Strategy Exploration in Empirical Games

Patrick R. Jordan[*] L. Julian Schvartzman[†] Michael P. Wellman
Computer Science & Engineering
University of Michigan
Ann Arbor, MI 48109-2121 USA
{prjordan,lschvart,wellman}@umich.edu

ABSTRACT

Empirical analyses of complex games necessarily focus on a restricted set of strategies, and thus the value of empirical game models depends on effective methods for selectively exploring a space of strategies. We formulate an iterative framework for strategy exploration, and experimentally evaluate an array of generic exploration policies on three games: one infinite game with known analytic solution, and two relatively large empirical games generated by simulation. Policies based on iteratively finding a beneficial deviation or best response to the minimum-regret profile among previously explored strategies perform generally well on the profile-regret measure, although we find that some stochastic introduction of suboptimal responses can often lead to more effective exploration in early stages of the process. A novel formation-based policy performs well on all measures by producing low-regret approximate formations earlier than the deviation-based policies.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Multiagent systems*

General Terms

Economics

Keywords

Empirical game theory, strategy exploration

1. INTRODUCTION

Often the most difficult obstacle to game-theoretic analysis of complex scenarios is developing a model of the game situation. In the *empirical game-theoretic analysis* (EGTA) approach [Wellman, 2006], expert modeling is augmented by empirical sources of knowledge: data obtained through real-world observations or (as emphasized here) outcomes of high-fidelity simulation. Simulation models employ procedural descriptions of strategic environments, which are often much easier to specify than declarative domain models. Prior work has developed an extensive EGTA methodology, where techniques from simulation, search, and statistics combine with game-theoretic concepts to characterize strategic properties of a domain.

A high-level view of the EGTA process is presented in Figure 1. The diagram highlights the iterative nature of EGTA. The basic step is simulation of a strategy profile (vector of strategies, one for each player), determining a payoff observation (i.e., a sample drawn from the outcome distribution induced by stochastic elements of the simulation environment), which gets added to the database of payoffs. Based on the accumulated data, we induce an empirical game model.

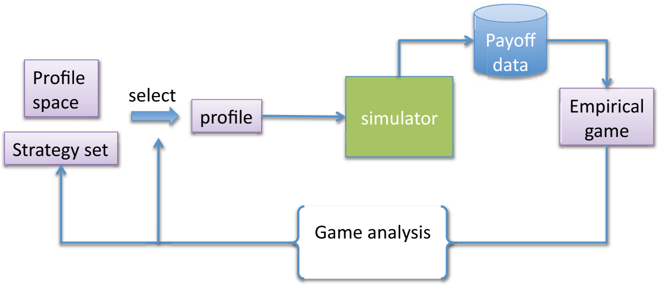

Figure 1: Dynamic game formulation through empirical game-theoretic analysis.

The EGTA process naturally supports a dynamic view of game formulation. Though the full strategy space allowed by the simulator may be large or infinite, due to computational constraints we can generally obtain direct outcome observations for a finite (and limited) set of profiles. Therefore, it makes sense to start from the most salient strategy candidates at first, incrementally adding candidates based on intermediate analysis results. For example, we might first solve a fairly restricted version of the game, admitting only a small slice of conceivable strategies. Based on these results, we could then generate additional strategy proposals to be added to the candidate set. Further simulation and analysis produces solutions for an expanded game, which then represents the starting point for subsequent rounds of refinement.

We focus here on one step in this process, namely the selection of strategies to add to the current candidate set: the

[*] Currently at Yahoo! Labs.

[†] Currently at Bank of America.

Cite as: Strategy Exploration in Empirical Games, Patrick R. Jordan, L. Julian Schvartzman, and Michael P. Wellman, *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. 1131-1138

method's *strategy exploration policy.* This problem bears a strong resemblance to the *profile selection problem* [Jordan et al., 2008] and the equilibrium finding algorithms proposed by Zinkevich et al. [2007]. In this work, we attempt to determine the optimal simulation sequence of strategy sets, instead of a sequence of profiles. Because simulating a set of strategies involves simulating all of its corresponding profiles, the *strategy exploration problem* is a special case of the profile selection problem.

2. EMPIRICAL GAMES

A **strategic game** $\Gamma = \langle N, (S_i), (u_i) \rangle$ consists of a finite set of players, N indexed by i; a non-empty set of strategies S_i for each player; and a utility function $u_i : \times_{j \in N} S_j \to \mathbb{R}$ for each player. A **symmetric game** satisfies $S_i = S_j$ and $u_i(\cdot) = u_j(\cdot)$ for every $i, j \in N$. For simplicity, we assume symmetric games in this paper, however extending the methods and analysis to non-symmetric games is straightforward. Let $\Gamma_{S \downarrow X}$ be a **restricted game** with respect to the **base game** Γ, where each player $i \in N$ in $\Gamma_{S \downarrow X}$ is restricted to playing strategies in $X_i \subseteq S_i$.

Each profile s is associated with the set of neighboring profiles that can be reached through a unilateral deviation by a player. The **unilateral deviation set** for player i and profile $s \in S$ is $\mathcal{D}_i(s) = \{(\hat{s}_i, s_{-i}) : \hat{s}_i \in S_i\}$, and the corresponding set unspecified by player is $\mathcal{D}(s) = \cup_{i \in N} \mathcal{D}_i(s)$.

Let $\Delta(\cdot)$ represent the probability simplex over a set. A **mixed strategy** σ_i is a probability distribution over strategies in S_i, with $\sigma_i(s_i)$ denoting the probability player i will play strategy s_i. The **mixed strategy space** for player i is given by $\Delta_i = \Delta(S_i)$. Similarly, $\Delta^\Gamma = \times_{i \in N} \Delta_i$ is the **mixed profile space**.

For a given player i, the **best-response correspondence** for a given profile σ is the set of strategies which yield the maximum payoff, holding the other players' strategies constant. The player i **best-response correspondence** for opponent profile $\sigma_{-i} \in \Delta(S_{-i})$ is

$$\mathcal{B}_i(\sigma_{-i}) = \underset{\hat{\sigma}_i \in \Delta_i}{\arg\max}\, u_i(\hat{\sigma}_i\ ,\ \sigma_{-i})$$

and for $\Delta \subseteq \Delta(S_{-i})$ is $\mathcal{B}_i(\Delta) = \times_{\sigma_{-i} \in \Delta} \mathcal{B}_i(\sigma_{-i})$. The **overall best-response correspondence** for profile $\sigma \in \Delta(S)$ is $\mathcal{B}(\sigma) = \times_{i \in N} \mathcal{B}_i(\sigma_{-i})$ and for $\Delta \subseteq \Delta(S)$ is $\mathcal{B}(\Delta) = \times_{\sigma \in \Delta} \mathcal{B}(\sigma)$. A **Nash equilibrium** (NE) is a profile $\sigma \in \Delta^\Gamma$ such that $\sigma \in \mathcal{B}(\sigma)$.

We use the symbols $\mathfrak{B}_i(\sigma_{-i})$ and $\mathfrak{B}_i(\Delta)$ to represent the pure-strategy variants of the best-response correspondences: $\mathfrak{B}_i(\sigma_{-i}) = \mathcal{B}_i(\sigma_{-i}) \cap S_i$ and $\mathfrak{B}_i(\Delta) = \mathcal{B}_i(\Delta) \cap S_i$. We also introduce symbols for the **pure-strategy best-response** to a set of profiles $X = \times_{i \in N} X_i$ where $\emptyset \subset X_i \subseteq S_i$: $\mathfrak{B}_i^\dagger(X_{-i}) = \mathfrak{B}_i\left(\Delta(X_{-i})\right)$ and $\mathfrak{B}^\dagger(X) = \times_{i \in N} \mathfrak{B}_i^\dagger(X_{-i})$. Under $\mathfrak{B}^\dagger(\cdot)$, each player's strategies are best responses to joint (correlated) mixtures over opponent strategies. A set of profiles $X \subseteq S$ is a **formation** [Harsanyi and Selten, 1988] if $\mathfrak{B}^\dagger(X) \subseteq X$; it is a **primitive formation** if no proper subset of X is a formation. We use the term *minimal formation* synonymously with primitive formation.

Strategy s_i' is an **improving deviation** for agent i with respect to profile σ if i would benefit by playing s_i' rather than its designated strategy in σ: $u_i(s_i', \sigma_{-i}) > u_i(s_i, \sigma_{-i})$. Let $\mathfrak{D}_i(\sigma)$ be the **set of improving deviations** with respect to σ for player i.

The regret measures described in this section quantifies the stability of strategies and profiles, respectively. A player's **regret**, $\delta_i(\sigma_i|\sigma_{-i})$, for playing strategy $\sigma_i \in \Delta_i$ against opponent profile $\sigma_{-i} \in \Delta(S_{-i})$ is $\max_{s_i \in S_i} u_i(s_i, \sigma_{-i}) - u_i(\sigma_i, \sigma_{-i})$. Finally, we use the regret of the constituent strategies to define the regret of a profile. The **regret of profile** $\sigma \in \Delta$, is the maximum gain from deviation from σ by any player. Formally, $\epsilon(\sigma) = \max_{i \in N} \delta_i(\sigma_i|\sigma_{-i})$. A Nash equilibrium σ has no regret, i.e., $\epsilon(\sigma) = 0$.

We can define a similar notions of regret for strategy sets. Let $\widehat{U}_i$ be the function that returns the best-response utilities of player i for each $\sigma_{-i} \in S_{-i}$ when player i's strategy set is limited to X_i. Thus, $\widehat{U}_i(\sigma_{-i}; X_i) = \max\{u_i(s_i, \sigma_{-i}) | s_i \in X_i\}$. For $\emptyset \subset X_i \subseteq S_i$, the *regret* of player i for having the restricted strategy set X_i against $\Delta(X_{-i})$ is $\delta_i(X_i|X_{-i}) = \max_{\sigma_{-i} \in \Delta(X_{-i})} \widehat{U}_i(\sigma_{-i}; S_i) - \widehat{U}_i(\sigma_{-i}; X_i)$. For $\emptyset \subset X \subseteq S$, the *regret* over all players for having restricted joint strategy set X is $\epsilon(X) = \max_{i \in N} \delta_i(X_i|X_{-i})$. A set of profiles $X \subseteq P$ is an ϵ-**formation** if $\epsilon(X) \leq \epsilon$.

3. STRATEGY EXPLORATION PROBLEM

The issue of strategy exploration is one facet of the broader problem of how to allocate simulation resources across the profile space. In the most general form, a policy for the resource allocation problem determines a *sequence of profiles* $\{s^{(j)}\}_{j=1}^k$ to simulate. Existing work on the *profile selection problem* (see Jordan et al. [2008] for a comprehensive review) reformulates the selection problem as a *search problem*, where the goal of search is to identify a low regret profile. In contrast to these models, the strategy exploration problem focuses on determining a *sequence of restricted-games* $\Gamma_{S \downarrow X^{(0)}}, \ldots, \Gamma_{S \downarrow X^{(k)}}$ to be simulated, where $X^{(0)} \subset \cdots \subset X^{(k)} \subseteq S$. Each restricted game $X^{(j+1)}$ is formed by adding an additional strategy to some player's restricted strategy set in $X^{(j)}$. Although fine-grained control of profile sampling is a more general perspective, we note that in practice dynamic modification of the strategy set is often deliberately controlled (usually manually), and is viewed as a significant and distinct decision. In typical studies reporting substantial empirical-game analyses [Kephart and Greenwald, 2002, Phelps et al., 2006, Wellman et al., 2007, 2008], the strategy set is hand-selected, and—though the underlying process is not always detailed in published reports—often extended iteratively in the course of the study. As each strategy is added, the analysis proceeds to explore (often but not always exhaustively) the expanded profile space. Since the profile space grows exponentially in strategies, and adding a strategy is an (implied) commitment to evaluate it adequately, strategies to add must be considered carefully.

We describe policies for the **revealed-payoff model** of observation [Jordan et al., 2008], in which each observation determines the true payoff for a designated pure-strategy profile. In this case, *simulating a restricted game* means observing the payoffs for each pure-strategy profile in $X^{(j)}$. Some of the policies we introduce require access to some payoffs outside of $X^{(j)}$ for combinations of a candidate strategy and the current equilibrium. Computing these payoffs will require some additional simulation, but far short of what would be entailed to fill out the profile space if the candidate is actually selected.[1] With the exception of the ran-

[1] This is true assuming that the support of the current equilibrium is much smaller than $X^{(j)}$.

dom (RND) strategy exploration policy, in the worst case the policies require knowledge of the payoffs for all single player deviations from profiles in $X^{(j)}$ to each of the remaining strategies. In other words, the utility functions for each player i must be defined over $S_i \times X_{-i}^{(j)}$ for the j^{th} step in the exploration. We define a concept that encapsulates this model of a game, called an *augmented restricted-game*. Let $\Gamma^{*}_{S\downarrow X}$ be an **augmented restricted-game** with respect to the *base game* Γ, where players in $\Gamma^{*}_{S\downarrow X}$ are restricted to playing profiles in $X \subseteq S$, however the utility function for each player i is a mapping $u_i : S_i \times X_{-i} \rightarrow \mathbb{R}$.

We can calculate the base-game regret of any profile in $\Gamma_{S\downarrow X}$ by calculating its restricted-game regret, if X is a formation. In addition, we know that our estimate can be understating the base-game regret by no more than ϵ, if X is an ϵ-formation. However, given the augmented restricted-game $\Gamma^{*}_{S\downarrow X}$, we can calculate the base-game regret for any profile in $\Gamma_{S\downarrow X}$. Furthermore, given $\Gamma^{*}_{S\downarrow X}$, we can calculate $\epsilon(X)$ without any additional profile observations. Therefore, we consider a variation of the strategy exploration problem that determines a sequence of *augmented restricted-games* $\Gamma^{*}_{S\downarrow X^{(0)}}, \ldots, \Gamma^{*}_{S\downarrow X^{(k)}}$ to be simulated, where $X^{(0)} \subset \cdots \subset X^{(k)} \subseteq S$.

How should we evaluate a candidate strategy exploration policy? Presumably, we are interested in solutions to the true game, and some strategies are more critical to determining these solutions than others. Thus, we seek policies that will introduce these strategies as early as possible. For example, if the true solution involves strategies S^* (e.g., a Nash equilibrium with support on S^*), we might evaluate a policy based on how many iterations it takes to cover this set. However, this approach treats finding a "solution" as an all-or-none matter, and fails to consider the usefulness of intermediate results. Therefore, we prefer a measure that captures degrees of quality of results at all steps of the iterative process. For this we appeal to the concepts of *regret* introduced previously.

To evaluate the quality of an empirical game model, we propose two evaluation metrics. In the first, we solve the model $\Gamma_{S\downarrow X}$ by employing our solution concept of choice (e.g., identifying a sample Nash equilibrium), and measure the regret of this solution profile with respect to Γ, the "true" or *base* game.[2] Intuitively, this captures the quality of the profile we would propose if we had to stop at the current iteration. A profile with regret ϵ constitutes an approximate, ϵ-Nash equilibrium, with $\epsilon = 0$ corresponding to exact equilibrium. All else equal, we consider profiles with smaller $\epsilon(s)$ to be more stable, and thus more plausible as plays of the actual game. Therefore, our objective is to find a minimum-regret profile.

Observe that a minimum-regret profile in $\Gamma^{*}_{S\downarrow X}$ may not be a Nash equilibrium in $\Gamma_{S\downarrow X}$. The second metric we consider measures $\epsilon(X)$ for the selected $\Gamma^{*}_{S\downarrow X}$. This metric is appropriate, for instance, when we care about the regret of a set of profiles (a set-valued solution). If $X \subset S$ is a formation, then, as analysts, we can restrict out attention to $\Gamma_{S\downarrow X}$ without risk of understating *any* regret values for the profile set X, not just solution profiles. Therefore, our objective is to find a minimum-regret ϵ-formation. Note that these objectives are not always aligned.

4. DEVIATION POLICIES

Consider the example two-player game presented in normal form in Table 1. There are four available strategies, $S = \{1, 2, 3, 4\}$. The strategy exploration problem asks in which order to introduce the strategies to our empirical game analysis. Introducing strategy 1 first, for example, would produce the solution profile $(1, 1)$ after the first iteration, which has a regret $\epsilon((1, 1)) = 3$.

	1	2	3	4
1	1,1	1,2	1,3	1,4
2	2,1	2,2	2,3	2,6
3	3,1	3,2	3,3	3,8
4	4,1	6,2	8,3	4,4

Table 1: An example symmetric two-player game of 4 strategies. Exploring strategies in the sequence (1,2,3,4) yields increasing regrets until the last step.

Note that regardless of the ordering, once $X = S$, equilibria in the restricted game and base game coincide, so regret is zero. Thus, we might expect that regret would tend to start high, and decrease progressively until reaching zero in the last step. This is not necessarily the case, however. For example, suppose we introduce strategies in the order (1,2,3,4). The sequence of regrets we observe would be (3,4,5,0), which increases monotonically until inevitably falling to zero at the end.

Thus, in the worst case it will be difficult to guarantee progress during intermediate steps of the EGTA process. Rather than dwell on this worst case, however, we consider it more useful to compare alternative exploration policies in *expectation*, with respect to random choices they may make. For example, consider the following possible exploration policies:

- Random (**RND**). Pick one of the remaining strategies with equal probability.
- Improving deviations only (**DEV**). Find a Nash equilibrium, σ, of the current restricted game, and choose a strategy uniformly from $\mathfrak{D}(\sigma) \setminus X$.
- Best response (**BR**). Find a Nash equilibrium, σ, of the current restricted game, and choose a strategy uniformly from $\mathcal{B}(\sigma) \setminus X$.

Note that DEV and BR build on analysis of the current restricted game; however, on the first iteration, DEV and BR choose randomly. These policies would also choose randomly if there are no improving deviations among the unexplored strategies — in which case we already have a equilibrium in Γ anyway.

We can evaluate each of these policies on the example game of Table 1. Since the game is so simple, we can calculate the expected regrets exactly, as shown in Table 2. From the table, we can see that expected regret does indeed decrease, under all three policies, as more strategies are explored. Moreover, limiting exploration to deviations (DEV)

[2] Of course, we cannot perform this evaluation in the context of an actual EGTA exercise, where the true game is unknown. All references to evaluation here are from the perspective of experimentally evaluating solutions to the strategy exploration problem.

dominates (at least as good in expectation at each step) random choice (RND), and picking best responses (BR) is the best of the three policies.

Step	Expected regret		
	RND	**DEV**	**BR**
1	3.000	3.000	3.000
2	2.333	1.375	0.000
3	1.250	0.208	0.000
4	0.000	0.000	0.000

Table 2: Expected regret under three exploration policies for the example game.

In addition to the three policies (RND, DEV, BR) introduced above, we consider the following exploration policies for analysis:

- Alternating (**BR+DEV**). Apply BR and DEV, in turn, on successive iterations.
- Softmax (**ST**). Find a Nash equilibrium σ of the current restricted game. Choose strategy s'_i from $\mathfrak{D}(\sigma)\backslash X$ with probability given by the softmax formula applied to deviation gains: $\alpha e^{(u(s'_i,\sigma_{-i})-u(\sigma))/\tau}$ where τ is the typical *temperature* parameter and α is a normalizing constant. Low values of τ mimic a best response (i.e., ST approximates BR), whereas $\tau \rightarrow \infty$ turns the selection equiprobable (i.e., ST approximates DEV).[3]

5. MINIMUM-REGRET PROFILES

Determining the profile with minimum regret in a restricted strategy space is fundamental to the modified best-response policy of the previous section and essential to evaluating policies on the first evaluation metric. We identify a minimum-regret constrained-profile (MRCP) by solving the following optimization problem:

$$\arg\min_{\sigma \in \Delta^{\Gamma^{*}_{S\downarrow X}}} \epsilon(\sigma).$$

This is a constrained optimization problem with a nonlinear, non-differentiable objective function and both inequality and equality constraints. Because the regret function is non-differentiable, standard optimization techniques that calculate the gradient of the Lagrangian do not apply. In lieu of gradient-based techniques, various *direct search algorithms* have been proposed to solve optimization problems where a gradient is not available or efficiently calculable. In practice, one of the most popular direct search algorithms is the *amoeba method* [Nelder and Mead, 1965]. This method iteratively refines a simplex in the search space until convergence or some fail condition is reached. Walsh et al. [2002] used the amoeba method to calculate Nash equilibria in an empirical game representing a continuous double auction scenario.

What about maintaining feasibility? When applying the amoeba method to the MRCP optimization problem, we have to reconcile the fact that the optimization problem is constrained and the amoeba method is an unconstrained optimization technique. However, if an iteration starts with a simplex where each vertex is within the feasible region, then we can modify the amoeba method such that we always end the iteration with a feasible simplex. To do this, we modify the reflect and expand steps of the original amoeba method to generate only feasible vertices. We use a binary search to select the maximum feasible reflection (α) and expansion (γ) scaling parameters, respectively, if the unmodified reflected or expanded vertex is infeasible. Using this approach, all the vertices of the simplex are feasible at the end of each iteration.

We make an additional observation. Minimum regret profiles may not have full support. In fact, minimum regret profiles will often have small support, a feature that is exploited by the equilibrium finding algorithms of Porter et al. [2008]. If the minimum regret profile does not have full support, then the inequality constraints on σ are active at the minimum. When the simplex approaches the boundary, it collapses to near zero measure. When this occurs, the amoeba algorithm may not recover to find a local minimum.

One way to deal with this problem is to search over restricted supports if the amoeba algorithm degenerates. The policies repeated compute the MRCP of different restricted games. The simplex will often degenerate to a support in which the MRCP has already been calculated. Using memoization we can dramatically improve performance.

6. FORMATION POLICIES

In Section 4 we defined an initial set of strategy exploration policies. Of those, the best-response (BR) policy is one of the most straightforward and intuitive. At the k^{th} iteration, pick a Nash equilibrium in $\Gamma_{S\downarrow X^{(k)}}$. Choose a best-response strategy from among the remaining strategies and use that strategy to construct $X^{(k+1)}$. The basic idea behind BR is successively refuting the current minimum-regret profile. This is a common theme among profile search algorithms like MRFS, EVI, and IGS [Jordan et al., 2008] . Our next strategy exploration policy is the natural extension of this idea to refuting the best ϵ-formation. The minimum-ϵ-maximum-τ (**MEMT**) strategy exploration policy chooses a strategy that maximizes the gain (τ) to deviating from a minimum-ϵ formation. The complete procedure is given in Algorithm 1. The algorithm works in two stages. The first stage selects a minimum-ϵ formation [Jordan, 2009]. The second stage selects a strategy that maximizes the τ of the minimum-ϵ formation.

The minimum-ϵ formation, denoted by $X_{\min}$, may have a τ-maximizing strategy, denoted by s_i, that is already in X. In this case, we place $X_{\min}$ in a tabu list, denoted by T. Subsequent calls to FIND-FORMATION will return the minimum-ϵ that is not in T or *NULL* if all formations are in T. This process continues until a τ-maximizing strategy is found that is not in X.

7. EXPERIMENTS

Our experimental approach follows the process we introduced in Section 4. We start with a base game Γ, and compare the results of applying various strategy exploration policies. We evaluate policies based on the two evaluation metrics given in Section 3: *expected minimum profile-regret* and *expected minimum formation-regret*.

To evaluate the expected minimum-regret, we select a minimum-regret profile after each iteration of a policy. We

[3] So that the temperature settings are meaningful across games, we employ normalized payoffs in computing gains.

Algorithm 1 MEMT($\Gamma^{*}_{S\downarrow X}$)

```
$T \leftarrow \emptyset$
$X_{\min} \leftarrow$ FIND-FORMATION($\Gamma^{*}_{S\downarrow X}, |X|, T$)
$\widehat{s_i} \leftarrow$ NULL
while $\widehat{s_i}$ is NULL and $X_{min}$ is not NULL do
    $s_i \leftarrow$ FIND-TAU($\Gamma^{*}_{S\downarrow X}, X_{\min}$)
    if $s_i \notin X$ then
        $\widehat{s_i} \leftarrow s_i$
    else
        $T \leftarrow T \cup \{X_{\min}\}$
        $X_{\min} \leftarrow$ FIND-FORMATION($\Gamma^{*}_{S\downarrow X}, |X|, T$)
if $\widehat{s_i}$ is NULL then
    $\widehat{s_i} \leftarrow$ FIND-TAU($\Gamma^{*}_{S\downarrow X}, X$)
return $\widehat{s_i}$
```

evaluate the policies on three games. The first is a two-player game based on the first-price sealed-bid auction (FPSB). This game has an infinite strategy space, but is convenient for analysis because we have known analytic forms for its payoff and best-response functions. Unfortunately, we do not have analytic forms for computing formation-regret and a MRCP, so we forgo analysis of the augmented policies in this scenario. Reeves [2005] extensively studied this game as a test for EGTA methods, and we build on his results to conduct our investigation of strategy exploration.

Our second test is the four-player empirical game generated in a recent study of CDA bidding strategies [Schvartzman and Wellman, 2009]. The empirical CDA game comprises 13 strategies, including strategies from the literature as well as some derived by reinforcement learning as part of the study.

Our final test is another empirical game, this one based on the Trading Agent Competition (TAC) Travel game [Wellman et al., 2007]. This version is a two-player model with 35 strategies, constructed manually with no explicit exploration policy.

The CDA and TAC games are most representative of domains we expect to subject to empirical game analysis. Our experiments in these domains are limited, however, to exploring subsets of those strategies actually introduced in the respective EGTA studies. The FPSB example provides the advantage of an infinite strategy space to explore experimentally, enabled by its relatively simple analytic form.

7.1 First-Price Sealed-Bid Auction

In a first-price sealed-bid auction with n players, each player i has a private valuation (type) t_i of a particular good, for which it submits a single bid a_i in a concealed manner. The highest bidder gets the good, and obtains a payoff equal to its valuation minus its bid. Other bidders obtain zero payoff. In case of a tie, the winner is chosen randomly among the highest bidders.

Following Reeves [2005], we consider a restricted version of the game with players limited to strategies that bid a constant fraction of their valuations. That is, agent i's strategy is defined by a *shading factor* $k_i \in [0, 1]$, such that it bids $a_i = k_i t_i$. Taking this restriction, and the assumption types are drawn $U[0, 1]$, yields a normal form game we designate FPSBn.

We exploit analytical results [Reeves, 2005] to identify deviations, best responses, and equilibria, and to calculate regret with respect to the true game. Of course, in an arbitrary scenario we would not generally have access to such convenient analytic methods. With an analytical form for the best-response correspondence, the BR algorithm behaves much like the *double oracle algorithm* [McMahan et al., 2003, Zinkevich et al., 2007].

We compare expected regret for the candidate policies enumerated above, applied to FPSB2. All policies start with a random strategy $k \sim U[0, 1]$, then on subsequent equilibria choose based on their stated criteria. We estimate expected regrets by sampling 10^6 exploration sequences for each method described above. Regrets in any given sequence are computed as the theoretically best response to the latest equilibrium found, given the strategies X explored thus far. For $n = 2$, we are able to establish that all equilibria are in fact symmetric, however we omit the proof due to space constraints. Therefore, we can limit attention to symmetric pure-strategy profiles in our search for Nash equilibria at each iteration. In case of multiple equilibria, we average their respective regrets with respect to the base game. For the ST (softmax) method, we uniformly generate sets of 100 deviating strategies to pick from (at each step), and consider temperatures $\tau \in \{.1, 1, 10\}$.

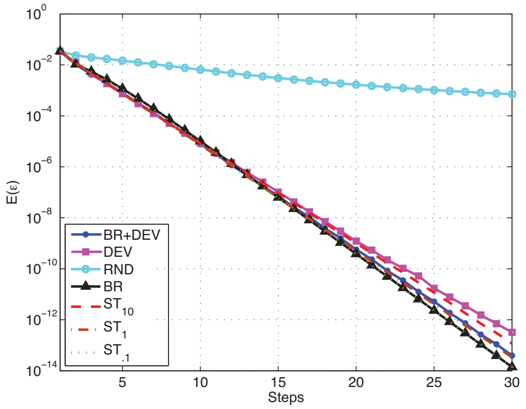

Figure 2: Expected regret in fpsb2 calculated by sampling 10^6 exploration sequences.

The results are shown in Figure 2. All methods that employ BR display comparable performance, and the worst method is clearly RND. Methods that employ random improving deviations (BR+DEV, DEV, ST_{10}, and ST_{1}) perform better in the early stages (steps 3-11), while those picking mostly best responses (BR and $\mathrm{ST}_{.1}$) catch up and perform slightly better thereafter. For steps 2–22, with the exception of a few comparisons,[4] all differences are statistically significant at the 0.05 level.

These results can be better understood by analyzing Figure 3, which shows regrets in FPSB2 after deviating from k_i to k_j. The surface spans only combinations such that k_j is an improving deviation from the profile where both players play k_i. The new equilibrium will have both playing k_j, thus the

[4] BR+DEV/DEV steps 8–9 ($p > .3$); BR+DEV/BR steps 2, 13–14 ($p > .1$); BR+DEV/ST_1 step 5 ($p = .13$); BR+DEV/$\mathrm{ST}_{.1}$ step 2 ($p = .2$); BR/ST_1 step 15 ($p = .06$); BR/$\mathrm{ST}_{.1}$ step 2 ($p = .2$); ST_{10}/ST_1 step 8 ($p = .06$); ST_{10}/$\mathrm{ST}_{.1}$ step 11 ($p = .12$); ST_1/$\mathrm{ST}_{.1}$ step 14 ($p = .13$)

height of the surface corresponds to the regret of that profile. The solid black line plots the best response as a function of k_i (projected onto the surface). The dotted (magenta) line represents the average deviation produced gain by the DEV policy. Above equilibrium, BR converges towards equilibrium in exactly one step, while DEV does so in expectation (solid black line overlaps dotted line exactly at $k_j = 0.5$). Below equilibrium, however, BR has a relatively slow convergence rate for $k_i < 0.4$, whereas DEV provides a much better expectation, which makes all methods using random improving deviations initially better. For $0.4 < k_i < 0.5$, BR and DEV (in expectation) become indistinguishable.

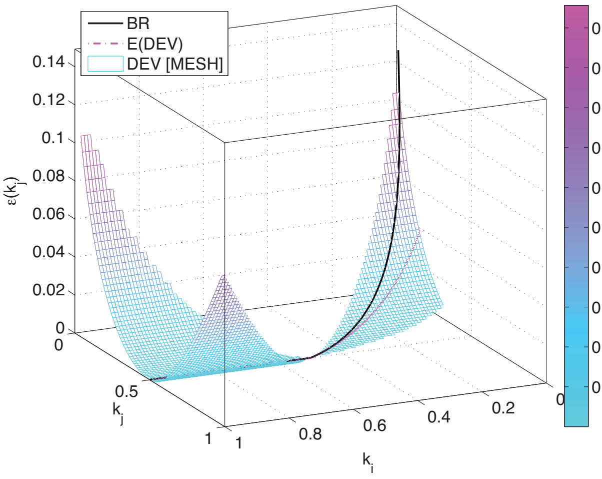

Figure 3: Regret in fpsb2 after deviating from k_i to k_j. Strategies that do not deviate are not shown.

7.2 Continuous Double Auction

The *continuous double auction* (CDA) [Friedman, 1993] is a simple and well-studied auction institution, employed commonly in commodity and financial markets. The "double" in its name refers to the fact that both buyers and sellers submit bids, and it is "continuous" in the sense that the market clears instantaneously on receipt of compatible bids. The CDA has also been widely employed in experimental economic studies, involving both human and software agents. Numerous papers have proposed novel bidding strategies for CDAs, accompanied by experimental comparisons to other known strategies. Some of the more prominent strategy families studied include "zero intelligence plus" [Cliff, 1998], "Gjerstad-Dickhaut" [Gjerstad and Dickhaut, 1998, Tesauro and Bredin, 2002], and "adaptive aggressiveness" [Vytelingum et al., 2008]. In most literature the comparison contexts (i.e., profiles of other-agent strategies in which featured strategies are evaluated) are selected by the experimenter. Exceptions include an early empirical game model [Walsh et al., 2002], and several studies that employ evolutionary search methods [Cai et al., 2007, Phelps et al., 2006].

In a recent EGTA study of a CDA game [Schvartzman and Wellman, 2009], the authors explored representative versions of all the prominent strategies from previous literature, and generated additional strategies using reinforcement learning. In total, the EGTA process iteratively considered 14 strategies: eight from the literature and six derived by reinforcement learning. The final empirical game model included evaluations for all four-player profiles over 13 of the strategies.[5] For purposes of the present study, we designate this model as the *base game*, and experimentally evaluate strategy exploration policies applied to these 13 strategies. This is of course a vast simplification of the actual infinite strategy space, but allows us to consider the implications of alternative orders that the strategies could be explored.

We evaluated expected regret as a function of number of steps by sampling 100 exploration sequences for each of DEV, RND, ST$_\tau$ for each starting strategy. The policies BR and MEMT require 13 sequences respectively, given that their exploration is deterministic after the random choice of starting strategy. For the restricted-game policies, we computed sample equilibria via replicator dynamics, evolving strategy populations until the corresponding symmetric mixed strategy has regret below 10^{-3}. In order to speed up computation, we seeded initial population proportions with the latest equilibrium mixture found in a given exploration sequence. We also cached equilibrium mixtures and MRCPs for repeated usage throughout the sampling process.

The results for the profile-regret analysis, presented in Figure 4, show that all methods employing improving deviations provide a similar expected regret, and clearly outperform RND. Different degrees of randomness in selecting beneficial deviations (τ) provided slightly different performance, and most variations of ST$_\tau$ resulted better than BR for steps 3–7. However, the augmented policy MEMT displayed the best performance. MEMT required 6 steps to reach the tolerance threshold in the worst case, but reached the threshold in 3.84 steps on average. This differs substantially with the improving-deviation policies results, where it took 8 steps to reach the 10^{-3} level. We also note, with the exception of MEMT, the figure displays the restricted-game variant of each policy, however the augmented version of each respective policy displayed quantitatively similar results.

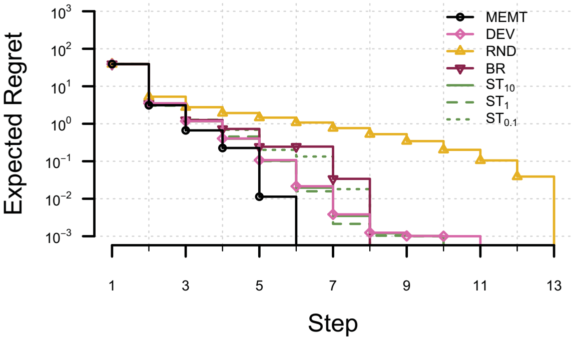

Figure 4: Expected minimum profile-regret in the empirical CDA game.

[5] The CDA scenario simulated was designed to be similar to prior studies, and employs 16 bidding agents. The authors couple the agents into groups of four to reduce the analysis to that of a four-player game [Wellman et al., 2005]. With 13 strategies there are a total of 1820 distinct profiles evaluated by simulation.

We evaluated two policies on the formation-regret measure: BR and MEMT. The BR policy is representative of the improving-deviation policies, whereas MEMT is a policy specifically design to minimize expected formation-regret. Like the profile-regret analysis, we found that MEMT and BR required 6 and 8 steps, respectively, to reach the tolerance threshold in the worst case, but reached the threshold in 3.84 and 4.30 steps on average. One explanation for this quick convergence is the existence of a 2-strategy primitive formation. This is the smallest formation that exists in the game and, therefore, the lower bound on the optimal number of steps to termination. If either support of the smallest primitive formation is the starting strategy, both policies terminate in two steps. Of the 13 starting strategies, 10 terminate in four steps or less for MEMT and 9 for BR, respectively. In cases where the best-response strategy is supporting the minimal profile-regret and the minimal formation-regret, both policies behave similarly. This occurred frequently in the CDA game, potentially due to its small primitive-formation.

7.3 TAC Travel Game

The original TAC market game, introduced in 2000, presented a challenge in the domain of travel shopping. In TAC/Travel, agents bid in three different kinds of auction mechanism (28 simultaneous auctions in all) to acquire flights, hotel rooms, and entertainment tickets to make trips for their clients. Years of competition and continued study led to numerous advances in trading agent strategy Wellman et al. [2007]. The University of Michigan team Walverine has been conducting an ongoing EGTA study of this game since 2004, with over 190,000 game instances in its data set at this writing. This exercise supported the selection of the Walverine version entered in 2004–06 tournaments, and has contributed in many ways to the development of EGTA methodology.

For the current experiment in exploration policy, we consider the two-player version of this empirical game (i.e., profiles with multiples of four agents playing any strategy). We further restrict consideration to 35 strategies for which we have evaluations of all combinations (630 profiles). We followed the same basic experimental procedure as for the CDA game described in the previous section. Results are presented in Figure 5. Unlike the results of the CDA experiment, MEMT and BR display similar quality. BR and MEMT require 7 and 8, respectively, steps to reach the tolerance threshold in the worst case, but reach the threshold in 4.74 and 4.86 steps on average, respectively.

Unlike the CDA game, the TAC/Travel game does not have a small primitive formation. In fact, the smallest primitive formation has 27 strategies. Figure 6 shows the expected minimum formation-regret averaged over the 35 traces up to a maximum of 16 steps. The dotted line shows the formation-regret of the optimal policy—the policy that yields the lowest regret at each step. The MEMT algorithm nearly converges in mean to the optimal policy after 15 steps, while BR is slower to converge with an expected regret of approximately 1.8 times that of MEMT after the same number of steps.

8. DISCUSSION

Our investigation of alternative strategy exploration policies provides evidence for several basic observations. First,

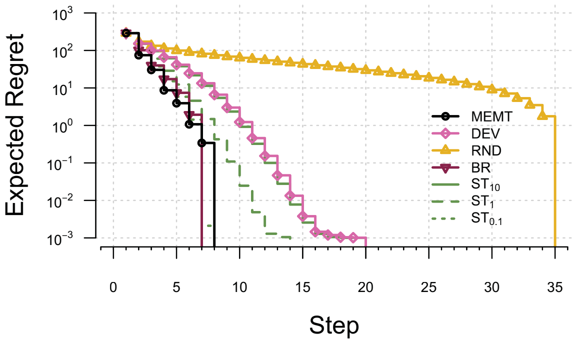

Figure 5: Expected minimum-profile regret in the empirical TAC/Travel game, logarithmic scale.

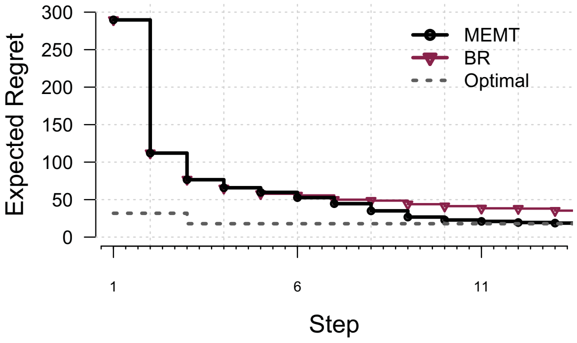

Figure 6: The expected minimum formation-regret in the empirical TAC/Travel game.

not surprisingly, any reasonable policy produces better candidate solutions as more strategies are explored. Second, considering only policies that provide a gain from deviation over the current equilibrium (or approximate formation) produces significant benefits over unrestricted selection. This leaves open the possibility that non-improving deviators with particular characteristics (e.g., complementarity with other known strategies) may be worthwhile. Third, formation-finding policies like MEMT perform well when small primitive formations exist and are comparable to the best improving-deviation policies when they do not.

Fourth, although best response is generally quite effective, there appears to be some advantage to exploring non-best response strategies, especially early in the process. As suggested by our analysis of the FPSB2 situation, exclusively introducing BR strategies may cause us to get stuck in a relatively unproductive region of profile space. Policies like MEMT can help overcome this problem.

Our experimental approach is limited by the need to know the true game in order to evaluate intermediate exploration results. We were nevertheless able to consider one game (FPSB2) with an infinite strategy space. The other two experimental sources were empirical games developed for distinct purposes, with discrete strategy sets developed manually or by employing reinforcement learning. Results were qualitatively similar, except that the expected regret curves were much smoother and more regular for the infinite game, FPSB2.

Unlike their restricted-game counterparts, augmented restricted-games allow for exact calculation of profile and formation regret with respect to a base game. Using restricted-game policies, we identified pathological cases where the regret of the selected profile increases as additional strategies are explored. With augmented restricted-game policies, this pathology cannot occur.

In general, computation of augmented restricted-games or an exact best response will not be possible for games of interest. We can view approaches that generate strategies via genetic algorithms [Phelps et al., 2006], reinforcement learning [Schvartzman and Wellman, 2009], or other heuristic optimization procedure as attempting to compute BR, perhaps succeeding only approximately. To the extent ST is a form of imperfect BR, this may be a rough model for what these approaches are accomplishing — though of course their degree of variance from BR is not as controlled. More direct evaluation of exploration policies based on heuristic optimization is difficult to perform in a domain-independent way, nevertheless such investigations may be a worthwhile direction for future work.

Finally, while the strategy exploration problem is a subclass of the profile selection problem, it has important practical uses for the profile *search* problem [Jordan et al., 2008], another distinct subclass of the profile selection problem. Foremost, in the profile search problem we (typically) attempt to identify minimum-regret pure-strategy profiles, however minimum-regret mixed-strategy profiles with small support are also valuable in EGTA. Unfortunately, mixed-strategy profiles dramatically change the search algorithms and, in some cases, require reasonable priors over all profile payoffs for the algorithms to be efficacious. On the other hand, reasoning about minimum-regret mixed-strategy profiles is a central part of the strategy exploration problem in which we do not require priors over the payoffs. If we are searching for approximate equilibria with small support (analogous to small formations), the strategy exploration problem is a natural extension of the profile search problem for set-valued solution concepts.

References

K. Cai, Jinzhong Niu, and Simon Parsons. Using evolutionary game-theory to analyse the performance of trading strategies in a continuous double auction market. In *Adaptive Agents and Multi-Agents Systems*, volume 4865 of *Lecture Notes in Computer Science*, pages 44–59. Springer, 2007.

D. Cliff. Evolving parameter sets for adaptive trading agents in continuous double-auction markets. In *Agents-98 Workshop on Artificial Societies and Computational Markets*, pages 38–47, Minneapolis, 1998.

D. Friedman. The double auction market institution: A survey. In D. Friedman and J. Rust, editors, *The Double Auction Market: Institutions, Theories, and Evidence*, pages 3–25. Addison-Wesley, 1993.

S. Gjerstad and J. Dickhaut. Price formation in double auctions. *Games and Economic Behavior*, 22:1–29, 1998.

John C. Harsanyi and Reinhard Selten. *A General Theory of Equilibrium Selection in Games*. MIT Press, June 1988.

Patrick R. Jordan. *Practical Strategic Reasoning with Applications in Market Games*. PhD thesis, University of Michigan, December 2009.

Patrick R. Jordan, Yevgeniy Vorobeychik, and Michael P. Wellman. Searching for approximate equilibria in empirical games. In *Seventh International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 1063–1070, Estoril, 2008.

Jeffrey O. Kephart and Amy R. Greenwald. Shopbot economics. *Autonomous Agents and Multiagent Systems*, 5: 255–287, 2002.

H. Brendan McMahan, Geoff Gordon, and Avrim Blum. Planning in the presence of cost functions controlled by an adversary. In *Twentieth Conference on Machine Learning*, 2003.

John Ashworth Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

Steve Phelps, M. Marcinkiewicz, Simon Parsons, and Peter McBurney. A novel method for automatic strategy acquisition in n-player non-zero-sum games. In *Fifth International Joint Conference on Autonomous Agents and Multi-Agent Systems*, 2006.

Ryan Porter, Eugene Nudelman, and Yoav Shoham. Simple search methods for finding a Nash equilibrium. *Games and Economic Behavior*, 63(2):642 – 662, July 2008.

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

L. Julian Schvartzman and Michael P. Wellman. Stronger CDA strategies through empirical game-theoretic analysis and reinforcement learning. In *Eighth International Joint Conference on Autonomous Agents and Multi-Agent Systems*, Budapest, May 2009.

G. Tesauro and J. L. Bredin. Strategic sequential bidding in auctions using dynamic programming. In *First International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 591–598, Bologna, 2002.

P. Vytelingum, D. Cliff, and N. R. Jennings. Strategic bidding in continuous double auctions. *Artificial Intelligence*, 172:1700–1729, 2008.

William E. Walsh, Rajarshi Das, Gerald Tesauro, and Jeffrey O. Kephart. Analyzing complex strategic interactions in multi-agent systems. In *AAAI-02 Workshop on Game-Theoretic and Decision-Theoretic Agents*, Edmonton, 2002.

Michael P. Wellman. Methods for empirical game-theoretic analysis (extended abstract). In *Twenty-First National Conference on Artificial Intelligence*, pages 1552–1555, Boston, 2006.

Michael P. Wellman, Daniel M. Reeves, Kevin M. Lochner, Shi-Fen Chen, and Rahul Suri. Approximate strategic reasoning through hierarchical reduction of large symmetric games. In *Twentieth National Conference on Artificial Intelligence*, pages 502–508, Pittsburgh, 2005.

Michael P. Wellman, Amy Greenwald, and Peter Stone. *Autonomous Bidding Agents: Strategies and Lessons from the Trading Agent Competition*. MIT Press, 2007.

Michael P. Wellman, Anna Osepayshvili, Jeffrey K. MacKie-Mason, and Daniel M. Reeves. Bidding strategies for simultaneous ascending auctions. *Journal of Theoretical Economics (Topics)*, 8(1), 2008.

Martin Zinkevich, Michael Bowling, and Neil Burch. A new algorithm for generating equilibria in massive zero-sum games. In *Twenty-Second Conference on Artificial Intelligence*, 2007.