Bootstrap Statistics for Empirical Games

Bryce Wiedenbeck
University of Michigan
btwied@umich.edu

Ben-Alexander Cassell
University of Michigan
bcassell@umich.edu

Michael P. Wellman
University of Michigan
wellman@umich.edu

ABSTRACT

Researchers often use normal-form games to model multi-agent interactions. When a game model is based on observational or simulated data about agent payoffs, we call it an *empirical game*. The payoff matrix of an empirical game can be analyzed like any normal-form game, for example, by identifying Nash equilibria or instances of other solution concepts. Given the game model's basis in sampled data, however, empirical game analysis must also consider sampling error and distributional properties of candidate solutions. Toward this end, we introduce bootstrap techniques that support statistical reasoning as part of the empirical game-theoretic analysis process. First, we show how the bootstrap can be applied to compute confidence bounds on the regret of reported approximate equilibria. Second, we experimentally demonstrate that applying bootstrapped regret confidence intervals can improve sampling decisions in simulation-based game modeling.

1. MOTIVATION

Traditional game-theoretic analysis is often confined to what is mathematically tractable for the game theorist, and as such is often restricted in scale (numbers of players, strategies) and complexity (dynamics, strategic interaction, information structures). When the extremes of abstraction required for analytic solution are too constraining, we may instead describe a strategic scenario in terms of an agent-based model (ABM), and attempt to build a game form by simulating that ABM [5, 6, 12, 18]. Alternatively, we may wish to build game models by observing real world play [4, 10]. In either case, payoff observations may be noisy, leading to uncertainty as to whether the game model accurately describes the true game. As such, we would like to quantify this uncertainty when conducting game-theoretic analysis on such games.

The approach of building and analyzing game models from data is known as *empirical game-theoretic analysis* (EGTA) [17]. Though some EGTA studies report statistical information (e.g., error bars on payoff estimates or significance of particular comparisons), there has been limited work to date on general methods for statistical reasoning about conclusions from empirical games. Instead, studies rely on extremely large sample size [18], or variance reduction techniques [12], to bolster confidence in results. Whereas more samples and less variance does increase confidence, we must quantify this confidence if results from EGTA studies are to be taken as serious scientific evidence for propositions of interest. In the absence of statistical guidance, we cannot tell whether practitioners are taking insufficient or excessive observations, or whether complex variance reduction measures are worth their cost, or most importantly, whether published results reflect fundamental properties of the games studied or artifacts of sampling variation.

The general dearth of statistical analysis in EGTA is understandable, given the difficulty of evaluating complex game-solution hypotheses in a traditional statistical framework. Little is typically known about payoff distributions prior to observing play, rendering parametric approaches inapplicable. Moreover, determining whether a profile satisfies a solution concept such as Nash equilibrium generally requires evaluating multiple statistical hypotheses about comparison of corresponding payoffs across potential deviations. This raises multiple-testing concerns [3], as well as other complexities due to the interdependencies among these comparisons.

This paper represents a first systematic effort to develop statistical methods generally applicable to game models derived from empirical data. Our approach is based on bootstrap techniques, which leverage the additional information available in observation data to characterize distributions over game-theoretic conclusions. We apply this approach to two problems. First, we present a bootstrap method to calculate statistical confidence bounds on reported equilibria. Second, we exploit this statistical information in an algorithm to improve sampling decisions.

2. BACKGROUND

2.1 Simulation-Based Games

In most applications of EGTA, a simulator acts as an oracle for player utility functions, taking as input a *strategy profile*—an assignment of one strategy to every player—and returning an observation of the payoff each player accrued from that profile in simulation. Since ABMs typically incorporate stochastic factors (uncertainty in environment and agent private information), the payoff-vector observation is actually a sample from some underlying distribution of payoff outcomes. Using the ABM simulator, we collect an observation set Θ by repeatedly sampling each strategy profile. We can then construct a game model $\mathcal{M}(\Theta)$ from the observation set. The most common way to construct a game

Appears in: *Alessio Lomuscio, Paul Scerri, Ana Bazzan, and Michael Huhns (eds.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France.*

model from simulation data is to build a normal-form payoff matrix, where each entry in the matrix corresponds to the mean of payoff observations of the corresponding profile. Researchers adopting the EGTA approach have used such models to perform standard game-theoretic calculations, such as derivation of Nash equilibria or finding dominant strategies [2, 12, 14]. They have also employed results of these game computations to estimate features of ABM outcomes in equilibrium, such as the average price of a security in a financial market [5].

Due to its statistical nature, it may be costly or impossible to identify exact Nash equilibria of the game underlying the simulator (hereafter referred to as the *true game*). Instead, it is often helpful to identify profiles that are close approximations to Nash equilibria. Quality of an approximate equilibrium is measured by regret, the largest gain any player can achieve through unilateral deviation. Formally, regret of a (potentially mixed-strategy) profile $\vec{\sigma}$ in a symmetric game is given by:

$$\epsilon(\vec{\sigma}) = \max_{\sigma \in \vec{\sigma}} \max_{s \in S} u(s, \vec{\sigma}_{-\sigma}) - u(\sigma, \vec{\sigma}),$$

where S is the set of pure strategies available to each player,[1] $u(\cdot)$ is the utility function, $\sigma \in \vec{\sigma}$ indicates that in profile $\vec{\sigma}$ some player is playing strategy σ and $\vec{\sigma}_{-\sigma}$ is the partial profile where the player using σ has been removed. By definition, a profile is a Nash equilibrium if and only if it has zero regret. We say a profile approximates Nash equilibrium at the level ϵ if the profile has regret less than or equal to ϵ.

Since empirical games are constructed from limited sampling of a noisy payoff-generating process, even exact equilibria of the empirical game may have non-negligible regret in the true game. Therefore, in order to draw conclusions about the true game, we wish to estimate the regret of a profile in the true game from payoff sample data. More specifically, we would like to state with statistical confidence whether or not a profile is an ϵ-Nash equilibrium of the true game through limited sampling of the simulator.

2.2 Bootstrap Statistics

The bootstrap is a computational method for estimating distributional information about a statistic computed on sample data [7]. Unlike classical statistical tests, it does not rest on explicit assumptions about the shape of the distribution. This feature is useful in reasoning about game solutions derived from sample data since nothing is initially known about the true payoff distributions. Furthermore, since the primary measure of interest, regret, is computed through taking a maximum rather than a sum or average, we cannot rely on the central limit theorem to fit a parametric statistical model to our sampling distribution.

The bootstrap treats a sample set as representative of the population and resamples the sample set to simulate drawing many samples from the population. If the original sample has size n, then each resample is a set of size n drawn with replacement from that sample. The statistic is then computed on each resample set, giving a bootstrap distribution for the statistic that can be used in place of a sampling distribution for the statistic. Of particular interest, the bootstrap distribution can be used to estimate confidence intervals for the statistic: the interval between the 2.5^{th} and 97.5^{th} percentile of the bootstrap distribution gives a two-sided 95% confidence interval, while the 5^{th} or 95^{th} percentiles give one-sided 95% confidence bounds.

2.3 Related Work

There has been limited work using bootstrapped statistics to analyze an agent-based model, let alone to analyze games. In contrast, discrete-event systems modeling has seen adoption of the bootstrap to analyze the output of simulation [8]. Axtell, et al. [1] suggested that a bootstrap approach may be necessary for determining if two agent-based models are equivalent, due to the complicated nature of such a hypothesis.

While the bootstrap has not previously been used in analyzing games, two methods for estimating true game regret from payoff sample data have been introduced. Reeves [15] proposed estimating the empirical distribution of regret of a profile by sampling game matrices from the space of possible matrices induced by assuming every payoff is independent and distributed normally with mean and variance equal to its sample mean and sample variance respectively. He recommended using the probability that a profile's regret is zero from the estimated empirical distribution of regret as a measure of the confidence that a profile is a Nash equilibrium. However, it is easy to demonstrate that mixed-strategy profiles, even true-game Nash equilibria, will have estimated regret of zero with negligible probability. Vorobeychik [16] presented a Bayesian framework for determining the posterior probability that a profile is an ϵ-Nash equilibrium of the true game from payoff sample data. The author proved tight probability bounds under the assumption that payoff observations are independent draws with Gaussian noise, and much weaker distribution-free bounds. Both of these works ignore the possibility that payoffs within a single observation of a profile may be correlated, and rely on distributional assumptions that cannot be guaranteed for small samples. No empirical evaluations have been presented for these methods, leaving open the question of their usefulness under more varied simulators and sampling designs.

3. METHODS

We propose two methods that apply bootstrapping to improve empirical game-theoretic analysis. First, we show how to estimate confidence intervals for the regret of strategy profiles, and in particular approximate Nash equilibria. We then show how such confidence intervals can be used to improve sample control decisions in simulation-based games.

3.1 Bootstrap Regret Estimates

Because empirical games are constructed from data, it is often unclear whether approximate Nash equilibria calculated in an empirical game correspond to Nash equilibria of the true game. Vorobeychik [16] proved that in the infinite-sample limit, empirical-game and true-game equilibria are identical, however, this does not rule out reporting spurious equilibria in empirical games. The bootstrap estimate of regret allows us to report confidence intervals for the true-game regret of empirical-game equilibria.

To compute a bootstrap distribution for the regret of a (mixed- or pure-strategy) profile $\vec{\sigma}$, we construct a large number of *bootstrap games* by simultaneously resampling

[1] We describe and test our methods using symmetric games to simplify exposition and because most EGTA studies have been on symmetric games. However, none of the techniques we describe depend on symmetry and all should be more broadly applicable.

the observations of every entry in the payoff matrix. For each bootstrap game, we construct a resampled observation set $\hat{\Theta}$ where for each payoff, we draw with replacement from its samples a resample set of equal size. We then construct a bootstrap game model $\mathcal{M}(\hat{\Theta})$ by applying the same procedure used to construct the empirical game (usually setting payoffs equal to sample averages).

Frequently, the payoffs for all strategies in a profile are observed in a single simulation. In this case, correlation between samples can be preserved by applying common indexing, where the resample is computed over indices $i \in \{1, \dots, n\}$. If i is drawn, then the i^{th} observation of each payoff is included in $\hat{\Theta}$. In the event that all payoffs in a game were sampled with common random numbers, correlations can similarly be preserved by common indexing across all payoffs in the game.

In each bootstrap game, we compute the regret of profile $\vec{\sigma}$, and across many bootstrap games, these values constitute a bootstrap distribution for the regret statistic. If $\vec{\sigma}$ is an equilibrium of the empirical game, we are generally interested in an upper bound on its true-game regret. The 95th percentile of the bootstrap distribution gives us a 95%-confidence upper bound for $\epsilon(\vec{\sigma})$. In Algorithm 1, we will also be interested in a two-sided confidence interval for $\epsilon(\vec{\sigma})$, which is estimated by the 2.5th and 97.5th percentiles of the bootstrap distribution.

3.2 Sampling Control

When game data is generated by a simulator, the practitioner may exert control over how many observations to gather of each profile. Historically, researchers have interwoven sampling and interim game-theoretic analysis to minimize the number of observations gathered in unproductive regions of profile space [13]. Similarly, it may be possible to reduce the number of observations taken even of relevant profiles while still delivering a baseline of statistical confidence by interweaving evaluation of regret confidence intervals with sampling. At the conclusion of a stage of sampling, a γ-level two-sided confidence interval on the regret of a profile expresses with greater than γ confidence that the profile is an ϵ-Nash equilibrium of the true game when the interval falls below ϵ; however, when the interval includes ϵ, we are unable to distinguish between having insufficient evidence that the claim is true and having sufficient evidence that the claim is false. Rather than report that we are uncertain whether or not a profile is an ϵ-equilibrium, we would often prefer to continue sampling until we have sufficient confidence to make a determination. To address this issue, we propose the Confidence-Interval-Based Stopping Rule (CIBSR) presented in Algorithm 1.

CIBSR is similar to the repeated confidence interval approach to terminating clinical trials proposed by Jennison and Turnbull [11], but utilizes the bootstrap to construct confidence intervals in place of parametric assumptions. The algorithm takes as arguments a candidate profile $\vec{\sigma}$, and any observations taken thus far Θ_{init}, and samples sequentially until there is sufficient evidence to decide whether or not the candidate profile is an ϵ-equilibrium of the true game. CIBSR is parameterized by the acceptable regret threshold ϵ, the confidence interval level γ, and the number of observations to gather of each relevant profile in each step x. CIBSR decides if a candidate is an ϵ-equilibrium by comparing the boundaries of a two-sided confidence interval[2] to ϵ, accepting the hypothesis that $\vec{\sigma}$ is an ϵ-equilibrium when the interval falls entirely within $[0, \epsilon]$, rejecting it when the interval falls entirely within (ϵ, ∞), and otherwise requesting further observations. Sampling under CIBSR is restricted to profiles that can affect the estimated regret distribution of the candidate. These profiles belong to either $\mathcal{S}(\vec{\sigma})$, the set of pure-strategy profiles that are realized with positive probability under the profile $\vec{\sigma}$, or $\bigcup_{\hat{\vec{\sigma}} \in \mathcal{D}(\vec{\sigma})} \mathcal{S}(\hat{\vec{\sigma}})$, the set of pure-strategy profiles that are realized with positive probability under some profile reachable from $\vec{\sigma}$ through a unilateral deviation to a pure strategy.

Algorithm 1 Confidence-Interval-Based Stopping Rule $(\vec{\sigma}, \Theta_{init}, \epsilon, \gamma, x)$

```
Require: $\vec{\sigma}$, the profile to evaluate
Require: $\Theta_{init}$, the observations used to identify $\vec{\sigma}$ as a candidate
Require: $\epsilon$, the acceptable approximation threshold
Require: $\gamma$, the confidence level to use
Require: $x$, the number of observations to take of each profile in each step
  $\Theta_{seq} \leftarrow \Theta_{init}$
  $[\epsilon_{left}, \epsilon_{right}] \leftarrow$ TWO-SIDED-REGRET-CI$(\vec{\sigma}, \Theta_{seq}, \gamma)$
  while $\epsilon_{left} < \epsilon$ and $\epsilon_{right} > \epsilon$ do
    Append $x$ observations of each profile $s \in \mathcal{S}(\vec{\sigma}) \cup (\bigcup_{\hat{\vec{\sigma}} \in \mathcal{D}(\vec{\sigma})} \mathcal{S}(\hat{\vec{\sigma}}))$ to $\Theta_{seq}$
    $[\epsilon_{left}, \epsilon_{right}] \leftarrow$ TWO-SIDED-REGRET-CI$(\vec{\sigma}, \Theta_{seq}, \gamma)$
  end while
  return $\epsilon_{right} \leq \epsilon$
```

Rather than simply wishing to determine if a particular profile is an equilibrium, practitioners may begin with no specific candidates, but sample from a simulator with the purpose of finding one or more equilibria of the true game. At any point in the sampling process, equilibria may be computed in the empirical game induced from the observations gathered thus far. As nothing is known about the payoff distributions prior to sampling, nor even which payoff distributions will be relevant for identifying equilibria of the game, practitioners typically sample sequentially according to rules of thumb, such as taking observations until the set of equilibria of the empirical game does not change with further sampling. Existing rules of thumb may reduce uncertainty in an indirect manner, but when such procedures terminate, no direct evidence can be provided of the regret of playing equilibria of the empirical game in the true game. Furthermore, since these stopping rules are typically very coarse heuristics, they may actually require more sampling than is necessary to have sufficient confidence that a profile is an ϵ-Nash equilibrium of the true game. We propose incorporating bootstrap confidence intervals into a sequential equilibrium finding procedure as in the Confidence-Interval-Based Equilibrium Finding (CIBEF) algorithm, presented in Algorithm 2.

At each step, CIBEF requests x additional observations of each profile and finds equilibria of the updated empirical game. For each equilibrium of the empirical game, a one-sided regret confidence interval at the γ-level is constructed,

[2] For all experiments we present, ONE-SIDED-REGRET-CI and TWO-SIDED-REGRET-CI implement the confidence interval methods described in Section 3.1.

Algorithm 2 Confidence-Interval-Based Equilibrium Finding (ϵ, γ, x)

```
Require: $\epsilon$, the acceptable approximation threshold
Require: $\gamma$, the confidence level to use
Require: $x$, the number of observations to take of each
  profile in each step
  $\Theta_{seq} \leftarrow \{\}$
  $E \leftarrow \{\}$
  while $E = \{\}$ do
    Append $x$ observations of each profile to $\Theta_{seq}$
    for $\vec{\sigma} \in \text{EQUILIBRIA}(\Theta_{seq})$ do
      if ONE-SIDED-REGRET-CI$(\vec{\sigma}, \Theta_{seq}, \gamma) \leq \epsilon$ then
        Append $\vec{\sigma}$ to $E$
      end if
    end for
  end while
  return $E$
```

and if the right-hand side of this interval is not greater than ϵ, the profile is appended to the set of equilibria. When one or more equilibria of the empirical game meet this criterion, sampling is terminated and the candidates meeting the criterion are returned.

4. EXPERIMENTS

4.1 Regret Bootstrap Experiments

For our bootstrap regret confidence intervals to be useful, we need to show that they are well-calibrated. We hypothesize that the 95th percentile of the sample-game bootstrap distribution of the regret of a candidate equilibrium provides an accurate 95% confidence bound for the true-game regret of that candidate. We are not aware of sufficient theoretical foundations to prove this hypothesis, so we test it experimentally.

4.1.1 Experimental Setup

Our hypothesis yields several testable predictions: most importantly, the confidence bound should be well-calibrated, namely the true-game regret of an equilibrium candidate should fall below the 95th percentile of the bootstrap distribution 95% of the time. We would also expect the other quantiles of the bootstrap distribution to be well-calibrated. In addition, confidence bounds should grow tighter as data is acquired, so the 95th percentile of the bootstrap distribution should shrink as the number of payoff observations grows. We also expect confidence bounds to be wider when data are more noisy, so the 95th percentile should grow as the variance of payoff samples grows.

We test these hypotheses by artificially generating true games and drawing samples from known noise distributions centered around each true game payoff. We then compute pure-strategy Nash equilibria and symmetric mixed-strategy Nash equilibria in the resulting empirical games[3], and use our bootstrap method to estimate the distribution of each candidate equilibrium's regret. These bootstrap estimates estimates are compared against the true-game regrets of the equilibrium candidates. Across a large number of randomly generated true games, our hypothesis predicts that k% of true-game regret values will fall below the k^{th} percentile of the empirical game's bootstrap regret distribution, especially when $k = 95$.

Game, Noise $\bar{z} = 100$	Size	.95 Frac. pure	.95 ϵ pure	.95 Frac. mixed	.95 ϵ mixed
uSym, normal	10	0.924	34.4	0.951	25.9
uSym, normal	100	0.947	1.5	0.955	6.3
uSym, bimodal	10	0.949	71.6	0.957	50.1
uSym, bimodal	100	0.935	13.5	0.949	12.7
Cgst, normal	10	0.928	20.6	0.966	18.1
Cgst, normal	100	0.972	0	0.941	1.8
CredNet, agg.	10	0.981	1.51	0.997	1.04
CredNet, agg.	100	0.971	0	0.927	0.23

Table 1: Bootstrap confidence intervals for regret.

Our experiments employ two classes of synthetic games: uniform symmetric games (uSym) and congestion games (Cgst), as well as one class of simulated game: credit network games (CredNet). To generate a true game from the uSym class, we draw a value from the distribution $U[0, 100]$ for each unique payoff in a symmetric game with $p \in \{2, 4\}$ players and $s \in \{2, 4, 6\}$ strategies. All results presented here use uSym games with 4 players and 4 strategies; results for other combinations of players and strategies are similar. To generate a true game from the Cgst class, we use 5 players and 3 strategies; each strategy s has a base value $v_b(s) \sim U[0, 3]$, a linear congestion cost $v_l(s) \sim U[0, 1]$, and a quadratic congestion cost $v_q(s) \sim [0, 1]$. The payoff to a player choosing strategy s is a function of the total number $n(s)$ of players choosing that strategy: $u(s) = v_b(s) - v_l(s)n(s) - v_q(s)(n(s))^2$. CredNet games are generated based on data from a simulator described by Dandekar, et al. [6]. In our initial experiments, we generated a CredNet game with 6 players, 6 strategies, and 2644 samples of each payoff, but found that it had particularly high variance; therefore we also generated a second data set with the same players and strategies called CredNet agg., where each of 1000 samples comes from 20 pre-aggregated runs of the simulator. The true game in our CredNet experiments is always the empirical game constructed using the full set of samples. To facilitate comparison of regret values across classes, we applied an affine transformation to rescale each uSym and Cgst payoff matrix to match range $[0, 100]$, which closely matches the payoff range of the CredNet true game.

Given a true game from the uSym or Cgst classes, we created noisy samples of each payoff by drawing from a known distribution centered at the true-game payoff and constructed empirical games from these sample sets. For Cgst, we added only normally distributed noise, but across uSym experiments we varied the noise distribution among normal, uniform, bimodal Gaussian mixture, and Gumbel. For both synthetic game classes, we varied the maximum-width parameter $\bar{z} \in \{0.1, 1, 10, 100\}$ over four orders of magnitude. For all experiments, we drew $z \sim U[0, \bar{z}]$ independently for every true-game payoff. For normally distributed noise, observations of each payoff are drawn from a normal distribution with variance z; for bimodal Gaussian, the z parameter controls the variance of the two Gaussians, which are spread apart by a random draw from $N(0, \bar{z})$; for uniform noise, z is equal to the half-width of the distribution; for Gumbel noise, z is the scale parameter. We also tested drawing noise from different models for different payoffs in the same game, and found similar results (not shown).

[3] In all of our experiments, mixed-strategy equilibria are computed using replicator dynamics [9].

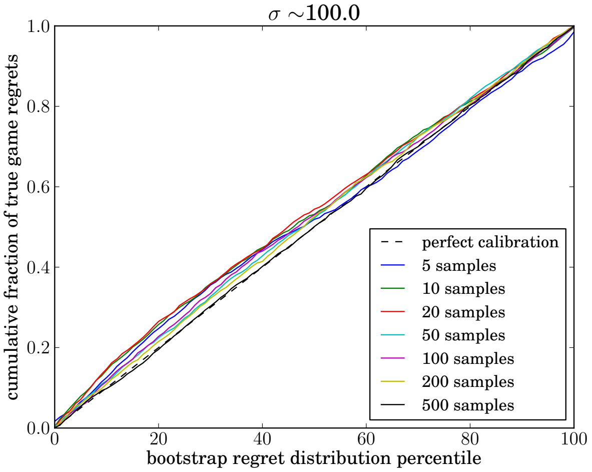

Figure 1: Bootstrap distribution calibration relative to the true game regret distribution in uSym games.

To create each sample game for the CredNet class, we selected a subsample without replacement out of the full set of simulator observations.

4.1.2 *Experimental Results*

For each combination of synthetic game class, number of players/strategies, and noise model, we generated 1000 true games. For each synthetic and simulated true game, we generated empirical games with sample sizes ranging from 5 to 500, and in each empirical game, we computed pure- and mixed-strategy Nash equilibria. We then computed bootstrap distributions for the regret of each equilibrium, which we compared to that equilibrium's true-game regret. Table 1 shows, in the ".95 Frac." columns, the fraction of true-game regrets that fell below the 95th percentile of the bootstrap distribution for a subset of game settings, noise models, and sample sizes. The data indicates that the 95th percentile of the regret bootstrap distribution provides a well-calibrated 95% confidence interval for true-game regret of equilibria computed in empirical games, especially in our synthetic games. Note that the 95th percentile of regret is sometimes very high, demonstrating the danger of reporting empirical game equilibria without statistical testing.

Table 1 also helps to support some of our secondary hypotheses: the ".95 ϵ" columns show the regret threshold growing tighter as the bootstrap gets more samples. Additional data (not shown) also demonstrates that the bootstrap confidence intervals are wider when variance is higher. Results for the other synthetic games, other noise magnitudes, and other numbers of samples are broadly similar.

While Table 1 shows good calibration for the 95% bootstrap confidence bound, we would hope that the whole bootstrap regret distribution, and not just the 95th percentile is well-calibrated. Figure 1 shows that for 4-player, 4-strategy uSym games with normal noise this is indeed the case: each curve shows the cumulative fractions of empirical game equilibria for which the true-game regret fell below each successive percentile of the bootstrap distribution. Because the curves closely track the 45° line, we conclude that on average the shape of the bootstrap distribution closely matches that of the sampling distribution for regret. Such plots for other synthetic games (not shown) appear similar: except with very small sample sizes or noise that completely swamps payoff variation (we ran one experiment with $\bar{z} = 1000$), calibration is consistently good. In our CredNet experiments, on the other hand, the cumulative fraction curves do not match the 45° line as closely and do not improve as much as the sample set grows. It is possible that this happens because we have only one true game for the CredNet class, from which each empirical game is a subsample. Unfortunately, within a reasonable sampling budget, we have no way to test this possibility. We show in Section 4.2.2 that we can still use the regret confidence interval from the bootstrap to make reasonable sampling decisions in CredNet games.

4.2 Bootstrap in Sample Control

Frequently, rules of thumb are used to decide when to terminate sampling in iterative applications of EGTA, such as sampling until the uncertainty in payoff estimates is below some threshold; however, the theoretical underpinnings of the bootstrap assume a fixed-sample statistical experiment, and as such may not deliver reliable inference when applied to EGTA applications where the sampling process is terminated based upon a feature of the gathered data. To empirically evaluate the usefulness of bootstrapping regret for iterative EGTA approaches, including its use in sample control, we conduct two sets of experiments. In the first set of experiments, we evaluate the accuracy of using bootstrapped regret confidence intervals to decide whether a mixed-strategy profile is an approximate equilibrium in the algorithm CIBSR. In the second set, we test the accuracy of bootstrapped regret confidence intervals at the termination of sampling with CIBEF and two commonly-applied rules of thumb to determine (i) whether the bootstrap approach is reliable when working with sequentially gathered data and (ii) whether CIBEF provides any reduction in sample costs or expected regret of candidate profiles when compared to existing rules of thumb.

4.2.1 *Labeling Equilibria*

To test the efficacy of CIBSR at labeling profiles as either ϵ-Nash equilibria (hereafter referred to as Eq) or not (hereafter referred to as Not-Eq), we measure the frequencies at which the algorithm correctly labels candidate profiles from synthetic or simulated game data, as in Section 4.1. For each game type considered, 1000 trials were run, where each trial consisted of labeling a single candidate profile, which may be either Eq or Not-Eq with respect to the true game. Similar to the experiments in Section 4.1, for synthetic games, each trial corresponds to a different randomly generated game, while a trial with simulation game data corresponds to a different random ordering of observations of a fixed data set, due to the cost of gathering additional simulation data. This means that for the simulated game trials, the set of approximate equilibria of the true game remains fixed across trials.

Selecting profiles randomly for evaluation would be an insufficiently stringent test for the algorithm, since Not-Eq profiles with high regret can be labeled with confidence with very few observations. Furthermore, such profiles would not merit application of statistical tools in practice, as a mixed-strategy profile is typically only of interest if it is believed to be a close approximation of equilibrium. We therefore construct candidate profiles by taking a small number of observations of each profile and computing an equilibrium of the current empirical game. These observations are then passed to the algorithm as Θ_{init}. With this procedure we are able to generate candidate profiles of either type, and Not-Eq instances will frequently be low regret, making correct labeling appropriately difficult.

For the synthetic games, candidates are selected after taking 5 observations of each profile. On each trial the algorithm is parameterized with a regret threshold $\epsilon = 0.05$ and step size of $x = 5$ observations. Each trial also has an observation cap of 1000 observations per profile, at which point, if the algorithm has not terminated, it labels the point as Eq if the median of the bootstrap distribution is below ϵ. For the credit network simulator, we explored both the unaggregated and aggregated data sets described in Section 4.1. Due to relatively high level of noise in the credit network data, candidates were chosen after 100 observations for the unaggregated data, and after 5 observations for the aggregated data.[4] Similarly, the algorithm is parameterized with $x = 100$ for the unaggregated data and $x = 5$ for the aggregated data, with observation caps set at the size of the full data set, 2644 and 1000 respectively. Experiments on these games were conducted for $\epsilon \in 0.05, 0.2, 0.5$, as different settings of ϵ lead to a different distribution of Eq and Not-Eq instances, and potentially change the difficulty of correct labeling. For all experiments presented here, the algorithm is parameterized with $\gamma = 0.95$. As the algorithm uses a two-sided γ-confidence interval and the final decision only depends on one of the boundaries of the interval, the algorithm terminates with a confidence level of 0.975.

Table 2 presents selected findings from these experiments. With the exception CredNet instances which use simulation data, all experiments reported here used normally distributed noise, with standard deviation given by σ in the table, to generate payoff observations. As in the prior experiments, for synthetic games $\sigma \in \{0.1, 1, 10, 100\}$ were evaluated, but only $\sigma = 1$ and $\sigma = 100$ are presented due to the similarity of the results. "Inst. Type" specifies whether the row refers to instances where the ground truth is Eq or Not-Eq. "#" specifies the number of instances out of the 1000 trials that were of the named type, while "Und. #" gives the number of trails for that type where the algorithm was undecided after reaching the observation cap imposed in the experiment. "Accuracy" lists first the fraction of labelings that were correct for the trials where the algorithm terminated with a confident decision, and second the fraction correct when the algorithm was forced to decide at the observation cap. "Agg. Acc." specifies the the fraction of labelings that were correct across all 1000 trials, including both Eq and Not-Eq instances.

Despite our specific choice of instances that were difficult to correctly label, CIBSR using the bootstrap method to construct confidence intervals delivered high levels of accuracy across all game types and parameter settings, with the lowest accuracy observed over a full set of trials being 0.922. On synthetic data, CIBSR delivered accuracy of at least 0.97 across all games and instance types. The credit network data proved more difficult, and despite overall high levels of accuracy, accuracy varied considerably between experiments and instance types. In particular, low regret thresholds made the algorithm less likely to label candidates as equilibria; this is reflected in the high accuracy for Non-Eq, low accuracy for Eq instances, and higher incidence of inconclusive results for Eq instances. The observation that the entire data set for the credit network game was frequently insufficient to have high confidence in declaring that a candidate profile was an ϵ-equilibrium reaffirms the difficulty of working with the CredNet data set.

Game	ϵ	Inst. Type	#	Und. #	Accuracy	Agg. Acc.
uSym	.05	Eq	624	22	.99, .77	.978
$\sigma = 1$		Not	376	21	.98, .76	
uSym	.05	Eq	240	7	.97, 1	.972
$\sigma = 100$		Not	760	11	.98 .36	
Cgst	.05	Eq	911	6	1, .67	.995
$\sigma = 1$		Not	89	1	.98, 1	
Cgst	.05	Eq	610	4	.99, .25	.984
$\sigma = 100$		Not	390	6	.98, .83	
CredNet	.05	Eq	96	84	0, .83	.974
		Not	904	250	1, 1	
CredNet	.2	Eq	642	558	.75, .90	.922
		Not	358	152	1, .99	
CredNet	.5	Eq	910	512	.99, .99	.985
		Not	90	56	.91, 1	
CredNet,	.05	Eq	428	122	.86, .98	.953
agg.		Not	572	64	.99, 1	
CredNet,	.2	Eq	774	156	.96, .95	.965
agg.		Not	226	18	.98, 1	
CredNet,	.5	Eq	997	128	.98, 1	.983
agg.		Not	3	1	1, 1	

Table 2: Sequential classification performance.

4.2.2 Sequential Equilibria Finding

The previous experiment demonstrated that the bootstrap method of constructing confidence intervals can successfully be used as terminating condition for a sequential statistical experiment without drastically biasing inferences drawn at the conclusion of the experiment. In contrast to the rules of thumb typically used to terminate iterative applications of EGTA, using a confidence-interval-based stopping rule yields statements about the likely values of the underlying regret; however, nothing precludes using the bootstrap method to estimate the regret distribution at the conclusion of an application of EGTA that used a rule of thumb to determine when to terminate sampling. In this section we conduct an experiment to evaluate the accuracy of the bootstrap method applied to the conclusion of sampling using two common rules of thumb, as well as CIBEF. Furthermore, we compare the performance of these rules in terms of average regret and number of observations requested.

The first rule of thumb considered is to cease sampling when all relevant payoff estimates demonstrate low variability; specifically, the stopping rule labeled SEM will request x further observations be made of each profile in each step until the estimated standard error in mean of each payoff is below a specified threshold ξ. The intuition behind this stopping rule is that reliable estimates of payoffs should lead to reliable inferences about the true game. The second rule we consider ceases sampling when the set of equilibria of the empirical game does not change with additional observations. This stopping rule, labeled EQC (for equilibrium comparison), will request x further observations be made for each profile until the sets of empirical game equilibria found in successive steps are equivalent. Distributions used

[4] Since each aggregated observation averages 20 observations, the candidate profiles for both experiments are constructed after the same number of simulations.

Game	Rule	Mean Obs.	Mean Regret	Median Regret	.95 Frac.
uSym	EQC	17.07	.0807	.0107	.94
$\sigma = 1$	SEM	5.02	.0073	.0014	.90
	CIBEF	5.08	.0081	.0018	.90
uSym	EQC	78.66	2.641	.6669	.96
$\sigma = 100$	SEM	1000.0	.5176	.1168	.96
	CIBEF	94.10	.9257	$1.98e{-}6$	.90
Cgst	EQC	10.04	.0006	$1.15e{-}7$	.92
$\sigma = 1$	SEM	5.01	.009	$2.26e{-}7$	.89
	CIBEF	5.01	.008	$2.31e{-}7$	.90
Cgst	EQC	20.06	.89	$1.44e{-}6$	.98
$\sigma = 100$	SEM	1000.00	.0771	$2.71e{-}6$	.94
	CIBEF	26.94	.1703	$1.58e{-}6$	.97
CredNet,	EQC	66.71	.0468	$1.03e{-}7$	.98
agg.	SEM	116.05	.0386	$1.96e{-}7$	.95
	CIBEF	11.10	.0295	$1.70e{-}7$	.99

Table 3: Stopping rule performance.

in mixed-strategy equilibria are considered to be equivalent if their Euclidean distance is below some threshold Δ. Both rules of thumb prescribe a stopping point at which equilibria of the empirical game are considered approximate equilibria of the true game, but provide no guarantee that the profiles that they identify are ϵ-equilibria for any particular ϵ.

In this experiment, a trial consists of the specified algorithm requesting observations from a synthetic or simulation-based game model until its stopping condition is triggered or an observation cap is reached, at which point it returns one or more equilibrium candidates. If the observation cap is reached, CIBEF returns the equilibrium of the empirical game with the lowest right-hand one-sided confidence bound, while SEM and EQC return all equilibria of the empirical game. For each candidate we record the regret of each candidate profile in the true game, as well as the number of observations taken of each profile prior to terminating the sampling procedure.[5] For each game model and algorithm, we ran 1000 trials, with each trial corresponding to a new random game for the synthetic game data, and new random reordering of the data for the simulated game data. In addition to metrics about the number of observations and regret of the selected profiles, we measure the average fraction of regret of the true games captured by the 95th-percentiles of bootstrapped regret distributions calculated at the termination of each trial, similar to Section 4.1. This measure gives an indication of the accuracy of using the bootstrap approach to give a confidence interval at the termination of sampling, when sampling is guided by these sample control algorithms.

Table 3 shows selected results from these experiments. "Mean Obs." refers to the average number of observations taken of each profile when the algorithm terminated. All trials were conducted with the same step sizes and observation caps as in Section 4.2.1. For the SEM stopping rule, the threshold ξ was set to 1.0, while for the EQC rule, a distance less than $\Delta = 0.01$ was considered sufficient to call two equilibrium candidates identical. For all experiments conducted with the CIBEF stopping rule, ϵ was set to 0.5 and γ was set to 0.95.

These results emphasize the general applicability of the bootstrap method of generating confidence intervals, even for sequential sampling experiments. The only experiments that resulted in substantial overconfidence on average, that is experiments in which the fraction of true game regrets captured by 95th percentile of the bootstrap regret distribution is less than 95%, were those settings in which sampling typically halted at the first opportunity. This outcome mirrors the results from Section 4.1, where we noted that for very small numbers of observations, the bootstrap method could be poorly calibrated. In contrast, some combinations of game models and stopping rules led to overly large confidence intervals, with greater than 95% of true game regrets being captured by the confidence interval method. In such cases the bootstrap method may be conservative in declaring candidates as equilibria, occasionally ruling out more profile candidates than is warranted by the expressed confidence level; however, using CIBEF an equilibrium meeting the confidence requirements will eventually be found.

In comparing the equilibrium-finding characteristics of the three stopping rules, our experiments show that CIBEF is typically comparable to and often an improvement over existing rules of thumb. For every game model considered, the median regret of the profiles returned by CIBEF was considerably below the approximation threshold of $\epsilon = 0.5$. Similarly, the mean regret was below the threshold for all but the uniform symmetric game model with the highest variance. Our data suggests that in most scenarios when CIBEF misidentifies a profile as an approximate Nash equilibrium it is still likely to have regret close to the threshold, but for high noise settings, profiles that are incorrectly returned may have significant regret. EQC nearly always yielded higher regret profiles than CIBEF, and performed particularly poorly in the synthetic game models with high variance. SEM frequently performed at the same level or better than CIBEF in terms of regret, but almost always terminated either immediately or at the observation cap. In terms of the number of observations taken prior to stopping, CIBEF was similar to SEM for low noise settings, but required many fewer observations for high noise settings. In contrast, CIBEF required fewer observations than EQC in low noise settings, in part due to EQC requiring two sampling steps prior to terminating, but required slightly more samples in noisier settings.

CIBEF outperformed EQC and SEM in terms of mean regret of returned candidates and the average number of observations taken prior to stopping on the aggregated credit network data. All rules performed excellently in terms of median regret, meaning that either CIBEF returned fewer non-equilibrium candidates or that the non-equilibrium candidates that it returned were closer approximations to equilibrium than the other two stopping rules. Here, CIBEF may benefit from often returning one candidate that is highly likely to be an equilibrium, rather than returning multiple candidates that may vary greatly in how well they approximate equilibria, as in EQC or SEM. As such, CIBEF can deliver significant savings in terms of sampling costs when finding only one equilibrium is acceptable. We were, however, unable to present results for the credit network game with unaggregated data, as this experiment proved too costly, particularly for CIBEF, as it must find equilibria and calculate confidence intervals for them in every sampling step. Though a potential detriment to CIBEF, in real applications of EGTA the cost of sampling will typically outweigh

[5] All algorithms considered here sample evenly from all profiles in a game's profile space, so "10 observations" refers to 10 observations taken of every value in the payoff matrix.

the cost of calculating confidence intervals, thus making the overhead of using CIBEF over EQC negligible.

5. CONCLUSION

Our experimental evidence demonstrates that the bootstrap method of confidence interval generation is approximately accurate for bounding the true-game regret of candidate equilibria in empirical games. Despite possible repeated testing concerns, our bootstrap method also proves approximately accurate in constructing confidence intervals on regret at the conclusion of sequential sampling procedures. Accuracy is lower in our experiments on credit network games than on randomly-generated games, but we believe that our the random game experiments may be more representative of EGTA in practice, due to limitations of conducting experiments with costly simulation data. Because our experiments show bootstrap confidence intervals to be accurate across multiple synthetic game classes, as well as across noise distributions and magnitudes, we recommend that practitioners of empirical game theory employ bootstrap methods to give regret bounds for reported equilibria. This recommendation stands even if bootstrap regret estimates are used to determine when to conclude sampling.

In addition, we provide evidence that the Confidence-Interval-Based Equilibrium Finding algorithm improves over previous EGTA experiment designs. Relative to rules of thumb for sample control, CIBEF can more consistently identify low regret equilibrium candidates, and often requires fewer observations. However, in games with particularly noisy payoffs, where misclassifications can be particularly egregious in terms of regret, we recommend that practitioners err on the side of caution and collect extra observations. Given the savings in observations CIBEF demonstrated, we believe that using bootstrapped confidence intervals on regret is a promising tool for sample control in this domain.

This work constitutes a first systematic effort to develop practical statistical methods for EGTA; future work could focus on developing theoretical foundations of applying the bootstrap to empirical games, and characterizing games for which the bootstrap approach is reliable. Additionally, there are other measures of interest to EGTA practitioners, such as social welfare, that may also benefit from using the bootstrap for statistical analysis. Other avenues of research include evaluating different bootstrap designs, and using information obtained through the bootstrap to guide more sophisticated sampling, such as profile exploration [13].

6. REFERENCES

[1] R. Axtell, R. Axelrod, J. M. Epstein, and M. D. Cohen. Aligning simulation models: A case study and results. *Computational and Mathematical Organization Theory*, 1(2):123–141, 1996.

[2] T. Baarslag, K. Fujita, E. H. Gerding, K. Hindriks, T. Ito, N. R. Jennings, C. Jonker, S. Kraus, R. Lin, V. Robu, and C. R. Williams. Evaluating practical negotiating agents: Results and analysis of the 2011 international competition. *Artificial Intelligence*, 198:73–103, 2013.

[3] R. Bender and S. Lange. Adjusting for multiple testing: When and how? *Journal of Clinical Epidemiology*, 54(4):343–349, 2001.

[4] T. F. Bresnahan and P. C. Reiss. Empirical models of discrete games. *Journal of Econometrics*, 48(1-2):57–81, 1991.

[5] B.-A. Cassell and M. P. Wellman. Asset pricing under ambiguous information: An empirical game-theoretic analysis. *Computational and Mathematical Organization Theory*, 18(4):445–462, 2012.

[6] P. Dandekar, A. Goel, M. P. Wellman, and B. Wiedenbeck. Strategic formation of credit networks. In *21st international conference on World Wide Web*, pages 559–568, 2012.

[7] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application*. Cambridge University Press, 1997.

[8] L. W. Friedman and H. H. Friedman. Analyzing simulation output using the bootstrap method. *Simulation*, 64(2):95–100, 1995.

[9] D. Fudenberg and D. K. Levine. *The Theory of Learning in Games*. MIT Press, 1998.

[10] X. A. Gao and A. Pfeffer. Learning game representations from data using rationality constraints. In *26th Conference on Uncertainty in Artifical Intelligence*, pages 185–192, Catalina Island, CA, 2010.

[11] C. Jennison and B. W. Turnbull. Repeated confidence intervals for group sequential clinical trials. *Controlled Clinical Trials*, 5(1):33–45, 1984.

[12] P. R. Jordan, C. Kiekintveld, and M. P. Wellman. Empirical game-theoretic analysis of the TAC supply chain game. In *Sixth International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 1188–1195, Honolulu, Hawaii, 2007.

[13] P. R. Jordan, Y. Vorobeychik, and M. P. Wellman. Searching for approximate equilibria in empirical games. In *Seventh International Conference on Autonomous Agents and Multi-Agent Systems*, pages 1063–1070, Estoril, Portugal, 2008.

[14] S. Phelps, P. McBurney, and S. Parsons. A novel method for automatic strategy acquisition and its application to a double-auction market game. *IEEE Transactions on Systems, Man, and Cybernetics: Part B*, 40:668–674, 2010.

[15] D. M. Reeves. *Generating Trading Agent Strategies: Analytic and Empirical Methods for Infinite and Large Games*. PhD thesis, University of Michigan, 2005.

[16] Y. Vorobeychik. Probabilistic analysis of simulation-based games. *ACM Transactions on Modeling and Computer Simulation*, 20(3), September 2010.

[17] M. P. Wellman. Methods for empirical game-theoretic analysis. In *21st National Conference on Artificial Intelligence*, pages 1552–1555, Boston, Massachusetts, 2006.

[18] M. P. Wellman, A. Osepayshvili, J. K. MacKie-Mason, and D. M. Reeves. Bidding strategies for simultaneous ascending auctions. *B. E. Journal of Theoretical Economics (Topics)*, 8(1), 2008.