Bounding Regret in Empirical Games

Steven Jecmen,[1] **Arunesh Sinha,**[2] **Zun Li,**[3] **Long Tran-Thanh**[4]
[1]Carnegie Mellon University, [2]Singapore Management University,
[3]University of Michigan, [4]University of Southampton
sjecmen@cs.cmu.edu, aruneshs@smu.edu.sg, lizun@umich.edu, l.tran-thanh@soton.ac.uk

Abstract

Empirical game-theoretic analysis refers to a set of models and techniques for solving large-scale games. However, there is a lack of a quantitative guarantee about the quality of output approximate Nash equilibria (NE). A natural quantitative guarantee for such an approximate NE is the regret in the game (i.e. the best deviation gain). We formulate this deviation gain computation as a multi-armed bandit problem, with a new optimization goal unlike those studied in prior work. We propose an efficient algorithm Super-Arm UCB (SAUCB) for the problem and a number of variants. We present sample complexity results as well as extensive experiments that show the better performance of SAUCB compared to several baselines.

Introduction

Real-world multi-agent interactions are often immensely complex and unstructured. These real-world problems are simply not amenable to theoretical analysis due to various complexities, such as stochastic and unknown utilities. This has led to the development of the area of empirical (or simulated) games (Wellman 2006; Tuyls et al. 2018), which have been used successfully to model and solve complex multi-agent game interactions in stock markets (Wang, Vorobeychik, and Wellman 2018) and cyber-security problems (Prakash and Wellman 2015). The main characteristic of such games is a simulator, which acts as an oracle for player utility functions, taking as input a strategy profile (the strategy of each player) and returning an observation of the utility each player obtained from that strategy profile in simulation. The simulator, in theory, allows one to fill the full game matrix with expected utilities. However, in practice, calls to the simulator are quite time consuming. Various techniques have been proposed in the literature showing how to use the simulator parsimoniously to compute approximate Nash equilibria (NE) of these complex games using double oracle or its adaptations (Lanctot et al. 2017).

Yet, applications of empirical games either do not provide a guarantee about the quality of the output approximate NE solution or do so in an ad-hoc manner. A natural quality measure is the most profitable deviation gain from the approximate NE, also called the regret in the underlying game. Further, given the cost involved in calling the simulator, the deviation gain computation must also make as few calls to the simulator as possible. Thus, there is a need to design principled methods to compute the regret of an approximate NE strategy profile in an empirical game.

Our *first contribution* is to set up the regret computation in empirical games as a stochastic multi-armed bandit problem and provide efficient solutions for this problem. However, unlike known bandit problems, our problem possesses unique characteristics due to the following factors: (a) we consider subsets of arms (called super-arms) which combine the rewards of underlying arms in a weighted sum and (b) our goal is to bound the reward of the best super-arm in an interval with high probability, which provides a high probability bound on the value of the best deviation. We emphasize that our goal is not to identify the most profitable deviation strategy but to just bound this deviation's gain, which results in substantially fewer samples as compared to approaches that aim to identify the best (super) arm. Our main approach is called Super-Arm UCB (SAUCB). While this problem, at first glance, may seem amenable to known techniques, we show via thorough experiments that simple approaches using prior methods require many more samples than SAUCB in practice. We also present three variants of SAUCB and obtain their sample complexities. Furthermore, we provide a lower bound for the sample complexity of a specific instance of this problem.

Our *second contribution* is an extensive set of experiments comparing SAUCB, its variants, and various baseline approaches. Our comparisons are for both synthetic data and for a large-scale example based on a well-known agent-based simulator of stock markets that has been used in recent papers in AI venues (Wang, Vorobeychik, and Wellman 2018). Among potential approaches in the literature, a recent work applies to our setting (Huang et al. 2018) (called COCI here based on the algorithm name); however, that work does not aim for the goal mentioned in (b) above. Simple approaches combining pure arm exploration and sampling from the mixed strategy also perform poorly. Thus, SAUCB is able to outperform all known approaches, including our own proposed variants, by a large amount. All the full and missing proofs in this paper and additional graphical results are in the appendix of the full version, which is available on the authors' webpages.

Problem Description

Empirical games are so complex that the utilities for players are unknown to begin with. A simulator takes as input a strategy profile and returns an observation of the payoffs of each player. In empirical game-theoretic analysis, this simulator is treated as a black-box payoff oracle, which allows any arbitrarily complex game to be analyzed. Since these games typically incorporate stochastic factors, the payoff observation vector is a sample from some underlying distribution of payoff outcomes. Thus, many simulation calls may be necessary to even estimate the expected payoff for a given strategy profile.

In this paper, for ease of presentation, we focus on symmetric games with n players. These are games in which all n players have the same strategy space $\mathcal{S}$. The outcome and hence the player payoffs depend only on how many players played each strategy; that is, a player's identity does not matter in the outcome of the game. It is well known that such games have a NE in symmetric strategies, i.e., a NE in which every player plays the same (possibly mixed) strategy. Note that our focus on symmetric games is without loss of generality, as the non-symmetric case would just require repeating the regret computation n times, once for each player.

We denote a symmetric mixed strategy as a K-dimensional vector p, where p_i denotes the probability of choosing $i \in \mathcal{S}$ and $K = |\mathcal{S}|$. Note that we consider only the space of symmetric strategies, so the strategy of every player is given by the same p. We wish to compute regret for a given fixed p, which will be implicit in all our notation henceforth. Let $U \subseteq \mathcal{S}$ be the support of p. In this work, we assume $|U| > 1$ to rule out the degenerate case when the mixed strategy is actually a pure strategy. Denote by D_i the random utility of a player when this player plays $i \in \mathcal{S}$ and others play according to p (called the deviating payoff). Again note that this payoff is not player-specific due to the symmetric nature of the game. Also, it can be readily inferred that the payoff of the mixed strategy p is given by $\sum_{j\in\mathcal{S}} p_j D_j$. Then, the regret R of the strategy profile p in a symmetric game is given by:

$$R = \max_{i\in S} E[R_i] \text{ where } R_i = D_i - \sum_{j\in\mathcal{S}} p_j D_j$$

R_i is the gain in payoff of deviating to strategy i.

Problem Statement: Our goal is to find an interval $[L_R, U_R]$ such that $R \in [L_R, U_R]$ with high probability and $U_R - L_R < W$ for some fixed W within the fewest samples possible.

Combinatorial Bounding Bandit Problem (CBBP): We set up the above problem in a stochastic bandit framework. However, the objective of this bandit problem is quite distinct from prior bandit problems; hence, we refer to this problem as the Combinatorial Bounding Bandit Problem (CBBP). There are K arms and an arm i has the random payoff D_i. Let $\mu_i = E[D_i]$. R_i, as described above, is a linear combination of payoffs of a subset of arms. We call this subset the super-arm (SA). For super-arm i this subset is $\{i\}\cup U$. Super-arm i has the random payoff R_i. Let $\nu_i = E[R_i]$. Then, following our problem statement, we wish to the find an interval $[L_R, U_R]$ such that the expected payoff of the best super arm lies in $[L_R, U_R]$ with high probability and $U_R - L_R < W$ for some fixed W within the fewest samples possible. Observe that this objective has not been studied in the literature (see related work for a detailed comparison). The combinatorial nature of this problem arises from the super-arm structure; we exploit this special structure in our solution.

Note that our sampling primitive is the deviating payoff D_i from a mixed-strategy profile. A call to the simulator returns a sample for an arm, so at every time step we sample from an arm and not from a super-arm. Thus, our sample complexity results count the number of times arms are pulled, not super-arms. We otherwise treat the simulator as a black box, in line with the standard methodology in empirical game-theoretic analysis.

Further Notation: Define $c_{k,i} = p_k$ if $k \neq i$ and $c_{k,i} = 1 - p_k$ otherwise. $c_{k,i}$ is a convenient shorthand to express $R_i = c_{i,i}D_i - \sum_{k\in\mathcal{S}\setminus i} c_{k,i}D_k$. Let $\widehat{D}_{i,t}$ and $\widehat{R}_{i,t}$ denote the empirical means of the samples of the random variables D_i (arm) and R_i (SA) respectively. Let i^* denote the index of the best super-arm, and let $T_i(t)$ denote the number of times arm i is pulled before time step t. In this work, following standard bandit literature, we assume that the distribution of D_i for any i is sub-gaussian with parameter g^2. Also, let $\Delta_i = \nu_{i^*} - \nu_i$. By the definition of super arms, $\Delta_i = \mu_{i^*} - \mu_i$ as well. As is standard in bandit literature we assume $\Delta_{i^*} = \Delta_{(1)} = \Delta_{(2)} \leq \Delta_{(3)} \leq \ldots \leq \Delta_{(K)}$ where (i) denotes the i^{th} best super arm. As a consequence, w.l.o.g., arms are ordered by means so that arm 1 has highest mean; this ordering is just for ease of notation and is not used by any algorithm. For readability, we will write Δ_i instead of $\Delta_{(i)}$.

Related Work

Empirical Games: There is a large body of work on empirical games (Wellman 2006; Jordan, Vorobeychik, and Wellman 2008; Jordan, Schvartzman, and Wellman 2010; Tuyls et al. 2018). However, as stated earlier, applications of these techniques either use ad-hoc methods to report the regret of approximate Nash equilibria or often do not report it (Wang, Vorobeychik, and Wellman 2018; Prakash and Wellman 2015).

Stochastic Bandits: In classical stochastic bandit problems, the goal is to design an efficient sampling algorithm to minimize the *cumulative regret* (Lattimore and Szepesvári 2018; Bubeck, Cesa-Bianchi, and others 2012; Auer, Cesa-Bianchi, and Fischer 2002). Cumulative regret is the cumulative reward difference between the optimal static strategy and the one realized by the designed algorithm, which is different from *regret* in the game-theoretic sense that we use here. We aim to bound the regret of the game in an interval, which translates to bounding the reward of the best SA in an interval. This makes our setting different from that of classical stochastic bandits.

Our problem is more related to *pure exploration* or *best arm identification* in multi-armed bandits (Audibert and Bubeck 2010; Bubeck, Munos, and Stoltz 2011). Originating from the problem of deriving PAC bounds for identifying ϵ-optimal arms (Even-Dar, Mannor, and Mansour 2006; 2002; Mannor and Tsitsiklis 2004), recently various efficient algorithms have been proposed for both the fixed budget setting (Gabillon et al. 2011; Gabillon, Ghavamzadeh, and Lazaric 2012; Karnin, Koren, and Somekh 2013) and the fixed confidence setting (Kalyanakrishnan et al. 2012; Mnih, Szepesvári, and Audibert 2008; Kaufmann, Cappé, and Garivier 2016). Even more relevant is the *combinatorial pure exploration problem* (CPE) (Chen et al. 2014; Gabillon et al. 2016; Chen et al. 2017). While this problem's goal is to identify the best super-arm (defined as a subset of single arms with additive rewards), our objective still differs as the values of our super-arms are weighted combinations of rewards of single arms. A recent work handles weighted combinations of rewards (Huang et al. 2018) in an algorithm called COCI which can be used for our problem, but COCI still aims to identify the best super-arm and not bound the highest reward in an interval. We compare to COCI in experiments. Overall, many methods for CPE do not apply to our problem, and those that apply (COCI) perform poorly.

Another line of work (Antos, Grover, and Szepesvári 2008; 2010; Carpentier et al. 2011) aims to estimate the mean values of the arms using an active learning approach. However, these works minimize the uniform mean square loss across *all arms* by controlling the empirical variance estimate, while our aim is to bound *only* the best super-arm mean. To the best of our knowledge, (Zhou, Li, and Zhu 2017) is the only work that studies multi-armed bandit problems in an empirical game setting. However, they consider the very different problem of identifying the pure strategy NE in a two player zero-sum game by efficiently querying different pure strategy profiles to construct the stochastic payoff bimatrix.

Approaches for Bounding Regret

Super-Arm UCB (SAUCB) is our primary solution for CBBP. SAUCB is specified in Algorithm 1. To begin with, we present a lower bound for the sample complexity followed by an simple approach that combines known methods to seemingly solve our problem but fails badly in practice.

Lower Bound: First, observe that given the number of pulls of each arm (that is, not as a random variable) we get that $\widehat{D}_{i,t}$ is sub-gaussian with mean μ_i and parameter $g^2/T_k(t)$. Next, the SA empirical payoff $\widehat{R}_{i,t}$ is a weighted sum of sub-gaussian random variables $\widehat{D}_{i,t}$. Hence $\widehat{R}_{i,t}$ has a sub-gaussian distribution with mean ν_i and parameter $g^2 \sum_{k\in\mathcal{S}} c_{k,i}^2/T_k(t)$. These results hold when the payoffs are normally distributed with variance g^2, which we use to prove the following:

Theorem 1. *Given W, there exists a problem instance for which the number of samples needed to get the true regret in an interval of width W with confidence $\mathrm{erf}(m/\sqrt{2})$ is at least*

$$T_{min} = K - 1 - |U| + 4\frac{g^2 m^2 \min_i \left\{\left(\sum_{k\in\mathcal{S}} c_{k,i}\right)^2\right\}}{W^2}.$$

erf is the error function associated with the normal distribution. With constant probability values and constant $|U|$, this lower bound for the problem instance is $\Omega(K + \frac{m^2}{W^2})$.

Proof Sketch. Choose the problem in which the arm payoffs are distributed normally with variance g^2. Then, we show that for a fixed number of time steps τ, there is an ideal proportion in which to distribute the samples among the arms in SA i. This is $T_k(\tau) \propto c_{k,i}\tau$, which can be readily seen to be the minimizer of $\sum_{k\in\mathcal{S}} \frac{c_{k,i}^2}{T_k(t)}$ and hence the variance of $\widehat{R}_{i,\tau}$. Using the ideal proportion, we calculate the minimum number of samples for any SA i to get the confidence width to be W with confidence $\mathrm{erf}(m/\sqrt{2})$ (i.e. the prob. of $\widehat{R}_{i,\tau}$ lying within m standard deviations of ν_i) as $\tau = 4\frac{g^2 m^2 (\sum_{k\in\mathcal{S}} c_{k,i})^2}{W^2}$. As i^* is not known, we take the worst case over SAs to get a lower bound. □

Simple Approach: One approach that is simple and apparently seems to address our problem is a modified pure exploration algorithm, which works as follows: since $R = \max_i E[D_i] - \sum_j p_j E[D_j]$, it would seem that estimating both (a) $\max_i E[D_i]$ (pure exploration) and (b) $\sum_j p_j E[D_j]$ (mixed-strategy utility estimation), each with $W/2$ bound width, would solve the overall problem of estimating R with bound width W. Indeed, we use this observation to propose a simple baseline. We use the successive elimination pure exploration algorithm (Even-Dar, Mannor, and Mansour 2006) (returning a bound width) along with a sampling-based estimation of mixed-strategy expected utility, which we together call *Modified SE*. This approach has the same asymptotic (in order notation) sample complexity as our SAUCB approach; however, the constants in the actual number of samples are higher. Thus, in practice this approach performs very poorly compared to SAUCB. Intuitively, samples are inefficiently allocated since the samples of arms in the best-arm identification portion may not be useful in narrowing the bound on the mixed-strategy payoff and vice versa, which is addressed in SAUCB. The details of Modified SE and its sample complexity are provided in the appendix.

SAUCB Algorithm

We start with a simple result where using Hoeffding's inequality for $\widehat{R}_{i,t}$ with fixed t, we can write

$$P\left[\widehat{R}_{i,t} + \sqrt{2g^2 \ln\frac{1}{\delta}\sum_{k\in\mathcal{S}}\frac{c_{k,i}^2}{T_k(t)}} \le \nu_i\right] \le \delta \text{ and}$$

$$P\left[\widehat{R}_{i,t} - \sqrt{2g^2 \ln\frac{1}{\delta}\sum_{k\in\mathcal{S}}\frac{c_{k,i}^2}{T_k(t)}} \ge \nu_i\right] \le \delta \quad (1)$$

Algorithm 1 SAUCB: A regret-bounding algorithm

Input: Mixed strategy p, width W, sub-gaussian param g^2

1: Pull each arm once, $t \leftarrow 0$, $B_{j,t} \leftarrow \sqrt{2g^2 \ln \frac{1}{\delta} \sum_{k \in \mathcal{S}} \frac{c_{k,j}^2}{T_k(t)}}$
2: **while** $w_t > W$ **do**
3: Increment t
4: $k \leftarrow \text{argmax}_{j \in \mathcal{S}} \widehat{R}_{j,t} + B_{j,t}$
5: $i \leftarrow \text{argmax}_{j \in \mathcal{S}} \left| \frac{\partial B_{k,t}}{\partial T_j(t)} \right|$
6: Pull arm i, increment count $T_i(t)$, update $\widehat{R}_{i,t}$
7: $k \leftarrow \text{argmax}_{j \in \mathcal{S}} \widehat{R}_{j,t} + B_{j,t}$
8: $w_t \leftarrow 2 * B_{k,t}$
9: **end while**
10: **return** interval $[\widehat{R}_{k,t} - B_{k,t}, \widehat{R}_{k,t} + B_{k,t}]$

The SAUCB algorithm selects the arm to pull in a hierarchical manner. The choice of a super-arm is driven by a UCB-style approach, but where the upper confidence width is based on the super-arm confidence bound as shown in Equation 1. In every round, the algorithm first chooses a super-arm k (line 4), and then chooses an arm i (line 5) within this super-arm such that this choice i leads to the largest reduction in the super-arm confidence width $B_{k,t}$. The magnitude of the derivative of $B_{k,t}$ w.r.t. $T_j(t)$ is used to decide the reduction that an arm j could provide (line 5), since this derivative will be 0 for arms outside the chosen super-arm. Thus, our approach of choosing an arm within each super-arm is guided by the goal of greedily reducing the confidence width of the chosen super-arm. Finally, the SA width is computed (line 8) and compared to the desired width (line 2) to decide whether to stop or not. We show the following sample complexity result.

Theorem 2. *In order to get an interval of width W containing the true regret with probability $1 - \alpha$, the total number of samples t required by SAUCB is bounded by*

$$t \leq K + 16g^2 \log \frac{1}{\delta} \left(\max \left(\frac{1 - p_1}{W^2}, \frac{p_1}{\max(W, \Delta_2)^2} \right) + \sum_{k \in \mathcal{S} \setminus 1} \max \left(\frac{1 - p_k}{\max(W, \Delta_k)^2}, \frac{p_k}{W^2} \right) \right)$$

where $\alpha = 2Kt_{max}\delta$ and $t_{max} = \frac{16Kg^2 \log \frac{1}{\delta}}{W^2}$. With constant probability values, t is $O(K + \frac{K \log \frac{1}{\delta}}{W^2})$. Furthermore, SAUCB uses a minimum of $\Omega(K + \frac{\log \frac{1}{\delta}}{W^2})$ samples.

Proof Sketch. First, each arm is sampled once, giving K samples. Suppose that at time t, suboptimal arm $k \neq 1$ is sampled when some super-arm i_t^* is chosen as the best super-arm. This arm will only be sampled if the width of the bound on i_t^* is greater than W: $2B_{i_t^*,t} > W$. In addition, since arm k is sampled, we have that for all arms $j \in \mathcal{S}$, $\left| \frac{\partial B_{i_t^*,t}}{\partial T_k(t)} \right| \geq \left| \frac{\partial B_{i_t^*,t}}{\partial T_j(t)} \right|$ which implies $T_j(t) \geq \frac{c_{j,i_t^*}}{c_{k,i_t^*}} T_k(t)$. Combining these, we get $T_k(t) \leq \left(16g^2 c_{k,i_t^*} \log \frac{1}{\delta} \right) / W^2$. If arm $k \neq 1$ is sampled when it is also chosen as the best super-arm ($i_t^* = k$), then $\widehat{R}_{k,t} + B_{k,t} \geq \widehat{R}_{1,t} + B_{1,t}$ which gives $T_k(t) \leq \left(16g^2 (1 - p_k) \log \frac{1}{\delta} \right) / \Delta_k^2$ where we again use the bound on the partial derivative to achieve the second inequality. Combining these results, since either $i_t^* = k$ or $i_t^* \neq k$ when k is sampled, $T_k(t) \leq 16g^2 \log \frac{1}{\delta} \max \left(\frac{1 - p_k}{\max(W, \Delta_k)^2}, \frac{p_k}{W^2} \right)$. We use analogous techniques to achieve a bound for arm 1: $T_1(t) \leq 16g^2 \log \frac{1}{\delta} \max \left(\frac{1 - p_1}{W^2}, \frac{p_1}{\max(W, \Delta_2)^2} \right)$. Summing over all arms, we achieve the required result. The complexity can be seen by noting that $\max \left(\frac{1 - p_k}{\max(W, \Delta_k)^2}, \frac{p_k}{W^2} \right) \leq \frac{1 - p_k}{W^2} + \frac{p_k}{W^2}$. The minimum number of samples is found in a manner similar to Thm. 1. □

The result above also reveals the difference from the best (super) arm identification problem. Typically in best arm identification problems, the sample complexity is a function of terms $1/\Delta_k^2$ for all k. However, here we see that these terms are clamped to W, as in $1/\max(W, \Delta_k)^2$. Thus, even if the top two super-arms are very close (i.e. Δ_2 is very small), our sample complexity is not as high as it would be for the best arm identification problem due to this clamping effect. Moreover, even when Δ_2 is large, the sample complexity depends on the maximum over $\frac{p_k}{W^2}$ and $\frac{1 - p_k}{\max(W, \Delta_k)^2}$, and hence W primarily determines the sample complexity, as can be seen in the order notation above. This also explains why we do better than the pure super-arm exploration algorithm COCI (Huang et al. 2018) in experiments.

While the upper bound above depends on K in a multiplicative manner, there are specific problem instances for which the upper bound matches the minimum number of samples for SAUCB. One such example is when the mixed strategy has a uniform distribution (i.e. all p_k are the same and $\Delta_k >> W$ for $k \in \mathcal{S} \setminus 1$). It can be readily checked that the second term reduces to $\sum_{k \in \mathcal{S} \setminus 1} p_k / W^2 = (1 - p_1)/W^2$. This makes the resulting upper bound for this instance $O(K + \frac{\log \frac{1}{\delta}}{W^2})$. Given this observation, the advantage of the SAUCB approach (e.g. as compared to the simple baseline Modified SE) is in better constants for the actual number of samples, leading to much better performance in practice as revealed in our experiments.

Variants

In order to provide a comprehensive comparison to alternatives, we also propose a few variants of the algorithm above. These variants serve as additional baselines that we compare SAUCB against. The main variation is in the concentration bound that is used for the upper confidence bound. For SAUCB, we have used Hoeffding's bound. Our first variant lil-SAUCB uses the law of the iterated logarithm (lil) bound (Jamieson et al. 2014; Jamieson and Nowak 2014). Unlike Hoeffding, the lil bound is time-uniform; that is, the lil bound holds for all timesteps (avoiding a naive union bound over time). While a number of other time-uniform concentration bounds exist in the literature (Huang et al.

2018; Zhao et al. 2016), in practice, the Hoeffding bound works much better for us than the lil bound (see experiments). Thus, we limit ourselves to just the Hoeffding bound and lil bound.

lil-SAUCB: This variant follows the same template as Alg. 1, except $B_{j,t} = \sum_{k \in \mathcal{S}} c_{k,j} b_{k,t}$ where

$$b_{k,t} = (1+\sqrt{\epsilon})\sqrt{\frac{2g^2(1+\epsilon)}{T_k(t)} \log \frac{\log\left(T_k(t)(1+\epsilon)\right)}{\delta}} \quad (2)$$

and $\epsilon \in (0,1)$, $\delta \in (0, \frac{\log(1+\epsilon)}{e})$ are chosen constants. We show the following sample complexity:

Theorem 3. *In order to get an interval of width W containing the true regret with probability $1-\alpha$, the total number of samples t taken by lil-SAUCB is bounded by*

$$t \leq K + \max\left[K_1 \frac{(1-p_1)^{2/7}}{W}, \min\left(K_1 \frac{p_1^{2/7}}{W}, K_2 \max_{j \in \mathcal{S}\setminus 1}\left(\left(\frac{p_1}{\Delta_j}\right)\left(1 + K_3 \frac{(1-p_j)^{5/7}}{p_1^{5/7}}\right)\right)\right)\right]^{63/26} + \sum_{k \in \mathcal{S}\setminus 1} \max\left[K_1 \frac{p_k^{2/7}}{W}, \min\left(K_1 \frac{(1-p_k)^{2/7}}{W}, K_2\left(\frac{(1-p_k)}{\Delta_k}\right)\left(1 + K_3 \frac{p_1^{5/7}}{(1-p_k)^{5/7}}\right)\right)\right]^{63/26}$$

where K_1, K_2, and K_3 are constants (defined in the appendix) that depend on ϵ, δ, g, and K. $\alpha = \frac{2K(2+\epsilon)}{\epsilon}\left(\frac{\delta}{\log(1+\epsilon)}\right)^{1+\epsilon}$. With constant probability values, t is $O\left(K + \frac{K^{22/13}}{W^{63/26}\delta^{33/52}\log^{9/13}\frac{1}{\delta}}\right)$.

Proof Sketch. We follow the same method as in the proof for Thm. 2, repeatedly using the following two inequalities for simplification:

$$\log x \geq 1/x \text{ for } x \geq 3 \text{ and } 3x^{1/3} \geq \log x \text{ for } x \geq 0\ .$$

Using the inequalities $2B_{i_t^*,t} > W$ and $\left|\frac{\partial B_{i_t^*,t}}{\partial T_k(t)}\right| \geq \left|\frac{\partial B_{i_t^*,t}}{\partial T_j(t)}\right|$, we can derive that arm k is only sampled if $T_k(t) < \left(K_1 \frac{c_{k,i_t^*}^{2/7}}{W}\right)^{63/26}$. We derive additional restrictions on the number of samples taken from arm $k \neq 1$ if it is sampled and chosen as the best SA at time step t, using $\widehat{R}_{k,t} + B_{k,t} \geq \widehat{R}_{1,t} + B_{1,t}$ to get $T_k(t) \leq \left(K_2\left(\frac{(1-p_k)}{\Delta_k}\right)\left(1 + K_3 \frac{p_1^{5/7}}{(1-p_k)^{5/7}}\right)\right)^{63/26}$ and on the number of samples taken from arm $k = 1$ if it is sampled and not chosen as the best SA at time step t (for some $j \neq 1$) using $\widehat{R}_{j,t} + B_{j,t} \geq \widehat{R}_{1,t} + B_{1,t}$ to get $T_1(t) \leq \left(K_2 \max_{j \in \mathcal{S}\setminus 1}\left(\left(\frac{p_1}{\Delta_j}\right)\left(1 + K_3 \frac{(1-p_j)^{5/7}}{p_1^{5/7}}\right)\right)\right)^{63/26}$. Combining these, we get the desired result. The complexity result follows from the fact that K_1 is $\Theta\left(\frac{K^{2/7}}{\delta^{11/42}\log^{2/7}\frac{1}{\delta}}\right)$. □

There are two other variants that arise from the way the super-arm is chosen. Until now, we used the bounds on the super-arms to choose a super-arm. However, there is a one-to-one correspondence between the super-arms and individual arms; thus, the super-arm choice can be driven by upper confidence bounds for the corresponding arms. In particular, the choice in line 4 (Alg. 1) is made using $\widehat{D}_{i,t} + b_{i,t}$ inside the argmax. For the Hoeffding bound variant, $b_{i,t} = \sqrt{\frac{2g^2 \log(1/\delta)}{T_i(t)}}$; for the lil bound variant, $b_{i,t}$ is given by Equation 2. We call these variants *single-SAUCB* and *single-lil-SAUCB* respectively. Note that since standard SAUCB uses a special construction of the Hoeffding bound that does not bound individual arms, *single-SAUCB* instead uses a linear combination of Hoeffding bounds on individual arms to bound the super-arms for determining the width of the regret bound. The following results provide the sample complexity for these algorithms.

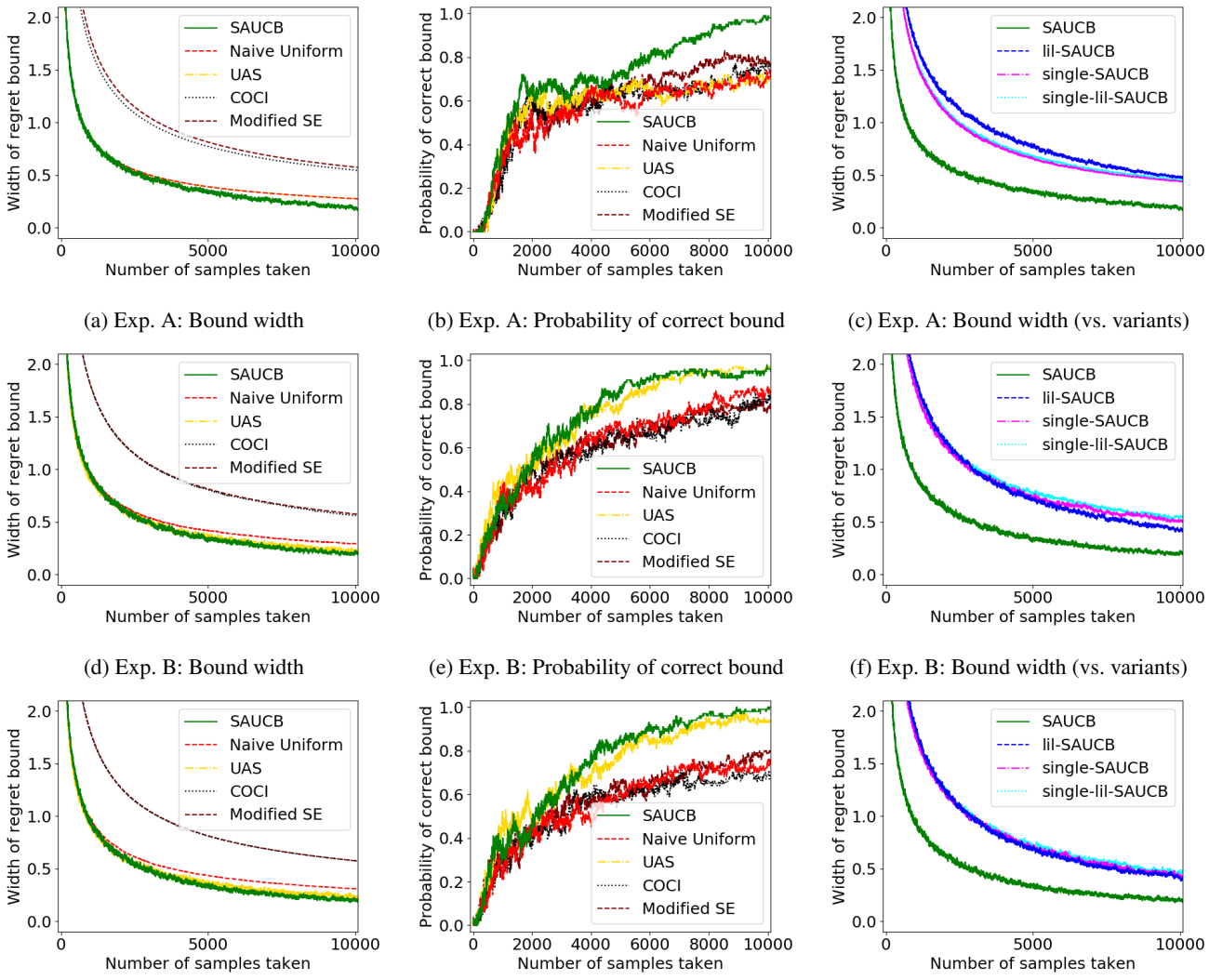
Theorem 4. *In order to get an interval of width W containing the true regret with probability $1-\alpha$, the total number of samples t taken by single-SAUCB is $O(K + \frac{K^{5/3}\log\frac{1}{\delta}}{W^2})$, where $\alpha = 2Kt_{max}\delta$ and $t_{max} = 8Kg^2\log\frac{1}{\delta}\left(\frac{(1+(K-1)^{1/3})}{W}\right)^2$.*

Theorem 5. *In order to get an interval of width W containing the true regret with probability $1-\alpha$, the total number of samples t taken by single-lil-SAUCB is $O\left(K + \frac{K^{22/13}}{W^{63/26}\delta^{33/52}\log^{9/13}\frac{1}{\delta}}\right)$, where $\alpha = \frac{2K(2+\epsilon)}{\epsilon}\left(\frac{\delta}{\log(1+\epsilon)}\right)^{1+\epsilon}$.*

Experiments

Baselines and Evaluation Metrics: We compare our approach against a number of baselines. We have experimentally found that the Hoeffding bound is the tightest in terms of width compared to the lil bound or the bound used in COCI. Therefore, for a fair comparison, we allow all the baseline algorithms to use Hoeffding bounds on the arms or super-arms. The baselines we compare against are (a) naive uniform: arms are sampled in a uniform distribution, (b) COCI: prior work by (Huang et al. 2018) which is a pure exploration algorithm for super-arms with weighted rewards, (c) UAS: modified SAUCB where the arms within a super-arm are not selected using the derivative values but aiming for every arm in the super-arm to be sampled equally often (that is, a uniform distribution within the super-arm), and (d) Modified SE: the simple approach that combines successive elimination (Even-Dar, Mannor, and Mansour 2006) and mixed-strategy sampling as described earlier.

We compare these baselines across three different criteria. First (1), for all experiments, we compare the bound width variation over the number of samples. Second (2), for the synthetic-data experiments, we compare the ground-truth probability of the true regret lying in a width-W bound around the empirical regret estimate at each time step; for SAUCB and each baseline, this probability is estimated as



(a) Exp. A: Bound width (b) Exp. A: Probability of correct bound (c) Exp. A: Bound width (vs. variants)

(d) Exp. B: Bound width (e) Exp. B: Probability of correct bound (f) Exp. B: Bound width (vs. variants)

(g) Exp. C: Bound width (h) Exp. C: Probability of correct bound (i) Exp. C: Bound width (vs. variants)

Figure 1: Results on synthetic data experiments

percentage of 100 runs in which the true regret was within a W width interval around the empirical regret. Third (3), for the large-scale experiments, we compare the total number of samples taken to achieve the required bound width in a single run.

We also compare against the variants of our algorithm: lil-SAUCB, single-SAUCB, and single-lil-SAUCB. Due to space constraints, we compare against these variants only for criterion (1) in the main paper. The comparison using criterion (2) for these variants is presented in the appendix.

Synthetic-Data Experiments

Our experiments follow the design in prior work (Audibert and Bubeck 2010). We use 20 arms with Bernoulli distributions, and set $W = 0.05$ and $\alpha = 0.05$. We present three setups Exp. A, Exp. B, and Exp. C in Table 1. In Exp. A, the top mean is not close to the next 5 means while the remaining 14 means are yet lower, and all arms are in support with equal probability. In Exp. B, the top two means are close and the remaining 18 arms have same means, and the top mean is in support while the second-best mean is not. In Exp. C, the means are exactly the same as in Exp. B, but the top mean is not in support while the second-best mean is.

Figs. 1a-1i show our results on synthetic data for the baselines as well as against our algorithm variants. The bound-width-variation-over-time results (averaged over 100 trials) in Figs. 1a, 1d, and 1g show that SAUCB is able to consistently achieve a lower bound width over all time steps as compared to the other baselines for all Exp. A, B, and C. Note that UAS performs identically to uniform in Exp. A, but performs almost as well as SAUCB in Exp. B and C. This can be explained since Exp. A has all arms in support of the mixed strategy, and so uniform sampling within the support for any super-arm is equivalent to uniformly sampling over

	Arm means
Exp. A	$[0.5, 0.42 \times 5, 0.38 \times 14]$
Exp. B	$[0.5, 0.48, 0.37 \times 18]$
Exp. C	$[0.5, 0.48, 0.37 \times 18]$
	Mixed strategy
Exp. A	$[0.05 \times 20]$
Exp. B	$[0.2, 0, 0.3, 0.3, 0.1, 0.1, 0 \times 14]$
Exp. C	$[0, 0.2, 0.3, 0.3, 0.1, 0.1, 0 \times 14]$

Table 1: Arm means and mixed strategy for Exp A, B, and C

all arms, which is much less effective than the derivative approach. Also, as explained earlier, COCI performs quite poorly primarily because its objective is to identify the best super-arm, not to bound it. Modified SE performs as poorly as COCI; in the appendix, we further dissect the reasons for this. As a result, we do not use COCI or Modified SE for the large-scale experiments.

Figs. 1b, 1e, and 1h show that SAUCB is able to get a correct bounding interval with probability 1 in fewer samples compared to the baselines for all Exp. A, B, and C. A point to note is that UAS has good performance in Exp. B initially; this is potentially because Exp. B contains the best arm in support and the support set is small, and hence UAS initially gets to sample the best arm often, when the empirical estimates of arm means are off. Figs. 1c, 1f, and 1i show that SAUCB consistently outperforms the variants by a large amount for Exp. A, B, and C in terms of the number of samples required to get a lower bound width. Thus, we choose SAUCB as our algorithm for the large-scale experiments from among the variants.

Large-Scale Experiments

For our large-scale experiments, we use an empirical game from past work in the domain of simulating and studying strategic behavior in stock markets (Wang, Vorobeychik, and Wellman 2018). The underlying strategic interaction in stock markets is extremely complicated and so cannot be described in a closed or compact form. Hence, the only interface to the system is the observation of outcomes from an agent-based simulator model of the stock market (Wellman and Wah 2017). The underlying game has many players, is dynamic and repeated, has partial observability of actions and state, as well as stochasticity. As stated earlier, empirical game-theoretic methods have been developed to solve such complex games; Wang, Vorobeychik, and Wellman (2018) calculate approximate NE using such techniques.

We examine two settings specified in this paper (the LSHN-K0 and MSMN-K0 settings with no spoofer), and bound the regret of the reported NE in each setting with $\alpha = 0.05$ and $W = 0.1$. An issue that arises in some such practical settings is that the sub-gaussian parameter is unknown. Thus, for these experiments, we pre-sample 10,000 deviating payoffs and clamp payoffs during the experiment to $[0, 1]$, taking anything above the 75th percentile as 1 and below the 25th percentile as 0; we do so because otherwise

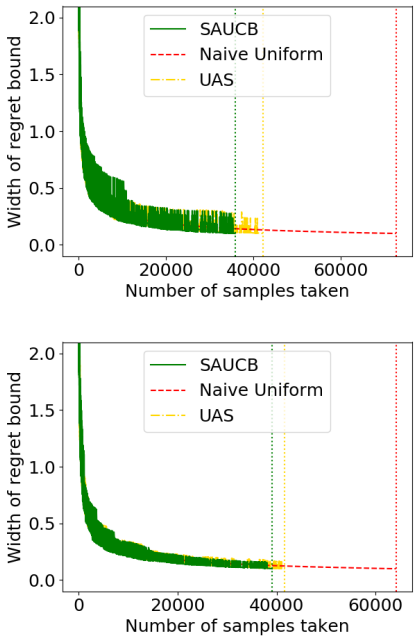


Figure 2: Bound width for LSHN-K0 (left) and MSMN-K0 (right) settings

normalizing the full range of payoffs (which have extremely high variance and very large outliers) results in a degenerate case where all payoffs are near-identical. Figure 2 shows our result, where the bound width varies a lot over time since we are showing only one run of the algorithm (as this is how SAUCB would be used in practice). The vertical lines show when different algorithms achieve the width W. As can be seen, SAUCB achieves the required bound in fewer samples than the baselines. The run-time for each run of an algorithm in these experiments is about 6 hours (on a 2.4GHz CPU) due to the time-consuming stock market simulator, as opposed to minutes for synthetic data.

Conclusion

We formulated a new kind of multi-armed bandit problem in order to provide quantitative regret guarantees for the approximate NE computed in empirical games. We proposed an algorithm SAUCB and some variants and analyzed these theoretically as well as experimentally in synthetic and stock-market empirical-game scenarios. We found SAUCB beat a wide range of alternate approaches quite convincingly. Overall, we hope that this work provides a basis for more principled guarantees about the equilibria output by various methods in the area of empirical games.

Acknowledgement

We thank Erik Brinkman and Michael P. Wellman for invaluable discussion towards the beginning of this work. Most of this work was done when Steven and Arunesh were at the University of Michigan, where the work was supported by US National Science Foundation under grant IIS-1741190.

References

Antos, A.; Grover, V.; and Szepesvári, C. 2008. Active learning in multi-armed bandits. In *ALT*, 287–302.

Antos, A.; Grover, V.; and Szepesvári, C. 2010. Active learning in heteroscedastic noise. *Theoretical Computer Science* 411(29-30):2712–2728.

Audibert, J.-Y., and Bubeck, S. 2010. Best arm identification in multi-armed bandits. In *COLT*.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47(2-3):235–256.

Bubeck, S.; Cesa-Bianchi, N.; et al. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* 5(1):1–122.

Bubeck, S.; Munos, R.; and Stoltz, G. 2011. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science* 412.

Carpentier, A.; Lazaric, A.; Ghavamzadeh, M.; Munos, R.; and Auer, P. 2011. Upper-confidence-bound algorithms for active learning in multi-armed bandits. In *ALT*, 189–203. Springer.

Chen, S.; Lin, T.; King, I.; Lyu, M. R.; and Chen, W. 2014. Combinatorial pure exploration of multi-armed bandits. In *NIPS*, 379–387.

Chen, L.; Gupta, A.; Li, J.; Qiao, M.; and Wang, R. 2017. Nearly optimal sampling algorithms for combinatorial pure exploration. In *COLT*, 482–534.

Even-Dar, E.; Mannor, S.; and Mansour, Y. 2002. Pac bounds for multi-armed bandit and markov decision processes. In *COLT*, 255–270. Springer.

Even-Dar, E.; Mannor, S.; and Mansour, Y. 2006. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research* 7(Jun):1079–1105.

Gabillon, V.; Ghavamzadeh, M.; Lazaric, A.; and Bubeck, S. 2011. Multi-bandit best arm identification. In *NIPS*, 2222–2230.

Gabillon, V.; Lazaric, A.; Ghavamzadeh, M.; Ortner, R.; and Bartlett, P. 2016. Improved learning complexity in combinatorial pure exploration bandits. In *AISTATS*, 1004–1012.

Gabillon, V.; Ghavamzadeh, M.; and Lazaric, A. 2012. Best arm identification: A unified approach to fixed budget and fixed confidence. In *NIPS*, 3212–3220.

Huang, W.; Ok, J.; Li, L.; and Chen, W. 2018. Combinatorial pure exploration with continuous and separable reward functions and its applications. In *IJCAI*, 2291–2297.

Jamieson, K., and Nowak, R. 2014. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *CISS*, 1–6.

Jamieson, K.; Malloy, M.; Nowak, R.; and Bubeck, S. 2014. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *COLT*, 423–439.

Jordan, P. R.; Schvartzman, L. J.; and Wellman, M. P. 2010. Strategy exploration in empirical games. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, 1131–1138.

Jordan, P. R.; Vorobeychik, Y.; and Wellman, M. P. 2008. Searching for approximate equilibria in empirical games. In *Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems*, 1063–1070.

Kalyanakrishnan, S.; Tewari, A.; Auer, P.; and Stone, P. 2012. Pac subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, 655–662.

Karnin, Z.; Koren, T.; and Somekh, O. 2013. Almost optimal exploration in multi-armed bandits. In *ICML*, 1238–1246.

Kaufmann, E.; Cappé, O.; and Garivier, A. 2016. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research* 17(1):1–42.

Lanctot, M.; Zambaldi, V.; Gruslys, A.; Lazaridou, A.; Tuyls, K.; Pérolat, J.; Silver, D.; and Graepel, T. 2017. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems*, 4190–4203.

Lattimore, T., and Szepesvári, C. 2018. Bandit algorithms. *preprint*.

Mannor, S., and Tsitsiklis, J. N. 2004. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research* 5(Jun):623–648.

Mnih, V.; Szepesvári, C.; and Audibert, J.-Y. 2008. Empirical bernstein stopping. In *Proceedings of the 25th international conference on Machine learning*, 672–679. ACM.

Prakash, A., and Wellman, M. P. 2015. Empirical game-theoretic analysis for moving target defense. In *Proceedings of the 2nd ACM Workshop on Moving Target Defense*, 57–65.

Tuyls, K.; Perolat, J.; Lanctot, M.; Leibo, J. Z.; and Graepel, T. 2018. A generalised method for empirical game theoretic analysis. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 77–85.

Wang, X.; Vorobeychik, Y.; and Wellman, M. P. 2018. A cloaking mechanism to mitigate market manipulation. In *IJCAI*, 541–547.

Wellman, M. P., and Wah, E. 2017. Strategic agent-based modeling of financial markets. *RSF: The Russell Sage Foundation Journal of the Social Sciences* 3(1):104–119.

Wellman, M. P. 2006. Methods for empirical game-theoretic analysis. In *AAAI*.

Zhao, S.; Zhou, E.; Sabharwal, A.; and Ermon, S. 2016. Adaptive concentration inequalities for sequential decision problems. In *Advances in Neural Information Processing Systems*, 1343–1351.

Zhou, Y.; Li, J.; and Zhu, J. 2017. Identify the nash equilibrium in static games with random payoffs. In *ICML*, 4160–4169.