# Exploring Large Strategy Spaces in Empirical Game Modeling

L. Julian Schvartzman and Michael P. Wellman

University of Michigan, Computer Science & Engineering, Ann Arbor, MI 48109-2121, USA

**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 equilibrium among previously explored strategies perform generally well, although we find that some stochastic introduction of suboptimal responses can often lead to more effective exploration in early stages of the process.

## 1 Introduction

Often the most difficult obstacle to game-theoretic analysis of complex scenarios is developing a model of the game situation in the first place. 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.

The iterative 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.

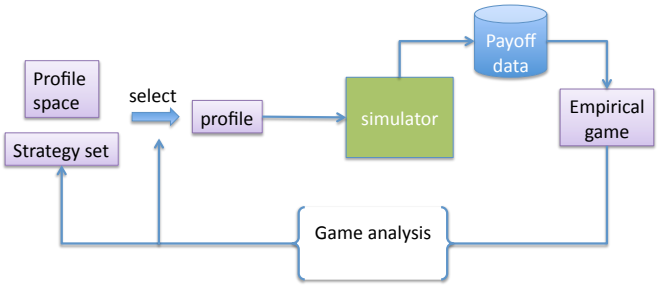


**Fig. 1.** Dynamic game formulation through empirical game-theoretic analysis.

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 method's *strategy exploration policy*.

## 2 Strategy Exploration Problem

### 2.1 Background

The issue of strategy exploration is one facet of the broader problem of how to allocate simulation resources across the profile space. Under the *noisy-payoff model* (Jordan et al., 2008), the basic step is simulating a profile (i.e., the "select" arrow in Figure 1), producing a noisy sample of the associated payoff vector. In the first study to directly address the problem, Walsh et al. (2003) proposed a method to choose the next profile to simulate based on a heuristic measure called expected confirmational value of information (ECVI). Jordan et al. (2008) presented a framework for evaluating profile-selection rules, and introduced an information-gain heuristic that outperformed ECVI in experiments. All of the above methods assume an initial set of samples for all profiles, in order for their heuristic or sensitivity measures to be well-defined.

Sureka and Wurman (2005) investigated a different formulation of the problem, where the basic step obtains the payoff of a profile (considered to be accurate rather than a noisy estimate), without reference to any prior samples of that profile. Under this *revealed-payoff model*, their TABU best-response search method is able to identify equilibria after exploring a fraction of the overall profile space. Jordan et al. (2008) proposed an alternative, called minimimum-regret-first search (MRFS), which required comparable search to identify equilibria but confirms approximate equilibria earlier in the process. Combining MRFS to generate initial estimates with information-gain search to refine conclusions given noisy payoffs addresses the problem end-to-end, and is most effective based on available evidence. However, we expect many improvements are possible, both generally and for particular classes of games.

In contrast to these models, the strategy exploration problem focuses on the basic step of adding a strategy to the candidate set (i.e., the leftmost up-arrow in Figure 1),

thus enlarging the profile space. 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.

In an empirical analysis of 4-player chess, Kiekintveld et al. (2006) started with a baseline set of strategies, and augmented the set with strategies derived by reinforcement learning (RL) against a set of the currently best candidates. The idea is that the RL process itself searches in a large space of strategies, but in a context where the other-agent strategies are kept fixed. This avoids the combinatorial consideration of profiles when improvements can be found more directly for the current context. Although in general strategy quality is context-dependent, one would typically expect many strategy variations to be relatively robust across (relevant) contexts.

We pursued this approach systematically in a recent analysis of strategies for bidding in CDAs (continuous double auctions) (Schvartzman and Wellman, 2009). Starting with representatives of the major CDA strategy proposals from the literature, we exhaustively evaluated profiles over this set, and iteratively added strategies derived by RL in the context of equilibria from the current empirical game. This exercise confirmed prior literature conclusions about the relative quality of known strategies, and successfully learned a sequence of new strategies that outperformed all of these. On convergence of the interleaved EGTA/RL process, the equilibria of the final empirical game were supported exclusively with learned strategies.

The CDA study demonstrated the effectiveness of RL for deriving stronger CDA trading strategies. The overall analysis was quite computationally intensive, which ultimately limited the number of strategies that could be explored. Thus, we are motivated to understand the effect of alternate policies for introducing strategies.

### 2.2 Problem Definition

We assume the underlying "true" game can be represented in normal form, with $n$ the number of players. Let $S_i$ denote the strategy set for player $i$, $\prod_{i=1}^{n} S_i$ the joint strategy set or *profile space*, and $u : \prod_{i=1}^{n} S_i \rightarrow \mathbb{R}^n$ the *payoff function*, describing the vector of payoffs associated with a given profile. The payoff to player $i$ for playing strategy $s_i \in S_i$ when others play joint strategy $s_{-i} \in \prod_{j \neq i} S_j$ is given by $u_i(s_i, s_{-i})$.

For a *symmetric game*, $S_i = S_j = S$ for all $i$ and $j$, and the payoff function is invariant to permutations of the players. For simplicity, we assume symmetric games in this paper, however extending the methods and analysis to non-symmetric games is straightforward. Assuming symmetric games enables us to focus on managing a single strategy set, and also allows us to simplify notation and description of search techniques.

Our problem formulation presumes an EGTA process as diagrammed in Figure 1. The empirical game model is also represented in normal form, with $\hat{S} \subseteq S$ the *current strategy set* (the same for each player, under symmetry). We abstract away from the profile sampling control problem by assuming the empirical game model is simply a projection of the true game onto the profile space induced by $\hat{S}$. The *strategy exploration problem*, then, boils down to choosing a new strategy, $\hat{s} \in S \setminus \hat{S}$, to add to the current strategy set, yielding the update $\hat{S} \leftarrow \hat{S} \cup \{\hat{s}\}$. A solution to this problem takes the form of a policy for choosing $\hat{s}$ based on analysis of the current empirical game. Implementing this policy within an EGTA process results in a sequence of strategies to be added, until all are explored, or (more realistically) we run out of time.

How should we evaluate a candidate strategy exploration policy? Presumably, we are interested in solutions to the true game, and some strategies are more important to such 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 concept of *regret*.

**Definition 1 (Regret)** *The* regret, $\varepsilon(s)$, *of a profile* $s = (s_1, \ldots, s_n)$, *is the maximum gain available to any player by deviating to another strategy:*

$$\varepsilon(s) = \max_i \max_{s'_i \in S_i} u_i(s'_i, s_{-i}) - u_i(s_i, s_{-i}).$$

To evaluate the quality of an empirical game model, we solve the model employing our solution concept of choice (e.g., identifying a sample Nash equilibrium), and measure the regret of this solution profile with respect to the *true* game.[1] Intuitively, this captures the quality of the profile we would propose if we had to stop at the current iteration. A profile with regret $\varepsilon$ constitutes an approximate, $\varepsilon$-Nash equilibrium, with $\varepsilon = 0$ corresponding to exact equilibrium. All else equal, we consider profiles with smaller $\varepsilon(s)$ to be more stable, and thus more plausible as plays of the actual game.

Evaluating a strategy exploration policy in these terms yields a sequence of regret values, one at each iteration. We seek policies providing lower regret for any given number of iterations. We conclude this section with some useful definitions.

**Definition 2 (Deviation)** *Strategy* $s'_i$ *is a* deviation *for agent* $i$ *with respect to profile* $s$ *if* $i$ *would benefit by playing* $s'_i$ *rather than its designated strategy in* $s$*:*

$$u_i(s'_i, s_{-i}) > u_i(s_i, s_{-i}).$$

**Definition 3 (Gain)** *The* gain *for agent* $i$ *from deviation* $s'_i$ *is the increase in payoff it obtains by switching from its designated strategy in* $s$*:*

$$u(s_i \rightarrow s'_i, s_{-i}) \equiv u(s'_i, s_{-i}) - u_i(s_i, s_{-i}).$$

[1] 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.

**Definition 4 (Best Response)** *Strategy $s_i'$ is a* best response *for agent $i$ with respect to profile $s$ if $i$ would maximize its payoff by playing $s_i'$. For all $s_i'' \in S_i$,*

$$u_i(s_i', s_{-i}) \geq u_i(s_i'', s_{-i}).$$

### 2.3 Example

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 $\varepsilon((1,1)) = 3$.

| | 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 $\hat{S} = S$, equilibria in the empirical game and true 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.
- Deviations only (DEV). Find a Nash equilibrium of the current empirical game, and pick one of the remaining deviating strategies with equal probability.
- Best response (BR). Find a Nash equilibrium of the current empirical game, and pick a best response among the remaining strategies.

Note that DEV and BR build on analysis of the current empirical game, and require access to some payoffs in the true game 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.[2] On the first iteration, there is no current empirical game, so DEV and

[2] This is true assuming that the support of the current equilibrium is much smaller than $\hat{S}$. In practice, simulations incurred during strategy selection should be cached for later use.

BR necessarily choose randomly. These policies would also choose randomly if there are no deviations among the unexplored strategies, in which case we already have a true equilibrium solution 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) 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.

## 3 Experimental Setup

Our experimental approach follows the process we illustrated by the simple example of Section 2.3. We start with a known game, and compare the results of applying various strategy exploration policies. In addition to the three policies (RND, DEV, BR) introduced above, we consider the following exploration policies:

- Alternating (BR+DEV). Apply BR and DEV, in turn, on successive iterations.
- Softmax (ST). Find a Nash equilibrium $s$ of the current empirical game, and let $D$ be the set of deviations among the remaining strategies. Pick strategy $s'_i$ from $D$ with probability given by the softmax formula applied to deviation gains:

$$\frac{e^{u(s_i \to s'_i, s_{-i})/\tau}}{\sum_{s''_i \in D} e^{u(s_i \to s''_i, s_{-i})/\tau}}, \tag{1}$$

  where $\tau$ is the typical *temperature* parameter. Low values of $\tau$ mimic a best response (i.e., ST approximates BR), whereas $\tau \to \infty$ turns the selection equiprobable (i.e., ST approximates DEV).[3]

We evaluate these 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. This game was previously and extensively studied as a

[3] So that the temperature settings are meaningful across games, we employ normalized payoffs in computing gains in (1).

test for EGTA methods by Reeves (2005), and we build on his results to conduct our investigation of strategy exploration.

Our second test is the four-player empirical game generated in our 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 our study. As noted above, this exercise motivated our present investigation. The RL operation can be viewed as an approximation of BR, and so our experiment allows us to consider alternative orderings.

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.

## 4 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 FPSB$n$.

The following analytical results are exploited in our experimental study to identify deviations, best responses, and equilibria, and to calculate regret with respect to the true game. (Of course, in an actual EGTA process we would not generally have access to such analytic assistance.)

**Theorem 1 (Reeves (2005), Appendix A.5).** *The* expected payoff *for a player choosing $k_i$ against everyone else playing $k$ in* FPSB$n$ *is:*

$$u(k_i,k) = \begin{cases} \dfrac{1}{2n} & \text{if } k_i = k = 0, \\ \dfrac{1-k_i}{n+1}\left(\dfrac{k_i}{k}\right)^{n-1} & \text{if } k_i \le k, \\ \dfrac{(1-k_i)((n+1)k_i^2-(n-1)k^2)}{2(n+1)k_i^2} & \text{otherwise.} \end{cases}$$

**Theorem 2 (Reeves (2005), Appendix A.6).** *The* best response *to everyone else playing $k$ in* FPSB$n$ *is:*

$$BR(k) = \begin{cases} \textit{undefined} & \textit{if } k = 0, \\ \xi & \textit{if } k < \dfrac{n-1}{n}, \\ \dfrac{n-1}{n} & \textit{if } k \geq \dfrac{n-1}{n}, \end{cases}$$

*where* $\xi \equiv$

$$\frac{\sqrt[3]{3}\left(k^2\left(n^2-1\right)\left(9n+\sqrt{3(n+1)\left((n-1)k^2+27(n+1)\right)}+9\right)\right)^{2/3}-3^{2/3}k^2\left(n^2-1\right)}{3(n+1)\sqrt[3]{k^2(n-1)\left(9n^2+18n+(n+1)^{3/2}\sqrt{3(n-1)k^2+81(n+1)}+9\right)}}.$$

**Theorem 3 (Reeves (2005), page 26).** *Pure strategy* $\dfrac{n-1}{n}$ *is the unique symmetric Nash equilibrium of* FPSB$n$.

**Corollary 1 (Gain from Deviation).** *Given two strategies $k_i < k_j$ in* FPSB2*, the gain obtained by deviating from one strategy to the other, while the second player maintains its choice, is:*

$$u(k_i \to k_j, k_j) = \frac{(k_i - k_j)(k_i + k_j - 1)}{3k_j}$$

$$u(k_j \to k_i, k_j) = \frac{(k_j - k_i)(k_i + k_j - 1)}{3k_j}$$

$$u(k_i \to k_j, k_i) = \frac{\left(k_i k_j - k_j - k_i + 3k_j^2\right)(k_i - k_j)}{6k_j^2}$$

$$u(k_j \to k_i, k_i) = \frac{\left(k_i k_j - k_j - k_i + 3k_j^2\right)(k_j - k_i)}{6k_j^2}$$

*Proof.* Follows almost directly from Theorem 1.

When strategies are restricted to a finite set (e.g., the current set in an EGTA process), Theorem 3 does not apply. For $n = 2$, we are able to establish that all equilibria are in fact symmetric.

**Theorem 4 (Equilibria in Restricted Game).** *In* FPSB2 *with players picking $k_i$ and $k_j$ from a finite set of strategies $0 < k \leq 1$, equilibrium is always symmetric.*

*Proof.* Omitted due to space constraints.

We now 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 $\hat{S}$ explored thus far. Per Theorem 4, 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 true game. For the ST (softmax) method, we uniformly generate sets of 100 deviating strategies to pick from (at each step), and consider temperatures $\tau \in \{.1, 1, 10\}$.

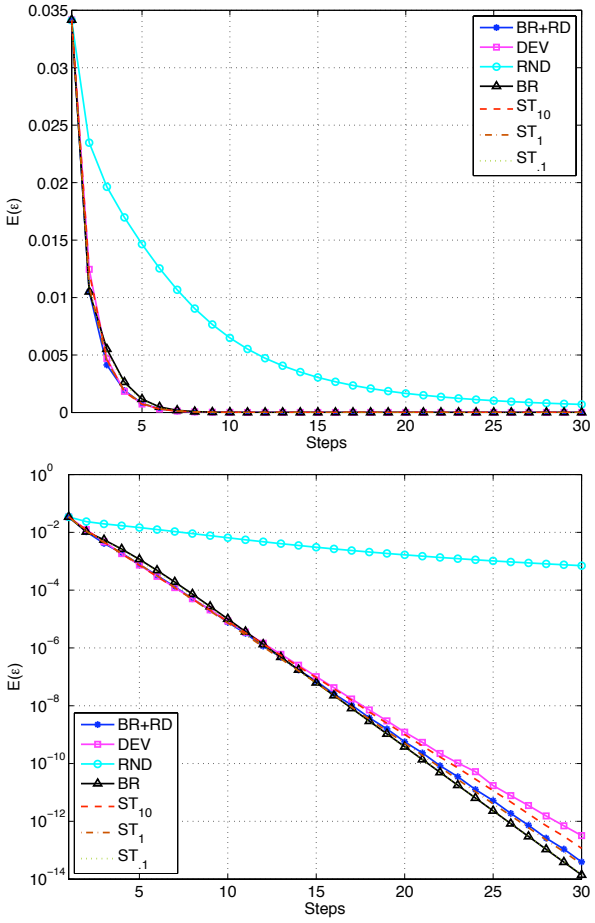


**Fig. 2.** Expected regret in FPSB2 calculated by sampling $10^6$ exploration sequences, plotted on linear (top) and logarithmic (bottom) scales.

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 devi-

ations (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 $ST_{.1}$) catch up and perform slightly better thereafter. For steps 2–22, all differences are statistically significant at the .05 level, with the exception of: 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+RD/$ST_{.1}$ step 2 ($p = .2$); BR/$ST_1$ step 15 ($p = .06$); BR/$ST_{.1}$ step 2 ($p = .2$); $ST_{10}$/$ST_1$ step 8 ($p = .06$); $ST_{10}$/$ST_{.1}$ step 11 ($p = .12$); $ST_1$/$ST_{.1}$ step 14 ($p = .13$).

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 a deviation from the profile where both players play $k_i$. From Theorem 4, the new equilibrium will have both playing $k_j$, thus the 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), as given by Theorem 2. The dotted (magenta, for color viewers) line represents the average deviation produced 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 deviations initially better. For $0.4 < k_i < 0.5$, BR and DEV (in expectation) become indistinguishable.

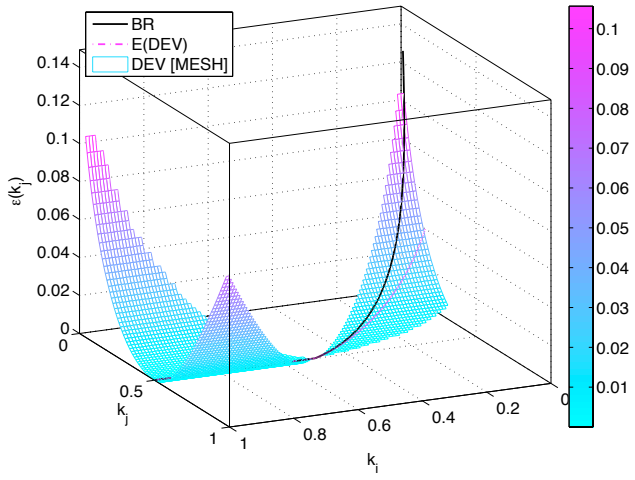


**Fig. 3.** Regret in FPSB2 after deviating from $k_i$ to $k_j$. Strategies that do not deviate are not shown.

## 5 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), we explored representative versions of all the prominent strategies from previous literature, and generated additional strategies using reinforcement learning. In total, our EGTA process iteratively considered 14 strategies: eight from the literature and six derived by RL. (We also learned seven additional strategies that were not explored because they failed to deviate.) Our final empirical game model included evaluations for all four-player profiles over 13 of the strategies.[4] For purposes of the present study, we designate this model as the "true 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 $10^6$ exploration sequences for each of DEV, RND, and $\mathrm{ST}_\tau$. BR required 13 sequences only, given that its exploration is deterministic after the random choice of starting strategy. We computed sample equilibria via replicator dynamics, evolving strategy populations until the corresponding symmetric mixed strategy has regret below 0.001. 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 for repeated usage throughout the sampling process.

The results, presented in Figure 4, show that all methods employing deviations provide a similar expected regret, and clearly outperform RND. Different degrees of randomness in selecting deviations ($\tau$) provided slightly different performance, and most variations of $\mathrm{ST}_\tau$ resulted better than BR for steps 3–7.

## 6 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

[4] The CDA scenario simulated was designed to be similar to prior studies, and employs 16 bidding agents. We 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.

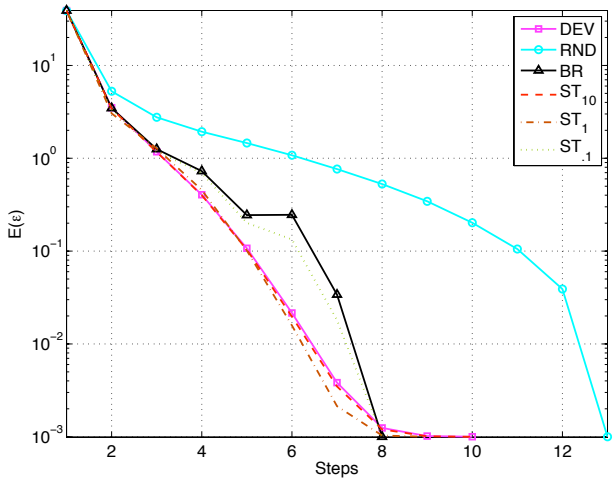


**Fig. 4.** Expected regret in the empirical CDA game, logarithmic scale.

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.

## 7 Discussion

Our investigation of alternative strategy exploration policies provides evidence for several basic observations. First, not surprisingly, any reasonable strategy produces better candidate solutions as more strategies are explored. Second, considering only strategies that deviate from the current equilibrium produces significant benefits over unrestricted selection. This leaves open the possibility that non-deviators with particular characteristics (e.g., complementarity with other known strategies) may be worthwhile, but we have not evaluated any exploration policies along these lines.

Third, although best response is generally quite effective, there appears to be some advantage to exploring non-best deviations, 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. In other words, we observe a form of explore-exploit tradeoff in selecting new strategies based

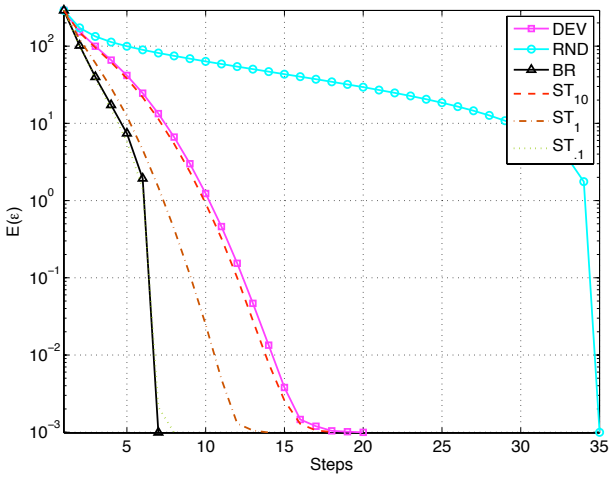


**Fig. 5.** Expected regret in the empirical TAC Travel game, logarithmic scale.

on our current model. Based on this experience, among the policies evaluated we recommend the softmax policy ST, perhaps annealing the temperature so that the policy approaches BR over time.

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. Our 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.

In general, computation of an exact best response, or exact implementation of softmax, 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.

## References

Cai, K., Niu, J., and Parsons, S. (2007). 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.

Cliff, D. (1998). 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, MN.

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

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

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

Kephart, J. O. and Greenwald, A. R. (2002). Shopbot economics. *Autonomous Agents and Multiagent Systems*, 5:255–287.

Kiekintveld, C., Wellman, M. P., and Singh, S. (2006). Empirical game-theoretic analysis of chat-uranga. In *AAMAS-06 Workshop on Game-Theoretic and Decision-Theoretic Agents*, Hakodate.

Phelps, S., Marcinkiewicz, M., Parsons, S., and McBurney, P. (2006). 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*, pages 705–712, Hakodate.

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

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

Sureka, A. and Wurman, P. R. (2005). Using tabu best-response search to find pure strategy Nash equilibria in normal form games. In *Fourth International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 1023–1029, Utrecht.

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

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

Walsh, W. E., Das, R., Tesauro, G., and Kephart, J. O. (2002). Analyzing complex strategic interactions in multi-agent systems. In *AAAI-02 Workshop on Game-Theoretic and Decision-Theoretic Agents*, Edmonton.

Walsh, W. E., Parkes, D., and Das, R. (2003). Choosing samples to compute heuristic-strategy Nash equilibrium. In *AAMAS-03 Workshop on Agent-Mediated Electronic Commerce*, Melbourne.

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

Wellman, M. P., Greenwald, A., and Stone, P. (2007). *Autonomous Bidding Agents: Strategies and Lessons from the Trading Agent Competition*. MIT Press.

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

Wellman, M. P., Reeves, D. M., Lochner, K. M., Cheng, S.-F., and Suri, R. (2005). Approximate strategic reasoning through hierarchical reduction of large symmetric games. In *Twentieth National Conference on Artificial Intelligence*, pages 502–508, Pittsburgh.