A Multiagent Model of Efficient and Sustainable Financial Markets

Betty Shea
University of British Columbia
sheaws@cs.ubc.ca

Mark Schmidt
University of British Columbia
schmidtm@cs.ubc.ca

Maryam Kamgarpour
University of British Columbia
maryamk@ece.ubc.ca

Abstract

In this paper, we introduce a model of a financial market as a multiagent repeated game where the players are market makers. We formalize the concept of market making and the parameters of the game. Our main contribution is a framework that combines game theory and machine learning methods. This approach allows us to consider markets on both a macro level, through game outcomes, and on a micro level, through the optimization efforts of players. Using simple equilibrium analysis, we show that our model explains situations where market outcomes are inefficient or unsustainable. We further apply our model to simulate market makers in the SP500 E-mini futures market and show that players learn to adapt their quotes to different market conditions.

1 Introduction

As the name suggests, the role of a market maker is to create a market for a financial asset i. A market maker accomplishes this by simultaneously providing buy and sell prices of i continuously throughout a trading day. In return for her readiness to transact i, a market maker hopes to profit from the difference between the price she sells and buys i.

A financial market is generally supported by multiple market makers. This lends itself to a multiagent game where players simultaneously maximize their own objectives, and where the payoff for each player is determined both by her own actions *and* those of other players. We model this using a repeated, coupled forward-reverse auction framework where the players are market makers.

Market making has been studied in both single agent settings ([8, 4]) and in multiagent setting ([3, 9]). Existing research uses (deep) reinforcement learning to study specific financial assets such as agent-dealer markets [3], Bitcoin [9] and corporate bonds [4]. The goal is usually to derive a strategy that predict future prices.

In contrast, our primary focus is to understand how the environment shapes the behaviour of market makers and, in turn, market dynamics and outcomes. The main contribution of this paper is to develop a model of financial markets that use game theory as a foundation for machine learning methods. This approach allows us to consider markets on a micro level, for example, by using players that employ online optimization or reinforcement learning strategies, and to analyze outcomes on a macro level using concepts of equilibria and welfare. To the best of our knowledge, combining game theory and optimization for financial policy design is a novel approach within the context of machine learning.

This paper begins by defining the model and measures of efficiency and sustainability. In Section 3, we use equilibrium analyses to show conditions under which markets may become inefficient or

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

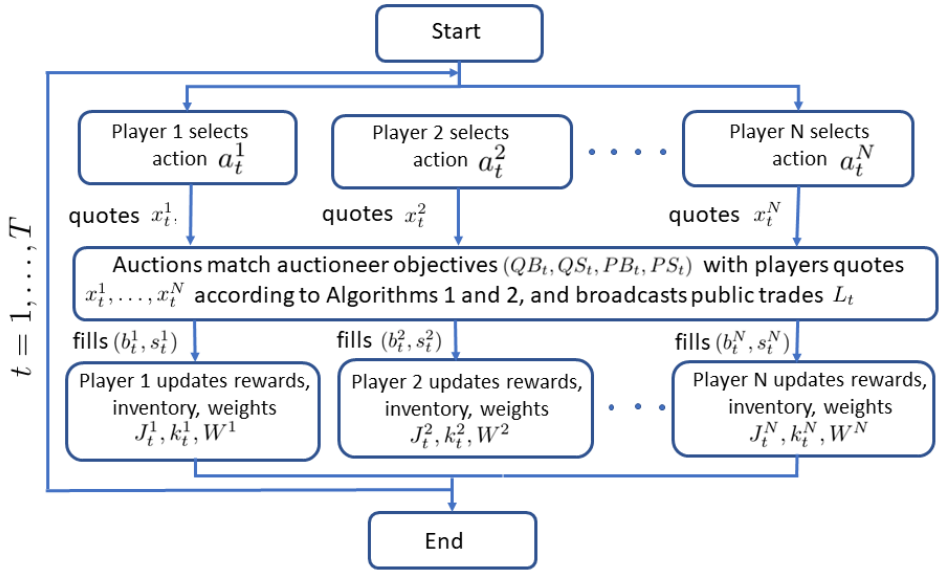

Figure 1: Repeated auction framework.

unsustainable. In Section 4, we provide experimental results using financial data to simulate players running a no-regret strategy in a bandit environment. Mechanism design is a natural application of our model and we discuss this in the context of future directions in Section 5.

2 Model definition

2.1 Framework

We propose a model of a market, in the form of a live order book, for a single financial asset i. This process is summarized in Figure 1. The model is a multiagent, repeated game of finite length where, at each time period $t \in \{1, \ldots, T\}$, a forward and a reverse auction occurs simultaneously. Let the set of players be $\mathcal{N} = \{1, \ldots, N\}$. The players represent the market makers and the auctioneer represents the demand to acquire and to dispose of i.

At every time period t, player j decides on an action $a_t^j \in \mathbb{R}$ that determines her quote x_t^j. The quote consists of a bid to buy i in the forward auction and an offer to sell i in the reverse auction. The combined quotes from all players feed into an allocation rule that matches the auctioneer's objectives to buy and sell i with the players' quotes. Each player receives her allocations. *Netted fills*, or offsetting buy and sell fills, contribute to her market making rewards and the remainder enters into her inventory. For example, if player j bids a quantity of five and offers a quantity of two, and if she is allocated three buys and two sells, her netted fills of two buys and two sells contribute to her market-making profit and the remainder of three buys enter into her inventory. Finally, each player updates her strategy for the next time period.

2.2 Environment

The fair value of asset i at every time step t, denoted V_t, is assumed to be given. At time t, the auctioneer's objectives for i are represented by four variables: the quantity it wants to buy and sell, QB_t and QS_t, and the maximum and minimum price it is willing to pay to transact PB_t and PS_t. We use the term *willingness to pay* to mean the amount in excess of fair value of the auctioneer's prices, i.e. $PB_t - V_t$ and $V_t - PS_t$.

The environment of the game is one of partial information. At time t, every player knows $V_{t'}$ for $t' = 1, \ldots, t$ but not the auctioneer's objective $(QB_{t'}, QS_{t'}, PB_{t'}, PS_{t'})$ for any t'. In addition, a player's action set, her inventory levels, her maximum trading size and the maximum amount of inventory she could carry is known only to her. The production cost for every player, however, is

known to all players and is assumed to be zero. At the conclusion of the auction, player j knows her own allocations, b_t^j and s_t^j, and the list of aggregated trades L_t. Suppose the auctions at time t results in trades occurring at d_t distinct prices. Then the list of market trades consist of d_t tuples $L_t = \left\{[L_{t,1}^{size}, L_{t,1}^{price}], \ldots, [L_{t,d_t}^{size}, L_{t,d_t}^{price}]\right\}$ where $L^{size} \in \mathbb{Z}_+$ and $L^{price} \in \mathbb{R}_+$.

2.3 Players

A player is governed by two parameters, her maximum inventory level k_{max}^j and her maximum trading size q^j. Her action set, denoted A^j, is finite and fixed. She starts with zero inventory and, at every iteration t of the repeated game, she holds some amount of i in inventory, denoted $k_t^j \in \mathbb{Z}$. She selects an action $a_t^j \in A^j$ and submits a simultaneous order to buy and to sell x_t^j based on her choice a_t^j. We refer to $x_t^j(a_t^j)$ as her *quote* for i. In essence, a player submits quantities that ensure that $|k_{t+1}^j|$ is at most k_{max}^j and that her expected profit per trade, her *edge*, is a_t^j. Formally

$$x_t^j(a_t^j) = (qb_t^j, qs_t^j, pb_t^j, ps_t^j) \tag{1}$$

where $qb_t^j = \min\{q^j, k_{max}^j - k_t^j\}$, $qs_t^j = \min\{q^j, k_{max}^j + k_t^j\}$, $pb_t^j = V_t - a_t^j$ and $ps_t^j = V_t + a_t^j$. A player *quotes tight* when a_t^j is chosen to be small and *quotes wide* when a_t^j is large.

The simultaneous decisions of all N players form the strategy profile $a_t = [a_t^j]_{j \in N}$. The auctioneer determines the outcome of the auction based on an allocation rule that takes $x_t(a_t)$ as an input.

After the auction, player j buys b_t^j and sells s_t^j units of i. We refer to b_t^j and s_t^j as her *fills* at time t. She then updates her rewards, inventory and strategy. The reward at time t consist of two quantities. Her market-making profit from netted fills, i.e. offsetting b_t^j and s_t^j, which is always positive

$$Jq_t^j(x_t(a_t)) = \min\{b_t^j, s_t^j\} \cdot (ps_t^j - pb_t^j) \tag{2}$$

The change in the value of her inventory which could be positive or negative

$$Jp_t^j(x_t(a_t)) = \begin{cases} \left(V_t - \overline{pb}_t^j\right) \cdot k_{t-1}^j, & k_{t-1}^j \geq 0 \\ \left(\overline{ps}_t^j - V_t\right) \cdot k_{t-1}^j, & k_{t-1}^j < 0 \end{cases} \tag{3}$$

where $\overline{pb}_t^j$ and $\overline{ps}_t^j$ represent the weighted average cost of acquiring inventory. For example, if player j adds 1 unit at price 100 to her inventory at $t = 1$ and 2 units at price 99 at $t = 2$, $\overline{pb}_2^j = 99 1/3$.

Above, (3) makes the simplifying assumption that inventory value is "marked to mid-market". In other words, were player j to liquidate her inventory at time t, she could do so at fair value V_t. Her per unit profit or loss would be given by the difference between her disposal price V_t and the average price, $\bar{pb}^j$ or $\bar{ps}^j$, of acquiring her inventory. Player j's reward at time t is therefore the sum of (2) and (3)

$$J_t^j(x_t(a_t)) = Jq_t^j(x_t(a_t)) + Jp_t^j(x_t(a_t)) \tag{4}$$

Fills that are not netted enter into player j's inventory with the update $k_t^j = k_{t-1}^j + b_t^j - s_t^j$.

As future values of V_t are unknown, inventory size $|k_t^j|$ represents risk to future rewards (through (3)) as well as risk to a player's future ability to quote (through (1)). Informally, balancing inventory size and immediate profits from netted trades is the risk-reward tradeoff facing a player.

We model player j's objective as minimizing regret $R^j(t) = \min_{a \in A^j} \sum_{t'=1}^t J^j(x_{t'}^{-j}, x^j(a))$ - $\sum_{t'=1}^t J^j(x_{t'}^{-j}, x_{t'}^j(a_{t'}^j))$ where x^{-j} denote the quotes of all players except player j. A player could run a no-regret strategy with weights W^j that represent a mixed strategy over her action set A^j.

2.4 Allocation policy

Our allocation policy follows financial industry standards of *best execution* which intuitively means that the policy tries to execute as many units of i as possible for as low a cost to the auctioneer as possible, *up to the auctioneer's willingness to pay*. Ties are broken through pro-rata allocation. This policy is summarized in algorithms 1 and 2.

Algorithm 1: Forward auction allocation at t

Input: qb^j and pb^j for $j \in \mathcal{N}, PS, QS$
Aggregate qb^j into unique prices $p_1, \ldots, p_d$
Sort pairs (p_i, q_i) by decreasing price
Initialize $QR = QS$; $i = 1$
while $i \leq d$, $p_i \geq PS$ *and* $QR > 0$ **do**
  if $QR \geq q_i$ **then**
    $b^j = qb^j$ for all j, $pb^j = p_i$
  else
    b^j = pro-rata(qb^j) for all j, $pb^j = p_i$
  $QR = QR - \min\{q_i, QR\}$; $i = i + 1$

Algorithm 2: Reverse auction allocation at t

Input: qs^j and ps^j for $j \in \mathcal{N}, PB, QB$
Aggregate qs^j by unique prices $p_1, p_2, \ldots$
Sort pairs (p_i, q_i) by increasing price
Initialize $QR = QB$; $i = 1$
while $i \leq d$, $p_i \leq PB$ *and* $QR > 0$ **do**
  if $QR \geq q_i$ **then**
    $s^j = qs^j$ for all j, $ps^j = p_i$
  else
    s^j = pro-rata(qs^j) for all j, $ps^j = p_i$
  $QR = QR - \min\{q_i, QR\}$; $i = i + 1$

2.5 Efficiency and Sustainability

An auctioneer's objective is to maximize efficiency. This equates to minimizing the cost of buying in the reverse auction and maximizing the gain from selling in the forward auction. In a repeated game, we need to also consider efficiency in future time periods. From (4), we see that a player's reward is negative if Jp_t^j is negative and outweighs Jt_t^j. A rational player faced with a game where expected rewards are negative would choose not to play. In real life, a market maker who loses too much money may have no choice but to cease operations. If a large enough number of players exit the game, future efficiency levels could be compromised. Therefore, both efficiency and sustaining players are reasonable social objectives.

Motivated by the above discussion, we define three measures of efficiency. The percentage of the auctioneer's quantity executed during the game

$$E_1 = \frac{\sum_{t=1}^{T} \sum_{j=1}^{N} \left(b_t^j + s_t^j \right)}{\sum_{t=1}^{T} \left(QB_t + QS_t \right)}. \tag{5}$$

The percentage of time where some trading occurred

$$E_2 = \frac{1}{T} |\{t \in \{1, \ldots, T\} : \sum_{j=1}^{N} (b_t^j + s_t^j) > 0\}| \tag{6}$$

The negative of the average price per unit of i executed above and below V_t

$$E_3 = -\frac{\sum_{t=1}^{T} \sum_{s=1}^{d_t} |V_t - L_{t,s}^{price}| \cdot L_{t,s}^{size}}{\sum_{t=1}^{T} \sum_{s=1}^{d_t} L_{t,s}^{size}} \tag{7}$$

Efficiency is higher with high values of E_1, E_2 and E_3.

We define sustainability S to equal the expected number of players with total rewards above some minimum value ϵ.

$$S = \mathbb{E}[|\{j \in \mathcal{N} : \sum_{t=1}^{T} J_t^j(x_t) \geq \epsilon\}|] \tag{8}$$

Sustainability could mean that S remains above some number required for markets to function effectively. In general, a higher value of S corresponds to higher levels of sustainability.

3 Equilibrium analysis

We illustrate a player's risk-reward tradeoff when faced with different market environments and relate this tradeoff to the welfare goals of sustainability and efficiency. We use a simplified version of our game that consists of only a single time period and two players, each with only a choice of two actions, operating in a full information environment.

We vary the circumstances that players face solely through changing the auctioneer's objectives (QB, QS, PB, PS). Our analysis demonstrate that

1. Low auction quantities alone are not sufficient to incentivize players to quote competitively.
2. An auctioneer with low willingness to pay can induce players to quote competitively.
3. Unbalanced buy and sell auction quantities can induce players to quote competitively.
4. Some environments lead to games that resemble a prisoner's dilemma game where the equilibrium point is mutually disadvantageous to players. In some cases, this equates to market makers choosing actions that lead to large negative rewards for all players.

3.1 Definitions

Denote the set of decisions of all N players by $A = A^1 \times \cdots \times A^N \subset \mathbb{R}^N$. The game is denoted by $\Gamma(\mathcal{N}, A, \{J^j\}_{j\in\mathcal{N}})$. A strategy profile $a \in A$ is a Nash equilibrium for Γ if and only if

$$J^j(x^{-j}(a^{-j}), x^j(a^j)) \geq J^j(x^{-j}(a^{-j}), x^j(\tilde{a}^j)) \quad \text{for all } \tilde{a}^j \in A^j, j \in \mathcal{N} \tag{9}$$

A Nash equilibrium a is said to be *admissible* if there is no other Nash equilibrium a' such that

$$J^j(x(\tilde{a})) \geq J^j(x(a)) \text{ for all } j \in \mathcal{N} \tag{10}$$

with at least one inequality strict [5].

Because there is only one time period in our equilibrium analysis, the inventory component of a player's reward (3) is always zero. Thus, for this section, we add a per unit inventory penalty k_p^j for any unmatched fills. In other words, for our analysis below, we replace the reward function (4) by

$$J^j(x(a)) = Jq^j(x(a)) + Jk^j(x(a)) \text{ where } Jk^j(x(a)) = |b^j - s^j| \cdot k_p^j \tag{11}$$

3.2 Games

For simplicity, assume that the two players P_1 and P_2 are identical with action sets $A^1 = A^2 = \{0.01, 0.02\}$, quote sizes $q^1 = q^2 = 100$ and inventory limits $k_p^1 = k_p^2 = 0.04$. Let the fair value of i be $V = 80.0$ and $\epsilon = 0$ in the measure of sustainability (8). We consider four games that have outcomes with various degrees of efficiency and sustainability. Table 2 contains the reward matrices of these games. Note that while the values in the reward matrices allow us to compare outcomes, the absolute numbers themselves are not meaningful.

Game 1 We start with a game that is advantageous to the players. In our full information setting, the auctioneer's objectives are known to all. Thus, if the auctioneer has a high willingness to pay, i.e. PB is much higher than V and PS much lower than V, and if it has sizeable quantity to execute, i.e. QB and QS are large compared to $\sum_j qs^j$ and $\sum_j qb^j$, players would prefer to quote wide.

For concreteness, let $QB = 500, QS = 200, PB = 80.02, PS = 79.98$. Using algorithms 1 and 2, and the modified reward function (11), we calculate b^j, s^j and J^j for $j \in \{1, 2\}$ as shown in table 1. The Nash equilibrium $a = (0.02, 0.02)$ corresponds to rewards (4.0, 4.0) and the welfare measures (5)-(8) are $E_1 = 0.286, E_2 = 1.0, E_3 = -0.02$ and $S = 2$. The outcome of this game shows that when demand is price insensitive and large, market inefficiencies could occur and players could make large profits.

Game 2 Now suppose that the auctioneer has the same willingness to pay as in game 1 but only has a small quantity to execute. Let $QB = 100, QS = 98, PB = 80.02, PS = 79.98$. From the rewards matrix, there are two Nash equilibria, $a = (0.01, 0.01)$ and $a' = (0.02, 0.02)$, but only the latter is admissible. The welfare measures (5) to (8) are $E_1 = 1.0, E_2 = 1.0, E_3 = -0.02$ and $S = 2$.

The outcome of this game is slightly more efficient because the percentage of market demand fulfilled is higher (reflected in $E2$). The per unit cost, however, remains high and the auctioneer still pays a price that is $V \pm 0.02$. Thus, we conclude that lower values of QB and QS may be insufficient to incentivize players to quote tight.

Games 3 Compared to game 1, the auctioneer's buy and sell quantities remain high but its willingness to pay is low. Let $QB = 500, QS = 200, PB = 80.01, PS = 79.99$. The only Nash equilibrium is $a = (0.01, 0.01)$ with value (2.0, 2.0) and welfare measures are

B	Player 1	Player 2
(0.01, 0.01)	$b^1 = 100, s^1 = 100, J^1 = 2.00$	$b^2 = 100, s^2 = 100, J^2 = 2.00$
(0.01, 0.02)	$b^1 = 100, s^1 = 100, J^1 = 2.00$	$b^2 = 100, s^2 = 100, J^2 = 4.00$
(0.02, 0.01)	$b^1 = 100, s^1 = 100, J^1 = 4.00$	$b^2 = 100, s^2 = 100, J^2 = 2.00$
(0.02, 0.02)	$b^1 = 100, s^1 = 100, J^1 = 4.00$	$b^2 = 100, s^2 = 100, J^2 = 4.00$

Table 1: Game 1 rewards calculation.

Game 1

P_1 \ P_2	0.01	0.02
0.01	(2.00, 2.00)	(2.00, 4.00)
0.02	(4.00, 2.00)	**(4.00, 4.00)**

Game 2

P_1 \ P_2	0.01	0.02
0.01	(0.94, 0.94)	(1.88, 0.00)
0.02	(0.00, 1.88)	**(1.92, 1.92)**

Game 3

P_1 \ P_2	0.01	0.02
0.01	**(2.00, 2.00)**	(2.00, 0.00)
0.02	(0.00, 2.00)	(0.00, 0.00)

Game 4

P_1 \ P_2	0.01	0.02
0.01	**(-1.06, -1.06)**	(1.88, -4.0)
0.02	(-4.0, 1.88)	(-0.08, -0.08)

Table 2: Payoff matrices for games 1 to 4.

$E_1 = 0.286, E_2 = 1.0, E_3 = -0.01$ and $S = 2$. Compared to game 1, the auctioneer's unwillingness to pay is sufficient for the equilibrium point to shift to a more efficient point. In addition, large values of PB and QB mean that players do not hold inventory and are profitable. Game 3 is an example of an outcome that is both sustainable and efficient.

Game 4 Compared to game 2, the auctioneer's willingness to pay remains high but it now wants to buy an additional 100 units of i. This creates an imbalance between QB and QS. Let $QB = 200, QS = 98, PB = 80.02, PS = 79.98$. The equilibrium is $a = (0.01, 0.01)$ with value (-1.06, -1.06). The outcome of this game achieves higher efficiency with $E_1 = 1.0, E_2 = 1.0, E_3 = -0.01$ and $S = 2$. Both players, however, suffered a loss. Thus, $S = 0$ which makes this game unsustainable.

It is interesting to note that this game resembles a prisoner's dilemma game. The point $a = (0.02, 0.02)$, which could be interpreted as acting cooperatively, is not a Nash equilibrium. Player 1 is incentivize to undercut the Player 2's quote for a reward of 1.88. By the same logic, Player 2 is also incentivized to quote tight. Cooperation, however, would have resulted in both players achieving a better reward.

We could ask when such an imbalance would likely occur in real markets. One extreme example is a 'market meltdown' where economic uncertainty, or panic, results in a situation where there are only sellers. This is a high risk scenarios for a typical market maker because inventory accumulates quickly and its value also changes quickly. Intuitively, the player should charge more by choosing to quote wide. Game 4, however, suggests that competition between players may not allow her to do so.

4 Experiments

We build on our equilibrium analysis by expanding the game to more players, actions and time periods. We accomplish this through simulating a bandit environment where players employ a no-regret strategy to learn an optimal action. Corresponding to a 'bankruptcy' situation, we require a player to exit the game when her cumulative rewards fall below a predetermined, negative-valued threshold. The goal remains to examine the relationship between efficiency and sustainability.

Our main conclusions are:

1. When the auctioneer's willingness to pay is low, players learn to quote tighter and player rewards decrease.
2. For every unit of increase in the auctioneer's willingness to pay, players increase their quote width by only a percentage of that value. This means that the actual price paid by the auctioneer is generally smaller than its willingness to pay.
3. When markets are volatile, fewer players survive and those remaining learn to quote wider.
4. A form of unplanned coordination may arise where players 'share' fills across time periods.

4.1 Simulation settings and data

Simulation uses historical price data of SP500 E-mini futures, denoted ES, traded on the Chicago Mercantile Exchange's Globex trading platform. We chose the asset ES because there are many participants in its market and it produces large amounts of financial data.

The data we use is a time series of two types of market events: trades in ES and updates to the best quote of ES. Trades are recorded as a trade price and a trade quantity. Quote updates are recorded as a bid price, an offer price, a bid quantity and an offer quantity. A data point is recorded when either of the event occurs except when updates occur within 250 milliseconds. In those cases, updates are aggregated.

We use data to set the fair value V_t and the auctioneer's quantities QB_t and QS_t. We do not, however, use data to set PB_t and PS_t. Theoretically, PB_t and PS_t represent what the auctioneer *would have liked* to trade. Historical volume data reflects what was actually traded and thus corresponds more to the sum of market maker fills. Instead, we treat PB_t and PS_t as a parameter to control the auctioneer's willingness to pay.

The game runs across two separate time periods: the month of March 2020 and the month of July 2020. Data from March coincides with a sell-off caused by the start of a pandemic, and data from July reflects a relatively quiet summer period. For every level of willingness to pay, we run five games of weekly duration in both time periods and average their results. We limit ourselves to data generated during 'normal' market hours between 9:30 and 16:00 ET and aggregate data into one minute intervals. Thus, each weekly game occurs over $T = 1950$ time periods.

The game consists of $N = 10$ identical players. Every player j has 12 actions in their action set $A^j = \{0.25, 0.50, 0.75, \ldots, 3.0\}$, quote size $q^j = 1000$ and inventory limit $k^j_{max} = 5000$. All players start with zero reward and inventory. If, at any point in the game, the total reward of player j falls below -50,000, she is forced to liquidate her inventory and exit the game.

All players employ the EXP3 algorithm (Auer et al., 2002) with a learning rate $\eta = \sqrt{\log |A^j|/T}$ to optimize their mixed strategy. We chose the EXP3 algorithm because it has been shown to be a no-regret strategy [1]. We use (2) as a player's objective. This satisfies the assumption in EXP3 that rewards lie within [0,1]. Inventory considerations do not affect the objective in this game. But because (2) is always positive, the event of 'bankruptcy' is purely driven by changes in inventory value.

4.2 Simulation results

Figure 2 summarizes our results. The charts show the effect of the auctioneer's willingness to pay, in the x-axes of both charts, and of market volatility, corresponding to using March versus July data, across various measures of efficiency and sustainability.

The red lines labelled `action` on the left chart show the strategies learned by the players by the end of the game. `action` is the expected value of a player's action at T, $\langle W^j_T, A^j \rangle$, averaged across all surviving players. The results show that, regardless of willingness to pay, players generally quote wider in a volatile period than in a quiet period. Holding market volatility constant, players learn to quote wider as the willingness to pay increases.

The blue lines labelled `txnCost` on the left chart correspond to the average price that the auctioneer paid above or below fair value to trade. By definition, `txnCost` is smaller than willingness to pay. Less intuitively, the results also show that when willingness to pay is high, `txnCost` can be significantly below both willingness to pay and `action`. The latter effect suggests that with enough

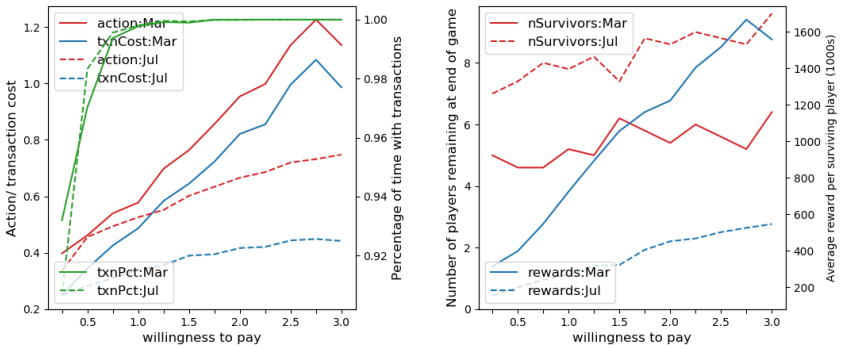

Figure 2: Simulation results varying willingness to pay and market volatility. The left chart plots measures of efficiency. The right chart plots measures of sustainability.

players playing a mixed strategy on their quote width, players end up taking turns quoting tight and 'share' fills across time periods. One interpretation of this is that a level of unplanned cooperation between players emerged by the end of the game.

The green lines labelled `txnPct` represent the percentage of time where some trading occur. Unsurprisingly, `txnPct` increases with increasing willingness to pay. The results also suggest that this effect is independent of market volatility.

On the right-hand side chart, the red lines labelled `nSurvivors` plot the number of players (out of 10) left in the game at time T. Games in March have fewer survivors than in July. Surprisingly, `nSurvivors` seems independent of willingness to pay.

Finally, the blue lines on the right chart labelled `rewards` plot the average cumulative reward of surviving players at time T. This value increases with higher values of willingness to pay and the absolute level of rewards is higher in the higher volatility environment. The primary reason for this is that trading volumes are also higher in March than in July. The loss of any non-survivor is roughly equal to the bankruptcy threshold of -50,000. This is much smaller than the average reward of surviving players and suggests that survivorship bias does not play a significant role.

5 Discussion and future direction

Our paper introduces a model of financial markets by casting market makers as players and non-market makers collectively as the auctioneer in a repeated game. We apply our model to examine the relationship between market efficiency and sustainability. Our equilibrium analysis suggests that outcomes depend on buy-sell imbalances in market demand and on the market's willingness to pay. Our simulation results expand on relationships between willingness to pay, market volatility, players' learned quote spreads, survivability and the cost to transact. The results also suggest that some level of cooperation could emerge without players explicitly seeking to cooperate.

Improvements could come from using more realistic models. For example, we could substitute reinforcement learning for online learning [2], change the allocation policy from pro-rata tie breaks to price-time allocations, or allow players to quote asymmetrically based on inventory levels.

The equilibrium analysis of game 4 shows that players could wind up in a version of prisoner's dilemma. A natural extension is to combine welfare measures (5)-(8) into a single objective and to consider solutions in terms of the price of anarchy or the price of stability [10]. This would bring us closer to mechanism design.

Our simulation suggests that there is value in allowing players to enter and exit the game. Existing literature on games with dynamic populations [7] may provide theoretical insights to our simulation.

Finally, a drawback in our analysis comes from either assuming a full information environment or a full bandit environment. Neither assumptions are realistic. Market makers, through experience, make informed guesses on the circumstances of other market makers and on market conditions. An improvement would be to move to a partial bandit environment [6].

References

[1] Auer, P., Cesa-Bianchi, N., Freund, Y. & Schapire, R.E. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48-77, 2002.

[2] Dworkin, L., Kearns, M. & Nevmyvaka, Y. Pursuit-evasion without regret, with an application to trading. *International Conference on Machine Learning*: 1521-1529, 2014.

[3] Ganesh, S. et al. Reinforcement learning for market making in a multi-agent dealer market. In *ArXiv:1911.05892*, 2019.

[4] Gueant, O. & Manziuk, I. Deep reinformcement learning for market making in corporate bonds: beating the course of dimensionality. In *ArXiv:1910.13205*, 2019.

[5] Hespanha, J. An introductory course in noncooperative game theory. 2007.

[6] Karaca, O., Sessa, P.G., Leidi, A. & Kamgarpour, M. No-regret learning from partially observed data in repeated auctions. In *arXiv:1912.09905*, 2019.

[7] Lykouris, T., Syrgkanis, V. & Tardos, É. Learning and efficiency in games with dynamic population. In *27th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, 120-129, 2016.

[8] Othman, A., Pennock, D., Reeves, D. & Sandholm, T. A practical liquidity-sensitive automated market maker. *ACM Transactions on Economics and Computation (TEAC)*, 2013.

[9] Patel, Y. Optimizing market making using multi-agent reinforcement learning. *In ArXiv:1812.10252*, 2019.

[10] Roughgarden, T. & Tardos, É. Introduction to the inefficiency of equilibria. In *Algorithmic Game Theory*, 443-460, 2007.