The Workshops of the Thirty-Second
AAAI Conference on Artificial Intelligence

Evaluating the Stability of Non-Adaptive Trading in Continuous Double Auctions: A Reinforcement Learning Approach

Mason Wright, Michael P. Wellman
University of Michigan
Ann Arbor, MI 48109
{masondw, wellman}@umich.edu

Abstract

The continuous double auction (CDA) is the predominant mechanism in modern securities markets. Despite much prior study of CDA strategies, fundamental questions about the CDA remain open, such as: (1) to what extent can outcomes in a CDA be accurately modeled by optimizing agent actions over only a simple, non-adaptive policy class; and (2) when and how can a policy that conditions its actions on market state deviate beneficially from an optimally parameterized, but simpler, policy like Zero Intelligence (ZI). To investigate these questions, we present an experimental comparison of the strategic stability of policies found by reinforcement learning (RL) over a massive space, or through empirical Nash-equilibrium solving over a smaller space of non-adaptive, ZI policies. Our findings indicate that in a plausible market environment, an adaptive trading policy can deviate beneficially from an equilibrium of ZI traders, by conditioning on signals of the likelihood a trade will execute or the favorability of the current bid and ask. Nevertheless, the surplus earned by well-calibrated ZI policies is empirically observed to be nearly as great as what a deviating reinforcement learner could earn, using a much larger policy space. This finding supports the idea that it is reasonable to use equilibrated ZI traders in studies of CDA market outcomes.

Introduction

The continuous double auction (CDA) is the preeminent security trading mechanism, accounting for trillions of dollars in transactions annually (NYSE 2017). In a CDA, buyers and sellers submit orders to the market, and any order that crosses the best-priced prior order of opposite type *clears*, producing a trade. Bidding in a CDA is a dynamic game of imperfect information, as trading agents do not know each others' valuations and generally do not observe all bids.

Despite the mechanism's prevalence, attempts at game-theoretic characterizations have generally been limited to highly stylized scenarios (Wilson 1987), or numeric solution of abstract models (Goettler, Parlour, and Rajan 2009). Many other research efforts aim to establish stylized facts about CDA market outcomes, based on simulation or analysis of rule-based traders in action (Wah and Wellman 2013; Chakraborty, Das, and Peabody 2015; Jovanovic and Menkveld 2016; Budish, Cramton, and Shim 2015). The literature also includes a progression of works, each presenting a novel policy for CDA trading agents and experimental evidence comparing it beneficially to less sophisticated policies from earlier papers (Cliff and Bruten 1997; Gjerstad and Dickhaut 1998; Tesauro and Das 2001; Tesauro and Bredin 2002; Schvartzman and Wellman 2009; Vytelingum, Cliff, and Jennings 2008).

Prior studies of heuristic strategies contribute to our understanding of the CDA mechanism, but results based on heuristic strategy profiles may be subject to doubt due to the possible strategic instability of these profiles. We can equilibrate over a class of heuristic strategies, but the question remains: how much gain is available by going beyond this class? In particular, one way to refine a strategy is to condition its actions on additional features, adapting its behavior by taking account of more state information. We seek to evaluate whether agents can benefit significantly by adopting more complex, adaptive policies, particularly as extending to such larger strategy spaces may be difficult or costly.

We present a systematic experimental study of the CDA, in which we derive trading policies via RL and empirical game-theoretic analysis (EGTA). We use as our baseline trading heuristic the Zero Intelligence (ZI) strategy, which has minimal capacity to adapt to market state. The version of ZI we employ has a few parameters, which we tune via EGTA to find approximate Nash-equilibrium mixtures within the baseline set. Against these policies we train more adaptive trading policies using Q-learning (Watkins and Dayan 1992). We conduct a statistically rigorous analysis of the benefit of conditioning a policy on market state, relative to the non-adaptive baseline. Results suggest the equilibrated non-adaptive CDA policies leave positive, but surprisingly modest, room for gain through complex conditioning on market state.

Prior Work: Heuristic CDA Strategies

A Zero Intelligence (ZI) trader sets its order price as a random surplus offset from its valuation, based on a uniform distribution from a specified range. ZI was introduced by Gode and Sunder (1993), to demonstrate how a CDA market's allocative efficiency approaches its optimum, even if all traders use such a simple strategy. The ZI policy model in various forms has been popular among experimental and analytical researchers alike, for its simplicity and ability to capture stylized facts of real markets or fit real-world financial data (Farmer, Patelli, and Zovko 2005;

Mike and Farmer 2008; Li and Das 2016). Several recent works have employed ZI traders in models of financial markets or prediction markets (Li and Das 2016; Wah, Wright, and Wellman 2017; Chakraborty, Das, and Peabody 2015; Wah, Lahaie, and Pennock 2016).

It has always been clear that ZI is not an optimal trading strategy. Cliff and Bruten (1997) showed that ZI tends to yield efficient allocations only if agents' aggregate supply and demand curves have equal slopes, and the authors proposed one of many strategic improvements on ZI, known as ZI Plus (ZIP). ZIP and other ZI successors such as GD and GDX (Gjerstad and Dickhaut 1998; Tesauro and Das 2001; Tesauro and Bredin 2002) and AA (Vytelingum, Cliff, and Jennings 2008; De Luca and Cliff 2011) adjust the surplus demanded by the agent during a run. Studies have shown such policies to be beneficial deviations from a single, fixed ZI policy (Tesauro and Das 2001; Walsh et al. 2002). Even stronger policies have been derived by RL, to deviate beneficially from a mixed strategy of GDX agents (Schvartzman and Wellman 2009). These studies have the limitation that they compare a new policy against a single, uncalibrated parameterization of ZI, which may be a straw man form of the ZI agent. The more relevant comparison, we argue, is to equilibrated ZI mixtures rather than to arbitrary ZI instances.

The prior work most similar to ours is a study by Schvartzman and Wellman (2009), which used RL to derive the strongest-yet trading strategy in a particular CDA setting. In that work, authors used Q-learning to derive a policy that deviates beneficially from other agents using a fixed ZI strategy, and showed their learned policies also deviate successfully from the strategies ZIP and GDX. Our work builds on the methods of Schvartzman and Wellman (2009) to serve a different goal. We employ RL in an attempt to characterize when and how CDA traders can benefit from conditioning their actions on market state. We compare equilibrated ZI mixed strategies to (approximate) best responses learned via RL, to measure the strategic effectiveness of calibrated ZI relative to more complex policies. In addition, we analyze the relative importance of features for learning proposed in prior work, through regression over experimentally learned policies.

Research Contributions

We investigate how and to what extent a trading policy in the CDA can improve by conditioning on market state, relative to a calibrated non-adaptive policy. We provide insight into the tradeoffs of using a non-adaptive policy (ZI) to model trading behavior in a CDA, and the relative importance of features for an adaptive trading agent.

Our main contributions are as follows:

- We evaluate the strategic stability of calibrated ZI policies against (approximate) best responses from Q-learning. Results show some surplus is lost by playing a calibrated ZI policy instead of adaptive policies, although the amount is small compared to the surplus lost by using non-equilibrated ZI parameters.
- We analyze the nature of adaptive trading policies that outperform ZI.
- Our findings suggest equilibrating over many ZIs yields diminished loss of surplus relative to what could be achieved by alternative ZI policies, but with small positive loss of surplus remaining with respect to what an adaptive agent could earn. Moreover, that the common practice of using equilibrated ZIs as an approximation of efficient agent behavior is likely acceptable.

CDA Market Model

Our study is based on a CDA market model, similar to those of prior studies by Wah and Wellman (2013; 2017). The market has a single security and many trading agents. The value of the security to an agent is defined as the sum of the agent's private value for the good (drawn from some random distribution), and the fundamental value, which evolves by a stochastic process. Agents can trade the security with one another through the CDA mechanism, by submitting limit orders to the market. Each agent can submit an order only in those time steps when it *arrives* at the market, as determined by a random (exponential) inter-arrival time process; upon each arrival, an agent is independently randomly assigned to buy or sell. Each agent's payoff from the CDA game is defined as the final fundamental value of its inventory, plus the cumulative private value of its inventory, plus its final cash holdings.

Our market model is populated by 17 trading agents, comprising 16 background traders and one market maker (MM). The MM maintains a *ladder* of buy and sell orders separated from the expected final fundamental value by a fixed spread, updated each time it arrives. The background traders act according to parameterized forms of the Zero Intelligence (ZI) policy.

Zero Intelligence

ZI is a simple strategy for CDA trading that has been shown to converge to efficient prices and allocations in many settings (Gode and Sunder 1993). Our variant of ZI, introduced by Wah et al. (2017), has three parameters: $\underline{d}$, $\overline{d}$, and $\eta \in (0, 1]$. At each arrival, a ZI agent places a limit order that demands a surplus equal to a random draw from $\mathcal{U}(\underline{d}, \overline{d})$. The exception is if the agent would earn at least η fraction of its randomly drawn surplus goal at the current quote; in that case, the agent opportunistically places an executable order at the quote instead.

Market Model Description

All 17 agents arrive at the market with independent inter-arrival times, drawn from an exponential distribution with rate λ_{BG} for background traders, λ_{MM} for the market maker. We let $\lambda_{BG} = 0.012$ and $\lambda_{MM} = 0.05$, with a game duration of $T = 2000$ time steps. Hence, each background trader arrives roughly every 83 time steps in expectation, the market maker every 20 time steps.

The fundamental value evolves as a mean-reverting random walk with zero-mean Gaussian noise and long-run mean μ.[1] At each time step, the fundamental value is updated, $r_t \leftarrow \kappa\mu + (1 - \kappa) \times r_{t-1} + \mathcal{N}(0, \sigma_s^2)$, where σ_s^2 is

[1] We take $\mu = 10^5$. The specific level does not matter, if the

the shock variance to public value, and κ is the mean reversion parameter. Throughout this study, we use $\kappa = 0.01$ and $\sigma_s^2 = 20000$. Given the observed fundamental at time t, the expected terminal fundamental value is

$$\hat{r}_t = \left(1 - (1-\kappa)^{T-t}\right)\mu + (1-\kappa)^{T-t} r_t.$$

ZI and MM agents use this estimate in determining their order prices.

Each background trader is assigned a private value vector at initialization, which lists the value of each additional security unit, given a current inventory. This vector has length 20, because the agent is restricted to hold (or owe) no more than 10 units. The vector is derived by sampling 20 values from $\mathcal{N}(0, \sigma_p^2)$, where σ_p^2 is the private value variance; the samples are sorted in non-increasing order, so that each agent's demand decreases with inventory. This study uses $\sigma_p^2 = 2 \times 10^7$.

When the market maker arrives, it cancels its existing orders and places a new ladder of orders, with 100 rungs each above and below the expected final fundamental. The ladder has orders centered around $\hat{r}_t$, at sell prices of $\hat{r}_t + 256 + 100 \times i$ and buy prices of $\hat{r}_t - 256 - 100i$, for $i \in \{0, \ldots, 99\}$.

When a background trader playing a ZI strategy arrives at the market, it cancels its previous order and is assigned with equal probability to place a buy or sell order. Suppose the agent has ZI parameters $\underline{d}, \overline{d}, \eta$. The agent computes the surplus it would obtain by trading immediately at the quote, which for a buyer is $\hat{r}_t + v_{i+1} - A$, or for a seller is $B - (\hat{r}_t + v_i)$, where A is the ask, B is the bid, and v_i is the agent's private value of unit i of inventory. The agent compares this surplus to ηs, where s is a random draw from $\mathcal{U}(\underline{d}, \overline{d})$. If the agent can obtain enough surplus, it transacts immediately. Otherwise, it places an order demanding the surplus goal s: either a buy order at $\hat{r}_t + v_{i+1} - s$ or a sell order at $\hat{r}_t + v_i + s$.

Each agent earns a payoff equal to the final fundamental value of its stock inventory, plus the cumulative private value of its stock inventory, plus its final cash holdings.

Definitions

Throughout this paper, we make use of many standard terms from game theory, defined here for completeness. By a *policy* or *pure strategy*, we mean a mapping from an agent's set of observation states to the (possibly stochastic) action the agent will take in each state. A *mixed strategy* is a probability distribution over pure strategies. A *profile* is an assignment of a strategy (pure or mixed) to each agent. A *symmetric* profile assigns the same strategy to each agent. A *Nash equilibrium* (NE) is a profile such that no agent can achieve a higher expected payoff by unilaterally deviating from its assigned strategy to any alternative strategy. The *regret* of a profile is the maximum over agents, of the maximum gain in expected payoff the agent can obtain, by deviating to any alternative strategy. By definition, a Nash equilibrium has zero regret.

In this paper, we call a profile an *equilibrium over strategy set* $\mathcal{S}$, if all pure strategies played with positive probability are in $\mathcal{S}$, and no agent can achieve higher expected payoff by deviating to a strategy in $\mathcal{S}$. We say a profile has *regret* x *with respect to strategy set* $\mathcal{S}'$, if x is the maximum any agent gains in expectation by deviating to a strategy in $\mathcal{S}'$.

evolving fundamental has negligible probability of hitting the zero lower bound.

Reinforcement Learning Methods

To investigate how much adaptive policies can increase payoffs relative to a well-calibrated ZI policy, and to investigate the nature of beneficially deviating policies, we need a way to search for beneficial deviations in a large search space. We tested several RL approaches, before settling on a variant of *Q-learning* that empirically worked well in our setting. We first fix the policies of all but one agent, which converts the trading game into a decision problem for the one strategic agent, then apply Q-learning.

Q-learning is a classical RL algorithm that provably converges to an optimal policy in finite Markov decision problems (MDPs) with bounded rewards, assuming a suitable learning rate sequence is used (Watkins and Dayan 1992). It works as follows. The learning agent progresses through a sequence of observing states s, getting rewards r, and taking actions a. The agent maintains an estimate of the Q-value of each state-action pair, $Q(s, a)$, which represents the expected value of taking action a in state s and playing optimally thereafter. On experiencing the sequence (s, a, r, s'), the agent performs a Q-learning value update,

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s', a')\right),$$

where α is the learning rate, and γ is the discount factor for future values. We set $\gamma = 0.9$ for learning, as a regularizer, but decay γ toward 1 as learning progresses. (The underlying game has no discounting.) We set $\alpha(s, a)$ to the reciprocal of the number of (s, a) observations to this point.

Learning Feature Set

Our learning agents use the following features in their state observations.

- P, the profit that would be obtained by trading immediately at the current price quote.
- V, the private value of the next unit to be traded.
- O, the *omega ratio*, estimated at recent trade prices, of the price X with respect to a threshold k defined at the next unit's valuation,

$$\frac{\mathbb{E}(X - k \mid X > k)\Pr(X > k)}{\mathbb{E}(k - X \mid X < k)\Pr(X < k)}.$$

- A, whether the action assigned to the player is buy or sell.
- D, the duration in time steps since the most recent trade.

To discretize the observations as is needed for Q-learning, we employ a *tile coding* system with a single tiling (Sherstov and Stone 2005; Schvartzman and Wellman 2009). That is, we set thresholds for the numerical features (P, O, V, and D), dividing each into three buckets, using threshold values chosen empirically in pilot simulations to provide evenly distributed observations over buckets.

ZI Regret Study

We designed an experiment to measure the regret of equilibrated static policies (ZI) with respect to either novel ZIs or an approximate best response over adaptive policies, derived via RL. As a sanity check, we wanted to show that our automated RL process could consistently find policies that outperformed the ZI baseline, as found in prior work (Schvartzman and Wellman 2009).

With a consistently effective learning process in hand, we sought to measure the strategic stability of equilibrated ZI mixed strategies with respect to approximate best responses derived via Q-learning. (By an *equilibrated* ZI mixed strategy, we mean a probability distribution over ZI strategy parameters exhibiting negligible empirical regret, relative to a fixed set of other ZI strategies.) We expected an arbitrarily chosen ZI pure strategy would have high regret with respect to a Q-learner or to other ZI strategies. More important, we hypothesized that as the set of ZI strategies is increased in size, the regret of the equilibrated mixed strategy with respect to either other ZI policies or a Q-learner will tend to diminish. The regret with respect to the other ZI policies necessarily approaches zero in the limit, but we expected there would remain a small positive regret with respect to a reinforcement learner. This regret represents the value of conditioning actions on market state in the CDA. If the measured regret is indeed small, this is evidence supporting the use of equilibrated ZI traders as a reasonable agent model.

Experiment Design

To measure the strategic stability of calibrated ZI strategies, we began with a set of 10 ZI parameterizations that were chosen heuristically for a combination of high fitness in our CDA environment and broad coverage of the space of viable policies. We then generated random subsets of our base strategy set, of several sizes, and used empirical game-theoretic analysis (EGTA) to find one or more symmetric Nash equilibria over each subset.[2] Next we challenged each distinct equilibrium mixed strategy of ZI policies, by training a Q-learning agent to deviate when all other agents play that mixed strategy. We also challenged each distinct equilibrium strategy with all 10 pure ZI strategies, to evaluate the regret with respect to the base strategy set.

ZI Strategy Set The 10 ZI strategies $(\underline{d}, \overline{d}, \eta)$ used in this study are as follows:

$$(0, 450, 0.5), (0, 600, 0.5), (90, 110, 0.5), (140, 160, 0.5),$$
$$(190, 210, 0.5), (280, 320, 0.5), (380, 420, 0.5), (380, 420, 1),$$
$$(460, 540, 0.5), (950, 1050, 0.5).$$

Henceforth, we will write a pure strategy as, for example, `280_320_.5`. A mixed strategy will be written as a series of ordered pairs of pure strategies and their probabilities, such as (`280_320_.5` $\times$ 0.1, `380_420_.5` $\times$ 0.9).

[2] As our games are finite and symmetric, symmetric NE necessarily exist. We numerically find approximate symmetric equilibria with negligible regret.

From this base set of 10 strategies, we randomly selected subsets of sizes two, five, or eight. Strategy subsets were selected uniformly randomly, rejecting duplicates. We used 30 distinct subsets of each size, in addition to the 10 subsets of one ZI strategy each, as well as the base set containing all 10 strategies. Overall, we conducted parallel experiments on 101 ZI strategy sets: 10 of size 1; 30 each of sizes 2, 5, and 8; and 1 of size 10.

EGTA Methods We used the methods of EGTA to find NE over each subset of ZI strategies. The essential EGTA process has been described at length in earlier studies (Wah, Lahaie, and Pennock 2016; Wah, Wright, and Wellman 2017), so we present only an overview. EGTA uses simulation to estimate the expected payoff for each agent in a strategy profile, and explores a space of profiles to identify approximate equilibria.

To test whether a mixed-strategy profile is an equilibrium, EGTA obtains payoff samples of each pure-strategy profile in the support of the mixed strategy (called the *subgame* of the support), as well as each pure-strategy profile where a single agent deviates to any other pure strategy. For example, if the 16 players in our game play strategies A and B with positive probability, it is necessary to sample payoffs of $i \in \{0, \ldots, 16\}$ agents playing A and the rest playing B; we then compute the expected payoff of the mixed strategy. Next, we would compute the payoffs for corresponding profiles where one agent deviates to any other pure strategy.

The number of samples required grows rapidly in the number of agents and strategies in support. To make this process tractable, we employ the *deviation-preserving reduction* (DPR) technique of (Wiedenbeck and Wellman 2012), to approximate our game of 16 players with a related 4-player game. We construct a reduced game's payoff table by running simulations of the full, 16-player game as follows. To estimate the payoff for a particular player in a 4-player reduced-game profile, we let that player control 1 agent in the 16-player simulation, while each of the other 3 players in the reduced game controls 5 agents in the full simulation.

We search for NE through a fully automated procedure that begins by testing whether each pure strategy in self-play is an equilibrium. The process then goes on to test equilibria over pairs of strategies, based on beneficial deviations found from the self-play profiles. Exploration continues, extending support size as necessary based on deviations found outside the current support. The process completes when an approximate NE is found with empirical regret less than a numerical tolerance, and all equilibrium candidates up to a current support size have been confirmed or refuted. For a given subgame, we use replicator dynamics (Taylor and Jonker 1978) and other numerical techniques to search for a symmetric NE over those strategies.

Pure-Strategy Regret Measurement We set out to accurately measure the regret of each equilibrium we found over a subset of ZI strategies, with respect to the base strategy set. This value serves as an empirical signal of how strategically stable a ZI mixed strategy is, with respect to the universe of all ZI strategies, if we believe that our base strategy set is sufficiently large and varied.

To estimate the regret of a mixed strategy M with respect to one of the 10 ZI strategies in the base strategy set, we run simulations where all but one agent plays a strategy sampled independently from M, but one agent deviates to that pure strategy. We explored the 10 pure strategies as if they were arms of a multi-armed bandit, seeking an upper bound on the regret with respect to the best arm. Initially, we sampled each strategy in turn, for $50{,}000$ simulations each. Thereafter, following each batch of 2500 simulations, we sampled a bootstrap distribution from the set of all payoffs of the current deviating strategy and selected the pure strategy with the highest 95th percentile for the mean payoff as the deviation to sample next. Thus, we obtained low-variance estimates for the upper confidence bound on those pure strategies that appeared to have a significant chance of being beneficial deviations. We terminated the process when either the greatest upper confidence bound became lower than the expected payoff of the equilibrium policy, or a total of $800{,}000$ simulations had been taken.

This pure-strategy regret measurement procedure allows us to evaluate the strategic stability of supposed NE, in case of approximation error caused by the player reduction of DPR. It is possible for a mixed strategy to be an exact Nash equilibrium in the *reduced game* of our DPR approximation, but not to be an equilibrium in the original game without player reduction, because the original game includes payoffs where non-round numbers of agents play each strategy, and with slightly different probability weights. Our regret measurement procedure employs the original, 16-player game, so it can sample payoffs where any number of background traders from 0 to 16 adopts each strategy in the equilibrium support. In this way, we empirically test whether the reduced-game NE are actual equilibria in the original game, and if not, how large their regret may be.

Q-Learning Regret Measurement We also aimed to measure the regret of each equilibrium over a subset of ZI strategies, against the best adaptive policy derived in a large policy space by Q-learning. This value gives a lower bound on how much better an adaptive agent can perform in the CDA than an agent that plays a fixed policy (ZI).

We conducted Q-learning against each equilibrium ZI mixed strategy as described above. To summarize and review the above, we performed a single run of Q-learning against each equilibrium, of 10^6 playouts in duration. Our exploration policy was ϵ-greedy, with $\epsilon = 0.1$. We modified conventional Q-learning based on empirical observations of the most useful special techniques in our setting: We truncated reward observations to ± 3000, added an artificial discount factor of 0.9 that decays to 1.0 with increasing iterations, used hand-tuned thresholds in each feature for observation bucketing, and used early stopping.

To measure the expected payoff of a policy from RL, we run our simulator with one agent playing the learned policy, and all others playing the baseline mixed strategy. We conduct at least 2×10^5 simulations per learned policy, and use the bootstrap to derive a confidence interval for the mean payoff. We then compare this payoff to the expected payoff of the baseline policy.

Results

In our ZI regret study, our automated policy learning method consistently found policies of greater expected value than the equilibrated ZI baselines, against which Q-learning was performed. However, the learned policies achieved only slightly greater payoff than those ZI equilibria that were derived from large sets of ZI pure strategies. The results suggest that there is a small but consistent advantage to conditioning actions on state in our CDA environment, relative to playing a well-calibrated mixed strategy of ZI policies. This benefit of an adaptive policy is small compared to the difference in expected payoff between a well-calibrated ZI strategy and a poorly chosen one.

Trend with respect to ZI Strategy Set Size

As the set of ZI strategies available to EGTA is augmented, we observe that the regret of the equilibrium mixed strategy over that set decreases, both with respect to our base set of ZI strategies, and to the adaptive strategy response produced by Q-learning. The regret with respect to other ZI strategies empirically is almost always lower than the regret with respect to adaptive strategies; or in other words, adaptive policies almost always achieve greater benefit in deviating from the ZI baseline than an alternative ZI strategy does.

In Fig. 1, we present the mean regret of each ZI subset's Nash equilibria, with respect to Q-learning (left) and with respect to the best-response ZI pure strategy (right). For example, in row 5 we display a marker for each equilibrium in each of the 30 ZI strategy subsets of size 5 that were randomly selected. In any row, each equilibrium is plotted with multiplicity equal to the number of strategy subsets of the appropriate size in which it occurs. With a line, we plot the mean regret of these equilibria for each strategy subset size.

Note in Fig. 1 how the regret of ZI equilibria grows smaller on average as the number of strategies equilibrated over increases from 1 to 10. This trend holds for regret with respect to Q-learning and with respect to the best-response ZI policy. The only exception to this trend is the small increase, from 4.4 to 4.7, in the mean regret with respect to Q-learning, from subset size 8 to size 10; this reversal may be due to noise in payoff sampling or the like. To provide a sense of scale in these payoff differences, we note that in the two Nash equilibria found over the base strategy set, the expected payoffs per background trader were 461.8 and 462.6.

This trend of ZI equilibrium regret growing smaller with increasing ZI strategy set size is supported by statistical hypothesis testing via the unpaired t-test. In these tests, we count each equilibrium's regret with a multiplicity equal to the number of strategy subsets in which it appears, similarly to the plot in Fig. 1. In the case of regret with respect to Q-learning, we find weak evidence (below statistical significance at 0.05 level) that the regret for size-one subsets is greater than size-two ($p = 0.14$), and strong evidence that regret for size-two is greater than size-five ($p = 10^{-8}$), and size-five is greater than size-eight ($p = 0.02$). In the case of regret with respect to ZI deviations, we find very similar hypothesis test results.

We also note in Fig. 1 that as the subset of ZI strategies equilibrated over is augmented, the regret of the ZI equilib-

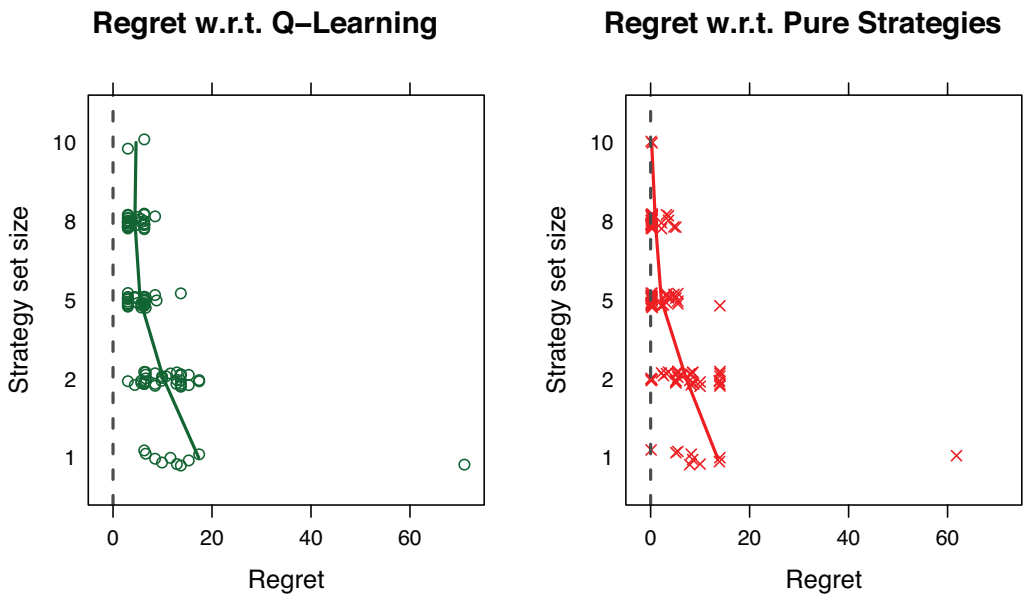

Figure 1: Regret of mixed-strategy Nash equilibria over ZI policy subsets. Each row corresponds to equilibria over ZI policy subsets of a given size. At left, regret is shown with respect to the Q-learning response; at right, with respect to the best-response ZI policy in the base policy set. Row means are plotted as lines.

rium with respect to the base strategy set approaches zero. This regret must be zero when the full strategy set is included, for a Nash equilibrium over the base strategy set cannot have any beneficial deviations within that set. For any subset of k strategies, drawn from a base set of N strategies, the likelihood of a zero-regret subset being selected is simply the likelihood of drawing a superset of the support of any Nash equilibrium of the base set. (The support of a Nash equilibrium is defined as the set of all pure strategies it uses with positive probability.)

Finally, observe how in Fig. 1 the regret of ZI equilibria with respect to Q-learning is always strictly positive, even as the number of ZI strategies equilibrated over becomes large. Indeed, the smallest regret of a ZI equilibrium with respect to Q-learning we find is 3.0, and the smallest mean regret for a subset size is 4.4, corresponding to strategy subsets of size 8. This suggests that there is a persistent benefit to adaptive policies, such as those we derive by RL in this study, relative to mixtures of ZI policies, even as those mixtures are calibrated over many parameterizations.

In Fig. 2, we present for each equilibrium the difference in regret between the response derived by Q-learning and the pure-strategy best response from the base strategy set. Each row corresponds to equilibria over subsets of ZI strategies of a certain size. Each equilibrium is plotted with a multiplicity equal to the number of strategy subsets in which it occurs.

We note that in almost all cases, our automated Q-learning procedure achieves more lift in payoff over the baseline than the ZI best response. In a few cases, it does not, likely due to insufficient iterations for Q-learning to converge, or the instability of Q-learning in the surface MDP of a POMDP. The increase in payoff improvement of Q-learning over ZI ranges from 3.0 to 4.4, over the various subset sizes, as shown by the solid line. These differences are statistically significant, based on paired t-tests, for subset sizes 1, 2, 5, and 8 ($p = 0.001$, 10^{-8}, 10^{-11}, and 10^{-13}, respectively). It is interesting that the lift of adaptive policies from Q-learning, relative to a non-adaptive ZI best response, appears roughly constant, even as the number of ZI strategies used for equilibration increases. This suggests a lingering benefit from conditioning actions on state, even against non-adaptive agents with carefully tuned parameters, providing a payoff gain of approximately 3.5.

Q-Learning Results

In the case of all 20 supposed equilibria, including the 10 pure strategies and 10 mixed strategies, Q-learning successfully discovered a beneficial deviation over the larger space of adaptive policies. The most common number of training iterations for the best policy was 10^6, which was the end of the training cycle. In several cases, however, an intermediate policy, corresponding to an earlier stopping time, produced a higher expected payoff.

In order to confirm that a learned policy had a payoff significantly greater than the equilibrium baseline, we played back the apparent best policy from a training run for 10^6 total playouts. We then took a bootstrap 95% confidence interval about the sample mean payoff, and in each case the lower bound thus obtained was greater than the mean payoff of the baseline profile.

Summary of Findings

In this series of experiments, our automated RL process consistently yielded a beneficial deviation, even against ZI

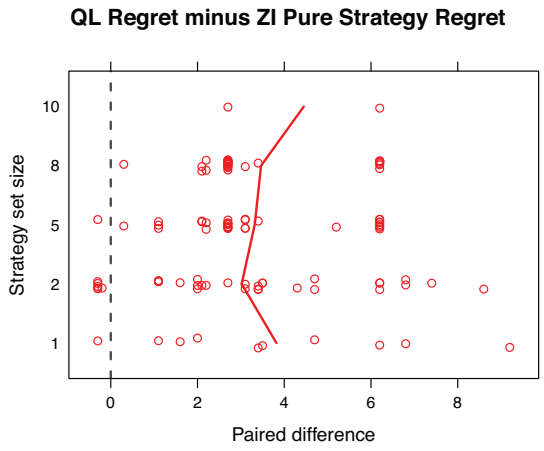

Figure 2: Paired difference in regret (w.r.t. Q-learning or ZI) of each mixed-strategy Nash equilibrium over a subset of ZI policies. We subtract the regret with respect to the best-response ZI policy in the base policy set, from regret with respect to the Q-learning response. Each row corresponds to equilibria over ZI policy subsets of a certain size. Row means are plotted as lines.

strategies that were equilibrated over the full base strategy set. However, the regret of equilibrated ZI policies grew lower, as the number of strategies used for equilibration was increased. The lift of an adaptive policy from Q-learning, relative to a ZI best response, appears to be almost constant on average, even as the strategic stability of the baseline ZI mixed strategy is increased. This benefit from policy adaptation is positive, but reasonably small, relative to the differences in payoffs between the deviating ZI policies we tested.

Summary of Learned Policies

We will take as a running example the pure ZI strategy equilibrium over the base strategy set, where all agents play `380_420_.5`. This example is chosen because it is an equilibrium over all the base strategies, so a Q-learner deviating successfully from it is making an improvement over any of its component pure strategies. Thus, it is an example of the benefit of policy adaptation over a fixed policy.

To study the range of successful adaptive deviations from `380_420_.5`, we performed 10 runs of Q-learning, selecting the best policy from each run. We analyzed the 10 resulting policies together, to find what they have in common to explain how they improve on the base strategies they are composed of. In Fig. 3, we present the distribution of mean surplus demanded by the adaptive agent, over the 10 policies derived by Q-learning against `380_420_.5`. By *mean surplus demanded*, we intend to say $\frac{\underline{d}+\overline{d}}{2}$, based on the parameters $\underline{d}$ and $\overline{d}$ of a ZI action. The Q-learner tends to demand slightly less surplus than the baseline ZI agents: Its mean surplus demanded is 382, compared to 400 for the others, and it demands strictly less mean surplus than the others in 56% of its arrivals. It demands strictly more mean surplus

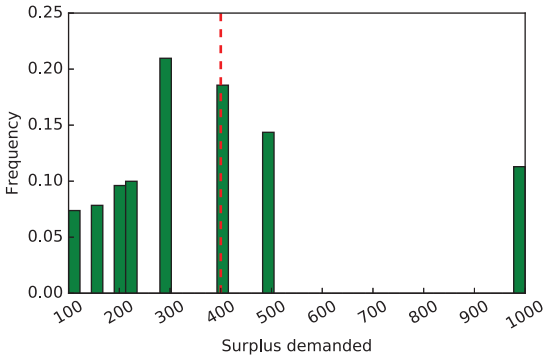

Figure 3: Histogram of the mean surplus demanded by 10 Q-learning derived policies, deviating from `380_420_.5`. Each state-action pair is weighted by the state's occurrence frequency. The mean surplus demanded by the equilibrium baseline policy is shown by the dotted red line.

than the others, perhaps opportunistically, in 26% of arrivals, demanding the same only 18% of the time. (In all of these figures, we weight each state-action pair of a learned policy by the state's occurrence frequency.)

Conclusion

We investigated the extent to which adaptive policies yield greater payoffs than non-adaptive, ZI policies in the CDA. Our work investigates whether a calibrated ZI strategy profile is a reasonable model for strategic behavior in the CDA. We provide insight into how a strategy that deviates from ZI can condition on market state to achieve greater surplus.

Recent works have employed CDA models to study high-frequency trading, prediction markets, frequent batch auctions, and market making. Many such studies rely on simple heuristic trading models, like ZI. It is thus vital to understand the strategic stability of non-adaptive, ZI strategy profiles, relative to adaptive policies for the CDA.

Our findings suggest traders can benefit from conditioning actions on state in the CDA, even against an equilibrated ZI profile. The magnitude of the regret of an equilibrated ZI profile, with respect to an adaptive deviating strategy, appears to be small, especially when ZI is equilibrated over many parameterizations.

References

Budish, E.; Cramton, P.; and Shim, J. 2015. The high-frequency trading arms race: Frequent batch auctions as a market design response. *Quarterly Journal of Economics* 130(4):1547–1621.

Chakraborty, M.; Das, S.; and Peabody, J. 2015. Price evolution in a continuous double auction prediction market with a scoring-rule based market maker. In *29th AAAI Conference on Artificial Intelligence*, 835–841.

Cliff, D., and Bruten, J. 1997. Zero is not enough: On the lower limit of agent intelligence for continuous double auction markets. Technical report, HP Laboratories.

De Luca, M., and Cliff, D. 2011. Human-agent auction interactions: Adaptive-aggressive agents dominate. In *22nd International Joint Conference on Artificial Intelligence*, 178–185.

Farmer, J. D.; Patelli, P.; and Zovko, I. I. 2005. The predictive power of zero intelligence in financial markets. *Proceedings of the National Academy of Sciences* 102:2254–2259.

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

Gode, D. K., and Sunder, S. 1993. Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality. *Journal of Political Economy* 101(1):119–137.

Goettler, R. L.; Parlour, C. A.; and Rajan, U. 2009. Informed traders and limit order markets. *Journal of Financial Economics* 93:67–87.

Jovanovic, B., and Menkveld, A. J. 2016. Middlemen in limit order markets.

Li, Z., and Das, S. 2016. An agent-based model of competition between financial exchanges: Can frequent call mechanisms drive trade away from CDAs? In *15th International Conference on Autonomous Agents and Multiagent Systems*, 50–58.

Mike, S., and Farmer, J. D. 2008. An empirical behavioral model of liquidity and volatility. *Journal of Economic Dynamics and Control* 32(1):200–234.

NYSE. 2017. NYSE Group Volume Records - Top 10 Years.

Schvartzman, L. J., and Wellman, M. P. 2009. Stronger CDA strategies through empirical game-theoretic analysis and reinforcement learning. In *8th International Conference on Autonomous Agents and Multiagent Systems*, 249–256.

Sherstov, A. A., and Stone, P. 2005. Function approximation via tile coding: Automating parameter choice. In *International Symposium on Abstraction, Reformulation, and Approximation*, 194–205.

Taylor, P. D., and Jonker, L. B. 1978. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences* 40(1-2):145–156.

Tesauro, G., and Bredin, J. L. 2002. Strategic sequential bidding in auctions using dynamic programming. In *1st International Joint Conference on Autonomous Agents and Multiagent Systems*, 591–598.

Tesauro, G., and Das, R. 2001. High-performance bidding agents for the continuous double auction. In *3rd ACM Conference on Electronic Commerce*, 206–209.

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

Wah, E., and Wellman, M. P. 2013. Latency arbitrage, market fragmentation, and efficiency: A two-market model. In *14th ACM Conference on Electronic Commerce*, 855–872.

Wah, E.; Lahaie, S.; and Pennock, D. M. 2016. An empirical game-theoretic analysis of price discovery in prediction markets. In *25th International Joint Conference on Artificial Intelligence*, 510–516.

Wah, E.; Wright, M. D.; and Wellman, M. P. 2017. Welfare effects of market making in continuous double auctions. *Journal of Artificial Intelligence Research* 59:613–650.

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*.

Watkins, C. J. C. H., and Dayan, P. 1992. Q-learning. *Machine Learning* 8(3-4):279–292.

Wiedenbeck, B., and Wellman, M. P. 2012. Scaling simulation-based game analysis through deviation-preserving reduction. In *11th International Conference on Autonomous Agents and Multiagent Systems*, 931–938.

Wilson, R. 1987. On equilibria of bid-ask markets. In Feiwel, G., ed., *Arrow and the Ascent of Modern Economic Theory*. MacMillan. 375–414.