Algorithmic Pricing and Liquidity in Securities Markets*

Jean-Edouard Colliard, Thierry Foucault, and Stefano Lovo

March 17, 2023

Abstract

We let "Algorithmic Market-Makers" (AMs), using Q-learning algorithms, choose prices for a risky asset when their clients are privately informed about the asset payoff. We find that AMs learn to cope with adverse selection and to update their prices after observing trades, as predicted by economic theory. However, in contrast to theory, AMs charge a mark-up over the competitive price, which declines with the number of AMs. Interestingly, markups tend to decrease with AMs' exposure to adverse selection. Accordingly, the sensitivity of quotes to trades is stronger than that predicted by theory and AMs' quotes become less competitive over time as asymmetric information declines.

Keywords: Algorithmic pricing, Market Making, Adverse Selection, Market Power, Reinforcement learning.

JEL classification: D43, G10, G14.

*Correspondence: colliard@hec.fr, foucault@hec.fr, lovo@hec.fr. All authors are at HEC Paris, Department of Finance, 1 rue de la Libération, 78351 Jouy-en-Josas, France. We are grateful to participants in "The Microstructure Exchange", the Microstructure Asia Pacific Online Seminar, and seminars at the University of Copenhagen, University Paris 1, and HEC Paris for helpful comments and suggestions. We thank Olena Bogdan, Amine Chiboub, Chhavi Rastogi and Andrea Ricciardi for excellent research assistance. This work was supported by the French National Research Agency (F-STAR ANR-17-CE26-0007-01, ANR EFAR AAP Tremplin-ERC (7) 2019), the Investissements d'Avenir Labex (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047), the Chair ACPR/Risk Foundation "Regulation and Systemic Risk", the Natixis Chair "Business Analytics for Future Banking".

Introduction

Firms (e.g., retailers, airlines, hotels, energy providers etc.) increasingly rely on algorithms to set the price of their products.[1] This evolution reflects efficiency and predictive gains of artificial intelligence but it generates new concerns, in particular about price discrimination and tacit collusion among algorithms (see MacKay and Weinstein (2022), CMA (2018), OECD (2017)).[2] Surprisingly, this worry has not been expressed for market makers in securities markets even though proprietary trading firms (market makers such as Citadel, Virtu, Jane Street, etc.) began using pricing algorithms at least two decades ago and now dominate liquidity provision in exchanges.[3]

Is this lack of concern justified? Is tacit collusion among pricing algorithms more difficult in securities markets? To study this question, we consider a framework in which "algorithmic market makers" (AMs) compete in prices and are at risk of trading with better informed investors. Our setting is, by design, very similar to standard models of market making with asymmetric information (in the spirit Glosten and Milgrom (1985) and Kyle (1985)). However, in contrast to these models, we assume that quotes are posted by AMs that set their quotes using Q-learning algorithms, a special type of reinforcement learning algorithm (often mentioned as the type of algorithms used for pricing decisions; see CMA (2018)). We focus on whether Q-learning algorithms cope with adverse selection, learn to account for the information contained in trades in choosing prices and whether their prices are competitive. To our knowledge, our paper is the first to analyze how Q-learning algorithms behave in the presence of asymmetric information (an important feature of trading in securities markets).

In our framework, AMs simultaneously post offers in response to clients' requests to buy one share of a risky asset. Clients' valuation for the asset is the sum of the payoff of the asset (a common value component) and a component specific to each client (a private valuation component). Clients

[1] For instance, Chen *et al.* (2016) find that more than 500 Amazon third-party sellers on Amazon marketplace were using algorithms to price their products.

[2] For instance, MacKay and Weinstein (2022) write: *"The explosion in the use of pricing algorithms over the past decade has sparked concerns about the effect on competition and consumers [...]."*.

[3] These firms are often referred to as "high-frequency market makers" because their algorithms (and hardware equipments) generate very frequent new orders (quotes, cancellations etc.). Menkveld (2013) finds that, in 2007-2008, a single high-frequency market maker accounts for about 15% of total trading volume in Dutch stocks (and more than 60% on one of the trading platforms for these stocks). Brogaard *et al.* (2015) find that fast traders on the Stockholm Stock Exchange are primarily market makers, who account for 83% of all limit orders on this exchange (and 44% of trading volume).

arrive sequentially and each one trades with the dealers posting the best offer in response to her request, provided that this offer is less than her valuation.[4] As clients' demand for the asset increases with the common value, dealers are exposed to adverse selection (they are more likely to sell the asset when its payoff is high than when it is low). In the baseline case, a new realization of the asset payoff is drawn after each client's arrival.

AMs behave as follows. In each trading round, each AM starts with an assessment of the expected profit associated with each possible price, picks a price (on a grid) based on this assessment, and updates its assessment of the expected profit associated with this price by taking a weighted average (with pre-specified weights) of its realized profit at the end of the trading round and its prior assessment of its expected profit. The assessment of the expected profit associated with other possible prices is unchanged. In each round t, the AM picks the price that generates the largest expected profit according to its assessment with a given probability of "exploitation". Otherwise, it "explores" by picking at random (with equal probability for each possible price) another price. Exploration enables the AM to receive feedback about the profit generated by a price and therefore to "learn" the expected profit associated with this price.

This iterative process is repeated over a large number of "episodes" (each made of one trading round), which collectively constitute one "experiment". In a given experiment, the set of parameters (e.g., the number of AMs, the distribution of the asset payoff, and the distribution of each client's private valuations) is constant across episodes and forms the "environment". For each environment considered in our analysis (i.e., for a fixed set of parameters), we run $10,000$ experiments, each made of $200,000$ episodes. In each experiment we record each AM's quote, the transaction price, and the trading volume (0, or 1) in each episode.

In early episodes of a given experiment, AMs "learn" the expected profit associated with each possible price, which leads to significant volatility in prices. After a large number of episodes, their pricing strategy eventually "converges" in most experiments, in the sense that AMs keep playing the same price over a large number of episodes. However, this "long run" price can vary

[4] In electronic securities markets (e.g., electronic limit order books markets used in most of the major stock markets in the world), market makers compete in prices ("à la Bertrand") with no room for product differentiation. Price priority is strictly enforced, which guarantees that clients' orders are filled at the best price, as assumed in our analysis.

from one experiment to another. Thus, our analysis focuses on the *empirical distribution* of final outcomes (e.g., transaction prices and dealers' profits) across experiments (holding the environment constant). We study how this distribution varies with parameters of the environment (in particular the intensity of adverse selection) and we systematically compare final outcomes to those predicted by economic theory. When there are multiple dealers, the outcomes predicted by theory (e.g., transaction prices and profits) are those corresponding to the Bertrand-Nash equilibrium of the environment considered in our simulations (accounting for the fact that market makers must post their quotes on a grid).[5] When there is a single dealer, predicted outcomes are those corresponding to the equilibrium of the environment in which the dealer behaves as a monopolist.

We observe several interesting regularities. First, when there is a single AM, it does not necessarily learn the monopoly price and its average quote is smaller than the monopoly price. In contrast, when there are two AMs, they charge a price above the competitive price on average (even the smallest price in the distribution of observed prices is well above the competitive price). This markup decreases with the number of AMs and becomes close to zero only with 10 AMs.

Second, in all environments, AMs learn how *not* to be adversely selected. That is, they charge prices that (more than) cover adverse selection costs. However, when their exposure to adverse selection increases (either because the volatility of the asset payoff increases or the variance of clients' private valuations decreases), AMs tend to choose prices that are *more* competitive (in particular, their realized bid-ask spreads are smaller on average). This is particularly striking when the variance of clients' private valuations increases. In this case, AMs' offers (and therefore transaction prices) increase, even though the competitive (Nash-Bertrand) price decreases because adverse selection costs decline. Overall these findings suggest that adverse selection interacts with the way Q-learning algorithms learn in non-intuitive ways.

In existing models (Kyle (1985) or Glosten and Milgrom (1985)), market makers are assumed to learn the asset payoff from the trading history (the "order flow") in a Bayesian way. For this reason, holding the asset payoff constant, these models predict that dealers eventually discover asset payoffs. For instance, dealers' pricing errors (the average squared difference between the asset payoff

[5] In particular, as is well-known, price discreteness can generate multiple Bertrand-Nash equilibria. We take this into account in our analysis.

and the transaction price) decrease over time (trades) on average (see, for instance, Glosten and Milgrom (1985)). To study whether AMs can also discover asset values, we extend our baseline setting to allow for two trading rounds per episode. We find that AMs behave qualitatively like a Bayesian learner would. That is, after a buy (no trade), they increase (reduce) their offer in the second trading round. Thus, price discovery takes place, even though AMs have no knowledge of the data generating process and their learning process is not Bayesian. However, observing the outcome of the first period brings new information to the algorithms, which then face less adverse selection in the second period. As adverse selection curbs algorithms' rent-seeking behavior, prices become less competitive on average in the second period. Moreover, this effect is stronger after observing a trade than after observing no trade in the first period. In this sense, AMs seem to overreact to trades relative to a Bayesian learner.

In summary, our findings have two main implications. First, algorithmic market makers settle on non competitive prices, even though they operate in an environment where economic theory predicts competitive outcomes. This echoes findings in Hendershott *et al.* (2011) and Brogaard and Garriott (2019). The former find empirically that algorithmic trading (AT) *increases* dealers' expected profits net of adverse selection costs (realized bid-ask spreads). Commenting on this result, they write (on p.4): *"This is surprising because we initially expected that if AT improved liquidity, the mechanism would be competition between liquidity providers."* Brogaard and Garriott (2019) study the effect of high frequency market makers' entry on one trading platform for Canadian stocks. They find that this entry triggers a decrease in bid-ask spreads. However, the entry of two competitors is not sufficient to obtain the competitive outcome, in contrast to what standard models of market making predicts. This pattern is exactly what we find when we compare average bid-ask spreads across environments that only differ by the number of AMs. Our second main implication is that adverse selection induces AMs to post quotes that are *more* competitive. In a cross-section of assets, this means that *realized bid-ask spreads* for AMs (a measure of dealers' expected profits net of adverse selection costs) should be smaller in assets that are more exposed to informed trading. For instance, they should be smaller for stocks than for Treasuries (high frequency market makers are active in both types of assets), even though adverse selection costs are larger for stocks. This

implication also holds dynamically: as adverse selection is resolved over time we expect AMs to quote less competitively, contrary to the predictions of standard asymmetric information models.

It is worth stressing that we do not claim that market-making algorithms necessarily behave as our AMs.[6] This does not mean however that the patterns uncovered by our analysis are unlikely to hold in reality.[7] Our approach is to make stylized assumptions on pricing algorithms to develop predictions about their effects on securities markets. In particular, our assumptions capture that (some) algorithms used in practice rely on experimentation ("trial and error") but eventually experiment less and less, as experimentation is costly. We believe these properties are reasonable in a financial context. As explained above, this approach delivers predictions that are quite different from those of standard economic models in the same environment. To decide which approach has more explanatory power, the next step (which is beyond the scope of this paper) would be to test these predictions empirically.

The rest of the paper proceeds as follows. In the next section, we position our contribution in the literature. Section 2, we present the economic environment analyzed in our paper. In Section 3, we study the case in which each episode has a single trading round. In this case, our analysis focuses on how AMs' behavior differ from two benchmarks: (a) competitive behavior (Nash-Bertrand equilibrium) and (b) monopolistic behavior (monopoly prices). In Section 4, we study whether AMs can discover asset fair values by considering the case in which each episode has two trading rounds. Section 5 concludes.

1 Contribution to the literature

Our paper is related to the emerging literature on algorithmic pricing and the possibility for algorithms to sustain non competitive outcomes. Calvano *et al.* (2020) show that Q-learning algorithms

[6] There is not much guidance on the actual design of market making algorithms in reality because market making firms do not disclose information on their algorithms. Securities trading firms strongly push back a regulator's attempt to require disclosure of their computer codes for surveillance purpose. See *"US regulator declares 'dead' moves to seize HFT code"*, Financial Times, October 14, 2017. In the EU, proprietary trading firms must make sure that they take steps to insure that their algorithms will not lead to disordely markets. However, they do not have to disclose their algorithms to regulators.

[7] The behavior of market makers in existing economic models is also highly stylised and, in contrast to our approach here, they are assumed to have a complete knowledge of their environment (e.g., the distribution of asset payoffs, traders' valuations etc.). Yet, these models have explanatory power for the behavior of security prices at high frequency (see, for instance, Glosten and Harris (1988) and the subsequent literature using price impact regressions).

can learn dynamic collusive strategies in a repeated differentiated Bertrand game. Asker *et al.* (2021) and Abada *et al.* (2022) show that supra competitive prices can be reached in this type of environment even if dynamic strategies are ruled out, through what Abada *et al.* (2022) call "collusion by mistake".[8] Cartea *et al.* (2022a) and Cartea *et al.* (2022b) study different families of reinforcement learning algorithms and develop new methods to study which ones may lead to non Nash behavior in a market-making environment.[9] Banchio and Skrzypacz (2022) find that Q-learning algorithms post less competitive bids in first price auctions than in second price auctions. In contrast to our setting, bidders and sellers have a fixed valuation for the auctioned good and bidders are not exposed to adverse selection in their setting (they consider private value auctions). In sum, in line with other papers, we find that pricing algorithms relying on Q-learning can lead to non competitive outcomes even when dynamic strategies are ruled out and when price setters compete in prices. However, new to the literature, we find that adverse selection tends to mitigate this issue.[10] Moreover, to our knowledge, we are the first to study price discovery with Q-learning algorithms (an issue specific to securities markets).

Our paper also contributes to the literature on algorithmic trading in securities markets. The theoretical literature on this issue (e.g., Biais *et al.* (2015), Budish *et al.* (2015), Menkveld and Zoican (2017), Baldauf and Mollner (2020), etc.) has mainly focused on how the increase in the speed with which algorithms can respond to information increases or reduces liquidity suppliers' exposure to adverse selection, using traditional workhorses models (Glosten and Milgrom (1985) or Kyle (1985)). Yet, O'Hara (2015) calls for the development of new methodologies to study the effects of algorithms in financial markets, writing that as a result of algorithmic trading: *"the data that emerge from the trading process are consequently altered [...] For microstructure researchers, I believe these changes call for a new research agenda, one that recognizes how the learning models used in the past are lacking [...]."* Our paper responds to this call. Instead of modeling algorithmic traders as Bayesian learners, with an omniscient knowledge of the environment in which they operate, we

[8] This idea is in line with an earlier literature in machine learning showing that games between Q-learning algorithms do not necessarily converge to a Nash equilibrium (Wunder *et al.*, 2010).

[9] In particular, Cartea *et al.* (2022b) show that using a finer pricing grid (a lower "tick size") reduces the scope for collusion.

[10] Another rather unique feature of our setting is that, in our setting, the demand faced by pricing algorithms is stochastic. See also Hansen *et al.* (2021) and Cartea *et al.* (2022b) other settings in which selling algorithms face a stochastic demand elasticity, but without adverse selection.

model them as Q-learning algorithms. These algorithms learn by trial and error with almost no prior knowledge of the environment, which represents the polar opposite of standard Bayesian learning. Moreover, Q-learning is relatively simple and transparent, which makes it a good candidate for a workhorse model of algorithmic interaction, much like the Glosten-Milgrom environment is a workhorse model of market-making. This approach generates strikingly different implications for those of canonical Bayesian-learning models. In particular, price competition does not guarantee a competitive outcome and, maybe even more surprisingly, increased adverse selection can reduce dealers' rents.

2 The economic environment

In this section, we provide a general description of the economic environment considered in our experiments (Section 3.3). We consider the market for one risky asset with $t = 1, 2, ..., T$ episodes (one can think of them as "trading days"). Quotes in this market are posted by N dealers who trade with clients. Each episode has $\bar{\tau}$ trading rounds and the asset payoff $\tilde{v}$ is realized at the end of the last trading round in a given episode. This payoff has a binary distribution, $\tilde{v} \in \{v_L, v_H\}$, with $v_L \leq v_H$ and $\mu := \Pr(\tilde{v} = v_H) = \frac{1}{2}$. We denote $\Delta v = v_H - v_L$. Realizations of the asset payoffs are independent across episodes. In the rest of this section, we describe traders' actions and realized profits in a given episode.

In each trading round τ, a new trader (the "client") arrives to buy one share of the asset. The buyer's valuation for the asset is $v_\tau^C = \tilde{v} + \tilde{L}_\tau$, where $\tilde{L}_\tau$ is i.i.d across trading rounds. Clients' private valuations are assumed to be normally distributed with mean zero and variance σ^2. The buyer observes her valuation for the asset and requests quotes from the dealers, who simultaneously respond by posting their offers $a(\tau) = \{a_{1\tau}, ..., a_{N\tau}\}$. The ask price $a_{n\tau}$ is the price at which dealer n is ready to sell at most one share in trading round τ. We denote by (i) $a_\tau^{min} = \min_n\{a_{n\tau}\}$ the smallest ask price, (ii) $\mathcal{D}_\tau$ the set of dealers posting this price and (iii) z_τ the number of dealers in $\mathcal{D}_\tau$. The client buys the asset if the best offer is less than her valuation for the asset ($a_\tau^{min} \leq v_\tau^C$) and, in this case, she splits her demand among the z_τ dealers posting this price. Otherwise she does not trade.

Let denote by $V(a(\tau), \tilde{L}_\tau, \tilde{v})$ the volume of trade in round τ. It equals 1 if the client buys the asset, i.e., if the client valuation $\tilde{v} + \tilde{L}_\tau$ is not smaller than the lowest price a_τ^{min}, and it is zero otherwise. Let denote by $Z(a_{n\tau}, a(\tau))$ the fraction of the τ^{th} round trade executed by dealer n, that is, $Z(a_{n\tau}, a(\tau)) = \frac{1}{z_\tau}$ if $a_{n\tau} = a_\tau^{min}$ and is zero otherwise. In trading round τ, dealer n's realized trading volume is:

$$I(a_{n,\tau}, a(\tau), \tilde{L}_\tau, \tilde{v}) := V(a(\tau), \tilde{L}_\tau, \tilde{v}) Z(a_{n\tau}, a(\tau)), \tag{1}$$

and his realized profit is:

$$\Pi(a_{n\tau}, a(\tau), \tilde{L}_\tau, \tilde{v}) := I(a_{n,\tau}, a(\tau), \tilde{L}_\tau, \tilde{v})(a_\tau^{min} - \tilde{v}), \tag{2}$$

Hence, dealer n's total realized profit in a given episode is:

$$\sum_{\tau=1}^{\overline{\tau}} \Pi(a_{n\tau}, a(\tau), \tilde{L}_\tau, \tilde{v}). \tag{3}$$

In our setting, holding prices constant, dealers are more likely to sell the asset when the asset payoff is high than when the asset payoff is low. Indeed, conditional on a realization of v, the likelihood that a trade occurs in trading round τ is:

$$D(a_\tau^{min}, v) := Pr(a_\tau^{min} \leq v + \tilde{L}_\tau) = 1 - G(a_\tau^{min} - v), \tag{4}$$

which increases with v as $D(a_\tau^{min}, v_H) > D(a_\tau^{min}, v_L)$. Thus, dealers are exposed to adverse selection: they are more likely to sell the asset when its payoff is high than when it is low.

Finally, we denote by $\overline{\Pi}(a, \mu_\tau) = N\mathbb{E}[\Pi(a, a, \tilde{L}, \tilde{v})]$ the dealers' expected aggregate profit when they all post the same price a, and attach probability μ_τ to the event $\tilde{v} = v_H$. This gives

$$\bar{\Pi}(a, \mu_\tau) := \mu_\tau D(a, v_H)(a - v_H) + (1 - \mu_\tau) D(a, v_L)(a - v_L) \tag{5}$$

In the rest of the paper, we study how AMs using Q-learning algorithms set their prices in such an environment. We consider two cases. In the first case, analyzed in Section 3, we consider an

environment in which $\overline{\tau} = 1$ (a single trading round per episode). Our focus in this case is on whether and how outcomes when prices are set by AMs differ from those obtained in two benchmarks: (i) the Nash-Bertrand equilibrium with multiple dealers (the competitive case) and (ii) the case in which dealers set their price to maximize their aggregate expected profit (the monopoly case). This comparison will help us to analyze how AMs exert market power and cope with adverse selection relative to rational Bayesian dealers. In the second case, analyzed in Section 4, we consider a dynamic environment in which $\overline{\tau} = 2$ (two trading rounds per episode) and we focus on price discovery (i.e., on how AMs adjust their quotes over time).

3 The Static Case ($\bar{\tau} = 1$)

In this section, we compare the pricing policies chosen by AMs using a Q-learning algorithm to equilibrium outcomes predicted by standard economic analysis in the environment described in Section 2 when there is a single trading round per episode. We refer to this case as the static case since, when we solve for dealers' equilibrium pricing policies in this case, they behave as if they were facing a static one-shot problem. We proceed in three steps. First, in Section 3.1, we derive the equilibrium outcomes in two benchmark cases (the monopolist and competitive cases). Then, in Section 3.2, we describe the Q-learning algorithms used by AMs to choose their pricing policy when $\bar{\tau} = 1$. Third, in Section 3.3, we compare the pricing policies chosen by AMs to those obtained in the benchmarks.

3.1 Benchmarks

Monopolist Case. In the monopolist case, in each episode, each dealer chooses her price, denoted a^m, to maximize dealers' aggregate expected profit. Recalling that $Pr(\tilde{v} = v_H) = \mu = 1/2$, a^m solves:

$$a^m \in arg\max_a \bar{\Pi}\left(a, \frac{1}{2}\right). \tag{6}$$

Economic theory predicts that this price should be the equilibrium outcome when $N = 1$.

Competitive Case. In the competitive case, dealers choose a price a^c such that each dealer's

expected profit is nil. That is,

$$a^c \ s.t. \ \bar{\Pi}\left(a^c, \frac{1}{2}\right) = 0. \tag{7}$$

When the set of prices is continuous, a^c is the Bertrand-Nash equilibrium of the game played by dealers in each trading round. This is the outcome predicted by economic theory when $N \geq 2$.

We explain how to obtain a^c and a^m in Appendices A.4 and A.3 and we find (numerically) that (i) the competitive price, a^c, increases with Δv and decreases with σ while (ii) the monopoly price increases with both Δv and σ. We provide a numerical example in Table 1 where we report a^m and a^c when $\Delta v = 4$ and $\sigma = 5$ ($v_H = 4$ and $v_L = 0$, so that $\mathbb{E}(\tilde{v}) = 2$), the baseline values of the parameters in our experiments. The table also reports the expected half-quoted spread, $\mathbb{E}(a - \mathbb{E}(\tilde{v})) = a - \mathbb{E}(\tilde{v})$, (the difference between the ask price posted by each dealer and the unconditional expected payoff of the asset) and the expected half realized spread, $\mathbb{E}(a - \tilde{v} \mid V = 1)$. In contrast to the average half quoted spread, the expected half realized spread measures the expected profit of a dealer conditional on a trade by the client. As this trade is more likely when v is high, this measure accounts for the adverse selection cost borne by the dealer. In fact the difference between average half quoted spread and average half realized spread is often used as a measure of adverse selection costs in empirical studies.[11] Last observe that the total expected profit of a dealer is the expected half realized spread times the probability that the client trades ($\bar{\Pi}(a, \mu) = \Pr(V = 1)\mathbb{E}(a - \tilde{v} \mid V = 1)$).

[Insert Table 1 about here]

When the dispersion of clients' private valuations (σ) increases or the volatility of the asset payoff (Δv) decreases, dealers' ask prices become lower in the competitive case because dealers' adverse selection costs decline. In contrast, when the dispersion of clients' private valuations increases, the monopolist offer becomes larger, despite the fact that their adverse selection costs decline. This reflects an increase in dealers' rents, as shown by the increase in the expected realized spread. The reason is that as σ increases, clients' demand becomes more inelastic, which as usual enables a monopolistic dealer to extract larger rents. In contrast, when Δv increases, the client's demand becomes more elastic and the adverse selection cost increases. As a result, the monopolist dealer

[11] See Foucault *et al.* (2013), ch. 2, for a description of various measures of bid-ask spreads in securities markets and their interpretation.

charges a larger price but she obtains smaller rents (the realized bid-ask spread declines).

3.2 Q-Learning Algorithms

3.2.1 Description of the Algorithms

We now describe the functioning of Q-learning algorithms in the environment described in Section 2 when $\bar{\tau} = 1$.[12] We use the same notations as in Section 2, unless otherwise stated. In contrast to the benchmark case, we restrict AMs to choose their quotes in a discrete and finite action set $\mathcal{A} = \{a_1, a_2 ... a_M\}$, where each a_m is a possible ask price.[13]

To each dealer n and episode t, we associate a so-called *Q-Matrix* $\mathbf{Q}_{n,t} \in \mathbb{R}^{M \times 1}$. In this section, $\mathbf{Q}_{n,t}$ is simply a column vector of size M. The m-th entry of the matrix is denoted $q_{m,n,t}$ where $q_{m,n,t}$ represents the estimate in episode t of the payoff that AM n expects from playing price a_m. The Q-learning algorithm is meant to refine the payoff estimates in $\mathbf{Q}_{n,t}$ over time, and to end up playing the action associated with the highest estimate.

More formally, the algorithms (AMs) play the game according to the following process. We first initialize the matrices $\mathbf{Q}_{n,0}$ with random values: Each $q_{m,n,0}$ for $1 \leq m \leq M$ and $1 \leq n \leq N$ is i.i.d. and follows a uniform distribution over $[\underline{q}, \overline{q}]$. Then, in each episode t, we do the following:

1. For each dealer n, we define $m^*_{n,t} = arg\max_m q_{m,n,t-1}$ the index associated with the highest value in matrix $\mathbf{Q}_{n,t-1}$, and we denote $a^*_{n,t} = a_{m^*_{n,t}}$ the *greedy price* of AM n in episode t. This is the price that seems to maximize the AM's static profit, according to the estimates available in episode t.

2. For each dealer n, with probability $\epsilon_t = e^{-\beta t}$ the AM "explores" by playing $a_{n,t} = a_{\tilde{m}_{n,t}}$, where $\beta > 0$ and $\tilde{m}_{n,t}$ is a random integer between 1 and M, all values being equiprobable. The price $a_{\tilde{m}_{n,t}}$ is thus a price taken randomly in $\mathcal{A}$. With probability $1 - \epsilon_t$, the dealer "exploits" and plays $a_{n,t} = a^*_{n,t}$, the greedy price. The random draws leading to exploring or exploiting are i.i.d. across all dealers in a given episode.

[12] See Calvano *et al.* (2020) for an introduction to Q-learning algorithms in the more complex case of infinite horizon problems. See also Sutton and Barto (2018) for an introductory textbook on this topic.

[13] This constraint is necessary because the algorithm must evaluate the payoff associated with each possible price.

3. We compute $a_t^{min} = \min_n a_{n,t}$ the *best ask* in episode t, z_t the number of AMs with $a_{n,t} = a_t^{min}$, and draw $\tilde{v}_t$ and $\tilde{L}_t$. Each dealer n then receives a profit equal to $\pi_{n,t} = \Pi(a_{n,t}, a(t), \tilde{L}_t, \tilde{v}_t)$, as given by (2).[14]

4. We update the Q-matrix of each dealer as follows, with $0 < \alpha < 1$:

$$\forall 1 \leq n \leq N, q_{m,n,t} = \begin{cases} \alpha\pi_{n,t} + (1-\alpha)q_{m,n,t-1} & \text{if } a_{n,t} = a_m \\ q_{m,n,t} & \text{if } a_{n,t} \neq a_m \end{cases} \tag{8}$$

5. We then repeat starting from stage 1, until the last episode T.

Intuitively, each Q-learning algorithm alternates between experimenting random prices, and playing the price that seems to lead to the highest payoff based on past plays. As the number of past episodes grows, information accumulates and there should be less value in experimenting. For this reason, the probability of experimenting decays over time (here exponentially, at rate β). This important parameter of the algorithm governs the trade-off between experimenting and exploiting. A second trade-off is how much should one react to one particular observation $\pi_{n,t}$, knowing that payoffs are stochastic. This is governed by the parameter α: if α is large the algorithm reacts quickly to new observations, but the estimates generated in the Q-matrix are unstable (consider the extreme case $\alpha \to 1$). Conversely, if α is small the estimates are stable but it will take a lot of experimentation to move the values of the Q-matrix towards accurate estimates of the expected payoffs associated with each price.

3.2.2 Convergence

There are many variants of the Q-learning algorithm, with different specifications for the experimentation probability ϵ_t and the updating rule (8). The one described in the previous section is common in practical applications and is also the one used in recent papers in the economic literature (e.g., Calvano *et al.* (2020)). We choose it for comparability with prior literature. This version of Q-learning does not satisfy the assumptions given in, e.g., Watkins and Dayan (1992),

[14] We index all variables by the episode counter and omits the trading round index, τ within an episode since $\tau = 1$ in each episode here.

Jaakkola *et al.* (1994), or Tsitsiklis (1994) to guarantee convergence. In fact, given the design of these algorithms and the environment in which they operate, Lemma 1 below shows that no matter t (that is, even when T becomes very large), with a probability that is bounded away from 0, the Q-matrix changes by an amount that is bounded away from 0. Thus, entries in the Q-matrix never converge.

Lemma 1. (Impossibility of convergence of the Q-matrix) *For any given t and $a_m \in \mathcal{A}$, if $a_{n,t} = a_m = a_t^{\min}$, then ,*

$$\Pr\left(|q_{m,n,t} - q_{m,n,t+1}| \geq \Delta_m^*\right) \geq P_m^*,$$

where

$$\Delta_m^* := \frac{\alpha}{2}\left(v_H - v_L + \left|a_m - \frac{v_H - v_L}{2}\right|\right),$$

and

$$P_m^* := \min\left\{\frac{1}{2N}D(a_m, v_L), 1 - \frac{1}{2}(D(a_m, v_L) + D(a_m, v_H))\right\}$$

For instance, consider the case $N = 1$ and suppose that the AM plays for T consecutive periods the same price a_m. Then, as T goes to infinity, the value of the Q-matrix $q_{m,t}$ does not converge in probability to $\overline{\Pi}(a_m, \mu)$, that is the actual monopolist dealer's expected profit when she sets a price of a_m (the metrics a monopolist would use to set his price in economic theory). Because the Q-matrix does not converge, the price that maximizes the Q-matrix needs not stay the same and will vary with probability 1 after sufficiently many episodes. Thus, one cannot expect the greedy price to remain stable, even when T becomes very large. This also implies that the greedy price observed in the final episode will vary across experiments.

In actual simulations, when α is small, most experiments give the impression of "converging" in the sense that, after a sufficiently large number of episodes, the price chosen by each AM stays constant for many periods (this is because Δ_m^* is linear in α). This is the meaning of "convergence" in many papers in the literature (e.g., Calvano *et al.* (2020)). Thus, following the literature, we say that an experiment has "converged" if all algorithms' actions have been constant over the last κT periods (e.g., $\kappa = 0.05$). Moreover, to describe the outcome of the interaction between AMs

we look at the distribution of prices after a large number T of episodes and across a large number K of experiments, focusing in particular on the mean of the distribution. When needed, we use a superscript k to denote the outcome of the k^{th} experiment. For instance, $a_{n,t}^{*k}$ is the greedy price of dealer n in episode t of experiment k.

Keeping these observations in mind, we set the parameters of our baseline simulations as follows. The parameters of the economic environment are the same as in Table 1: $\Delta v = 4$, $\sigma = 5$, $v_H = 4$, and $v_L = 0$. In addition, the set of available prices $\mathcal{A}$ is all integers between 1 and 15 included. We initialize the Q-matrices with values between $\underline{q} = 3$ and $\overline{q} = 6$ so that all values of the initial Q-matrix are above the maximal payoff a dealer can get in a given period.[15] There are $K = 10,000$ experiments, $T = 200,000$ episodes per experiment, and in all experiments we set $\beta = 0.0008$ and $\alpha = 0.01$. This means that the algorithm chooses to experiment 1249.5 times in expectation, and hence "tries" each price about 100 times on average. As $\alpha = 0.01$, this frequency of experimentation is enough (in expectation) to override the initial values of the Q-matrix.[16] Finally, we set $\kappa = 0.05$, so that an experiment is said to "converge" if algorithms' actions have been unchanged for the last 10,000 episodes.

3.3 Results

In this section, we report the main results of our experiments. We first consider the monopoly case ($N = 1$) and duopoly case ($N = 2$), holding other parameters to their baseline values (Sections 3.3.1 and 3.3.2). In particular, we compare the distribution of final prices in these cases to their equilibrium values when $N = 1$ (monopoly price) and $N = 2$ (duopoly case), accounting for the fact that dealers must position their prices on a grid. Given this constraint, the theoretical monopoly price is $a^m = 7$ and there are two possible Nash-Bertrand equilibria ($a^c = 3$ or $a^c = 4$).

[15] This specification is common in the literature on Q-learning to guarantee that all actions are chosen sufficiently often to overcome the initial values of the Q-matrix. Indeed, as long as $q_{m,n,t}$ is larger than the maximal payoff the agent can realize, action m will necessarily be picked again because all the cells associated with actions that are played eventually fall below the maximal payoff.

[16] Note that Q-learning algorithms are meant for situations in which agents have no prior knowledge of the environment. Hence, there is no basis on which one could optimize the algorithm, e.g., by picking the "best" values of α and β. Rather, these values and the rules used by the algorithm must be seen as parameters.

3.3.1 A single AM ($N = 1$) behaves more competitively than in theory

Consider the case in which $N = 1$ first. Panel (a) of Figure 1 reports the evolution of the greedy price $a_{1,t}^{*k}$ over episodes, averaged over the $10,000$ experiments while Panel (b) reports the distribution over the K experiments of the final greedy price, $a_{1,T}^{*k}$ (whether convergence takes place or not). Panel (a) suggests that, on average the greedy price converges as the number of episodes becomes large. However, only 73.64% of the experiments converge (as defined in Section 3.2.2) and the final greedy price is heterogeneous across experiment as shown in Figure 1. The average final greedy price, $a_{1,T}^{*k}$, across all experiments is 6.16. However, there is substantial heterogeneity across experiments. As panel (b) shows, while most experiments ultimately reach a greedy price of 6, a substantial number reach 5 or 7, and in a few cases even 8 or 9. This dispersion in final outcomes across experiments is due to the environment being stochastic. Even though the values of the Q-matrix are not very sensitive to individual observations (remember that $\alpha = 0.01$), there is still a significant probability to obtain sufficiently many "bad draws" resulting in zero demand with prices of 6 or 7 to lead to a Q-matrix with a greedy price of 5 or 8, even though the monopolist's payoff is maximized at 7.

[Insert Fig. 1 here.]

A striking feature of this experiment is that, most of the times (in more than 75% of all experiments), the algorithm fails to learn the theoretical (optimal) monopoly price 7 even though T is large. Moreover, this failure is not random: On average, the final price posted by the algorithm is below 7. The modal price is 6, and the algorithm is more likely to set a price of 5 than a price of 8, even though playing 8 would give a higher expected profit than playing 5. The reason is that the updating rule (8) is biased against actions giving a high payoff with a low probability, such as choosing a high price. This effect is more pronounced when α is larger, but still significant even with the low value of α we are using (in unreported results, we checked that the average final greedy price indeed decreases in the parameter α).[17]

[17] To understand this point, imagine there are only two actions a_1 and a_2. Action a_1 gives a sure payoff π_1, whereas a_2 gives a payoff $\pi^+ > \pi_1$ with probability p, and $\pi^- < \pi_1$ with probability $1-p$, with $\pi_2 = p\pi^+ + (1-p)\pi^- > \pi_1$. If both actions are played many times, the expectation of $q_{2,t}$ associated with a_2 will converge to π_2. However, as noted in Lemma 1, the random variable $q_{2,t}$ itself does not converge pointwise. Instead, the distribution of $q_{2,t}$ converges to a non-degenerate distribution. A simple example is the case $\alpha = 1$: then $q_{2,t}$ will be equal to $\pi^+ > \pi_1$ with probability p and $\pi^- < \pi_1$ with probability $1-p$. Hence, the Q-learning algorithm will mistakenly pick action 1 as the greedy action with probability $1-p$.

One might be tempted to interpret this failure to learn the optimal price with probability 1 as a deficiency of the algorithm. However, this class of algorithms is not explicitly designed to learn the optimal price. Rather they seek to reach a certain balance between "exploring" and "exploiting". For instance, one could reach a final outcome closer to the monopoly price by choosing smaller values of α and β. However, doing so would be at the cost of playing suboptimal prices for more periods (so that the average profit of the AMs over all episodes might be smaller).

In any case, an important conclusion from the single dealer case is that the Q-learning algorithm used by the AMs in our experiments is not by itself biased towards high prices. If anything, the single dealer case shows that the opposite happens. This makes the non competitive final outcomes observed in the duopoly case (see next section) more striking.

3.3.2 Two AMs do not suffice to obtain Bertrand-Nash outcomes

Now consider the duopoly case ($N = 2$). The starting values of the Q-matrices, $\mathbf{Q}_{1,0}$ and $\mathbf{Q}_{2,0}$, for each AM as well as all the subsequent random draws for the two AMs, are drawn independently of each other (except the client's demand). Panel (a) of Figure 2 reports the evolution of the greedy price $a_{n,t}^{*k}$ for each AM over T episodes, averaged over the K experiments. Panel (b) reports the distribution of the final greedy price $a_{n,T}^{*k}$ for each AM over the K experiments.

[Insert Fig. 2 here.]

As can be seen in Figure 2, the AMs' quotes converge more quickly in the duopoly case than in the monopoly case. Convergence is also more frequent: In 94.18% of the experiments, the quote posted by each AM has converged after 200,000 episodes (vs., only 73.64% when $N = 1$). Moreover, in all experiments with convergence, the AMs end up posting the same price ($a_{1,T}^{*k} = a_{2,T}^{*k}$). However, this price is rarely one of the two Bertrand-Nash equilibrium prices (3 or 4 due to price discreteness). Indeed, we observe $a_{1,T}^{*k} = a_{2,T}^{*k} = 4$ in 5.57% of experiments only, and we never observe $a_{1,T}^{*k} = a_{2,T}^{*k} = 3$. As Panel (b) of Figure 2 shows, a majority of experiments (more than 60%) converge to a price of 5, about 20% converge to a price of 6, and some to 7 (the monopoly price) or 8. Thus, on average, the prices posted by the two competing AMs are far above the Bertrand-Nash equilibrium price.

The reason for this seemingly collusive outcome is different from the one in Calvano *et al.* (2020), because our setup precludes dynamic strategies (quotes cannot be contingent on past competitors' quotes in our set-up). Its origin seems closer to that in Asker *et al.* (2022) who also find that prices set by Bertrand competitors using Q-learning are above competitive prices (in an environment without adverse selection). In the first episodes, both AMs are experimenting with a high probability. AM 1 for instance is gradually learning how to best respond to AM 2. However, most of the time, AM 2 chooses a random price since the likelihood of experimentation is high in early episodes. The best response to AM 2 is actually for AM 1 to play $a = 6$. As AM 1 plays 6 more and more often (since the likelihood of experimentation declines over time), AM 2 should in principle learn that her best response is then to play $a = 5$ (in an undercutting process typical of Bertrand competition). However, because both AMs experiment less and less often over time, this undercutting process will typically not last long enough to reach the Bertrand outcome. For instance, both AMs may have reached a price of only 5 when the probability of experimenting ever again becomes very small. If for both AMs playing 3 or 4 did not prove profitable in the past (when the other AM was playing differently), then the AMs appear "stuck" with supra-competitive prices.[18] Our next step is to study how the probability that this happens depends on the parameters of the model, and in particular on the degree of adverse selection.

3.3.3 Adverse selection tends to make AMs' quotes more competitive

In this section we study how the outcomes of the simulations vary when we change the parameters of the economic environment, in particular the degree of adverse selection. For each set of parameters, in each experiment k and episode t we compute the following four variables (which correspond to empirically observable quantities):

1. **The trading volume** V_t^k, which is equal to 1 if a trade happens and 0 otherwise.

2. **The quoted spread** QS_t^k, which is the best ask minus the asset's ex ante expected value:

$$QS_t^k = a_t^{\min,k} - \mathbb{E}[\tilde{v}]. \tag{9}$$

[18] See Abada *et al.* (2022) for a comprehensive analysis and discussion of this issue. Wunder *et al.* (2010) show that even in a simple prisoner's dilemma Q-learning algorithms may not reach the Nash equilibrium.

3. **The realized spread** RS_t^k, which is:

$$RS_t^k = a_t^{\min,k} - v_t^k. \tag{10}$$

The realized spread is computed only when there is a trade. It measures the profit actually realized by the AM with the best quote, given the actual value v_t^k of the asset. Its average value over trades is a standard measure of dealers' expected profits per share in the literature (see Section 3.1).

We then compute the average across the K experiments of these three quantities in the last episode. That is, we compute:

$$\overline{V} = \frac{\sum_{k=1}^{K} V_T^k}{K} \tag{11}$$

$$\overline{QS} = \frac{\sum_{k=1}^{K} QS_T^k}{K} \tag{12}$$

$$\overline{RS} = \frac{\sum_{k=1}^{K} V_T^k RS_T^k}{\sum_{k=1}^{K} V_T^k}. \tag{13}$$

$$\tag{14}$$

[Insert Fig. 3 here.]

Panels a) and b) in Figure 3 show the effect of a change in σ (the variance of clients' private valuation) and Δv (the volatility of the asset payoff) on the average trading volume, the average quoted spread, and the average realized spread in the case with a single AM (dashed line), two AMs (plain line) and in the Bertrand-Nash equilibria (dotted lines).

As explained previously, an increase in σ reduces dealers' exposure to adverse selection and the elasticity of clients' demand to dealers' price. In the benchmark case (see Table 1), the first effect reduces adverse selection costs. For this reason, the quoted spread in the Bertrand-Nash equilibria decreases (weakly due to price discreteness) with σ. However, surprisingly, the opposite pattern is observed for the quoted spread posted by AMs: As σ increases, the two AMs post less competitive quotes. In fact, the effect of σ on AM's quotes is similar to its effect on the monopoly price (red dashed line in 3). In this case, like a monopolist, the competing AMs seem to take advantage of

the decrease in the client's demand elasticity to charge larger markups and thereby obtain larger expected profits (as shown by the evolution of the average realized spread).[19]

Thus, surprisingly, a decrease in adverse selection makes the quotes posted by AMs less competitive. The effect of Δv on AMs' quotes (Panel b) conveys a similar message. As Δv decreases from 8 to 4, AMs' exposure to adverse selection decreases. However, as shown by the evolution of AMs' realized spread, their rents increase, exactly as in the monopolist case. When Δv keeps decreasing (from 4 to 1), AMs rents decrease but in a way similar to what is observed in theory for the monopolist.[20] In sum, competing AMs react to a decline in adverse selection (an increase in σ or a decrease in Δv) in a way qualitatively similar to a monopolist rather than Bertrand competitors.

Panel (c) of 3 shows the effect of an increase in the number of AMs (from 1 to 10). As the number of AMs increases, AMs' quotes become closer to the Bertrand-Nash equilibria. Thus, AMs' rents (realized bid-ask spreads) decline. This pattern may seem intuitive. However, in theory it takes only two dealers to obtain the Bertrand-Nash equilibrium. Thus, economic theory predicts that bid-ask spreads and dealers' rents should decline when N increases from 1 to 2 but that a further increase in N should have no effect. Empirical findings regarding the effects of high frequency market makers' entry on bid-ask spreads, reported in Brogaard and Garriott (2019) (discussed in the introduction), are more consistent with the patterns obtained for AMs than those predicted by the Bertrand-Nash equilibrium.

3.3.4 Welfare implications of algorithmic market-making

Spread measures do not immediately translate into welfare measures. In particular, the realized spread RS measures a market-maker's realized profit (and hence, cost for the client) conditionally on a trade, but does not take into account the probability that this trade occurs. To further investigate the consequences of AMs for total welfare in the economy and its distribution between market-makers and buyers, we compute the levels of welfare, consumer surplus, and firm profits achieved with AMs and compare them to their counterparts in the competitive benchmark.

[19] The decline in the client's demand elasticity explains why trading volume increases with σ in the experiments, despite the fact that AMs charge a larger price to their client.

[20] AMs' rents also evolve in a way similar to that observed in one of the two Nash Bertrand equilibria (dotted purple line) but opposite to that in the other one (yellow dashed line). We think that the pattern observed in first case is due to price discreteness and will therefore not be robust with a finer grid, in contrast to other patterns.

For a given best ask a, total welfare can be computed as:

$$W(a) = \Pr(\tilde{v} + \tilde{L} \geq a)\mathbb{E}[\tilde{L}|\tilde{v} + \tilde{L} \geq a]. \tag{15}$$

In words, welfare in this model is driven by the liquidity shocks $\tilde{L}$, which create gains from trade between buyers and market-makers. Welfare is always lower when the ask price increases, and as a result even in the competitive case as increase in adverse selection lowers welfare. Welfare can be further decomposed into consumer surplus CS and producer surplus PS:

$$\begin{aligned} CS(a) &= \Pr(\tilde{v} + \tilde{L} \geq a)\mathbb{E}[\tilde{L} + \tilde{v} - a|\tilde{v} + \tilde{L} \geq a], & (16) \\ PS(a) &= \Pr(\tilde{v} + \tilde{L} \geq a)\mathbb{E}[a - \tilde{v}|\tilde{v} + \tilde{L} \geq a]. & (17) \end{aligned}$$

Based on the results of the experiments, we compute the average realized values of W, CS, and PS, and show in Fig. 4 how they vary with Δv and σ.

[Insert Fig. 4 here.]

We observe that an increase in σ leads to an increase in profits, due to both the AMs behaving less competitively (realized spreads increase) and demand elasticity being lower. However, because this elasticity is low, high prices have a lower impact on the probability that a trade is realized, and conditionally on a trade the gains are also higher. As a result, consumer surplus and total welfare also increase with σ. An increase in Δv has a somewhat ambiguous impact on realized spreads but it reduces profits, consumer surplus, and hence total welfare.

Overall, the comparative statics of welfare and profit with respect to σ and Δv are the same in the two benchmarks and with a duopoly of AMs, the levels reached with AMs being in between the monopoly benchmark and the competitive benchmark.

4 Price Discovery ($\bar{\tau} = 2$)

Models of trading with asymmetric information in financial markets are often used to study the process by which market participants discover asset fundamental values ("price discovery"). In these models, trades convey information about an asset payoff (because some trades come from informed investors). Using this information, uninformed traders (e.g., dealers) update their beliefs about this payoff in a Bayesian way. Via this dynamic learning process, over time, prices get closer to the asset value (see, for instance, Glosten and Milgrom (1985) or Easley and O'Hara (1992)).

In this section, we study whether AMs can also discover asset fundamental values ($\tilde{v}$ in our setting). To do so, we consider the case with two trading rounds ($\bar{\tau} = 2$), following the same steps as when $\bar{\tau} = 1$. That is, in Section 4.1, we first explain how to derive equilibrium prices in our two benchmarks (the monopoly case and the Bertrand-Nash equilibrium). Then, we explain how Q-learning algorithms work in this environment (Section 4.2). Finally we present the results in Section 4.3.

4.1 Benchmarks: Learning the Fundamental Value

When $\bar{\tau} = 2$, dealers can learn information about $\tilde{v}$ from the trading outcome at date 1. Thus, their beliefs regarding the payoff of the asset evolve over time. As is standard in models of trading with asymmetric information, in the benchmark monopoly and competitive cases, we assume that dealers update their beliefs in a Bayesian way. At the end of the first trading round in a given episode, there are two possible trading histories (H_1): (i) a trade at price a_1^{min} ($H_1 = \{1, a_1^{min}\}$) or (ii) no trade ($H_1 = \{0, a_1^{min}\}$). In the first case, dealers' Bayesian beliefs about the likelihood that $v = v_H$ is (remember that dealers' prior belief about this event is 1/2):

$$\mu_2(1, a_1^{min}) := Pr(v = v_H \mid H_1 = \{1, a_1^{min}\}) = \frac{D(a_1^{min}, v_H)}{D(a_1^{min}, v_H) + D(a_1^{min}, v_L)}, \tag{18}$$

where $D(a, v)$, given by (4), is the probability that the client buys the asset at price a when the asset value if v. In the second case, dealers' Bayesian beliefs about the likelihood that $v = v_H$ is:

$$\mu_2(0, a_1^{min}) := Pr(v = v_H \mid H_1 = \{0, a_1^{min}\}) = \frac{1 - D(a_1^{min}, v_H)}{2 - (D(a_1^{min}, v_H) + D(a_1^{min}, v_L))}. \tag{19}$$

It is easily checked that $\mu_2(1, a_1^{min}) > \mu_2(0, a_1^{min})$ if (and only if) $\Delta v > 0$. That is, Bayesian dealers should revise their beliefs about the expected payoff of the asset upward after a trade (buy) at date 1 and downward after no trade at date 1.

Given these observations, one expects the monopoly price and the competitive price (the Nash-Bertrand equilibrium price) to be larger (smaller) in the second trading round than in the first if there is a trade (no trade) in the first trading round. Table 2 shows that this is the case for the parameters of our experiments. In addition, in the competitive case, the difference between dealers' ask prices when there is a trade and when there is no trade increases with the informativeness of the order flow in the first period (i.e., increases with Δv and decreases with σ). In addition, Table 2 shows that, in the competitive experiment, the ask price posted by dealers in the second period is smaller on average than in the first period (that is, $\mathbb{E}[a_2^c] - a_1^c \leq 0$). This reflects the fact that as time passes, the informational asymmetry between dealers and their clients decline since dealers learn information about the asset payoff. Thus, they face less adverse selection and therefore across all possible realizations of v and the trading history at date 1, their ask price should be closer to the asset unconditional value in the second period than in the first period. In Section 4.3, we study whether AMs' quotes satisfy these properties or not. This is a way to study whether AMs learn to discover the asset payoff, even though they are not programmed to be Bayesian, as competitive dealers do in the benchmark case.

[Insert Table 2 here.]

Table 2 also shows that, as in the case with one trading round (and for the same reasons), (i) the competitive and the monopoly prices increase with the volatility of the asset (Δv) in each trading round, (ii) the competitive prices in each trading round decrease with the dispersion of clients' private valuations (σ) and (iii) the monopoly prices in each trading round increase with this

dispersion.

Last, observe that, in the competitive case, the quotes posted by dealers in the first trading round are identical to those obtained when there is a single trading round (compare Tables 1 and 2). In contrast, the monopolist price in the first trading round differs from that obtained when there is a single a trading round. This reflects the fact that, in choosing her price in the first trading round, a monopolist accounts for the effect of this price on her expected trading profit in the first trading round *and* her expected trading profit in the second trading round via the effect of her choice on her belief about the asset payoff given the first period outcome (trade/no trade).[21]

4.2 Q-Learning Algorithms

In this section, we explain how we adapt the Q-learning algorithms described in Section 3.2 to the case in which episodes have two trading rounds. The algorithms will keep track in each episode of the "state" they are in, and will play an action depending on the state. More specifically, for each AM n, we define $(N+3)$ states, denoted s_n, as follows: (i) $s_n = \emptyset$ in the first trading round; (ii) $s_n = NT$ in the second trading round if no trade takes place in the first; (iii) $s_n \in \mathcal{S} = \left\{0, \frac{1}{N}, \frac{1}{N-1}, ..., \frac{1}{2}, 1\right\}$ is the number of shares sold by AM n if a trade took place in period 1 (depending on how many AMs shared the market). Each AM then relies on a Q-matrix $\mathbf{Q}_{n,t} \in \mathbb{R}^{M\times(N+3)}$, in which each line corresponds to a different price and each column to a state, ordered as in the previous paragraph. We denote $q_{m,s,n,t}$ the (m,s) entry of matrix $\mathbf{Q}_{n,t}$.

We then modify the process described in Section 3.2.1 as follows. For any experiment k, we initialize the matrices $\mathbf{Q}_{n,0}$ with random values: Each $q_{m,s,n,0}$ (for $1 \leq m \leq M$, $1 \leq n \leq N$, and $s \in \mathcal{S}$) is i.i.d. and follows a uniform distribution over $[\underline{q}, \overline{q}]$. Then, in each episode t, we do the following:

Period 1:

1. For each AM n, we define $m_{n,t}^{1,*} = arg\max_m q_{m,\emptyset,n,t-1}$ the index associated with the highest value in matrix $\mathbf{Q}_{n,t-1}$ in state $s = \emptyset$ (the first period), and we denote $a_{n,t}^{1,*} = a_{m_{n,t}^{1,*}}$ the

[21] One can show that $\mu_2(1, a_1^{min})$ and $\mu_2(0, a_1^{min})$ increase with a_1^{min}. Thus, by choosing a high a_1^{min}, the monopolist improves the informational content of a trade at date 1 but it reduces the informational content of observing no trade.

corresponding greedy price.

2. For each AM n, with probability $\epsilon_t = e^{-\beta t}$ the AM "explores" by playing $a^1_{n,t} = a_{\tilde{m}^1_{n,t}}$, where $\beta > 0$ and $\tilde{m}^1_{n,t}$ is a random integer between 1 and M, all values being equiprobable. With probability $1 - \epsilon_t$, the AM "exploits" and plays the greedy price $a^1_{n,t} = a^{1,*}_{n,t}$. The random draws leading to exploring or exploiting are i.i.d. across all AMs in a given trading round of a given episode.

3. We compute $a^{1,min}_t = \min_n a^1_{n,t}$, and draw $\tilde{v}_t$ and $\tilde{L}_{1,t}$. This determines the position $I^1_{n,t}$ taken by each AM in period 1 and the state $s_{n,t}$ it will be in when period 2 starts. Formally, denote $\mathcal{D}^1_t$ the set of AMs who quote $a^{1,min}_t$ and z^1_t the size of this set. Then, if $\tilde{v}_t + \tilde{L}_{1,t} \geq a^{1,min}_t$ we have $I^1_{n,t} = s_{n,t} = \frac{1}{z^1_t}$ for every $n \in \mathcal{D}^1_t$, and $I^1_{n,t} = s_{n,t} = 0$ for $n \notin \mathcal{D}^1_t$. If $\tilde{v}_t + \tilde{L}_{1,t} < a^{1,min}_t$ then $I^1_{n,t} = 0$ and $s_{n,t} = NT$ for every n.

4. We update the first column of the Q-matrix of each AM as follows:

$$\forall 1 \leq n \leq N, q_{m,\emptyset,n,t} = \begin{cases} \alpha[a^1_{n,t} I^1_{n,t} + \max_{m'} q_{m',s_{n,t},n,t-1}] + (1-\alpha) q_{m,\emptyset,n,t-1} & \text{if } a^1_{n,t} = a_m \\ q_{m,\emptyset,n,t-1} & \text{if } a_{n,t} \neq a_m \end{cases} \tag{20}$$

Period 2:

1. At the beginning of period 2 we know the state $s_{n,t}$ in which AM n finds itself. We define $m^{2,*}_{n,t} = arg\max_m q_{m,s_{n,t},n,t-1}$ the index associated with the highest value in matrix $\mathbf{Q}_{n,t-1}$ in state $s = s_{n,t}$, and we denote $a^{2,*}_{n,t} = a_{m^{2,*}_{n,t}}$ the corresponding greedy price.

2. With probability ϵ_t the AM plays a random price $a^2_{n,t}$, following the same process as in period 1. With probability $1 - \epsilon_t$, the AM plays $a^2_{n,t} = a^{2,*}_{n,t}$.

3. We compute $a^{2,min}_t = \min_n a^2_{n,t}$ and draw $\tilde{L}_{2,t}$. This determines the position $I^2_{n,t}$ taken by each AM in period 2, following the same rules as in period 1.

4. For each AM n, we only update the column corresponding to state $s_{n,t}$, as follows:

$$\forall 1 \leq n \leq N, q_{m,s_{n,t},n,t} = \begin{cases} \alpha[a^2_{n,t} I^2_{n,t} - \tilde{v}_t(I^1_{n,t} + I^2_{n,t})] + (1-\alpha) q_{m,s_{n,t},n,t-1} & \text{if } a^2_{n,t} = a_m \\ q_{m,s_{n,t},n,t-1} & \text{if } a^2_{n,t} \neq a_m \end{cases} \tag{21}$$

Q-learning algorithms were initially designed to solve dynamic stochastic optimization problems (both finite and infinite horizon), and are thus in principle well suited to optimizing prices in this environment. The Q-matrix is defined in such a way that each algorithm can in principle learn to play a different price in period 2 depending on the "state", that is, depending on whether there was a trade in period 1. Note that, in addition, the state needs to include the amount sold by the AM in period 1. Indeed, as v_t is only revealed in period 2, the Q-matrix can record only at the end of period 2 what was the actual cost of selling some units of the asset in period 1.[22]

4.3 Results

To study price discovery with AMs using the Q-learning algorithms described in the previous section, we proceed exactly as in Section 3.3. In particular, we use the same parameter values for K (number of experiments), T (number of episodes per experiments), α and β. For brevity, we only focus on the case with two AMs ($N = 2$). We measure price discovery by AMs (i.e., whether AMs' quotes reflect information about the asset payoff contained in the first period trade) by computing the magnitude of the average price reaction to the observation of a trade vs. no trade (across experiments with the same environment). Formally, defining $V_t^{\tau,k}$ the total trading volume in trading round τ of episode t in experiment k, we compute:

$$Discovery = \frac{\sum_{k=1}^K V_T^{1,k}[a_T^{2,min,k} - a_T^{1,min,k}]}{\sum_{k=1}^K V_T^{1,k}} - \frac{\sum_{k=1}^K (1 - V_T^{1,k})[a_T^{2,min,k} - a_T^{1,min,k}]}{\sum_{k=1}^K (1 - V_T^{1,k})}. \tag{22}$$

[22] Using inventory levels as the state variable is common in other applications of Q-learning, in particular in dynamic pricing and revenue management. See, e.g., Rana and Oliveira (2014) for a recent example. The list of states used by the algorithms is an important parameter of the model. The list could be even richer (e.g., conditioning on prices in period 1 as well), or coarser (not distinguishing states NT and 0).

The variable *Discovery* is the empirical counterpart, in our experiments, of the difference between the ask price in the second period when there is a trade and when there is no trade in the benchmark cases. In these cases, this difference is always positive because dealers become more optimistic about the asset payoff after observing a buy in the first trading round than after observing no buy (see Table 2).

We also want to study whether price discovery induces dealers to charge lower markups relative to their expectation of the asset payoff because it reduces informational asymmetries, as is observed when dealers are competitive in the benchmark case ($\mathbb{E}(a_2^c) < a_1^c$; see Table 2). To this end, we compute the average difference (denoted $Difference$) between the ask price posted in the second trading round and the ask price posted in the first trading round across experiments:

$$Difference = \frac{\sum_{k=1}^{K}[a_T^{2,min,k} - a_T^{1,min,k}]}{K}. \tag{23}$$

If AMs behave as in the competitive benchmark, $Difference$ should be negative. If it is not and $Discovery > 0$, this indicates that (i) price discovery takes place but (ii) AMs take advantage of the reduction in informational asymmetries to charge less competitive prices, in line with our observations in the static case.

Figure 5 plots $Discovery$ and $Difference$ for different values of σ and Δv. In addition, we plot the highest and lowest values these quantities can take across the several Nash equilibria of the game, the monopoly benchmark, and the competitive benchmark with continuous prices.

[Insert Fig. 5 here.]

First, we observe that for all values of the parameters, $Discovery$ is positive. Thus, AMs learn to quote higher prices when a trade occurred in period 1 than when a trade did not occur. Hence, Q-learning algorithms are able to learn from past trades and contribute to price discovery. However, the algorithms seem to significantly "overshoot". That is, the difference in the prices posted by AMs following a buy or no buy in the first trading round is always larger than that predicted in the most competitive Nash-Bertrand equilibrium (the dashed dotted line), given price discreteness. This indicates that the difference in posted prices following a trade or no trade in the first trading

round is in part driven by deviations from competitive prices.

Second, we observe that $Difference$ is always positive, that is on average the algorithms use a higher price in the second trading round than in the first. This is in stark contrast to the competitive benchmark in which, at least if the tick size were zero, $Difference$ should be negative (as shown in Table 2 and the dashed green line in Figure 5 in the case of "Difference").

A mechanism that explains both results is, as in the static case, that adverse selection curbs the market power of algorithms. In our set-up, observing the trading outcome in the first trading round always reduces informational asymmetries between dealers and clients in the benchmark case. Thus, dealers' adverse selection cost is smaller in the second trading round. This decrease in adverse selection leads the AMs to settle on using less competitive prices, as we already observed in the static case. In addition, adverse selection is reduced more after a trade than after no trade (observing a trade is less likely ex ante, hence is more informative when it happens). Thus, AMs tend to charge larger markups after a trade than after no trade, explaining why on average $Difference$ is positive insead of negative as in the competitive case.

These results give interesting insights into how competition between algorithms can be spotted in the data. The first result implies that quotes will tend to over-react to order flow, potentially generating more long-term reversal. The second result implies that spreads tend to widen as adverse selection is resolved over time, whereas in competitive environments the opposite should occur (see, e.g., Glosten and Putnins (2020)).

5 Conclusion

We study the interaction of market-makers using Q-learning algorithms in a standard microstructure environment a la Glosten and Milgrom (1985). We show that this provides a natural workhorse model to study the role of algorithms in securities markets, and how their behavior may differ from what is predicted by standard theory. We find that, despite their simplicity and the challenge of an environment with adverse selection, algorithms behave in a realistic way: their quotes reflect adverse selection costs and they update their quotes in response to the observed order flow. However, their behavior is markedly different from what standard theory predicts. In particular, their quotes tend

to be above the competitive level, and to become less competitive over time as adverse selection gets resolved. More generally, our analysis shows that the interaction between algorithms is significantly affected by the presence and extent of adverse selection, suggesting that securities markets are a quite specific and particularly interesting application of recent research on competition between algorithms.

References

Abada, I., Lambin, X. and Tchakarov, N. (2022). *Collusion by Mistake: Does Algorithmic Sophistication Drive Supra-Competitive Profits?* Working paper. 6, 17

Asker, J., Fershtman, C. and Pakes, A. (2021). *Artificial intelligence and pricing: The impact of algorithm design*. Tech. rep., National Bureau of Economic Research. 6

—, — and — (2022). Artificial intelligence, algorithm design, and pricing. *AEA Papers and Proceedings*, **112**, 452–56. 17

Baldauf, M. and Mollner, J. (2020). High-frequency trading and market performance. *The Journal of Finance*, **75** (3), 1495–1526. 6

Banchio, M. and Skrzypacz, A. (2022). Artificial intelligence and auction design. *Available at SSRN 4033000 9*. 6

Biais, B., Foucault, T. and Moinas, S. (2015). Equilibrium fast trading. *Journal of Financial Economics*, **116** (2), 292–313. 6

Brogaard, J. and Garriott, C. (2019). High-frequency trading competition. *Journal of Financial and Quantitative Analysis*, **54** (4), 1469–1497. 4, 19

—, Hagströmer, B., Nordén, L. and Riordan, R. (2015). Trading fast and slow: Colocation and liquidity. *The Review of Financial Studies*, **28** (12), 3407–3443. 1

Budish, E., Cramton, P. and Shim, J. (2015). The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response *. *The Quarterly Journal of Economics*, **130** (4), 1547–1621. 6

Calvano, E., Calzolari, G., Denicolo, V. and Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, **110** (10), 3267–97. 5, 11, 12, 13, 17

Cartea, Á., Chang, P., Mroczka, M. and Oomen, R. C. (2022a). *AI driven liquidity provision in OTC financial markets*. Working paper. 6

—, — and Penalva, J. (2022b). *Algorithmic Collusion in Electronic Markets: The Impact of Tick Size*. Working paper. 6

Chen, L., Mislove, A. and Wilson, C. (2016). An empirical analysis of algorithmic pricing on amazon marketplace. In *Proceedings of the 25th international conference on World Wide Web*, pp. 1339–1349. 1

CMA (2018). Pricing algorithms. pp. 3–62. 1

Easley, D. and O'Hara, M. (1992). Time and the process of security price adjustment. *The Journal of Finance*, **47** (2), 577–605. 21

Foucault, T., Marco, P. and Ailsa, R. (2013). *Market Liquidity: Theory, Evidence, and Policy*. Oxford: Oxford University Press. 10

Glosten, L. and Putnins, T. (2020). *Welfare Costs of Informed Trade*. Working paper. 27

Glosten, L. R. and Harris, L. E. (1988). Estimating the components of the bid/ask spread. *Journal of financial Economics*, **21** (1), 123–142. 5

— and Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, **14** (1), 71–100. 1, 3, 4, 6, 21, 27

Hansen, K. T., Misra, K. and Pai, M. M. (2021). Frontiers: Algorithmic collusion: Supra-competitive prices via independent algorithms. *Marketing Science*, **40** (1), 1–12. 6

Hendershott, T., Jones, C. M. and Menkveld, A. J. (2011). Does algorithmic trading improve liquidity? *The Journal of Finance*, **66** (1), 1–33. 4

Jaakkola, T., Jordan, M. I. and Singh, S. P. (1994). On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, **6** (6), 1185–1201. 13

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, **53** (6), 1315–1335. 1, 3, 6

MacKay, A. and Weinstein, S. (2022). *Dynamic Pricing Algorithms, Consumer Harm, and Regulatory Response*. Working paper. 1

Menkveld, A. and Zoican, M. (2017). Need for speed? exchange latency and liquidity. *Review of Financial Studies*, **30** (4), 1188–1228. 6

Menkveld, A. J. (2013). High frequency trading and the new market makers. *Journal of financial Markets*, **16** (4), 712–740. 1

OECD (2017). Algorithms and collusion: Competition policy in the digital age. pp. 1–72. 1

O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics*, **116** (2), 257–270. 6

Rana, R. and Oliveira, F. S. (2014). Real-time dynamic pricing in a non-stationary environment using model-free reinforcement learning. *Omega*, **47**, 116–126. 25

Sutton, R. and Barto, A. (2018). *Reinforcement Learning: An Introduction*. Cambridge (Mass.): MIT Press. 11

Tsitsiklis, J. (1994). Asynchronous stochastic approximation and q-learning. *Machine Learning*, **16**, 185–202. 13

Watkins, C. and Dayan, P. (1992). Q-learning. *Machine Learning*, **8**, 279–292. 12

Wunder, M., Littman, M. L. and Babes, M. (2010). Classes of multiagent q-learning dynamics with epsilon-greedy exploration. In *ICML*, pp. 1167–1174. 6, 17

A Appendix

A.1 Tables

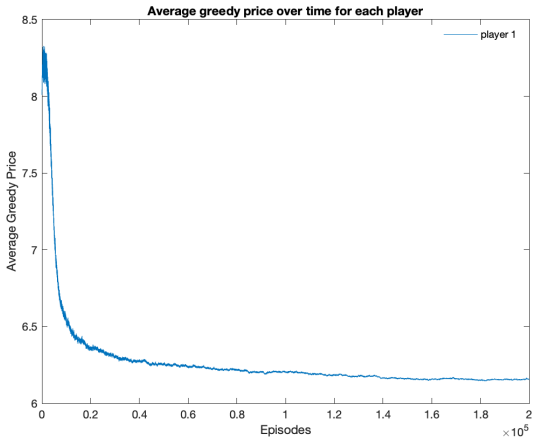
Panel A					
σ	0.5	1	3	5	7
Competitive Case					
a^c	4.00	4.00	3.24	2.68	2.47
Quo. Spread	2.00	2.00	1.24	0.68	0.47
Real. Spread	0	0	0	0	0
Monopoly					
a^m	4.37	4.69	5.68	6.54	7.03
Quo. Spread	2.37	2.69	3.68	4.54	5.03
Real. Spread	0.03	0.09	0.32	0.68	1.03
Panel B					
Δv	0	2	4	6	8
Competitive Case					
a^c	2	2.16	2.68	3.65	5.02
Quo. Spread	0	0.16	0.68	1.65	3.02
Real. Spread	0	0	0	0	0
Monopoly					
a^m	5.75	5.94	6.54	7.66	9.11
Quo. Spread	3.75	3.94	4.54	5.66	7.11
Real. Spread	0.82	0.78	0.69	0.57	0.47

Table 1: Predicted Outcomes in the Benchmark Cases, $\bar{\tau} = 1$. Prices are continuous. Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$ ($v_H = 4$ and $v_L = 0$). Panel A: $\Delta v = 4$. Quotes have been rounded up to two decimals (which explains why they are equal when $\sigma = 0.5$ and $\sigma = 1$). Panel B: $\sigma = 5$.

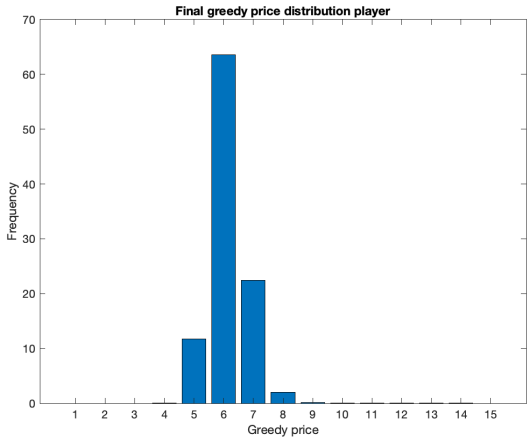
Panel A					
σ	0.5	1	3	5	7
Competitive Case					
a_1^c	4.00	4.00	3.24	2.68	2.47
$a_2^c, V_1 = 1$	4.00	4.00	3.82	3.26	2.92
$a_2^c, V_1 = 0$	4.00	4.00	2.44	2.08	2.02
$\mathbb{E}(a_2^c)$	4.00	4.00	2.96	2.62	2.45
Monopoly					
a_1^m	4.38	4.75	5.65	6.53	7.8
$a_2^m, V_1 = 1$	4.38	4.75	6.2	7.33	8.47
$a_2^m, V_1 = 0$	4.38	4.75	5.45	6.28	7.59
$\mathbb{E}(a_2^m)$	4.38	4.75	5.65	6.53	7.8
Panel B					
Δv	0	2	4	6	8
Competitive Case					
a_1^c	2	2.16	2.68	3.65	5.03
$a_2^c, V_1 = 1$	2	2.5	3.26	4.6	5.87
$a_2^c, V_1 = 0$	2	1.8	2.08	2.45	3.67
$\mathbb{E}(a_2^c)$	2	2.09	2.62	3.42	4.61
Monopoly					
a_1^m	5.76	5.94	6.53	7.61	9.09
$a_2^m, V_1 = 1$	5.76	6.2	7.33	8.61	9.73
$a_2^m, V_1 = 0$	5.76	5.87	6.28	7.26	8.86
$\mathbb{E}(a_2^m)$	5.76	5.93	6.49	7.53	9.01

Table 2: Predicted Outcomes in the Benchmark Cases, $\bar{\tau} = 2$. Prices are continuous. Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$ ($v_H = 4$ and $v_L = 0$). Panel A: $\Delta v = 4$. Quotes have been rounded up to two decimals (which explains why they are equal when $\sigma = 0.5$ and $\sigma = 1$). Panel B: $\sigma = 5$. In each case, $I_1 = 1$ if a trade takes place at date 1 and $I_1 = 0$ otherwise.

A.2 Figures

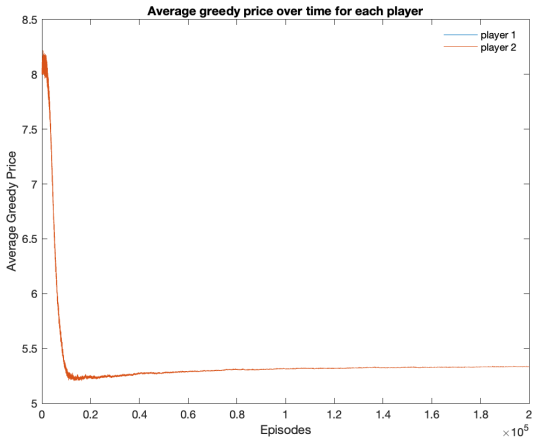


(a) Average greedy price as a function of time.

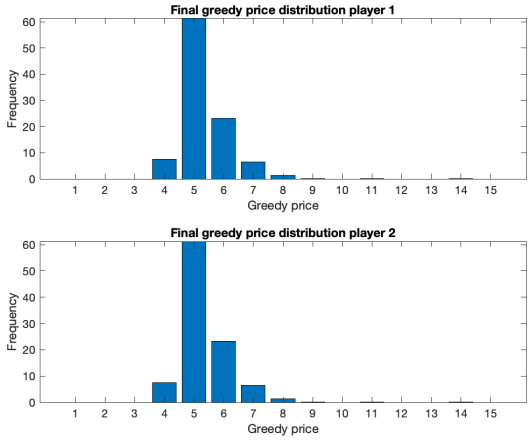


(b) Distribution of the final greedy price.

Figure 1: A single AM - Baseline Parameters.
Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\sigma = 5$, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$ ($v_H = 4$, $v_L = 0$ and $\Delta v = 4$)

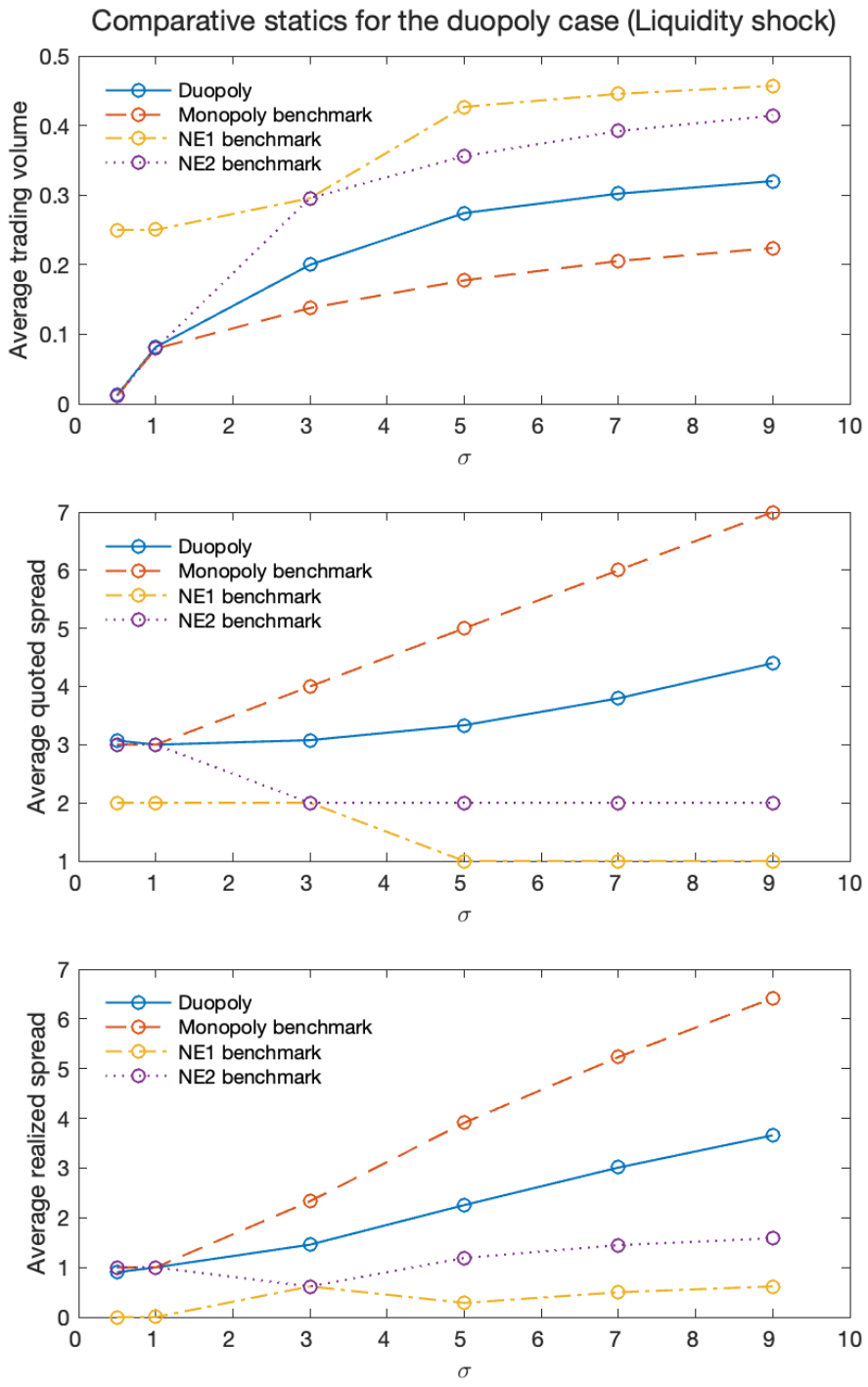


(a) Average greedy price of both AMs as a function of time.

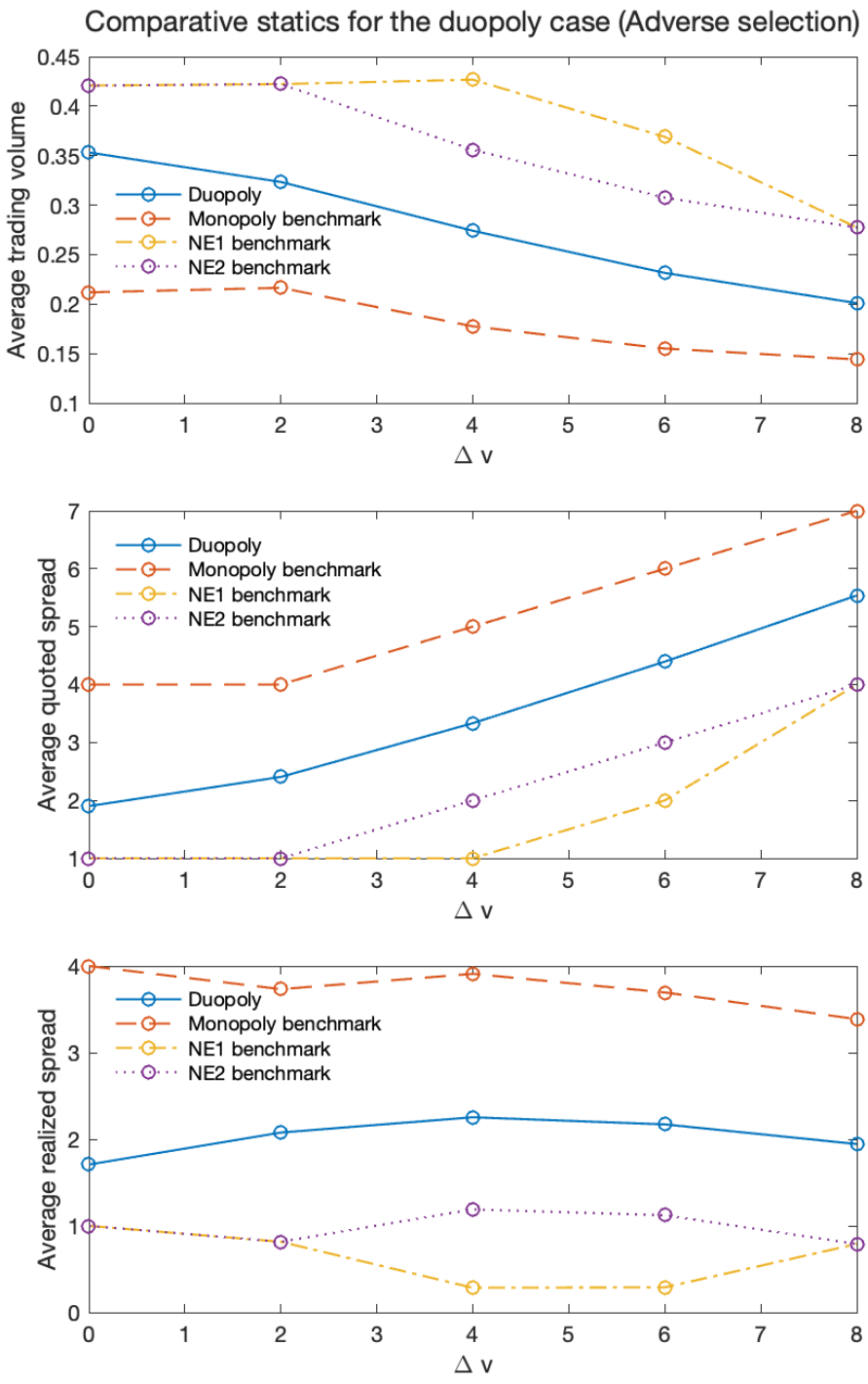


(b) Distribution of the final greedy price of both AMs.

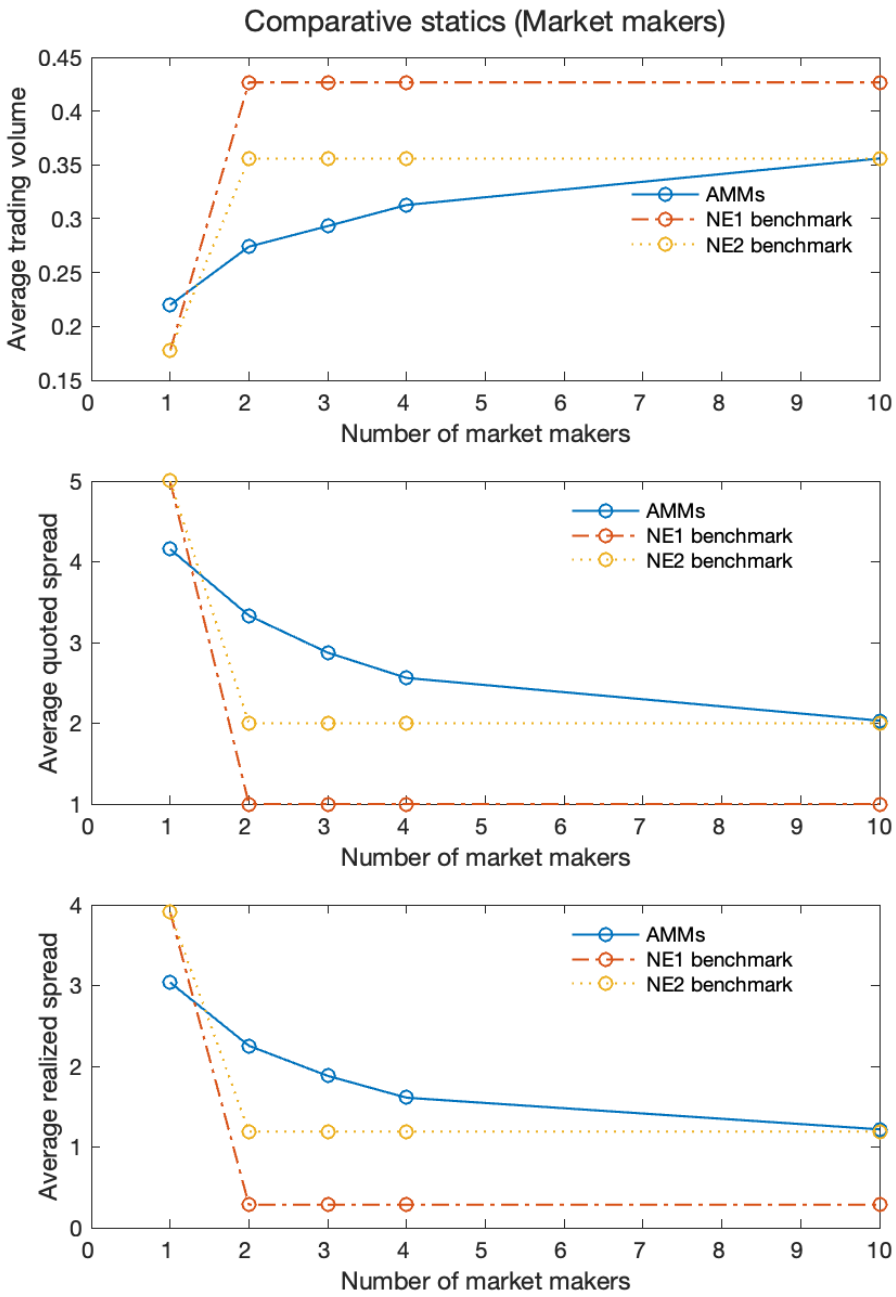
Figure 2: Duopoly of AMs - Baseline Parameters.
Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\sigma = 5$, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$ ($v_H = 4$, $v_L = 0$ and $\Delta v = 4$)



(a) Dispersion of Clients' Private Valuations (σ)
Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$ ($v_H = 4$, $v_L = 0$ and $\Delta v = 4$)

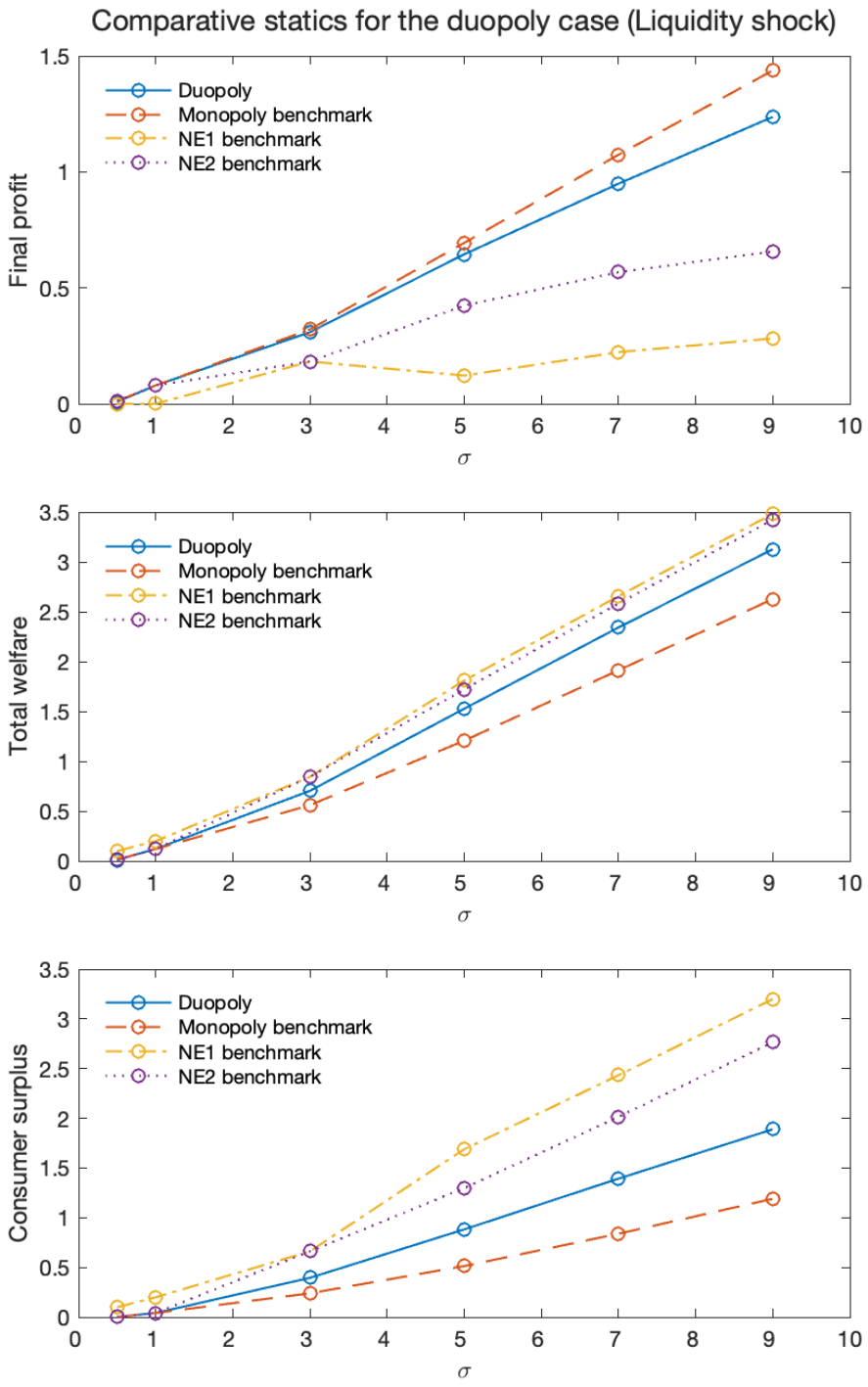


(b) Volatility of the Asset Payoff (Δv)
Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\sigma = 5$, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$

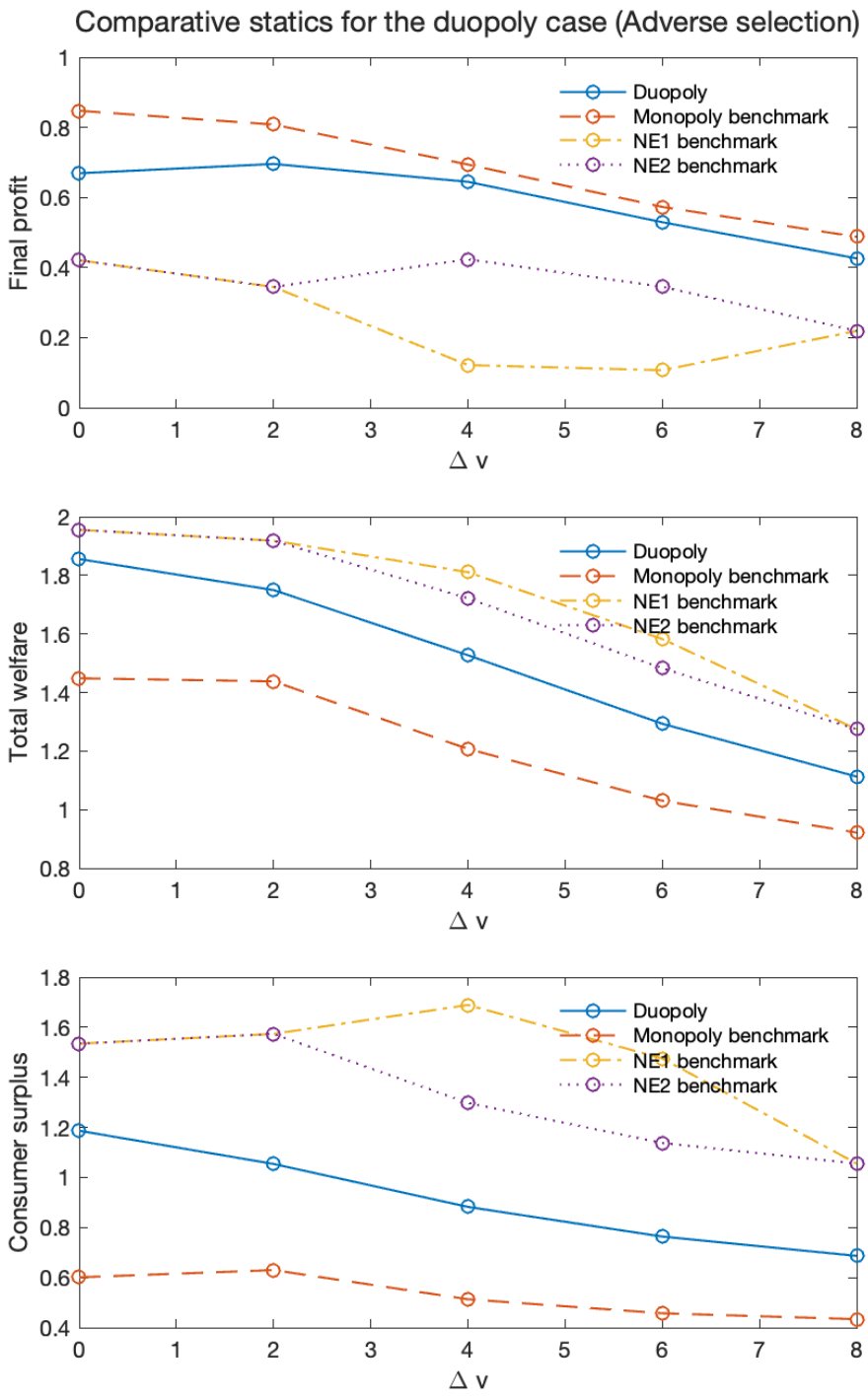


(c) Number of AMs (N)
Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\sigma = 5$, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$ ($v_H = 4$, $v_L = 0$ and $\Delta v = 4$)

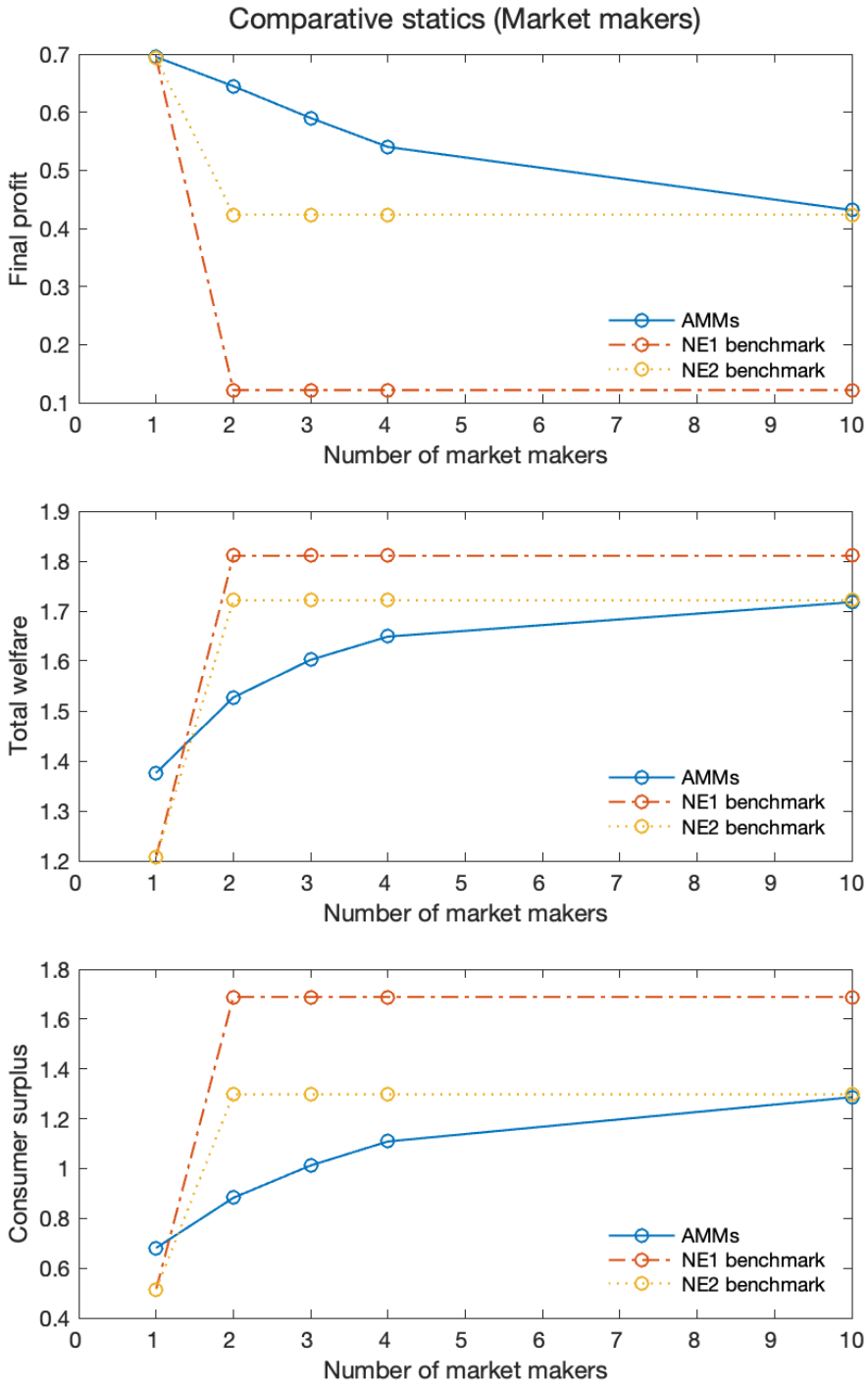
Figure 3: Comparative statics



(a) Dispersion of Clients' Private Valuations (σ)
Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$ ($v_H = 4$, $v_L = 0$ and $\Delta v = 4$)

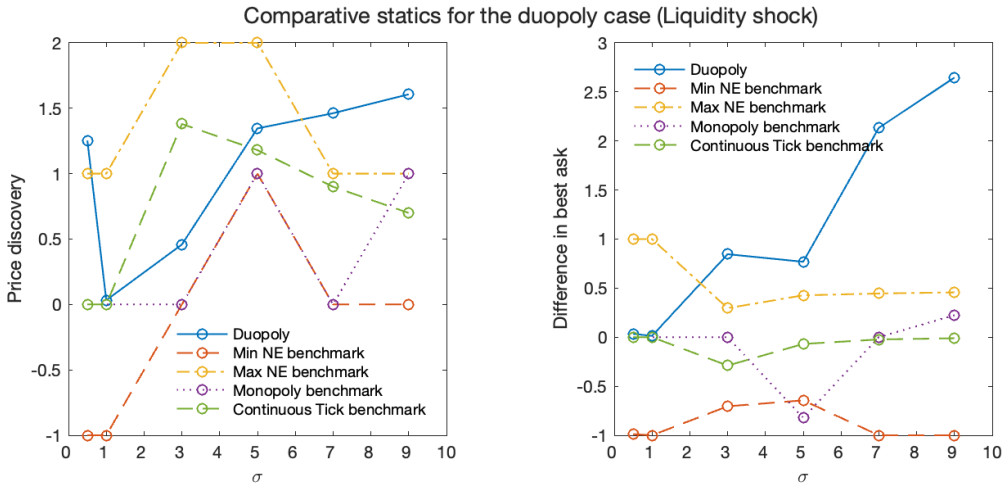


(b) Volatility of the Asset Payoff (Δv)
Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\sigma = 5$, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$

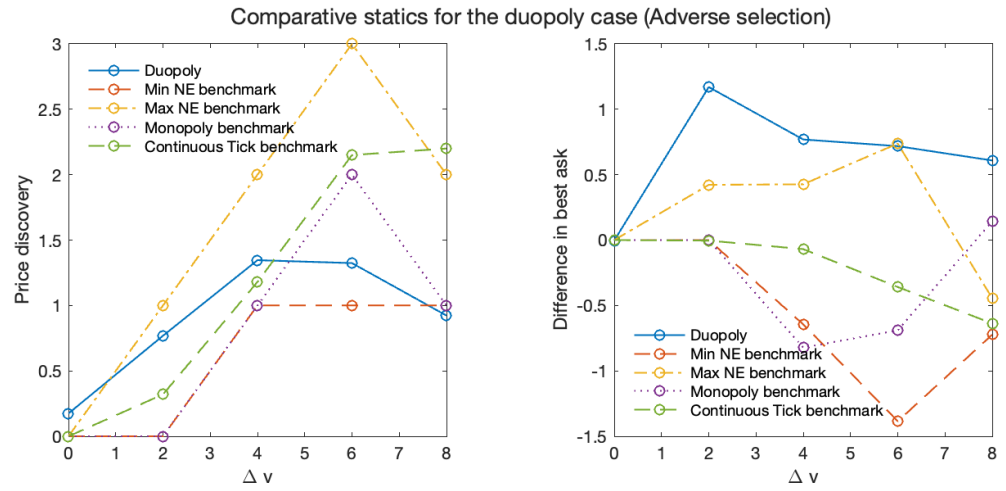


(c) Number of AMs (N)
Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\sigma = 5$, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$ ($v_H = 4$, $v_L = 0$ and $\Delta v = 4$)

Figure 4: Comparative statics



(a) Dispersion of Clients' Private Valuations (σ)
Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$ ($v_H = 4$, $v_L = 0$ and $\Delta v = 4$)



(b) Volatility of the Asset Payoff (Δv)
Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\sigma = 5$, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$

Figure 5: Comparative statics

A.3 Derivation of the Competitive Price

In this section, we explain how to compute the competitive price in a given trading round for given dealers' beliefs about the distribution of the asset payoff. We do so in the general case (for any $\overline{\tau}$) so that our results apply in particular when $\overline{\tau} = 1$ and $\overline{\tau} = 2$.

Let $V_\tau = I(a(\tau), \tilde{L}_\tau, \tilde{v}) \in \{0, 1\}$ denote the realized trade in period τ and let $I_{n\tau} = V_\tau Z(a_{n\tau}, a(\tau)) \in [0, 1]$ the trade executed by dealer n in round τ. Let H_τ denote the trading history (the observation of clients' trading decisions and the best quotes until trading round τ). That is, $H_\tau = \{(V_i, a_i^{min})\}_{\{i=1,2,...\tau,\}}$ for $\tau \geq 0$ and $H(0) = \emptyset$. The trading history contains information about the asset payoff. Indeed, holding the best quote constant, a client is more likely to buy the asset when $\tilde{v}$ is large than when $\tilde{v}$ is low. Let $\mu(H_{\tau-1})$ be dealers' estimate of the probability that $\tilde{v} = v_H$ at the beginning of trading round τ (with $\mu(0) = \mu = \frac{1}{2}$), given the trading history.

In a Nash-Bertrand equilibrium, in the τ^{th} trading round, all dealers posts the same price a_τ^c such that their expected profit is zero. This happens only if the expected profit of the dealer posting the lowest price is nil among all dealers. Thus, a_τ^c solves:

$$\overline{\Pi}(a_\tau^c, \mu(H_{\tau-1})) = \mu(H_{\tau-1})D(a_\tau^c, v_H)(a_\tau^c - v_H) + (1 - \mu(H_{\tau-1}))D(a_\tau^c, v_L)(a_\tau^c - v_L) = 0. \quad \text{(A.1)}$$

We deduce that:

$$a_\tau^c = \mathbb{E}(\tilde{v} \mid H_{\tau-1}) + \frac{\mu(H_{\tau-1})(1 - \mu(H_{\tau-1}))(v_H - v_L))(D(a_\tau^c, v_H) - D(a_\tau^c, v_L))}{\mu(H_{\tau-1})D(a_\tau^c, v_H) + (1 - \mu(H_{\tau-1}))D(a_\tau^c, v_L)}. \quad \text{(A.2)}$$

The competitive price is the smallest solution to this equation. Observe that it is equal to dealers' expectation of the asset payoff conditional on their information at the beginning of trading round j plus a markup (since $D(a_\tau^c, v_H) - D(a_\tau^c, v_L) = G(^c - v_H) - G(a^c - v_L) > 0$). This markup increases with dealers' uncertainty about the asset payoff at the beginning of trading round τ (measured by $\mu(H_{\tau-1})(1 - \mu(H_{\tau-1}))(v_H - v_L)$).

There is no analytical solution to (A.2). However, one can easily solve it numerically for specific parameter values. To solve (numerically) for the competitive price in the first trading round, we just replace $\mu(H_{\tau-1})$ by $\mu = 1/2$ in (A.2) (dealers' prior at the beginning of an episode). To solve

for the competitive price in the second trading round after a trade in the first trading round, we replace $\mu(H_1)$ by $\mu_2(1, a_1^c)$ (given in (18) in the text) in (A.2). To solve for the competitive price in the second trading round after no trade in the first trading round, we replace $\mu(H_1)$ by $\mu_2(0, a_1^c)$ (given in (19) in the text) in (A.2). Also, note that the probability that a trade occurs in trading round τ is $Pr(V_\tau = 1) = \mu(H_{\tau-1})D(a_\tau^c, v_H) + (1 - \mu(H_{\tau-1}))D(a_\tau^c, v_L)$. Hence, one also gets that:

$$a_\tau^c = \mathbb{E}(v \mid H_{\tau-1}, V_\tau = 1). \tag{A.3}$$

That is, the competitive price is the expected payoff of the asset conditional on the beginning of the trading history up trading round τ and the occurrence of a trade in trading round τ.

A.4 Derivation of the Monopolist's Prices

For any given belief $\mu \in [0, 1]$ that a monopolist might have about $\tilde{v}$ in a given round τ, if the monopolist sets a price of a, his expected payoff from that trading round is equal to $\bar{\Pi}(a, \mu_{\tau-1})$. Let $a^m(\mu)$ solve:

$$a^m(\mu) \in Arg \max_a \bar{\Pi}(a, \mu). \tag{A.4}$$

That is, $a^m(\mu)$ is the price that maximizes the monopolist dealer's expected payoff in round τ, given his belief μ. If the monopolist plays price $a^m(\mu)$, his round τ expected payoff is equal to

$$\Pi^*(\mu) := \bar{\Pi}(a^m(\mu), \mu)$$

Monopolist case with one trading round ($\bar{\tau} = 1$). Because the initial belief is $\mu = \frac{1}{2}$, when there is a single trading round, the monopolist sets a price of $a^m\left(\frac{1}{2}\right)$.

Monopolist case with two trading rounds ($\bar{\tau} = 2$). We now consider the optimal pricing policy of a monopolist dealer when there are two trading rounds. To do so, we proceed by backward induction.

In the second and last round, the price that the monopolist will choose if at the beginning of the second period his belief is μ_2 must be equal to $a^m(\mu_2)$ leading to a second period payoff of $\Pi^*(\mu_2)$

Let consider now the monopolist total payoff from the perspective of period 1. If he sets a first period price of a, then his posterior belief μ_2 is equal to $\mu_2(1,a)$ (given in (18) in the text) in (A.2) if the first period client buys, whereas $\mu_2 = \mu_2(0,a)$ (given in (18) in the text) if the first period client does not buy. Hence the monopolist will set his first period price a_1^m equal to the price a maximizing his total payoff

$$\underbrace{\bar{\Pi}\left(a, \frac{1}{2}\right)}_{\text{First round payoff}} + \underbrace{Pr(a)\Pi^*(\mu_2(1,a)) + (1 - Pr(a))\Pi^*(\mu_2(0,a))}_{\text{second round payoff}}, \tag{A.5}$$

where $Pr(a) := \frac{1}{2}(D(a, v_H) + D(a, v_L))$ is the probability that a trade takes place at date 1 if the monopolist chooses price a at this date. Thus, in choosing her price at date 1, the monopolist accounts for the effect of this price on her expected profit on the trade at date 1 and her continuation value.

When $\bar{\tau} = 2$, we obtain the benchmark price at date 2 in the monopoly case by solving numerically (A.4) (both when there is a trade at date 1 and when there is no trade) and the benchmark price at date 1 by maximizing (A.5).

A.5 Proof of Lemma 1

Fix a price a_m and a dealer n. Suppose that at episode t the dealer's price is $a_{n,t} = a_m$ and it is the lowest price among dealers, i.e. $a_{n,t} = a_m = a_t^{\min}$. Then three outcomes are possible: either the dealer does not trade, the dealer sells the asset worth v_H, or the dealer sells the asset worth v_L. In all cases the Q-matrix is updated. If the dealer does not trade then $\pi_{n,t} = 0$ and $q_{m,n,t+1} = (1-\alpha)q_{m,n,t+1}$, implying

$$|q_{m,n,t} - q_{m,n,t+1}| = \alpha |q_{m,n,t}|$$

If the dealer trades then $q_{m,n,t+1} = \alpha(a_m - \tilde{v}) + (1-\alpha)q_{m,n,t+1}$, and thus

$$|q_{m,n,t} - q_{m,n,t+1}| = \alpha |a_m - v_H - q_{m,n,t}|$$

if $\tilde{v} = v_H$, and

$$|q_{m,n,t} - q_{m,n,t+1}| = \alpha |a_m - v_L - q_{m,n,t}|$$

if $\tilde{v} = v_L$. Denote $\Delta_m(q) := \alpha \max\{|q|, |a_m - v_H - q|, |a_m - v_L - q|\}$ the maximum possible value of that $|q_{m,n,t} - q_{m,n,t+1}|$ can take given that $q_{m,n,t} = q$. Note that

$$\min_q \Delta_m(q) = \alpha \max\left\{\frac{a_m - v_L}{2}, \frac{v_H - a_m}{2}, \frac{v_H - v_L}{2}\right\} = \frac{\alpha}{2}\left(v_H - v_L + \left|a_m - \frac{v_H - v_L}{2}\right|\right) = \Delta_m^*$$

In words, no matter the value of $q_{m,n,t}$, at least one of the three possible outcomes mentioned above leads to $|q_{m,n,t} - q_{m,n,t+1}| \geq \Delta_m^*$. Thus the probability that $|q_{m,n,t} - q_{m,n,t+1}| \geq \Delta_m^*$ cannot be smaller than the smallest of the probabilities of these three events.

Now, given $a_{n,t} = a_m = a_t^{\min}$, the probability that the dealer sells the asset worth v_H, is at least $\frac{1}{2N} D(a_m, v_H)$. The probability that the dealer sells the asset worth v_L, is at least $\frac{1}{2N} D(a_m, v_L) < \frac{1}{2N} D(a_m, v_H)$. The probability that the dealer does not trade is $1 - \frac{1}{2}(D(a_m, v_L) + D(a_m, v_H))$, hence the expression for P_m^*. Q.E.D.