

Algorithmic Pricing and Liquidity in Securities Markets*

Jean-Edouard Colliard, Thierry Foucault, and Stefano Lovo

UNDER REVISION

March 7, 2025

Abstract

We let “Algorithmic Market Makers” (AMs), using Q-learning algorithms, determine prices for a risky asset in a standard market making game. We observe that AMs effectively adapt to adverse selection. However, AMs charge a markup over the competitive price. We show that this markup is larger when AMs’ receive noisier feedback about the true average payoff of their actions. For this reason, AMs’ markups are larger when there is no adverse selection, other things equal. We also show that AMs leave unique footprints on market outcomes. For instance, their quoted spread is more sensitive to an increase than to a decrease in adverse selection costs.

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 Sabrina Buti, Sylvain Chatherine, Alex Chincio, Winston Dou, Vincent Glode, Itay Goldstein, Terrence Hendershott, Yan Ji, Anton Lines, Lin Peng, Nick Roussanov, Mao Ye, Bart Yueshen, Chaojun wang, Yajun Wang, participants in “The Microstructure Exchange”, the Microstructure Asia Pacific Online Seminar, the 2022 Oxford Artificial Intelligence and Financial Markets Workshop, the 2023 NYU Stern Microstructure Conference, the 2023 Western Finance Association Meetings, the 2023 European Finance Association Meetings, the 2023 Financial Markets Liquidity Conference, the 2023 Luiss Finance Workshop, the 2023 CFM-Imperial conference, the ESCP Workshop on Competition policy in direct financial markets, and seminar participants at Aalto University, Bank of England, Bank of France, Baruch College, Bundesbank, Cornell University, CRESE, Frankfurt School of Management, HEC Paris, HKUST, Hong-Kong University, Keio University, Paris School of Economics, Peking University, Tokyo University, University College London, University of Copenhagen, University Paris 1, and Wharton for helpful comments and suggestions. We thank Olena Bogdan, Amine Chiboub, Pietro Fadda, 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” and the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 101018214). All rights reserved for Jean-Edouard Colliard, Thierry Foucault, and Stefano Lovo.

Introduction

Prices are increasingly set by algorithms in financial markets. For instance, Brogaard *et al.* (2014) and Chaboud *et al.* (2019) find, respectively, that about 42% and 60% of trades in stocks and currencies in their sample take place at prices set by algorithms. In U.S treasury markets, principal trading firms (PTFs), which also rely on algorithms, account for 21% of total trading volume (Brain *et al.* (2018)). Even in residential real-estate, pricing algorithms are now used (by intermediaries such as OpenDoor or Offerpad) to make cash offers to homeowners (Buchak *et al.* (2019)).

Until recently, these algorithms were rule-based. However, progress in Artificial Intelligence raises the possibility of using self-learning algorithms for securities trading, similar to those that have proved very successful in other contexts (e.g., in playing games such as chess or Go). This evolution raises questions about the effects of such AI-powered algorithms on market efficiency, liquidity, and stability.¹ To address these questions, it is necessary to develop insights on algorithms' behavior in financial markets (Goldstein *et al.* (2021)). This is a necessary step to better predict and explain their effects on market outcomes.²

Our paper contributes to this research agenda using an experimental approach. Specifically, we let Algorithmic Market Makers (AMs) using Q-learning algorithms play a market-making game similar to that in Glosten and Milgrom (1985) and study how they set their quotes in various market conditions (e.g., with and without adverse selection).³ To identify aspects of AMs' behavior that are anomalous from an economic viewpoint, we systematically compare AMs' quotes in our experiments to those predicted by the standard economic analysis of the game, which we refer to as the "Glosten-Milgrom prices."

In the baseline version of the market making game one client wishes to buy one share of a risky asset and requests quotes from 2 market makers. Market makers are uninformed about the payoff of the asset and simultaneously respond with an offer. The client buys if the best offer is less than

¹For instance, former SEC chairman Gary Gensler has expressed concern that AI-powered algorithms for trading could trigger the next financial crisis. See *The S.E.C.'s Chief Is Worried About A.I.*, The New-York Times, August 7, 2023

²Goldstein *et al.* (2021) note that *"Just as insights into human behavior from the psychology literature spawned the field of behavioral finance, so can insights into algorithmic behavior (or the psychology of machines) spawn an analogous blossoming of research in algorithmic behavioral finance."*

³We choose this specification because Q-learning is a foundational model for reinforcement learning algorithms (one important class of self-learning algorithms) and the Glosten and Milgrom (1985) is a canonical model of market making with asymmetric information.

her valuation for the asset, which is the sum of the payoff of the asset and a private valuation (the client’s “liquidity shock”), and does not trade otherwise. Thus, holding market makers’ prices constant, the client is more likely to buy the asset when its payoff is high than when its payoff is low. Market makers are therefore exposed to adverse selection. We assume that market makers face a long but finite sequence (1 million) of clients. After each client arrival, the asset pays off and the market makers receive their realized profit.⁴

Importantly, market makers have no prior knowledge of the primitives of the game (e.g., the distribution of the asset payoff, the distribution of the client’s demand, the number of market makers...). To learn how to set prices, they therefore use Q-learning algorithms. More specifically, each time a new client requests a quote, each algorithmic market maker (AM) either picks a price randomly in a fixed set or picks the “greedy price”, i.e., the price which, according to the AM’s past experience, is associated with the highest profit estimate (or “Q-value”). After the client’s decision is made, each AM updates the Q-value of the price it chose by taking a weighted average of its *realized profit* with the client and the Q-value of this price just before the client’s decision.

The Q-learning algorithm is therefore an iterative method to learn the profits associated with a set of possible actions (here AMs’ prices). This method embeds two key principles of any reinforcement learning algorithm. First, the algorithm receives “signals” (in our case realized profits) about the expected value of an action each time it chooses this action, and uses this signal to update its assessment of the value of the action for the decision maker. The sensitivity of the algorithm to new signals is controlled by one parameter (the so called “learning rate”). Second, to obtain a signal about the value of an action, the algorithm must try that action (experiment). The rate at which the algorithm experiments is controlled by another parameter (the “experimentation rate”). Q-learning algorithms are often designed to have a decaying experimentation rate over time. Indeed, at a given iteration, exploration enables the algorithm to learn new information about the value of an action, at the cost of not choosing the action with the highest Q-value. Intuitively, in a stationary environment, learning new information becomes less valuable over time since, with experience, the algorithm’s estimates of the profit associated with each action should become more accurate. Thus, it becomes less valuable to experiment over time. In our baseline experiments, as

⁴A new realization of the asset payoff is drawn after each client’s arrival. Moreover, these realizations and clients’ liquidity shocks are i.i.d. Thus, each request is exactly a repetition of the static market making game.

is usual in the literature, we exogenously set AMs’ learning and experimentation rates. However, as explained below, we also consider an extension in which the market makers designing the AMs can choose these parameters.

To study how the parameters of the environment (e.g., the volatility of the asset payoff or the dispersion of clients’ private valuation) affect AMs’ prices, we run experiments with different parameterizations. For a given environment, the path of prices chosen by an AM is stochastic because (a) the client’s decision is stochastic, (b) the asset payoff is stochastic and (c) the prices chosen by AMs are stochastic (due to experimentation). Consequently, the long run Q-value of each price and therefore the prices eventually chosen by the AMs are also stochastic. Thus, for each parameterization of the market making game, we run 1,000 different simulations (experiments) and we focus on the average long-run outcomes across these experiments. These outcomes can be seen as representative of the AMs’ behavior after their training phase.

In general, we observe that AMs eventually settle on the same price. Thus, after the training phase, they typically share the client’s demand. Interestingly, their “long run price” reflects their exposure to adverse selection: AMs’ quoted spread (the difference between their long run price and the unconditional expected payoff of the asset) is larger in environments with adverse selection than in environments without, other things equal. However, AMs’ long run price is significantly above the Glosten and Milgrom price, which means that AMs make positive profits on average. As a result, AMs’ long run price is not informationally efficient.⁵ Moreover, the difference between AMs’ long run price and the Glosten and Milgrom price (AMs’ average realized spread) increases with the dispersion of clients’ liquidity shocks. Thus, AMs are less competitive when the dispersion of clients liquidity shock is larger.

AMs’ supra-competitive prices stem from the fact that the AMs imperfectly learn to undercut each other.⁶ The reason is that AMs receive noisy signals about the average profit they can obtain when they undercut. For instance, suppose that both AMs settle on the same price p_0 above the Glosten and Milgrom price. Now suppose that AM1 experiments by undercutting p_0 by one tick at $p_0 - tick$. On average, this action is profitable since p_0 is above the Glosten and Milgrom price.

⁵The Glosten and Milgrom price is the expected payoff of the asset conditional on a trade taking place. Thus, it incorporates all available public information when a trade takes place and is therefore informationally efficient.

⁶As explained in detail in Section 4, our experiments are designed in such a way that there is no scope for tacit collusion between AMs. Thus, tacit collusion is not the reason why AMs prices are supracompetitive.

However, on a given instance, AM1 may be “unlucky”, if the client decides not to trade because her valuation is too low (AM1’s profit is then zero), or if the client trades but the realization of the asset payoff is large (AM1’s profit is then small or even negative). In these cases, AM1 receives a “negative” signal about the value of undercutting. Intuitively, the AM is more likely to learn fast that undercutting is profitable if (i) on average the signal is strong (the net gain of undercutting is large) and (ii) the noise is small (the variance of the net gain from undercutting is small).

We show that this simple heuristic goes a long way in explaining AMs’ behavior. First, it explains why in the early stage of AMs’ training, we observe a decline in prices and why this process eventually stops before AMs reach competitive prices. Indeed, when prices are high, the average gain from undercutting is relatively large compared to the noise. However, as prices gradually become lower, undercutting has a smaller signal to noise ratio. Hence, it requires more experiments for AMs to realize that undercutting is indeed profitable. But, as explained previously, the AMs’ experimentation rate decays over time. Thus, at some point, the AMs simply don’t experiment enough to discover that undercutting is profitable.

This issue is more acute when the dispersion of clients’ private valuations is large or when there is no adverse selection because the variance of AMs’ profits at a given price is larger in these cases. Thus, the signal received by the AMs is noisier in these environments, which explains why they stabilize on prices at which the gain from undercutting is even larger (that is, prices that are even less competitive).

Given this mechanism, we find that AMs’ long run price is more competitive (closer to the Glosten and Milgrom price) when their experimentation rate decays at a smaller rate. Then why don’t dealers choose a very high experimentation rate for the AMs? To answer this question, we consider an extension in which the market makers choose AMs’ learning and experimentation rates and show that each finds those considered in our baseline experiments optimal (in a sense that we make precise in the extension). The market makers do not engage in a race to the top for their experimentation rate for two reasons. First, as explained previously, experimenting for a longer duration is costly: It requires taking actions that have relatively low Q-value and might thus have truly low expected payoff. Second, if one market maker deviates by increasing its experimentation rate, he increases his average profit for a while since its AM is more likely to undercut the other AM when it is truly profitable to do so. However, eventually, this leads the other AM to also reduce

its price, making both AMs worse off.

Overall, these results show that AMs' supra competitive prices are a feature of their learning process, not a bug. Armed with this understanding of AMs' behavior, we make a series of predictions about the effects of AMs on trading outcomes. First, we observe that entry of additional AMs makes their long run prices closer to the Glosten-Milgrom price. Indeed, as the number of AMs increases, the signal-to-noise ratio from undercutting increases as well: The average gain from undercutting gets larger while the variance of this gain gets smaller.⁷

Second, we show that a reduction in the tick size can result in larger quoted and realized spreads. The reason is the following. When the tick size is reduced, an AM gets a larger increase in average profit when it undercuts a price at which AMs are tied up. This effect raises the signal-to-noise ratio from undercutting and therefore works to make AMs' prices more competitive. However, when the tick size is reduced, AMs' choice set becomes larger. Thus, AMs' experimentation capacity is spread out over a larger number of actions and, as a result, AMs experiment each price a lower number of times. This effect slows down AMs' ability to learn the value of undercutting prices at which the AMs are tied. When the tick size becomes small enough this effect dominates and AMs' prices become less competitive.

Third, AMs' reaction to symmetric shocks on adverse selection costs is asymmetric: They raise their price when the shock is positive and stay put when the shock is negative.⁸ To highlight this phenomenon, after the AMs' training phase, we change the volatility of the asset by a fixed amount and we assume that new clients arrive. When the volatility of the asset increases, AMs face larger adverse selection costs. We observe that even though AMs do not experiment any more, they quickly raise their price. The reason is that the AMs keep updating the Q-value of the price on which they have settled. As they receive lower profits (since the adverse selection cost has increased), they revise the Q-value of the price on which they have settled downwards and, at some point, this Q-value becomes smaller than the Q-value of another, higher price, to which they switch. In contrast, when the volatility of the asset decreases, AMs face smaller adverse selection cost. Thus, they

⁷Interestingly, this pattern is also found empirically by Brogaard and Garriott (2019) who study the effects of entry of high-frequency market makers on the liquidity of Canadian stocks. As noted by Brogaard and Garriott (2019), it cannot be explained by the standard economic analysis of the market-making game, which predicts that two dealers are sufficient to obtain the competitive outcome.

⁸This asymmetry has the flavor of the "Rockets and Feathers" observed in some product markets. See Peltzman (2000).

receive higher profits and revise the Q-value of the price on which they have settled upwards. Thus, a negative shock on adverse selection reinforces AMs' assessment that the price learned during the training phase has the highest Q-value, while a positive shock weakens this assessment.

Last, we show that AMs' learning process affects their price dynamics following trades. For this, we consider an extension of the market making game in which AMs receive requests from two clients, in sequence, before the asset payoff is realized. In this extension, the Glosten and Milgrom price for the second client is larger than for the first client if the latter buys and smaller if the latter does not trade. Moreover, on average, the price charged to the second client is smaller than for the first client.⁹ We observe that, as the Glosten and Milgrom prices, AMs' quotes increase after a buy from the first client and decrease when the first client does not. However, AMs overreact in the first case and underreact in the second case, relative to the Glosten and Milgrom prices. As a result, AMs charge even less competitive prices for the second client than for the first, so that on average their price in the second period is larger than in the first period. These patterns stem from the fact that AMs' learning problem for the second client is more complex than for the first. Indeed, each AM can make its price for the second client contingent on the outcome with the first client (e.g., buy/no buy). For each outcome, AMs have fewer opportunities to learn how to do so than they have to learn how to set a price for the first client.¹⁰ Thus, AMs' estimates of the average profits they can obtain by undercutting are less accurate than with the first client.

In the next section, we position our contribution in the literature. Section 2 presents the market making game and its Nash equilibrium. Section 3 describes the Q-learning algorithms used by AMs in our experiments and how we design our experiments. In Section 4, we report our experimental findings and propose an interpretation for AMs' behavior in the experiments. This interpretation suggests several testable implications that we present in Section 5. In Section 6, we endogenize the choice the parameters of the Q-learning algorithms by dealers using them. Section 7 concludes. Some formal derivations are in the appendix and an online appendix provides additional results.

⁹These properties are well-known. They follow from the fact the first client's decision conveys information about the asset payoff and, in Glosten and Milgrom (1985), market makers account for this information (in a Bayesian way) in setting their quotes for the second client.

¹⁰For instance, suppose the first client buys 40% of the time. Then, AMs have only 400,000 (600,000) opportunities to learn how to set price for the second client after a first client's buy (no buy). In contrast, they have 10^6 opportunities to learn from how to set price with the first client.

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.¹¹ [Calvano *et al.* \(2020\)](#) show that Q-learning algorithms can learn dynamic collusive strategies in a repeated differentiated Bertrand game. [Asker *et al.* \(2023\)](#) and [Abada *et al.* \(2022\)](#) show that supra competitive prices can be reached in this type of environment even if collusive strategies (via dynamic punishment strategies) are ruled out theoretically, through what [Abada *et al.* \(2022\)](#) call “collusion by mistake”.¹² [Banchio and Skrzypacz \(2022\)](#) find that Q-learning algorithms post less competitive bids in first price auctions than in second price auctions. [Banchio and Mantegazza \(2022\)](#) show how reinforcement learning can be approximated with a continuous time system of differential equations. In contrast to our setting, in these models, player’s payoff are deterministic and the only source of noise an algorithm faces in estimating its action’s payoffs comes from the stochastic play of the other algorithms. For example, in [Banchio and Skrzypacz \(2022\)](#), 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 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 mitigates this issue.¹³ To our knowledge, we are the first to study how market makers using Q-learning algorithms behave in the presence of adverse selection.¹⁴ [Dou *et al.* \(2023\)](#) study how informed traders using Q-learning algorithms behave in a [Kyle \(1985\)](#)’s environment. Their analysis and ours are complementary: We focus on market makers’ pricing behavior while [Dou *et al.* \(2023\)](#) focus on informed investors’ order submission strategies. Interestingly, they find that, in noisier environments, informed investors

¹¹Regulators have expressed concerns about this possibility in online retailers’ markets (see [MacKay and Weinstein \(2022\)](#), [Competition Market Authority \(2018\)](#), [OECD \(2017\)](#)). We are not aware of similar concerns expressed for securities markets so far.

¹²This 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](#)). See also [Waltman and Kaymak \(2008\)](#) for an application to Cournot competition.

¹³Another uncommon feature of our setting is that the demand faced by pricing algorithms is stochastic. See also [Hansen *et al.* \(2021\)](#), [Cartea *et al.* \(2022b\)](#), or [Wilk \(2022\)](#) for other settings in which selling algorithms face a stochastic demand elasticity, but without adverse selection.

¹⁴[Cont and Xiong \(2023\)](#) and [Guéant and Manziuk \(2019\)](#) study how market makers using reinforcement algorithms set prices in the face of inventory holding costs. However, there is no adverse selection in their framework.

behave less competitively (submit orders of smaller sizes and get larger average profits). This observation echoes our finding that an increase in the variance of AMs’ profits (e.g., due to an increase in the dispersion of their clients’ liquidity demands) leads AMs to settle on less competitive prices. As in their set-up, this finding is related to the collusion by mistake (or, using the terminology of Dou *et al.* (2023), “artificial stupidity”) phenomenon. 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 converge to non Nash behavior in a market making environment.

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 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. As explained in the introduction, this approach generates strikingly different predictions for those of canonical Bayesian-learning models.

At a more abstract level, our paper is related to an extent stream of literature on reinforcement learning in decision environments and games. In our setup, dealers do not know their strategic environment and treat the choice of which prices to use as a bandit problem. They are not able to formulate Bayesian priors over their environment and have to pick actions based on experience following a certain process, as in, e.g., Easley and Rustichini (1999). An important feature of our process is that dealers do not experiment all actions forever. This implies that in the long-run they

will have correct estimates of the payoffs associated with the actions they take, the actions they take will appear optimal given their payoff estimates, but estimates for the actions that are not selected in the long-run may be incorrect. This feature is not specific to our approach, but is a well-known property of algorithms for the bandit problem, even in the Bayesian case where a solution is known (Gittins, 1979), and of more general reinforcement learning environments (Easley and Kiefer, 1988).

Applied to games, this feature implies that players may not converge to a Nash equilibrium. An extent literature (surveyed, e.g., in Fudenberg and Levine (1998)) studies various reasonable learning processes and whether they converge to the Nash equilibrium. An important concept in that literature is the self-confirming equilibrium (Fudenberg and Levine, 1993): players behaving as statisticians may converge to a situation where each player behaves optimally given her beliefs about payoffs, these beliefs are correct for the strategies that are actually played, but the players have wrong estimates for the payoffs of deviations, because by definition they don't play them. This type of phenomenon is exactly why our algorithms do not converge to the Nash equilibrium, which again is something well understood in the literature on learning in games.¹⁵

These two streams of literature are mostly interested in comparing different algorithms and processes to each other and to the Nash equilibrium benchmark. Our goal is different. Assuming that today's financial algorithms can be modeled as following one such process, we are asking how they react to changes in their economic environment, a comparative statics question. We show that the insights coming from the theoretical literature have a number of interesting new empirical implications for the literature on algorithmic and high-frequency trading, as well as normative implications (specifically, on how the tick size can affect the competitiveness of algorithms).¹⁶

2 The Market Making Game

In this section, we describe the market making game played by algorithmic market makers in our experiments and we derive the “Glosten-Milgrom price” obtained in the Nash equilibrium of this game. These prices form a natural and useful benchmark for understanding algorithmic market

¹⁵Dou *et al.* (2023) elaborate more on the relation between Q-learning and self-confirming and experience-based equilibria.

¹⁶Pouget (2007) and Banchio and Skrzypacz (2022) go further and show how different trading mechanisms that would be equivalent with rational traders can lead to different outcomes with reinforcement learning algorithms, calling for research on market design specifically for markets populated by such algorithms.

makers' behavior.

2.1 The Market Making Game with Adverse Selection

One investor (“client”) wants to buy one share of a risky asset.¹⁷ The asset payoff, \tilde{v} , has a binary distribution, $\tilde{v} \in \{v_L, v_H\}$, with $\Delta_v := v_H - v_L \geq 0$ and $\mu := \Pr(\tilde{v} = v_H)$. This payoff is realized before trading starts and is only disclosed after trading has taken place or not.

The client privately knows her own valuation for the asset, that is equal to $\tilde{v}^C = \tilde{v} + \tilde{L}$, where \tilde{L} is normally distributed with mean zero and variance σ^2 , and is independent from \tilde{v} . We refer to \tilde{L} as *the client's liquidity shock* and denote its c.d.f by $G(\cdot)$. The distribution of \tilde{v}^C is therefore a mixture of two normal distributions with means v_L or v_H , respectively, as shown in Figure 1.

[INSERT FIGURE 1 ABOUT HERE]

After observing her valuation, the client requests quotes from N dealers, who simultaneously respond by posting a price (a_n for dealer n) at which they are willing to sell up to one share of the asset. We denote $\bar{a} = \{a_n\}_{1 \leq n \leq N}$ the vector of prices, $a^{\min} := \min_n \{a_n\}$ the best offer (i.e., the lowest price), and N^{\min} the number of dealers posting this offer. The asset payoff is disclosed to dealers after the client's decision (buy/no buy).

The client buys if and only if the best offer is less than her valuation ($a^{\min} \leq \tilde{v}^C$). Let $V(a^{\min}, \tilde{v}^C)$ be the client's realized demand (volume of trade). It is 1 if the client buys the asset and 0 otherwise. Dealer n 's realized trading volume is

$$I(a_n, \bar{a}, \tilde{v}^C) := V(a^{\min}, \tilde{v}^C) Z(a_n, \bar{a}), \quad (1)$$

where $Z(a_n, \bar{a}) = \frac{1}{N^{\min}}$ if $a_n = a^{\min}$ (the client's demand is split equally among the dealers posting the best offer) and $Z(a_n, \bar{a}) = 0$ otherwise. Hence, dealer n 's realized profit is

$$\Pi(a_n, \bar{a}, \tilde{v}^C, \tilde{v}) := I(a_n, \bar{a}, \tilde{v}^C)(a^{\min} - \tilde{v}). \quad (2)$$

Importantly, in this game, dealers are exposed to adverse selection. Indeed, holding the best offer constant, the client is more likely to buy the asset when its payoff is low than when its payoff is

¹⁷We only consider the case in which the client is a buyer. This simplifies the analysis without changing the economics of the problem.

large. To see this, let $D(a^{\min}, v) := \Pr(a^{\min} \leq \tilde{v}^C \mid \tilde{v} = v)$ be the probability that the client buys the asset when the payoff of the asset is v . We have

$$D(a^{\min}, v) := \Pr(a^{\min} \leq \tilde{v}^C \mid \tilde{v} = v) = \Pr(a^{\min} \leq v + \tilde{L}) = 1 - G(a^{\min} - v). \quad (3)$$

Thus, holding the best price constant, $D(a^{\min}, v_H) > D(a^{\min}, v_L)$ (see Figure 1).

2.2 The Market Making Game without Adverse Selection

As our experiments are designed to understand how variations in dealers' exposure to adverse selection affects their prices, it is useful to have a benchmark market-making game in which there is no adverse selection. We design a market-making game identical to the one described in Section 2.1, with the exception that the clients' valuation is $\tilde{v}^C = \tilde{w}^C + \tilde{L}$, where \tilde{w}^C and \tilde{v} are i.i.d. In this case, there is no adverse selection since the client's decision to buy does not depend on \tilde{v} . Thus, the likelihood that the client buys the asset is the same whether the asset payoff is high or low. Nevertheless, this likelihood is the same as when there is adverse selection because the unconditional distribution of the client's valuation for the asset is exactly the same in both cases.¹⁸

The last point is important. It enables us to compare experimental outcomes with and without adverse selection, holding all others parameters of the environment unchanged (e.g., the distribution of clients' valuations). To see why, consider an alternative, such as increasing the standard deviation of liquidity shocks, σ , in the adverse selection case. Such an increase reduces adverse selection, because it reduces the difference between the likelihood of a buy when $v = v_H$ and when $v = v_L$ (the red area in Figure 1). However, an increase in σ also alters the entire distribution of the client's valuation for the asset and, consequently, the likelihood of a buy at a given price. Therefore, differences in outcomes across experimental treatments with different values of σ cannot be solely attributed to variations in adverse selection.

2.3 Glosten-Milgrom Benchmark

We benchmark the outcomes of our experiments against the "Glosten-Milgrom price" obtained in the Nash equilibrium of the market-making game, both with and without adverse selection. Our

¹⁸The likelihood of a buy is $\mathbb{E}_\mu(V(a^{\min}, \tilde{v}^C)) := \mu D(a^{\min}, v_H) + (1 - \mu) D(a^{\min}, v_L)$ in either case since \tilde{w}^C has the same distribution as \tilde{v} .

goal is not to test whether our algorithmic market makers can learn to play these prices. Instead, we aim at identifying behaviors exhibited by algorithms that rely on trial-and-error methods to set prices, which deviate from the predictions of standard economic analysis of the market-making game. This approach allows us to uncover patterns and implications that are unique footprints of these algorithms.

From (2), dealer n 's expected profit, $\bar{\Pi}(a_n, \bar{a}; \mu) := \mathbb{E}_\mu(\Pi(a_n, \bar{a}, \tilde{v}_\tau^C, \tilde{v}))$, is

$$\bar{\Pi}(a_n, \bar{a}; \mu) = Z(a_n, \bar{a})[\mu D(a^{\min}, v_H)(a^{\min} - v_H) + (1 - \mu)D(a^{\min}, v_L)(a^{\min} - v_L)], \quad (4)$$

which can be written as:

$$\bar{\Pi}(a_n, \bar{a}; \mu) = \underbrace{Z(a_n, \bar{a})\mathbb{E}_\mu(V(a^{\min}, \tilde{v}^C))}_{\text{Dealer's expected trading volume}} \left[\underbrace{(a^{\min} - \mathbb{E}_\mu(\tilde{v}))}_{\text{Quoted spread}} - \underbrace{\Delta_v \frac{(1 - \mu)\mu\Delta_D(a^{\min})}{\mathbb{E}_\mu(V(a^{\min}, \tilde{v}^C))}}_{\text{Adverse selection cost}} \right], \quad (5)$$

where $\mathbb{E}_\mu(\tilde{v}) := \mu v_H + (1 - \mu)v_L$, and $\mathbb{E}_\mu(V(a^{\min}, \tilde{v}^C)) := \mu D(a^{\min}, v_H) + (1 - \mu)D(a^{\min}, v_L)$ are the expected trading volume, and the likelihood of a buy, respectively, and $\Delta_D(a^{\min}) := D(a^{\min}, v_H) - D(a^{\min}, v_L)$. The term in brackets in (5) is dealer n 's expected profit per share conditional on a trade.

Let a^* be the lowest price such that if $a^{\min} = a^*$ then dealers obtain zero expected profits ($\bar{\Pi}(a^*, \bar{a}; \mu) = 0$). From (5), we deduce

$$a^* = \mathbb{E}_\mu(\tilde{v} \mid \tilde{v}^C > a^*) = \mathbb{E}_\mu(\tilde{v}) + \underbrace{\Delta_v \frac{(1 - \mu)\mu\Delta_D(a^*)}{\mathbb{E}_\mu(V(a^*, \tilde{v}^C))}}_{\text{Adverse selection cost}}, \quad (6)$$

The zero expected price is the expected payoff of the asset conditional on the client buying the asset, exactly as in [Glosten and Milgrom \(1985\)](#). For this reason, we call it the Glosten-Milgrom price.¹⁹ In our setting, using the standard Bertrand logic, a^* is the unique Nash equilibrium of the market making game.²⁰

¹⁹The Glosten-Milgrom price is the solution of a fixed point problem (6) for which there is no closed-form solution given our specification of $G(\cdot)$. This problem always has at least one solution (when there are more than one, the Glosten-Milgrom price is the smallest root of (6)). See [Appendix A.2](#).

²⁰The Bertrand logic is as follows. Suppose instead that there is a Nash equilibrium at a price $a > a^*$. Then, if one dealer deviates by undercutting by an infinitesimal amount this price, he increases by $\frac{N-1}{N}$ her expected profit since he captures the entire demand while the decline in her expected profit per share is infinitesimal.

The Glosten-Milgrom price has several important properties. First, the expected (half) quoted spread, $\overline{QS} := \mathbb{E}(a^{\min} - \tilde{v})$, is strictly positive and just equal to dealers’ adverse selection cost. Second, the expected (half) realized spread, $\overline{RS} := \mathbb{E}(a^{\min} - \tilde{v} \mid \tilde{v}^C > a^{\min})$, is equal to zero. The expected realized spread differs from the expected quoted spread because it is computed using realizations of $(a^{\min} - \tilde{v})$ (the realized profits of dealers’ posting the best price) *only* when trades happen ($\tilde{v}^C > a^{\min}$). It measures the expected profit per share traded for a dealer (the term in bracket in (5) and is often used to this end by empiricists. Third, as shown in Figure 2 and proved formally Appendix A.2, the expected quoted spread (or, equivalently, adverse selection costs) increases with the volatility of the asset payoff and decreases with the variance of the investors’ liquidity shocks. Last, the Glosten-Milgrom price does not depend on the number of competing dealers, N . This is the standard result that two competitors are sufficient when firms producing a homogeneous product with identical costs compete based on prices (Tirole, 1988).

[INSERT FIGURE 2 ABOUT HERE]

In the case without adverse selection, the client’s decision to buy the asset is uninformative since \tilde{v}^C is independent from \tilde{v} . Thus, $a^* = \mathbb{E}_\mu(\tilde{v} \mid \tilde{v}^C > a^*) = \mathbb{E}_\mu(\tilde{v})$. Therefore, in this case, the expected quoted spread (and therefore expected realized spreads as well) is zero in equilibrium, for all values of the parameters.

3 Algorithmic Market Makers

3.1 The Problem

We now consider dealers who must play the market-making game for a number T of “episodes”. An episode consists of only one trading round and realizations of the asset payoffs and client valuations are independent across episodes. We assume that dealers are risk-neutral and only care about total (non discounted) payoffs. Hence, denoting $\pi_{n,t}$ the realized profit of dealer n in episode t , if the dealer were able to form a rational expectation about $\pi_{n,t}$ she would look for a pricing policy from $t = 1$ to $t = T$ that maximizes her total expected profit, that is:

$$\mathbb{E} \left[\sum_{t=1}^T \pi_{n,t} \right]. \tag{7}$$

In particular, because the market-making game is finitely repeated and has a unique Nash equilibrium, rational dealers would play the Glosten-Milgrom price in each episode.

Instead, we assume that dealers have minimal prior knowledge about their environment. Namely, at date 0, each dealer n only knows that she will have to select a price in a set \mathcal{A} for each of the T episodes, and she will only observe her realized profit $\pi_{n,t}$ at the end of each episode t . She knows neither the structure of the trading environment, nor the structure of competition, nor more generally the stochastic mapping from the prices she sets into the payoffs she receives. Thus, each dealer is unable to compute the expectation (7). Given this (lack of) information, from the perspective of the dealer, choosing prices amounts to a multi-armed bandit problem: she can try different prices (“arms”) to estimate the average payoffs associated with each price. Possibly, these average payoffs vary over time.

We assume that each dealer approaches this problem using a reinforcement learning algorithm. While there are various types of reinforcement learning algorithms, they all share a common core principle: Decision-makers learn to optimize their behavior through experimentation. In our context, the process involves trying a price, observing the resulting profit, updating estimates of the average profit associated with each price, and iterating further.

Over time, this iterative process allows the decision-maker to refine his estimate of the average payoff corresponding to each price. However, it also introduces a fundamental trade-off between experimentation and exploitation—a dilemma central to the bandit problem. Exploitation involves selecting the action with the highest estimated average payoff, whereas experimentation involves choosing an action with a lower estimated payoff to gain more information. Although experimentation is necessary for learning, it carries a cost, as it may lead to actions that yield genuinely low payoffs. For this reason, reinforcement learning algorithms usually reduce the frequency of experimentation over time, as the informational benefits of experimentation naturally diminish.

3.2 Q-Learning Algorithms

To focus the analysis on a simple form of reinforcement learning, we assume that dealers use Q-learning algorithms. Indeed, Q-learning is the foundation of more sophisticated reinforcement

learning algorithms and is one popular approach to solve multi-armed bandits problems.²¹ We refer to such dealers as Algorithmic Market Makers (AMs).

We restrict AMs to choose their quotes in $\mathcal{A} = \{a_1, a_2 \dots a_M\}$, where each a_m is a possible ask price.²² We choose this price grid so that the expected payoff of the asset, the Glosten-Milgrom price, and the monopoly price (the one maximizing $\bar{\Pi}(a, a; \mu)$ when $N = 1$) are all in the range $[a_1, a_M]$ (see below).

The Q-learning algorithm used by each AM works as follows. To each AM n and episode t , we associate a so-called *Q-Matrix* $\mathbf{Q}_{n,t} \in \mathbb{R}^{M \times 1}$, which is simply a column vector of size M .²³ The m^{th} entry of the matrix, denoted $q_{m,n,t}$, represents the estimate by AM n , in episode t , of the profit from playing price a_m . For each AM, we initialize $\mathbf{Q}_{n,0}$ with random values. Specifically, for each AM n and each price index m , $q_{m,n,0}$ has a uniform distribution over $[\underline{q}, \bar{q}]$ and is i.i.d across prices and AMs.

The Q-learning algorithm specifies i) how an AM chooses its price in every episode t , and ii) how an AM’s Q-matrix evolves over time given the prices it chose and the resulting realized payoff in a given episode. This specification relies on two parameters (common to all AMs), $\alpha \in (0, 1)$ and $\beta > 0$, and a probability $\epsilon_t := e^{-\beta t}$. Given this parameterization, we iterate the following three steps for each episode t between 1 and T :

1. Action: We first determine the behavior of each AM in episode t . For each AM 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 denote by $a_{n,t}^* := a_{m_{n,t}^*}$ the *greedy price* of this AM, that is, the price which according to the AM’s previous trades yields the largest estimated profit. With probability $1 - \epsilon_t$, AM n takes an “exploitation” action: it plays the greedy price. With probability ϵ_t , it takes an “exploration” action: the AM draws a random integer $\tilde{m}_{n,t}$ between 1 and M (all values being equiprobable) and quotes $a_{n,t} = a_{\tilde{m}_{n,t}}$. Thus, $a_{\tilde{m}_{n,t}}$ is chosen randomly in \mathcal{A} . Whether to explore and which price to

²¹See Sutton and Barto (2018) for an introductory textbook on reinforcement learning and the application of Q-learning to multi-armed bandits problems.

²²This constraint is necessary because the algorithm must evaluate the average profit associated with each possible price. Thus, the set of prices cannot be continuous.

²³In general, the Q-matrix of an agent has S columns, each corresponding to a state realized at the beginning of each episode that can affect the average payoff obtained by the agent with a given action. If there is no such state, $S = 1$, which is the case considered here. In particular, we do not allow AMs to condition the choice of their price on their past trading history to be as close as possible to the market making game considered in Section 2.1.

try are drawn independently across dealers. We denote $\bar{a}_t = (a_{1,t}, a_{2,t} \dots a_{n,t})$ the vector of prices quoted by all AMs in episode t (i.e., for the t^{th} client) and we record $a_t^{\min} = \min_n \{a_{n,t}\}$ the best offer in episode t .

2. Feedback: We then determine the realized profit for each AM in a way that reflects the true nature of the market making game described in Section 2. Nature draws the asset payoff \tilde{v}_t , the client’s liquidity shock \tilde{L}_t and, in the case without adverse selection, w_t^C , as described in Sections 2.1 and 2.2. We then determine the client’s valuation, \tilde{v}_t^C as described in these sections. If $v_t^C \geq a_t^{\min}$, we record a trade at price a_t^{\min} and otherwise we record the absence of trade. In either case, each AM n receives a profit equal to $\pi_{n,t} = \Pi(a_{n,t}, \bar{a}_t, \tilde{v}_t^C, \tilde{v}_t)$, as given by (2). In particular, the AMs quoting a_t^{\min} share the profit (or loss) from selling the asset (while others get zero). Moreover, if no trade takes place, all dealers receive a profit of zero.

3. Update: Each AM updates its Q-matrix as follows:

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

In words, after playing action m the AM updates the associated value in the Q-matrix and inputs a weighted average of the observed payoff and the previous value. The values associated with other actions do not change.

The Q-learning algorithm uses only two parameters to control the trade-off between “exploring” and “exploiting” common to all reinforcement learning algorithms (see Section 3.1) and how to update AMs’ estimates of their expected profit at each price:

- The parameter β controls the speed at which ϵ_t decays over time, so that AMs explore a lot in early episodes and end up exploiting with a probability close to 1 in later episodes. The logic is as follows. Experimentation is potentially costly since it means that the AM posts a price which, according to its current estimate, does not yield the highest expected profit. In early episodes, it makes sense to pay this cost because early estimates are unreliable anyway, and so experimenting with a new price may uncover a more profitable action. As the number of past episodes grows, information accumulates and the learning gain in experimenting becomes smaller. Intuitively, the algorithm should therefore gradually shift from exploring to exploiting over time. This is governed

by β : a larger β means that the shift to exploiting will occur faster.

– The parameter α controls the sensitivity of the AMs’ estimates to new observations. The higher is α , the higher is the impact of a new realization of the profit obtained by an AM at a given price on its profit estimate at this price. Importantly, the AM’s profit at a given price is random because both the asset payoff and the client’s decision to trade are random (see Section 4.2.1). Thus, even if all AMs keep playing the same prices, a too large α leads to unstable estimates (consider the extreme case $\alpha \rightarrow 1$). Conversely, if α is small, the entries of the Q-matrix are more stable but learning is slower (in the extreme case in which $\alpha \rightarrow 0$ there is no learning). Moreover, if the environment is not stationary (which is the case in our setup due to the interaction between AMs), a low α will give too much weight to less relevant payoff realizations in the distant past.

Parameters α and β are fixed parameters (sometimes called “hyper-parameters”): They do not change over time. We explain how we specify these parameters in our experiments in the next section and discuss how agents designing Algorithmic Market Makers might choose them in Section 6.

Last, there are many variants of the Q-learning algorithm, with different specifications for the experimentation probability ϵ_t and the updating rule (8), and more sophisticated classes of reinforcement learning algorithms. We choose a simple Q-learning algorithm for comparability with recent literature in finance and economics, and because it features in a simple and transparent way the main properties of reinforcement learning algorithms more generally.

3.3 Experimental Design

In this section, we describe how we design our experiments. This design is guided by two important considerations.

First, for a given parameterization of the market-making game $(\Delta_v, \sigma, \mu, \mathbb{E}_\mu(\tilde{v}))$, two different runs of T episodes can lead to different outcomes, even when starting from the same deterministic initial Q-matrix. Indeed, the profit, π_t , an AM actually receives in a given episode t with a given price is stochastic. It depends on the realization of the asset payoff in this episode, \tilde{v}_t , and the client’s valuation, \tilde{v}_t^C . Hence, in a given history, an AM can be “lucky” with a certain price and end up choosing this price very often, whereas in a different history, the same AM is unlucky with this same price and hence plays differently. To address this issue, we run a large number K of experiments

consisting of T episodes each, holding the parameters of the market-making game constant, and we focus on the distribution of outcomes (e.g., the average and the standard deviation of quoted spreads) across these experiments.

Second, the Q-learning algorithms that we use do not converge to a constant action as the number of episodes T grows large (see Appendix A.5 for a formal analysis).²⁴ The intuition is as follows. Suppose that this is not true. That is, after some period, AMs play the same greedy price a_m forever and, to simplify, that AMs do not experiment anymore. At this price, the likelihood that the client does not trade for the next T' episodes is always strictly positive, because $\Pr(\tilde{v}^C < a_m) > 0$ in our setting. In the absence of trade over the next T' episodes, the AMs' estimate of their profit at price a_m will decline. As this estimate can become arbitrarily close to zero with a positive probability for T' large enough, there is always another price that can become the greedy price with a positive probability - a contradiction.²⁵ In sum, in our setting, there is no price that can be a greedy price forever, no matter how long is the training period. To address this issue, we choose a large value of T and focus on the average value of different variables in episode T , across K experiments. We check that T is large enough that the distribution of these variables has converged (more on this below). In particular, we focus on the long run average behavior of the AMs, which is stable.²⁶

After experimenting with different parameterizations to address the two issues above, we settled on the following baseline parameterization. The parameters of the market making game are the same as in Figure 2: $\Delta_v = 4$, $\sigma = 5$, $v_H = 4$, $v_L = 0$, $\mu = 0.5$, and $N = 2$ (two AMs). In addition, AMs can choose all prices between 1.1 and 14.9 included on a grid with a tick size of 0.1 (139 prices in total). This specification makes sure that the zero expected profit prices are in the range

²⁴Watkins and Dayan (1992), Jaakkola *et al.* (1994), or Tsitsiklis (1994) study conditions under which Q-learning converges to the optimal action. These conditions are not met in our setup, for three reasons: (i) convergence to the optimal action requires the algorithms to experiment an infinite number of times, whereas our specification of ϵ_t leads to a finite expected number of experimentations; (ii) the updating rule needs to be such that the weight given to each additional observation goes to zero as T goes to infinity, whereas (8) always gives a constant weight α to the latest observation; (iii) the environment needs to be stationary, which is not the case in a multi-agent problem in which each agent changes its strategy over time. It is possible to change the algorithm to avoid problems (i) and (ii), at the cost of losing comparability with the recent literature using Q-learning algorithms in economics and finance. We do this in Online Appendix A.4. We still observe a distance with the predictions of the Glosten-Milgrom benchmark, due to problem (iii).

²⁵In Appendix A.5, we also show formally that the values of each AM's Q-matrix cannot converge to a single point.

²⁶Other papers in the literature take a different approach and wait for the algorithms to keep the same action for a large number of episodes before ending each experiment. That is, each experiment has potentially a different T . We do not follow this approach as it can in principle be misleading in a stochastic setup, see the Online Appendix OA.5. However, we observe that in most experiments the algorithms have indeed taken the same action for a large number of periods, so that this difference in approaches is likely inconsequential in practice.

of possible prices for all specifications considered in our experiments. We initialize the Q-matrices with random values following a uniform distribution between $\underline{q} = 3$ and $\bar{q} = 6$, so that all values of the initial Q-matrix are above the maximal payoff a dealer can get in a given period.²⁷ We run $K = 1,000$ experiments, each with $T = 1,000,000$ episodes. In all experiments we set $\alpha = 0.01$ and $\beta = 8.10^{-5}$. Given this specification, the AMs choose to experiment 12,500 times in expectation, and hence “explore” each price around 90 times on average.²⁸

For each set of parameters, in episode t of experiment k we record the minimum ask price $a_t^{min,k}$ and the realized asset value v_t^k . We then compute the following variables:

1. **The quoted spread** QS_t^k , which is the best offer minus the expected payoff of the asset:

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

2. **The realized spread** RS_t^k , which is the best offer minus the realized payoff of the asset and is computed only when there is a trade:

$$RS_t^k = a_t^{min,k} - v_t^k. \quad (10)$$

We first check that the distribution of the best ask price $a_t^{min,k}$ has converged. To do so, we record the empirical distribution of $a_t^{min,k}$ across the K experiments, for different values of t . We then test the null hypothesis that the samples $\{a_t^{min,k}\}_{k=1\dots K}$ and $\{a_T^{min,k}\}_{k=1\dots K}$ come from the same distribution, using a Kolmogorov-Smirnov test. After $t = 500,000$ episodes the p-value of the test is 0.94. For $t = 700,000$ and $t = 900,000$ the p-value is above 0.9999. In short, after 500,000 episodes the distributions of prices at various horizons become statistically indistinguishable from each other.

We then compute the average over the K experiments of these variables in the last episode ($t = T$).²⁹ The average quoted spreads and average realized spreads are empirical estimates of the

²⁷This 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. See in particular [Asker et al. \(2023\)](#). Indeed, as long as $q_{m,n,t}$ is larger than the maximal payoff the agent can obtain, action m will necessarily be picked again because all the cells associated with actions that are played eventually fall below the maximal payoff.

²⁸Each price will be played many more times due to the initialization of the Q-matrix, and in addition a price will be played with some probability when it becomes the greedy price.

²⁹The average realized spread is: $\frac{\sum_{k=1}^K V_T^k RS_T^k}{\sum_{k=1}^K V_T^k}$. That is, it is computed only when a trade occurs.

expected quoted spread, \overline{QS} , and the expected realized spread, \overline{RS} (defined in Section 2).

As AMs must post their quotes on a grid, the Glosten-Milgrom price might not be on the grid (and therefore AMs’ realized spread cannot be exactly zero). Moreover, if the tick size is large enough and the number of dealers small enough, the market making game can have two Nash equilibria in pure strategies and one equilibrium in mixed strategy (see Appendix A.6 for more details). Thus, when we report the results from our simulations (Section 4), we always compare to, and report, the quoted and realized spreads in the least competitive pure-strategy Nash equilibrium. In any case, as the tick size in our experiments is small, the difference between the Glosten-Milgrom price and the price in the least competitive Nash equilibrium is small.

4 Algorithmic Market Makers’ Behavior

In this section, we describe how AMs behave in our experiments (Section 4.1), with a focus on their long-run behavior, that is, after their “training” is supposed to be over. We then propose an explanation for this behavior (Section 4.2).

4.1 Experimental Findings

We first report, in Figure 3 (Panel A), the distribution of the greedy price in the last episode in the baseline case, with $\Delta_v = 4$ and $\sigma = 5$ (in all 1,000 experiments both AMs have the same greedy price in the last episode). In this case, the Glosten-Milgrom price is $a^* = 2.68$ and is therefore not exactly on the grid of possible prices. In the least competitive Nash equilibrium, dealers post a price of 2.8 (about 1 tick above the Glosten-Milgrom price). As the figure shows, AMs’ quotes vary across experiments (standard deviation of 0.73) and, in all experiments, the greedy price is above the Glosten-Milgrom price.

[INSERT FIGURE 3 ABOUT HERE]

The modal greedy price in the last episode is 4.60 and the mean is 4.97. Since 2.8 is the least competitive Nash equilibrium, at any price above 2.8 an AM would be strictly better off by undercutting its competitor. For instance, consider the case in which both AMs settle on a price of 5. At this price, in the baseline case, each AM obtains a true expected profit of 0.30. However,

each AM could obtain a greater expected profit, of 0.59, by undercutting its competitor by one tick (posting a price of 4.90). The AMs do not learn this.³⁰

Panel B of Figure 3 shows the evolution over episodes 1 to T of the average greedy price (averaged over the K experiments). In the first part of the learning process (roughly the 20,000 first episodes), the average greedy price decreases and then stabilizes at 4.97. Thus, initially at least, AMs seem to learn to lower their price to attract more clients. However, as their experimentation rate decays, they have fewer opportunities to learn and their assessment of their profit at each price changes less from one client to the next. As a result, in a given experiment, the greedy price tends to stabilize. Yet, it varies across experiments (the figure shows the evolution of the greedy prices one standard deviation away from the average) because the trading history is not the same.

In Panel A of Figure 4, we study the effect of the dispersion in clients' liquidity shocks σ on AMs' average quoted spread. To this end, we run $K = 1,000$ experiments for different values of σ ranging from 1 to 9 (other parameters are as in the baseline case), both in the adverse-selection case and the no-adverse-selection case ($18,000 = 2 \times 1,000 \times 9$ experiments overall). For each value of σ , we then compute and plot the average quoted spread \overline{QS} in each case. We also plot the quoted spread in the Glosten-Milgrom benchmark, with and without adverse selection.

Consider the adverse-selection case first. For all values of σ , the average quoted spread in this case is largely above the Glosten-Milgrom quoted spread. Strikingly, this is also the case when there is no adverse selection. These observations confirm for a broader set of parameters that AMs settle on non-competitive prices, failing to learn that they could increase their expected profit by undercutting their competitor at these prices.

Moreover, we observe that, in the adverse-selection case, the average quoted spread increases with σ , the dispersion of clients' liquidity shocks. This pattern is strikingly different from the Glosten-Milgrom benchmark in which the quoted spread decreases with this dispersion, even after accounting for the fact that AMs' quotes are constrained to be on a specific price grid (see the dotted-dashed lines on the figure). Interestingly, this pattern is also observed when there is no adverse selection. In sum, an increase in the dispersion of clients' liquidity shocks affects AMs'

³⁰Of course, by playing a price of 4.90, the AM may well eventually induce its competitor to post another price, say 4.90, at which they will both be worse off. However, nothing in the AM's design allows for this type of forward-looking reasoning (in particular, as AMs cannot condition their prices on the past trading history, they cannot learn that undercutting might generate a loss in future profits by triggering a drop in their competitor's price).

quoted spreads in a way that cannot be explained by the standard economic analysis of the market-making game.

In Panel B of Figure 4, we report the average realized spreads \overline{RS} for different values of σ . The figure shows in another way that AMs do not post competitive quotes: Their average profits per trade (average realized spread) are far above zero and those in the Glosten-Milgrom benchmark. Interestingly, AMs learn to cope with adverse selection since their average realized spread is positive for all values of σ . Thus, their average quoted spread exceeds adverse selection costs on average, which explains why AMs' average quoted spreads are larger when there is adverse selection than when there is not. However, AMs' average realized spreads are *smaller* with adverse selection than without, all else equal. Thus, adverse selection induces AMs to behave *more* competitively (charge smaller markups relative to costs).

The negative effect of adverse selection on dealers' rents contrasts sharply with the Glosten-Milgrom benchmark and is unexpected. To our knowledge, no existing models of competing risk-neutral dealers predict a negative relationship between dealers' quoted spreads and adverse selection.³¹

Last, AMs' average realized spreads increase with the dispersion of clients' liquidity shocks. Thus, AMs get larger rents, whether there is adverse selection or not, when the dispersion of clients' liquidity shocks increases. This finding is again at odds with the Glosten-Milgrom benchmark, even after accounting for price discreteness. It suggests again that it becomes more difficult for AMs to learn to undercut when the dispersion of clients' liquidity shocks gets larger and therefore adverse selection costs are smaller.

[INSERT FIGURE 4 ABOUT HERE]

In Figure 5, we consider the effect of the volatility of the asset payoff, Δ_v . Panel A shows that the average quoted spread increases with the asset volatility in the adverse-selection case, as in the Glosten-Milgrom benchmark. However, in contrast to this benchmark, this is also the case in the

³¹Liu and Wang (2016) examine a model featuring a monopolist dealer who serves both informed and uninformed investors. They show that an increase in the precision of a public signal about informed investors' private information (a decrease in adverse selection) can lead the monopolist dealer to widen the bid-ask spread. As explained by Liu and Wang (2016), this adjustment allows the monopolist dealer to mitigate adverse selection risk by redirecting some trades with informed investors to uninformed ones. This cannot happen in our setting because, at any given time, only one client enters the market. Therefore, the negative relationship between adverse selection and AMs' rents in our experiments must arise from a different mechanism.

no-adverse-selection case. Thus, the positive relationship between the asset volatility and AMs' quoted spread cannot just be due to the fact that adverse selection costs increase with the volatility of the asset payoff. Another mechanism must be at play.

Panel B of Figure 5 illustrates that AMs' average profit per trade (realized spread) increases with the volatility of the asset payoff. However, this increase is less pronounced under adverse selection. Similar to the observations in Figure 4, AMs' average realized spread is smaller in the adverse selection case across all values of Δ_v , despite their average quoted spread being larger. Overall, Figure 5 conveys a message consistent with Figure 4: AMs adapt to adverse selection, and adverse selection drives them to become more competitive.

[INSERT FIGURE 5 ABOUT HERE]

In sum, four facts emerge from our experiments:

1. AMs learn to not be adversely selected. In all environments considered in our experiments, their long-run average quoted spread is large enough to cover their adverse selection cost (average realized spreads are positive). Moreover, AMs charge larger quoted spreads in the case with adverse selection than in the case without.
2. AMs do not fully learn to undercut each other when it is theoretically profitable to do so. They settle on prices well above the Glosten-Milgrom price (their average realized spreads are strictly positive). This means that each AM could obtain a larger expected profit by undercutting its competitor. However, it fails to learn this.
3. AMs' prices are more competitive (average realized spreads are smaller) when there is adverse selection than when there is not.
4. AMs' prices are less competitive—whether there is adverse selection or not—when the dispersion of clients' liquidity shocks or the volatility of the asset payoff increase.

In Section OA.4 of the Online Appendix, we run additional simulations to assess the robustness of these observations when varying the hyperparameters (α, β) . Overall, our findings are robust to variations in these parameters.³²

³²We consider 3×3 configurations, with 3 values for β ($5 \cdot 10^{-6}$, $8 \cdot 10^{-5}$, $3.2 \cdot 10^{-4}$) and 3 values of α . We observe

As explained previously, the last 3 facts cannot be explained by the standard economic analysis of the market making game presented in Section 2. For instance, the standard analysis implies that AMs’ quoted spreads should decrease with the dispersion of clients’ liquidity shocks and that their average realized spreads should be zero whether there is adverse selection or not.

One possibility could be that AMs learn how to play a collusive equilibrium sustained by dynamic punishment strategies, as found in [Calvano *et al.* \(2020\)](#) and subsequent papers ([Dou *et al.*, 2023](#)). However, this explanation cannot hold in our case: while algorithms play the market making game many times with different clients in our experiments, they cannot condition their action on the trading history (in particular past prices and trade outcomes in the previous episode), unlike in [Calvano *et al.* \(2020\)](#). Therefore, they cannot execute strategies similar to punishment strategies in repeated games. The crucial point here is the absence of any observable state on which the algorithm can condition, not the fact that the algorithm maximizes the one-shot profit of the game instead of the discounted value of future profits.³³

Consequently, the puzzling outcomes we observe must rather stem from the way algorithms learn to behave. In the next section, we highlight two features of their learning process that we think explain the outcomes that we cannot explain with standard tools from economics.

4.2 Interpretation

Our experiments indicate that AMs “leave money on the table”, in the sense that, even after a relatively long learning period, they fail to undercut their competitor when it would theoretically be profitable to do so. This behavior is more apparent, in the sense that dealers’ markups are larger, when there is no adverse selection, the dispersion of clients’ liquidity shocks is larger, or the volatility of the asset payoff is larger.

the same patterns for all configurations (see Figures [OA.2](#), [OA.3](#), and [OA.4](#)). That is: (i) quoted spreads are not competitive (realized spreads are far above zero) and (ii) they become less competitive as σ increases. Moreover, for each value of σ , AMs’ average realized spreads are larger when there is no adverse selection (the red curve is above the blue curve for realized spreads). The patterns for quoted spreads are less pronounced for $\alpha = 0.1$ (in this case, the average quoted spread is almost insensitive to σ).

³³The updating rule (8) is adequate for what [Sutton and Barto \(2018\)](#) call an “episodic task”: an optimization problem with a clear beginning and end, here a one-period game. Other papers in the literature typically use the rule $q_{m,n,t} = \alpha[\pi_{n,t} + \gamma \max_{m'} q_{m',n,t}] + (1 - \alpha)q_{m,n,t-1}$, which is meant for computing the value of an action in an infinite horizon problem like an infinitely repeated game. We can of course implement such a rule in our experiments and ask each dealer to maximize the discounted value of all future episodes. As long as the algorithm cannot condition on past history this makes no difference. However, because of the new term $\gamma \max_{m'} q_{m',n,t}$ there is less update after each episode, this makes learning slower and the final greedy price higher, keeping all other parameters constant.

In Section 4.2.1, we show that these instances share a common characteristic: They each correspond to environments in which, other things equal, the variance of dealers’ gains from undercutting their competitor is large. As a result, it is more difficult for AMs to accurately assess the average profit that could be achieved by improving prices. This simple intuition provides a unified explanation for why outcomes are systematically less competitive in some specifications of the market making game. In Section 4.2.2, we show that another factor hinders AMs’ ability to accurately estimate the benefit of lowering their prices, namely the fact that their competitors’ experimentation creates non-stationarity in the environment they face. Finally, in Section 4.2.3, we discuss the role of the choice of the experimentation rate.

4.2.1 More Volatile Profits Hinder Competition

To build up intuition, we start with an example. Suppose $v_H = 4$, $v_L = 0$, $\mu = 0.5$, and $\sigma = 5$ (as in the baseline case) and assume that the AMs eventually settle on a price of 5 (the modal price in the experiments). At this price, the true expected profit for the AM is 0.3. If instead, one of the two AMs (henceforth AM1) undercuts by choosing a price of 4.9, its true expected profit is larger and equal to 0.59. If AM1 knew this, as assumed in the calculation of the Glosten-Milgrom equilibrium, it would undercut. However, in our experiments AM1 never observes the expected profit it obtains at a given price. Rather, it observes realized profits and estimates the profitability of each price from these realizations.

In our example, AM1’s possible realized profits when it undercuts are (i) 0 if the client decides not to trade, (ii) 0.9 if the client buys and $\tilde{v} = v_H = 4$, (iii) 4.9 if the client buys and $\tilde{v} = v_L = 0$. The standard deviation of AM1’s profit is then 1.71 if it undercuts, which is large compared to the expected profit of 0.59. Learning that this expected profit is indeed larger than AM1’s (noisy) assessment of its average profit at a price of 5.0 may require experimenting price 4.9 for a large number of episodes.

This learning problem repeats itself at each price. If AM2 uses an initial price of a_2 , AM1 will “explore” and try many different prices. Prices above a_2 will always give a payoff of zero and be gradually eliminated. Prices below a_2 are not guaranteed to give a positive profit, but eventually AM1 will learn to play a price $a_1 < a_2$. As AM2 now makes zero profit, it will eventually try other prices. Again those above a_1 will gradually be eliminated, and AM2 will eventually learn

to undercut AM1, etc. This undercutting process resembles the familiar process of elimination of dominated strategies. However, it is noisy and gradual, and cannot go on forever, because the algorithms are not programmed to experiment for an infinite number of periods. How many times they undercut each other, and hence how low the final price is, then depends on how quickly they learn to undercut. As argued above, a more noisy profit from undercutting will make learning more difficult, and thus lead to higher long-run prices.

The previous example and reasoning suggests the following heuristic. With a slight abuse of notation, in the case $N = 2$ denote $\Pi(a_1, a_2)$ the (random) profit of AM1 if AM1 quotes a_1 and AM2 quotes a_2 . The expected gain for AM1 from undercutting AM2 is $\mathbb{E}[\Pi(a_2 - tick, a_2) - \Pi(a_2, a_2)]$, and the variance of this gain is $\mathbb{V}[\Pi(a_2 - tick, a_2)] + \mathbb{V}[\Pi(a_2, a_2)]$. We posit that, for a given expected gain from undercutting, a change in the parameterization of the market-making game that increases the variance impairs the AMs' learning to undercut, and eventually leads to less competitive outcomes. Conversely, for a given variance, a change in the parameters that increases the expected gain from undercutting should lead to more competitive outcomes.

Note that, while this simple heuristic may seem intuitive, only the sign but not the magnitude of $\mathbb{E}[\Pi(a_2 - tick, a_2) - \Pi(a_2, a_2)]$ would play a role with rational and risk-neutral players. The variance would play no role.

[INSERT FIGURE 9 ABOUT HERE]

This simple heuristic goes a long way in explaining observations in our experiments. Consider the observation that the adverse selection case leads to more competitive outcomes than the no adverse selection case. Figure 9 shows the distribution of AM1's realized profit when $a_2 = 5$ and AM1 undercuts AM2 by one tick ($a_1 = 4.9$) in both cases. The likelihood that the client does not buy at price $a_1 = 4.9$ is the same whether there is adverse selection or not (we have designed the no adverse selection case for this purpose; see Section 2.2). However, conditionally on a trade taking place, a small realized profit (0.9) is more likely than a large profit (4.9) when there is adverse selection. This is simply because adverse selection raises the likelihood of adverse outcomes for the AMs. As a result, adverse selection shifts profits toward zero and reduces the variance of AMs' profits. In the case considered in Figure 9, the variance of AM1s' profit at $a_1 = 4.9$ is 2.93 when there is no adverse selection versus 1.81 when there is.

Thus, adverse selection reduces the variance of profits for AM1 when it undercuts AM2. Our heuristic implies that this can lead to more competitive outcomes, which is indeed what we observe experimentally.³⁴ Moreover, this property holds for any price posted by AM1 and parameterization of the market making game, holding a_2 constant. Thus, our heuristic can explain why AMs' rents are lower with adverse selection in all the parameterizations we used in our experiments.

To understand how the parameterization of the market making game affects the variance of AM1's profit from undercutting, we compute $\mathbb{V}[\Pi(a_1, a_2)]$ analytically in Appendix A.3 for each pair of prices (a_1, a_2) . We denote this variance by $\text{Var}_{n.as}^1(a_1, a_2)$ when there is no adverse selection and $\text{Var}_{as}^1(a_1, a_2)$ when there is. We establish the following properties.

First, for a given parameterization of the environment, $\text{Var}_j^1(a_1, a_2)$ ($j \in \{as, n.as\}$) is larger when $a_1 < a_2$ than when $a_1 = a_2$ for a_1 close enough to a_2 . The reason is simply that when AM1 matches AM2's offer, profits are split between the 2 AMs, which makes their variance smaller relative to their variance at a price slightly better than a_2 . Thus, intuitively, learning the true average payoff obtained when matching a_2 requires experimenting less than for learning the true average payoff obtained by undercutting a_2 's offer. Moreover, $\text{Var}_j^1(a_1, a_2) = 0$ ($j \in \{as, n.as\}$) when $a_1 > a_2$ since the client never trades with AM1 at such prices. Thus, learning not to post prices above a_2 should be much quicker than learning to post prices below a_2 . This pushes AMs to settle on identical prices above competitive levels, exactly what we observe (see Figure 3).

Second, for $a_1 \leq a_2$, $\text{Var}_{as}^1(a_1, a_2) < \text{Var}_{n.as}^1(a_1, a_2)$. Thus, for any parameterization of the market making game, the variance of AM1s' profit when it undercuts or matches AM2's offer (holding a_2 constant) is smaller in the environment with adverse selection, as claimed above.

Third, for $a_1 \leq a_2$, $\text{Var}_{n.as}^1(a_1, a_2)$ and $\text{Var}_{as}^1(a_1, a_2)$ increase with the variance of clients' liquidity shocks. Indeed, such an increase raises the likelihood of a trade and therefore increases the dispersion of realized profits, as explained previously. Last, an increase in the volatility of the asset payoff (Δ_v) increases the variance of AM1s' profits in the absence of adverse selection because it increases the range of possible realized profit for AM1 when the client decides to buy the asset (the difference between realized profits in the high and low realizations for v is Δ_v). This effect is weaker in the case with adverse selection because an increase in the volatility of the asset payoff also raises the

³⁴The expected gain from undercutting is larger when there is no adverse selection, which goes in the other direction. Hence, in this comparison it appears that the variance effect dominates.

cost of adverse selection, which as explained before tends to reduce the dispersion of profits.

Thus, our heuristic can explain our experimental observations: less competitive outcomes (larger average realized spreads) in environments without adverse selection (Figures 4 and 5), with larger values of σ (Figure 4), and with larger values of Δ_v (see Figure 5). On that last figure, our heuristic also explains why Δ_v has a weaker positive effect on AMs’ average realized spreads in experiments with adverse selection.

4.2.2 Strategic Uncertainty Hinders Competition

To establish the previous properties, we computed the variance of AM1’s profit holding a_2 constant. However, in the experiments, AM2 changes its price randomly over time, due to idiosyncrasies in its trading history and experimentation choices. As a result, holding the parameterization of the market-making game constant, the distribution of AM1’s profits at a given price is non-stationary.

We refer to this additional layer of learning complexity as “strategic uncertainty”. Intuitively, strategic uncertainty should slow down AMs’ ability to learn to undercut as well. One way to test this conjecture is to run experiments in which the price set by one AM is constant over time and compare the outcomes to those obtained when both AMs use Q-learning algorithms.

We have done so in the baseline case fixing AM2’s offer at 5.0 in every period, i.e., about the level of the average greedy price after T episodes in our baseline experiments (see Figure 3). We report the results of these experiments in the Appendix A.4. As conjectured, we observe that AM1 learns to undercut more quickly. Indeed, it takes “only” 46,043 episodes for the average greedy price for AM1 over $K = 1,000$ experiments to reach 4.9 and, after $T = 1,000,000$ episodes, the modal greedy price for AM1 is 4.9 (Figure 12). Thus, strategic uncertainty also slows down learning for the AMs.

4.2.3 The Role of Experimentation

The previous findings do not imply that Q-learning algorithms cannot learn to be competitive. They just mean that learning to be competitive is slow due to strategic uncertainty, especially when profits are more volatile. To overcome these obstacles, algorithms should experiment more intensively and for a longer duration. For instance, we show in the Online Appendix OA.2, that if AMs’ experimentation rate (ϵ_t) never falls below some threshold then outcomes are much more

competitive.

However, it does not follow that, in reality, agents designing the AMs will choose high experimentation rates. Indeed, experimenting is costly: By definition, while experimenting, the algorithms might pick actions that are truly dominated (deliver low average payoffs). As a result, AMs’ average total payoff can decrease with higher experimentation rate. We provide an example in the Online Appendix (Section OA.3).

Yet, one may anticipate that, in choosing the hyperparameters of their algorithms, dealers might engage in a race to the bottom, anticipating that if they do not experiment much, their competitor will and will eventually learn to capture all demand. This argument requires a meta-model of hyper-parameter choices by AM. We study such a model in Section 6 and show that this intuition does not hold. Instead, we do find that dealers are likely to “stick” to parameters with relatively low experimentation.

5 Implications

In this section, we derive additional testable implications, focusing on patterns that specifically arise from the way AMs learn to set their quotes, in the sense that they differ from the Glosten-Milgrom benchmark. Our goal is to further identify the “signature” that reinforcement learning should leave in the data.

5.1 Entry and Liquidity

Brogaard and Garriott (2019) find empirically that average bid-ask spreads gradually decline with entry of new high frequency market makers, for a sample of Canadian stocks (see their figure 1 for instance). They observe that this pattern is difficult to explain with standard models of price competition since these models imply that, for homogeneous products, two price competitors are sufficient to reach competitive prices.

[INSERT FIGURE 6 ABOUT HERE]

As shown in Figure 6, in line with Brogaard and Garriott (2019)’s empirical findings, we also observe that AMs’ average quoted and realized spreads also decline gradually with the number of

AMs.³⁵ This pattern cannot be explained by the standard economic analysis of the market-making game, which predicts that the entry of additional dealers has no effect once two dealers compete. However, this pattern aligns with our interpretation of AMs’ behavior in Section 4.2.1. Specifically, as the number of AMs increases, the difference in expected gains from undercutting becomes larger, all else being equal, because AMs’ profits—when tied at the same price—are divided among more dealers. Additionally, the variance of this profit also decreases for the same reason. Consequently, the AMs’ speed of learning increases and outcomes are more competitive.

5.2 Tick Size and Competition

Another interesting question is how algorithms react to changes in the tick size. The tick size in stock markets has been the subject of heated debates, both in the U.S and in Europe and was reduced recently in the U.S from one penny to half a penny. In this section, we show that reducing the tick size does not necessarily lead AMs to set more competitive prices.³⁶

Specifically, we conduct experiments with the baseline parameters and different values of the tick size $tick \in \{0.01, 0.05, 0.10, 0.50, 1.00\}$. The range of the price grid remains the same as in the baseline experiments (in which $tick = 0.10$), with prices ranging from $1.00 + 1 \times tick$ to $15.00 - 1 \times tick$. For each tick size, we conduct $K = 1,000$ experiments (with adverse selection) and report the average values of the quoted spread and the realized spread in the final episode $T = 10^6$ in Figure 7 (panels A and B).

[INSERT FIGURE 7 ABOUT HERE]

Panel A shows that, holding the experimentation rate constant ($\beta = 8.10^{-5}$), AMs’ average quoted and realized spreads are hump-shaped in the tick size.³⁷ In particular, as the tick size declines from 1 to 0.1, AMs’ average quoted spreads increase while they decrease (by a small amount) when the tick size is reduced further.

We think that this pattern is due to two forces. First, as the tick size is reduced, the expected benefit of undercutting by one tick is larger because the undercutter gets a large increase in market

³⁵This graph is obtained by repeating our experiments with various number of AMs, ranging from 3 to 10.

³⁶Cartea *et al.* (2022b) also studies the effect of reducing the tick size in a market making game between algorithms. However, they consider a set-up without adverse selection and in which products sold by dealers are differentiated.

³⁷In contrast, a reduction in the tick size reduces both the quoted and realized spreads in the Glosten-Milgrom equilibrium (dashed lines in Figure 7).

share (the expected demand from a client) at the cost of an increasingly smaller reduction in its price. This is in line with the standard intuition for why reducing the tick size leads to more competitive outcomes with price competition.³⁸

Second, holding AMs' experimentation rate constant, AMs have more opportunities to experiment with a given price when the tick size is large since their choice set is smaller. As a result, it takes them fewer episodes to eliminate high prices and learn to undercut their competitor.

The first force increases the expected benefit of undercutting when the tick size is reduced, which leads to more competitive prices. The second force works in the opposite direction. To show this, we run experiments in which we increase the AMs' experimentation rate when the tick size is reduced, in such a way that the average number of times each AM experiments a price on the grid is identical across treatments with different tick sizes.³⁹

In this case, we observe (see Panels C and D in Figure 7) that the average quoted and realized spreads decline when the tick size is reduced. Thus, when we neutralize the second force, the first one leads to more competitive outcomes, as predicted. This finding highlights again the importance of the choice of the AMs' hyper-parameters: The non-monotonic effect of the tick size on dealers' spreads stem from the fact that their experimentation rate does not adjust when the tick size is changed.

Thus, if dealers using pricing algorithms do not re-parametrize their algorithms when the tick size changes, one could observe an increase in average spreads when the tick size is reduced. This insight is new and important for policy-making since it implies that a too small tick size is unlikely to minimize trading costs for liquidity demanders and maybe more important, it implies that the choice of the tick size depends on how algorithms are parametrized.

³⁸While intuitive, in the standard game-theoretic treatment of the Bertrand game with a positive tick, only ordinal comparisons between expected profits play a role to determine whether undercutting is profitable. That is, competitors undercut each other as long as it increases expected profits, no matter how small the increase is.

³⁹To do so, we rely on the following heuristic. For a given *tick*, the total number of entries in each AM's Q-matrix is $(14/\text{tick}) - 1$. For a given β , the expected number of episodes with experimentation is $\sum_{t=1}^{\infty} e^{-\beta t} = \frac{e^{-\beta}}{1-e^{-\beta}}$. Thus, the expected number of times each AM is going to experiment a given price is $\frac{e^{-\beta}}{1-e^{-\beta}} \times \frac{\text{tick}}{14-\text{tick}}$. When we vary the tick size, we vary β so that this quantity remains equal to 89.92, its value in the baseline case ($\text{tick} = 0.1$ and $\beta = 8.10^{-5}$).

5.3 Rockets and Feathers

In this section we show that AMs’ reaction to symmetric shocks to their adverse selection cost is asymmetric: They raise their spreads fast after an increase in their cost while not changing them after a decrease. This behavior is similar to the “rocket and feathers” phenomenon, that is, the tendency for output prices to react asymmetrically to shocks to input prices in some markets.⁴⁰

More specifically, we conduct the following experiment. We first run a simulation with $T_1 = 1,000,000$ episodes, with the baseline parameters (in particular, $\Delta_v = 4$). Then, for episodes between $T_1 + 1$ and $T_1 + 1000$, we simulate a temporary shock to adverse selection by changing Δ_v to a different value Δ'_v (adverse selection costs increases with Δ_v ; see Figure 2). Afterwards, we revert to the initial value of Δ_v and continue the simulation until $T = 2,000,000$ episodes. We use three values of Δ'_v : $\Delta'_v = \Delta_v = 4$ (“Placebo” value) ; $\Delta'_v = 7$ (positive adverse selection shock) ; $\Delta'_v = 1$ (negative adverse selection shock). In each case we conduct $K = 1,000$ experiments.

[INSERT FIGURE 8 ABOUT HERE]

Panel A of Figure 8 shows the dynamics of the quoted spread over the 2 million episodes for the positive adverse selection shock, averaging over the K experiments. Panel B zooms in around the shock in $T_1 + 1$, and shows episodes from $T_1 - 1000$ to $T_1 + 10000$. We observe on Panel B that the AMs quickly adjust their spreads upwards only a few hundred episodes after the shock. Namely, the spread increases from an average of 3 to a value slightly below 4. When the shock is over the AMs gradually decrease their spreads again.

Given the design of our experiments, there is virtually no experimentation after episode T_1 . Yet, the AMs keep adjusting their Q-matrix because the learning parameter, α , is constant. For this reason, they quickly learn to increase their spreads following a positive shock to adverse selection costs. Indeed, at the price they were playing for many episodes before T_1 , they are now obtaining smaller profits. As a result, the Q-value of this price is decreasing after T_1 and it does so faster if α is large (see the updating rule (8)). For instance, suppose $\alpha = 0.01$. One-hundred episodes after T_1 , the q-value of the price at which AMs eventually settled before T_1 only account for $(1 - \alpha)^{100} = 0.99^{100} \simeq 37\%$. In other words, 73% of this value is determined by the realization of AMs’ profits after T_1 . This is typically sufficient to learn that the price used before T_1 is no

⁴⁰See, for instance, [Peltzman \(2000\)](#) for product markets and [Green et al. \(2010\)](#) for securities markets.

longer profitable and switch to a higher price which has a higher Q-value. This ability of Q-learning algorithms to react to a negative temporary shock is one reason why the learning rate, α , is often constant in practice.

To verify that the patterns in Panel A of Figure 8 are truly driven by the shock to Δ_v , Panel C of Figure 8 plots the evolution of the quoted spread when Δ_v remains unchanged from T_1 to T (a placebo test). We observe that, in this case, the quoted spread remains constant on average between T_1 and T . Thus, the rapid increase in AMs' spreads when Δ_v rises is indeed driven by the increase in adverse selection costs.

Finally, Panel D considers the case in which Δ_v drops to 1, that is, a negative shock to adverse selection. In contrast to what happens with a positive shock, we observe no change in quoted spreads following this shock, as in the placebo case.

Thus, the reaction of AMs' quotes to symmetric shocks to their cost of adverse selection is asymmetric. The reason is as follows. First, as explained above, following a transient positive shock on Δ_v , the Q-value of the price on which AMs settled before T_1 quickly *decreases* and at some point another, higher, price dominates. In contrast, following a transient negative shock on Δ_v , the Q-value of the price on which AMs settled before T_1 quickly *increases*, so that the shock *reinforces* AMs' choice to post this price.

Of course, one could consider other shocks that cause variations in AMs' profits. We conjecture that the same asymmetry would be observed. The reason is that, by their very nature, reinforcement learning algorithms are more likely to select actions associated with higher profits. This characteristic creates an asymmetry between negative and positive profit shocks, leading to a quick reaction to positive shocks but a muted or delayed reaction to negative shocks. This observation suggests another testable implication: algorithms should adjust their prices more quickly following profit-reducing shocks than after symmetric profit-enhancing ones.

5.4 Bid-Ask Spread Dynamics

So far, we have considered an environment in which AMs play T independent market-making games. These games are independent because the realizations of the asset payoffs across episodes are i.i.d. As a result, clients' demands are independent across episodes. In this section, we relax this assumption by introducing a market-making game with two clients who arrive sequentially.

This version of the market-making game is more complex since the first client’s decision conveys information about the asset payoff. This information can then be used to set prices for the second client. This extension of our baseline setup enables us to study concepts such as price discovery and price impact. In particular, we will show that AMs using reinforcement learning lead to novel predictions regarding the dynamics of prices in the presence of adverse selection.

For this analysis, we assume that before the asset payoff is revealed, dealers receive orders from two different buyers who arrive sequentially in periods $\tau = 1$ and $\tau = 2$. The valuation of the buyer in period τ is $\tilde{v}_\tau^C = \tilde{v} + \tilde{L}_\tau$, where \tilde{L}_1 and \tilde{L}_2 are independent and normally distributed with mean zero and variance σ^2 .

As a benchmark, we first derive the two-period Glosten-Milgrom prices. Denote the market makers’ belief about the likelihood that $v = v_H$ prior to the arrival of the τ^{th} client by μ_τ . Thus, $\mu_1 = \mu$. At the end of the first trading round, there are two possible trading histories: (i) a trade at price a_1^{\min} , giving a belief we denote $\mu_2(1, a_1^{\min})$; (ii) no trade at price a_1^{\min} , giving a belief we denote $\mu_2(0, a_1^{\min})$. Conditionally on μ_τ , one can derive dealers’ expected profits in periods $\tau = 1$ and $\tau = 2$ exactly as in the one-period case. The competitive price in period τ , a_τ^* , is given by (6) with $\mu = \mu_\tau$. The unique Nash equilibrium of the two period market making game is such that, in each period, at least two AMs post a_τ^* .

We show in the Appendix OA.1 that these Glosten-Milgrom prices satisfy two properties: (i) $a_2^* > a_1^*$ if there is a trade in $\tau = 1$, and $a_2^* < a_1^*$ otherwise and (ii) $\mathbb{E}[a_2^*] < a_1^*$. The first property reflects the fact that the first client’s decision contains information about v since, other things equal, a buy is more likely if $v = v_H$ than if $v = v_L$. Thus, as dealers are bayesian, they update their beliefs about v upward after observing a buy from the first client and downward after observing no trade ($\mu_2(1, a_1^{\min}) > \mu_2(0, a_1^{\min})$). Consequently, the model implies that the Glosten-Milgrom price increases after a buy and decrease otherwise. This well-known property of the Glosten-Milgrom equilibrium has received a lot of attention in the literature and is one of the foundation of so called price impact regressions (regression of price changes on order flow; see [Glosten and Harris \(1988\)](#)). The second property means that the ask price gets closer to the unconditional expectation of \tilde{v} over time, or in other words the spread charged by the market-maker decreases. Over time market-makers are learning and are, in expectation, less exposed to adverse selection.⁴¹

⁴¹See [Glosten and Putnins \(2020\)](#) for a study of the welfare implications of this point.

We now study how AMs set their quotes, as we did in Section 3. The key difference is that each episode features two clients arriving sequentially with the same common value, \tilde{v} . To allow the algorithms to react to the occurrence of a trade in period 1, we let them keep track in each episode of the “state” they are in, and let them play an action that depends on the state.⁴² For brevity, here, we just outline how we program the algorithms in the 2-player case. A more precise and general treatment is given in Appendix OA.1.

For each AM_{*n*} ($n \in \{1, 2\}$) and episode t , we denote $s_{n,t} \in \{\emptyset, NT, 0, \frac{1}{2}, 1\}$ the state the algorithm finds itself in. The states are defined as follows: (i) $s_n = \emptyset$ in the first period; (ii) $s_n = NT$ in the second period if “No Trade” took place in the first; (iii) $s_n = 0$ in the second period if there was a trade in the first period, but AM_{*n*} did not trade; (iv) $s_n = \frac{1}{2}$ in the second period if there was a trade in the first period, and both AMs shared the market; (v) $s_n = 1$ in the second period if there was a trade in the first period, and AM_{*n*} sold one share.

This partition of the state space implies that each algorithm keeps track both of (i) whether a trade took place (which is important to analyze the impact of order flow on prices) and (ii) of its inventory after period 1 (e.g., $s_n = \frac{1}{2}$ indicates a short position of $-\frac{1}{2}$ for AM_{*n*}). The latter is important: As \tilde{v} is realized only at the end of the second period, the algorithm cannot know how profitable the first-period trade was before the end of the second period. Hence, the algorithm needs to keep track of its inventory, and learn what is the value of being in a state with a short position vs. a state with a zero inventory.⁴³ To do this, each AM relies on a Q-matrix $\mathbf{Q}_{n,t} \in \mathbb{R}^{M \times 5}$, in which each line corresponds to a different price and each column to a state, ordered as in the previous paragraph.

We then run simulations as in Section 3.2 and compute the average prices across $K = 1,000$ experiments in the last episode $T = 10^6$. As we are interested in testing whether AMs learn to react to the order flow, we aggregate states $s_{1,T} \in \{0, \frac{1}{2}, 1\}$ into a “Trade” state and denote \bar{a}_2^T the average best quote in period 2 conditionally on a trade occurring in period 1. Symmetrically, \bar{a}_2^{NT} denotes the average best quote in period 2 if no trade occurred in period 1. Finally, \bar{a}_1 denotes the

⁴²Q-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. See [Leach and Madhavan \(1993\)](#) for the analysis of a monopolist’s optimal behavior in the two-period market making game.

⁴³Using 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 an 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).

average best quote in period 1 and \bar{a}_2 the average best quote in period 2 (unconditionally, that is, whether a trade occurred or not in period 1).

Figure 10 plots \bar{a}_1 , \bar{a}_2^T , and \bar{a}_2^{NT} for different values of σ . We observe that $\bar{a}_2^T > \bar{a}_1 > \bar{a}_2^{NT}$, as in the Glosten-Milgrom benchmark. However, after a buy, AMs overreact relative to the Glosten-Milgrom prices. For instance, when $\sigma = 5$, the Glosten-Milgrom price should increase from 2.8 to 3.4 after a buy. Instead, in the experiments, the average price increases from 4.80 to 5.80. Hence, AMs charge even larger markups for the second client. Moreover, in the Glosten-Milgrom benchmark this revision should become smaller when the dispersion of clients' liquidity shocks increases since the first's client buy is then less informative. Instead, the revision becomes larger in the experiments. Conversely, when no trade occurs in the first period, the second-period price is almost equal to the first-period price while it is significantly smaller in the Glosten-Milgrom benchmark. Thus, AMs underreact to the absence of trade relative to the Glosten-Milgrom prices.

These patterns imply that AMs extract even larger rents on average from the second period client than the first and these rents increase with the dispersion of liquidity shocks, as in the one period case. This can be seen on Figure 11, which plots the average best quote in the second period \bar{a}_2 , unconditionally on whether a trade occurred in period 1. While in the Glosten-Milgrom benchmark prices should on average decrease in the second period, the opposite happens in the experiments.

[INSERT FIGURES 10 and 11 ABOUT HERE]

The fact that AMs need time to learn noisy payoffs is again important to understand the distance between the experimental results and the Glosten-Milgrom benchmark. There are two effects. The first effect is that the variance of \tilde{v} in period 2 conditional on the outcome of period 1 is smaller than the unconditional variance in our setting.⁴⁴ This effect should reduce AMs' rents in the second period and therefore make their prices closer to the Glosten-Milgrom prices than in the first period. However, there is a countervailing effect, which seems to dominate in our experiments: the algorithms have fewer opportunities to learn about the average profits of their actions for the second client than for the first client. Indeed, remember that the AMs learn state by state. They face the "period 1" state in each of the 10^6 episodes, but in period 2 they can be in 4 different states. For

⁴⁴Indeed, the conditional variance of the asset payoff in period τ is $Var_\tau(\tilde{v}) = \mu_\tau(1 - \mu_\tau)\Delta_v^2$. As $\mu_1 = 0.5$ in our experiments, we have $Var_2(\tilde{v}) < Var_1(\tilde{v})$.

any first period price, the probability of a trade is less than 50% and decreases with the price (e.g., to 30% for $a_1 = 4.80$). Thus, the AMs have more than 500,000 episodes to learn how to set their quotes after no trade, whereas after a trade they have fewer than 500,000 episodes to learn, and the learning is split across three different states. This makes it particularly difficult for AMs to learn to undercut each other after a trade. As in the one-period case, this feature (lacks of experimentation) leads to high prices.

These experimental results give insights into how competition between AMs can be spotted in the data. They imply that quotes will tend to over-react to the order flow (here a buy). This means that the change in prices following a buy or a sell should partially revert. Such patterns have been found for long in existing empirical studies and are usually attributed to order processing costs or inventory holding costs for market makers. Our experiments suggest that they could become more prevalent as quotes are posted by algorithms, reflecting algorithms’ imperfect learning of the benefits of undercutting. More generally, quoted spreads and realized spreads should tend to widen after histories that are more rarely observed, or even simply over time.⁴⁵

6 Choosing Algorithms’ Hyper-Parameters

AMs’ long-run prices do not form a Nash equilibrium, meaning dealers are “leaving money on the table.” However, this observation is based on exogenous values of the hyperparameters α and β . A natural question is whether it holds if dealers choose these parameters themselves. Dealers may try to parameterize their algorithms to undercut their opponents, potentially leading to a competitive outcome. In this section we analyze this possibility and show why it does not hold.

6.1 Dealers do not choose competitive algorithms

We run experiments where dealers (AMs’ designers) select the hyperparameters $\alpha \in \{\alpha_l, \alpha_m, \alpha_h\}$ and $\beta \in \{\beta_l, \beta_m, \beta_h\}$, with $\alpha_l = 0.001, \alpha_m = 0.01, \alpha_h = 0.1; \beta_l = 5 \cdot 10^{-6}, \beta_m = 8 \cdot 10^{-5}, \beta_h = 3.2 \cdot 10^{-4}$. Dealers can deviate from the baseline (α_m, β_m) .

We examine which of the 81 parameter pairs are “stable,” as defined below. As noted in Section 3, dealers coding Q-learning algorithms have limited knowledge of the trading environment and

⁴⁵This type of behavior might lead to sudden evaporation of liquidity after events that have been rarely encountered by algorithms and potentially explain flash crashes.

can only learn expected profits through experimentation. We run the following experiment. Both dealers initially use (α_m, β_m) for $K = 1,000$ experiments, recording total profit over T episodes. Then, dealer 1 deviates to (α', β') for K' experiments (dealer 2 does not deviate) and records total profit. A statistical test is conducted under the null hypothesis that (α', β') yields the same expected profit as (α_m, β_m) .

Results, detailed in Online Appendix OA.6, show that deviating to α_l or α_h never rejects the null (p-values ≈ 1), with lower payoffs (0.1034 per episode for α_l vs. 0.2834 for (α_m, β_m)). Keeping α_m but switching to β_l or β_h yields slightly higher payoffs but low statistical significance. With $K' = 100$, the p-value is 0.5. As K' increases, the p-value rises for (α_l, β_m) , indicating unprofitability, and falls to 0.06 for (α_h, β_m) at $K' = 500$, but rises to 0.29 at $K' = 1000$. Even after 1,000 tests, no alternative significantly increases profit, even at a weak 0.25 confidence level. If anything, slower learning (higher β , lower α) is preferable.

Thus, dealers do not parameterize their algorithms competitively. While higher α and lower β may boost short-term profit by undercutting opponents, the long-term effect is negative. Once AM1 undercuts AM2, AM2 makes no profit and eventually adjusts, forcing AM1 to share demand at lower prices. This long-run effect outweighs short-term gains, preventing competitive hyperparameter selection.⁴⁶

6.2 Discussion

Our conclusion from this exercise is that both dealers choosing (α_m, β_m) is a “stable” parameterization, in the sense that additional experimentation does not give any dealer a reason to adopt alternative parameters. In fact, even we the modelers cannot tell whether dealer 1 would be better off in expectation by adopting (α_m, β_h) . Moreover, as modelers we can conduct such an experimentation, but the agents themselves cannot, for two reasons. First, if dealer 1 has an incentive to experiment with alternative hyperparameters, then so does dealer 2, and each dealer cannot observe the algorithm of his opponent. This can only add considerable noise to the learning of “better” hyperparameters. Second, like in the original game, experimenting comes at the cost of using random hyperparameters, likely to lead to lower payoffs, and it is not optimal to experiment

⁴⁶Abada *et al.* (2022) finds a similar result when comparing Q-learning to an actor-critic algorithm, where the latter learns to undercut Q-learning but ultimately results in lower payoffs.

forever. Both arguments reinforce our point that both dealers have little reason to deviate from the baseline hyperparameters (α_m, β_m) .⁴⁷

Given our initial assumption that the dealers have no information about their environment, it is very difficult, and beyond the scope of this paper, to say more about the equilibrium choice of hyperparameters by the players. We conclude this section with a brief discussion of two proposals that have been made in the literature.

A first natural possibility is to assume that the agents choose α and β by using a second-layer of Q-learning. This solves the problem mentioned above, that in reality agents would have to experiment different values of α and β , with the other players experimenting at the same time. A limitation of this approach is that this second layer also needs hyperparameters, so that the same problem simply repeats itself at a higher level. See [Dou *et al.* \(2023\)](#) for an example of this approach.

A second approach is to look for a “Nash equilibrium” in hyperparameters. However, this requires the assumption that the players somehow know the exact expected profits for every choice of hyperparameters. This can be the case if the agents are able to conduct “offline” experiments, that is, they are able to conduct the same simulations as we do. This is the approach followed for instance in [Abada *et al.* \(2022\)](#). However, one then needs to explain why the agents restrict themselves to using Q-learning or other algorithms if they actually have complete information about the game that they play (in particular, why don’t they play a Nash equilibrium then?). [Compte \(2023\)](#) also follows this approach, with the assumption that agents know the entire structure of the game, but are for some reason (e.g., regulation, rules of the market place or trading platform) constrained to using a certain family of algorithms. While this is certainly a realistic assumption in some contexts, in our application to financial markets it is not clear why such a constraint would be present.

7 Conclusion

We study the prices posted by market makers using Q-learning algorithms in a standard market making game with adverse selection (similar to [Glosten and Milgrom \(1985\)](#)) and compare them

⁴⁷For completeness, we repeat the analysis for all 81 possible pairs of hyperparameters chosen by the dealers. We find 4 other pairs that are stable in the same sense (at the 0.25 confidence level): both dealers choose (α_m, β_l) , both dealers choose (α_m, β_h) , dealer 1 chooses (α_m, β_h) and dealer 2 chooses (α_m, β_m) , and conversely.

to the Glosten-Milgrom prices, obtained in the Nash equilibrium of the market making game. We find that, despite their simplicity and the challenge of an environment with adverse selection, our algorithmic market makers (AMs) behave in a realistic way: their quoted spreads reflect adverse selection costs and they update their quotes in response to the observed order flow. However, they also deviate from the Glosten-Milgrom prices in many important ways. In particular, their quoted spread are larger than the competitive spreads and their rents increase when adverse selection costs decrease. Moreover, they over-react to the order flow.

We argue that these findings stem from the fact that AMs receive a noisy feedback about the average profit of their actions (because of uncertainty in their client’s demand and the asset payoff) and this noise is larger when adverse selection is less intense. In response, AMs should experiment more in noisier environments. However, our experiments suggest that this would require very long training periods and that it may not even be optimal for agents designing AMs to do so (because experimentation is costly).

Overall, our results suggest that securities markets are a quite specific and particularly interesting application of recent research on competition between pricing algorithms. In particular, they raise the possibility that these algorithms may not lead to more competitive outcomes in assets that are risky but less exposed to adverse selection. They also suggest that these algorithms could be significantly less competitive when facing states that they rarely encounter (which may explain why variations in liquidity have become more extreme with the rise of algorithmic pricing). Future research could consider the robustness of our conclusions when more complex algorithms are used or when they are used in conjunction with some prior “model of the world”.

The fact that, in any given interaction, each AMM does not play the one-shot Nash best response to the quote posted by the other AMM might suggest that the algorithm would not survive when facing a rational Bayesian agent who does so. While an in-depth exploration of the question of humans vs. machines is beyond the scope of this paper, in the Online Appendix [OA.7](#), we show that a rational forward-looking dealer has no incentive to consistently undercut an algorithmic competitor, as this would trigger price experimentation, mutual undercutting, and drive the game to a zero-profit Nash equilibrium. The human would gain more by encouraging a collusive play, ensuring positive aggregate profit while capturing most of it. This aligns with [Werner \(2024\)](#), who finds that in human-machine Bertrand competition, humans adopt the machine’s collusive behavior

rather than forcing it to undercut.

References

- ABADA, I., LAMBIN, X. and TCHAKAROV, N. (2022). *Collusion by Mistake: Does Algorithmic Sophistication Drive Supra-Competitive Profits?* Working paper. 7, 38, 39
- ASKER, J., FERSHTMAN, C. and PAKES, A. (2023). The impact of artificial intelligence design on pricing. *Journal of Economics & Management Strategy*, **forthcoming**, 1–29. 7, 19
- BALDAUF, M. and MOLLNER, J. (2020). High-frequency trading and market performance. *The Journal of Finance*, **75** (3), 1495–1526. 8
- BANCHIO, M. and MANTEGAZZA, G. (2022). *Adaptive Algorithms and Collusion via Coupling*. Working paper. 7
- and SKRZYPACZ, A. (2022). Artificial intelligence and auction design. *Available at SSRN 4033000* 9. 7, 9
- BIAIS, B., FOUCAULT, T. and MOINAS, S. (2015). Equilibrium fast trading. *Journal of Financial Economics*, **116** (2), 292–313. 8
- BRAIN, D., DE POOTER, M., DOBREV, D., FLEMING, M., JOHANSSON, P., JONES, C., KEANE, F., PUGLIA, M., REIDERMAN, L., RODRIGUES, T. and OR, S. (2018). Unlocking the Treasury Market through TRACE. *FED Notes*. 1
- BROGAARD, J. and GARRIOTT, C. (2019). High-frequency trading competition. *Journal of Financial and Quantitative Analysis*, **54** (4), 1469–1497. 5, 29
- , HENDERSHOTT, T. and RIORDAN, R. (2014). High-Frequency Trading and Price Discovery. *The Review of Financial Studies*, **27** (8), 2267–2306. 1
- BUCHAK, G., MATVOS, G., PISKORSKI, T. and SERU, A. (2019). *Why is Intermediating Houses so Difficult? Evidence from iBuyers*. Tech. rep., NBER, Working Paper 28252. 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. 8
- CALVANO, E., CALZOLARI, G., DENICOLO, V. and PASTORELLO, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, **110** (10). 7, 24
- CARTEA, A., CHANG, P., MROZKA, M. and OOMEN, R. (2022a). Ai-driven liquidity provision in otc financial markets. *Quantitative Finance*, **22** (12), 2171–2204. 8
- CARTEA, Á., CHANG, P. and PENALVA, J. (2022b). *Algorithmic Collusion in Electronic Markets: The Impact of Tick Size*. Working paper. 7, 8, 30
- CHABOUD, A., DAO, A. and VEGA, C. (2019). *What makes HFTs tick? Tick size changes and information advantage in a market with fast and slow traders*. Tech. rep., Available at SSRN: <https://ssrn.com/abstract=3407970>. 1
- COMPETITION MARKET AUTHORITY (2018). Pricing algorithms. pp. 3–62. 7
- COMPTE, O. (2023). *Q-based Equilibria*. Working paper. 39
- CONT, R. and XIONG, W. (2023). Dynamics of market making algorithms in dealer markets: Learning and tacit collusion. *Mathematical Finance*, **forthcoming**. 7
- DOU, W., GOLDSTEIN, I. and JI, Y. (2023). *AI-Powered Trading, Algorithmic Collusion, and Price Efficiency*. Tech. rep., Available at SSRN: <https://ssrn.com/abstract=4452704>. 7, 8, 9, 24, 39
- EASLEY, D. and KIEFER, N. M. (1988). Controlling a stochastic process with unknown parameters. *Econometrica*, **56** (5), 1045–1064. 9

- and RUSTICHINI, A. (1999). Choice without beliefs. *Econometrica*, **67** (5), 1157–1184. 8
- FUDENBERG, D. and LEVINE, D. (1998). *The Theory of Learning in Games*. Cambridge (Mass.): MIT Press. 9
- and LEVINE, D. K. (1993). Self-confirming equilibrium. *Econometrica*, **61** (3), 523–545. 9
- GITTINS, J. C. (1979). *Bandit Processes and Dynamic Allocation Indices*. Tech. Rep. 2. 9
- GLOSTEN, L. and PUTNINS, T. (2020). *Welfare Costs of Informed Trade*. Working paper. 34
- GLOSTEN, L. R. and HARRIS, L. E. (1988). Estimating the components of the bid/ask spread. *Journal of financial Economics*, **21** (1), 123–142. 34
- 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, 6, 8, 12, 39
- GOLDSTEIN, I., SPATT, C. S. and YE, M. (2021). Big Data in Finance. *The Review of Financial Studies*, **34** (7), 3213–3225. 1
- GREEN, R. C., LI, D. and SCHÄFFROFF, N. (2010). Price discovery in illiquid markets: Do financial asset prices rise faster than they fall? *The Journal of Finance*, **65** (5), 1669–1702. 32
- GUÉANT, O. and MANZIUK, I. (2019). Deep reinforcement learning for market making in corporate bonds: Beating the curse of dimensionality. *Applied Mathematical Finance*, **26** (5), 387–452. 7
- 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. 7
- 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. 18
- KYLE, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, **53** (6), 1315–1335. 7, 8
- LEACH, J. C. and MADHAVAN, A. (1993). Price experimentation and security market structure. *Review of Financial Studies*, **6** (2), 375–404. 35
- LIU, H. and WANG, Y. (2016). Market making with asymmetric information and inventory risk. *Journal of Economic Theory*, **163**, 73–109. 22
- MACKAY, A. and WEINSTEIN, S. (2022). *Dynamic Pricing Algorithms, Consumer Harm, and Regulatory Response*. Working paper. 7
- MENKVELD, A. and ZOICAN, M. (2017). Need for speed? exchange latency and liquidity. *Review of Financial Studies*, **30** (4), 1188–1228. 8
- OECD (2017). Algorithms and collusion: Competition policy in the digital age. pp. 1–72. 7
- O’HARA, M. (2015). High frequency market microstructure. *Journal of Financial Economics*, **116** (2), 257–270. 8
- PELTZMAN, S. (2000). Prices rise faster than they fall. *Journal of Political Economy*, **108** (3), 466–502. 5, 32
- POUGET, S. (2007). Adaptive traders and the design of financial markets. *The Journal of Finance*, **62** (6), 2835–2863. 9
- 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. 35

- SUTTON, R. and BARTO, A. (2018). *Reinforcement Learning: An Introduction*. Cambridge (Mass.): MIT Press. 15, 24
- TIROLE, J. (1988). *The Theory of Industrial Organization*. Cambridge, Massachussets: MIT Press. 13
- TSITSIKLIS, J. (1994). Asynchronous stochastic approximation and q-learning. *Machine Learning*, **16**, 185–202. 18
- WALTMAN, L. and KAYMAK, U. (2008). Q-learning agents in a cournot oligopoly model. *Journal of Economic Dynamics and Control*, **32** (10), 3275–3293. 7
- WATKINS, C. and DAYAN, P. (1992). Q-learning. *Machine Learning*, **8**, 279–292. 18
- WERNER, T. (2024). *Algorithmic and Human Collusion*. Tech. rep., Max PlankInstitute fir Human Development D`usseldorf Institute for Competition Economics. 40
- WILK, E. (2022). *Pricing Under Pressure: The Effect of Signal Corruption on the Gameplay of Pricing Algorithms*. Working paper. 7
- WUNDER, M., LITTMAN, M. L. and BABES, M. (2010). Classes of multiagent q-learning dynamics with epsilon-greedy exploration. In *ICML*, pp. 1167–1174. 7

Appendix

A.1 Figures

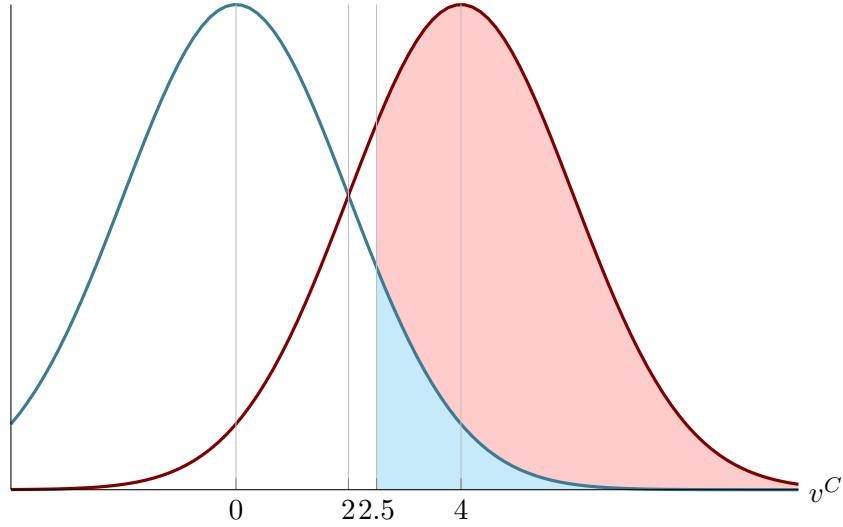


Figure 1: **Distribution of the client's valuation, \tilde{v}^C .** Parameter values are $v_H = 4$, $v_L = 0$, $\mu = \frac{1}{2}$, and $\sigma = 5$. If dealers' best offer is $a^{min} = 2.5$, the likelihood that the client buys the asset is given by (i) the blue area when $v = v_H = 4$ and (ii) the red plus blue area when $v = v_H = L$. The unconditional likelihood of a buy is therefore half the red area plus the blue area.

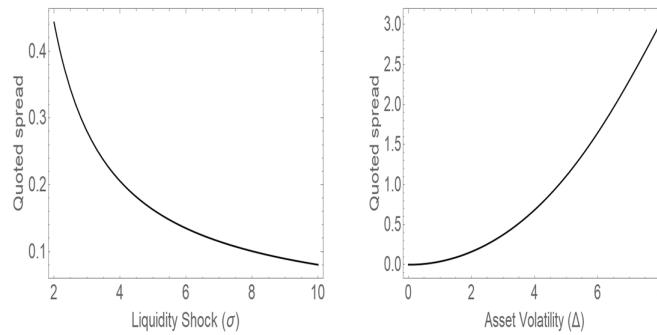
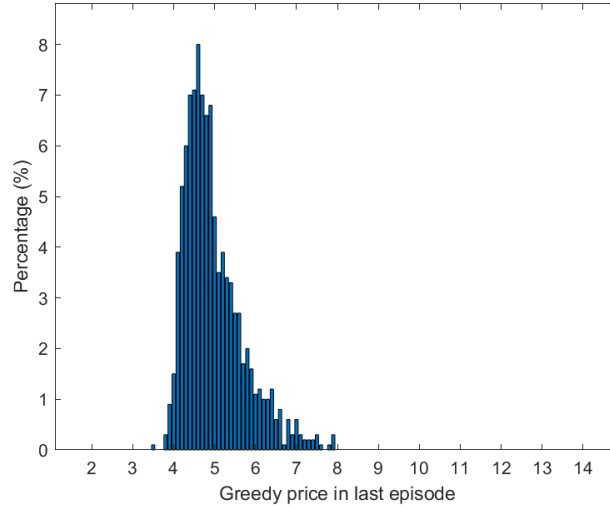


Figure 2: **Glosten-Milgrom Benchmark.** The figure shows the equilibrium quoted spread in the Glosten-Milgrom benchmark as a function of the variance of clients' liquidity shocks, σ (L.H.S) and the volatility of the asset payoff, Δ_v (R.H.S). Baseline parameters are $E_\mu(v) = 2$, $\mu = \frac{1}{2}$, $\Delta_v = 4$, $\sigma = 5$.

Panel A: Distribution of the greedy price of AM 1 in the last episode.

This panel shows a histogram of the greedy price of AM 1 in episode T : For each possible price a between 1.10 and 14.90 the bar indicates the percentage of the 1,000 experiments conducted in which $a_{1,T}^* = a$.



Panel B: Dynamics of the average greedy price of AM 1 for episodes 1 to T .

This graph shows for each episode t the average of AM 1's greedy price $a_{1,t}^*$ across the 1,000 experiments conducted. As a measure of dispersion, we also compute the standard deviation of $a_{1,t}^*$ across experiments and plot the average of $a_{1,t}^*$ plus/minus one standard deviation (with a 500-episode moving average for better readability).

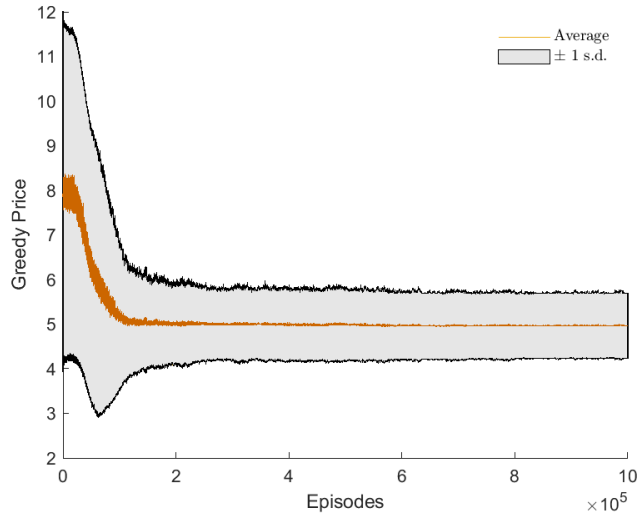
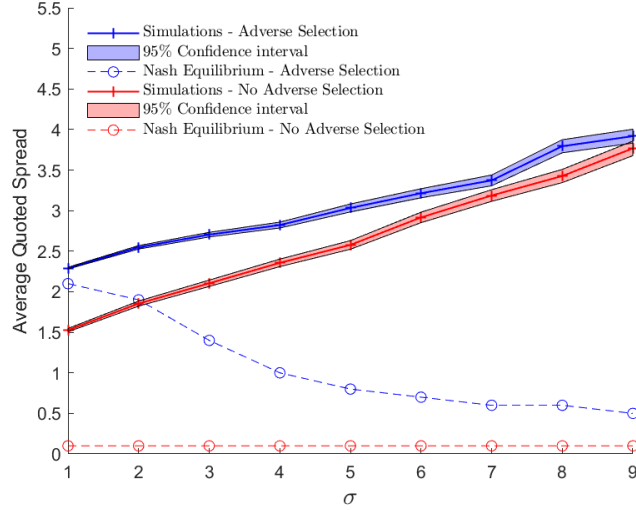


Figure 3: **Greedy price of AM 1 in the adverse-selection case, baseline parameters:** $\sigma = 5$, $\Delta_v = 4$, $N = 2$, $\mu = \frac{1}{2}$, $\mathbb{E}(v) = 2$, $T = 1,000,000$, and $K = 1,000$.

Panel A: Average Quoted Spread.

This graph plots the average over 1,000 experiments of the quoted spread, with 95% confidence intervals, both in the adverse-selection case and the no-adverse-selection case. The graph additionally plots the values of the quoted spread in both cases, in the Glosten-Milgrom benchmark of Section 2.3 (accounting for price discreteness).



Panel B: Average Realized Spread.

This graph plots the average over 1,000 experiments of the realized spread, with 95% confidence intervals, both in the adverse-selection case and the no-adverse-selection case. The graph additionally plots the values of the realized spread in both cases, in the Glosten-Milgrom benchmark of Section 2.3 (accounting for price discreteness).

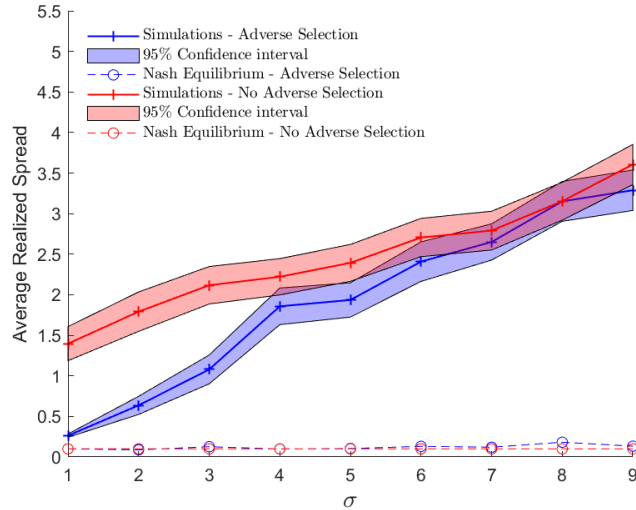
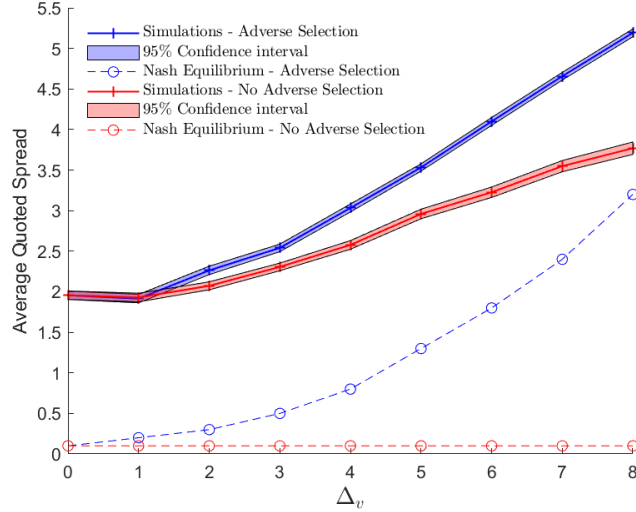


Figure 4: Average Quoted Spread $\bar{Q}S$ and Average Realized Spread $\bar{R}S$ in the adverse-selection case and the no-adverse-selection case, for different values of the dispersion of clients' liquidity shocks σ . The other parameters are $\Delta_v = 4$, $N = 2$, $\mu = \frac{1}{2}$, $\mathbb{E}(v) = 2$, $T = 1,000,000$, and $K = 1,000$.

Panel A: Average Quoted Spread.

This graph plots the average over 1,000 experiments of the quoted spread, with 95% confidence intervals, both in the adverse-selection case and the no-adverse-selection case. The graph additionally plots the values of the quoted spread in both cases, in the Glosten-Milgrom benchmark of Section 2.3 (accounting for price discreteness).



Panel B: Average Realized Spread.

This graph plots the average over 1,000 experiments of the realized spread, with 95% confidence intervals, both in the adverse-selection case and the no-adverse-selection case. The graph additionally plots the values of the realized spread in both cases, in the Glosten-Milgrom benchmark of Section 2.3 (accounting for price discreteness).

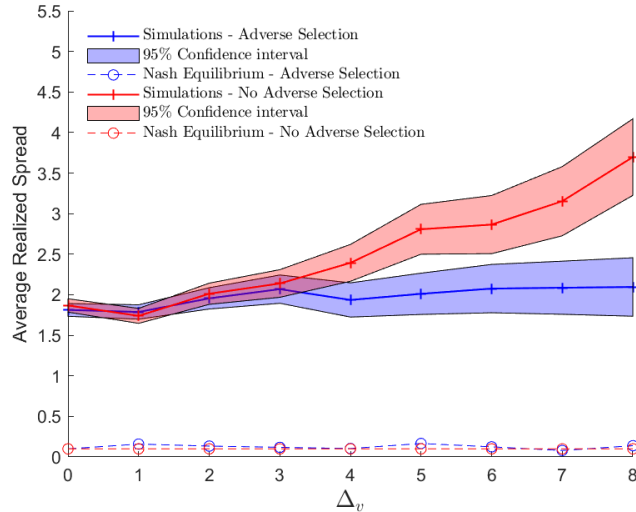
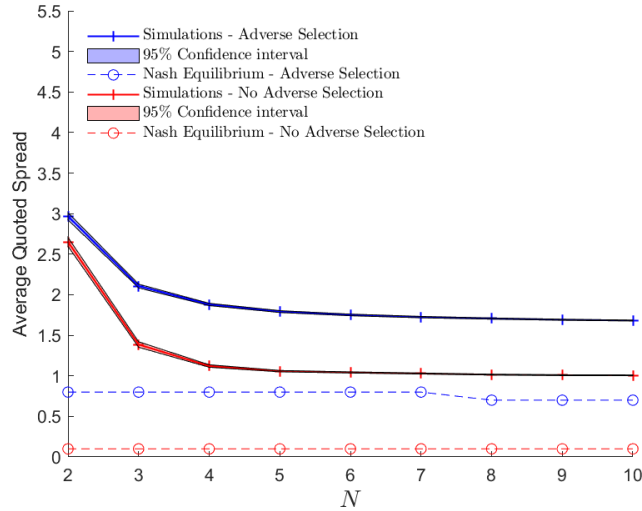


Figure 5: **Average Quoted Spread $\bar{Q}S$ and Average Realized Spread $\bar{R}S$ in the adverse-selection case and the no-adverse-selection case, for different values of the asset volatility Δ_v .** The other parameters are $\sigma = 5$, $N = 2$, $\mu = \frac{1}{2}$, $\mathbb{E}(v) = 2$, $T = 1,000,000$, and $K = 1,000$.

Panel A: Average Quoted Spread.

This graph plots the average over 1,000 experiments of the quoted spread, with 95% confidence intervals, both in the adverse-selection case and the no-adverse-selection case. The graph additionally plots the values of the quoted spread in both cases, in the Glosten-Milgrom benchmark of Section 2.3 (accounting for price discreteness).



Panel B: Average Realized Spread.

This graph plots the average over 1,000 experiments of the realized spread, with 95% confidence intervals, both in the adverse-selection case and the no-adverse-selection case. The graph additionally plots the values of the realized spread in both cases, in the Glosten-Milgrom benchmark of Section 2.3 (accounting for price discreteness).

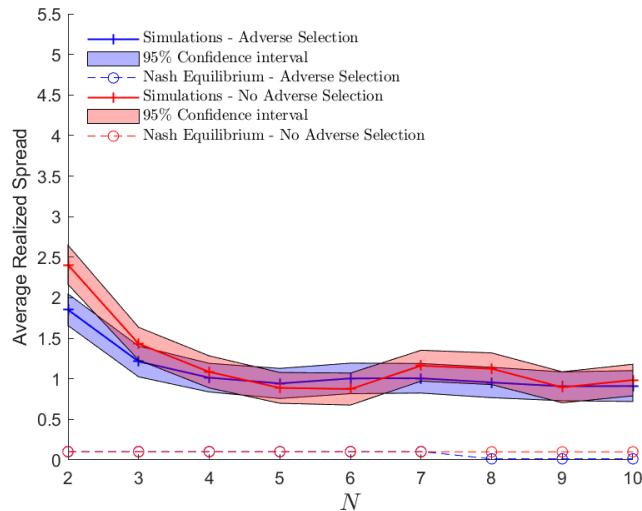
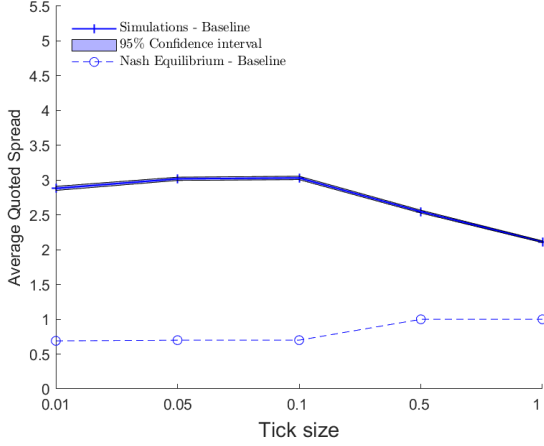
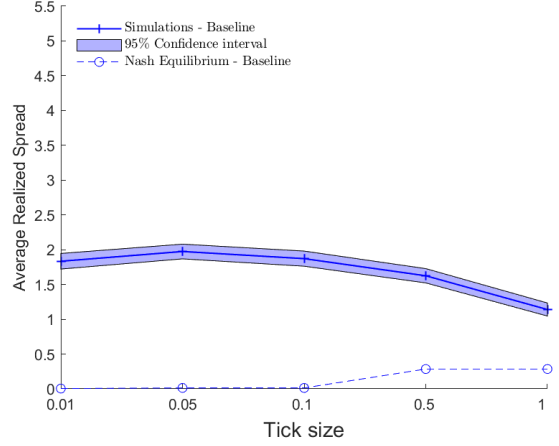


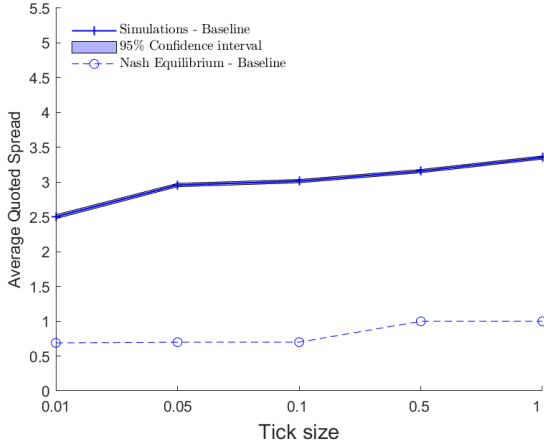
Figure 6: **Average Quoted Spread $\bar{Q}S$ and Average Realized Spread $\bar{R}S$ in the adverse-selection case and the no-adverse-selection case, for different values of the number N of AMs.** The other parameters are $\sigma = 5$, $\Delta_v = 4$, $\mu = \frac{1}{2}$, $\mathbb{E}(v) = 2$, $T = 1,000,000$, and $K = 1,000$.



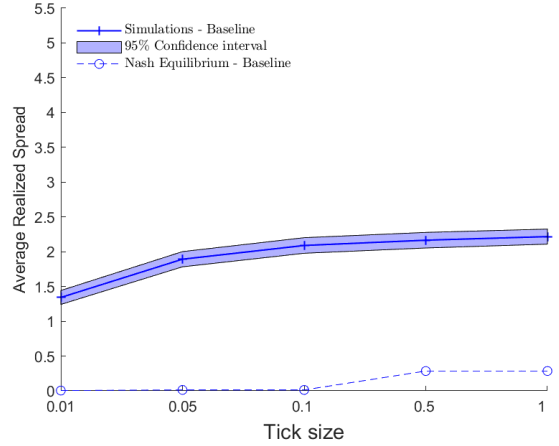
Panel A: Average Quoted Spread $\bar{Q}S$



Panel B: Average Realized Spread $\bar{R}S$

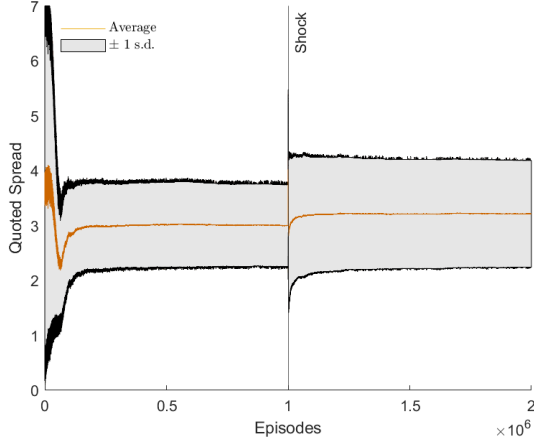


Panel C: Average Quoted Spread $\bar{Q}S$, adjusted β

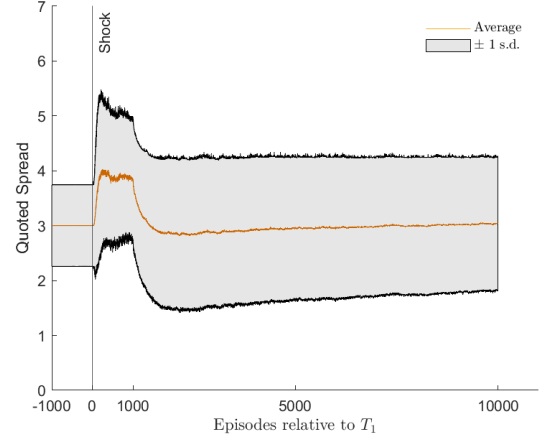


Panel D: Average Realized Spread $\bar{R}S$, adjusted β

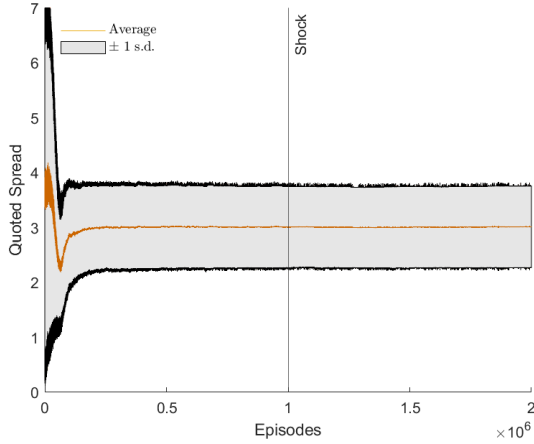
Figure 7: **Average Quoted Spread $\bar{Q}S$ and Average Realized Spread $\bar{R}S$ in the adverse-selection case, for different values of the tick size.** We use the baseline parameters $\sigma = 5$, $\Delta_v = 4$, $N = 2$, $\mu = \frac{1}{2}$, $\mathbb{E}(v) = 2$. We report the average over $K = 1,000$ experiments of the quoted spread (Panel A and Panel C) and the realized spread (Panel B and Panel D) in episode $T = 10^6$. In Panels A and B we use the baseline value $\beta = 8.10^{-5}$. In Panels C and D we adjust β to the tick size so that the average number of experimentations per price in the grid is constant (see the main text).



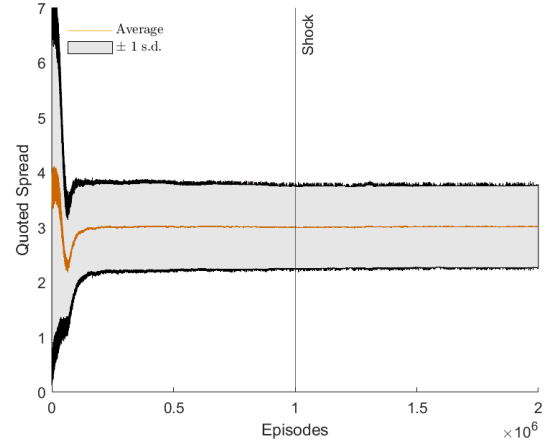
Panel A: $\Delta'_v = 7$



Panel B: $\Delta'_v = 7$, zoom on the shock period



Panel C: $\Delta'_v = \Delta_v = 4$



Panel D: $\Delta'_v = 1$

Figure 8: Dynamics of quoted spreads after a shock on adverse selection. Each panel shows the dynamics of the quoted spread, averaged over $K = 1,000$ experiments. We use the baseline parameters $\sigma = 5$, $\Delta_v = 4$, $N = 2$, $\mu = \frac{1}{2}$, $\mathbb{E}(v) = 2$. Between episodes $T_1 + 1$ and $T_1 + 1,000$, with $T_1 = 10^6$, there is a shock to Δ_v , whose value is changed to Δ'_v . Δ'_v is equal to 7 in Panel A and B (positive adverse selection shock), 1 in Panel D (negative adverse selection shock), and remains equal to 4 in panel C (Placebo). Panel B zooms on episodes between $T_1 - 1,000$ and $T_1 + 10,000$, while all other panels show all episodes between 1 and $T = 2 \cdot 10^6$.

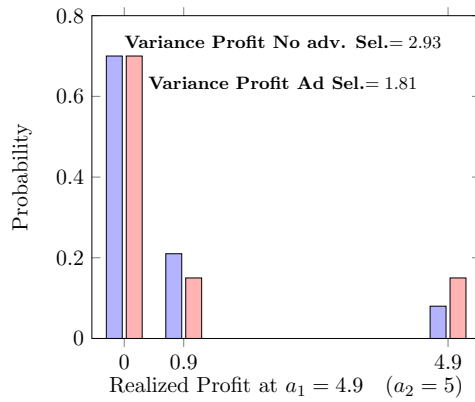


Figure 9: **Distribution of profits with and without adverse selection.** The figure compares the distribution of realized profits for AM 1 when it posts a price of 4.9 while AM 2 posts a price of 5 in the baseline case ($\mu = 0.5, v_H = 4, v_L = 0$) in the case without adverse selection (red) and the case with adverse selection (blue) when $\sigma = 5$. AM 1's realized profit can be 0 (the client does not trade), 0.9 (the client buys and the asset payoff is $v_H = 4$) or 4.9 (the client buys and the asset payoff is $v_L = 0$).

This graph plots the average over 1,000 experiments of the first-period and second-period prices, with 95% confidence intervals. The graph additionally plots the values of these prices in the Glosten-Milgrom benchmark of Section ?? (accounting for price discreteness).

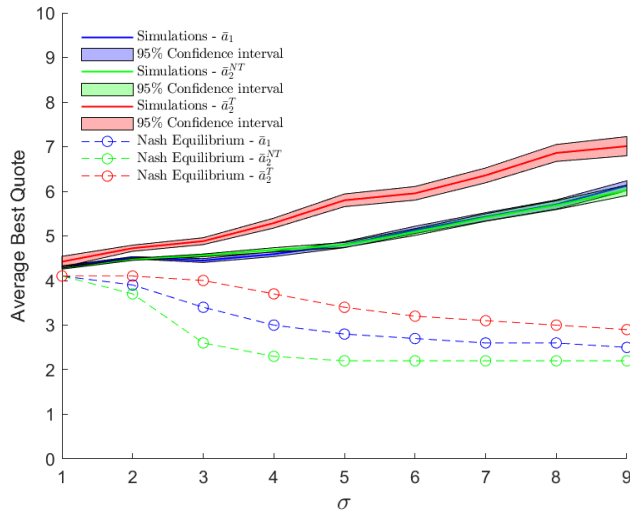


Figure 10: Average first-period price \bar{a}_1 and second-period price after a trade \bar{a}_2^T and after no trade \bar{a}_2^{NT} , for different values of the dispersion of clients' liquidity shocks σ . The other parameters are $\Delta_v = 4$, $N = 2$, $\mu = \frac{1}{2}$, $\mathbb{E}(v) = 2$, $T = 1,000,000$, and $K = 1,000$.

This graph plots the average over 1,000 experiments of the first-period and second-period price, with 95% confidence intervals. The graph additionally plots the values of these prices in the Glosten-Milgrom benchmark of Section ?? (accounting for price discreteness).

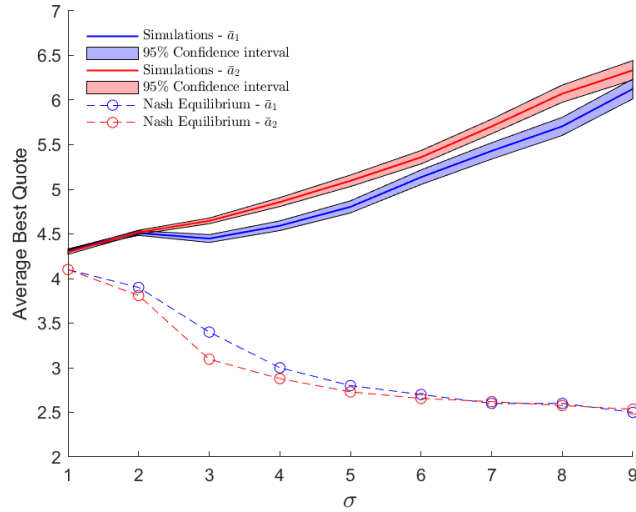


Figure 11: **Average first-period price \bar{a}_1 and average second-period price \bar{a}_2 , for different values of the dispersion of clients' liquidity shocks σ .** The other parameters are $\Delta_v = 4$, $N = 2$, $\mu = \frac{1}{2}$, $\mathbb{E}(v) = 2$, $T = 1,000,000$, and $K = 1,000$.

A.2 The Glosten-Milgrom Equilibrium

We just study the equilibrium with adverse selection since the equilibrium without adverse selection is straightforward. We show that, as claimed in the text, (i) the Glosten-Milgrom equilibrium always exists and (ii) the Glosten-Milgrom price increases with δ_v and decreases with σ in equilibrium.

As explained in the text, the Glosten-Milgrom price solves:

$$a^* = E_\mu(\tilde{v} \mid \tilde{v}^C > a^*), \quad (\text{A.1})$$

Define $F(a; \sigma, \Delta_v) := a - E_\mu(\tilde{v} \mid \tilde{v}^C > a)$. The Glosten-Milgrom price is the smallest root of:

$$F(a^*; \sigma, \Delta_v) = 0. \quad (\text{A.2})$$

We first show that there is always a solution to (A.2). Thus, the Glosten-Milgrom price always exists in our setting.

The Glosten-Milgrom price always exists. Let $\Pr_\mu(\tilde{v} = v_H \mid \tilde{v}^C > a)$ be the probability that the asset payoff is high ($v = v_H$) conditional on a trade, given dealers' beliefs (μ) about the payoff of the asset. We have

$$E_\mu(\tilde{v} \mid \tilde{v}^C > a) = \Pr_\mu(\tilde{v} = v_H \mid \tilde{v}^C > a)v_H + (1 - \Pr_\mu(\tilde{v} = v_H \mid \tilde{v}^C > a))v_L. \quad (\text{A.3})$$

Therefore, as $E_\mu(\tilde{v}) = \mu v_H + (1 - \mu)v_L$, we have

$$E_\mu(\tilde{v} \mid \tilde{v}^C > a) - E_\mu(\tilde{v}) = [\Pr_\mu(\tilde{v} = v_H \mid \tilde{v}^C > a) - \mu](v_H - v_L). \quad (\text{A.4})$$

It follows that:

$$F(a; \sigma, \Delta_v) = a - E_\mu(\tilde{v}) + (\Pr_\mu(\tilde{v} = v_H \mid \tilde{v}^C > a) - \mu)(v_H - v_L), \quad (\text{A.5})$$

Standard calculations yield:

$$\Pr_\mu(\tilde{v} = v_H \mid \tilde{v}^C > a) = \frac{D(a, v_H)}{\mu D(a, v_H) + (1 - \mu)D(a, v_L)}\mu, \quad (\text{A.6})$$

where $D(a, v)$ is defined in (3). As $D(a, v_L) > 0$, $\Pr_\mu(\tilde{v} = v_H \mid \tilde{v}^C > a) < 1$ when $\mu < 1$ and a finite.

Observe that (i) $F(\cdot)$ is continuous, (ii) $F(a; \sigma, \Delta_v) < 0$ for any $a \leq E_\mu(\tilde{v})$ and (iii) $F(v_H; \sigma, \Delta_v) > 0$ since $\Pr_\mu(\tilde{v} = v_H \mid \tilde{v}^C > a) < 1$ for all finite a (in particular $a = v_H$). Thus, there is at least one solution in the interval $(E_\mu(v), v_H)$ to (A.2). If there are multiple solutions, the competitive one is the smallest. Observe that as $F(E_\mu(\tilde{v}); \sigma, \Delta_v) < 0$, the Glosten-Milgrom price must be such that:

$$\frac{\partial F(a^*; \sigma, \Delta_v)}{\partial a} \Big|_{a=a^*} > 0. \quad (\text{A.7})$$

If it were not the case, there would be another solution to (A.2) in $(E_\mu(v), a^*)$. A contradiction since a^* is the smallest solution to (A.2).

The Glosten-Milgrom price decreases with σ . We deduce from (A.2) that

$$\frac{\partial a^*}{\partial \sigma} = - \frac{\frac{\partial F}{\partial a} \Big|_{a=a^*}}{\frac{\partial F}{\partial \sigma} \Big|_{a=a^*}}. \quad (\text{A.8})$$

As $\frac{\partial F}{\partial a} \Big|_{a=a^*} > 0$, we have that $\frac{\partial a^*}{\partial \sigma} < 0$ if and only if $\frac{\partial F}{\partial \sigma} > 0$. We now show that this is the case.

Observe, using (A.5), that $\frac{\partial F}{\partial \sigma} > 0$ iff $\Pr_\mu(\tilde{v} = v_H \mid \tilde{v}^C > a^*)$ decreases with σ . Using (A.6), we obtain

$$\frac{\partial \Pr_\mu(\tilde{v} = v_H \mid \tilde{v}^C > a^*)}{\partial \sigma} = \mu \left[\frac{\frac{\partial D(a^*, v_H)}{\partial \sigma} E_\mu(D(a^*, \tilde{v})) + \frac{\partial E_\mu(V(a^*, \tilde{v}^C))}{\partial \sigma} D(a^*, v_H)}{(E_\mu(V(a^*, \tilde{v}^C)))^2} \right]. \quad (\text{A.9})$$

It follows, after simplifying the numerator of the previous expression, that $\frac{\partial \Pr_\mu(\tilde{v} = v_H \mid \tilde{v}^C > a^*)}{\partial \sigma}$ has the same sign as

$$D(a^*, v_L) \frac{\partial D(a^*, v_H)}{\partial \sigma} - D(a^*, v_H) \frac{\partial D(a^*, v_L)}{\partial \sigma}.$$

Now remember that $D(a^*, v) = 1 - G(a^* - v)$ where $G(\cdot)$ is the c.d.f of a Gaussian variable with mean zero and variance σ^2 . It follows that $\frac{\partial D(a^*, v)}{\partial \sigma} = (\sqrt{2\pi}\sigma^2)^{-1} \exp(-\frac{(a^*-v)^2}{2\sigma^2})(a^* - v)$. Hence, the previous expression is negative since $a^* \in (v_L, v_H)$. Hence, a^* decreases with σ .

The Glosten-Milgrom price increases with Δ_v . We can proceed in the same way for analyzing the effect of Δ_v on a^* . The same reasoning as before shows that a^* increases with Δ_v if and only if $\Pr_\mu(\tilde{v} = v_H \mid \tilde{v}^C > a^*)$ increases with Δ_v . After some algebra, one obtains that $\frac{\partial \Pr_\mu(\tilde{v} = v_H \mid \tilde{v}^C > a^*)}{\partial \Delta_v}$ has the same sign as:

$$D(a^*, v_L) \frac{\partial D(a^*, v_H)}{\partial \Delta_v} - D(a^* - v_H) \frac{\partial D(a^*, v_L)}{\partial \Delta_v}.$$

Now remember that (i) $D(a^*, v) = 1 - G(a^* - v)$ where $G(\cdot)$ is the c.d.f of a Gaussian variable with mean zero and variance σ^2 and (ii) $v_H = \mu + \frac{\Delta_v}{2}$ and $v_L = \mu - \frac{\Delta_v}{2}$. It follows that $\frac{\partial D(a^*, v_H)}{\partial \Delta_v} > 0$

while $\frac{\partial D(a^*, v_L)}{\partial \Delta_v} < 0$. We deduce that $\frac{\partial \Pr_\mu(\tilde{v}=v_H | \tilde{v}^C > a^*)}{\partial \Delta_v} > 0$. Hence, a^* increases with Δ_v .

A.3 The Variance of AMs' Profits

To simplify notations, in this section, we define: $p(a) := \mathbb{E}_{\frac{1}{2}}(V(a, \tilde{v}^C)) = \frac{D(a_1, v_H)}{2} + \frac{D(a_1, v_L)}{2}$. We first consider the case in which $a_1 < a_2$.

A.3.1 The case $a_1 < a_2$.

No Adverse Selection Case.

Consider the case without adverse selection first. The distribution of AM 1's profit $\Pi(a_1, a_2, \tilde{v}^C, \tilde{v})$ when $a_1 < a_2$ (the case assumed in the text) is as follows:

1. $(a_1 - v_H)$ with probability $\frac{D(a_1, v_H)}{4} + \frac{D(a_1, v_L)}{4} = \frac{p(a)}{2}$.
2. $(a_1 - v_L)$ with probability $\frac{D(a_1, v_H)}{4} + \frac{D(a_1, v_L)}{4} = \frac{p(a)}{2}$.
3. 0 with probability $1 - p(a)$.

Denote by $\bar{\Pi}_{n.as} = p(a_1)(a_1 - \mathbb{E}_{\frac{1}{2}}(v))$, AM 1's expected profit in this case (remember again that $a_1 < a_2$). By definition (index *n.as* refers to "no adverse selection"):

$$\text{Var}_{n.as}(\Pi(a_1, a_2)) = \mathbb{E}((\Pi(a_1, a_2) - \bar{\Pi}_{n.as})^2). \quad (\text{A.10})$$

That is:

$$\text{Var}_{n.as} = p(a_1)(a_1 - \mathbb{E}_{\frac{1}{2}}(v) - \bar{\Pi}_{n.as})^2 + p(a_1)\frac{\Delta_v^2}{4} + \bar{\Pi}_{n.as}^2 - 2p(a_1)\bar{\Pi}_{n.as}(a_1 - \mathbb{E}_{\frac{1}{2}}(v)) \quad (\text{A.11})$$

Hence, as $\bar{\Pi}_{n.as} = p(a_1)(a_1 - \mathbb{E}_{\frac{1}{2}}(v))$, we deduce that:

$$\text{Var}_{n.as} = p(a_1)(1 - p(a_1))(a_1 - \mathbb{E}_{\frac{1}{2}}(v))^2 + p(a_1)\frac{\Delta_v^2}{4}, \quad (\text{A.12})$$

Adverse Selection Case.

Now consider the case with adverse selection. The distribution of AM 1's profit is then as follows:

1. $(a_1 - v_H)$ with probability $\frac{D(a_1, v_H)}{2}$.

2. $(a_1 - v_L)$ with probability $\frac{D(a_1, v_L)}{2}$

3. 0 with probability $1 - p(a_1)$.

Observe that, holding a_1 constant, the likelihood of a trade is the same, equal to $p(a_1)$, whether there is adverse selection or not. However, as discussed in the text and shown in 9, adverse selection shifts the distribution of profits conditional on a trade to the left because $(a_1 - v_H) < (a_1 - v_L)$ and $D(a_1, v_H) > p(a_1) > D(a_1, v_L)$.

More formally, denote $\bar{\Pi}_{as}$ AM1's expected profit with adverse selection (' as ') when it quotes $a_1 < a_2$ (this is given by (5) with $Z = 1$). By definition:

$$\text{Var}_{as}(\Pi(a_1, a_2)) = E((\Pi(a_1, a_2) - \bar{\Pi}_{as})^2). \quad (\text{A.13})$$

This is:

$$\text{Var}_{as}(\Pi(a_1, a_2)) = E((\Pi(a_1, a_2) - \bar{\Pi}_{n.as})^2) - 2E((\Pi(a_1, a_2) - \bar{\Pi}_{n.as})(\bar{\Pi}_{n.as} - \bar{\Pi}_{as})) + (\bar{\Pi}_{n.as} - \bar{\Pi}_{as})^2. \quad (\text{A.14})$$

That is, since by definition $\bar{\Pi}_{as} = E(\Pi(a_1, a_2))$ in the case we are considering,

$$\text{Var}_{as}(\Pi(a_1, a_2)) = E((\Pi(a_1, a_2) - \bar{\Pi}_{n.as})^2) - (\bar{\Pi}_{n.as} - \bar{\Pi}_{as})^2. \quad (\text{A.15})$$

Using the fact that $\Delta_D := D(a_1, v_H) - D(a_1, v_L)$, we deduce:

$$\begin{aligned} \text{Var}_{as} &= \frac{p(a_1)}{2}(a_1 - v_H - \bar{\Pi}_{n.as})^2 + \frac{p(a_1)}{2}(a_1 - v_L - \bar{\Pi}_{n.as})^2 + \\ &(1 - p(a_1))\bar{\Pi}_{n.as}^2 - \frac{\Delta_D}{4}[(a_1 - v_L - \bar{\Pi}_{n.as})^2 - (a_1 - v_H - \bar{\Pi}_{n.as})^2] - (\bar{\Pi}_{n.as} - \bar{\Pi}_{as})^2, \end{aligned} \quad (\text{A.16})$$

and therefore:

$$\text{Var}_{as} = \text{Var}_{n.as} - \frac{\Delta_D}{4}[(a_1 - v_L - \bar{\Pi}_{n.as})^2 - (a_1 - v_H - \bar{\Pi}_{n.as})^2] - (\bar{\Pi}_{n.as} - \bar{\Pi}_{as})^2, \quad (\text{A.17})$$

The second term is negative because $\Delta_D > 0$ and $a_1 - v_H < a_1 - v_L$. Thus, $\text{Var}_{n.as} < \text{Var}_{as}$. Moreover, after some algebra, we can rewrite the previous equation as

$$\text{Var}_{as} = \text{Var}_{n.as} - \frac{\Delta_D \Delta v}{2}[(a_1 - E_{\frac{1}{2}}(v)) - \bar{\Pi}_{n.as}] - (\bar{\Pi}_{n.as} - \bar{\Pi}_{as})^2. \quad (\text{A.18})$$

This can be simplified further by observing that:

$$\bar{\Pi}_{n.as} = p(a_1)(a_1 - \mathbb{E}_{\frac{1}{2}}(v)), \quad (\text{A.19})$$

$$\bar{\Pi}_{as} = p(a_1)(a_1 - \mathbb{E}_{\frac{1}{2}}(\tilde{v} \mid \tilde{v}^C > a)), \quad (\text{A.20})$$

and, therefore,

$$\bar{\Pi}_{n.as} - \bar{\Pi}_{as} = p(a_1)(\mathbb{E}_{\frac{1}{2}}(\tilde{v} \mid \tilde{v}^C > a) - \mathbb{E}_{\frac{1}{2}}(v)) = \frac{\Delta_D \Delta_v}{4}, \quad (\text{A.21})$$

where last equality follows from eq.(6) for $\mu = 0.5$ and the definition of $p(a_1)$ (remember that that, holding prices constant, the likelihood of a trade ($p(a_1)$) is the same in the adverse selection case as in the case without).

Thus, we can rewrite eq.(A.18) as:

$$\text{Var}_{as} = \text{Var}_{n.as} - \left[\left(\frac{\Delta_D \Delta_v}{2} \right) ((a_1 - \mathbb{E}_{\frac{1}{2}}(v))(1 - p(a_1)) + \frac{\Delta_D \Delta_v}{8}) \right]. \quad (\text{A.22})$$

To analyze the effect of σ on $\text{Var}_{n.as}$, observe first that $p(a_1) = \mathbb{E}_{\mu}(V(a, \tilde{v}^C))$ increases with σ for $a_1 \geq \mathbb{E}_{\mu}(v)$. Indeed:

$$\frac{\partial \mathbb{E}_{\mu}(V(a, \tilde{v}^C))}{\partial \sigma} = \mu \frac{\partial D(a, v_H)}{\partial \sigma} + (1 - \mu) \frac{\partial D(a, v_L)}{\partial \sigma}. \quad (\text{A.23})$$

As $D(a, v) = 1 - G(a - v)$ and $G(\cdot)$ is the c.d.f of a Gaussian variable with mean zero and variance σ^2 , we have $\frac{\partial D(a, v)}{\partial \sigma} = (\sqrt{2\pi}\sigma^2)^{-1} \exp(-\frac{(a-v)^2}{2\sigma^2})(a - v)$. As $a - v_H < a - v_L$, we deduce that:

$$\frac{\partial \mathbb{E}_{\mu}(V(a, \tilde{v}^C))}{\partial \sigma} = \mu \frac{\partial D(a, v_H)}{\partial \sigma} + (1 - \mu) \frac{\partial D(a, v_L)}{\partial \sigma} > (\sqrt{2\pi}\sigma^2)^{-1} \exp(-\frac{(a - v_L)^2}{2\sigma^2})(a - \mathbb{E}_{\mu}(v)) > 0, \quad (\text{A.24})$$

for $a > \mathbb{E}_{\mu}(v)$. Thus, for $a > \mathbb{E}_{\mu}(v)$, $\mathbb{E}_{\mu}(V(a, \tilde{v}^C))$ is maximal when σ goes to infinity and therefore $\mathbb{E}_{\mu}(V(a, \tilde{v}^C)) < \frac{1}{2}$ (since $D(a, v)$ goes to $\frac{1}{2}$ when σ goes to infinity). It follows from (A.12) that $\text{Var}_{n.as}$ increases with σ . A similar reasoning shows that Var_{as} increases with Δ_v .

Now consider the effect of σ on Var_{as} . The first term in eq.(A.22) ($\text{Var}_{n.as}$) increases with σ (as shown before) while the second term in brackets decreases with σ for $a_1 \geq \mathbb{E}_{\frac{1}{2}}(v)$ (the relevant case in our experiments) since Δ_D decreases with σ and $p(a_1)$ increases with σ . As this term is multiplied by -1 , we deduce that Var_{as} also increases with σ .

A.3.2 The case $a_1 \geq a_2$.

When $a_1 = a_2$, the analysis is identical to that followed when $a_1 < a_2$. The only difference is that realized profits are shared equally between the two dealers. Thus, the expressions for $Var_{n.as}(a_1, a_2)$ and $Var_{as}(a_1, a_2)$ are those given in eq.(A.12) and eq.(A.22) respectively divided by $\frac{1}{4}$. Thus, when the tick is small enough, for $a_1 = a_2 - tick$, $Var_j(a_1 - tick) > Var_j(a_1, a_2)$ for $j \in \{n.as, as\}$. Moreover, for $a_1 \geq a_2$, $Var_{n.as}(a_1, a_2) < Var_{as}(a_1, a_2)$ and both variances increase with σ .

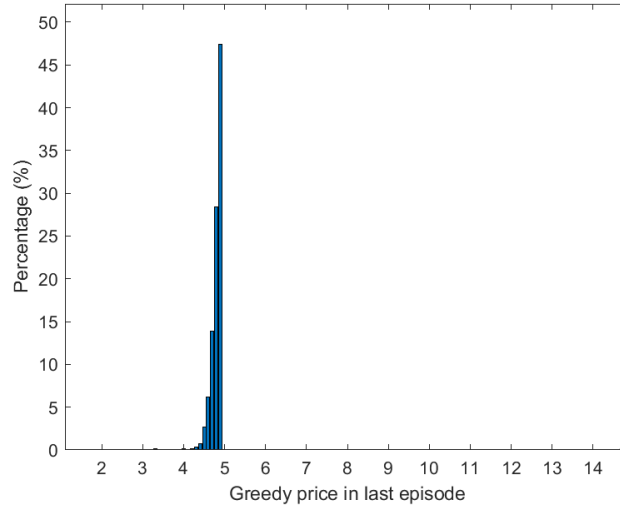
When $a_1 > a_2$, AM1 never trades and therefore $Var_{n.as}(a_1, a_2) = Var_{as}(a_1, a_2) = 0$.

A.4 The role of Strategic Uncertainty

As explained in the text, to assess the effect of strategic uncertainty on the competitiveness of AMs prices, we re-run the baseline experiments (Figure 3), assuming that AM2's price is fixed at $a_2 = 5.0$ in every episode. Figure 12 replicates Figure 3 in this case, with a histogram of the greedy price of AM 1 in episode T , and a plot of how the average greedy price of AM 1 evolves over episodes.

Panel A: Distribution of the greedy price of AM 1 in the last episode.

This panel shows a histogram of the greedy price of AM 1 in episode T : For each possible price a between 1.10 and 14.90 the bar indicates the percentage of the 1,000 experiments conducted in which $a_{1,T}^* = a$.



Panel B: Dynamics of the average greedy price of AM 1 for episodes 1 to T .

This graph shows for each episode t the average of AM 1's greedy price $a_{1,t}^*$ across the 1,000 experiments conducted. As a measure of dispersion, we also compute the standard deviation of $a_{1,t}^*$ across experiments and plot the average of $a_{1,t}^*$ plus/minus one standard deviation (with a 500-episode moving average for better readability).

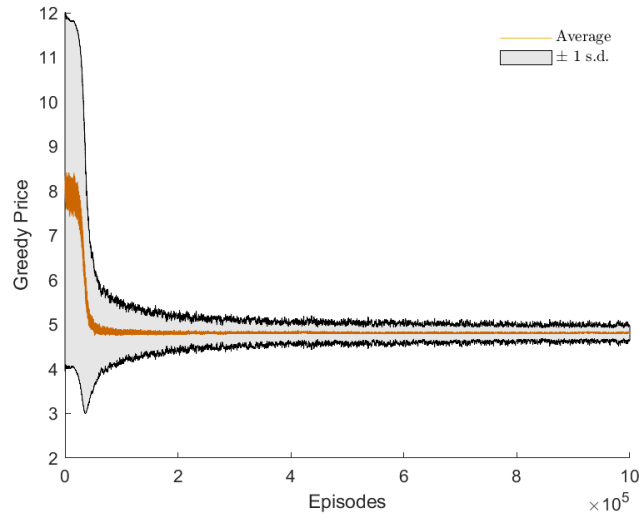


Figure 12: **Greedy price of AM 1 when AM 2 plays a constant price:** adverse-selection case, baseline parameters $\sigma = 5$, $\Delta_v = 4$, $N = 2$, $\mu = \frac{1}{2}$, $\mathbb{E}(v) = 2$, $T = 1,000,000$, and $K = 1,000$. AM 2 plays a constant price of 5.0 in every episode, while AM 1 uses a Q-learning algorithm with $\alpha = 0.01$ and $\beta = 0.00008$.

A.5 Convergence

As explained in the text, the environment in which AMs operate implies that the Q-matrices do not converge to a single value. More precisely, suppose AMs keep playing the same price profile $a \in \mathcal{A}^N$ at every episode t . Let a_m be the best price in a , and suppose it is played by AM n . Let $q_{m,n,t}$ denote the m -th entry in AM n 's Q-matrix at time t , i.e., the value that at time t , AM n attaches to playing price a_m . We show that for any t ,

$$\exists \Delta_q > 0, \text{ and } \epsilon > 0 \text{ s.t. } \Pr(|q_{m,n,t} - q_{m,n,t+1}| \geq \Delta_q) \geq \epsilon.$$

That is, each entry of the Q-matrix cannot converge in probability to a single value.⁴⁸ Thus, no matter how large is the number of episodes T , there is a strictly positive probability bounded away from 0 that the Q-matrix of the dealers posting the best price in episode t changes by more than a fixed amount $\Delta_q > 0$.

Formally, let define

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

that is strictly positive, as long as $v_H \neq v_L$ or $a_m \neq \frac{v_H + v_L}{2}$. Let

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

that is strictly positive because for any finite a_m and $v \in \{v_L, v_H\}$, one has $0 < D(a, v) < 1$.

Lemma 1. *For any given t and $a_m \in \mathcal{A}$, if $a_{n,t} = a_m = a_t^{\min}$, then,*

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

Proof. 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. $a_{nt} = a_m = a_t^{\min}$. 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

⁴⁸ $q_{m,n,t}$ converges in probability to a real number $q \in \mathbb{R}$ if for any $\epsilon > 0$, one has $\lim_{t \rightarrow \infty} \Pr(|q_{m,n,t} - q| \geq \epsilon) = 0$.

$q_{m,n,t+1} = (1 - \alpha)q_{m,n,t}$, implying

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

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

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

whereas if $\tilde{v} = v_L$,

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

Denote $\Delta_m(q) := \alpha \max\{|q|, |a_m - v_H - q|, |a_m - v_L - q|\}$. This is the maximum possible value that $|q_{m,n,t} - q_{m,n,t+1}|$ can take, given that $q_{m,n,t} = q$. Note that three situations are possible.

If $a_m < v_L$, then

$$\Delta_m(q) = \begin{cases} \alpha(-q + v_H + a_m) & \text{for } q \leq \frac{v_H - a_m}{2} \\ \alpha q & \text{for } q > \frac{v_H - a_m}{2} \end{cases}$$

that implies

$$\Delta_m(q) \geq \frac{\alpha(v_H - a_m)}{2} \geq \frac{\alpha(v_H - v_L)}{2}, \frac{\alpha(v_L - a_m)}{2}$$

If $v_L \leq a_m \leq v_H$, then

$$\Delta_m(q) = \begin{cases} \alpha(-q + v_H - a_m) & \text{for } q \leq \frac{v_H + v_L}{2} - a_m \\ \alpha(q - v_L + a_m) & \text{for } q > \frac{v_L + v_H}{2} - a_m \end{cases}$$

that implies

$$\Delta_m(q) \geq \frac{\alpha(v_H - v_L)}{2} \geq \frac{\alpha(v_H - a_m)}{2}, \frac{\alpha(v_L - a_m)}{2}$$

If $a_m > v_H$, then

$$\Delta_m(q) = \begin{cases} -\alpha q & \text{for } q \leq \frac{v_L - a_m}{2} \\ \alpha(q - v_L + a_m) & \text{for } q > \frac{v_L - a_m}{2} \end{cases}$$

that implies

$$\Delta_m(q) \geq \frac{\alpha(a_m - v_L)}{2} \geq \frac{\alpha(v_H - v_L)}{2}, \frac{\alpha(v_H - a_m)}{2}$$

Hence we can write

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

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.

A.6 Nash Equilibria

In this section we analyze the Nash equilibria of the one-shot game when market makers are constrained to choose prices from a finite grid (a positive tick size). We first show that when the tick size is small enough or the number N of market makers is large enough the game has a unique Nash equilibrium. However it is possible that for N relatively small and tick size relatively large the game has more than one pure Nash equilibrium. Namely for the value of the parameters in the range of our experiment we find that the game has either 1 or 2 pure Nash equilibria. We show that for $N = 2$, if the game has 2 pure Nash equilibria then it also has one mixed strategy equilibrium where market makers independently randomize their quotes over the two prices that form the two pure equilibria.

Denote $\Pi(a)$ the expected payoff of a monopolistic market maker who sets a price a . Namely

$$\Pi(a) = \mu D(a, v_H)(a - v_H) + (1 - \mu) D(a, v_L)(a - v_L)$$

Let a^* be the smallest solution of the equation $\Pi(a) = 0$. This is the equilibrium price in the game where market makers can choose their prices on the real line. We know that a^* exists because $\Pi(a)$ is continuous in a , strictly negative for $a < v_L$ and strictly positive for $a > v_H$.

Let us now consider the game with $N > 1$ market makers that have to choose their prices on a grid \mathcal{A} , and denote δ the tick size. Without loss of generality we can assume that a^* is not on the price grid, i.e., $a^* \notin \mathcal{A}$.

Lemma 2. *In the game with N market makers, an action profile $\{a_1, a_2, \dots, a_N\} \in \mathcal{A}^N$ is a pure Nash equilibrium if and only if the following two conditions are satisfied,*

1. *All market makers set the same price $a \in \mathcal{A}$*
2. *The price a is such that $\Pi(a) \geq 0$ and*

$$\frac{1}{N}\Pi(a) \geq \max_{a' \in \mathcal{A}, a' < a} \{\Pi(a')\} \quad (\text{A.25})$$

Proof. Sufficient condition: suppose that all market makers except market maker i set a price equal to a , and that a satisfies condition (A.25). Let us consider the best response of market maker i . The expected payoff from playing $a' = a$ is $\frac{1}{N}\Pi(a) \geq 0$, as the market maker i has to share the payoff $\Pi(a)$ with the other $n - 1$ market makers. The expected payoff from undercutting the other market makers by playing some $a' < a$ is $\Pi(a') \leq \frac{1}{N}\Pi(a)$, by condition (A.25). Whereas the expected payoff from from playing some $a'' > a$ is $0 \leq \frac{1}{N}\Pi(a)$, as no client trades with market maker i . Hence $a_i = a$ for all i is a pure Nash equilibrium.

Necessary condition: Suppose the action profile $\{a_1, a_2, \dots, a_N\} \in \mathcal{A}^N$ forms a Nash equilibrium, and let a^{\min} be the lowest offered price. If $\Pi(a^{\min}) < 0$, then the MM playing a^{\min} gets a negative profit, and he has a profitable deviation by setting any $a' > v_H$. Hence, because prices belong to a discrete grid without loss of generality it must be that $\Pi(a^{\min}) > 0$. If there is a market maker i such that $a_i \neq a^{\min}$, then $a_i > a^{\min}$ and the market maker's payoff is nil. But then market maker i has a profitable deviation by playing a^{\min} that provides him with a fraction of the strictly positive payoff $\Pi(a^{\min})$. Hence all market makers must play a^{\min} and get a payoff of $\frac{1}{N}\Pi(a^{\min})$. If there is $a' < a^{\min}$ such that $\Pi(a') > \frac{1}{N}\Pi(a^{\min})$ then playing such a' would be a profitable deviation. Hence conditions 1. and 2. in the Lemma are necessary conditions for an equilibrium. \square

Denote $\hat{a} \in \mathcal{A}$ the smallest price in the grid larger than a^* . Formally

$$\hat{a} = \min\{a \in \mathcal{A}, s.t. \Pi(a) \geq 0\}.$$

Note that if the price grid δ is small enough, then \hat{a} is the price on the grid closest to a^* weakly greater than a^* . Then we have

Corollary 1. *If N is large enough or δ is small enough then all MMs playing \hat{a} is the unique Nash equilibrium of the game.*

Proof. Let first show that playing \hat{a} , is an equilibrium. Because of the definition of \hat{a} , all $a' < \hat{a}$ on the grid \mathcal{A} provide strictly negative payoff, and generically we have $\Pi(\hat{a}) > 0$. Thus \hat{a} satisfies condition 2 in the Lemma 2 and thus playing a is a Nash equilibrium.

Let now show that if N is large or δ small, then there are no other equilibria. Suppose that there is another equilibrium, where all MMs play $a \neq \hat{a}$. Then $\Pi(a) > 0$ and hence $a > \hat{a}$. If a player deviates from this equilibrium and plays \hat{a} instead, then his payoff is $\Pi(\hat{a}) > \frac{1}{N}\Pi(a)$ for $N > \pi(a)/\pi(\hat{a})$. Thus for N large enough playing $a \neq \hat{a}$ cannot be an equilibrium as it violates condition (A.25).

Now fix N and suppose that there is an equilibrium where $a > \hat{a}$, and consider the deviation to the largest price on the grid that is smaller than a . This is $a' = a - \delta$. If a MM deviates and plays a' , then his payoff is $\Pi(a - \delta)$ that tends to $\Pi(a)$ as δ goes to 0. Thus for δ small enough and $n > 1$, $\Pi(a') > \frac{1}{N}\Pi(a)$, and thus a cannot be a Nash equilibrium as it violates condition (A.25). \square

Because the δ we use in our experiments is relatively large, for N relatively small we find that depending on the value of the parameters, the game has either 1 or 2 pure Nash equilibria. In the next Lemma we show that if the 2-payer game has two Nash equilibria, then it also has a third equilibrium in mixed strategies.

Lemma 3. *Suppose that for $N = 2$, there are two pure Nash equilibria: \hat{a} and $a > \hat{a}$. Then the game also has a mixed strategy equilibrium where market makers independently randomize between setting a price of $a_i = a$ with probability $\eta = \frac{\Pi(\hat{a})}{\Pi(a) - \Pi(\hat{a})}$ and $a_i = \hat{a}$ with the complementary probability.*

Proof. Note first that both a and \hat{a} must satisfy condition (A.25). Namely condition (A.25) applied to \hat{a} implies $\Pi(\hat{a}) > 0$, and if applied to a , implies $\Pi(\hat{a}) < \frac{1}{2}\Pi(a)$. These two inequalities imply that $0 < \eta < 1$.

Not that if the other MM j plays a and \hat{a} with probability η and $1 - \eta$, respectively, then MM i is indifferent between playing a or \hat{a} . This because η is the solution of the following indifference condition

$$\underbrace{\eta\Pi(\hat{a}) + \frac{1}{2}(1 - \eta)\Pi(\hat{a})}_{\text{expected payoff from playing } \hat{a}} = \underbrace{\frac{1}{2}\eta\Pi(a)}_{\text{expected payoff from playing } a}$$

Both actions lead to an expected payoff of

$$\Pi^{mix} = \frac{\Pi(a)\Pi(\hat{a})}{2(\Pi(a) - \Pi(\hat{a}))} > 0$$

Let show that by unilaterally deviating to any $a' \notin \{\hat{a}, a\}$ MM i cannot gain more than Π^{mix} . For $a' < \hat{a}$ the deviation payoff is strictly negative, by definition of \hat{a} . For $\hat{a} < a' < a$ the deviation payoff is

$$\eta\Pi(a') = \frac{\Pi(a')\Pi(\hat{a})}{\Pi(a) - \Pi(\hat{a})} \leq \frac{\Pi(a)\Pi(\hat{a})}{2(\Pi(a) - \Pi(\hat{a}))} = \Pi^{mix},$$

where the inequality follows from condition (A.25) applied to equilibrium a , that implies $\frac{1}{2}\Pi(a) \geq \Pi(a')$. For $a' > a$, MM does not trade and gets $0 < \Pi^{mix}$. \square

Online Appendix to “Algorithmic Pricing and Liquidity in Securities Markets”

Jean-Edouard Colliard, Thierry Foucault, and Stefano Lovo

This Online Appendix provides additional robustness tests and experiments.

OA.1 The two-period case

OA.1.1 Glosten-Milgrom benchmark

As mentioned in the text, in the two-period Glosten-Milgrom benchmark the second-period quotes are given by (6), with $\mu = \mu_\tau$, where μ_τ is the probability that $\tilde{v} = v_H$ conditionally on past history. More formally, we define H_1 the history in the first period, as follows: (i) if there was a trade at price a_1^{\min} then $H_1 = \{1, a_1^{\min}\}$; (ii) if the best quote was a_1^{\min} but no trade occurred then $H_1 = \{0, a_1^{\min}\}$. We denote $\mu_2(H_1)$ the market makers’ Bayesian beliefs about the likelihood that $v = v_H$ after observing H_1 . We have:

$$\mu_2(1, a_1^{\min}) := \Pr(v = v_H \mid H_1 = \{1, a_1^{\min}\}) = \frac{D(a_1^{\min}, v_H)\mu_1}{\mathbb{E}_{\mu_1}(V(a_1^{\min}, \tilde{v}_1^C))} \quad (\text{OA.1})$$

$$\mu_2(0, a_1^{\min}) := \Pr(v = v_H \mid H_1 = \{0, a_1^{\min}\}) = \frac{(1 - D(a_1^{\min}, v_H))\mu_1}{1 - \mathbb{E}_{\mu_1}(V(a_1^{\min}, \tilde{v}_1^C))}. \quad (\text{OA.2})$$

We claim two results about these prices in the main text:

First, it is easily checked that $\mu_2(1, a_1^{\min}) > \mu_1 > \mu_2(0, a_1^{\min})$ if (and only if) $\Delta_v > 0$. That is, Bayesian market makers revise their estimate of the expected payoff of the asset upwards after a buy in period 1 and downwards after no trade.

Second, the ask prices increases over time in expectation: $\mathbb{E}(a_2^{\min}) \leq a_1^{\min}$. The proof has two steps: (i) we show that the ask price (6) is a concave function of the dealers’ belief μ ; (ii) we show that, given this concavity, the expectation of the next ask price is below the level of the current ask price.

(i) Let $D(a, v)$ denote the probability that a client buys at price a given the fundamental value of the asset is v . This function is decreasing in a and increasing in v . For a given belief μ , the dealers’ aggregate expected profit when the best ask is a is:

$$\Pi(a, \mu) := \mu(a - v_H)D(a, v_H) + (1 - \mu)(a - v_L)D(a, v_L) \quad (\text{OA.3})$$

The equilibrium price denoted $a(\mu)$ is the smallest a such that

$$\Pi(a, \mu) = 0 \tag{OA.4}$$

From the implicit function theorem we have that

$$a'(\mu) = -\frac{\partial \Pi / \partial \mu}{\partial \Pi / \partial a} > 0 \tag{OA.5}$$

and

$$a''(\mu) = -\frac{(\partial \Pi / \partial a)(\partial \Pi^2 / \partial \mu^2) - (\partial \Pi / \partial \mu)(\partial \Pi^2 / \partial \mu \partial a)}{(\partial \Pi / \partial a)^2} = \frac{(\partial \Pi / \partial \mu)(\partial \Pi^2 / \partial \mu \partial a)}{(\partial \Pi / \partial a)^2} \tag{OA.6}$$

where the second equality follows from the fact that $\Pi(a, \mu)$ is linear in μ . Because $\partial \Pi / \partial \mu < 0$, we have that $a''(\mu) < 0$ only if

$$\frac{\partial \Pi^2}{\partial a \partial \mu} > 0 \tag{OA.7}$$

That is

$$D(a, v_H) + (a - v_H) \frac{\partial D(a, v_H)}{\partial a} > D(a, v_L) + (a - v_L) \frac{\partial D(a, v_L)}{\partial a}. \tag{OA.8}$$

We have $D(a, v_H) > D(a, v_L)$, $(a - v_H) \frac{\partial D(a, v_H)}{\partial a} > 0$ because $a < v_H$, and $(a - v_L) \frac{\partial D(a, v_L)}{\partial a} < 0$ because $a > v_L$. Together these conditions imply that (OA.8) is positive, which is a sufficient condition for $a''(\mu) < 0$. This concludes the proof that $a(\mu)$ is concave in μ .

(ii) The expected best ask in period 2 can be written as:

$$\mathbb{E}[a_2^{\min}] = \Pr(H_1 = \{1, a_1^{\min}\})a(\mu_2(1, a_1^{\min})) + \Pr(H_1 = \{0, a_1^{\min}\})a(\mu_2(0, a_1^{\min})) = \mathbb{E}[a(\mu_2)]. \tag{OA.9}$$

Since $a(\cdot)$ is concave, by Jensen's inequality we have:

$$\mathbb{E}[a(\mu_2)] \leq a(\mathbb{E}[\mu_2]). \tag{OA.10}$$

Finally, by the law of iterated expectations, $\mu_1 = \mathbb{E}[\tilde{v}] = \mathbb{E}[\mathbb{E}[\tilde{v}|H_1]] = \mathbb{E}[\mu_2]$. Thus, we obtain $\mathbb{E}[a(\mu_2)] \leq a(\mathbb{E}[\mu_2]) = a(\mu_1) = a_1^{\min}$, which concludes the proof.

OA.1.2 Experiments with Q-learning algorithms

We formally define the algorithms and the process we simulate in the two-period case.

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) $s_n \in \mathcal{S} = \left\{0, \frac{1}{N}, \frac{1}{N-1}, \dots, \frac{1}{2}, 1\right\}$ 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 point (iii). 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 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 $[\underline{q}, \bar{q}]$. 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 $s = \emptyset$, and we denote $a_{n,t}^{1,*} = a_{m_{n,t}^{1,*}}$ the corresponding greedy price.
2. For each AM n , with probability $\epsilon_t = e^{-\beta t}$ the AM “explores”: it draws a random integer $\tilde{m}_{n,t}^1$ between 1 and M , all values being equiprobable, and plays $a_{n,t}^1 = a_{\tilde{m}_{n,t}^1}$. With probability $1 - \epsilon_t$, the AM “exploits” and plays the greedy price $a_{n,t}^1 = a_{n,t}^{1,*}$. 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 $a_t^{1,min} = \min_n \{a_{n,t}^1\}$, and draw \tilde{v}_t and $\tilde{L}_{1,t}$. This determines the position $I_{n,t}^1$ 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}_t^1 the set of AMs who quote $a_t^{1,min}$ and z_t^1 the size of this set. Then, if $\tilde{v}_t + \tilde{L}_{1,t} \geq a_t^{1,min}$ we have $I_{n,t}^1 = s_{n,t} = \frac{1}{z_t^1}$ for every $n \in \mathcal{D}_t^1$, and $I_{n,t}^1 = s_{n,t} = 0$ for $n \notin \mathcal{D}_t^1$. If $\tilde{v}_t + \tilde{L}_{1,t} < a_t^{1,min}$ then $I_{n,t}^1 = 0$ and $s_{n,t} = NT$ for every n .
4. We update the first column of the Q-matrix of each AM n as follows:

$$q_{m,\emptyset,n,t} = \begin{cases} \alpha [a_{n,t}^1 I_{n,t}^1 + \max_{m'} q_{m',s_{n,t},n,t-1}] + (1 - \alpha) q_{m,\emptyset,n,t-1} & \text{if } a_{n,t}^1 = a_m \\ q_{m,\emptyset,n,t-1} & \text{if } a_{n,t}^1 \neq a_m \end{cases} \quad (\text{OA.11})$$

Period 2:

1. At the beginning of period 2 we know the state $s_{n,t}$ in which AM n finds itself. We define $m_{n,t}^{2,*} = \arg \max_m q_{m,s_{n,t},n,t-1}$ the index associated with the highest value in matrix $\mathbf{Q}_{n,t-1}$ in state $s = s_{n,t}$, and we denote $a_{n,t}^{2,*} = a_{m_{n,t}^{2,*}}$ the corresponding greedy price.
2. With probability ϵ_t the AM plays a random price $a_{n,t}^2$, following the same process as in period 1.
 1. With probability $1 - \epsilon_t$, the AM plays $a_{n,t}^2 = a_{n,t}^{2,*}$.
3. We compute $a_t^{2,min} = \min_n a_{n,t}^2$ and draw $\tilde{L}_{2,t}$. This determines the position $I_{n,t}^2$ 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:

$$\forall 1 \leq n \leq N, q_{m,s_{n,t},n,t} = \begin{cases} \alpha[a_{n,t}^2 I_{n,t}^2 - \tilde{v}_t(I_{n,t}^1 + I_{n,t}^2)] + (1 - \alpha)q_{m,s_{n,t},n,t-1} & \text{if } a_{n,t}^2 = a_m \\ q_{m,s_{n,t},n,t-1} & \text{if } a_{n,t}^2 \neq a_m \end{cases} \quad (\text{OA.12})$$

The way updating works in the 2-period case is best understood backwards. (OA.12) is the updating in period 2 when the state is $s_{n,t}$. At the end of period 2, we know the quantities $I_{n,t}^1$ and $I_{n,t}^2$ sold by AM n in periods 1 and 2, respectively. We count the revenues $a_{n,t}^2 I_{n,t}^2$ generated by the period-2 sale, and subtract the cost $\tilde{v}_t(I_{n,t}^1 + I_{n,t}^2)$ of having sold $I_{n,t}^1 + I_{n,t}^2$ units worth \tilde{v}_t each. (OA.11) is the updating done in period 1. The reward recorded by the algorithm has two components. First, the revenues $a_{n,t}^1 I_{n,t}^1$ from selling $I_{n,t}^1$ units. As already mentioned, in period 1 the value of \tilde{v}_t is still unknown and cannot be deducted from the revenues, this will be done at the end of period 2 only. To keep track of this cost, and following the standard specification of Q-learning, we add the term $\max_{m'} q_{m',s_{n,t},n,t-1}$: this term is the value associated with moving to state $s_{n,t}$ in period 2, which as we just saw incorporates the cost of selling the asset. For instance, if AM n sells one unit in period 1 we have $I_{n,t}^1 = 1$ and revenues of $a_{n,t}^1 \times 1$ are recorded in the first column of the Q-matrix. In addition, AM n will start period 2 in state $s_{n,t} = 1$, and the expected value of this state is $\max_{m'} q_{m',1,n,t-1}$. This value takes into account that in this state AM n starts with an inventory of 1, which will have a cost of \tilde{v}_t .

We repeat this process for $T = 10^6$ episodes, after which the experiment ends. We then repeat the entire process for $K = 1,000$ experiments. For the last episode T of experiment k , we denote $a_\tau^{min,k}$ the best quote and V_τ^k the realized volume in period $\tau \in \{1, 2\}$. We then define:

$$\bar{V}_2 = \frac{\sum_{k=1}^K V_2^k}{K} \quad (\text{OA.13})$$

$$\bar{a}_1 = \frac{\sum_{k=1}^K a_1^{\min,k}}{K} \quad (\text{OA.14})$$

$$\bar{a}_2 = \frac{\sum_{k=1}^K a_2^{\min,k}}{K} \quad (\text{OA.15})$$

$$\bar{a}_2^T = \frac{\sum_{k=1}^K a_2^{\min,k} V_2^k}{K \bar{V}_2} \quad (\text{OA.16})$$

$$\bar{a}_2^{NT} = \frac{\sum_{k=1}^K a_2^{\min,k} (1 - V_2^k)}{K(1 - \bar{V}_2)}. \quad (\text{OA.17})$$

Thus, \bar{a}_1 is the average best quote in period 1 across the K experiments, \bar{a}_2 the average best quote in period 2 across the K experiments, \bar{a}_2^T is the average best quote in period 2 conditionally on a trade occurring in period 1 (irrespective of who traded), and \bar{a}_2^{NT} in the average best quote in period 2 conditionally on no trade occurring in period 1.

OA.2 Infinite experimentation and empirical average

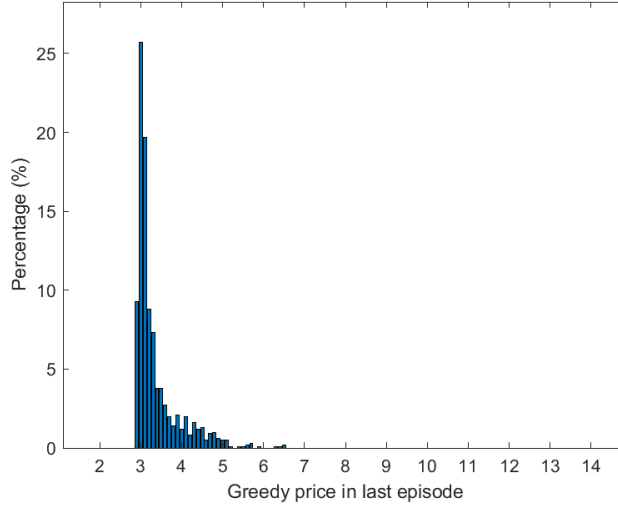
We run the same experiments as in Figure 3, with the same parameters, but we change the parameterization of both algorithms. We now have $\epsilon_t = 0.05 + 0.95 \exp^{-\beta t}$: for early episodes the experimentation probability ϵ_t will be high and then decrease exponentially like in the baseline case, but it will converge towards 0.05 instead of 0. Thus, in the long-run the algorithms will still experiment once every 20 episodes on average. Moreover, we change the updating rule (8) so that now the entries in the Q-matrix correspond to the empirical average of the profit obtained with each price. Formally, denoting $\nu_{m,n,t}$ the number of times price m has been tried by AM n before episode t , we update $q_{m,n,t}$ as:

$$q_{m,n,t} = \begin{cases} \frac{\pi_{n,t} + \nu_{m,n,t} q_{m,n,t-1}}{1 + \nu_{m,n,t}} & \text{if } a_{n,t} = a_m \\ q_{m,n,t} & \text{if } a_{n,t} \neq a_m \end{cases} \quad (\text{OA.18})$$

We initialize each Q-matrix as in the baseline case, and start with $\nu_{m,n,1} = 1$ for every m and n . Figure OA.1 replicates Figure 3 in that case, with a histogram of the greedy price of AM 1 in episode T , and a plot of how the average greedy price of AM 1 evolves over episodes.

Panel A: Distribution of the greedy price of AM 1 in the last episode.

This panel shows a histogram of the greedy price of AM 1 in episode T : For each possible price a between 1.10 and 14.90 the bar indicates the percentage of the 1,000 experiments conducted in which $a_{1,T}^* = a$.



Panel B: Dynamics of the average greedy price of AM 1 for episodes 1 to T .

This graph shows for each episode t the average of AM 1's greedy price $a_{1,t}^*$ across the 1,000 experiments conducted. As a measure of dispersion, we also compute the standard deviation of $a_{1,t}^*$ across experiments and plot the average of $a_{1,t}^*$ plus/minus one standard deviation (with a 500-episode moving average for better readability).

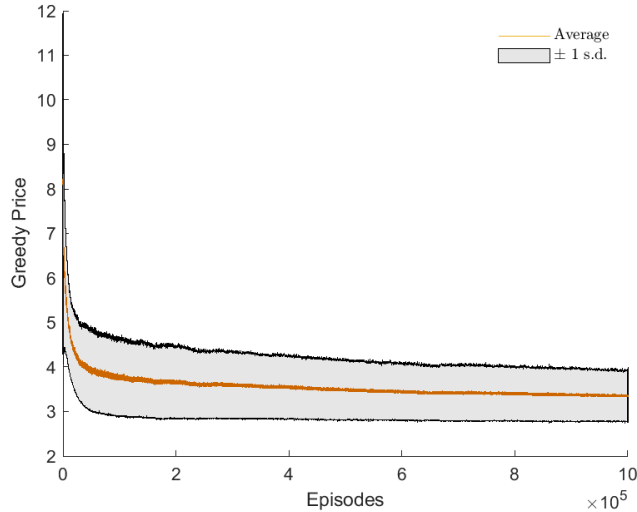
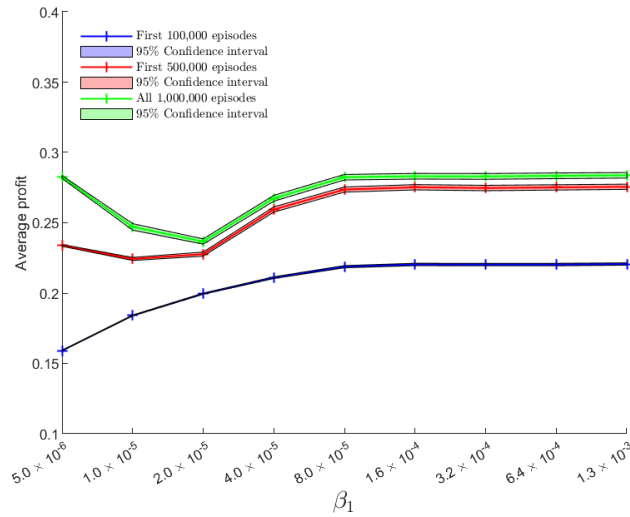


Figure OA.1: **Greedy price of AM 1 when AM 1 and AM 2 keep experimenting in the long-run:** adverse-selection case, baseline parameters $\sigma = 5$, $\Delta_v = 4$, $N = 2$, $\mu = \frac{1}{2}$, $\mathbb{E}(v) = 2$, $T = 1,000,000$, and $K = 1,000$. Both AMs use $\epsilon_t = 0.05 + 0.95 \exp^{-\beta t}$ and the Q-matrix records the empirical average of the profit obtained with each price in past episodes.

OA.3 Experimentation is costly



This figure shows AM1’s average profit per client over the 100,000 first episodes (blue), 500,000 first episodes (red) and all episodes (green) for various values of β_1 (AM1’s experimentation rate), holding AM2’s experimentation rate constant at the baseline value for AMs’ experimentation rates in our experiments ($\beta = 8.10^{-5}$). Observe that a decrease in β_1 relative to the baseline reduces AM1’s average profit per client. This reflects the fact that experimentation is costly (it requires to take actions that are truly dominated).

OA.4 Robustness to alternative values of α and β

To test the robustness of our results to the choice of α and β , we run simulations for $K = 1,000$ experiments under the baseline parameters, but with different choices of α and β for both algorithms. We consider $\alpha \in \{\alpha_l, \alpha_m, \alpha_h\}$ and $\beta \in \{\beta_l, \beta_m, \beta_h\}$, with $\alpha_l = 0.001, \alpha_m = 0.01, \alpha_h = 0.1$; $\beta_l = 5.10^{-6}, \beta_m = 8.10^{-5}, \beta_h = 3.2.10^{-4}$. Table OA.1 gives, for each possible pair of choices, the average best quote in episode T across the K experiments.

		AM 2								
		(α_l, β_l)	(α_l, β_m)	(α_l, β_h)	(α_m, β_l)	(α_m, β_m)	(α_m, β_h)	(α_h, β_l)	(α_h, β_m)	(α_h, β_h)
AM 1	(α_l, β_l)	5.3713	5.4242	5.3859	4.2640	4.3353	4.3381	3.9211	3.9852	3.9824
	(α_l, β_m)	5.4242	5.7346	5.7070	4.1941	4.8067	4.8024	3.8814	4.0062	4.0008
	(α_l, β_h)	5.3859	5.7070	5.6899	4.1924	4.8282	4.8359	3.8850	3.9939	3.9997
	(α_m, β_l)	4.2640	4.1941	4.1924	4.1133	4.0979	4.1443	3.9844	4.0808	4.0745
	(α_m, β_m)	4.3353	4.8067	4.8282	4.0979	5.0333	5.0449	3.8685	4.4044	4.4114
	(α_m, β_h)	4.3381	4.8024	4.8359	4.1443	5.0449	5.0749	3.8609	4.4280	4.4305
	(α_h, β_l)	3.9211	3.8814	3.8850	3.9844	3.8685	3.8609	3.9953	4.0122	4.0141
	(α_h, β_m)	3.9852	4.0062	3.9939	4.0808	4.4044	4.4280	4.0122	4.1543	4.1839
	(α_h, β_h)	3.9824	4.0008	3.9997	4.0745	4.4114	4.4305	4.0141	4.1839	4.3404

Table OA.1: This matrix gives the average value of the best price across all experiments in the last episode. For instance, if AM1 chooses (α_m, β_l) and AM2 chooses (α_l, β_l) , the best price in the last episode is 4.26. Hyper-Parameters: $\alpha_l = 0.001, \alpha_m = 0.01, \alpha_h = 0.1$; $\beta_l = 5.10^{-6}, \beta_m = 8.10^{-5}, \beta_h = 3.2.10^{-4}$.

Note that in all cells of Table OA.1 the values of σ and Δ_v are equal to their baseline values and the tick size is 0.1. Thus, the least competitive Nash equilibrium price is 2.68 in each case. Our baseline hyperparameters (α_m, β_m) for both AMs give an average price of 5.03. If both AMs have the same hyperparameters, the average price ranges more generally between 4.00 and 5.37. If one includes cells with asymmetric hyperparameters for the two algorithms, the minimum price achieved is 3.86. In all cases this is far above the Nash equilibrium price.

We then reproduce Fig. 4 for all choices of hyperparameters where both dealers use the same values of α and β in the sets defined above. The results are in Fig. OA.2 to OA.4. The figures show in particular that the realized spread is higher in the no adverse selection case than in the case with adverse selection, across all values of α , β , and σ .

OA.5 Waiting for the experiment to “converge” can be misleading

In this section we explain why we choose to run experiments in which algorithms interact for a large but fixed number T of episodes, instead of waiting for the algorithms to play the same actions for a certain number of times, as is done in other papers in the literature.

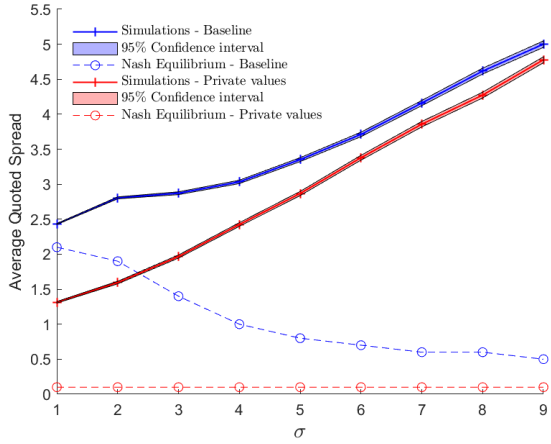
Consider the following two procedures for the numerical experiments:

- Fixed stopping time procedure: the algorithms play for a fixed number T of episodes.
- Random stopping time procedure: the algorithms play until they have both taken the same action for κ episodes in a row, then the procedure stops. The final episode is denoted \tilde{T} .

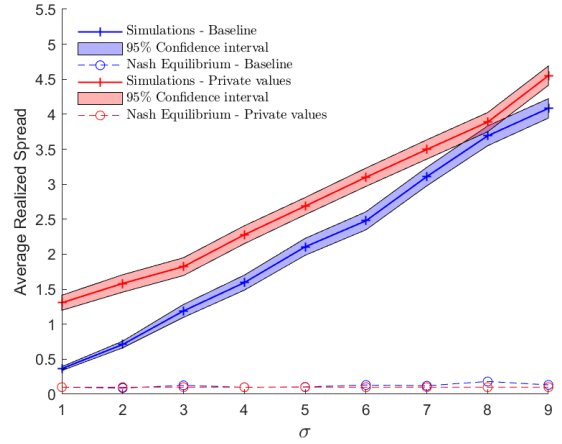
The random stopping procedure is in principle the appropriate thing to do if we know theoretically that the algorithms will eventually converge, in the sense that with probability 1 they will both play the same actions for every period after some random period. Then one can wait for the same actions to be repeated a large number of times κ , and if κ is large enough it is likely that the algorithms have indeed converged.

However, as we showed in Section A.5, the probability that our Q-learning algorithms converge in this sense is zero: there is a probability of 1 that an AM will change its optimal action if one waits for long enough. Then, the random stopping procedure implies that we are conditioning experimental observations on a specific path having been taken in the experiment. This may in principle bias the results.

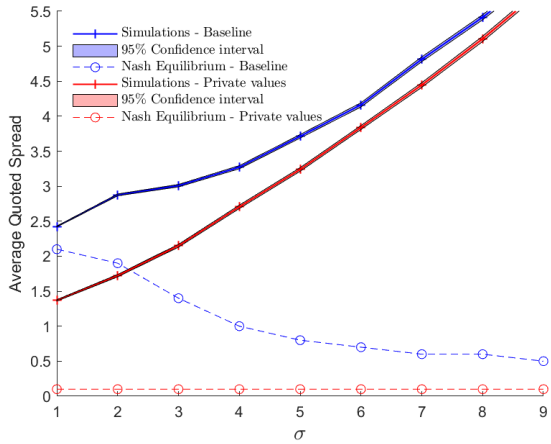
To better understand this point, we consider a very simple example in which the correct quantity to estimate can be computed theoretically. Assume there is only one Q-learning algorithm that can



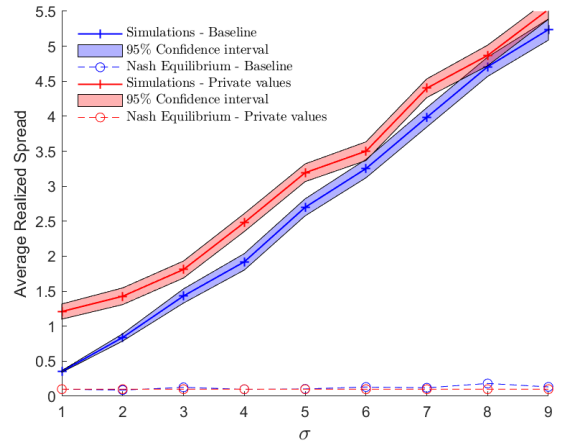
Quoted spread, $\alpha = 0.001, \beta = 5 \times 10^{-6}$



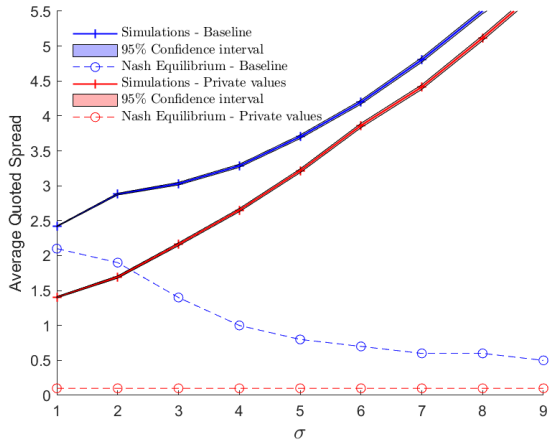
Realized spread, $\alpha = 0.001, \beta = 5 \times 10^{-6}$



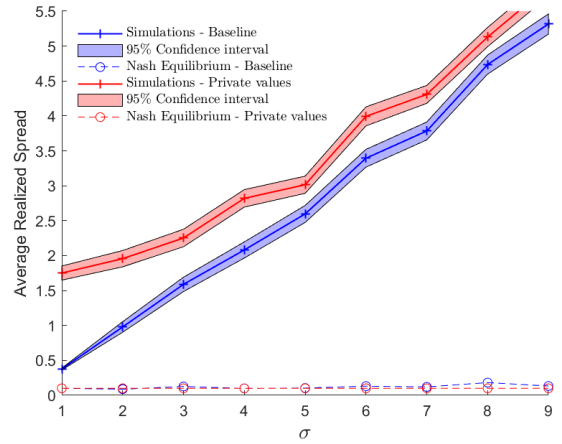
Quoted spread, $\alpha = 0.001, \beta = 8 \times 10^{-5}$



Realized spread, $\alpha = 0.001, \beta = 8 \times 10^{-5}$

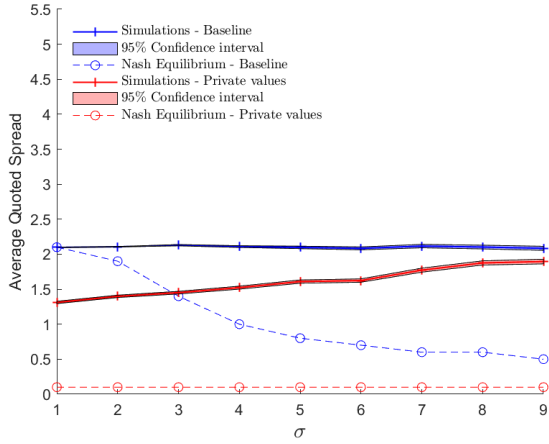


Quoted spread, $\alpha = 0.001, \beta = 3.2 \times 10^{-4}$

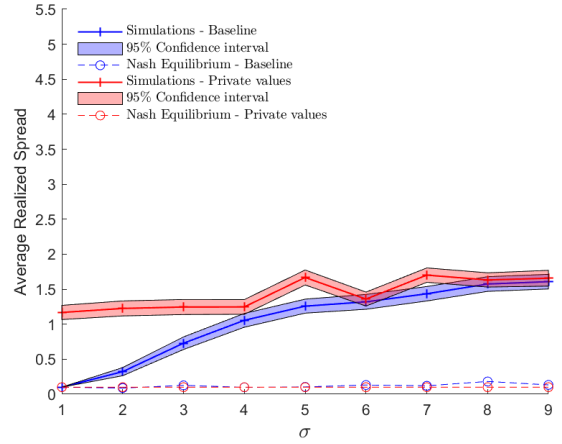


Realized spread, $\alpha = 0.001, \beta = 3.2 \times 10^{-4}$

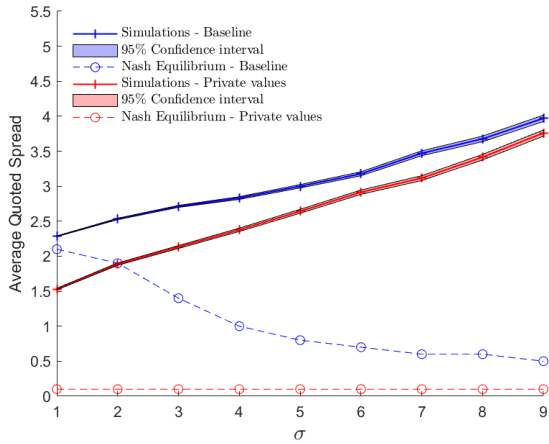
Figure OA.2: Average quoted spread and realized spread in the last episode, for different values of σ . Each line corresponds to a different parameterization of the algorithms, with a low value of α (0.001) and different values of β .



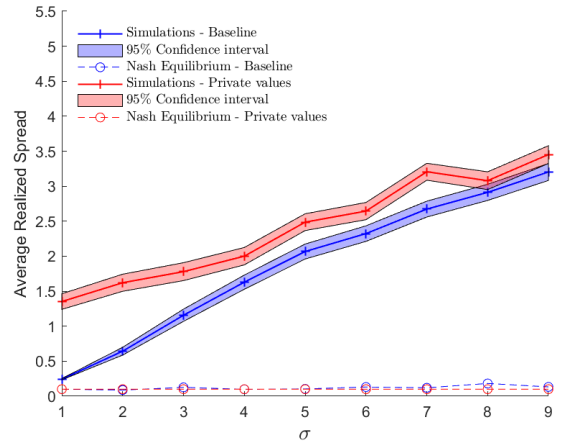
Quoted spread, $\alpha = 0.01, \beta = 5 \times 10^{-6}$



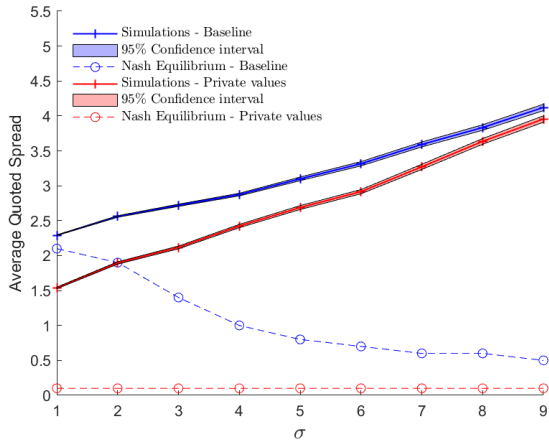
Realized spread, $\alpha = 0.01, \beta = 5 \times 10^{-6}$



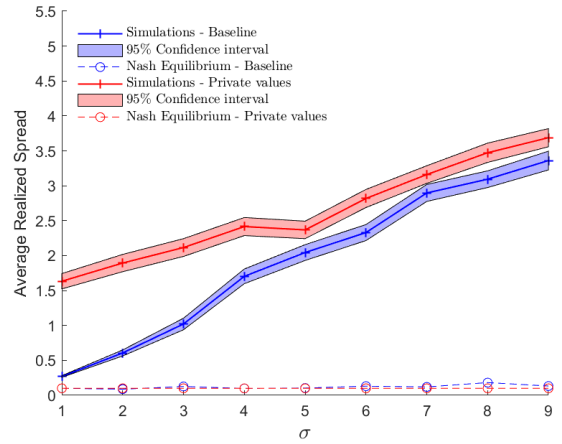
Quoted spread, $\alpha = 0.01, \beta = 8 \times 10^{-5}$



Realized spread, $\alpha = 0.01, \beta = 8 \times 10^{-5}$

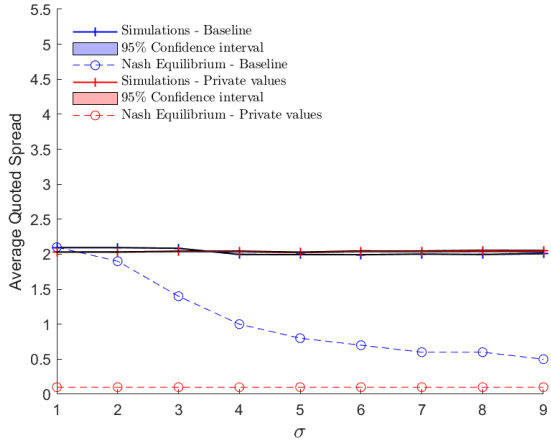


Quoted spread, $\alpha = 0.01, \beta = 3.2 \times 10^{-4}$

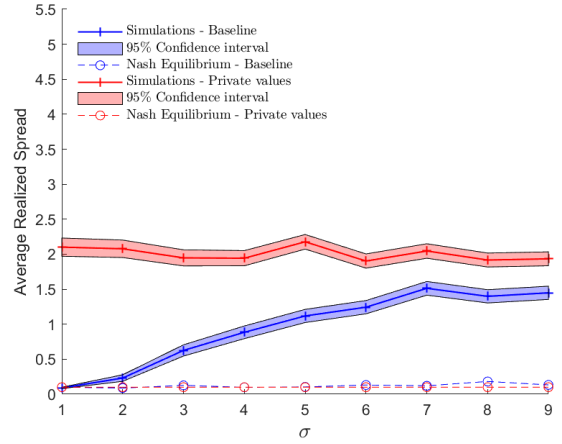


Realized spread, $\alpha = 0.01, \beta = 3.2 \times 10^{-4}$

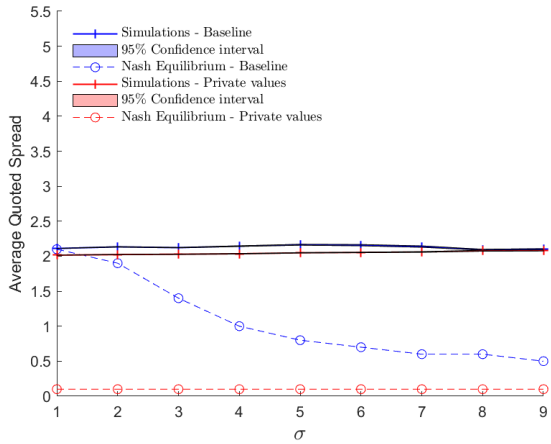
Figure OA.3: Average quoted spread and realized spread in the last episode, for different values of σ . Each line corresponds to a different parameterization of the algorithms, with the baseline value of α (0.01) and different values of β .



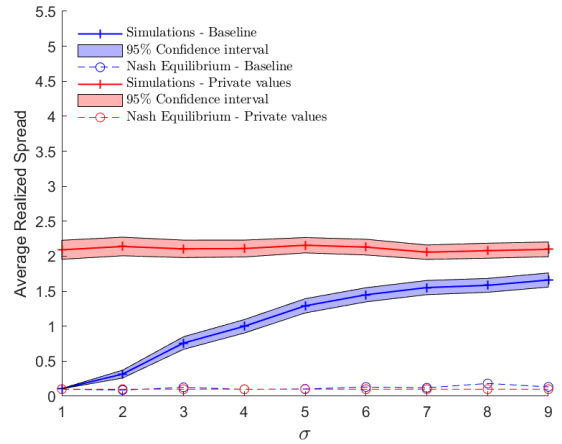
Quoted spread, $\alpha = 0.1, \beta = 5 \times 10^{-6}$



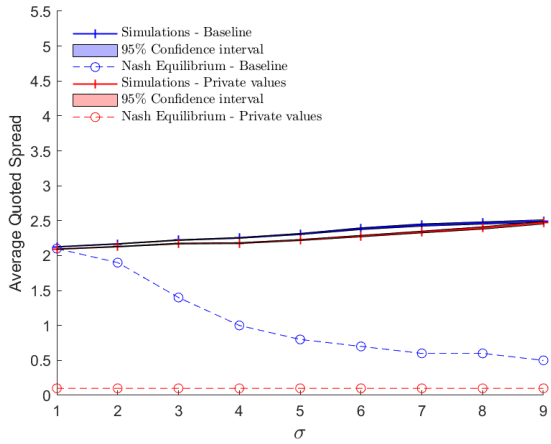
Realized spread, $\alpha = 0.1, \beta = 5 \times 10^{-6}$



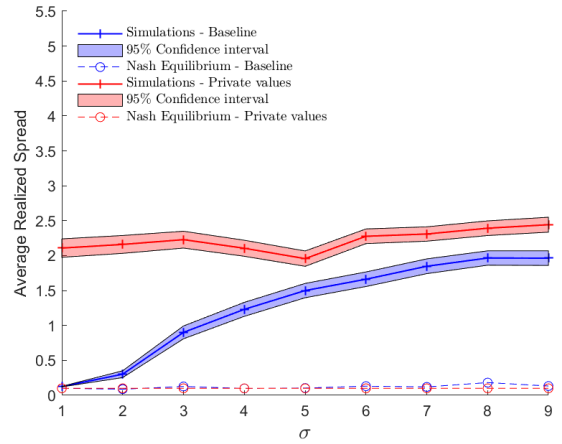
Quoted spread, $\alpha = 0.1, \beta = 8 \times 10^{-5}$



Realized spread, $\alpha = 0.1, \beta = 8 \times 10^{-5}$



Quoted spread, $\alpha = 0.1, \beta = 3.2 \times 10^{-4}$



Realized spread, $\alpha = 0.1, \beta = 3.2 \times 10^{-4}$

Figure OA.4: Average quoted spread and realized spread in the last episode, for different values of σ . Each line corresponds to a different parameterization of the algorithms, with a high value of α (0.1) and different values of β .

take two actions a_1 and a_2 . Action a_i gives a payoff π_i^h with probability p_i , and $\pi_i^l = 0$ with probability $1 - p_i$. Assume $\pi_1^h > \pi_2^h$. The algorithm does not experiment (or the probability of experimentation decays exponentially, so that in the long-run it becomes null), and updates with a rule similar to (8), with $\alpha = 1$.

Because $\alpha = 1$, the Q-matrix can only take four values:

$$Q_1 = \begin{pmatrix} \pi_1^h \\ \pi_2^h \end{pmatrix}, Q_2 = \begin{pmatrix} 0 \\ \pi_2^h \end{pmatrix}, Q_3 = \begin{pmatrix} \pi_1^h \\ 0 \end{pmatrix}, Q_4 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \quad (\text{OA.19})$$

Given that $\pi_1^h > \pi_2^h$, when the Q-matrix is Q_2 the algorithm will play a_2 . With probability p_2 the next value of the Q-matrix will be Q_2 again, and with probability $1 - p_2$ it will be Q_4 . Similarly, when the Q-matrix is Q_3 the algorithm will play a_1 , then the next value will be Q_3 with probability p_1 and otherwise Q_4 . When the Q-matrix is Q_4 the algorithm will play a_1 with probability $1/2$, leading to either Q_3 or Q_4 , and a_2 with probability $1/2$, leading to either Q_2 or Q_4 . Note that the only state of the Q-matrix that can lead to Q_1 is Q_1 itself, and only with a probability lower than 1. Hence, in the long-run the probability that the Q-matrix is Q_1 is zero.

The Q-matrix then follows a Markov process with 3 states Q_2 , Q_3 , and Q_4 , and the transition probabilities just described. It is easy to compute the stationary probability of each state, and then the stationary probability that the algorithm plays a_1 is:

$$\Pr(a = a_1) = \frac{1 - p_2}{2 - p_1 - p_2}. \quad (\text{OA.20})$$

Now we can test how each procedure will estimate $\Pr(a = a_1)$. We take $p_1 = 0.1$ and $p_2 = 0.9$, which gives $\Pr(a = a_1) = 0.1$. In words, the algorithm will constantly alternate between a_1 and a_2 , but in the long-run it will play a_1 10% of the time and a_2 90% of the time.

To implement the fixed stopping time procedure, we take $T = 50,000$. We simulate T periods for $K = 1,000$ experiments, and we record the percentage of experiments in which the algorithm plays a_1 or a_2 in the last episode.

To implement the random stopping time procedure, we let the algorithm run for 50,000 episodes, and then wait until the algorithm has played the same action for 100 episodes. We then stop the algorithm and record the action played in the last episode. We run $K = 1,000$ experiments and record the percentage of experiments in which the algorithm plays a_1 or a_2 in the last episode.

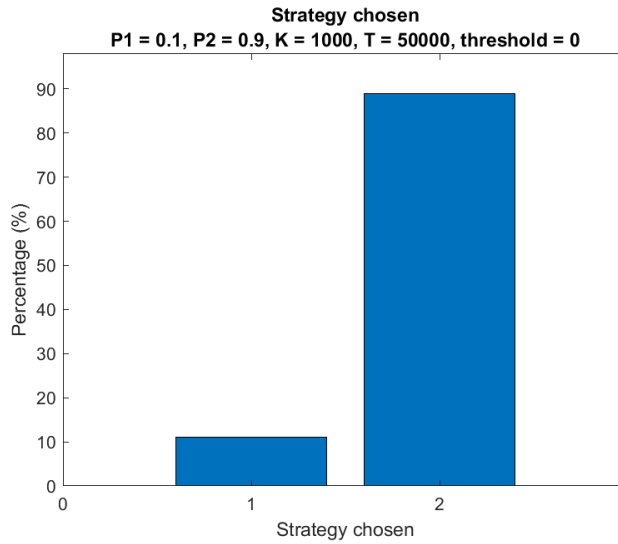
Figure OA.5 shows the outcome of our experiments. On Panel A we see that, using the fixed

stopping time procedure, the percentage of experiments that end with action a_1 is very close to the theoretical value of 10%. On Panel B instead, with the random stopping time procedure the percentage of experiments that end with action a_1 is 0%, so that the estimate of $\Pr(a = a_1)$ is significantly biased downwards.

The reason for this bias is that the second procedure conditions the observation on having the same action taken 100 times in a row. Conditionally on being in state Q_2 and playing a_2 , the probability of remaining in Q_2 is 0.9. The probability to remain in Q_2 for 100 episodes in a row is $0.9^{100} \simeq 2.65 \times 10^{-5}$, so that on average it will take $1/(2.65 \times 10^{-5}) \simeq 37,648$ repetitions of a sequence of 100 episodes to observe a constant action. For action a_1 , the probability of remaining in Q_3 is only 0.1, and the probability to remain in Q_3 for 100 episodes in a row is $0.1^{100} = 10^{-100}$, which is virtually zero. Hence, the random stopping time procedure picks up very particular histories, heavily biased towards action a_2 .

This example is clearly extreme and meant only for illustration. With lower values of α and actions that are less different we do not expect the two procedures to lead to radically different results. However, given that the random stopping procedure is in principle biased and is also typically much more computationally intensive, we recommend using the fixed stopping time procedure instead.

Panel A: Fixed stopping time procedure.



Panel B: Random stopping time procedure.

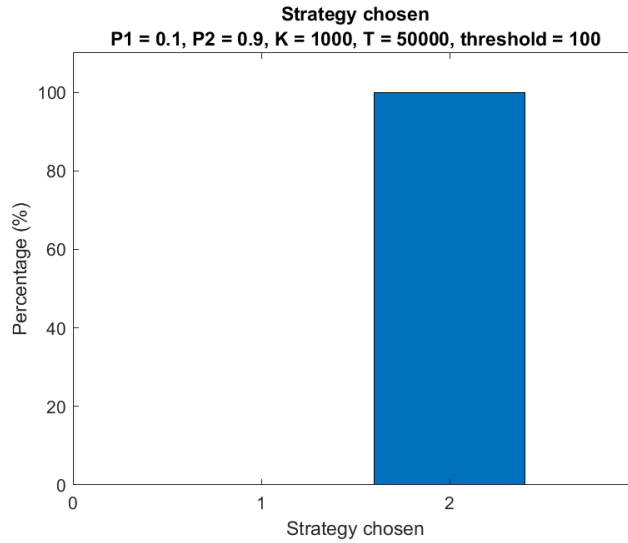


Figure OA.5: Percentage of experiments ending with actions a_1 and a_2 , using either the fixed stopping time procedure or the random stopping time procedure.

OA.6 Choice of α and β

We detail the statistical tests we conduct in Section 6. We denote $\theta_n = (\alpha_n, \beta_n)$ the hyperparameters chosen by dealer n . We take the perspective of dealer 1 and consider her total profit over T episodes as in eq. (7). For better readability we scale by the number T of episodes, which is a constant. If

dealer 1 uses hyperparameters θ_1^* and dealer 2 uses θ_2^* , we denote the expected per-episode profit of player 1 as:

$$u(\theta_1^*, \theta_2^*) = \mathbb{E} \left(\frac{1}{T} \sum_{t=1}^T \pi_{1,t} \right). \quad (\text{OA.21})$$

To study the stability of our baseline hyperparameters, $\theta_1^* = \theta_2^* = (\alpha_m, \beta_m)$, we do the following:

- First, we run $K = 1,000$ experiments in which both dealers use these hyperparameters. We denote $u_{1,k}^*$ the average per-episode profit obtained by dealer 1 in experiment $k \in \{1, \dots, K\}$ and compute:

$$\bar{u}^* = \frac{1}{K} \sum_{k=1}^K u_{1,k}^* \quad (\text{OA.22})$$

$$\bar{s}^* = \sqrt{\frac{1}{K-1} \sum_{k=1}^K (u_{1,k}^* - \bar{u}^*)^2}. \quad (\text{OA.23})$$

\bar{u}^* is dealer 1's estimate of her expected profits, and \bar{s}^* is her estimate of the standard deviation of this profit.

- Second, we assume that for another K' experiments, dealer 1 uses alternative hyperparameters θ' , while dealer 2's hyperparameters stay constant. We denote $u_{1,k'}$ the average per-episode profit obtained by dealer 1 in experiment $k' \in \{K+1, \dots, K+K'\}$ and compute:

$$\bar{u}' = \frac{1}{K'} \sum_{k'=K+1}^{K+K'} u_{1,k'} \quad (\text{OA.24})$$

$$\bar{s}' = \sqrt{\frac{1}{K'-1} \sum_{k'=K+1}^{K+K'} (u_{1,k'} - \bar{u}')^2}. \quad (\text{OA.25})$$

The question we ask is whether, based on K experiments with $\theta_1 = \theta_1^*$ and K' experiments with $\theta_1 = \theta'$, dealer 1 will have some statistical basis for preferring the new hyperparameters θ' to θ_1^* . We assume dealer 1 conducts a Welch test as follows. Compute:

θ'	$K' = 100$	$K' = 200$	$K' = 300$	$K' = 400$	$K' = 500$	$K' = 600$	$K' = 700$	$K' = 800$	$K' = 900$	$K' = 1000$
(α_l, β_l)	1	1	1	1	1	1	1	1	1	1
(α_l, β_m)	1	1	1	1	1	1	1	1	1	1
(α_l, β_h)	1	1	1	1	1	1	1	1	1	1
(α_m, β_l)	0.51	0.63	0.83	0.87	0.93	0.9	0.94	0.95	0.91	0.89
(α_m, β_h)	0.5	0.37	0.28	0.16	0.06	0.06	0.19	0.17	0.3	0.29
(α_h, β_l)	1	1	1	1	1	1	1	1	1	1
(α_h, β_m)	1	1	1	1	1	1	1	1	1	1
(α_h, β_h)	1	1	1	1	1	1	1	1	1	1

Table OA.2: Deviations from (α_m, β_m) : This table gives the p-values for a test of the null hypotheses that (α_m, β_m) and θ' give the same expected payoff to dealer 1, against the alternative hypothesis that θ' is more profitable, for all possible θ' and different values of K' .

$$\Delta \bar{u} = \bar{u}' - \bar{u}^* \quad (\text{OA.26})$$

$$s_{\Delta \bar{u}} = \sqrt{\frac{s'^2}{K'} + \frac{s^{*2}}{K}} \quad (\text{OA.27})$$

$$\nu = \frac{\left(\frac{s'^2}{K'} + \frac{s^{*2}}{K}\right)^2}{\frac{s'^4}{K'^2(K'-1)} + \frac{s^{*4}}{K^2(K-1)}} \quad (\text{OA.28})$$

$$t = \frac{\Delta \bar{u}}{s_{\Delta \bar{u}}}. \quad (\text{OA.29})$$

Under the null hypothesis H_0 that $u(\theta_1^*, \theta_2^*) = u(\theta', \theta_2^*)$, t should follow a Student distribution with ν degrees of freedom. Denoting F_ν the associated cdf, we can compute p the p-value of the test of H_0 against the alternative that θ' leads to higher payoffs, $u(\theta', \theta_2^*) > u(\theta_1^*, \theta_2^*)$, as $p = 1 - F_\nu(t)$. Table ?? below reports the p-values we obtain for different θ' and different values of K' .⁴⁹

As we observe in the table, for all values of K' dealer 1 can definitely reject the possibility that changing α is optimal. For β there is considerable uncertainty, as the average profits obtained with β_l, β_m , and β_h are very close to each other. For $K' = 100$ for instance the p-values are 50% for both potential “deviations”, which means that the three payoffs are basically undistinguishable. As K' increases, dealer 1 can more and more safely conclude that using a lower β does not bring superior profits. Changing for a higher β is less clear. For $K' = 500$ and $K' = 600$, dealer 1 will conclude that the null hypothesis that θ' is not more profitable is rejected at the 10% level. However, for higher values of K' the p-value increases again, up to around 30%.

We then repeat this exercise for all the 81 possible hyperparameter pairs (θ_1^*, θ_2^*) . For each of these 81 pairs, we look at the 8 possible deviations of each player, and record in each case the p-value of the test of the null hypothesis that the deviation is not more profitable. In Table ??, we

⁴⁹Note that for higher values of K' we simply ran additional simulations. This means that the 100 experiments used for the case $K' = 100$ are the first half of the 200 experiments used for $K' = 200$.

	(α_l, β_l)	(α_l, β_m)	(α_l, β_h)	(α_m, β_l)	(α_m, β_m)	(α_m, β_h)	(α_h, β_l)	(α_h, β_m)	(α_h, β_h)
(α_l, β_l)	0	0	0	0	0	0	0	0	0
(α_l, β_m)	0	0	0	0	0	0	0	0	0
(α_l, β_h)	0	0	0	0	0	0	0	0	0
(α_m, β_l)	0	0	0	1	0	0	0	0	0
(α_m, β_m)	0	0	0	0	0.29	0.4	0	0	0
(α_m, β_h)	0	0	0	0	0.4	0.59	0	0	0
(α_h, β_l)	0	0	0	0	0	0	0	0	0
(α_h, β_m)	0	0	0	0	0	0	0	0	0
(α_h, β_h)	0	0	0	0	0	0	0	0	0

Table OA.3: “Stable” choices of hyperparameters: For each pair of hyperparameter choices, the table gives the lowest p-value of the test that deviating to another set of hyperparameters does not lead to higher profits, across both players and all possible alternative choices of hyperparameters. A value of 0 means that there exists a deviation such that the null hypothesis that the deviation is not profitable can be rejected at a level close to 0%. A value of 1 means that no deviation allows to reject the null hypothesis at a level lower than 100%.

report the lowest p-value we found for each pair (θ_1^*, θ_2^*) . All tests were conducted with $K' = 1000$. We see in this table that 5 configurations in total are “stable” at the 0.25 confidence level, while all others are rejected at any level.

OA.7 Human vs Machine

In this Appendix we explore some possible outcomes from the interaction between an AMM and a human MM that we refer to as a non-algorithmic market maker (NAMM). Namely can a AMM who faces a NAMM survive despite not choosing what would be the rational best response to the other MM quote? The answer is likely to depend a lot on the knowledge of the non-algorithm about the environment. Formally, we consider the case with one algorithm and one non algorithm. The algorithm behaves like an AM in our set-up. We assume that the non-algorithm has rational expectations: she knows the price posted by the AM in each period and she knows the expected payoff she can obtain at each price. We first consider the case in which the non-algorithm adopts a myopic predatory strategy. That is he chooses, for each client, the price that maximizes her one-shot expected profit given the AMs’ prices. We refer to this rule as being myopic and show that this eventually induce the AMM to cycle its price above the one chosen by the NAMM. This allows the NAMM to make strictly positive profit and brings the AMM profits to 0.

We then show that the NAMM can achieve strictly higher payoff by adopting a less myopic strategy, that induces the AMM to post quotes just above the monopoly price and let it trade times to time. This leads to monopolist aggregate profit and allows the NAMM to reaping most- but not all- of it.

Benchmark As an example, we have ran a simulation, in a simplified setting, where the grid has all prices between 2 and 8, with a tick size of 1 and there are 2 competing market-makers. Given our parameterization, the Nash equilibrium price is unique and equal to 3, $a^m = 7$ and $a^- = 2$. Figure OA.6 reports the distribution of prices in the benchmark case with two AMs. Both AMs settle on a price of 4, one tick above the Nash equilibrium. Their average profit in the last episode is 0.1.

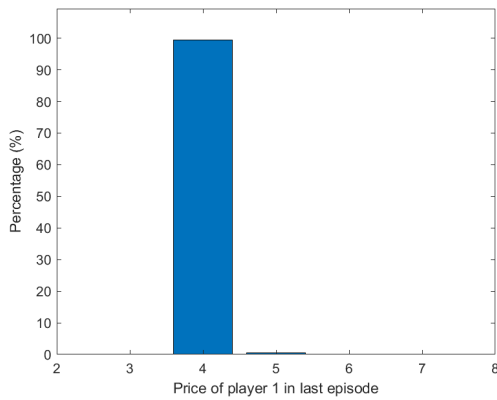
Myopic NAMM In Figure OA.7, we consider the case in which the first market-maker is a non-algorithm and follows the myopic strategy described previously. We observe that after about 10 thousands iterations, the AM stops trading because it is always undercut by the NAMM. Then, the AM’s price cycles through all prices between 4 and 8 and thus the final distribution of this price across experiments is roughly uniform (see Panel c of Figure OA.7). Thus, even though the AM keeps posting prices, it does not participate to trading. In fact all orders are executed by the NAMM, who on average earns a profit of 0.5 (see the R.H.S panel in panel a) of Figure OA.7). As the figure shows, the NAMM’s strategy pays-off because it enables the NAMM to trade at prices above 4 (the price obtains when both participants are AMs). In fact, across experiment, the price posted by the NAM is roughly uniformly distributed in the range $[4, 8]$. In this situation the AMM chance of getting a trade negligible.

Non-myopic NAMM We now show that the NAMM has one strategy that dominates the myopic strategy and that leaves some small profits to the AMM. Intuitively, the NAMM can “train” the AMM to post high prices and leave it sufficient profits so that the AMM does not ‘realize’ that choosing a lower price could yield larger average profits. For any given round t , the strategy is as follows. Let a^{m+} be the first price on the grid above the monopoly price a^m . If the best greedy price (and hence the posted by the AMM) is equal to a^{m+} , and the Q-value of a^{m+} , $Q_t(a^{m+})$, is such a^{m+} remains the greedy price even if the NAMM undercut by posting a^m , then the NAM post a^m . If instead by posting a^m the NAMM would induce in AMM a greed price below a^{m+} , then the NAMM posts a^{m+} so that he shares profits with the AMM.

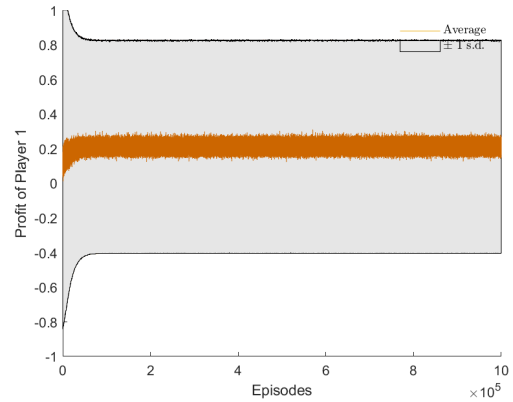
Intuitively, this strategy “teaches” the AM to play a price just above the monopoly price at a relatively low cost (the NAMM just needs to share profits at a price of a^{m+} from time to time). Figure OA.8 shows that for the parameter values considered in our simulations, this strategy works well. It results in a situation in which the NAMM ends up posting the monopoly price of 7 most of the time while AM2 posts a price of $a^{m+} = 8$ most of the time across experiments in the last

episode.⁵⁰ As a result, the NAMM obtains an average profit of 0.76 on average in the last episode, which is strictly above the average profit of the myopic strategy describes above. The AM is not completely excluded from the market and makes a small profit (see Figure OA.8).

In sum, this analysis shows that if a non-algorithm has rational expectations, it can drive out the algorithm from the market with a predatory strategy. However, doing so is dominated by a strategy that trains the algorithm to post high prices by sharing profits, from time to time, with the algorithm.



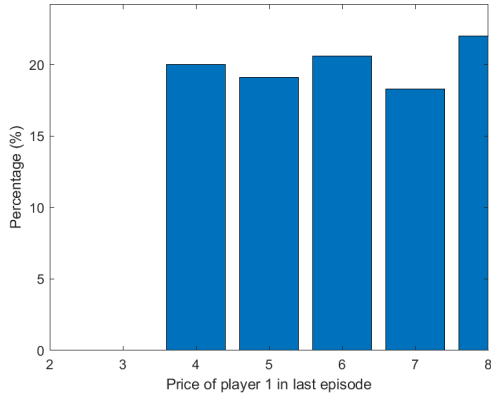
(a) AMMs quotes



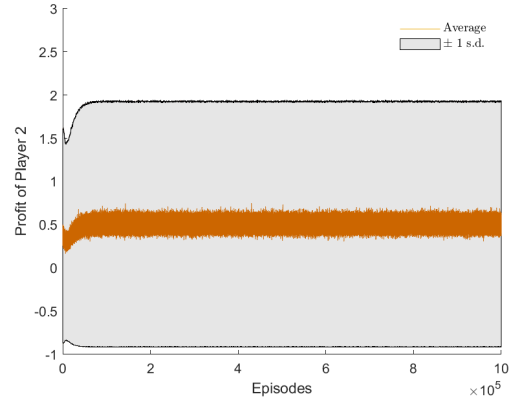
(b) AMMs profits

Figure OA.6: Final price and the evolution of payoffs for two competing AMMs.

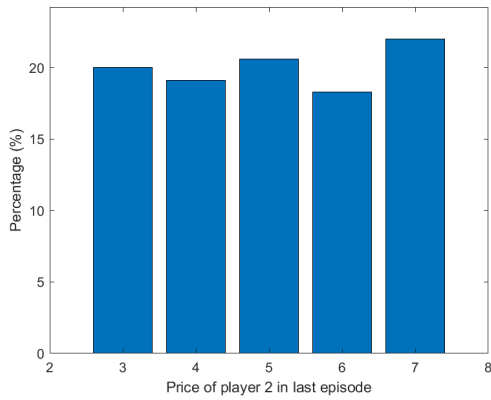
⁵⁰In the last episode, the AMM plays 8 in 992 out of 1,000 experiments while the NAMM plays 7 (and gets the monopoly profit) in 988 out of 1,000 experiments.



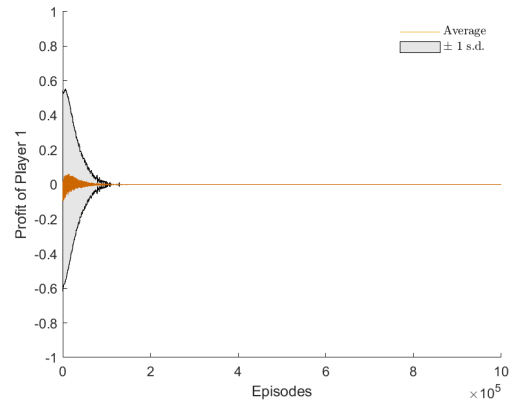
(a) NAMM 1 quotes



(b) NAMM 1 profit

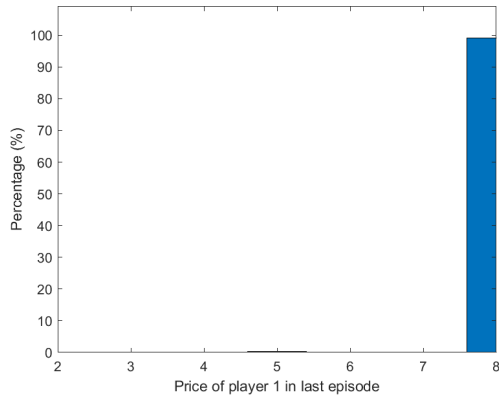


(c) AMM 2 quotes

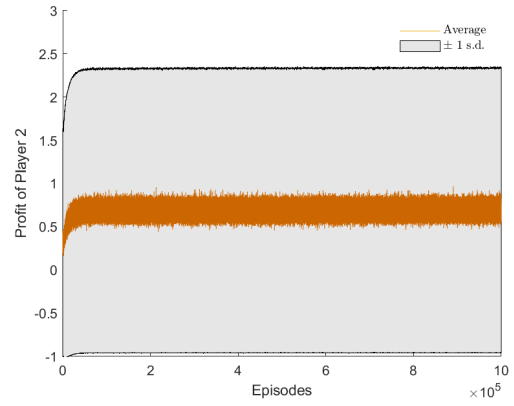


(d) AMM 2 profits

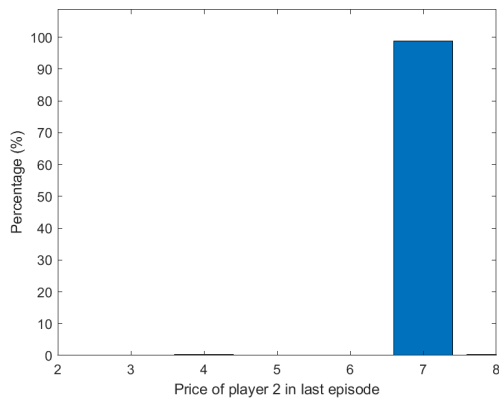
Figure OA.7: Final price and the evolution of payoffs for a myopic NAMM and AMM .



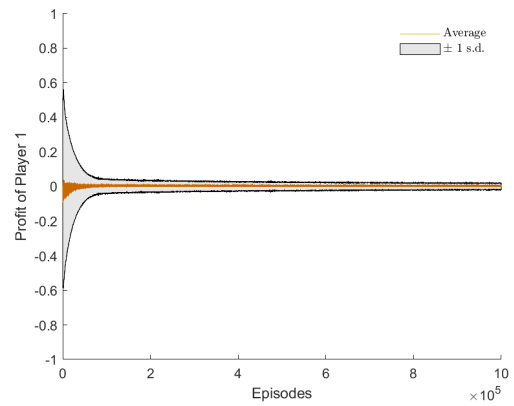
(a) NAMM 1 quotes



(b) NAMM 1 profit



(c) AMM 2 quotes



(d) AMM 2 profits

Figure OA.8: Final price and the evolution of payoffs for a myopic NAMM and AMM .