Evaluating the Stability of Non-Adaptive Trading in Continuous Double Auctions

Mason Wright & Michael P. Wellman
University of Michigan

ABSTRACT

The continuous double auction (CDA) is the predominant mechanism in modern securities markets. Many agent-based analyses of CDA environments rely on simple non-adaptive trading strategies like Zero Intelligence (ZI), which (as their name suggests) are quite limited. We examine the viability of this reliance through empirical game-theoretic analysis in a plausible market environment. Specifically, we evaluate the strategic stability of equilibria defined over a small set of ZI traders with respect to strategies found by reinforcement learning (RL) applied over a much larger policy space. RL can indeed find beneficial deviations from equilibria 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 the adaptive strategies can earn, despite their much more expressive policy space. Our findings generally support the use of equilibrated ZI traders in CDA studies.

KEYWORDS

game theory; reinforcement learning; auctions

ACM Reference Format:
Mason Wright & Michael P. Wellman. 2018. Evaluating the Stability of Non-Adaptive Trading in Continuous Double Auctions. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 9 pages.

1 INTRODUCTION

The continuous double auction (CDA) is the preeminent security trading mechanism, accounting for trillions of dollars in transactions annually [16]. 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 other's 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 [33], or numeric solution of abstract models [10]. Many other research efforts aim to establish stylized facts about CDA market outcomes, based on simulation or analysis of rule-based traders in action [2, 3, 11]. 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 [5, 8, 18, 23–25].

Prior studies of heuristic strategies contribute to our understanding of the CDA mechanism, but results based on heuristic strategies 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 accounting for more state information. We seek to evaluate whether agents can benefit significantly by adopting more complex, adaptive policies, as extending to such larger strategy spaces may be difficult or costly. Recent work suggests the outcomes of economic simulation studies can be biased by which learning modality agents use, and reinforcement learning (RL) is a useful tool for exploring how the learning environment affects equilibrium results [13].

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 severely limited ability to adapt to market state. The version of ZI we use has a few parameters, which we tune via EGTA to find approximate Nash-equilibrium mixtures in the baseline set. Against these policies we train more adaptive trading policies using Q-learning [30]. 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 conditioning on market state.

We also conduct a detailed analysis on the nature of our learned CDA trading strategies, employing regression trees [18]. We then classify conditions where learned policies demand more or less surplus from trade than the ZI mixed-strategy baseline.

Additionally, we present simple analytical arguments on the conditions under which conditioning a CDA trading policy on market state can be beneficial. A further empirical analysis shows that market state indicates to a trading agent, most critically, the likelihood at which an order at some price will be executed. The most useful signals for the agent appear to be the recent order history and the current bid and ask.

1.1 Prior Work: Heuristic CDA Strategies

A Zero Intelligence 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 [9], 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 [7, 14, 15]. Several recent works have employed ZI traders in models of financial markets or prediction markets [3, 14, 26, 28].

Recently, Wah and Wellman [27] used ZI traders in a model of latency arbitrage between two markets, where an equilibration

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden. © 2018 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

process (EGTA) was used to find Nash-equilibrium parameters for the ZI agents, with or without a latency arbitrageur present. Li and Das [14] also used a ZI model to compare the CDA mechanism to frequent batch auctions. Wah et al. [28] used ZI again to study the effect of market makers on other traders in a CDA. Other groups have commonly employed ZI traders for similar ends, for instance Chakraborty et al. [3] used a ZI model to study the effectiveness of a market making strategy for prediction markets, as did Wah et al. [26] and, earlier, Othman [17].

It has always been clear that ZI is not an optimal trading strategy. Cliff and Bruten [5] 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 [8, 23, 24] and AA [6, 25] adjust the surplus demanded by the agent during a run, based on the prices of recent trades. Studies have shown such policies to be beneficial deviations from a single, fixed ZI policy [24, 29]. Even stronger policies have been derived by RL, to deviate beneficially from a mixed strategy of GDX agents [18]. 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 [18], which employed Q-learning in a particular CDA setting to derive beneficial deviations from profiles using a fixed ZI strategy, as well as more sophisticated strategies like ZIP and GDX. Our work builds on these methods 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. In addition, we analyze the relative importance of features for learning proposed in prior work, through regression over experimentally learned policies.

1.2 Research Contributions

- We evaluate the strategic stability of equilibrated ZI policies against (approximate) best responses from Q-learning. Some surplus is lost by forgoing adaptation, but the amount is small compared to the loss from non-equilibration of parameters.
- We study the nature of adaptive policies that outperform ZI.
- We provide intuition for when an adaptive agent can deviate beneficially from a ZI policy baseline, and offer insight into the tradeoffs facing such an agent.
- Our findings suggest equilibrating over many ZI policies yields a profile where no alternative ZI policy can earn significantly greater surplus, but an adaptive policy can still earn a small positive amount more. When this holds, the common practice of using equilibrated ZIs as an approximation of efficient behavior may be considered acceptable.

2 CDA MARKET MODEL

Our study is based on a CDA market model, similar to those of prior works by Wah et al. [27, 28]. The market has a single security and many traders. The security's value to an agent is 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 trade the security with one another via the CDA mechanism, by submitting limit orders to the market. Each agent can submit an order only in time steps when it *arrives* at the market, as determined by a random (exponential) inter-arrival time process; at 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 ZI policy.

2.1 Zero Intelligence

ZI is a simple strategy for CDA trading that can converge to efficient prices and allocations in many settings [9]. Our variant of ZI, introduced by Wah et al. [28], 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 uniformly drawn from interval $[\underline{d}, \overline{d}]$. 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.

2.2 Market Model Description

Our market model includes several adjustable parameters, such as the number of time steps per simulation and the degree of mean reversion in the security's fundamental value. We selected these parameters' values based on experience with similar models from prior studies [27, 28], in an effort to ensure a reasonable level of trading would occur in typical simulation runs, and simulations could be completed quickly enough to allow many epochs of RL.

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, $r_t \leftarrow \kappa\mu + (1-\kappa)r_{t-1} + \mathcal{N}(0, \sigma_s^2)$, where σ_s^2 is the fundamental shock variance, and κ the mean reversion parameter. Throughout this study, we take $\kappa = 0.01$ 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 setting their order prices.

Each background trader is assigned a private value vector, in which element θ_i gives the value of an additional security unit, given a current inventory i. This vector has length 20, because the

[1] We take $\mu = 10^5$. The specific level does not matter, as long as the evolving fundamental has negligible probability of hitting the zero lower bound.

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 takes $\sigma_p^2 = 2 \times 10^7$.

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

When a background trader playing a ZI strategy arrives, it cancels its previous order and is assigned with equal probability to buy or sell. 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 $\hat{r}_t + v_{i+1} - \alpha$, or for a seller is $\beta - (\hat{r}_t + v_i)$, where α is the ask, β 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 $s \sim U[\underline{d}, \overline{d}]$. 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 background trader earns a payoff equal to the final value of its inventory (fundamental plus private value), plus its final cash holdings. More formally, let I_t^j denote the inventory (stock holdings) of agent j at time t, and c_t^j its cash holdings. The final payoff for a trader j with positive final inventory ($I_T^j > 0$) is

$$\mathcal{U}^j = I_T^j r_T + \sum_{i=0}^{I_T^j - 1} \theta_i^j + c_T^j.$$

The payoff for nonpositive inventory is computed similarly.

For reinforcement learning of trading strategy, the learning agent l is trained on an interim reward signal $\mathcal{R}_t^l$ that it receives after each action, plus a final signal $\mathcal{R}_T^l$ it receives at the end. If the learning agent l arrives at time $t + k$, and its previous arrival had been at time t, the interim reward signal is given by

$$\mathcal{R}_{t+k}^l = \left(I_{t+k}^l \hat{r}_{t+k} - I_t^l \hat{r}_t\right) + \left(c_{t+k}^l - c_t^l\right) + \left(\Theta_{t+k}^l - \Theta_t^l\right),$$

where Θ_t^l is the cumulative private value of l's inventory at time t. These interim rewards are structured such that their sum must equal the trader's overall payoff: $\sum_t \mathcal{R}_t^l = \mathcal{U}^l$.

3 DEFINITIONS

We use many standard terms from game theory, defined here for completeness. By an agent *policy* or player *pure strategy*, we mean a mapping from the set of observation states to the (possibly stochastic) action taken 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 player. A *symmetric* profile assigns the same strategy to each player. A *Nash equilibrium* (NE) is a profile such that no player can achieve a higher expected payoff by unilaterally deviating from its assigned strategy to any alternative. The *regret* of a profile is the maximum over players, of the maximum gain in expected payoff obtainable by deviating to any alternative strategy. A Nash equilibrium thus has zero regret.

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 player 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 player gains in expectation by deviating to a strategy in $\mathcal{S}'$.

4 REINFORCEMENT LEARNING METHODS

To learn improved background trader policies, we first fix the policies of all but one agent, which converts the trading game into a decision problem for the one strategic agent. This agent can then use RL to search for an approximate best response to the policies of the others. We tested several RL approaches[2] before settling on a variant of *Q-learning* [30] that empirically worked well in our setting. The Q-learning agent progresses through a sequence of observing states s, getting reward $\mathcal{R}$, 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 $(s, a, \mathcal{R}, s')$, the agent performs a Q-learning value update,

$$Q(s, a) \leftarrow (1 - \rho) Q(s, a) + \rho \left(\mathcal{R} + \gamma \max_{a'} Q(s', a')\right),$$

where ρ 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 $\rho(s, a)$ 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.

Note that the omega ratio seeks to measure the recent favorability of the market, in terms of the expected upside in proportion to the expected downside of purchasing a unit of stock. To discretize the observations for Q-learning, we employ a *tile coding* system with a single tiling [18, 19]. That is, we threshold the numerical features (P, O, V, and D), dividing each into three buckets, using boundary values chosen empirically in pilot simulations to provide evenly distributed observations over buckets.

The action set of the Q-learning agent is the same as the ZI strategy set available to the other background traders. That is, the learner trains a policy that maps each observation state to one of the 10 ZI strategies listed below in Section 5.1.

[2] In particular, we also tried Sarsa with eligibility traces [21] and POMCP [20], with neither producing results better than Q-learning for our problem.

5 ZI REGRET STUDY

We designed an experiment to measure the regret of equilibrated static policies (ZI) with respect to either alternative ZIs or 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 [18].

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.

We began with a set of 10 ZI policies, selected heuristically for high fitness and broad coverage. We 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 in each subset.[3] Next we challenged each ZI equilibrium strategy, by training a Q-learner against other agents playing that mixed strategy. We also challenged each distinct equilibrium strategy with each of our 10 pure ZI strategies, to evaluate regret with respect to the base strategy set.

5.1 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 write a pure strategy as, for example, `280_320_.5`. A mixed strategy is a set of ordered pairs of pure strategies and their probabilities, such as $\{$`280_320_.5` $\times$ 0.1, `380_420_.5` $\times$ 0.9$\}$.

As noted above, we selected these particular ZI parameterizations with the goal of having broad coverage of the space of reasonably high-fitness strategies. We had previously observed the relative strategic stability of many ZI strategies in a small pilot study, as well as in prior work that used a similar market model [28].

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 singleton subsets, and the set of all 10. 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.

[3] As our games are finite and symmetric, symmetric NE necessarily exist [4]. We numerically find approximate symmetric equilibria with negligible regret.

5.2 EGTA Methods

The essential EGTA process has been described at length elsewhere [12, 26, 28, 29, 31], so we present only an overview. EGTA employs simulation to estimate the expected payoff for each agent in a strategy profile, and explores a space of profiles to identify approximate equilibria. We used the methods of EGTA to find NE over each subset of ZI strategies. A total of 20 distinct equilibria were identified across these subsets, including the 10 pure profiles that are equilibria for the respective singleton sets.

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 (i.e., profiles played with positive probability; we term the collection a *subgame*), 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 sample count 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 [32], approximating our game of 16 players with a related 4-player game. We construct the 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 [22] and other numerical techniques to search for a symmetric NE over those strategies.

5.3 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 lets us evaluate the strategic stability of supposed NE, in case of approximation error caused by player reduction. It is possible for a mixed strategy to be an exact Nash equilibrium in the reduced game of our DPR approximation, but not an equilibrium in the original game, because the distribution of profiles in the original game includes many not reflected in the reduced game. Our procedure evaluates the regret of reduced-game NE with respect to the original game, that is, it samples profiles where any number of background traders from 0 to 16 adopts each strategy in the equilibrium support.

5.4 Q-Learning Regret Measurement

We 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 provides a lower bound on how much improvement can be obtained through adaptive trading relative to the ZI equilibria.

We conducted Q-learning against each equilibrium ZI profile as described above. In summary, we performed a single run of Q-learning against each equilibrium, of 10^6 playouts. Our exploration policy was ϵ-greedy, with $\epsilon = 0.1$. We modified conventional Q-learning based on what we found to be useful tricks in our setting: We truncated reward observations to ±3000, used 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.

Note that the policy space used by our Q-learner is a strict superset of the ZI policies in the base strategy set. This means that the Q-learner can in theory deviate at least as successfully against any opponent profile as the best fixed response from the base strategy set. However, a Q-learner may not always match or outperform the ZI best response, due to insufficient time to converge, or problems converging in a POMDP.

6 RESULTS

In our experiments, the RL method consistently found policies of greater expected value than the equilibrated ZI baselines. 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 between a well-calibrated ZI strategy and a poorly chosen one.

6.1 Studying Effects of ZI Strategy Set Size

As the set of ZI strategies available to EGTA is augmented, 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.

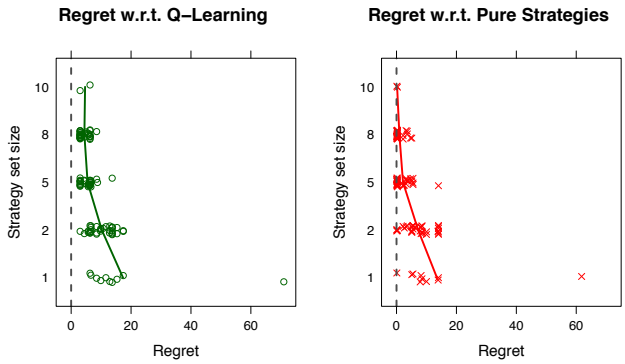

Figure 1: Regret of equilibria over ZI policy subsets. Each row comprises equilibria over ZI policy subsets of a given size. Left: regret w.r.t. Q-learning response; right: regret w.r.t. best-response ZI policy of full policy set. Solid lines present row means.

Fig. 1 presents the regrets of each ZI subset's NEs, 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 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 equilibrium 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.

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 equilibrated over many parameterizations.

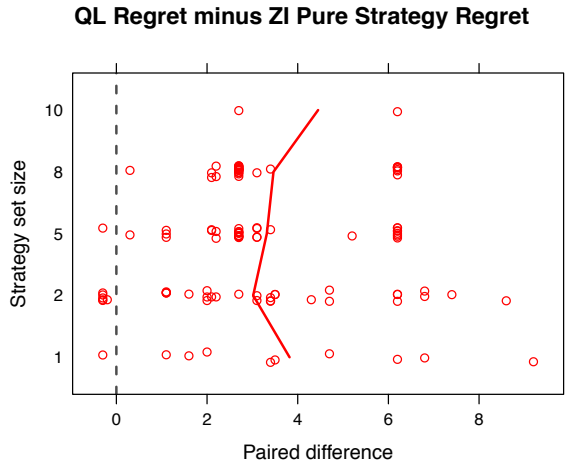

Figure 2: Paired difference in regret (w.r.t. Q-learning or ZI) for each equilibrium, in ZI policy subsets of various sizes. Regret w.r.t. ZI is subtracted from regret w.r.t. the Q-learning response. Each row comprises equilibria over ZI policy subsets of the given size. Solid lines present row means.

Fig. 2 presents 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 Q-learning procedure achieves more lift in payoff over the baseline than the ZI best response. In

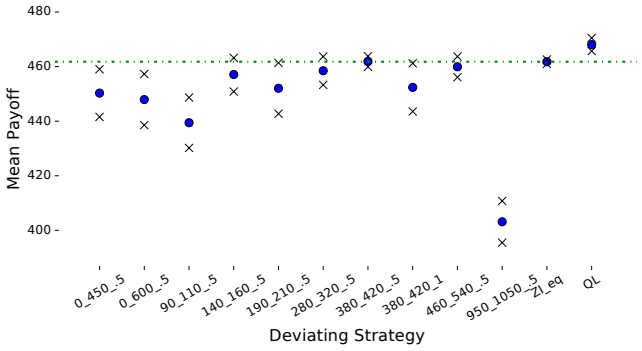

Figure 3: Empirical payoff per deviation from a base strategy set equilibrium. Each column is a deviating strategy: blue dot is sample mean payoff, crosses are 95% confidence interval. `ZI_eq` is the equilibrium strategy (`380_420_.5`). `QL` is the learned response. Dotted line is equilibrium payoff.

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 mean 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.

6.1.1 Deviation Payoffs of an Example Equilibrium. Let us examine the range of payoffs for an agent unilaterally deviating from an equilibrium over ZI strategies, to pure strategies in the base strategy set. We take as our example the pure-strategy Nash equilibrium over the base strategy set, where all background traders play `380_420_.5`.

In Fig. 3, we present the empirical distribution of payoffs for each unilateral deviation from this equilibrium profile. The crosses in each column indicate a 95% confidence interval for the mean, derived via bootstrap sampling. Because we automatically collect more samples for pure-strategy deviations that have higher upper confidence bounds, as in the UCB-1 method of sampling [1], some confidence intervals are narrower than others.

Note how all ZI strategies except `380_420_.5` have sample mean payoffs lower than the equilibrium payoff, as expected. The equilibrium action of `380_420_.5` has a mean payoff on resampling almost exactly equal to the independent estimate for the equilibrium payoff, shown as `ZI_eq`. This indicates we have likely collected enough samples per policy that sampling error is under control.

Finally, observe that the expected payoff of the adaptive policy derived from Q-learning (`QL`) is significantly higher than the equilibrium payoff or any pure-strategy deviation, at 468.1 over an equilibrium payoff of 461.8. This gain from deviating to an adaptive

policy, however, is smaller than the loss that would occur by deviating to several inferior ZI pure strategies. Thus, in this example we can see that the payoff gain from conditioning on state as this Q-learner does is moderate, relative to the benefit of choosing a well-calibrated ZI strategy instead of an arbitrary one.

6.1.2 Q-Learning Performance against ZI Strategy Sets. For all 20 supposed equilibria over ZI strategies, including the 10 pure strategies and 10 mixed strategies, Q-learning successfully discovered a beneficial deviation over the larger space of adaptive policies. 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.

6.1.3 Summary of Effects of ZI Strategy Set Size. In this series of experiments, our automated RL process consistently yielded a beneficial deviation, even against ZI 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.

6.2 Intuition Behind Successful Policy Adaptation

In our model CDA environment, there are several useful features of market state that are unknown to the agent when it arrives at the market: (a) the pure strategy drawn by each other agent from a publicly known symmetric mixed strategy; (b) the private value vector of each other agent; (c) the inventory of each other agent; and (d) any orders beyond the best bid and ask in the limit-order book. There are other unobserved aspects of state, such as previous fundamental values and agent arrival times, but because the fundamental value and arrival processes are memoryless, these state features are of no value in choosing an action.

We say a policy in the CDA *conditions on state*, if the probability distribution over surplus demanded differs, based on the history of the agent's observations and actions. If a policy does not condition on state, the agent has the same probability distribution over surplus demanded, regardless of the observed market state.

An agent in our model can benefit through observations, only if its observations help it to predict the orders of the other agents, which are the only environment events that affect its payoff besides the final fundamental value. The agent can be intuitively viewed as attempting to predict the likelihood its order will transact, as a function of the surplus demanded. Some strategy proposals in the literature, such as GD [8], attempt to maximize expected revenue based on such a probabilistic prediction. Similarly, we expected that in our analysis of learned policies, we would find the agent to condition its actions on state in a way that appears to demand more surplus in favorable conditions.

6.3 Isolating the Effect of ZI Greediness

Recall that the η parameter of our version of ZI tunes the agent's tendency to immediately trade at a quote that offers a corresponding fraction of its desired surplus, in lieu of placing a limit order demanding the full surplus. Thus, η can be viewed as a "greediness" parameter, where $\eta \ll 1$ means the agent greedily accepts quotes with surplus far below its desired amount, while $\eta = 1$ means the agent always demands the full amount.

Setting $\eta < 1$ thus provides a rudimentary way for agents to condition their behavior on market state. We performed a follow-up experiment to remove the effect of $\eta < 1$ from our results. In this experiment, all background traders other than the learning agent play variants of the 10 ZI strategies listed in Section 5.1, modified such that $\eta = 1$. This led to one duplicate, yielding a new set of 9 ZI strategies. The learning agent retained the original 10 ZI strategies as its available actions.

We used EGTA as before to find a Nash equilibrium over the strategies of the non-learning agents. In equilibrium, all ZI agents used the strategy `190_210_1`. Note that the mean surplus demanded of this equilibrium is only 200, much lower than the 390 or 400 of the equilibria with $\eta = 0.5$ available.[4] This decrease is likely due to how $\eta = 0.5$ allows agents to place high-demand limit orders, yet also take opportunities to trade greedily; with $\eta = 1$, agents must demand lower surplus at equilibrium.

Over three independent runs of Q-learning, learners achieved a mean gain in expected payoff of 10.9, relative to the equilibrium payoff of background traders. This is a much higher deviation gain than learners averaged against the flexible-η equilibrium, which was just 4.7. This larger gain by the adaptive agent suggests that ZI with $\eta = 0.5$ is able to realize much of the benefit of adaptive behavior in the CDA, in spite of the simplicity of this conditionality. If η is fixed at 1, however, the room for gains from adaptivity are significantly greater.

6.4 Analysis of Learned Policies

We 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 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. 4, we present the distribution of mean surplus demanded by the adaptive agent, over the 10 policies derived by Q-learning against `380_420_.5`. The Q-learner tends to demand slightly less surplus than the baseline ZI agents: 379 on average, compared to 400 for the ZI, and it demands strictly less surplus than the ZI in 56% of its arrivals. It demands strictly more mean surplus than the others, perhaps opportunistically, in 25% of arrivals, and the same amount 19% of the time.

[4] The *mean surplus demanded* of a ZI strategy is $(\underline{d} + \overline{d})/2$.

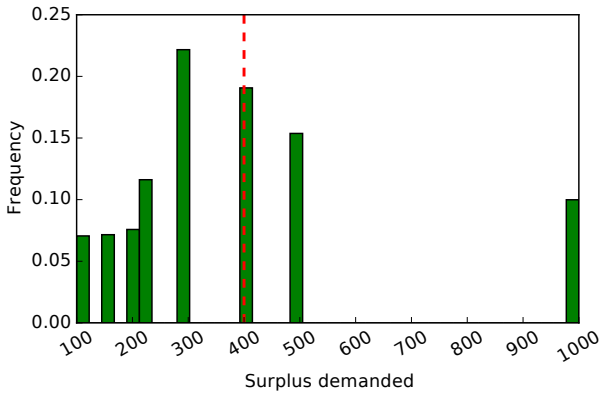

Figure 4: Histogram of the mean surplus demanded by 10 Q-learned policies, deviating from `380_420_.5`. Each state-action pair is weighted by the state's occurrence frequency. The mean surplus demanded by the equilibrium baseline policy is shown by the dotted red line.

6.4.1 Machine Learning for Policy Analysis. It has been noted that tabular policies are difficult to understand intuitively [18]. To draw useful insights into the adaptive policies for CDA trading from Q-learning, we use machine learning techniques that can help to summarize complicated policies.

Here we analyze the set of policies from Q-learning against a baseline strategy of `380_420_.5`, as above, but without the feature D, which did not appear useful in a greedy feature-elimination study (not shown). We use regression and classification methods to summarize or draw insight from the set of policies learned in this setting, over multiple independent runs of Q-learning.

Our first effort was to predict the mean surplus demanded by the learner's action in a given market state. Market state is determined by the four features P, V, O, and A. Recall that A, the action type in $\{buy, sell\}$, is a binary feature, while the other features are partitioned into a low, medium, or high bucket. We encode the binary feature as 1 for buy, 0 for sell. We encode ternary features as 0 for low, 1 for medium, 2 for high.

Least squares regression yields a fairly poor fit, with an MAE of 159.7. (We weight each state-action pair according to its state's occurrence frequency, for both learning and evaluation.) The learned coefficients are $P = 74$, $O = 43$, $A = 29$, $V = 4$. This agrees with the results of our greedy feature elimination study (not shown), that P is by far the most critical feature for learning, and O is also important. However, it suggests that V, the private value of the next unit traded, has little linear effect on surplus demanded.

Due to the poor linear fit, we tested decision tree regression, splitting to reduce MSE, with a maximum depth of 3. We found normalized feature importances of $P = 0.58$, $O = 0.34$, $A = 0.01$, $V = 0.07$. Again, P is the most important feature, with O second, though now V is third. These results further support the idea that P is the most useful feature for an adaptive agent in the CDA.

Beyond relative feature importances, our results give insight into the directional effect of each feature on surplus demanded. It appears, from the high positive coefficient in linear regression, that a high P indicates that the agent can earn a large surplus by trading immediately, so it may be an opportune time to demand more surplus than usual.

Similar to our decision tree regression study, we performed decision tree classification, classifying the mean surplus demanded by each state's action as less than, equal to, or greater than the mean surplus demand of the equilibrium ZI (in this case, 400). Classifier splits are chosen based on the Gini coefficient. We achieved a weighted zero-one loss of 0.34 with a decision tree of depth 3. The relative feature importances in this tree were almost identical to those of the decision tree regressor.

A simple, depth-2 decision tree classifier predicts the adaptive agent will demand less surplus than the baseline, if neither P nor O is high. Only with high P (immediate surplus available) or high O (recent transaction prices) will the agent demand higher surplus than the baseline. Deeper decision trees yield higher accuracy, but their rule sets do not appear as interpretable in this case.

6.4.2 Summary of Learned Policy Analysis. We discussed the intuition behind what market information is most useful to trading agents in our market model, as well as how agents trade off between the likelihood an order will transact and the surplus demanded. We showed that against a running example ZI equilibrium profile, successful adaptive agents typically demand slightly less surplus than the baseline agents, but occasionally demand much more. Adaptive agents benefit most from conditioning on the features P (immediate surplus available), O (signal of recent transaction prices), and V (private value of next unit traded). Adaptive agents tend to demand more surplus when P or O is high.

7 CONCLUSION

We investigated the extent to which adaptive policies yield greater payoffs than non-adaptive, ZI policies at equilibrium in the CDA. We thus addressed whether a calibrated ZI strategy profile is a reasonable model for strategic behavior in the CDA.

Our findings suggest traders can benefit from conditioning actions on state in the CDA, even against an equilibrated ZI profile. But the size 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. The extension of ZI to support immediate trading when a fraction η of demanded surplus is available appears to be pivotal for this finding, as the gains to adaptivity are much greater when strategies are restricted to $\eta = 1$. With this limited amount of conditionality, our results support the use of equilibrated ZI in CDA studies.

Further, we provided insight into how a strategy that deviates from ZI can condition on market state to achieve greater surplus. It seems the most useful state features for adaptive CDA traders are the immediate surplus available, the recent history of transaction prices, and the private value of the next unit traded. These signals can help the agent to determine how to trade off the likelihood of an order transacting against the amount of surplus requested. In particular, adaptive agents appear to benefit from demanding greater surplus when the current bid or ask is favorable, otherwise demanding less surplus to increase the likelihood of trading at all.

REFERENCES

[1] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine Learning* 47, 2-3 (2002), 235–256.

[2] Eric Budish, Peter Cramton, and John Shim. 2015. The high-frequency trading arms race: Frequent batch auctions as a market design response. *Quarterly Journal of Economics* 130, 4 (2015), 1547–1621.

[3] Mithun Chakraborty, Sanmay Das, and Justin Peabody. 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.

[4] Shih-Fen Cheng, Daniel M. Reeves, Yevgeniy Vorobeychik, and Michael P. Wellman. 2004. Notes on equilibria in symmetric games. In *AAMAS-04 Workshop on Game-Theoretic and Decision-Theoretic Agents*.

[5] Dave Cliff and Janet Bruten. 1997. *Zero is not enough: On the lower limit of agent intelligence for continuous double auction markets.* Technical Report. HP Laboratories.

[6] Marco De Luca and Dave Cliff. 2011. Human-agent auction interactions: Adaptive-aggressive agents dominate. In *22nd International Joint Conference on Artificial Intelligence*. 178–185.

[7] J. Doyne Farmer, Paolo Patelli, and Ilija I. Zovko. 2005. The predictive power of Zero Intelligence in financial markets. *Proceedings of the National Academy of Sciences* 102 (2005), 2254–2259.

[8] Steven Gjerstad and John Dickhaut. 1998. Price formation in double auctions. *Games & Economic Behavior* 22 (1998), 1–29.

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

[10] Ronald L. Goettler, Christine A. Parlour, and Uday Rajan. 2009. Informed traders and limit order markets. *Journal of Financial Economics* 93 (2009), 67–87.

[11] Boyan Jovanovic and Albert J. Menkveld. 2016. *Middlemen in limit order markets.* Technical Report. New York University and VU University Amsterdam.

[12] Marc Lanctot, Vinicius Zambaldi, Audrūnas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. 2017. A unified game-theoretic approach to multiagent reinforcement learning. In *Thirty-First Annual Conference on Neural Information Processing Systems*.

[13] Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-agent reinforcement learning in sequential social dilemmas. In *16th International Conference on Autonomous Agents and Multi-Agent Systems*. 464–473.

[14] Zhuoshu Li and Sanmay Das. 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.

[15] Szabolcs Mike and J. Doyne Farmer. 2008. An empirical behavioral model of liquidity and volatility. *Journal of Economic Dynamics and Control* 32, 1 (2008), 200–234.

[16] NYSE. 2017. NYSE Group Volume Records - Top 10 Years. (2017). http://www.nyxdata.com/nysedata/asp/factbook/viewer_edition.asp

[17] Abraham Othman. 2008. Zero-intelligence agents in prediction markets. In *7th International Conference on Autonomous Agents and Multiagent Systems*. 879–886.

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

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

[20] David Silver and Joel Veness. 2010. Monte-Carlo planning in large POMDPs. In *Advances in Neural Information Processing Systems*. 2164–2172.

[21] Satinder Singh and Richard Sutton. 1996. Reinforcement learning with replacing eligibility traces. *Machine learning* 22, 1-3 (1996), 123–158.

[22] Peter D. Taylor and Leo B. Jonker. 1978. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences* 40, 1-2 (1978), 145–156.

[23] Gerald Tesauro and Jonathan L. Bredin. 2002. Strategic sequential bidding in auctions using dynamic programming. In *1st International Joint Conference on Autonomous Agents and Multiagent Systems*. 591–598.

[24] Gerald Tesauro and Rajarshi Das. 2001. High-performance bidding agents for the continuous double auction. In *3rd ACM Conference on Electronic Commerce*. 206–209.

[25] Perukrishnen Vytelingum, Dave Cliff, and Nicholas R. Jennings. 2008. Strategic bidding in continuous double auctions. *Artificial Intelligence* 172 (2008), 1700–1729.

[26] Elaine Wah, Sébastien Lahaie, and David M. Pennock. 2016. An empirical game-theoretic analysis of price discovery in prediction markets. In *25th International Joint Conference on Artificial Intelligence*. 510–516.

[27] Elaine Wah and Michael P. Wellman. 2016. Latency arbitrage in fragmented markets: A strategic agent-based analysis. *Algorithmic Finance* 5 (2016), 69–93.

[28] Elaine Wah, Mason Wright, and Michael P. Wellman. 2017. Welfare effects of market making in continuous double auctions. *Journal of Artificial Intelligence Research* 59 (2017), 613–650.

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

[30] Christopher J. C. H. Watkins and Peter Dayan. 1992. Q-learning. *Machine Learning* 8, 3-4 (1992), 279–292.

[31] Michael P. Wellman. 2016. Putting the agent in agent-based modeling. *Autonomous Agents and Multi-Agent Systems* 30 (2016), 1175–1189.

[32] Bryce Wiedenbeck and Michael P. Wellman. 2012. Scaling simulation-based game analysis through deviation-preserving reduction. In *11th International Conference on Autonomous Agents and Multiagent Systems*. 931–938.

[33] Robert Wilson. 1987. On equilibria of bid-ask markets. In *Arrow and the Ascent of Modern Economic Theory*, G. Feiwel (Ed.). MacMillan, 375–414.