A Novel Individually Rational Objective In Multi-Agent Multi-Armed Bandits: Algorithms and Regret Bounds

Aristide C. Y. Tossou
Chalmers University
Gothenburg, Sweden
yedtoss@gmail.com

Christos Dimitrakakis
University of Oslo/Chalmers University
christos.dimitrakakis@gmail.com

Jaroslaw Rzepecki
Microsoft Research
Cambridge, UK
jaroslaw.rzepecki@microsoft.com

Katja Hofmann
Microsoft Research
Cambridge, UK
katja.hofmann@microsoft.com

ABSTRACT

We study a two-player stochastic multi-armed bandit (MAB) problem with different expected rewards for each player, a generalisation of two-player general sum repeated games to stochastic rewards. Our aim is to find the egalitarian bargaining solution (EBS) for the repeated game, which can lead to much higher rewards than the maximin value of both players. Our main contribution is the derivation of an algorithm, UCRG, that achieves simultaneously for both players, a high-probability regret bound of order $\tilde{O}\left(T^{2/3}\right)$ after any T rounds of play. We demonstrate that our upper bound is nearly optimal by proving a lower bound of $\Omega\left(T^{2/3}\right)$ for any algorithm. Experiments confirm our theoretical results and the superiority of UCRG compared to the well-known explore-then-commit heuristic.

KEYWORDS

multi-armed bandits, egalitarian bargaining solution, safety, individual rationality

ACM Reference Format:
Aristide C. Y. Tossou, Christos Dimitrakakis, Jaroslaw Rzepecki, and Katja Hofmann. 2020. A Novel Individually Rational Objective In Multi-Agent Multi-Armed Bandits: Algorithms and Regret Bounds. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 9 pages.

1 INTRODUCTION

Multi-agent systems are ubiquitous in many real-life applications such as autonomous drones, games, computer networks, etc. Agents acting in such systems are usually modeled as self-interested, aiming to maximize their own individual utility. We focus on stochastic two-player general-sum repeated games, a setting which captures the key challenges faced when interacting in a multi-agent system. We consider the case where in each round, the two players (the *agent* and its *opponent*) simultaneously select actions and then each obtain a numerical reward. The goal of each player is to maximize its individual accumulated reward over multiple rounds. Thus, the problem can be seen as an instance of the multi-agent multi-armed bandit problem, where the reward obtained by each agent depends on all agents' actions.

An agent in such a game will behave differently depending on what it assumes about the opponent. Powers et al. [31] propose rigorous criteria that characterise behaviours: (1) *Safety*: against any opponent, the average reward is close to the maximin; (2) *Individual Rationality*: in self-play, the average reward is Pareto efficient[1] and individually not below the maximin. However, Individual Rationality is not well defined, since due to various *folk theorems* [30], for infinite-horizon undiscounted reward, every outcome that is feasible and individually not below the maximin can be realized as a Nash Equilibrium of the repeated game. In short, the set of Individually Rational outcomes may be infinite. So the main question is which outcome should one aim for and why?

For two agents, this question has received a lot of attention in the *Bargaining Problem* [29]. This is a game where, if the agents play without any bargaining, then their baseline utility is achieved at the so-called *disagreement point*. However, by bargaining, they can reach an agreement that will give them higher utility. Many solutions to the problem have been proposed (Nash [29], Egalitarian [21], Utilitarian [39], Kalai–Smorodinsky [22]), based on axiomatic properties of the corresponding solution concept.

In this paper, we strengthen the Individual Rational criterion proposed by Powers et al. [31] and require the agents to be close to the *unique* solution of a Bargaining Problem, with the disagreement point being the maximin of both players. We also pick the Egalitarian Bargaining Solution (EBS) since, as opposed to the other solutions, it has been shown [21] to be connected to some *fairness* and *equality* concepts, and in particular to one of the Rawls' notions of justice [33]. EBS also enjoys strong mathematical properties. On top of the *individual rationality criterion*, it also satisfies *independence of irrelevant alternatives* (i.e. eliminating choices that were irrelevant does not change the choices of the agents), *individual monotonicity* (if a player has better options in one game compared to another game, then that player should get a weakly-better value in the game with better options) and (importantly) *uniqueness*.

Related work. There is a growing interest in the multi-agent multi-armed bandit problem. Many of the works [3, 9, 16, 23, 35] have focused on maximizing *social welfare*, i.e. the *sum of rewards*

[1] i.e., it is impossible for one agent to change to a better policy without making the other agent worse off.

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

over all agents. However, this may not make sense under individual rationality since it is possible for an agent to obtain lower than what it could have obtained without cooperation regardless of the strategies other agents follow [8]. We illustrate this issue in Example 1.1.

A second line of research has focused on single-stage equilibria such as single-stage Nash Equilibrium or single-stage Correlated Equilibrium [2, 5, 6, 11, 36]. Aiming for single-stage equilibria for *repeated games* is problematic [8] since the agents can usually obtain individually much larger rewards by cooperating. Moreover, unlike this paper, many previous works [8, 26, 28, 31, 38] focus on deterministic rewards. Therefore their work does not deal with uncertainty about the rewards. As we show experimentally in this paper, a well-known heuristic that spends some initial rounds to explore and learn about the rewards is inferior to our proposed algorithm.

Other works consider discounted rewards, effectively decreasing the effect of future actions [12, 15, 17, 18, 24, 25, 27, 32, 41]. In contrast, we consider the case of infinite-horizon average rewards. Works such as [2, 5] provide the notion of "no-regret" which is orthogonal to our setting. Indeed, their notion of regret is not related to a lack of information about the rewards.

Brafman and Tennenholtz [7], Wei et al. [40] tackle online learning for a generalization of repeated games called *stochastic games*. However, they consider zero-sum games where the sum of the rewards of both players for any joint-action is always 0. In our case, we look at the general sum case where no such restrictions are placed on the rewards.

Our work is also related to multi-objective multi-armed bandits [14] by considering the joint-actions as arms controlled by a single player. Typical work on multi-objective multi-armed bandits tries to find any solution that is as close as possible to the Pareto frontier. However, not all Pareto efficient solutions are acceptable as illustrated by Example 1.1. Instead, our work shows that a specific Pareto efficient solution (the EBS) is more desirable.

Contributions. In this work,

- We strengthen the Individual Rationality criterion [31] that agents in a multi-agent system should aim for. We do this by requiring the agents to be close to the *unique* solution of a Bargaining Problem with disagreement the maximin of both players.
- We propose using the EBS due to its connection to fairness, justice and equality contrarily to other Bargaining solutions.
- We show that the EBS can be achieved by a stationary policy that has non-zero probability on at most two joint-actions (Proposition 4.1). We also show that this EBS policy gives an equal amount above the maximin (called advantage) of both players except in degenerate cases where one player is already receiving its maximum advantage (Proposition 4.2).
- We present a learning algorithm UCRG (Upper Confidence for Repeated Games) that can achieve the EBS in self-play for two player multi-armed bandit problems with stochastic rewards from a distribution unknown to both players. We derived a high probability upper bound of $\tilde{O}\left(T^{2/3}\right)$[2] for UCRG's regret (the difference between the value it achieved and that of an optimal EBS policy) after any number of T rounds. Importantly, our upper bound holds individually for both players and for an unknown T. Also, our bound is not asymptotic and holds for any finite total number of rounds T (Theorem 4.3).
- We derive a lower bound on the regret for any learning algorithm by giving an example game in which any algorithm would have to suffer $\Omega\left(T^{2/3}\right)$ regret demonstrating that our upper bound is optimal up to poly-logarithmic factors (Theorem 4.5).
- We present an exact polynomial-time algorithm that can compute an EBS for a game with known deterministic rewards in Equation (6).
- We perform experiments that validate our theoretical bounds and show our approach achieves a smaller regret compared to a well-know heuristic (Section 6).

Paper organization. The paper is organized as follows: Section 2 presents formally our setting, assumptions, as well as key definitions needed to understand the remainder of the paper. Section 3 shows a description of our algorithm while section 4 contains its analysis as well as the lower bound. We conclude in section 7 with an indication about future works.

Example 1.1 (Comparison of the EBS value to other concepts). In Table 1, we present a game and give the values achieved by the single-stage *NE*, and Correlated Equilibrium [15] (*Correlated*); maximizing the sum of rewards (*Sum*), and a Pareto-efficient solution (*Pareto*). In this game, the maximin value is $(\frac{3}{10}, \frac{3}{10})$. *Sum* plays the pair (C, D) which leads to $\frac{1}{10}$ for the first player, much lower than its maximin. *Pareto* is also similarly problematic. Consequently, it is not enough to converge to any Pareto solution since that does not necessarily guarantee rationality for both players. Both *NE* and *Correlated* fail to give the players a value higher than their maximin while the EBS (computed using Equation (6)) shows that a high value $(\frac{23}{25}, \frac{23}{25})$ is achievable by playing (C, D) and (D, C) with appropriate probabilities. A conclusion similar to this example can also be made for all non-trivial zero-sum games.

	C	D
C	$\frac{4}{5}, \frac{4}{5}$	$\frac{1}{10}, \frac{9}{5}$
D	$\frac{9}{5}, 0$	$\frac{3}{10}, \frac{3}{10}$

(a) Game

Maximin	EBS	NE	Sum	Correlated	Pareto
$\frac{3}{10}, \frac{3}{10}$	$\approx \frac{23}{25}, \frac{23}{25}$	$\frac{3}{10}, \frac{3}{10}$	$\frac{1}{10}, \frac{9}{5}$	$\frac{3}{10}, \frac{3}{10}$	$\frac{9}{5}, 0$

(b) Comparison of solutions

Table 1: Comparison of the EBS to other concepts

[2] We used $\tilde{O}$ to hide logarithmic factors.

2 BACKGROUND AND PROBLEM STATEMENT

We focus on two-player multi-armed bandit problems. At round t, both players select and play a joint action $a_t = (a_t^i, a_t^{-i})$ from a finite set $\mathcal{A} = \mathcal{A}^i \times \mathcal{A}^{-i}$. Then, they receive rewards $(r_t^i, r_t^{-i}) \in [0,1]^2$ generated from a fixed but unknown bounded distribution depending on their joint action. The actions and rewards are then revealed to both players. We assume the first agent to be under our control and the second agent to be the opponent. We would like to design algorithms such that our agent's cumulative rewards are as high as possible. The opponent can have one of two types known to our agent[3]: (1) *self-player* (another independently run version of our algorithm) or (2) arbitrary (i.e any possible opponents with no access to the agent's internal randomness).

To measure performance, we compare our agent to an oracle that has full knowledge of the distribution of rewards for all joint-actions. The oracle then plays like this: (1) in self-play, both agents compute before the game starts the egalitarian bargaining solution and play it; (2) against any other arbitrary opponent, the oracle plays the policy ensuring the maximin value.

Our goal is to design algorithms that have low expected regret against this oracle after any number of T rounds, where regret is the difference between the value that the oracle would have obtained and the value that our algorithm actually obtained. Next, we formally define the terms that describe our problem setting.

Definition 2.1 (Policy). A policy π^i in a repeated game for player i is a mapping from each possible history to a distribution over actions. That is: $\forall t \geq 0, \pi^i : \mathcal{H}_t \to \Delta\mathcal{A}^i$ where t is the current round and $\mathcal{H}_t$ is the set of all possible histories of joint-actions up to round t.

A policy is called *stationary* if it plays the same distribution in each round. It is called *deterministic stationary* if it plays the same action in each round.

Definition 2.2 (Joint-Policy). A joint policy (π^i, π^{-i}) is a pair of policies, one for each player $i, -i$ in the game. In particular, this means that the probability distributions over the actions of both players are independent. When each component policy is *stationary*, we call the resulting joint policy *stationary* and similarly for *deterministic stationary*.

Definition 2.3 (Correlated-Policy). Any joint-policy where player actions are not independent is *correlated*[4]. A correlated policy π specifies a probability distribution over joint-actions known by both players: $\forall t \geq 0, \pi : \mathcal{H}_t \to \Delta\mathcal{A}$.

In this paper, when we refer to a policy π without any qualifier, we will mean a correlated-policy, which is required for the egalitarian solution. When we refer to π^i and (π^i, π^{-i}) we will mean the components of a non-correlated joint-policy.

[3]Our work is trivially extended to unknown type by checking if the opponent is *self*.
[4]For example through a public signal.

2.1 Solution concepts

In this section, we explain the two solution concepts we aim to address: safety–selected as the maximin value and individual rationality selected as achieving the value of the EBS. We start from the definition of value for a policy.

Definition 2.4 (Value of a policy). The value $V^i(\pi)$ of a policy π for player i in a repeated game M is defined as the infinite horizon undiscounted expected average reward given by:

$$V_M^i(\pi) = \limsup_{T\to\infty} \frac{1}{T}\,\mathbb{E}\left(\sum_{t=1}^{T} r_t^i \;\middle|\; \pi, M\right).$$

We use $V_M = (V_M^i, V_M^{-i})$ to denote values for both players and drop M when clear from the context.

Definition 2.5 (Maximin value). The maximin policy π_{SV}^i for player i and its value SV^i are such that:

$$\pi_{\mathrm{SV}}^i = \operatorname*{argmax}_{\pi^i} \min_{\pi^{-i}} V^i(\pi^i, \pi^{-i}), \qquad \mathrm{SV}^i = \max_{\pi^i} \min_{\pi^{-i}} V^i(\pi^i, \pi^{-i}).$$

where $V^i(\pi^i, \pi^{-i})$ is the value for player i playing policy π^i while all other players play π^{-i}.

Definition 2.6 (Advantage game and Advantage value). Consider a repeated game between two players i and $-i$ defined by the joint-actions $\mathcal{A} = \mathcal{A}^i \times \mathcal{A}^{-i}$ and the random rewards r drawn from a distribution $R : \mathcal{A} \to \Delta\mathbb{R}^2$. Let $\mathrm{SV} = (\mathrm{SV}^i, \mathrm{SV}^{-i})$ be the maximin value of the two players. The *advantage game* is the game with (random) rewards r_+ obtained by subtracting the maximin value of the players from r. More precisely, the advantage game is defined by: $r_+(a) = r(a) - \mathrm{SV}\ \ \forall a \in \mathcal{A}$. The value of any policy in this advantage game is called advantage value.

Definition 2.7 (EBS in repeated games). Consider a repeated game between two players i and $-i$ with maximin value $\mathrm{SV} = (\mathrm{SV}^i, \mathrm{SV}^{-i})$. A policy π_{Eg} is an EBS if it satisfies the following two conditions: (1) it belongs to the set Π_{Eg} of policies maximizing the minimum of the advantage value for both players. (2) it maximizes the value of the player with the highest advantage value.

More formally, for any vector $x = (x^1, x^2) \in \mathbb{R}^2$, let $L : \mathbb{R}^2 \to \mathbb{R}^2$ be a permutation of x such that $L^1(x) \leq L^2(x)$. For any $x \in \mathcal{R}^2, y \in \mathcal{R}^2$ let's define a lexicographic maximin ordering $\geq_\ell$ on $\mathbb{R}^2$ as:

$$x \geq_\ell y \iff \Big(L^1(x) > L^1(y)\Big) \vee \Big(L^1(x) = L^1(y) \wedge L^2(x) \geq L^2(y)\Big).$$

A policy π_{Eg} is an EBS [5] if: $V(\pi_{\mathrm{Eg}}) - \mathrm{SV} \geq_\ell V(\pi) - \mathrm{SV}\ \forall \pi$.

We call EBS value the value $V_{\mathrm{Eg}} = V(\pi_{\mathrm{Eg}})$ and $V_+(\pi_{\mathrm{Eg}}) = V(\pi_{\mathrm{Eg}}) - \mathrm{SV}$ will be used to designate the egalitarian advantage.

2.2 Performance criteria

We can now define precisely the two criteria we aim to optimize.

Definition 2.8 (Safety Regret). The safety regret for an algorithm Λ playing for T rounds as agent i against an arbitrary opponent π^{-i} with no knowledge of the internal randomness of Λ is defined by:

$$\mathrm{Regret}_T(\Lambda, \pi^{-i}) = \sum_{t=1}^{T} \mathrm{SV}^i - r_t^i.$$

[5]Also corresponds to the leximin solution to the Bargaining problem [4].

Definition 2.9 (Individual Rational Regret). The individual rational regret for an algorithm Λ playing for T rounds as agent i against itself Λ' identified as $-i$ is defined by:

$$\text{Regret}_T(\Lambda, \Lambda') = \max\left\{\sum_{t=1}^{T} V^i_{\text{Eg}} - r^i_t, \sum_{t=1}^{T} V^{-i}_{\text{Eg}} - r^{-i}_t\right\}.$$

3 METHODS DESCRIPTION

Generic structure. Before we detail the safe and individual rational algorithms, we will describe their general structure. The key challenge is how to deal with uncertainty, the fact that we do not know the rewards. To deal with this uncertainty, we use the standard principle of *optimism in the face of uncertainty* [20]. It works by **a)** constructing a set of *statistically plausible games* containing the true game with high probability through a confidence region around estimated mean rewards, a step detailed in section 3.1; **b)** finding within that set of plausible games, the one whose EBS policy (called *optimistic*) has the highest value, a step detailed in section 3.2; **c)** playing this optimistic policy until the start of an artificial *epoch* where a new epoch starts when the number of times any joint-action has been played is doubled (also known as the *doubling trick*), a step described in Jaksch et al. [20] and summarized by Algorithm 1.

3.1 Construction of the plausible set

At epoch k, our construction is based on creating a set $\mathcal{M}_k$ containing all possible games with expected rewards $\mathbb{E}\, r$ such that,

$$\mathcal{M}_k = \{r : |\mathbb{E}\, r^i(a) - \bar{r}^i_k(a)| \leq C_k(a) \;\&\; \mathbb{E}\, r^i(a) \leq 1 \;\; \forall i, a\} \quad (1)$$

$$C_k(a) = \sqrt{\frac{\ln 1/\delta_k}{1.99 N_{t_k}(a)}}.$$

where t_k is the number of rounds played up to episode k, $N_{t_k}(a)$ is the number of times action a has been played up to round t_k, $\bar{r}_k(a)$ is the empirical mean reward observed up to round t_k and δ_k is an adjustable probability. The plausible set can be used to define the following upper and lower bounds on the rewards of the game:

$$\hat{r}^i_k(a) = \bar{r}^i_k(a) + C_k(a), \qquad \check{r}^i_k(a) = \bar{r}^i_k(a) - C_k(a).$$

We denote $\hat{M}$ the game with rewards $\hat{r}$ and $\check{M}$ the game with $\check{r}$. Values in those two games are resp. denoted $\hat{V}$, $\check{V}$. We used $C_k(\pi)$, $C_k(\pi^i, \pi^{-i})$ to refer to the bounds obtained by a weighted (using π) average of the bounds for individual action. When clear from context, the subscript k is dropped.

3.2 Optimistic EBS policy

3.2.1 Problem formulation. Our goal is to find a game $\tilde{M}_k$ and a policy $\tilde{\pi}_k$ whose EBS value is near-optimal simultaneously for both players. In particular, if we refer to the true but unknown game by M and assume that $M \in \mathcal{M}_k$ we want to find $\tilde{M}_k$ and $\tilde{\pi}_k$ such that:

$$V_{\tilde{M}_k}(\tilde{\pi}_k) \geq_\ell V_{M'}(\pi')$$
$$\forall \pi', M' \in \mathcal{M}_k \mid \Pr\left\{V_{M'}(\pi') \geq V_M(\pi_{\text{Eg}}) - (\epsilon_k, \epsilon_k)\right\} = 1 \quad (2)$$

where $\geq_\ell$ is defined in Definition 2.7 and ϵ_k is a small configurable error.

Note that the condition in (2) is required (contrarily to single-agent games [20]) since in general, there might not exist a game in $\mathcal{M}_k$ that achieves the highest EBS value simultaneously for both players. For example, one can construct a case where the plausible set contains two games with EBS value (resp) $(\frac{1}{2} + \epsilon, \frac{1}{2} + \epsilon)$ and $(\frac{1}{2}, 1)$ for any $0 < \epsilon < 1$ (See Table 2). This makes the optimization problem (2) significantly more challenging than for single-agent games since a small ϵ error in the rewards can lead to a large (linear) regret for one of the players. This is also the root cause for why the best possible regret becomes $\Omega(T^{2/3})$ rather than $\Omega(\sqrt{T})$ typical for single-agent games. We refer to this challenge as the *small ϵ-error large regret* issue.

3.2.2 Solution. To solve (2), **a)** we set the optimistic game $\tilde{M}_k$ as the game $\hat{M}$ in $\mathcal{M}_k$ with the highest rewards $\hat{r}$ for both players. Indeed, for any policy π' and game $M' \in \mathcal{M}_k$, one can always get a better value for both players by using $\hat{M}$; **b)** we compute an advantage game corresponding to $\tilde{M}_k$ by estimating an optimistic maximin value for both players, a step detailed in paragraph 3.2.3; **c)** we compute in paragraph 3.2.4 an EBS policy $\tilde{\pi}_{k,\text{Eg}}$ using the advantage game; **d)** we set the policy $\tilde{\pi}_k$ to be $\tilde{\pi}_{k,\text{Eg}}$ unless one of three conditions explained in paragraph 3.2.5 happens. Algorithm 2 details the steps to compute $\tilde{\pi}_k$ and to correlate the policy, players play the joint-action minimizing their observed frequency of played actions compared to $\tilde{\pi}_k$ (See function Play() of Algorithm 1).

3.2.3 Optimistic Maximin Computation. Satisfying (2) implies finding a value $\check{\text{SV}}$ with:

$$\text{SV}^i - \epsilon_k \leq \check{\text{SV}}^i \leq \text{SV}^i + \epsilon_k \quad \forall i \quad (3)$$

where SV^i is the maximin value of player i in the true game M. To do so, we return a lower bound value for the optimistic maximin policy $\hat{\pi}^i_{\text{SV}_k}$ of player i. We begin by computing in polynomial time[6] the (stationary) maximin policy for the game $\hat{M}$ with the largest rewards. We then compute the (deterministic, stationary) best response policy $\check{\pi}^{-i}_{\hat{\text{SV}}}$ using the game $\check{M}$ with the lowest rewards. The detailed steps are available in Algorithm 3. This results in a lower bound on the maximin value satisfying (3) as proven in Lemma 5.1.

3.2.4 Computing an EBS policy. Armed with the optimistic game and the optimistic maximin value, we can now easily compute the corresponding optimistic *advantage game* whose rewards are denoted by $\hat{r}_+$. An EBS policy $\tilde{\pi}_{k,\text{Eg}}$ is computed using this *advantage game*. The key insight to do so is that the EBS involves playing a single deterministic stationary policy or combining two deterministic stationary policies (Proposition 4.1). Given that the number of actions is finite we can then just loop through each pair of joint-actions and check which one gives the best EBS score. The score (justified by the proof of Proposition 4.2) to use for any two joint-actions a and a' is: $\text{score}(a, a') = \min_{i \in \{1,2\}} w(a, a') \cdot \hat{r}^i_+(a) + (1 - w(a, a')) \cdot r^i_+(a')$ with w as follows:

$$w(a, a') = \begin{cases} 0, & \text{if } \hat{r}^i_+(a) \leq \hat{r}^{-i}_+(a) \text{ and } \hat{r}^i_+(a') \leq \hat{r}^{-i}_+(a') \\ 1, & \text{if } \hat{r}^i_+(a) \geq \hat{r}^{-i}_+(a) \text{ and } \hat{r}^i_+(a') \geq \hat{r}^{-i}_+(a') \\ \frac{\hat{r}^{-i}_+(a') - \hat{r}^i_+(a')}{(\hat{r}^i_+(a) - \hat{r}^i_+(a')) + (\hat{r}^{-i}_+(a') - \hat{r}^{-i}_+(a))} & \text{otherwise .} \end{cases} \quad (4)$$

[6] For example by using linear programming [1, 13].

And the policy $\tilde{\pi}_{k,\text{Eg}}$ is such that

$$\tilde{\pi}_{k,\text{Eg}}(a_{\text{Eg}}) = w(a_{\text{Eg}}, a'_{\text{Eg}}); \quad \tilde{\pi}_{k,\text{Eg}}(a'_{\text{Eg}}) = 1 - w(a_{\text{Eg}}, a'_{\text{Eg}}); \quad (5)$$

$$a_{\text{Eg}}, a'_{\text{Eg}} = \underset{a \in \mathcal{A},\, a' \in \mathcal{A}}{\text{argmax}} \ \text{score}(a, a'). \quad (6)$$

3.2.5 Policy Execution. We always play the optimistic EBS policy $\tilde{\pi}_{k,\text{Eg}}$ unless one of the following three events happens:

- *The probable error on the maximin value of one player is too large.* Indeed, the error on the maximin value can become too large if the weighted bound on the actions played by the maximin policies is too large. In that case, we play the action causing the largest error.
- *The* small ϵ-error large regret *issue is probable*: Proposition 4.2 implies that the *small ϵ-error large regret* issue may only happen if the player with the lowest ideal advantage value (the maximum advantage under the condition that the advantage of the other player is non-negative) is receiving it when playing an EBS policy. This allows Algorithm 2 to check for this player and play the action corresponding to its ideal advantage as long as the other player is still receiving ϵ_k-close to its EBS value (Line 6 to 15 in Algorithm 2).
- *The probable error on the EBS value of one player is too large* This only happens if we keep not playing the EBS policy due to the *small ϵ-error large regret* issue. In that case, the error on the EBS value used to detect the *small ϵ-error large regret* issue might become too large making the check for the *small ϵ-error large regret* issue irrelevant. In that case, we play the action of the EBS policy responsible for the largest error.

4 THEORETICAL ANALYSIS

Before we present theoretical analysis for the learning algorithm, we discuss the existence and uniqueness of the EBS value, as well as the type of policies that can achieve it.

Properties of the EBS. Fact 1 implies that any (optimal) value achievable can be achieved by a stationary (correlated-) policy; allowing us to restrict our attention to stationary policies. Fact 2 means that the EBS always exists and is unique; providing us with a good benchmark to compare against.

Fact 1 (Achievable values for both players). *Any achievable value $V = (V^i, V^{-i})$ for the players can be achieved by a stationary correlated-policy.*

Sketch. We first show that the value for joint-actions exists and is unique. Then, similarly to [29], we consider the convex hull of the set of values for joint-actions and show that this convex hull corresponds exactly to the set of all achievable values. Since we can achieve any point of the convex hull with a stationary policy, this concludes the proof. □

Fact 2 (Existence and Uniqueness of the EBS value for stationary policies). *If we are restricted to the set of stationary policies, then the EBS value defined in Definition 2.7 exists and is unique.*

Sketch. [19] already proved that the egalitarian value as defined in Definition 2.7 always exists and is unique for any bargaining problem that is convex, closed, of non-empty Pareto frontier and non-degenerate. We then proved this fact, by showing that the repeated game we consider implies a bargaining problem satisfying those properties. □

Algorithm 1 UCRG

Definitions: $N_k(a)$ denotes the number of rounds action a has been played in episode $k - N_k$ the number of rounds episode k has lasted $- t_k$ the number of rounds played up to episode $k - N_{t_k}(a)$ the number of rounds action a has been played up to round $t_k - \bar{r}^i_k(a)$ the empirical average rewards of player i for action a up to round t_k.
Input: For each episode k, we are given two numbers ϵ_k and δ_k.
Initialization: Let $t \leftarrow 1$. Set $N_k, N_k(a), N_{t_k}(a)$ to zero for all $a \in \mathcal{A}$.

for episodes $k = 1, 2, \ldots$ **do**
  $t_k \leftarrow t$
  $N_{t_{k+1}}(a) \leftarrow N_{t_k}(a) \quad \forall a$
  $\hat{r}^i_k(a) = \bar{r}^i_k(a) + C_k(a), \qquad \check{r}^i_k(a) = \bar{r}^i_k(a) - C_k(a) \quad \forall a, i$ with C_k computed using δ_k as in (1).
  $\tilde{\pi}_k \leftarrow$ OptimisticEgalitarianPolicy$(\bar{r}_t, \hat{r}_k, \check{r}_k, \epsilon_k)$

  Execute policy $\tilde{\pi}_k$:
  do
    Let $a_t \leftarrow$ Play$(\tilde{\pi}_k)$, play it and observe r_t
    $N_k \leftarrow N_k + 1 \quad N_k(a_t) \leftarrow N_k(a_t) + 1$
    $N_{t_{k+1}}(a_t) \leftarrow N_{t_{k+1}}(a_t) + 1$ and Update $\bar{r}_k(a_t)$
    $t \leftarrow t + 1$
  while $N_k(a_t) \leq \max\{1, N_{t_k}(a)\}$
end for

function Play(π)
  Let a_t the action a that minimizes $\left|\pi(a) - \frac{N_k(a)}{N_k}\right|$
  Ties are broken in favor of the player with the lowest, then in favor of the lexicographically smallest action.
return a_t
end function

The following Proposition 4.1 strengthens the observation in Fact 1 and establishes that a weighted combination of at most two joint-actions can achieve the EBS value. This allows for an efficient algorithm that can just loop through all possible pairs of joint-actions and check for the best one. However, given any two joint-actions one still needs to know how to combine them to get an EBS value. This question is answered by proposition 4.2.

Proposition 4.1 (On the form of an EBS policy). *Given any two-player repeated game, the EBS value can always be achieved by a stationary policy with non-zero probability on at most two joint-actions.*

Sketch. We follow the same line of reasoning used in [26] by showing that the EBS value lies on the outer boundary of the convex hull introduced in the proof of Fact 1. □

Algorithm 2 Optimistic EBS Policy Computation

function OptimisticEgalitarianPolicy($\bar{r}, \hat{r}, \check{r}, \epsilon_k$)
  $\hat{\pi}^i_{SV_k}, \check{\pi}^{-i}_{S\hat{V}_k}, S\check{V}^i_k$ = OptMaximin($\bar{r}, \hat{r}, \check{r}, i$)
  Let $\hat{r}^i_+(a) = \hat{r}^i(a) - S\check{V}^i_k \ \ \forall (i,a)$
  Compute the EBS policy $\hat{\pi}_{k,\text{Eg}}$ using (6) and $\hat{r}^i_+$
  Let $\hat{\pi}_k \leftarrow \hat{\pi}_{k,\text{Eg}}$
  ($\forall i$, from the set of actions with positive advantage, ϵ_k-close to the EBS advantage of $-i$, find the one maximizing i advantage)

$$\tilde{\mathcal{A}}_i = \{a \mid \hat{r}^i_+(a) + \epsilon_k \geq \hat{V}^i_+(\hat{\pi}_{k,\text{Eg}}) \wedge \hat{r}^i_+(a) \geq 0\} \quad \forall i \in \{1,2\}$$
$$\hat{a}_i = \underset{a \in \tilde{\mathcal{A}}_{-i}}{\text{argmax}}\, \hat{r}^i_+(a) \quad \forall i \in \{1,2\}$$

  (Look for the player i whose advantage for action $\hat{a}_i$ is larger than the EBS advantage of i)

$$\tilde{\mathcal{P}} = \{i \in \{1,2\} \mid \hat{r}^i_+(\hat{a}_i) > \hat{V}^i_+(\hat{\pi}_{k,\text{Eg}})\}$$

  (If there is a player i whose advantage for $\hat{a}_i$ is better than the EBS advantage, play $\hat{a}_i$)
  if $\tilde{\mathcal{P}} \neq \varnothing$ **then**
    $\tilde{p} = \text{argmax}_{i \in \tilde{\mathcal{P}}}\, \hat{r}^i_+(\hat{a}_i)\,, \quad \hat{\pi}_k \leftarrow \hat{a}_{\tilde{p}}$
  end if
  (If potential error on the EBS value is too large, play the responsible action.)
  if $2C_k(\hat{\pi}_{k,\text{Eg}}) > \epsilon_k$ **then**
    Let $\hat{a}_{k,\text{Eg}} = \text{argmax}_{a \in \mathcal{A} | 2C_k(a) > \epsilon_k}\, \hat{\pi}_{k,\text{Eg}}(a)$
    $\hat{\pi}_k \leftarrow \hat{a}_{k,\text{Eg}}$
  end if
  (If potential error on the maximin value is too large, play the responsible action.)
  for each $i \in \{1,2\}$ **where** $2C_k(\hat{\pi}^i_{SV_k}, \check{\pi}^{-i}_{S\hat{V}_k}) > \epsilon_k$ **do**
    Let $\hat{a}_{SV_k} = \text{argmax}_{a \in \mathcal{A} | 2C_k(a) > \epsilon_k}\, \hat{\pi}^i_{SV_k}(a) \cdot \check{\pi}^{-i}_{S\hat{V}_k}(a)$
    $\hat{\pi}_k \leftarrow \hat{a}_{SV_k}$
  end for
  return $\hat{\pi}_k$
end function

Proposition 4.2 (Finding an EBS policy). *Let us call the ideal advantage value V^i_{+I} of a player i, the maximum advantage that this player can achieve under the restriction that the advantage value of the other player is non-negative. More formally: $V^i_{+I} = \max_{\pi | V^{-i}_+(\pi) \geq 0} V^i_+(\pi)$. The egalitarian advantage value for the two players is exactly the same unless there exists an EBS policy that is deterministic stationary where at least one player (necessarily including the player with the lowest ideal advantage value) is receiving its ideal advantage value.*

Proof. From proposition 4.1 we can achieve the EBS value by combining at most two deterministic stationary policies. We will prove this proposition (4.2) for any two possible deterministic stationary policies (by considering a repeated game with only the corresponding joint-actions available), which immediately means that Proposition 4.2 is also true for the EBS value in the full repeated game.

Consider any two deterministic stationary policy of advantage values $((x^1_1, x^2_1), (x^1_2, x^2_2))$. We will now show how to compute the weight $w = \text{argmax}_w \min_{i \in \{1,2\}} w * x^i_1 + (1-w)x^i_2$.

Case 1: $x^1_1 \leq x^2_1$ and $x^1_2 \leq x^2_2$. This basically means that the advantage value of player 2 is always higher or equal than that of the player 1. So the minimum is maximized by playing the policy maximizing the value of player 1. So, $w = 0$ and we have a single deterministic stationary policy where the player with the lowest ideal advantage receives it.

Case 2: $x^1_1 \geq x^2_1$ and $x^1_2 \geq x^2_2$. This is essentially *Case 1* with the role of players 1 and 2 exchanged. Here $w = 1$.

If both *Case 1* and *Case 2* do not hold, it means that for the first policy, one player receives an advantage value strictly greater than that of the other player while the situation is reversed for the second policy. Without loss of generality, we can assume this player is 1 (if this is not the case, we can simply switch the id of the policy) which leads to *Case 3*.

Case 3: $x^1_1 > x^2_1$ and $x^1_2 < x^2_2$ In this case, the optimal w is such that $w = \frac{x^2_2 - x^1_2}{(x^1_1 - x^1_2) + (x^2_2 - x^2_1)}$. This weight w is clearly between the open interval $]0, 1[$. This means that we have exactly two distinct policies. Plugging in the weight shows that the advantage value of both players is the same, which completes the proof. □

Algorithm 3 Optimistic Maximin Policy Computation

function OptMaximin($\bar{r}, \hat{r}, \check{r}, i$)
  Calculate i's optimistic maximin policy: $\hat{\pi}^i_{SV_k} = \text{argmax}_{\pi^i} \min_{\pi^{-i}} \hat{V}^i(\pi^i, \pi^{-i})$
  Find the best response: $\check{\pi}^{-i}_{S\hat{V}_k} = \text{argmin}_{\pi^{-i}} \check{V}^i(\hat{\pi}^i_{SV_k}, \pi^{-i})$
  Get a lower bound on the maximin value: $S\check{V}^i_k = \min_{\pi^{-i}} \check{V}^i(\hat{\pi}^i_{SV_k}, \pi^{-i}) = \check{V}^i(\hat{\pi}^i_{SV_k}, \check{\pi}^{-i}_{S\hat{V}_k})$
  return $\hat{\pi}^i_{SV_k}, \check{\pi}^{-i}_{S\hat{V}_k}, S\check{V}^i_k$
end function

Regret Analysis. The following theorem 4.3 gives us a high probability upper bound on the regret in self-play against the EBS value, a result achieved without the knowledge of T.

Theorem 4.3 (Individual Rational Regret for Algorithm 1 in self-play). *After running Algorithm 1 in self play with $\delta_k = \delta / B_{t_k}$ and $\epsilon_k = (2|\mathcal{A}| \ln(1/\delta_k)/(1.99 t_k))^{1/3}$ where $B_t = 16|\mathcal{A}| \ln t + 2|\mathcal{A}| + 1$ for any rounds $T \geq |\mathcal{A}|$, with probability at least $1 - \delta$, $\delta > 0$, the individual rational regret (definition 2.9) for each player*

is upper bounded as:

$$\begin{aligned} Regret_T &\leq 5\sqrt[3]{|\mathcal{A}|\ln(B_T/\delta)}\cdot T^{2/3} + 2|\mathcal{A}|\log_2(8T/|\mathcal{A}|) \\ &\quad + \sqrt{T\ln(B_T/\delta)}\cdot\left(\sqrt{1/2}+\sqrt{12|\mathcal{A}|}\right)+\sqrt{T} \\ &= O\left(5\sqrt[3]{|\mathcal{A}|\ln(\ln(T)/\delta)}\cdot T^{2/3}\right). \end{aligned}$$

Sketch. The structure of the proof follows that of [20]. More precisely, as the algorithm is divided into epochs, we first show that the regret bound within an epoch is sub-linear. We then combine those per-epoch regret terms to get a regret for the whole horizon simultaneously. Both of these regrets are computed with the assumption that the true game M is within our plausible set. We then conclude by showing that this is indeed true with high probability.

To prove the regret within an epoch, the key step is to prove that the value of policy $\tilde{\pi}_k$ returned by Algorithm 2 in our plausible set is ϵ-close to the EBS value in the true model (optimism). In our case, we cannot always guarantee this optimism. Our proof identifies the concerned cases and shows that they cannot happen too often. Then for the remaining cases, we show that we can guarantee the optimism with an error of $4\epsilon_{t_k}$: the combination of Lemma 5.2 and Lemma 5.1 is crucial for this. □

By definition of EBS, Theorem 4.3 also applies to the safety regret. However, in Theorem 4.4, we show that the optimistic maximin policy enjoys near-optimal safety regret of $O\left(\sqrt{|\mathcal{A}^i|T\ln(\ln(T)/\delta)}\right)$.

Theorem 4.4 (Safety Regret of policy $\hat{\pi}^i_{\mathrm{SV}_k}$ from Algorithm 3). *Assume that in Algorithm 1, player i executes policy $\hat{\pi}^i_{\mathrm{SV}_k}$ (as computed by Algorithm 3) instead of $\tilde{\pi}_k$ with $\delta_k = \frac{\delta}{16|\mathcal{A}^i|\ln t_k + 2|\mathcal{A}^i|+1}$ while replacing any computation on joint-action a by an equivalent computation on single-action a^i. After any rounds $T \geq |\mathcal{A}^i|$ against any opponent, then with probability at least $1-\delta$, $\delta > 0$, the safety regret (definition 2.8) of this policy is upper-bounded by:*

$$Regret_T \leq \sqrt{\frac{T}{2}\ln\left(\frac{16|\mathcal{A}^i|\ln(1.3T)}{\delta}\right)}\cdot\left(4+\sqrt{24|\mathcal{A}^i|}\right)+\sqrt{T}.$$

Sketch. The proof works similarly to that of Theorem 4.3 by observing that here we can always guarantee optimism when the true game M is within our plausible set. Indeed, for any opponent policy π_o^{-i}, we have: $\hat{\pi}^i_{\mathrm{SV}_k} = \mathrm{argmax}_{\pi^i}\min_{\pi^{-i}}\hat{V}^i(\pi^i,\pi^{-i})$ and $\hat{V}(\hat{\pi}^i_{\mathrm{SV}_k},\pi_o^{-i}) \geq \max_{\pi^i}\min_{\pi^{-i}}\hat{V}^i(\pi^i,\pi^{-i}) \geq \max_{\pi^i}\min_{\pi^{-i}}V^i(\pi^i,\pi^{-i}) = \mathrm{SV}^i$. □

Lower bounds for the individual rational regret. Here we establish a lower bound of $\Omega\left(T^{2/3}\right)$ for any algorithm trying to learn the EBS value. This shows that our upper bound is optimal up to logarithm-factors. The key idea in proving this lower bound is the example illustrated in Table 2. In that example, the rewards of the first player are all $\frac{1}{2}$ and the second player has an ideal value of 1. However, 50% of the time, a player cannot realize its ideal value due to an ϵ-increase in a single joint-action for both players. The main intuition behind the proof of the lower bound is that any algorithm that wants to minimize regret can only try two things **(a)** detect whether there exists a joint-action with an ϵ or if all rewards of the first player are equal, or **(b)** always ensure the ideal value of the second player. To achieve **(a)** any algorithm needs to play all joint-actions for $\frac{1}{\epsilon^2}$ times. Picking $\epsilon = T^{-1/3}$ ensures the desired lower bound. The same ϵ would also ensure the same lower bound for an algorithm targeting only **(b)**.

	a_1^2	a_2^2	$\cdots$	$a^2_{\lvert\mathcal{A}^2\rvert}$
a_1^1	(0.5, 1)	(0.5, 0.5)	$\cdots$	(0.5, 0.5)
a_2^1	(0.5[+ϵ], 0.5[+ϵ])	(0.5, 0.5)	$\cdots$	(0.5, 0.5)
$\vdots$	$\vdots$	$\vdots$	$\cdots$	(0.5, 0.5)
$a^1_{\lvert\mathcal{A}^1\rvert}$	(0.5, 0.5)	(0.5, 0.5)	$\cdots$	(0.5, 0.5)

Table 2: Lower bounds example. The rewards are generated from a Bernoulli distribution whose parameter is specified in the table. The first value in parentheses is the one for the first player while the other is for the second player. Here, ϵ is a small constant defined in the proof.

Theorem 4.5 (Lower bounds). *For any algorithm Λ, any natural numbers $|\mathcal{A}^1| \geq 2$, $|\mathcal{A}^2| \geq 2$, $T \geq |\mathcal{A}^1|\times|\mathcal{A}^2|$, there is a general sum game with $|\mathcal{A}| = |\mathcal{A}^1|\times|\mathcal{A}^2|$ joint-actions such that the expected individual rational regret of Λ after T steps is at least $\Omega\left(T^{2/3}\frac{|\mathcal{A}|^{1/3}}{4}\right)$.*

Proof Sketch. The proof is inspired by the one for bandits in Theorem 6.11 of [10]. We used our game in Table 1 and then compute the optimal egalitarian solution for this game based on the possible values of ϵ. □

5 TECHNICAL LEMMAS

Lemma 5.1 (Pessimism and Optimism of the maximin value). *For any player i and epoch k for which the true model M is within our* plausible set $\mathcal{M}_k$, *the maximin value computed satisfies:*

$$\mathrm{SV}^i - 2C_k(\hat{\pi}^i_{\mathrm{SV}_k}, \check{\pi}^{-i}_{\mathrm{S}\hat{V}_k}) \leq \mathrm{S}\check{V}^i_k \leq \mathrm{SV}^i.$$

Proof. By definition (See Algorithm 3), we have:

$$\hat{\pi}^i_{\mathrm{SV}_k} = \operatorname*{argmax}_{\pi^i}\min_{\pi^{-i}}\hat{V}^i(\pi^i,\pi^{-i}), \tag{7}$$

$$\check{\pi}^{-i}_{\mathrm{S}\hat{V}_k} = \operatorname*{argmin}_{\pi^{-i}}\check{V}^i(\hat{\pi}^i_{\mathrm{SV}_k},\pi^{-i}), \tag{8}$$

$$\mathrm{S}\check{V}^i_k = \min_{\pi^{-i}}\check{V}^i(\hat{\pi}^i_{\mathrm{SV}_k},\pi^{-i}) = \check{V}^i(\hat{\pi}^i_{\mathrm{SV}_k},\check{\pi}^{-i}_{\mathrm{S}\hat{V}_k}). \tag{9}$$

Pessimism of the maximin value. We have:

$$\mathrm{SV}^i = \max_{\pi^i}\min_{\pi^{-i}}V^i(\pi^i,\pi^{-i}) \geq \min_{\pi^{-i}}V^i(\hat{\pi}^i_{\mathrm{SV}_k},\pi^{-i}) \tag{10}$$

$$\geq \min_{\pi^{-i}}\check{V}^i(\hat{\pi}^i_{\mathrm{SV}_k},\pi^{-i}) = \mathrm{S}\check{V}^i_k. \tag{11}$$

Optimism of the maximin value. We have:

$$\mathrm{S}\check{V}^i_k = \check{V}^i(\hat{\pi}^i_{\mathrm{SV}_k},\check{\pi}^{-i}_{\mathrm{S}\hat{V}_k}) \tag{12}$$

$$= \hat{V}^i(\hat{\pi}^i_{\mathrm{SV}_k},\check{\pi}^{-i}_{\mathrm{S}\hat{V}_k}) - 2C_k(\hat{\pi}^i_{\mathrm{SV}_k},\check{\pi}^{-i}_{\mathrm{S}\hat{V}_k}) \tag{13}$$

$$\geq \min_{\pi^{-i}}\hat{V}^i(\hat{\pi}^i_{\mathrm{SV}_k},\pi^{-i}) - 2C_k(\hat{\pi}^i_{\mathrm{SV}_k},\check{\pi}^{-i}_{\mathrm{S}\hat{V}_k}) \tag{14}$$

$$= \hat{V}^i(\hat{\pi}^i_{\mathrm{SV}_k},\hat{\pi}^{-i}_{\mathrm{SV}_k}) - 2C_k(\hat{\pi}^i_{\mathrm{SV}_k},\check{\pi}^{-i}_{\mathrm{S}\hat{V}_k}) \tag{15}$$

5 TECHNICAL LEMMAS

$$= \max_{\pi^i} \hat{V}^i(\pi^i, \hat{\pi}^{-i}_{\mathrm{SV}_k}) - 2C_k(\hat{\pi}^i_{\mathrm{SV}_k}, \check{\pi}^{-i}_{\mathrm{S}\hat{V}_k}) \quad (16)$$

$$\geq \max_{\pi^i} V^i(\pi^i, \hat{\pi}^{-i}_{\mathrm{SV}_k}) - 2C_k(\hat{\pi}^i_{\mathrm{SV}_k}, \check{\pi}^{-i}_{\mathrm{S}\hat{V}_k}) \quad (17)$$

$$\geq \max_{\pi^i} \min_{\pi^{-i}} V^i(\pi^i, \pi^{-i}) - 2C_k(\hat{\pi}^i_{\mathrm{SV}_k}, \check{\pi}^{-i}_{\mathrm{S}\hat{V}_k}) \quad (18)$$

$$= \mathrm{SV}^i - 2C_k(\hat{\pi}^i_{\mathrm{SV}_k}, \check{\pi}^{-i}_{\mathrm{S}\hat{V}_k}). \quad (19)$$

□

Lemma 5.2 (Optimism of the advantage game). *This lemma proves that the advantage value for any policy π in our optimistic model is greater than in the true model. For any policy π, player i and epoch k for which the true model M is within our* plausible set *$\mathcal{M}_k$: $\tilde{V}^i_+(\pi) \geq V^i_+(\pi)$.*

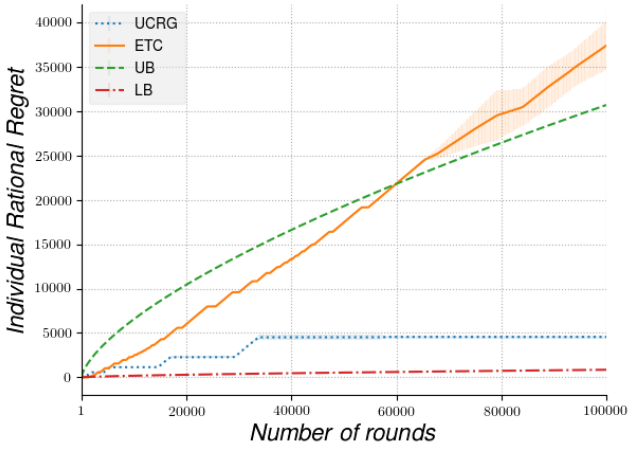
Proof. We have:

$$\tilde{V}^i_+(\pi) = \tilde{V}^i(\pi) - \check{\mathrm{SV}}^i_k \geq V^i(\pi) - \check{\mathrm{SV}}^i_k \geq V^i(\pi) - \mathrm{SV}^i = V^i_+(\pi).$$

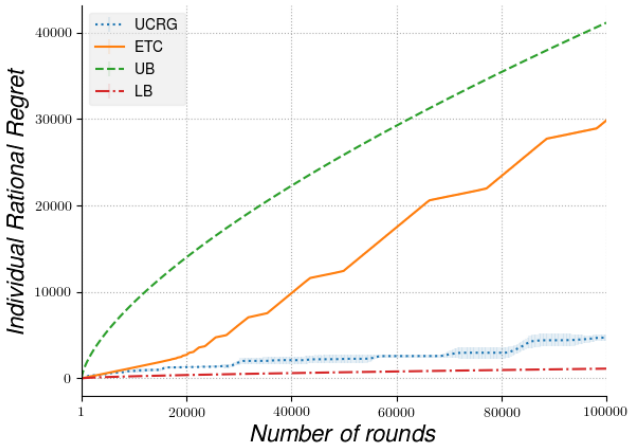
where the second inequality comes from Lemma 5.1. □

6 EXPERIMENTS

We compared our solution UCRG to a heuristic well-known as Explore-Then-Commit (ETC). ETC plays each action for m' rounds. After that, ETC uses the estimated empirical game to compute an EBS policy (as in (6)) which is subsequently played for the remaining rounds. The doubling trick (Algorithm 1) is used to deal with unknown T. Since single-player multi-armed bandits (MAB) are a special case of our setting, we pick the exploration parameter m' of ETC as the minimax optimal value for ETC-like policies in MAB [37]. In particular, we pick $m' = (A \log 1/\delta_k)^{1/3}(2t_k)^{2/3}$ with δ_k as in Theorem 4.3. In the experiments, we also compared our regret with the theoretical lower bound derived in Theorem 4.5 (LB in Figure 1) and the theoretical upper bound derived in Theorem 4.3 (UB in Figure 1). In Figure 1a, we use the worst-case game shown in Table 2 with two actions. In Figure 1b, we use the generalized rock-paper-scissors [34] scaled in $[\frac{1}{4}, \frac{3}{4}]$. The probability of error δ is set to 0.01 and the horizon T to 10^5. Figures 1a and 1b confirm the validity of our theoretical bounds. Figure 1a shows that the naive ETC heuristic can obtain a linear-regret. And Figure 1b shows that even in simpler games, our algorithm UCRG still outperforms ETC.



(a) $M(\epsilon) = \begin{bmatrix} 0.5 & 0.5 \\ 0.5+\epsilon & 0.5 \end{bmatrix}$ $M^i = M(0); M^{-i} = M(0.5)$



(b) $M^i = M^{-i} = \begin{bmatrix} 5/8 & 1/4 & 3/4 \\ 3/4 & 5/8 & 1/4 \\ 1/4 & 3/4 & 5/8 \end{bmatrix}$

Figure 1: Average Individual Rational Regret with standard error using 50 trials in self-play for UCRG, ETC, our lower and upper bound (LB & UB resp). Rewards are drawn from Bernoulli distributions with means as shown by the matrices M.

7 CONCLUSION AND FUTURE DIRECTIONS

In this paper, we illustrated a situation in which typical solutions for self-play in repeated games, such as single-stage equilibria or sum of rewards, are not appropriate. We propose the use of an egalitarian bargaining solution (EBS) which guarantees each player to receive no less than their maximin value. We analyze the properties of EBS for repeated games with stochastic rewards and derive an algorithm that achieves a near-optimal finite-time regret of $\tilde{O}(T^{2/3})$ with high probability. We are able to conclude that the proposed algorithm is near-optimal, since we prove a matching lower bound up to logarithmic-factors. Although our results imply a $\tilde{O}(T^{2/3})$ safety regret (i.e. compared to the maximin value), we also show that a component of our algorithm guarantees the near-optimal $\tilde{O}(\sqrt{T})$ safety regret against arbitrary opponents.

Our work illustrates an interesting property of the EBS which is: it can be achieved with sub-linear regret by two individually rational agents who are uncertain about their utility. We wonder if other solutions to the Bargaining Problem such as the Nash Bargaining Solution or the Kalai–Smorodinsky Solution also admit the same property. Since the EBS can be realised as an equilibrium, another intriguing question is whether one can design an algorithm that converges naturally to the EBS against some well-defined class of opponents. Finally, a natural and interesting future direction for our work is its extension to stateful games such as Markov games.

REFERENCES

[1] Ilan Adler. 2013. The equivalence of linear programs and zero-sum games. *International Journal of Game Theory* 42, 1 (2013), 165–177.

[2] Bikramjit Banerjee and Jing Peng. 2004. Performance bounded reinforcement learning in strategic interactions. In *AAAI*, Vol. 4. 2–7.

[3] Ilai Bistritz and Amir Leshem. 2018. Distributed multi-player bandits-a game of thrones approach. In *Advances in Neural Information Processing Systems*. 7222–7232.

[4] Walter Bossert and Guofu Tan. 1995. An arbitration game and the egalitarian bargaining solution. *Social Choice and Welfare* 12, 1 (1995), 29–41.

[5] Michael Bowling. 2005. Convergence and no-regret in multiagent learning. In *Advances in neural information processing systems*. 209–216.

[6] Michael Bowling and Manuela Veloso. 2001. Convergence of Gradient Dynamics with a Variable Learning Rate. In *In Proceedings of the Eighteenth International Conference on Machine Learning*. 27–34.

[7] Ronen I Brafman and Moshe Tennenholtz. 2002. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research* 3, Oct (2002), 213–231.

[8] Ronen I Brafman and Moshe Tennenholtz. 2003. Efficient learning equilibrium. In *Advances in Neural Information Processing Systems*. 1635–1642.

[9] Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. 2019. Delay and cooperation in nonstochastic bandits. *The Journal of Machine Learning Research* 20, 1 (2019), 613–650.

[10] Nicolo Cesa-Bianchi and Gábor Lugosi. 2006. *Prediction, learning, and games*. Cambridge university press.

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

[12] Jacob W Crandall and Michael A Goodrich. 2011. Learning to compete, coordinate, and cooperate in repeated games using reinforcement learning. *Machine Learning* 82, 3 (2011), 281–314.

[13] George B Dantzig. 1951. A proof of the equivalence of the programming problem and the game problem. *Activity analysis of production and allocation* 13 (1951), 330–338.

[14] Madalina M Drugan and Ann Nowe. 2013. Designing multi-objective multi-armed bandits algorithms: A study. *learning* 8 (2013), 9.

[15] Amy Greenwald and Keith Hall. 2003. Correlated-Q Learning. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning (ICML'03)*. AAAI Press, 242–249.

[16] Eshcar Hillel, Zohar S Karnin, Tomer Koren, Ronny Lempel, and Oren Somekh. 2013. Distributed exploration in multi-armed bandits. In *Advances in Neural Information Processing Systems*. 854–862.

[17] Junling Hu and Michael P Wellman. 2003. Nash Q-learning for general-sum stochastic games. *Journal of machine learning research* 4, Nov (2003), 1039–1069.

[18] Junling Hu, Michael P Wellman, et al. 1998. Multiagent reinforcement learning: theoretical framework and an algorithm.. In *ICML*, Vol. 98. Citeseer, 242–250.

[19] Haruo Imai. 1983. Individual monotonicity and lexicographic maxmin solution. *Econometrica: Journal of the Econometric Society* (1983), 389–401.

[20] Thomas Jaksch, Ronald Ortner, and Peter Auer. 2010. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research* 11, Apr (2010), 1563–1600.

[21] Ehud Kalai. 1977. Proportional solutions to bargaining situations: interpersonal utility comparisons. *Econometrica: Journal of the Econometric Society* (1977), 1623–1630.

[22] Ehud Kalai, Meir Smorodinsky, et al. 1975. Other solutions to Nash's bargaining problem. *Econometrica* 43, 3 (1975), 513–518.

[23] Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. 2016. Distributed cooperative decision-making in multiarmed bandits: Frequentist and Bayesian algorithms. In *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE, 167–172.

[24] Michael L Littman. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*. Elsevier, 157–163.

[25] Michael L Littman. 2001. Friend-or-foe Q-learning in general-sum games. In *ICML*, Vol. 1. 322–328.

[26] Michael L. Littman and Peter Stone. 2003. A Polynomial-time Nash Equilibrium Algorithm for Repeated Games. In *Proceedings of the 4th ACM Conference on Electronic Commerce (EC '03)*. ACM, New York, NY, USA, 48–54. https://doi.org/10.1145/779928.779935

[27] Michael L Littman and Csaba Szepesvári. 1996. A generalized reinforcement-learning model: Convergence and applications. In *ICML*, Vol. 96. 310–318.

[28] Enrique Munoz de Cote and Michael L. Littman. 2008. A Polynomial-time Nash Equilibrium Algorithm for Repeated Stochastic Games. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, Corvallis, Oregon, 419–426.

[29] John F Nash Jr. 1950. The bargaining problem. *Econometrica: Journal of the Econometric Society* (1950), 155–162.

[30] Martin J Osborne and Ariel Rubinstein. 1994. *A course in game theory*.

[31] Rob Powers, Yoav Shoham, and Thuc Vu. 2007. A general criterion and an algorithmic framework for learning in multi-agent systems. *Machine Learning* 67, 1 (01 May 2007), 45–76.

[32] Haiyan Qiao, Jerzy Rozenblit, Ferenc Szidarovszky, and Lizhi Yang. 2006. Multi-agent learning model with bargaining. In *Proceedings of the 2006 winter simulation conference*. IEEE, 934–940.

[33] John Rawls. 1971. *A theory of justice*. Harvard university press.

[34] Yuzuru Sato, Eizo Akiyama, and J Doyne Farmer. 2002. Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences* 99, 7 (2002), 4748–4751.

[35] Shahin Shahrampour, Alexander Rakhlin, and Ali Jadbabaie. 2017. Multi-armed bandits in multi-agent networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2786–2790.

[36] Satinder Singh, Michael Kearns, and Yishay Mansour. 2000. Nash Convergence of Gradient Dynamics in General-Sum Games. In *In Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*. Morgan, 541–548.

[37] Paul N Somerville. 1954. Some problems of optimum sampling. *Biometrika* 41, 3/4 (1954), 420–429.

[38] Jeff L Stimpson and Michael A Goodrich. 2003. Learning to cooperate in a social dilemma: A satisficing approach to bargaining. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. 728–735.

[39] William Thomson. 1981. Nash's bargaining solution and utilitarian choice rules. *Econometrica: Journal of the Econometric Society* (1981), 535–538.

[40] Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. 2017. Online Reinforcement Learning in Stochastic Games. 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.). 4987–4997.

[41] Martin Zinkevich, Amy Greenwald, and Michael L. Littman. 2006. Cyclic Equilibria in Markov Games. In *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. C. Platt (Eds.). MIT Press, 1641–1648.