m_2} denote a 2 player game with respectively m_1 and m_2 strategies for each player.

PROPOSITION 4. *For sets of games* Γ_{m_1,m_2}, $\max(m_1, m_2) > 2$, *we have VoL*$(\Gamma_{m_1,m_2}) = +\infty$.

PROOF. Consider for $\epsilon < \frac{1}{2}$ the game

$$\begin{array}{c} \\ T \\ B \end{array} \begin{array}{c} \begin{array}{ccc} L & C & R \end{array} \\ \left(\begin{array}{ccc} 1-\epsilon, 1-\epsilon & 2\epsilon, \frac{3\epsilon}{2} & 2\epsilon, \frac{1}{2} \\ \frac{1}{2}, 2\epsilon & \epsilon, 1-\epsilon & 1, 2\epsilon \end{array}\right) \end{array}$$

The game admits three NE: (L, B), $((0, 1), (2/3, 0, 1/3))$ and $(2/3, 1/3), (0, 1-\epsilon, \epsilon)$. Of the three, the latter has the lowest social cost, equal to $1/3 + o(\epsilon)$, where $o(\epsilon) \longrightarrow_{\epsilon \to 0} 0$.

We can define the following correlated equilibrium π:

$$\begin{array}{c} \\ T \\ B \end{array} \begin{array}{c} \begin{array}{ccc} L & C & R \end{array} \\ \left(\begin{array}{ccc} 0 & 1-\frac{5\epsilon}{2} & \epsilon \\ \epsilon & 0 & \epsilon/2 \end{array}\right) \end{array}$$

The best social cost in a correlated equilibrium will be lower than that of π, which is $o(\epsilon)$. We also have that the best social cost in a CCE will be lower than that of a CE.

Thus taking $\epsilon \to 0$, we obtain an unbounded VoL. □

Since the set of CE $\subseteq$ CCE, we can again extend some results from previous papers to the latter set.

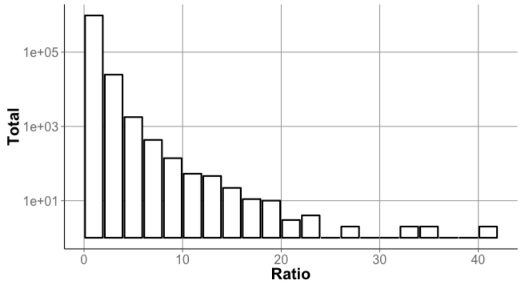

Figure 3: Histogram of ratios best NE/best CCE (VoL) obtained over 10^6 simulations for 3×3 games.

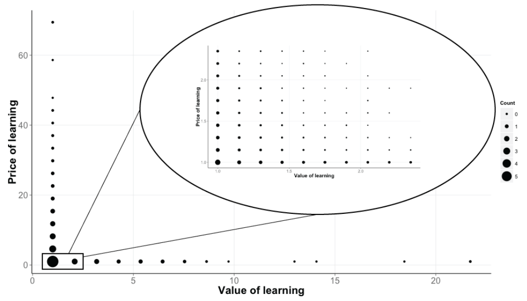

Figure 4: 2D histogram of VoL and PoL over 10^6 simulations for 3×3 games. The count legend is to be interpreted as a power of ten (so count of 5 is 10^5). We zoomed in the portion $[1, 2.5]^2$ to show finer results.

PROPOSITION 5. *For games Γ_{m_1,m_2}, $\max(m_1, m_2) > 2$, we have PoL$(\Gamma_{m_1,m_2}) = +\infty$.*

PROOF. Since CE $\subseteq$ CCE, the social cost of the worst CCE is higher than that of the worst CE. By [8] we have that PoM $= +\infty$, hence PoL $= +\infty$. □

We run a number of simulations to see how VoL is distributed for random games (figure 3). We have also included a 2D histogram (figure 4) showing (VoL, PoL) for a number of generated games. Some sampled games have high VoL and some high PoL but not both, indicating a competitive relationship between the two quantities.

8. CONCLUSION

No-regret learning, due to its simplicity to implement in multi-agent settings, has seen considerable exposure in the literature of the last decade. The convergence of play to the set of coarse correlated equilibria is one property that makes these learning algorithms useful in practice. But if we look closer, it is not clear where this convergence leads the play. We have first shown that we can steer it using a somewhat unnatural algorithm to any NE of the one-shot game, while maintaining the no-regret property. In the next sections, we have understood better how the class of CCE relates to no-regret dynamics, and to the smaller class of CE. This lead us to define more general measures of the price of anarchy: if it is hard to predict where the play following no-regret dynamics will go, we are at least able to give some PoA bounds on the resulting payoffs. This section is concluded with experimental results that show a concentration of small ratios, indicating a closeness to NE payoffs. The question of the Value of Learning for 2×2 games is left open, with our proven lower bound of 3/2, which we believe to be tight.

REFERENCES

[1] E. Anshelevich, A. Dasgupta, J. Kleinberg, É.. Tardos, T. Wexler, and T. Roughgarden. The price of stability for network design with fair cost allocation. In *Foundations of Computer Science (FOCS)*, pages 295–304. IEEE, 2004.

[2] I. Ashlagi, D. Monderer, and M. Tennenholtz. On the value of correlation. *Journal of Artificial Intelligence Research*, pages 575–613, 2008.

[3] R. J. Aumann. Subjectivity and correlation in randomized strategies. *Journal of mathematical Economics*, 1(1):67–96, 1974.

[4] S. Barman and K. Ligett. Finding any nontrivial coarse correlated equilibrium is hard. In *ACM Conference on Economics and Computation (EC)*, 2015.

[5] A. Blum, E. Even-Dar, and K. Ligett. Routing without regret: On convergence to nash equilibria of regret-minimizing algorithms in routing games. *Theory of Computing*, 6(1):179–199, 2010.

[6] A. Blum, M. Hajiaghayi, K. Ligett, and A. Roth. Regret minimization and the price of total anarchy. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 373–382. ACM, 2008.

[7] M. Bowling. Convergence and no-regret in multiagent learning. *Advances in neural information processing systems*, 17:209–216, 2005.

[8] M. Bradonjic, G. Ercal-Ozkaya, A. Meyerson, and A. Roytman. On the price of mediation. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 315–324. ACM, 2009.

[9] R. I. Brafman and M. Tennenholtz. Efficient learning equilibrium. *Artificial Intelligence*, 159(1):27–47, 2004.

[10] V. Conitzer and T. Sandholm. Awesome: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. *Machine Learning*, 67(1-2):23–43, 2007.

[11] C. Daskalakis, P. W. Goldberg, and C. H. Papadimitriou. The complexity of computing a Nash equilibrium. *SIAM J. Comput.*, 39(1):195–259, 2009.

[12] D. P. Foster and R. V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1):40–55, 1997.

[13] A. Greenwald and A. Jafari. A general class of no-regret learning algorithms and game-theoretic equilibria. In *Learning Theory and Kernel Machines*, pages 2–12. Springer, 2003.

[14] S. Hart and Y. Mansour. The communication complexity of uncoupled nash equilibrium procedures. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 345–353. ACM, 2007.

[15] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[16] R. Kleinberg, G. Piliouras, and É. Tardos. Multiplicative updates outperform generic no-regret learning in congestion games. In *ACM Symposium on Theory of Computing (STOC)*, 2009.

[17] E. Koutsoupias and C. H. Papadimitriou. Worst-case equilibria. In *STACS*, pages 404–413, 1999.

[18] M. L. Littman and P. Stone. A polynomial-time nash equilibrium algorithm for repeated games. *Decision Support Systems*, 39(1):55–66, 2005.

[19] J. Nash. Non-cooperative games. *Annals of mathematics*, pages 286–295, 1951.

[20] T. Roughgarden. Intrinsic robustness of the price of anarchy. In *Proc. of STOC*, pages 513–522, 2009.

[21] Y. Shoham, R. Powers, and T. Grenager. If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, 171(7):365–377, 2007.

[22] H. Young. *Strategic Learning and Its Limits*. Arne Ryde memorial lectures. Oxford University Press, 2004.