Algorithmic Pricing and Liquidity in Securities Markets*

Jean-Edouard Colliard, Thierry Foucault, and Stefano Lovo

November 20, 2025

Abstract

We study "Algorithmic Market Makers" (AMs) that use Q-learning algorithms to set prices for a risky asset. We find that while AMs successfully adapt to adverse selection, they struggle to learn competitive pricing strategies. This failure is driven by limited experimentation and noisy feedback regarding the profitability of undercutting a competitor. Consequently, an increase in AMs' profit volatility tends to result in less competitive market outcomes. These features leave identifiable patterns: for example, AMs earn higher rents in the absence of adverse selection, and their bid-ask spreads respond asymmetrically to symmetric shocks to their costs.

Keywords: Algorithmic pricing, Market Making, Adverse Selection, Competition, 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 Catherine, Alex Chinco, Winston Dou, Vincent Glode, Itay Goldstein, Terrence Hendershott, Yan Ji, Anton Lines, Lin Peng, Nick Roussanov, Chaojun Wang, Yajun Wang, Liyan Yang, Mao Ye, Bart Yueshen, participants to "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, the 2025 IESE Banking Initiative Workshop, and seminar participants at Aalto University, Bank of England, Bank of France, Baruch College, Bayes Business School, Bundesbank, Cornell University, CRESE, ESMA, Frankfurt School of Management, HEC Paris, Hong-Kong University, HKUST, 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, Lucie Bois, 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).

Introduction

Prices in securities markets are increasingly set by algorithms. For instance, Brogaard *et al.* (2014) and Chaboud *et al.* (2025) find, respectively, that about 42% and 60% of trades in stocks and currencies in their sample take place at prices set by algorithms. In US treasury markets, principal trading firms (PTFs), which also rely on algorithms, account for 21% of total trading volume (Brain *et al.*, 2018). Until recently, these algorithms relied on predefined rules. However, progress in Artificial Intelligence (AI) enables using self-learning pricing algorithms–similar to those used for autonomous gameplay (e.g., chess or Go) and self-driving cars. Reflecting this shift, mentions of AI tools in patent applications related to algorithmic and high-frequency trading have risen significantly since 2015 (IMF, 2024).

This evolution raises concerns about the impact of AI-powered algorithms on market efficiency, liquidity, and stability.[1] To assess whether such concerns are well-founded, it is essential to develop a deeper understanding of algorithmic behavior in financial markets, in order to better predict and explain their influence on market outcomes (Goldstein *et al.*, 2021).[2] Our paper contributes to this research agenda through an experimental approach.

In our experiments, AI-powered algorithms play a market making game similar to the one proposed by Glosten and Milgrom (1985). These algorithms exhibit two key features of reinforcement learning: (i) they update the estimated value of an action based on the observed outcomes when that action is taken, and (ii) they engage in limited experimentation. The main finding from our experiments is that such algorithms fail to learn competitive pricing strategies due to noisy feedback regarding the profitability of undercutting a competitor. As a result, changes in the game's parameters that increase the volatility of algorithmic profits lead to less competitive prices, and hence wider spreads. We show that this mechanism leaves identifiable footprints, that is, patterns that are not predicted by standard solutions of the game.

In the baseline version of the market making game, two dealers receive quote requests from clients who wish to buy one share of a risky asset. Dealers are uninformed about the payoff of the

[1] For instance, former SEC Chairman Gary Gensler has expressed concern that AI-driven trading algorithms 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.

[2] Goldstein *et al.* (2021) write: *"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."*

asset and simultaneously respond with an offer. Each client buys if the best offer is less than 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 dealers' prices constant, a client is more likely to buy the asset when its payoff is high than when its payoff is low. Dealers are therefore exposed to adverse selection. After each client arrival, the asset pays off, and dealers receive their realized profit—that is, the difference between the selling price and the asset's payoff if the dealer sold the asset, and zero otherwise.

Importantly, dealers 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 dealers...). They respond to clients' requests using Algorithmic Market Makers (AMs), which learn to set prices using Q-learning, a foundational model for reinforcement learning algorithms. More specifically, for each new request, each algorithmic market maker 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 before the client's decision.

Q-learning is therefore an iterative method to make decisions and learn the profits associated with a set of possible actions (here, AMs' prices). It embeds two key principles of any reinforcement learning algorithm. First, each new action generates new data (in our case realized profits) used by the algorithm to update its assessment of the value of this action for the decision maker. This update is controlled by the so called "learning rate", the weight on new data in the algorithm's update.

Second, to get data about an action, the algorithm must try it. This is the reason why the algorithm is programmed to experiment, that is, sometimes choose an action different from the greedy one. Such experimentation is the cost to pay to learn that another action might in fact have even more value for the decision maker. 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 becomes more accurate. Thus, experimentation is limited: Q-learning algorithms are in general designed to have a decaying experimentation rate over time.

We assume that dealers train their algorithms over a long but finite sequence (one million) of

clients. For a given parametrization of the market making game (e.g., the volatility of the asset payoff or the dispersion of clients' private valuations), the price for a given 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 parametrization of the market making game, we run $1,000$ different simulations (experiments) and we focus on the average long-run outcomes across these experiments. We interpret these outcomes as representative of the AMs' behavior after their training phase.

In general, we observe that, during their training phase, AMs begin by lowering their prices, as if they were undercutting each other. However, this process eventually stops and the AMs settle on the same price well above the "Glosten-Milgrom price" (the price predicted by the standard economic analysis of the market making game). Thus, AMs learn not to be adversely selected but they do not fully learn to be competitive. Moreover, we observe that AMs' average profit conditional on a client's buy (their average realized spread) is larger when the dispersion of clients' liquidity shocks increases or when there is no adverse selection (everything else equal).

This behavior cannot be attributed to AMs learning to play a collusive equilibrium, as the design of our experiments rules out the possibility of tacit collusion. Instead, we explain it by the very structure of the AMs' learning process. AMs learn to undercut by observing the *realized* profit from doing so, rather than the true expected profit. Even when the expected profit is high, the realized profit may be low due to randomness in market outcomes. In other words, AMs receive noisy feedback about the value of undercutting, which hinders their learning.

For example, suppose both AMs converge on the same price p_0 above the Glosten-Milgrom price. If AM_1 experiments by undercutting slightly, quoting $p_0 - tick$, the action is profitable on average since p_0 exceeds the competitive price. However, in a given instance, AM_1 may be "unlucky": the client can choose not to trade because her valuation is too low, resulting in zero profit, or the client trades, but the asset's realized payoff turns out to be high, yielding a small or even negative profit. In such cases, AM_1 receives a misleading negative signal about the value of undercutting.

Intuitively, AM_1 is more likely to learn quickly that undercutting is profitable when (i) the average signal is strong (i.e., the net gain from undercutting is large), and (ii) the noise is low (i.e., the variance of the net gain is small). This simple heuristic goes a long way in explaining

the AMs' behavior. First, it explains why we observe a decline in prices in the early stage of the AMs' training, and why this process eventually stops before they reach competitive prices. Indeed, when prices are high, the average gain from undercutting is relatively large compared to the noise. Thus, in the early phase of their training, the AMs quickly undercut each other. However, as prices become lower, undercutting has a smaller signal to noise ratio. Hence, it requires more experiments for the AMs to realize that undercutting is indeed profitable. But, as explained previously, the AMs' experimentation rate decays over time. At some point, their experimentation rate is simply too small for them to learn the value of undercutting further.

This issue becomes more pronounced when the dispersion of clients' private valuations is high or when there is no adverse selection, as both conditions increase the variance of AMs' profits at a given price. As a result, the feedback received by the AMs becomes noisier, which impairs their ability to accurately assess the profitability of undercutting. This explains why, in such environments, AMs tend to stabilize at prices where the gain from undercutting is even larger, that is, at prices that are even less competitive.

Consistent with this mechanism, we find that AMs' long-run prices are more competitive (closer to the Glosten-Milgrom price) when their experimentation rate decays more slowly. This naturally raises the question: why don't dealers simply choose a very high experimentation rate for their AMs? To address this, we extend our framework by allowing dealers to select both the learning and experimentation rates for their respective AMs. We show that, under this extension, there is no race to the top in experimentation rates, for two reasons. First, as previously discussed, prolonged experimentation is costly: it involves taking actions with relatively low Q-values, which may correspond to genuinely low expected payoffs. Second, if one dealer increases his AM's experimentation rate, he boosts his profits only temporarily because the competing AM eventually adjusts its pricing in response, ultimately driving both AMs toward lower prices and reducing profits for both.[3]

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 their effects on trading outcomes.

[3]Importantly, the AMs keep learning even when they stop experimenting. Indeed, they keep updating the Q-value of the greedy action since their learning coefficient is fixed. Thus, if AM_1 undercuts AM_2, the latter will eventually discover that the price on which it settled has a low Q-value and will at some point match or undercut the price set by AM_1. Thus, AM_1's boost in profits due to a greater ability to undercut when it is profitable to do so is only transient.

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.

Second, we predict and confirm experimentally that a reduction in the tick size can result in larger quoted and realized spreads. The reason is the following. On the one hand, 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. On the other hand, the AMs' choice set becomes larger when the price grid is finer. Thus, the AMs' (fixed) experimentation capacity is spread out over a larger number of actions and, as a result, they experiment each price a smaller number of times. This effect slows down AMs' ability to learn the value of undercutting prices at which they are tied. When the tick size becomes small enough this effect dominates and AMs' prices become less competitive. This finding shows how standard policy recommendations for market design might be altered when prices are set by self-learning algorithms.

Third, AMs' responses to symmetric shocks in their expected profits after the training phase are asymmetric: they increase their bid-ask spreads in response to negative shocks but do not adjust them downward in response to positive shocks. This asymmetry arises because a positive shock (e.g., a decrease in adverse selection costs) increases the AMs' estimated average payoff associated with their long-run greedy price, reinforcing their preference for this price. In contrast, a negative shock lowers their estimate of the profit at the long-run greedy price. As this revised estimate falls below the Q-value of a higher price, the AMs eventually switch to quoting the higher price.

Finally, we show that the AMs' learning process influences the dynamics of prices before the asset pays off. That is, we consider an extension of the market making game in which AMs sequentially receive requests from two clients before the asset payoff is realized. In this setting, we allow the AMs' quotes for the second client to depend on the trading outcome with the first client. We observe that AMs increase their quotes following a purchase by the first client and decrease them when the first client does not trade, mirroring the direction of adjustment in Glosten-Milgrom prices. However, relative to these prices, AMs tend to overreact in the former case and underreact in the latter. Consequently, AMs offer even less competitive prices to the second client than to the first, so that,

in sharp contrast to the Glosten-Milgrom prices, their price in the second period exceeds that in the first on average.

These patterns stem from the fact that learning becomes more difficult for AMs as the number of states on which they can condition their actions becomes larger. Indeed, for each possible outcome in the first period (buy/no buy), AMs have fewer opportunities to learn how to set their price in the second period than in the first. For instance, if the first client buys 40% of the time then AMs have only $400,000$ $(600,000)$ opportunities to learn how to set the price for the second client after a first client's buy (no buy). In contrast, they have 10^6 opportunities to learn how to set the price for the first client. This effectively reduces AMs' ability to learn through experimentation in the second period, which leads to less competitive prices, for the same reasons as in the one period case. More generally, this finding suggests that AMs' spreads should be less competitive in states that are rarely observed.

In the next section, we position our contribution within the existing literature. Section 2 presents the market making game. Section 3 describes the Q-learning algorithm and outlines our experimental design. In Section 4, we report our experimental findings and explain the AMs' behavior in the experiments. This explanation leads to several testable implications, which we present in Section 5. In Section 6, we endogenize the AMs' hyperparameters. Section 7 concludes. Formal derivations are provided in the Appendix, and additional results are available in the Online Appendix.

1 Contribution to the Literature

Our paper is related to the emerging literature on algorithmic pricing (e.g., Calvano *et al.* (2020), Klein (2021)). This literature has focused on product markets (see, e.g., OECD (2017)). To our knowledge, we are the first to study how Q-learning algorithms for market making behave in an environment with uncertainty about demand and costs, due to asset volatility and adverse selection.[4] These are key features of trading in securities markets, and they drive our new implications on algorithmic market making.

One branch of the literature on algorithmic pricing shows that Q-learning algorithms can learn

[4]Cont and Xiong (2024) and Guéant and Manziuk (2019) study theoretically how market makers using reinforcement algorithms set prices in the face of inventory holding costs. However, there is no adverse selection in their framework. Hansen *et al.* (2021), Cartea *et al.* (2022b), or Wilk (2022) study selling algorithms that face a stochastic demand but without adverse selection.

to play collusive strategies (e.g., Calvano *et al.* (2020)). Such strategies are ruled out in our experiments. Our finding that these algorithms can fail to learn to undercut is in line with what Abada *et al.* (2024) call "collusion by mistake" (see also Asker *et al.* (2024)). Dou *et al.* (2023) study how informed traders using Q-learning algorithms behave in a Kyle (1985) environment. Their analysis and ours are complementary. They study under which conditions Q-learning algorithms fail to learn to collude, while our experiments focus on why algorithms can fail to learn to compete. Moreover, we focus on market makers' pricing behavior while Dou *et al.* (2023) focus on informed investors' order submission strategies.

Our paper also contributes to the broader literature on algorithmic trading in securities markets. The theoretical literature on this issue (e.g., Biais *et al.* (2015), Budish *et al.* (2015), Foucault *et al.* (2016), Menkveld and Zoican (2017), Baldauf and Mollner (2020)) has mainly focused on how the speed with which algorithms respond to information affects liquidity suppliers' exposure to adverse selection, using traditional workhorse models (Glosten and Milgrom (1985) or Kyle (1985)).

In contrast to this literature, our approach is experimental. In the same way in which economists use experiments to understand human behavior and how it can differ from standard economic analysis, we run experiments to understand algorithmic behavior in a controlled environment.[5] The Nash equilibrium of the market making game played by algorithms serves as a benchmark to identify "surprising" outcomes, that is, experimental outcomes that are are not predicted by the standard economic analysis of the game. This approach contributes to the research agenda described in Goldstein *et al.* (2021).

Our experimental findings suggest that modeling algorithmic behavior calls for new learning models (as suggested by O'Hara (2015)). These could build on insights from the literature on reinforcement learning in decision environments and games. For instance, an important feature of our algorithms is that they do not experiment all actions forever. Thus, in the long-run, AMs have correct estimates of the payoffs associated with the actions they take and these actions appear optimal given AMs' estimates. However, their estimates for the actions that are not selected in the long-run can be incorrect. This feature is a well-known property of algorithms for the bandit

[5]In experimental economics, researchers assume that the behavior of human subjects in the experiments is representative of human behavior more generally. Similarly, we assume that the behavior of Q-learning algorithms is representative of reinforcement learning algorithms in general because, as explained in the introduction, Q-learning embeds two key features of reinforcement learning algorithms: (i) adjustment of the value of an action based on outcomes when this action is taken and (ii) limited experimentation.

problem, even in the Bayesian case where a solution is known (Gittins, 1979), and of more general reinforcement learning environments (Easley and Kiefer, 1988). Theoretical analyses of the bandit problem could therefore prove useful to understand algorithms' behavior in securities markets.

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 do not play them. This type of phenomenon is related to why our algorithms fail to undercut each other down to the competitive price. This suggests that using concepts from the literature on learning in games could prove powerful to understand algorithmic behavior.[6]

In any case, our objective is not to study which algorithms or behavioral processes converge to Nash equilibria (as, for instance, Cartea *et al.* (2022a) and Cartea *et al.* (2022b)). Rather, assuming that current pricing algorithms can be modeled using Q-learning, we focus on how these algorithms respond to changes in their economic environment–a comparative statics question. This is a useful step to understand how to design markets in the presence of algorithms, as highlighted by our analysis of the effect of the tick size.[7]

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. This price is a useful benchmark for interpreting algorithmic market makers' behavior.

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

[7] In the same vein, Pouget (2007) and Banchio and Skrzypacz (2022) 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.

2.1 The Market Making Game with Adverse Selection

One investor ("client") wants to buy one share of a risky asset.[8] 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, 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 $G(.)$. 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, 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 (or trading volume). It is 1 if the client buys the asset and 0 otherwise. Dealer n's realized trading volume is

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

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

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

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 high than when its payoff

[8] We restrict attention to the case in which the client is a buyer. This simplification makes the analysis more tractable without altering the underlying economics of the problem. In the benchmark model, extending the framework to accommodate a selling client is straightforward. However, when prices are set by algorithms, introducing two-sided market making would require additional assumptions about how the algorithms are programmed (e.g., whether they set bid and ask prices jointly or independently). To keep the experiments clear and focused, we sidestep this complication.

is low. 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 its payoff is v. We have

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

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

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 without adverse selection. Hence, 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.[9]

The last point is important. It allows us to compare experimental outcomes with and without adverse selection, holding all other 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. This

[9] The 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 $\tilde{v}$.

approach allows us to uncover unique footprints of algorithms relying on trial-and-error methods to set prices, that is, patterns that are not predicted by economic analysis of the market making game.

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

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

which can be written as:

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

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 asset value, 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 conditionally 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

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

The zero expected profit 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.[10] In our setting, using the standard Bertrand logic, a^* is the unique Nash equilibrium of the market making game.[11]

The Glosten-Milgrom price has several important properties. First, the expected (half) quoted

[10] The Glosten-Milgrom price is the solution of the fixed point problem (6) for which there is no closed-form solution given our specification of $G(.)$. 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.

[11] The Bertrand logic is as follows. Suppose (to be contradicted) 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 $\frac{N-1}{N}$ his expected profit since he captures the entire demand while the decline in his expected profit per share is infinitesimal. Hence, "$a > a^*$" cannot be an equilibrium.

spread, $\overline{QS} := \mathbb{E}(a^{\min} - \tilde{v})$, is strictly positive and just equal to the 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 brackets in (5)) and is often used to this end by empiricists. Third, as shown in Figure 2 and proved formally in Appendix A.2, the expected quoted spread (or, equivalently, the adverse selection cost) 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.

[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 realized spread are 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) profits. 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 $\pi_{n,t}$, he would look for a pricing policy from $t = 1$ to $t = T$ that maximizes his 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 he will have to select a price in a set $\mathcal{A}$ for each of the T episodes, and he will only observe his realized profit $\pi_{n,t}$ at the end of each episode t. He knows neither the structure of the trading environment, nor that of the competition, nor, more generally, the stochastic mapping from the prices he sets to the payoffs he 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: he can try different prices ("arms") to estimate the average payoffs associated with each price. Possibly, these average payoffs vary over time.[12]

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.[13] We refer

[12] See, e.g., Easley and Rustichini (1999) for a formal treatment of the problem of a decision-maker who is not able to formulate Bayesian priors over his environment and has to pick actions based on experience.

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

to such dealers as Algorithmic Market Makers (AMs).

We restrict AMs to choosing their quotes in $\mathcal{A} = \{a_1, a_2 ... a_M\}$, where each a_m is a possible ask price.[14] 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.[15] The m^{th} entry of the matrix, denoted $q_{m,n,t}$, represents the estimate by the n^{th} AM (henceforth 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 $[\underline{q}, \overline{q}]$ 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): (i) the learning coefficient, $\alpha \in (0, 1)$ and (ii) the experimentation rate, $\beta > 0$, which determines the probability $\epsilon_t := e^{-\beta t}$. Given this parametrization, 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 estimates 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 try are drawn independently across dealers. We

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

[15] In 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.

denote $\bar{a}_t = (a_{1,t}, a_{2,t}...a_{n,t})$ the vector of prices quoted by all AMs in episode t 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, $\tilde{w}_t^C$, as described in Sections 2.1 and 2.2. We then determine the client's valuation, 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 $\pi_{n,t} = \Pi(a_{n,t}, \bar{a}_t, v_t^C, v_t)$, as given by (2). In particular, the AMs quoting $a_t^{\min}$ share the profit (or loss) from selling the asset, while the 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:

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

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.[16]

Parameters α and β are fixed parameters (sometimes called "hyper-parameters"). The parameter β controls the speed at which ϵ_t decays over time. A larger β means that ϵ_t decays faster and AMs switch more quickly to exploiting. The parameter α controls the sensitivity of the AMs' estimates to new observations. The literature on Q-learning cautions that a too large α leads to unstable estimates (consider the extreme case $\alpha \to 1$), whereas a too small α makes learning slower (in the extreme case in which $\alpha \to 0$ there is no learning).

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 rein-

[16]The 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.

forcement 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

Our experimental design is guided by two important considerations. First, for a given parametrization of the market making game $(\Delta_v, \sigma, \mu, \mathbb{E}_\mu(\tilde{v}))$, 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.4 for a formal analysis).[17] 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.[18]

In our baseline experiments, we choose the parameters of the market making game as: $\Delta_v = 4$,

[17] Watkins 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 OA.2. We still observe a distance with the predictions of the Glosten-Milgrom benchmark, due to problem (iii).

[18] Other 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.4. 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.

$\sigma = 5$, $v_H = 4$, $v_L = 0$, $\mu = 0.5$, and $N = 2$ (same parameters as in Figure 2). 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 Glosten-Milgrom prices are in the range 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 $\overline{q} = 6$, so that all values of the initial Q-matrix are above the maximal payoff a dealer can get in a given period.[19] 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.[20]

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 check that the distribution of $a_t^{min,k}$ has converged using a Kolmogorov-Smirnov test.[21] We then compute the empirical analogs of the quoted spread $\overline{QS}$ and the realized spread $\overline{RS}$ defined in Section 2.3 by recording in the last episode $QS_T^k = a_T^{min,k} - \mathbb{E}[\tilde{v}]$ and $RS_T^k = a_T^{min,k} - v_T^k$, respectively, and then taking the average over the K experiments.[22]

The Glosten-Milgrom price might not be on the grid the AMs have to use, in which case the 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.5). 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 price that can be played in a 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.

[19] This specification is common in the literature on Q-learning to guarantee that all actions are chosen sufficiently often to overcome the initial values of the Q-matrix. See in particular Asker *et al.* (2024). 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.

[20] Each 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.

[21] We test the null hypothesis that the samples $\{a_t^{min,k}\}_{k=1...K}$ and $\{a_T^{min,k}\}_{k=1...K}$ come from the same distribution, using a Kolmogorov-Smirnov test. After $t = 500,000$ episodes the p-value of the test is 0.9995. 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.

[22] The 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.

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 (in all 1,000 experiments both AMs have the same greedy price in the last episode, but this price differs from one experiment to another). As the figure shows, AMs' quotes vary across experiments (standard deviation of 0.73). The modal greedy price in the last episode is 4.60 and the mean is 4.97. In all experiments, the greedy price is above the Glosten-Milgrom price ($a^* = 2.68$) and the least competitive Nash equilibrium, 2.8 (about 1 tick above the Glosten-Milgrom price). At any price above 2.8, an AM can therefore, in theory, obtain a strictly larger expected profit 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.[23]

[INSERT FIGURE 3 ABOUT HERE]

Panel B of Figure 3 shows the evolution over episodes 1 to T of the average greedy price (averaged over the K experiments) and the greedy prices one standard deviation away from the average (remember that the greedy prices vary across experiments due to randomness in the trading history). In the first part of the learning process (roughly the $20,000$ first episodes), the average greedy price decreases. Thus, in the early phase of their training, AMs learn to reduce their price to attract clients and increase their market share. However, over time, the greedy price declines less and less and eventually stabilizes at a price (4.97 on average across experiments) above the competitive price.

[23] Of course, by playing a price of 4.90, the AM may 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).

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 with and without adverse selection ($18,000 = 2 \times 1,000 \times 9$ experiments overall). For each value of σ, we also plot the average quoted spread $\overline{QS}$ and the quoted spread in the Glosten-Milgrom benchmark (dashed lines in Figure 4).

Consider the adverse selection case first. For all values of σ, the average quoted spread in this case is largely above the Glosten-Milgrom benchmark. 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 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 either decreases with σ (adverse selection case) or is nil for all values of σ (no adverse selection case).[24]

Last, we observe that AMs' quoted spread is larger in the adverse selection case than in the no adverse selection case, other things equal. This means that AMs learn to adjust their prices to adverse selection.

In Panel B of Figure 4, we report the average realized spreads $\overline{RS}$ for different values of σ. We observe that AMs' average realized spread is positive and increases with σ. Thus, AMs become less competitive when the dispersion of clients' liquidity shocks becomes larger. This finding is again at odds with the Glosten-Milgrom benchmark, even after accounting for price discreteness. It suggests that it is more difficult for AMs to learn to undercut when the dispersion of clients' liquidity shocks gets larger. Furthermore, Panel B of Figure 4 shows that AMs' average realized spreads are *smaller* with adverse selection than without, all else equal. Thus, even though AMs stop undercutting each other at higher prices when there is adverse selection (Panel A), their prices are in fact closer to their costs in this case.

[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

[24] This is the case even after accounting for price discreteness (see the dotted-dashed lines on the figure).

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 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.

We again observe that AMs adjust their prices to adverse selection: For all values of Δ_v, the quoted spread is larger with adverse selection than without. However, Panel B of Figure 5 shows that AMs' average realized spread is always smaller in the adverse selection case across all values of Δ_v. This again suggests that adverse selection helps dealers to learn to undercut prices when they are above costs. Finally, we observe that the AMs' average realized spread increases with the volatility of the asset payoff, whether there is adverse selection or not. This suggests that an increase in the volatility of the asset payoff also hinders AMs' ability to learn to undercut.

[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 earn smaller rents (their average realized spreads are smaller) when there is adverse selection than when there is not.

4. AMs earn larger rents–whether there is adverse selection or not–when the dispersion of clients' liquidity shocks or the volatility of the asset payoff increase.

A natural question is whether these facts are robust to the choice of the hyperparameters (α, β). We show that this is the case in Section OA.3 of the Online Appendix where we run additional simulations with different values for (α, β).[25]

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. Moreover, we are not aware of models with competing risk-neutral dealers predicting a negative relationship between dealers' realized spreads and adverse selection costs, as observed in Panel B of Figures 4 and 5.[26]

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.

Instead, the algorithms fail to learn to compete. Indeed, they "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 failure is more pronounced when there is no adverse selection, the dispersion of clients' liquidity shocks is larger, or the volatility of the asset payoff is larger. The next section provides an explanation for these patterns.

4.2 Interpretation

In Section 4.2.1, we show that a simple heuristic provides a unified explanation for our observations. Namely, a larger volatility of AMs' profits hinders their learning to be competitive. This volatility

[25] We 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 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 or Δ_v increase (sometimes, weakly). Moreover, for each value of σ or Δ_v, AMs' average realized spreads are larger when there is no adverse selection (sometimes, the difference is not statistically significant).

[26] Liu and Wang (2016) examine a model with 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 costs) can lead the monopolist dealer to widen the bid-ask spread. Market power is key for this mechanism.

stems from the economic environment and is reinforced by strategic uncertainty (Section 4.2.2). Finally, we discuss how the choice of the experimentation rate affects this logic (Section 4.2.3).

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 long run greedy price in the experiments). At this price, the true expected profit for the AMs is 0.3. If one AM (henceforth AM_1) undercuts by one tick (a price of 4.9), its true expected profit is $0.59 > 0.3$. However, to learn that undercutting is profitable, AM_1 must try to do so. When it does, it does not observe that undercutting yields an expected profit of 0.59. Rather, as feedback, AM_1 receives the profit realized by undercutting.

In our example, AM_1's realized profit if it undercuts is (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 $\tilde{v} = v_L = 0$. These events happen with probabilities 0.703, 0.214, and 0.081, respectively. Thus, the variance of AM_1's realized profit is 1.78, which is large compared to the expected profit of 0.59. Moreover, if instead AM_1 matches AM_2's price, the variance of its realized profit is 0.45, with an expected profit of 0.3.[27] Intuitively, given the large dispersion in profits around the expected value, learning that quoting 4.9 yields a higher average profit than quoting 5 ("learning to undercut") requires experimenting with undercutting over a large number of episodes.[28]

This learning problem repeats itself at each price. If AM_2 uses an initial price of a_2, AM_1 will "explore" and try many different prices. Prices above a_2 give a payoff of zero with certainty and are therefore eliminated relatively quickly. In contrast, as just explained, prices below a_2 are not guaranteed to give a larger profit than $a_1 = a_2$, even if they do on average. However, with sufficiently many trials, AM_1 will eventually learn to play a price $a_1 < a_2$. When this happens, AM_2 makes zero profit and eventually uses other prices. Again those above a_1 are gradually eliminated, and AM_2 eventually learns to undercut AM_1, etc.

This process resembles the familiar process of elimination of dominated strategies. However,

[27] In this case, AM_1's realized profit is (i) 0 if the client decides not to trade, (ii) 0.5 if the client buys and $\tilde{v} = v_H = 4$, and (iii) 2.5 if the client buys and $\tilde{v} = v_L = 0$. These events occur with probabilities 0.71, 0.214, and 0.07, respectively.

[28] Intuitively, AM_1's learning problem is similar to that of a statistician who must determine whether the difference between the unobserved expected values of two variables is positive. This is more difficult (requires more data) when the difference in expected values of the variables is small relative to the sum of the variances of the variables.

it is gradual because AMs get noisy feedback about their average performance at a given price. Moreover, it becomes less and less effective because the rate at which AMs experiment decays exponentially with time.

This reasoning suggests the following heuristic to predict when AMs' prices will remain high relative to AMs' costs. With a slight abuse of notation, in the case $N = 2$ denote $\Pi(a_1, a_2)$ the (random) profit of AM_1 if AM_1 quotes a_1 and AM_2 quotes a_2. The expected gain for AM_1 from undercutting AM_2 is $\mathbb{E}[\Pi(a_2 - tick, a_2) - \Pi(a_2, a_2)]$, and the variance of this gain is $\mathbb{V}[\Pi(a_2 - tick, a_2)] + \mathbb{V}[\Pi(a_2, a_2)]$. Intuitively, holding a_2 constant, undercutting a_2 provides a signal on the expected gain from undercutting and the ratio of the expected gain from undercutting to the variance of this gain is akin to the signal-to-noise ratio. Thus, we posit that, for a given expected gain from undercutting, a change in the parametrization of the market making game that increases the variance of this gain impairs AMs' ability to learn to undercut, and eventually leads to less competitive outcomes (i.e., larger average realized spreads). Conversely, for a given variance, a change in the parameters that increases the expected gain from undercutting should lead to more competitive outcomes.[29]

This simple heuristic goes a long way in explaining the effects of a change in the environment (e.g., a change in σ) on the prices eventually chosen by the AMs in our experiments. To show this, 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 $\text{Var}_{n.as}(a_1, a_2)$ when there is no adverse selection and $\text{Var}_{as}(a_1, a_2)$ when there is. We establish the following properties.

First, for given parameters, $\text{Var}_j(a_1, a_2) = 0$ ($j \in \{as, n.as\}$) when $a_1 > a_2$ since the client never trades with AM_1 at such prices. Thus, an AM quickly learns to lower its price when it's outbid. In contrast, learning whether it is more profitable to set $a_1 = a_2$ or $a_1 < a_2$ is difficult. This difficulty depends on the expected gain from undercutting, $\mathbb{E}[\Pi(a_2 - tick, a_2) - \Pi(a_2, a_2)]$, and the variances $\text{Var}_j(a_1, a_2)$ and $\text{Var}_j(a_2, a_2)$, which are both strictly positive. In particular, as prices become closer to the competitive level, the gain from undercutting becomes small and estimating that undercutting is profitable is more difficult. Given limited experimentation, this leads AMs to settle on identical prices above competitive levels, exactly as we observe (see Figure 3).

[29] While this simple heuristic may seem intuitive, note that with rational and risk-neutral players, 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, and the variance would play no role.

Second, we establish that, for $a_1 \leq a_2$, $\text{Var}_{as}(a_1, a_2) < \text{Var}_{n.as}(a_1, a_2)$. Thus, for any parametrization of the market making game, the variance of AM_1's profit when it undercuts or matches AM_2's offer (holding a_2 constant) is smaller in the environment with adverse selection. The reason is that adverse selection increases the likelihood of obtaining profits close to zero for the AMs, as shown in Figure 6. As the variance of AMs' profit is smaller with adverse selection, our heuristic implies that AMs' realized spreads should be smaller on average, which is indeed what we observe experimentally (see Section 4.1).[30]

[INSERT FIGURE 6 ABOUT HERE]

Third, for $a_1 \leq a_2$, $\text{Var}_{n.as}(a_1, a_2)$ and $\text{Var}_{as}(a_1, a_2)$ increase with the variance of clients' liquidity shocks, σ, and the volatility of the asset payoff, Δ_v. Indeed, an increase in σ raises the likelihood of a trade and therefore the dispersion of realized profits while an increase in Δ_v increases the range of possible realized profit for AM_1 when there is a trade.[31] Thus, our heuristic implies that AMs' average realized spreads should be larger when σ or Δ_v are larger, which is consistent with our observations (see Figures 4 and 5). It is worth stressing that the expected gain from undercutting also increases with the variance of clients' liquidity shocks, σ, at any price a_2. The variance effect seems to dominate in our simulations and the long run greedy price increases therefore with σ.

4.2.2 Strategic Uncertainty Hinders Competition

To present the heuristic explanation for our experimental findings, we have considered the variance of one AM's profit in the market making game, holding the price of the other AM constant. However, in the experiments, each AM changes its price randomly over time, due to idiosyncrasies in its trading history and experimentation choices. As a result, holding the parametrization of the market making game constant, the distribution of one AM's profits at a given price is non-stationary.

We refer to this source of variation for AMs' profits at a given price as "strategic uncertainty". Intuitively, strategic uncertainty makes learning the expected gain from undercutting even more difficult because it is an additional source of noise, above and beyond that generated by the volatility

[30] The 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.

[31] The effect of Δ_v on the variance of AMs' profits is weaker when there is adverse selection because an increase in the volatility of the asset payoff also raises the cost of adverse selection, which, as explained before, tends to reduce the dispersion of profits.

of the asset and the client's valuation. It therefore slows down dealers' ability to learn the average profit they can obtain at a given price. Hence, our heuristic implies that if we turn off this source of noise then the AMs' long-run price should be more competitive.

To test this prediction, we run experiments in which the price set by AM_2 is constant and equal to 5.0 in every period, the level of the average greedy price after T episodes in our baseline experiments (see Figure 3). As predicted, we observe (see Figure 7) that AM_1 learns to undercut more quickly. Indeed, it takes "only" 46,369 episodes for the average greedy price for AM_1 over $K = 1,000$ experiments to reach 4.9 and, after $T = 1,000,000$ episodes, the modal greedy price for AM_1 is 4.9. Thus, strategic uncertainty also slows down learning for the AMs.

[INSERT FIGURE 7 ABOUT HERE]

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 profit volatility. To overcome these obstacles, algorithms could experiment more intensively and for a longer time. 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.

In Section 6.1, we endogenize the choice of hyperparameters by the dealers. In particular, we show that they are likely to "stick" to parameters with relatively low experimentation, because more experimentation eventually leads the other algorithm to undercut for longer, reducing profits. Thus, allowing dealers to choose the hyperparameters of their algorithms does not prevent them from failing to be competitive.

5 Implications

In this section, we derive additional testable implications. We focus on patterns that are specific to AMs, in the sense that they are not predicted by the Glosten-Milgrom benchmark. Our goal is to further identify the "signature" that reinforcement learning algorithms should leave in the data. One could test these implications in equity markets where algorithmic market makers dominate liquidity supply (see Brogaard *et al.* (2014), Baron *et al.* (2019) or Aquilina *et al.* (2021)). Also,

the market making game considered in our experiments resembles electronic Request for Quotes (RFQs) systems used in bond markets (see Hendershott and Madhavan (2015)). Interestingly, some firms (e.g., Overbond) now provide AI tools to automate dealers' bids in these systems.[32] Thus, data from RFQs could be another way to test the predictions developed in this section.

5.1 Entry

Brogaard and Garriott (2019) empirically find that average bid-ask spreads gradually decline with the entry of new high-frequency market makers. This pattern cannot be explained by the standard model of price competition considered in Section 2.1. However, it is consistent with our interpretation of AMs' behavior. Indeed, as the number of AMs increases, the expected gains from undercutting become larger–ceteris paribus–because AMs' profits, when tied at the same price, are divided among more dealers. Additionally, the variance of these profits also decreases for the same reason. Consequently, our heuristic (see Section 4.2) implies that an increase in the number of AMs accelerates their learning to undercut and, therefore, leads to more competitive outcomes, all else being equal.

To test this conjecture, we run experiments varying the number of competing AMs from 2 to 10 and report the results in Figure 8. As predicted, we find that AMs' average quoted and realized spreads decline gradually with the number of AMs.

[INSERT FIGURE 8 ABOUT HERE]

In real markets, this pattern could also arise due to inventory costs, because a larger number of dealers allows for more risk sharing. Hence, testing our mechanism with field data will require controlling for inventory costs, e.g., the level of dealers' inventories. This level plays no role in our analysis because AMs are not penalized for risk-taking.

5.2 Tick Size

Our heuristic for AMs' behavior also implies that reducing the tick size should have an ambiguous effect on the competitiveness of their quotes.[33] On the one hand, a smaller tick size increases the

[32] See "*Overbond AI To Automate Up To Half of RFQs*," MarketsMedia, February 23, 2021.

[33] Cartea *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.

average gain from undercutting, since an undercutter can double its market share (when there are two competitors) with a smaller price improvement. This is the standard intuition for why reducing the tick size tends to result in more competitive outcomes under price competition.

On the other hand, a smaller tick size also increases the number of possible quoting options available to each AM. As a result, holding the AMs' experimentation rate constant, each individual price is explored less frequently. As explained in Section 4.2.3, this feature reduces the effectiveness of their learning to undercut. Therefore, the net effect of a tick size reduction on AMs' realized spreads can be positive, contrary to the standard intuition.

To test this conjecture, we run 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 9 (panels A and B).

[INSERT FIGURE 9 ABOUT HERE]

As predicted, 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. In particular, as the tick size declines from 1 to 0.05, AMs' average realized spreads increase while they decrease (by a small amount) when the tick size is reduced further. In contrast, a reduction in the tick size reduces both the quoted and realized spreads in the Glosten-Milgrom benchmartk (dashed lines in Figure 9), as per the standard intuition.

To show that this pattern is due to insufficient experimentation, 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.[34] We report the results of these experiments in Figure 9 (Panels C and D). We observe (Panel C) that the average quoted spread declines when the tick size is reduced, as per

[34] To do so, we rely on the following heuristic. For a given $tick$, the total number of entries in each AM's Q-matrix is $(14/tick) - 1$. 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 $\frac{e^{-\beta}}{1-e^{-\beta}} \times \frac{tick}{14-tick}$. When we vary the tick size, we vary β so that this quantity remains equal to 89.92, its value in the baseline case ($tick = 0.1$ and $\beta = 8.10^{-5}$).

the standard intuition. This is also the case for the realized spread (Panel D).

The choice of tick size is an important policy issue in securities markets. For example, the SEC's decision to reduce the tick size for US equity markets–from one penny to half a penny–sparked a heated debate between policymakers and industry participants.[35] Our findings suggest that accounting for AMs' behavior is essential when predicting the effects of changes in tick size and, more broadly, in market design. These effects may depend, in particular, on the extent to which the algorithms used by market participants are reparametrized following such changes, as illustrated in Figure 9.

5.3 Asymmetric Reactions to Symmetric Shocks

By design, reinforcement learning algorithms are more likely to select actions associated with higher perceived profits–that is, the greedy action is more likely to be chosen than any other. This characteristic creates an asymmetry in AMs' responses to profit-reducing and profit-enhancing shocks following changes in the environment after their training phase (e.g., an increase in adverse selection costs). Specifically, profit-enhancing shocks reinforce AMs' incentive to continue posting the price they learned during training, whereas profit-reducing shocks may eventually prompt them to switch to a different price, as these shocks gradually reduce their estimated profits.

To illustrate this point, we consider a shock to the AMs' adverse selection costs after their training phase. 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 AMs' adverse selection cost by changing Δ_v to a different value Δ'_v. 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 relative to the baseline) ; $\Delta'_v = 1$ (negative adverse selection shock). In each case we conduct $K = 1,000$ experiments.

[INSERT FIGURE 10 ABOUT HERE]

Panel A of Figure 10 shows the dynamics of the quoted spread over the 2 million episodes for the positive adverse selection (profit-reducing) shock, averaging over the K experiments. Panel

[35] See "SEC cuts tick size for stock market trades to a half penny", *Financial Times*, September 18, 2024.

B zooms in around the shock in $T_1 + 1$, and shows episodes from $T_1 - 1000$ to $T_1 + 10000$. As expected, 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 shock that increases 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 accounts for $(1-\alpha)^{100} = 0.99^{100} \simeq 37\%$ of the new Q-value. In other words, 63% 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 longer profitable and switch to a higher price with a higher Q-value.

To verify that the patterns in Panel A of Figure 10 are truly driven by the shock to Δ_v, Panel C of Figure 10 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 decrease in adverse selection costs. In contrast to what happens when the shock increases adverse selection, we observe no change in quoted spreads following this profit-enhancing shock, as in the placebo case.

Thus, in line with our conjecture, AMs' reaction to symmetric shocks to their cost of adverse selection is asymmetric. Indeed, following a transient increase in 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 decrease in 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.

5.4 Bid-Ask Spread Dynamics

As explained in Section 4.2.3, AMs' failure to learn to be competitive stems from the fact that experimentation is limited. Intuitively, this problem becomes more acute when AMs can condition their actions on more states, for instance due to the arrival of new information. A natural case is when dealers can condition their prices on past trades. Limited experimentation implies that, in this case, prices will become less competitive as time passes.

To check this conjecture, 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.[36]

We then study how AMs set their quotes, as we did in Section 3. The key difference is that each episode now features two clients arriving sequentially with the same common value, $\tilde{v}$, instead of one (the $\tilde{v}$ across episodes are i.i.d, as in Section 3). 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 play an action that depends on the state. 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

[36] Clients cannot choose the timing of their trades. As a result, each client has a weakly dominant strategy: to buy if and only if the price is below her valuation, regardless of her beliefs about how AMs set their prices. If clients were able to choose when to trade, we would require a theory explaining how they form such beliefs in order to determine their submission strategies. This would introduce a layer of complexity well beyond the scope of the paper.

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.[37] 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 in $\mathcal{A}$ 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 for the first and the second clients across $K = 1,000$ experiments in the last episode $T = 10^6$. 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 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).

As in the one client case analyzed in the previous sections, we use the Glosten-Milgrom prices in each period as a benchmark. We derive these prices in the Appendix OA.1 and show that they 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' 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 in the Glosten-Milgrom benchmark, they update their beliefs about v upward after observing a buy from the first client and downward after observing no trade. The second property means that, on average, the ask price in the second period is closer to the unconditional expectation of $\tilde{v}$. In other words, in the Glosten-Milgrom benchmark, the spread charged by the dealer decreases over time on average. The reason is that dealers learn from past transactions and are therefore, in expectation, less exposed to adverse selection as time passes.[38]

Figure 11 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: The distance between $\bar{a}_2^T$ and the first period price is significantly larger than in the Glosten-Milgrom benchmark and increasingly so as the dispersion of clients' liquidity shocks (σ) increases. In contrast, when no

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

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

trade occurs in the first period, AMs underreact: The distance between $\bar{a}_2^{NT}$ and the first period price is significantly smaller (in fact it is almost zero) than in the Glosten-Milgrom benchmark.

These patterns imply that AMs extract even larger rents on average from the second client than the first. As shown by Figure 12, this effect is so strong that the average price for the second client across all states (trade and no trade) is strictly larger than the average price for the first client, in contrast to the Glosten-Milgrom benchmark's prediction ($\mathbb{E}[a_2^*] < a_1^*$; see dashed lines in Figure 12).

[INSERT FIGURES 11 and 12 ABOUT HERE]

Thus, AMs' failure to learn to be competitive is even stronger for the second client than for the first. The reason is that the algorithms have fewer opportunities to learn about the average profits of their actions for the second client than for the first client. Indeed, the AMs learn state by state. They face the "period 1" state in each of the 10^6 episodes. In contrast, they face each of the 4 possible states in "period 2" much less frequently. Moreover, these states are not observed with the same frequency. Indeed, for 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 3 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 (lack of experimentation) leads to high prices.

More generally, these experimental results imply that quoted spreads and realized spreads should tend to be large after histories that are more rarely observed.

6 Robustness

AMs' long-run prices do not form a Nash equilibrium, meaning dealers are "leaving money on the table" and fail to learn to compete. As in any such situation, this raises the question of whether this possibility is robust to competition from more sophisticated agents. We consider two possibilities. First, in Section 6.1 we let dealers choose the hyper-parameters of their algorithms, so that they can, for instance, choose a high experimentation rate to eventually outbid their competitor. Second,

in Section 6.2, we study how an AM fares against a highly sophisticated agent with full knowledge of the game and the AM's behavior. The question is whether such an agent would drive out of the market AMs if they compete inefficiently.

6.1 Choosing Algorithms' Hyperparameters

In this section, we endogenize the choice of AMs' hyper-parameters (α, β) by the dealers using these algorithms. Doing so is not straightforward and there is no standard approach to this question in the economic literature yet (more on this at the end of this section). Indeed, dealers cannot compute ex-ante the expected total profit obtained with a given AM since they are supposed to have no knowledge of the environment. If they could, they would not use AMs in the first place.

To address this issue, we use the following approach. In a first step, we assume that both dealers initially use (α_m, β_m) for $K = 1,000$ experiments and record their average total profit per period and the standard deviation of this profit, across the K experiments. Then, we assume that for K' experiments, dealer 1 deviates to another parametrization, (α', β') (dealer 2 does not deviate) and records his average total profit per period and the standard deviation of this profit.[39] Finally, to decide whether the deviation has been profitable, dealer 1 conducts a statistical test under the null hypothesis that the parametrization (α', β') yields the same expected profit as the parametrization (α_m, β_m). If there is no profitable deviation in this statistical sense, we say that the parametrization (α_m, β_m) is stable. We provide a detailed explanation of this process in Online Appendix OA.5.

Practically, we allow dealer 1 to select $\alpha \in \{\alpha_l, \alpha_m, \alpha_h\}$ and $\beta \in \{\beta_l, \beta_m, \beta_h\}$, with $\alpha_l = 0.001, \alpha_m = 0.01, \alpha_h = 0.1$; $\beta_l = 5.10^{-6}, \beta_m = 8.10^{-5}, \beta_h = 3.2.10^{-4}$. Thus, dealer 1 has 8 possible deviations. Moreover, we consider 10 values in $K^{'}$ ranging from 100 to 1,000. As shown in Online Appendix OA.5, only (α_m, β_h) could seem a profitable deviation to dealer 1, and only if he experiments this deviation relatively few times ($K' \leq 500$). With more experimentation, one cannot reject the hypothesis that such a deviation is in fact not profitable (p-value of 0.29 for $K' = 1000$). In any case, deviating by choosing a higher experimentation rate (lower β) than dealer 2 is clearly not profitable.

We then repeat this exercise by considering the 81 possible configurations for the pairs (α, β) chosen by both dealers. That is, for each possible pair, we ask whether one dealer can find it

[39] Considering unilateral deviations by one AM (here AM_1) is sufficient since both AMs face the same environment.

profitable to deviate to another pair in the sense defined above (in this case, we only consider $K^{'} = 1,000$ for brevity). We find that only 5 configurations of pairs (α, β) for each dealer are stable at the 0.25 confidence level. This means that, for these 5 pairs, there exists no alternative parametrization for which the dealer can reject the hypothesis that the current and alternative parametrizations are equally profitable, at the 25% confidence level. For all other pairs, the null hypothesis can be rejected at the 0.1% level. The pair (α_m, β_m) is among the 5 stable pairs. The "most" stable configuration seems to be that in which both AMs choose $\alpha = \alpha_m$ and $\beta = \beta_l$: if both AMs start with a low experimentation rate ($\beta = \beta_l$), none has an incentive to deviate.

In sum, we do not find support for the idea that competition on the choice of AMs' hyperparameters should fix AMs' failure to learn to be competitive. Intuitively, while a higher α and a lower β can boost short-term profit by undercutting opponents, the long-term effect is negative. Indeed, once AM_1 undercuts AM_2, AM_2 makes no profit and eventually adjusts, forcing AM_1 to share demand at lower prices. This long-run effect outweighs short-term gains, which makes the average profit per period from choosing a high experimentation rate significantly smaller or not significantly larger than those obtained with a high experimentation rate.

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. As noted previously, the literature on this question is still very scarce. To the best of our knowledge, two approaches have been proposed so far. The first is to assume that the agents choose α and β by using a second-layer of Q-learning (Dou *et al.*, 2023). This approach is natural. However, a limitation is that the second layer also needs hyperparameters, so that the same problem repeats itself at a higher level.

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.* (2024). 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. 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.

6.2 Competition from an Omniscient Player

The fact that, in any given interaction, each AM does not play the one-shot Nash best response to the quote posted by the other AM might suggest that the algorithm would not survive when facing a rational Bayesian agent who does so. While an in-depth exploration of this question is beyond the scope of this paper, in the Online Appendix OA.6, we show that a rational forward-looking dealer ("an omniscient player") has no incentive to consistently undercut an algorithmic competitor. The omniscient player gains more by encouraging a collusive play, ensuring positive aggregate profit while capturing most of it.

7 Conclusion

We have conducted experiments in which Q-learning algorithms act as market makers in a setting similar to Glosten and Milgrom (1985). We find that, despite their simplicity and the challenges posed by an environment with adverse selection, our algorithmic market makers (AMs) exhibit realistic behavior: their quoted spreads reflect adverse selection costs, and they adjust their quotes in response to observed order flow. However, AMs do not learn to undercut each other down to competitive price levels. This failure arises from two factors: the noisy feedback they receive about the average profitability of their actions and the fact that, by design, their experimentation is limited.

These features help to explain and predict the behavior of algorithmic market makers. For example, they imply that AMs earn higher rents when their profits are more volatile (such as when clients' liquidity shocks are more dispersed or when there is no adverse selection), which is consistent with our experimental findings. Furthermore, they suggest that a finer price grid (i.e., a smaller tick size) can lead AMs to post less competitive prices; that AMs are likely to set wider spreads in rarely observed states; and that their bid-ask spreads should respond asymmetrically to symmetric shocks in their profits. We also confirm these implications experimentally.

Future research could test our predictions using field data. It could also extend our analysis in several directions. First, in our experiments, only dealers use Q-learning algorithms. In other papers (Dou *et al.* (2023) and Yang (2025)), only informed investors use Q-learning algorithms. One could consider the case in which both dealers and clients use such algorithms, e.g., with clients using Q-learning algorithms to choose the timing of their submission strategy, or whether to post limit or market orders. Second, in our experiments, quotes are posted only by algorithms, whereas in actual security markets algorithms still coexist with humans. An interesting avenue for future research is therefore to run experiments in which prices are set both by reinforcement learning algorithms and humans.[40]

[40] Werner (2024) considers experiments with humans and pricing algorithms in product markets, assuming that pricing algorithms interact with humans after being trained in markets without humans. He finds that the presence of humans does not necessarily make outcomes more competitive.

References

Abada, I., Lambin, X. and Tchakarov, N. (2024). Collusion by mistake: Does algorithmic sophistication drive supra-competitive profits? *European Journal of Operational Research*, **318** (3), 927–953. 7, 34

Aquilina, M., Budish, E. and O'Neill, P. (2021). Quantifying the High-Frequency Trading Arms Race. *The Quarterly Journal of Economics*, **137** (1), 493–564. 25

Asker, J., Fershtman, C. and Pakes, A. (2024). The impact of artificial intelligence design on pricing. *Journal of Economics & Management Strategy*, **33** (2), 276–304. 7, 17

Baldauf, M. and Mollner, J. (2020). High-frequency trading and market performance. *The Journal of Finance*, **75** (3), 1495–1526. 7

Banchio, M. and Skrzypacz, A. (2022). Artificial intelligence and auction design. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC '22, New York, NY, USA: Association for Computing Machinery, pp. 30–31. 8

Baron, M., Brogaard, J., Hagstromer, B. and Kirilenko, A. (2019). Risk and return in high-frequency trading. *Journal of Financial and Quantitative Analysis*, **54** (3), 993–1024. 25

Biais, B., Foucault, T. and Moinas, S. (2015). Equilibrium fast trading. *Journal of Financial Economics*, **116** (2), 292–313. 7

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. 26

—, Hendershott, T. and Riordan, R. (2014). High-Frequency Trading and Price Discovery. *The Review of Financial Studies*, **27** (8), 2267–2306. 1, 25

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. 7

Calvano, E., Calzolari, G., Denicolo, V. and Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, **110** (10). 6, 7, 21

Cartea, A., Chang, P., Mroczka, 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. 6, 8, 26

Chaboud, A. P., Dao, A., Vega, C. and Zikes, F. (2025). What Makes HFTs Tick? Tick Size Changes and Information Advantage in a Market with Fast and Slow Traders. *Management Science*, **71** (1), 553–583. 1

Compte, O. (2023). *Q-based Equilibria*. Working paper. 34

Cont, R. and Xiong, W. (2024). Dynamics of market making algorithms in dealer markets: Learning and tacit collusion. *Mathematical Finance*, **34** (2), 467–521. 6

Dou, W., Goldstein, I. and Ji, Y. (2023). *AI-Powered Trading, Algorithmic Collusion, and Price Efficiency*. Working paper. 7, 8, 21, 34, 36

Easley, D. and Kiefer, N. M. (1988). Controlling a stochastic process with unknown parameters. *Econometrica*, **56** (5), 1045–1064. 8

— and Rustichini, A. (1999). Choice without beliefs. *Econometrica*, **67** (5), 1157–1184. 13

Foucault, T., Kozhan, R. and Tham, W. W. (2016). Toxic arbitrage. *The Review of Financial Studies*, **30** (4), 1053–1094. 7

Fudenberg, D. and Levine, D. (1998). *The Theory of Learning in Games*. Cambridge (Mass.): MIT Press. 8

— and Levine, D. K. (1993). Self-confirming equilibrium. *Econometrica*, **61** (3), 523–545. 8

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, **41** (2), 148–177. 8

Glosten, L. and Putnins, T. (2020). *Welfare Costs of Informed Trade*. Working paper. 31

Glosten, L. R. 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, 7, 11, 35

Goldstein, I., Spatt, C. S. and Ye, M. (2021). Big Data in Finance. *The Review of Financial Studies*, **34** (7), 3213–3225. 1, 7

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. 6

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

Hendershott, T. and Madhavan, A. (2015). Click or call? auction versus search in the over-the-counter market. *The Journal of Finance*, **70** (1), 419–447. 26

IMF (2024). *Global financial stability report: Steadying the course: Uncertainty, artificial intelligence, and financial stability*. Report. 1

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. 16

Klein, T. (2021). Autonomous algorithmic collusion: Q-learning under sequential pricing. *The RAND Journal of Economics*, **52** (3), 538–558. 6

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, **53** (6), 1315–1335. 7

Liu, H. and Wang, Y. (2016). Market making with asymmetric information and inventory risk. *Journal of Economic Theory*, **163**, 73–109. 21

Menkveld, A. and Zoican, M. (2017). Need for speed? exchange latency and liquidity. *Review of Financial Studies*, **30** (4), 1188–1228. 7

OECD (2017). Algorithms and collusion: Competition policy in the digital age. pp. 1–72. 6

O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics*, **116** (2), 257–270. 7

Pouget, S. (2007). Adaptive traders and the design of financial markets. *The Journal of Finance*, **62** (6), 2835–2863. 8

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. 31

Sutton, R. and Barto, A. (2018). *Reinforcement Learning: An Introduction*. Cambridge (Mass.): MIT Press. 13, 15

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

Watkins, C. and Dayan, P. (1992). Q-learning. *Machine Learning*, **8**, 279–292. 16

Werner, T. (2024). *Algorithmic and Human Collusion*. Working paper. 36

Wilk, E. (2022). *Pricing Under Pressure: The Effect of Signal Corruption on the Gameplay of Pricing Algorithms*. Working paper, California Institute of Technology. 6

Yang, H. (2025). *AI Coordination and Self-Fulfilling Financial Crises*. Working paper. 36

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 $\sigma = 5$. This figure displays the distributions of clients' valuations for $v = v_H$ (red) and $v = v_L$ (blue). The former distribution is shifted to the right relative to the latter, and the two distributions partially overlap. If dealers' best offer is $a^{min} = 2.5$, the likelihood that the client buys the asset is given by (i) the area (in blue) to the right of 2.5 and under the blue curve when $v = v_L = 0$, and (ii) the area (in blue and red) to the right of 2.5 and under the red curve when $v = v_H = 4$.

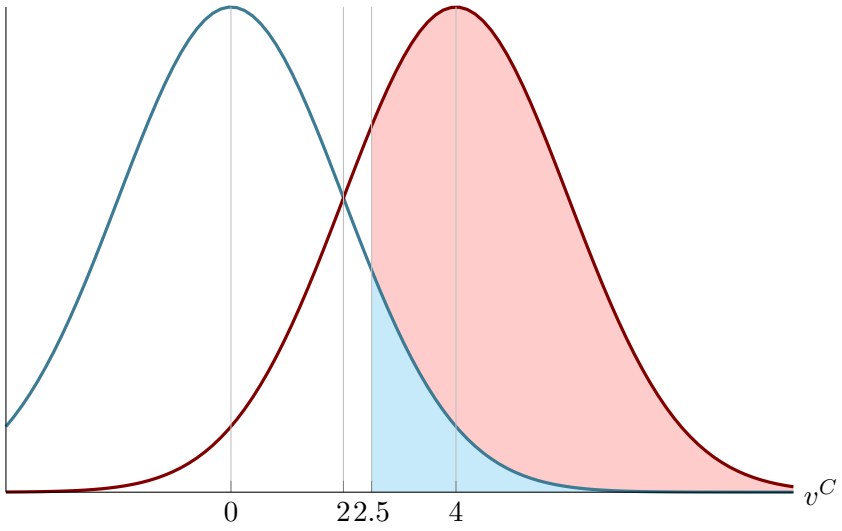


Figure 2: Glosten-Milgrom Benchmark. Parameters: $\mathbb{E}_\mu(v) = 2$, $\mu = \frac{1}{2}$, $\Delta_v = 4$, $\sigma = 5$. The figure shows the equilibrium quoted spread in the Glosten-Milgrom benchmark. The left panel shows that the quoted spread is a decreasing function of the variance of clients' liquidity shocks, σ. The right panel shows that the quoted spread is an increasing function of the volatility of the asset payoff, Δ_v.

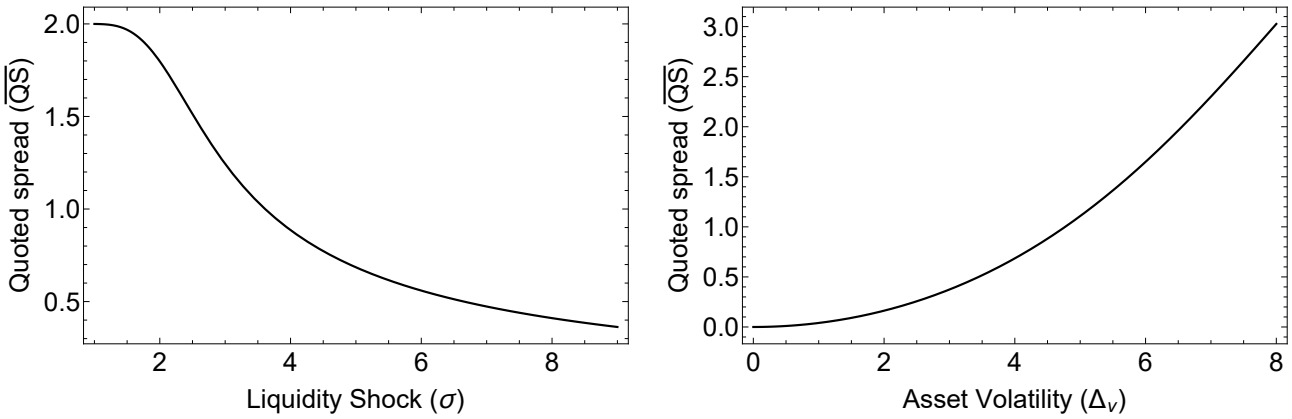
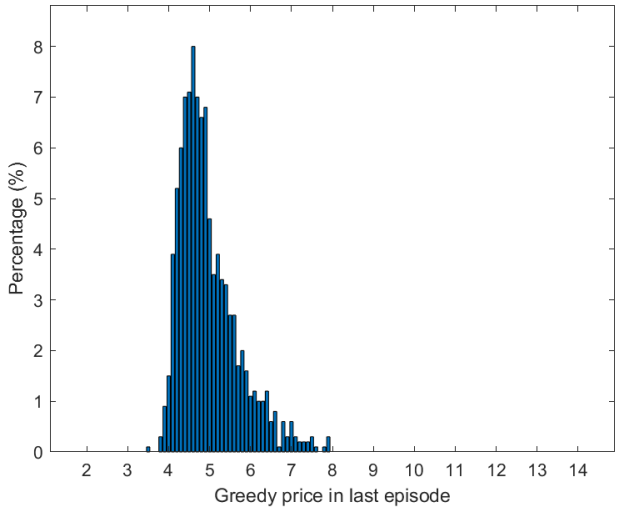


Figure 3: Greedy price of AM_1 in the adverse selection case. Parameters: $\sigma = 5$, $\Delta_v = 4$, $N = 2$, $\mu = \frac{1}{2}$, $\mathbb{E}(v) = 2$, $T = 1,000,000$, and $K = 1,000$.

Panel A: 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 $a^*_{1,T} = a$. The mode of the distribution is 4.60, the mean 4.97, and the distribution is positively skewed.



Panel B: Dynamics of the average greedy price of AM_1 for episodes 1 to T.

This graph shows for each episode t between 1 and 1,000,000 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). Greedy prices start from an average of about 8 and converge to around 5 after 200,000 episodes. The standard deviation is about 4 at the start and decreases to around 1 after 200,000 episodes.

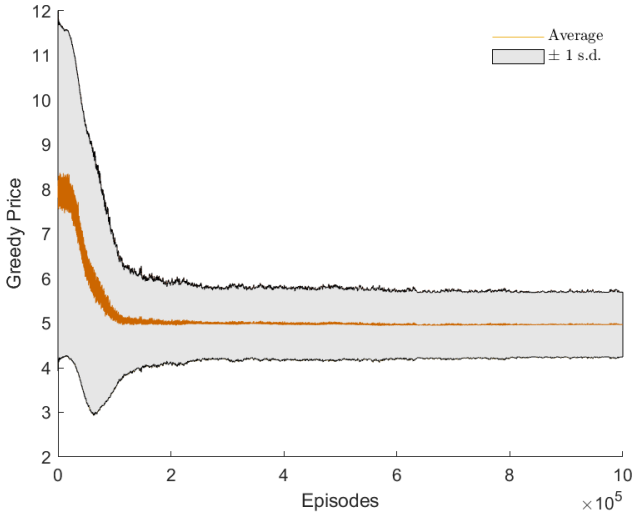
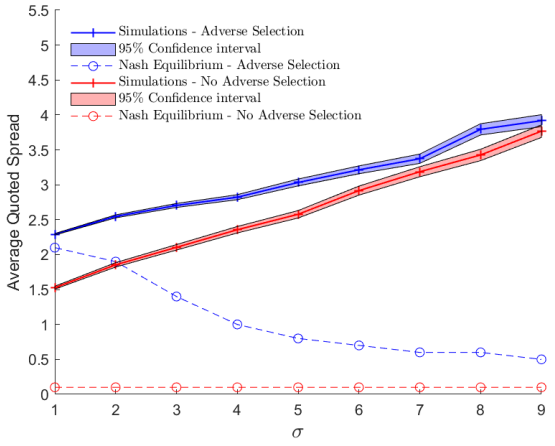


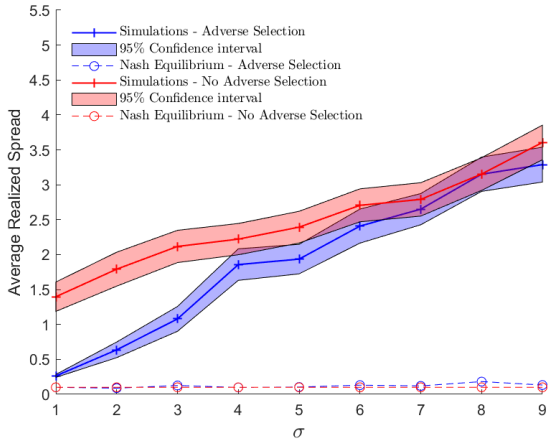
Figure 4: Average Quoted Spread $\overline{QS}$ and Average Realized Spread $\overline{RS}$ in the adverse selection case and the no adverse selection case, for different values of the dispersion of clients' liquidity shocks σ. Other parameters: $\Delta_v = 4$, $N = 2$, $\mu = \frac{1}{2}$, $\mathbb{E}(v) = 2$, $T = 1,000,000$, and $K = 1,000$.

Panel A: 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). Quoted spreads in the Glosten-Milgrom benchmark are nil in the no adverse selection case, positive and decreasing with σ in the adverse selection case. Simulated quoted spreads are strictly above their Glosten-Milgrom values and increase with σ, and are higher in the adverse selection case than in the no adverse selection case.

Panel B: 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). Realized spreads in the Glosten-Milgrom benchmark are nil in the no adverse selection case and in the adverse selection case. Simulated realized spreads are strictly above their Glosten-Milgrom values and increase with σ, and are higher in the no adverse selection case than in the adverse selection case.



Panel A: Average Quoted Spread $\overline{QS}$

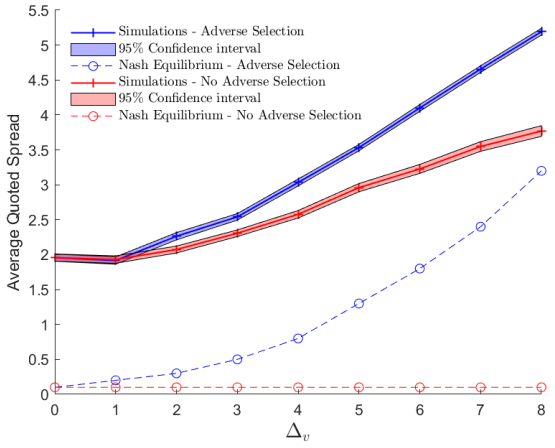


Panel B: Average Realized Spread $\overline{RS}$

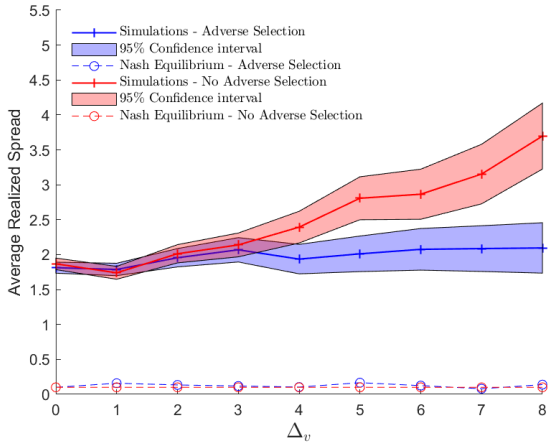
Figure 5: Average Quoted Spread $\overline{QS}$ and Average Realized Spread $\overline{RS}$ in the adverse selection case and the no adverse selection case, for different values of the asset volatility Δ_v. Other parameters: $\sigma = 5$, $N = 2$, $\mu = \frac{1}{2}$, $\mathbb{E}(v) = 2$, $T = 1,000,000$, and $K = 1,000$.

Panel A: 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). Quoted spreads in the Glosten-Milgrom benchmark are nil in the no adverse selection case, positive and increasing with Δ_v in the adverse selection case. Simulated quoted spreads are strictly above their Glosten-Milgrom values and increase with Δ_v, and are higher in the adverse selection case than in the no adverse selection case.

Panel B: 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). Realized spreads in the Glosten-Milgrom benchmark are nil in the no adverse selection case and in the adverse selection case. Simulated realized spreads are strictly above their Glosten-Milgrom values and increase with Δ_v, and are higher in the no adverse selection case than in the adverse selection case.



Panel A: Average Quoted Spread $\overline{QS}$



Panel B: Average Realized Spread $\overline{RS}$

Figure 6: 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 5.0 in the baseline case ($\mu = 0.5$, $v_H = 4$, $v_L = 0$; $\sigma = 5$) in the case without adverse selection (red) and the case with adverse selection (blue). AM_1's realized profit can be 0 (the client does not trade, probability of about 70%), 0.9 (the client buys and the asset payoff is $v_H = 4$, probability of about 20% or 15% with or without adverse selection, respectively) or 4.9 (the client buys and the asset payoff is $v_L = 0$, probability of about 10% or 15% with or without adverse selection, respectively).

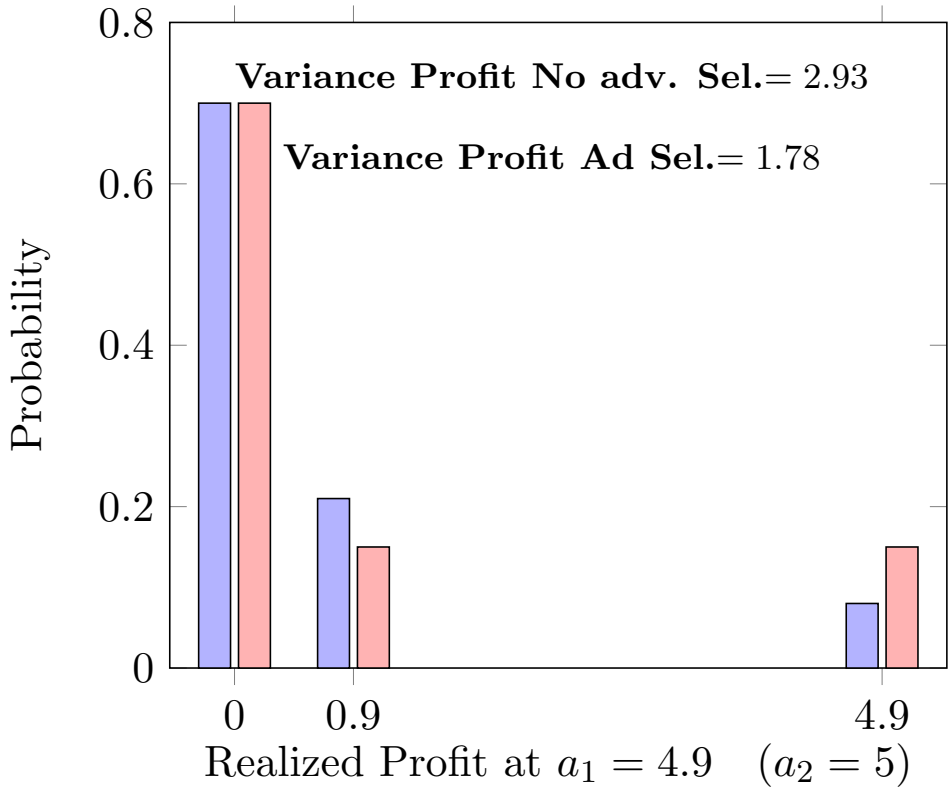
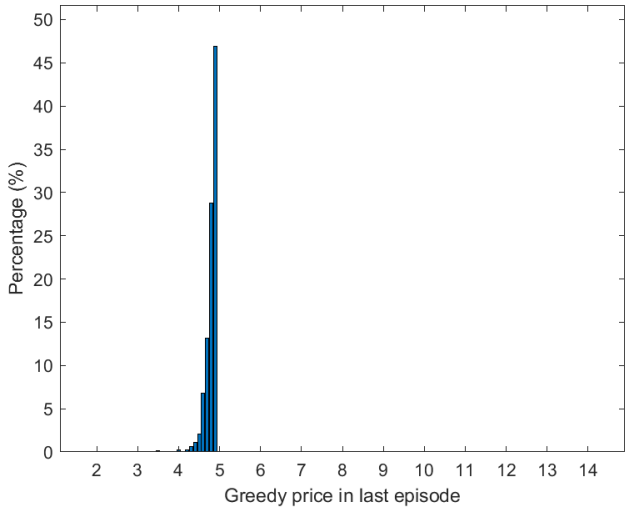


Figure 7: Greedy price of AM_1 when AM_2 plays a constant price in the adverse selection case. Parameters: $\sigma = 5$, $\Delta_v = 4$, $N = 2$, $\mu = \frac{1}{2}$, $\mathbb{E}(v) = 2$, $T = 1,000,000$, and $K = 1,000$. AM_2 plays a constant price of 5.0 in every episode, while AM_1 uses a Q-learning algorithm with $\alpha = 0.01$ and $\beta = 0.00008$.

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 $a^*_{1,T} = a$. All prices are included between 4 and 5, with close to 50% of the realizations at a price of 4.9, the mode of the distribution.



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). Greedy prices start from an average of about 8 and converge to around 5 after 50,000 episodes. The standard deviation is about 4 at the start and converges to less than 0.25 after 600,000 episodes.

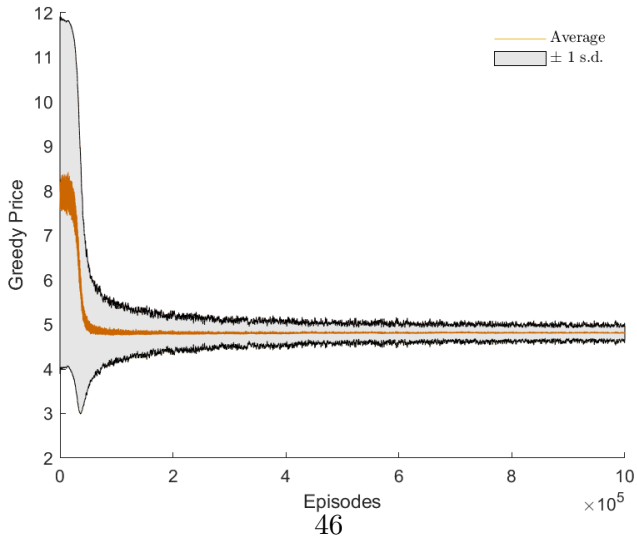
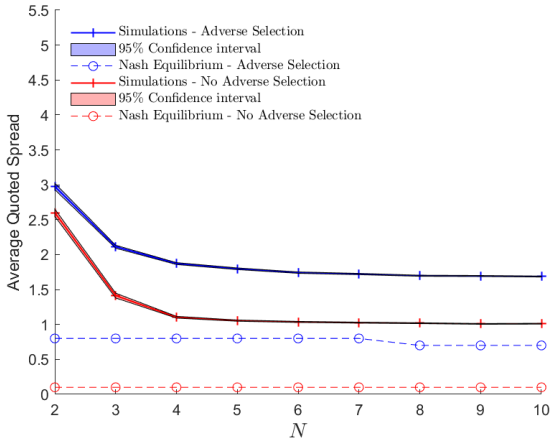


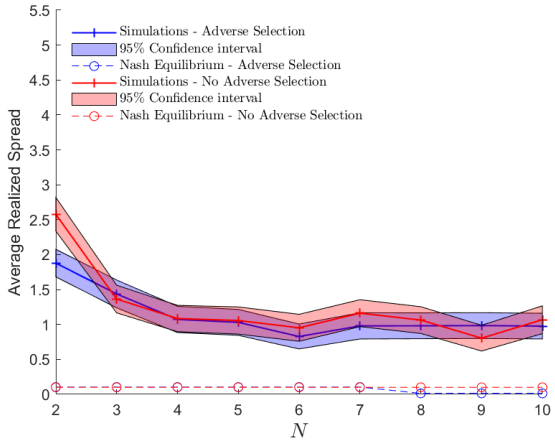
Figure 8: Average Quoted Spread $\overline{QS}$ and Average Realized Spread $\overline{RS}$ in the adverse selection case and the no adverse selection case, for different values of the number N of AMs. Other parameters: $\sigma = 5$, $\Delta_v = 4$, $\mu = \frac{1}{2}$, $\mathbb{E}(v) = 2$, $T = 1,000,000$, and $K = 1,000$.

Panel A: 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). Quoted spreads in the Glosten-Milgrom benchmark are nil in the no adverse selection case, positive and constant in N in the adverse selection case. Simulated quoted spreads are strictly above their Glosten-Milgrom values and decrease with N, and are higher in the adverse selection case than in the no adverse selection case.

Panel B: 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). Realized spreads in the Glosten-Milgrom benchmark are nil in the no adverse selection case and in the adverse selection case. Simulated realized spreads are strictly above their Glosten-Milgrom values and decrease with N. They are higher in the no adverse selection case than in the adverse selection case for $N = 2$, and not statistically different from each other for $N > 2$.

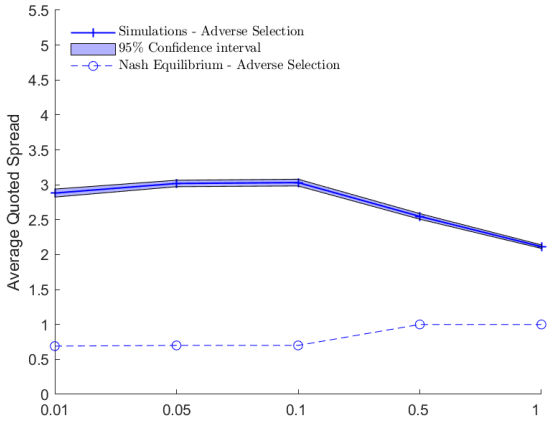


Panel A: Average Quoted Spread $\overline{QS}$

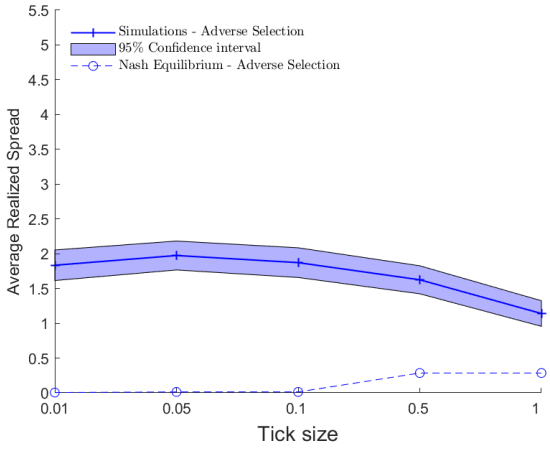


Panel B: Average Realized Spread $\overline{RS}$

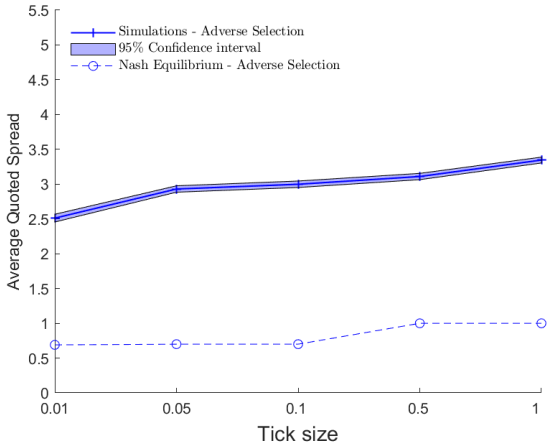
Figure 9: Average Quoted Spread $\overline{QS}$ and Average Realized Spread $\overline{RS}$ in the adverse selection case, for different values of the tick size. Other parameters: $\sigma = 5$, $\Delta_v = 4$, $N = 2$, $\mu = \frac{1}{2}, \mathbb{E}(v) = 2$. We report the average over $K = 1,000$ experiments of the quoted spread (Panel A and Panel C) and the realized spread (Panel B and Panel D) in episode T. 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). The graphs additionally plot the values of the quoted spread and the realized spreads in the Glosten-Milgrom benchmark of Section 2.3 (accounting for price discreteness). Simulated spreads are strictly above their Glosten-Milgrom values. In Panels A and B they decrease with the tick size, whereas in Panels C and D they increase with the adjusted tick size.



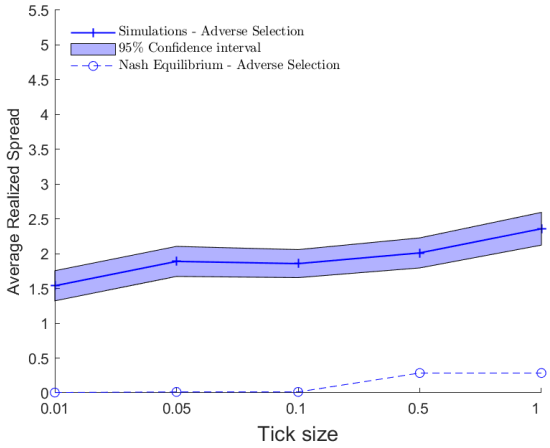
Panel A: Average Quoted Spread $\overline{QS}$



Panel B: Average Realized Spread $\overline{RS}$

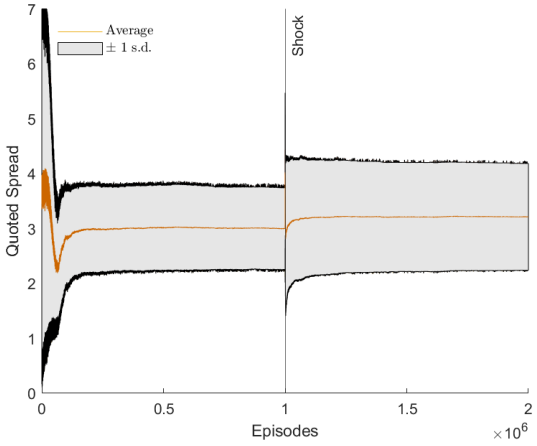


Panel C: Average Quoted Spread $\overline{QS}$, adjusted β

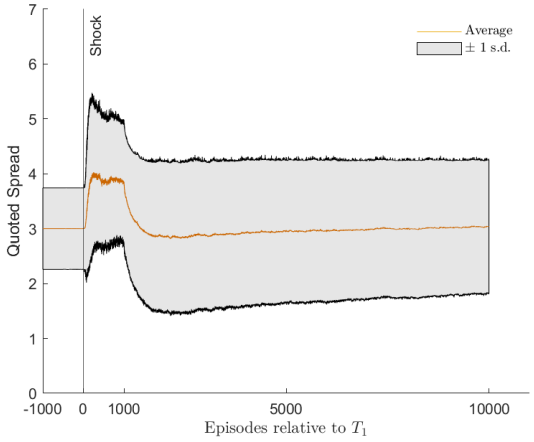


Panel D: Average Realized Spread $\overline{RS}$, adjusted β

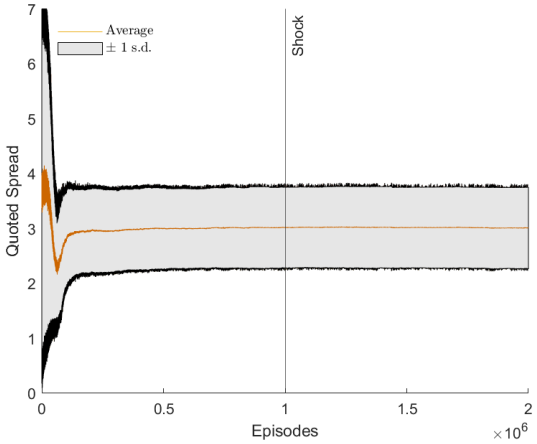
Figure 10: Dynamics of quoted spreads after a shock on adverse selection. Parameters: $\sigma = 5$, $\Delta_v = 4$, $N = 2$, $\mu = \frac{1}{2}, \mathbb{E}(v) = 2$. Each panel shows the dynamics of the quoted spread, averaged over $K = 1,000$ experiments. Between episodes $T_1 + 1$ and $T_1 + 1,000$, with $T_1 = 10^6$, there is a shock to Δ_v, the value of which is changed to Δ_v'. Δ_v' is equal to 7 in Panel A and B (high adverse selection shock), 1 in Panel D (low 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 $T = 2.10^6$. Panels A and B shows that after the shock the quoted spread jumps upwards and then stabilizes at a higher level than before the shock. Panels C and D show that quoted spreads are unaffected by the shock.



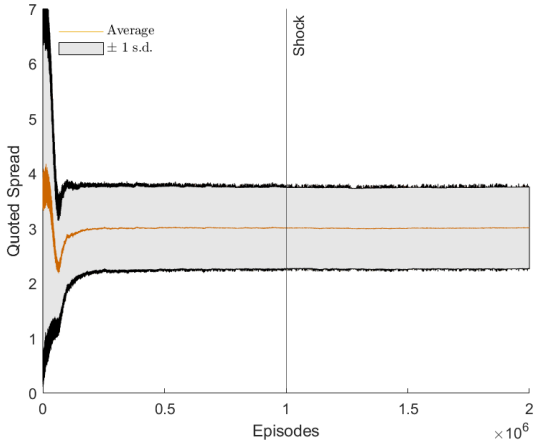
Panel A: $\Delta_v' = 7$



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



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



Panel D: $\Delta_v' = 1$

Figure 11: 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. Other parameters: $\Delta_v = 4$, $N = 2$, $\mu = \frac{1}{2}$, $\mathbb{E}(v) = 2$, $T = 1,000,000$, and $K = 1,000$.

This graph plots the average over 1,000 experiments of the first-period and second-period prices, with 95% confidence intervals. The graph additionally plots the values of these prices in the Glosten-Milgrom benchmark of Section OA.1 (accounting for price discreteness). All prices decrease with σ in the Glosten-Milgrom benchmark. Simulated prices are strictly above their Glosten-Milgrom values and increase with σ. In the Glosten-Milgrom benchmark, for a given σ second-period prices after a trade are higher than first-period prices, which are higher than second-period prices after no trade. In the simulations, second-period prices after a trade are the highest, and the difference between first-period prices and second-period prices after no trade is not significant.

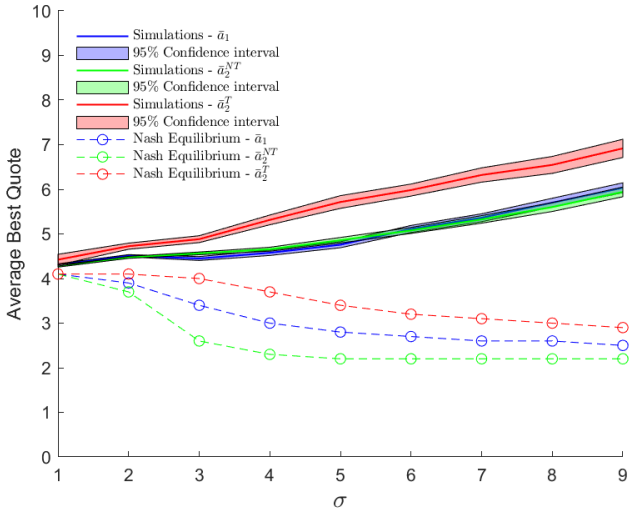
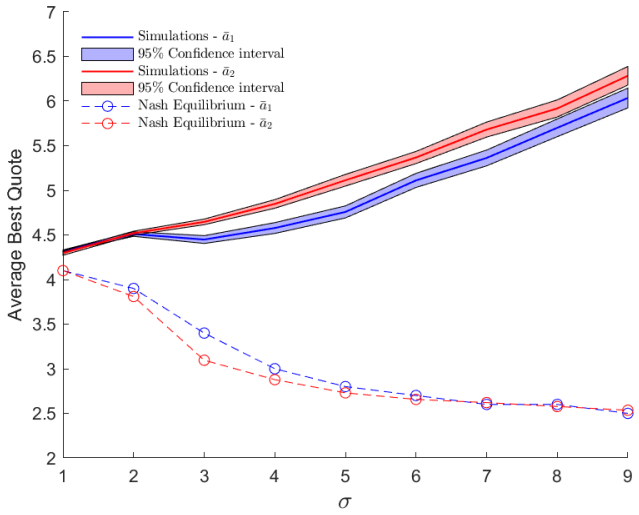


Figure 12: Average first-period price $\bar{a}_1$ and average second-period price $\bar{a}_2$, for different values of the dispersion of clients' liquidity shocks σ. Other parameters: $\Delta_v = 4$, $N = 2$, $\mu = \frac{1}{2}$, $\mathbb{E}(v) = 2$, $T = 1,000,000$, and $K = 1,000$.

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 OA.1 (accounting for price discreteness). Average prices in both periods decrease with σ in the Glosten-Milgrom benchmark, and increase with σ in the simulations. Prices are on average lower in the second period than in the first period in the Glosten-Milgrom benchmark, whereas the opposite obtains in the simulations.



A.2 The Glosten-Milgrom Equilibrium

We only 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:

$$a^* = \mathbb{E}_\mu(\tilde{v} \mid a^* \leq \tilde{v}^C), \tag{A.1}$$

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

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

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 $\Pr_\mu(\tilde{v} = v_H \mid a \leq \tilde{v}^C)$ be the probability that the asset payoff is high ($v = v_H$) conditional on a trade at price a, given dealers' beliefs (μ) about the payoff of the asset. We have

$$\mathbb{E}_\mu(\tilde{v} \mid a \leq \tilde{v}^C) = \Pr_\mu(\tilde{v} = v_H \mid a \leq \tilde{v}^C)v_H + (1 - \Pr_\mu(\tilde{v} = v_H \mid a \leq \tilde{v}^C))v_L. \tag{A.3}$$

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

$$\mathbb{E}_\mu(\tilde{v} \mid a \leq \tilde{v}^C) - \mathbb{E}_\mu(\tilde{v}) = [\Pr_\mu(\tilde{v} = v_H \mid a \leq \tilde{v}^C) - \mu](v_H - v_L). \tag{A.4}$$

It follows that:

$$F(a; \sigma, \Delta_v) = a - \mathbb{E}_\mu(\tilde{v}) - (\Pr_\mu(\tilde{v} = v_H \mid a \leq \tilde{v}^C) - \mu)(v_H - v_L), \tag{A.5}$$

Standard calculations yield:

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

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

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

$$\frac{\partial F(a^*;\sigma,\Delta_v)}{\partial a} \mid_{a=a^*} > 0. \tag{A.7}$$

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

The Glosten-Milgrom price decreases with σ. Using the implicit function theorem, we deduce from (A.2) that

$$\frac{\partial a^*}{\partial \sigma} = -\frac{\frac{\partial F}{\partial a} \mid_{a=a^*}}{\frac{\partial F}{\partial \sigma} \mid_{a=a^*}}. \tag{A.8}$$

From (A.7), we know that $\frac{\partial F}{\partial a} \mid_{a=a^*} > 0$. Thus, $\frac{\partial a^*}{\partial \sigma} < 0$ if and only if $\frac{\partial F}{\partial \sigma} \mid_{a=a^*} > 0$.

To show that this is the case, observe, using (A.5), that $\frac{\partial F}{\partial \sigma} \mid_{a=a^*} > 0$ iff $\Pr_\mu(\tilde{v} = v_H \mid \tilde{v}^C > a^*)$ decreases with σ. Using (A.6):

$$\frac{\partial \Pr_\mu(\tilde{v} = v_H \mid a^* \leq \tilde{v}^C)}{\partial \sigma} = \mu(1-\mu)\left[\frac{D(a^*, v_L)\frac{\partial D(a^*, v_H)}{\partial \sigma} - D(a^*, v_H)\frac{\partial D(a^*, v_L)}{\partial \sigma}}{(\mu D(a, v_H) + (1-\mu)D(a, v_L))^2}\right]. \tag{A.9}$$

Now remember that $D(a^*, v) = 1 - G(a^* - v)$ where $G(.)$ 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, $\frac{\partial \Pr_\mu(\tilde{v}=v_H \mid a^* \leq \tilde{v}^C)}{\partial \sigma} < 0$ since $a^* \in (v_L, v_H)$ and therefore a^* decreases with σ.

The Glosten-Milgrom price increases with Δ_v. We can proceed in the same way to analyze the effect of Δ_v on a^*. The same reasoning as before shows that a^* increases with Δ_v if and only if $\Pr_\mu(\tilde{v} = v_H \mid a^* \leq \tilde{v}^C)$ increases with Δ_v. After some algebra, one obtains that $\frac{\partial \Pr_\mu(\tilde{v}=v_H \mid a^* \leq \tilde{v}^C)}{\partial \Delta_v}$ 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 $G(.)$ 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 $\frac{\partial \Pr_\mu(\tilde{v}=v_H \mid \tilde{v}^C > a^*)}{\partial \Delta_v} > 0$. Hence, a^* increases with Δ_v.

A.3 The Variance of AMs' Profits

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

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 $a_1 < a_2$ 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 that $a_1 < a_2$). By definition (index $n.as$ refers to "no adverse selection"):

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

That is:

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

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

$$\text{Var}_{n.as}(\Pi(a_1, a_2, \tilde{v}^C, \tilde{v})) = p(a_1)(1 - p(a_1))(a_1 - \mathbb{E}_{\frac{1}{2}}(v))^2 + p(a_1)\frac{\Delta_v^2}{4}, \tag{A.12}$$

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 Figure 6, 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 by $\bar{\Pi}_{as}$, AM_1's expected profit with adverse selection ('as')'when $a_1 < a_2$ (this is given by (5) in the text with $Z = 1$). By definition:

$$\text{Var}_{as}(\Pi(a_1, a_2, \tilde{v}^C, \tilde{v})) = \mathbb{E}((\Pi(a_1, a_2, \tilde{v}^C, \tilde{v}) - \bar{\Pi}_{as})^2), \tag{A.13}$$

which is (using Π to denote $\Pi(a_1, a_2, \tilde{v}^C, \tilde{v})$ for shortening the equation):

$$\begin{aligned} \text{Var}_{as}(\Pi) &= \mathbb{E}\big[(\Pi - \bar{\Pi}_{n.as})^2\big] - 2\,\mathbb{E}\big[(\Pi - \bar{\Pi}_{n.as})(\bar{\Pi}_{n.as} - \bar{\Pi}_{as})\big] + (\bar{\Pi}_{n.as} - \bar{\Pi}_{as})^2 \\ &= \mathbb{E}\big[(\Pi(a_1, a_2, \tilde{v}^C, \tilde{v}) - \bar{\Pi}_{n.as})^2\big] - (\bar{\Pi}_{n.as} - \bar{\Pi}_{as})^2, \end{aligned} \tag{A.14}$$

where the second line follows from $\bar{\Pi}_{as} = \mathbb{E}(\Pi)$ (by definition).

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

$$\begin{aligned} \text{Var}_{as} = \frac{p(a_1)}{2}(a_1 - v_H - \bar{\Pi}_{n.as})^2 + \frac{p(a_1)}{2}(a_1 - v_L - \bar{\Pi}_{n.as})^2 + \\ (1 - p(a_1))\bar{\Pi}_{n.as}^2 - \frac{\Delta_D}{4}[(a_1 - v_L - \bar{\Pi}_{n.as})^2 - (a_1 - v_H - \bar{\Pi}_{n.as})^2)] - (\bar{\Pi}_{n.as} - \bar{\Pi}_{as})^2, \end{aligned} \tag{A.15}$$

and therefore:

$$\text{Var}_{as} = \text{Var}_{n.as} - \frac{\Delta_D}{4}[(a_1 - v_L - \bar{\Pi}_{n.as})^2 - (a_1 - v_H - \bar{\Pi}_{n.as})^2)] - (\bar{\Pi}_{n.as} - \bar{\Pi}_{as})^2, \tag{A.16}$$

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

$$\text{Var}_{as} = \text{Var}_{n.as} - \frac{\Delta_D \Delta_v}{2}[(a_1 - \mathbb{E}_{\frac{1}{2}}(v)) - \bar{\Pi}_{n.as}] - (\bar{\Pi}_{n.as} - \bar{\Pi}_{as})^2. \tag{A.17}$$

This can be simplified further by observing that:

$$\bar{\Pi}_{n.as} = p(a_1)(a_1 - \mathbb{E}_{\frac{1}{2}}(v)), \quad \textit{and} \quad \bar{\Pi}_{as} = p(a_1)(a_1 - \mathbb{E}_{\frac{1}{2}}(\tilde{v} \mid \tilde{v}^C > a)). \tag{A.18}$$

Hence:

$$\bar{\Pi}_{n.as} - \bar{\Pi}_{as} = p(a_1)(\mathbb{E}_{\frac{1}{2}}(\tilde{v} \mid \tilde{v}^C > a)) - \mathbb{E}_{\frac{1}{2}}(v)) = \frac{\Delta_D \Delta_v}{4}, \tag{A.19}$$

where the last equality follows from (6) for $\mu = 0.5$ and the definition of $p(a_1)$ (remember 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 (A.17) as:

$$\text{Var}_{as} = \text{Var}_{n.as} - \left[\frac{\Delta_D \Delta_v}{2}\left((a_1 - \mathbb{E}_{\frac{1}{2}}(v))(1 - p(a_1)) + \frac{\Delta_D \Delta_v}{8}\right)\right]. \tag{A.20}$$

To analyze the effect of σ on $\text{Var}_{n.as}$, observe first that $p(a_1) = \mathbb{E}_{\frac{1}{2}}(V(a, \tilde{v}^C))$ increases with σ for $a_1 \geq \mathbb{E}_{\frac{1}{2}}(v)$. Indeed:

$$\frac{\partial \mathbb{E}_{\frac{1}{2}}(V(a, \tilde{v}^C))}{\partial \sigma} = 0.5\left(\frac{\partial D(a, v_H)}{\partial \sigma} + \frac{\partial D(a, v_L)}{\partial \sigma}\right). \tag{A.21}$$

As $D(a, v) = 1 - G(a - v)$ and $G(.)$ 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:

$$\frac{\partial \mathbb{E}_{\frac{1}{2}}(V(a, \tilde{v}^C))}{\partial \sigma} = 0.5\left(\frac{\partial D(a, v_H)}{\partial \sigma} + \frac{\partial D(a, v_L)}{\partial \sigma}\right) > (\sqrt{2\pi}\sigma^2)^{-1} exp(-\frac{(a - v_L)^2}{2\sigma^2})(a - \mathbb{E}_{\frac{1}{2}}(v)) > 0, \tag{A.22}$$

for $a > \mathbb{E}_{\frac{1}{2}}(v)$. Thus, for $a > \mathbb{E}_{\frac{1}{2}}(v)$, $p(a_1)$ increases with σ. Hence, $p(a_1)$ is maximal when σ goes to infinity and therefore $p(a_1) < \frac{1}{2}$ (since $D(a, v)$ goes to $\frac{1}{2}$ when σ goes to infinity). It follows from (A.12) that $Var_{n.as}$ increases with σ. A similar reasoning shows that $Var_{n.as}$ increases with Δ_v.

Now consider the effect of σ on Var_{as}. The first term in (A.20) ($\text{Var}_{n.as}$) 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.20) 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 $a_1 > a_2$, AM_1 never trades and therefore $Var_{n.as}(a_1, a_2) = Var_{as}(a_1, a_2) = 0$.

A.4 Convergence

Consider the event that the greedy price remains constant at a_{m^*} after a certain date t, forever. We show below that this event has a probability equal to zero.

Observe that the Q-value $q_{m,t}$ of any price a_m above v_H must be at least equal to $q_{m,0} \times \alpha^t$. This corresponds to the case in which a_m would have been played in each period, and no trade occurred, which is worse than trading, as trading brings at least $a_m - v_H > 0$. As $q_{m,0} > 0$ in our specification, $q_{m,t} > 0$ too. Since a_{m^*} is the greedy price, it has the highest Q-value, and hence $q_{m^*,t}$ is greater than all other $q_{m,t}$, and hence strictly positive. Moreover, it is bounded above by $\bar{q}$ (corresponding to the profit realized when selling at the maximal price on the grid when $v = v_L$). As there are always several prices above v_H in our experiments, the second-largest value of the Q-matrix, denoted q', is necessarily strictly positive as well.

Now, imagine that from t onward, clients do not buy at the greedy price for δ episodes. Then at date $t + \delta$, the Q-value of the greedy price is $q_{m^*,t+\delta} = q_{m^*,t} \times \alpha^\delta \leq \bar{q} \times \alpha^\delta$. Consider δ large enough so that $\bar{q} \times \alpha^\delta < q'$. After such a sequence of no trades, a_{m^*} will hence no longer be the greedy price. Such a sequence always happens with a strictly positive probability p' because the likelihood of no trade at any price is always strictly positive. Conversely, the probability that a_{m^*} is still the greedy price after δ periods is lower than $(1 - p') > 0$. In this occurrence $q_{m^*,t+\delta} \leq \bar{q}$. Then, repeating the same reasoning, we conclude that the probability that a_{m^*} is still the greedy price in $t + 2\delta$ is $(1 - p')^2$. Iterating, we deduce that that the probability that the greedy price remains constant at a_{m^*} forever is zero.

A.5 Nash Equilibria

In this section, we analyze the Nash equilibria of the one-shot game when dealers are constrained to choose prices from a finite grid (i.e., there is a positive tick size). While the Glosten-Milgrom equilibrium is generically unique when dealers can choose any price on the real line, this uniqueness does not necessarily hold when prices must be chosen from a grid $\mathcal{A}$. In particular, for some of our experimental settings, the game admits exactly two pure strategy equilibria and at least one mixed strategy equilibrium. Our goal in this section lies in identifying the highest price that can be played in equilibrium.

More formally, we say that a price $a \in \mathcal{A}$ is a *possible pure equilibrium price* if there is a pure strategy equilibrium in which dealers play a. It is useful to first characterize the set of pure strategy

equilibria. To this end, for any price $a \in \mathcal{A}$, let denote with $\bar{\Pi}(a;\mu)$ the expected payoff of a monopolist dealer when setting a price a:

$$\bar{\Pi}(a;\mu) := \mu D(a, v_H)(a - v_H) + (1-\mu)D(a, v_L)(a - v_L).$$

Lemma 1. *In the game with N dealers, an action profile $\{a_1, a_2, \dots a_N\} \in \mathcal{A}^N$ is a pure Nash equilibrium if and only if*

1. *All dealers set the same price $a \in \mathcal{A}$.*
2. *The price a satisfies $\bar{\Pi}(a;\mu) \geq 0$ and*

$$\frac{1}{N}\bar{\Pi}(a;\mu) \geq \max_{a' \in \mathcal{A},\, a' < a} \{\bar{\Pi}(a';\mu)\} \tag{A.23}$$

Proof. **Sufficiency:** Suppose that all dealers except dealer i set a price equal to a, and that a satisfies condition (A.23). Consider the best response of dealer i. The expected payoff from playing $a' = a$ is $\frac{1}{N}\bar{\Pi}(a;\mu) \geq 0$, as dealer i shares the payoff $\bar{\Pi}(a;\mu)$ with the other $N-1$ dealers. The expected payoff from undercutting the others by playing some $a' < a$ is $\bar{\Pi}(a';\mu) \leq \frac{1}{N}\bar{\Pi}(a;\mu)$, by condition (A.23). On the other hand, the expected payoff from playing some $a'' > a$ is $0 \leq \frac{1}{N}\bar{\Pi}(a;\mu)$, since no client trades with dealer i. Hence, $a_i = a$ for all i constitutes a pure Nash equilibrium.

Necessity: 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 price offered. If $\bar{\Pi}(a^{\min};\mu) < 0$, then the dealer offering $a^{\min}$ earns a negative profit and can profitably deviate by setting any $a' > v_H$. Since prices lie on a discrete grid, we can assume without loss of generality that $\bar{\Pi}(a^{\min};\mu) > 0$.

If there exists a dealer i such that $a_i \neq a^{\min}$, then $a_i > a^{\min}$ and his payoff is zero. But then dealer i can profitably deviate by switching to $a^{\min}$, thereby obtaining a share of the strictly positive payoff $\bar{\Pi}(a^{\min};\mu)$. Hence, all dealers must play $a^{\min}$ and receive a payoff of $\frac{1}{N}\bar{\Pi}(a^{\min};\mu)$.

If there exists $a' < a^{\min}$ such that $\bar{\Pi}(a';\mu) > \frac{1}{N}\bar{\Pi}(a^{\min};\mu)$, then playing a' would constitute a profitable deviation. Hence, conditions 1 and 2 in the Lemma are necessary for equilibrium. □

We can now prove the main result of this section. Denote by $\hat{a}$ the highest among the possible pure equilibrium prices.

Lemma 2. *If a is a price played with positive probability in a Nash equilibrium, then $a \leq \hat{a}$.*

Proof. The result is obvious for any price a played in a pure strategy equilibrium by definition of $\hat{a}$. Let us now consider any given mixed strategy equilibrium, and let $a^\star$ denote the highest price in the support of this equilibrium. We prove by contradiction that $a^\star \leq \hat{a}$.

Suppose $a^\star > \hat{a}$. Since we are considering a mixed strategy equilibrium, for any given dealer n, the equilibrium payoff must be equal to the expected gain from playing $a^\star$. Because $a^\star$ is the highest price in the support, the dealer's equilibrium payoff is:

$$\frac{\eta}{N}\bar{\Pi}(a^\star;\mu)$$

where $\eta < 1$ is the equilibrium probability that all other dealers play $a^\star$.

If instead the dealer plays any price $a' < a^\star$, he will receive $\bar{\Pi}(a';\mu)$ at least when all the other dealers play $a^\star$. Hence, his expected payoff is at least $\eta\bar{\Pi}(a';\mu)$. Since $a^\star$ is in the support of the mixed equilibrium, deviating to $a' < a^\star$ cannot be a profitable deviation and hence we must have $\eta\bar{\Pi}(a';\mu) \leq \frac{\eta}{N}\bar{\Pi}(a^\star;\mu)$ for any $a' < a^\star$. This inequality implies that $a^\star$ satisfies condition (A.23), and thus there exists a pure strategy equilibrium in which all dealers play $a^\star$. Since $a^\star$ is a pure equilibrium price, it must be that $a^\star \leq \hat{a}$, contradicting $a^\star > \hat{a}$. □

Online Appendix for "Algorithmic Pricing and Liquidity in Securities Markets"

Jean-Edouard Colliard, Thierry Foucault, and Stefano Lovo

This Online Appendix provides additional results and robustness tests.

OA.1 The two-period case

OA.1.1 Glosten-Milgrom benchmark

As a benchmark, we derive the two-period Glosten-Milgrom prices. Denote the dealers' belief about the likelihood that $v = v_H$ prior to the arrival of the τ^{th} client by μ_τ. Thus, $\mu_1 = \mu$. 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 dealers' Bayesian beliefs about the likelihood that $v = v_H$ after observing H_1. We have:

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

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

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^*.

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, as claimed in the main text, Bayesian dealers revise their estimate of the expected payoff of the asset upwards after a buy in period 1 and downwards after no trade.

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

Step 1. 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:

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

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:

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

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

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.

Step 2. 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 $a(.)$ is concave, by Jensen's inequality we have:

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

Finally, because Bayesian belief are martingales, we have $\mathbb{E}[\mu_2] = \mu_1$. 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) $s_n \in \mathcal{S} = \left\{0, \frac{1}{N}, \frac{1}{N-1}, ..., \frac{1}{2}, 1\right\}$ 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 $[\underline{q}, \overline{q}]$. Then, in each episode t, we do the following:

Period 1:

1. For each AM_n, we define $m_{n,t}^{1,*} = arg\max_m q_{m,\emptyset,n,t-1}$ the index associated with the highest value in matrix $\mathbf{Q}_{n,t-1}$ in state $s = \emptyset$, 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 that 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:

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

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. 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:

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

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:

$$\bar{V}_2 = \frac{\sum_{k=1}^{K} V_2^k}{K} \tag{OA.13}$$

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

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

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

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

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 parametrization 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 averages 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:

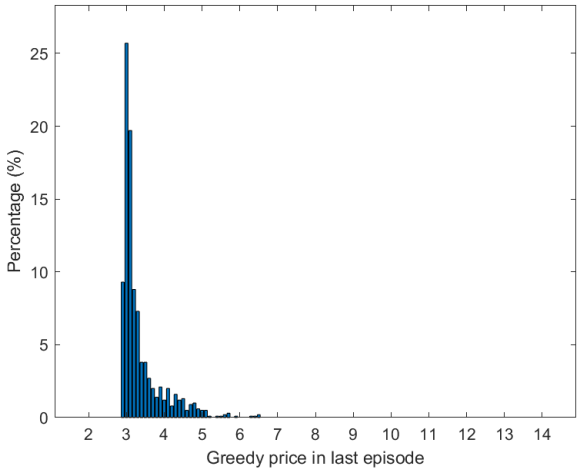
$$q_{m,n,t} = \begin{cases} \frac{\pi_{n,t} + \nu_{m,n,t} q_{m,n,t-1}}{1 + \nu_{m,n,t}} & \text{if } a_{n,t} = a_m \\ q_{m,n,t} & \text{if } a_{n,t} \neq a_m \end{cases} \tag{OA.18}$$

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.

Figure OA.1: Greedy price of AM_1 when AM_1 and AM_2 keep experimenting in the long-run. Adverse-selection case and 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.

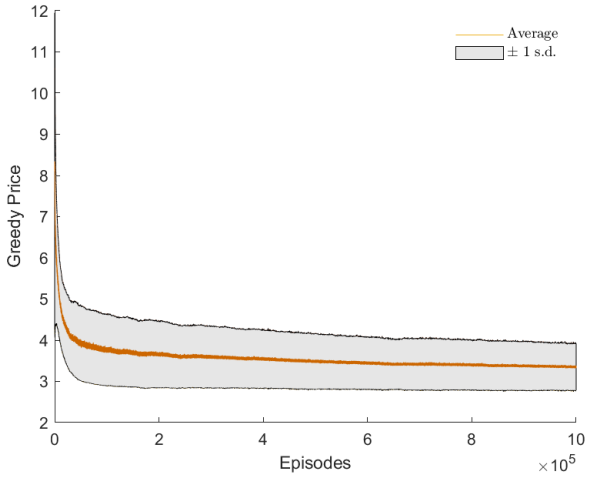
Panel A: Distribution of the greedy price of AM_1 in the last episode.

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 $a^*_{1,T} = a$. The mode of the distribution is 3.0, and all prices are between 2.9 and 6.5.



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). Greedy prices start from an average of about 8 and converge to around 4 after 200,000 episodes. The standard deviation is about 4 at the start and decreases to around 1 after 200,000 episodes.



OA.3 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.

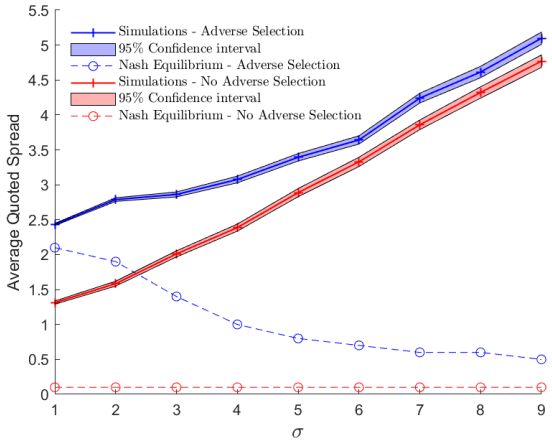
Note that in all cells of Table OA.1 the values of σ and Δ_v are equal to their baseline values and the tick size is 0.1. Thus, the least competitive Nash equilibrium price is 2.68 in each case. Our baseline hyperparameters (α_m, β_m) for both AMs give an average price of 5.03. If both AMs have the same 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.

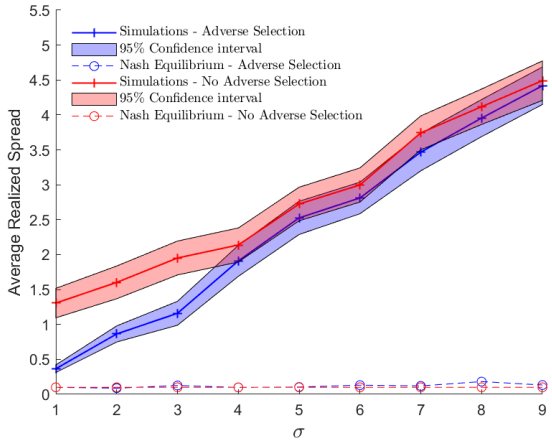
Table OA.1: Final price for different hyperparameters. This matrix gives the average value of the best price across all experiments in the last episode. For instance, if AM_1 chooses (α_m, β_l) and AM_2 chooses (α_l, β_l), the best price in the last episode is 4.26. Hyperparameters: $\alpha_l = 0.001, \alpha_m = 0.01, \alpha_h = 0.1$; $\beta_l = 5.10^{-6}, \beta_m = 8.10^{-5}, \beta_h = 3.2.10^{-4}$.

		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)	5.37	5.42	5.39	4.26	4.34	4.34	3.92	3.99	3.98
	(α_l, β_m)	5.42	5.73	5.71	4.19	4.81	4.8	3.88	4.01	4
	(α_l, β_h)	5.39	5.71	5.69	4.19	4.83	4.84	3.88	3.99	4
	(α_m, β_l)	4.26	4.19	4.19	4.11	4.1	4.14	3.98	4.08	4.07
	(α_m, β_m)	4.34	4.81	4.83	4.1	5.03	5.04	3.87	4.4	4.41
	(α_m, β_h)	4.34	4.8	4.84	4.14	5.04	5.07	3.86	4.43	4.43
	(α_h, β_l)	3.92	3.88	3.88	3.98	3.87	3.86	4	4.01	4.01
	(α_h, β_m)	3.99	4.01	3.99	4.08	4.4	4.43	4.01	4.15	4.18
	(α_h, β_h)	3.98	4	4	4.07	4.41	4.43	4.01	4.18	4.34

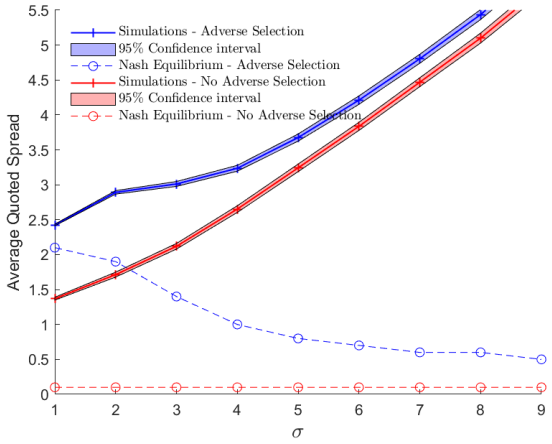
Figure OA.2: Average quoted spread and realized spread in the last episode, for different values of σ. Each line corresponds to a different parametrization of the algorithms, with a low value of α (0.001) and different values of β. All panels display the same features: Simulated quoted spreads are increasing in σ, higher in the adverse selection case than in the no adverse selection case, higher in the simulations than in the Glosten-Milgrom benchmark. The same applies for the simulated realized spreads, except that they are higher in the no adverse selection case than in the adverse selection case.



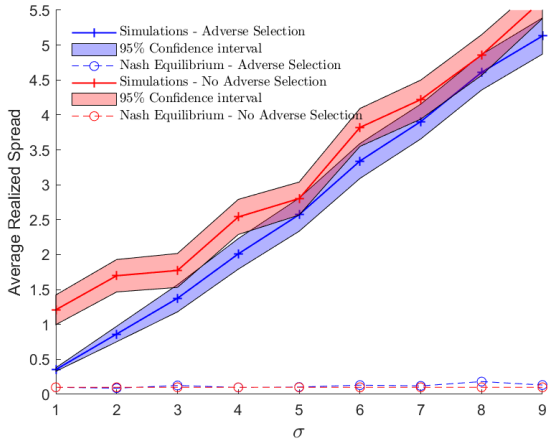
Quoted spread, $\alpha = 0.001, \beta = 5 \times 10^{-6}$



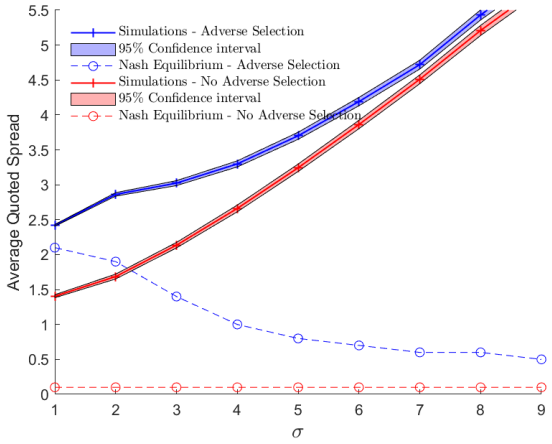
Realized spread, $\alpha = 0.001, \beta = 5 \times 10^{-6}$



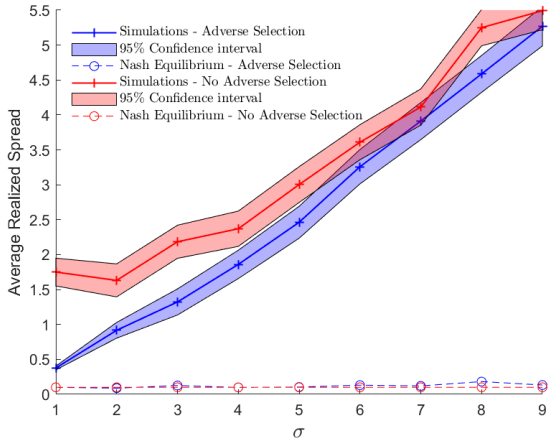
Quoted spread, $\alpha = 0.001, \beta = 8 \times 10^{-5}$



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

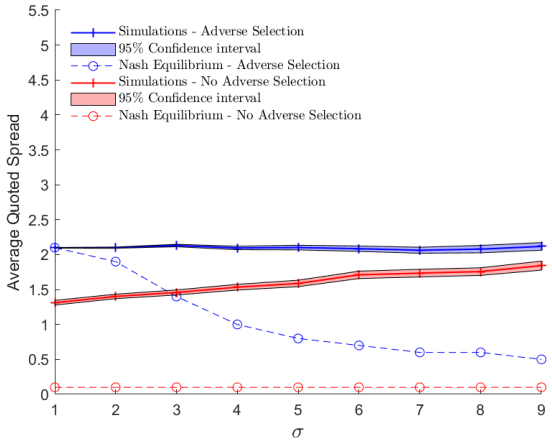


Quoted spread, $\alpha = 0.001, \beta = 3.2 \times 10^{-4}$

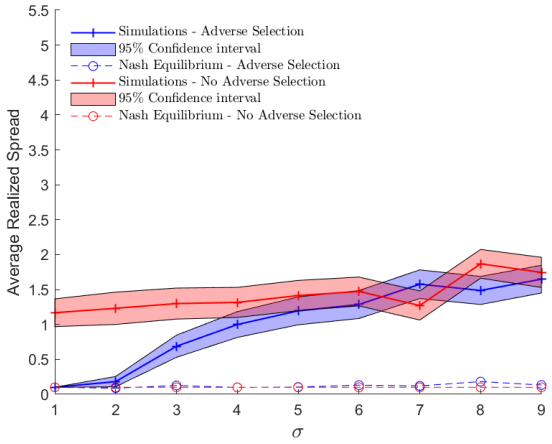


Realized spread, $\alpha = 0.001, \beta == 3.2 \times 10^{-4}$

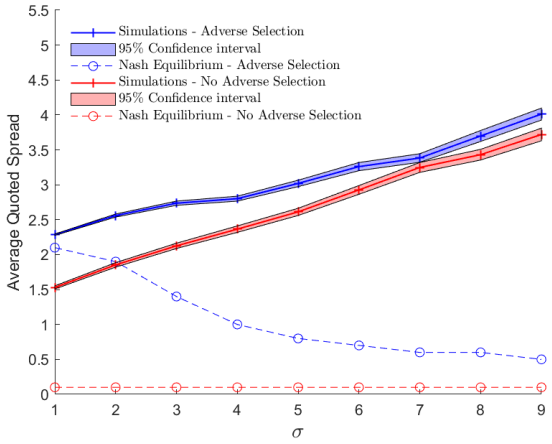
Figure OA.3: Average quoted spread and realized spread in the last episode, for different values of σ. Each line corresponds to a different parametrization of the algorithms, with the baseline value of α (0.01) and different values of β. All panels display the same features: Simulated quoted spreads are increasing in σ, higher in the adverse selection case than in the no adverse selection case, higher in the simulations than in the Glosten-Milgrom benchmark. The same applies for the simulated realized spreads, except that they are higher in the no adverse selection case than in the adverse selection case.



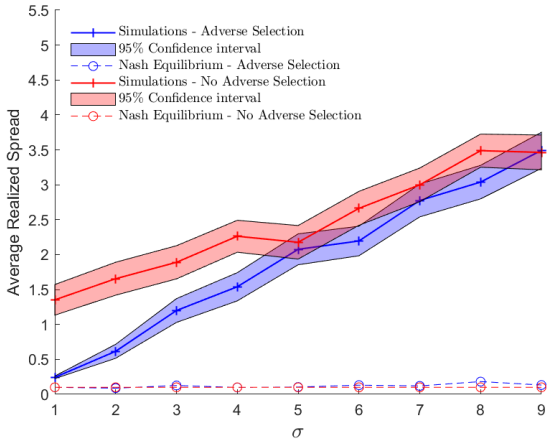
Quoted spread, $\alpha = 0.01, \beta = 5 \times 10^{-6}$



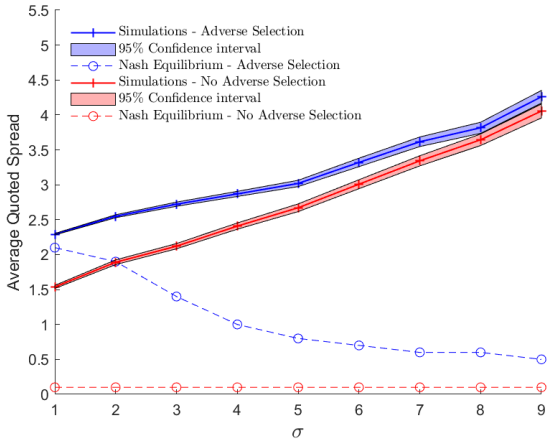
Realized spread, $\alpha = 0.01, \beta = 5 \times 10^{-6}$



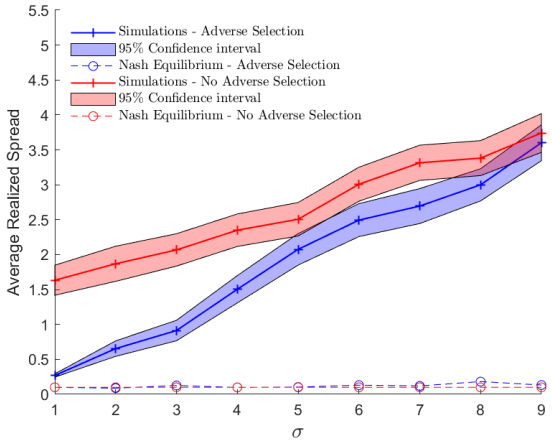
Quoted spread, $\alpha = 0.01, \beta = 8 \times 10^{-5}$



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

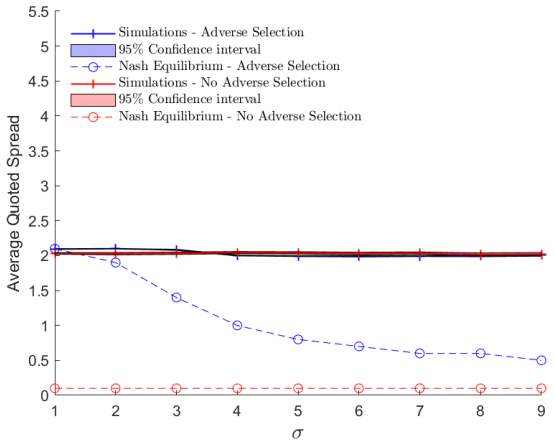


Quoted spread, $\alpha = 0.01, \beta = 3.2 \times 10^{-4}$

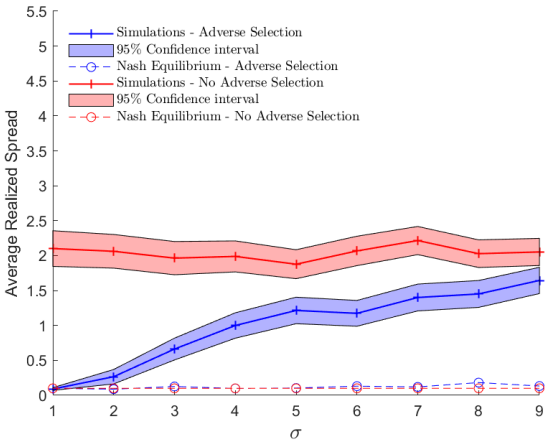


Realized spread, $\alpha = 0.01, \beta == 3.2 \times 10^{-4}$

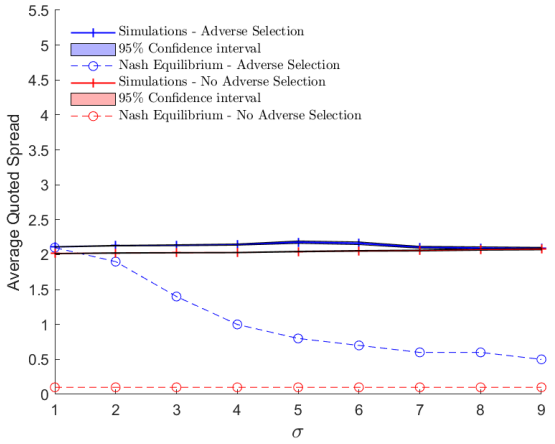
Figure OA.4: Average quoted spread and realized spread in the last episode, for different values of σ. Each line corresponds to a different parametrization of the algorithms, with a high value of α (0.1) and different values of β. All panels display the same features: Simulated quoted spreads are flat or slightly increasing in σ, equal or slightly higher in the adverse selection case than in the no adverse selection case, higher in the simulations than in the Glosten-Milgrom benchmark. Simulated realized spreads are flat with respect to σ in the no adverse selection case, increasing in σ in the adverse selection case, higher in the former case, and higher than in the Glosten-Milgrom benchmark.



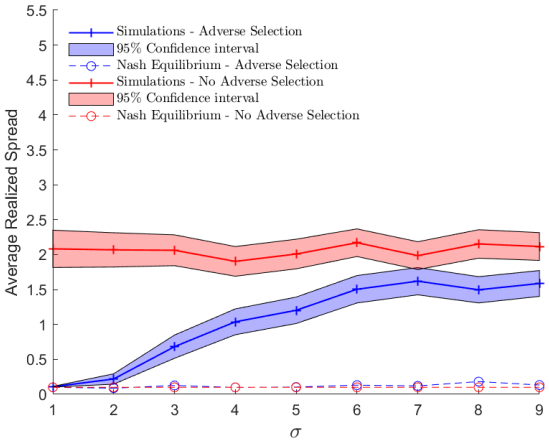
Quoted spread, $\alpha = 0.1, \beta = 5 \times 10^{-6}$



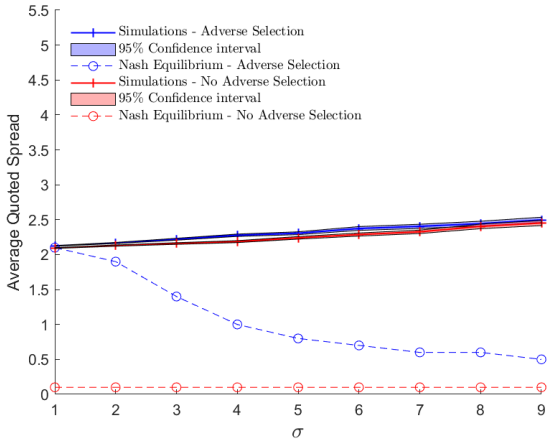
Realized spread, $\alpha = 0.1, \beta = 5 \times 10^{-6}$



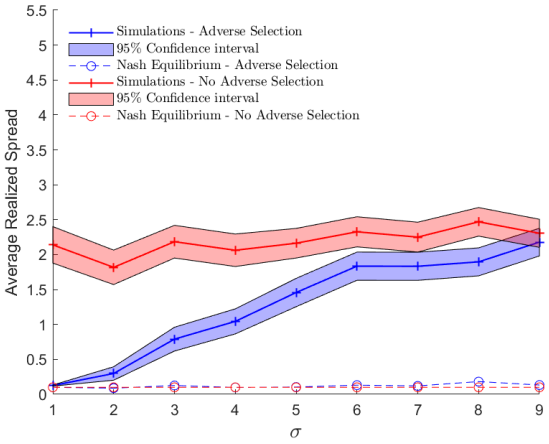
Quoted spread, $\alpha = 0.1, \beta = 8 \times 10^{-5}$



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



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



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

OA.4 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 several 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 time 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.4, 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 greedy 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 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:

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

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:

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

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 episodes 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.

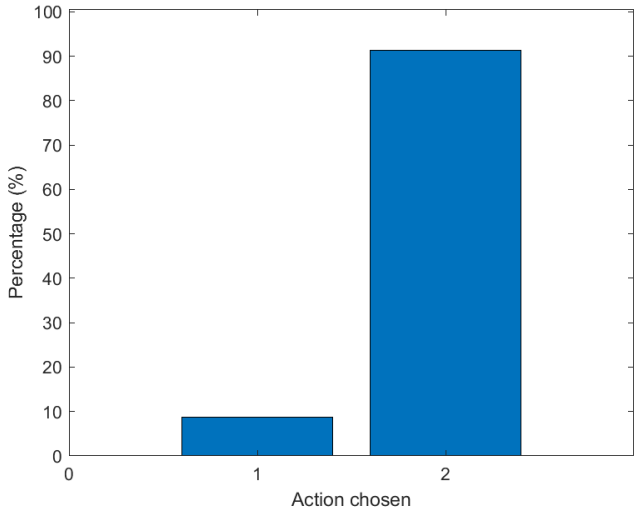
The reason for this bias is that the second procedure conditions the observation on having the same action taken 100 times in a row. Conditionally on being in state Q_2 and playing a_2, the probability of remaining in Q_2 is 0.9. The probability to remain in Q_2 for 100 episodes in a row is $0.9^{100} \simeq 2.65 \times 10^{-5}$, 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.

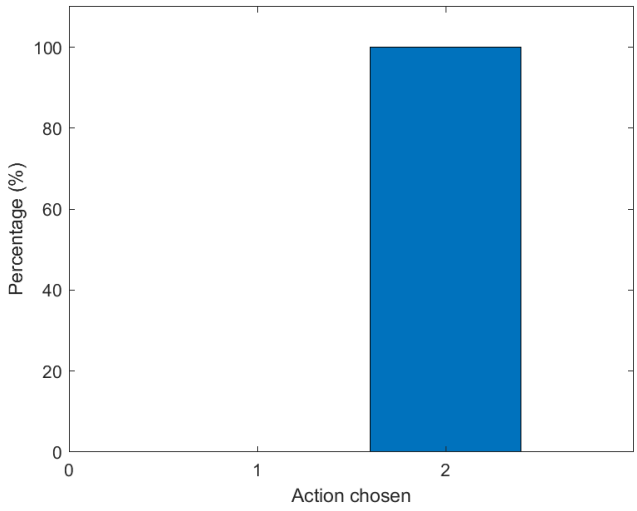
This example is clearly extreme and meant only for illustration. With lower values of α and actions that are less different we do not expect the two procedures to lead to radically different results. However, given that the random stopping procedure is in principle biased and is also typically much more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

Figure OA.5: Percentage of experiments ending with actions a_1 and a_2, fixed stopping time and random stopping time procedures. Action a_1 gives π_1^h with probability 0.1 and 0 otherwise, and a_2 gives $\pi_2^h < \pi_1^h$ with probability 0.9, and 0 otherwise. A Q-learning algorithm chooses between a_1 and a_2, with $\alpha = 1$ and $\beta = 0.0008$. With the fixed time procedure (Panel A), the experiment stops after $T = 50,000$ episodes. With the random time procedure (Panel B), the experiment stops after $T = 50,000$ episodes and the algorithm playing the same action 100 times in a row. Panel A shows that action a_1 is picked 10% of the time with the fixed stopping time procedure, and Panel B that it is never picked with the random stopping time procedure.

Panel A: Fixed stopping time procedure.



Panel B: Random stopping time procedure.



OA.5 Choice of α and β

We detail the statistical tests we conduct in Section 6.1. We denote $\theta_n = (\alpha_n, \beta_n)$ the hyperparameters chosen by dealer n. We take the perspective of dealer 1 and consider his total profit over T episodes as in (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 dealer 1 as:

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

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 $k \in \{1, ...K\}$ and compute:

$$\bar{u}^* = \frac{1}{K}\sum_{k=1}^{K} u_{1,k}^* \tag{OA.22}$$

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

$\bar{u}^*$ is dealer 1's estimate of his expected profits, and $\bar{s}^*$ is his 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-period profit obtained by dealer 1 in experiment $k' \in \{K+1, \ldots, K+K'\}$ and compute:

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

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

We ask 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:

$$\Delta\bar{u} = \bar{u}' - \bar{u}^* \tag{OA.26}$$

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

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

$$t = \frac{\Delta\bar{u}}{s_{\Delta\bar{u}}}. \tag{OA.29}$$

Under the null hypothesis H_0 that $u(\theta_1^*, \theta_2^*) = u(\theta', \theta_2^*)$, the statistics 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 OA.2 reports the p-values we obtain for different θ' and different values of K'.[41]

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'.

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)	0.79	0.76	0.67	0.88	0.89	0.92	0.93	0.94	0.91	0.89
(α_m, β_h)	0.2	0.02	0.08	0.05	0.06	0.13	0.1	0.18	0.35	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

As we observe in the table, for all values of K' dealer 1 can definitely reject the possibility that changing α is profitable. 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 any K', dealer 1 can relatively safely conclude that using a lower β does not bring superior profits. Using a higher β is less clear. Dealer 1 seems to have been particularly"lucky" with (α_m, β_h) in experiments 100 to 200, so that with $K' = 200$ dealer 1 would reject the hypothesis that (α_m, β_h) is not more profitable than (α_m, β_m), with a p-value of 2%. However, for higher values of K' the p-value increases again, up to around 30%.

[41] Note 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 $K' = 200$.

Thus, in any case, if dealers had an incentive to change their parameters this would be towards experimenting less, not more, which would lead to even less competitive prices than under our baseline specification.

We then repeat this exercise for all 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 OA.3, we 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.

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%.

	(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

OA.6 Machine against Omniscient Market Maker

In this Appendix we explore some possible outcomes from the interaction between an AM and a more sophisticated market maker. Namely, can an AM survive such competition despite not choosing what would be the rational best response to the other market maker's quote?

The answer is likely to depend a lot on the knowledge of the non-algorithm about the environment. We consider the worst-case scenario and assume that an AM is facing an "Omniscient Market Maker" (OM). The AM is programmed as in our baseline model, and the OM is rational and has full information: she knows the price posted by the AM in each period and she knows the expected payoff she can obtain at each price. We will show that it is not optimal for the OM to play so as to completely exclude the AM from the market, which shows that, even in this extremely unfavorable

situation, the AM survives sophisticated competition.

We first consider an "Exclusion" strategy such that the AM trades with probability zero in the long-run and is therefore "excluded" from the market. We pick the least costly way for the OM to ensure this long-term exclusion. Then, we consider an "Accommodation" strategy, which aims at "teaching" the AM to play just above the monopoly price, by giving it a small profit, which enables the OM to obtain the monopoly profit most of the time. We then provide an experimental example in which "accommodation" is more profitable for the OM than "exclusion".

Benchmark. As an example, we run 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 parametrization, the Nash equilibrium price a^* is unique and equal to 3, and the monopoly price is $a^m = 7$. 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.

Exclusion. We then assume that the OM follows "Exclusion": (i) if the AM plays a price $a \leq a^*$, then the OM plays a^* ; (ii) if the AM plays a price $a \in (a^*, a^m]$, then the OM undercuts by one tick; (iii) if the AM plays a price $a > a^m$, then the OM plays the monopoly price a^m. This strategy ensures that the AM is always excluded from the market in the long-run, due to (ii). Moreover, (i) and (iii) ensure that this is the least costly way for the OM to guarantee that the AM is always excluded.

Figure OA.7 shows the outcome of the interaction between an AM and an OM following "Exclusion". We observe that after around 10,000 episodes, the AM stops trading because it no longer plays prices below a^*, and is hence always undercut. 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 (Panel B). Thus, even though the AM keeps posting prices, it does not participate to trading. In fact all orders are executed by the OM, who on average earns a profit of 0.53 (Panel D). As the figure shows, the OM's strategy pays off because it enables the OM to trade at high prices. In fact, across experiments, the price posted by the OM is roughly uniformly distributed in the range $[3, 7]$.

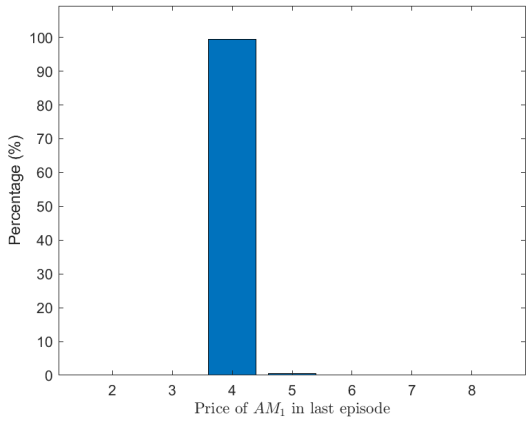
Accommodation. For any episode t, the strategy is as follows. Let a^{m+} be the first price on the grid above the monopoly price a^m. If the AM plays a^{m+}, and the Q-value of a^{m+} is such that a^{m+} remains the greedy price even if the OM undercuts by posting a^m, then the OM posts a^m. If

instead by posting a^m the OM would induce a greedy price below a^{m+}, then the OM posts a^{m+} so that she shares her profit with the AM.

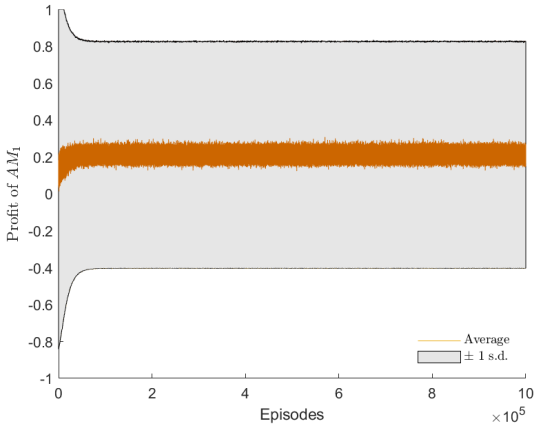
Intuitively, this strategy "teaches" the AM to play a price just above the monopoly price at a relatively low cost (the OM just needs to share profits at a price of a^{m+} from time to time). Figure OA.8 shows that for the parameter values considered in our simulations, this strategy works well. In most experiments, in the final episode, the OM posts the monopoly price of 7 while the AM posts a price of $a^{m+} = 8$.[42] As a result, the OM obtains an average profit of 0.76 on average in the last episode, which is strictly above the average profit of 0.53 obtained with the exclusion strategy described above. The AM is not completely excluded from the market and makes a small profit.

In conclusion, this example shows that, even if an AM faced competition from an omniscient market maker, the AM would still be active in the long-run, because it is not in the omniscient market maker's interest to drive the AM out of the market.

Figure OA.6: Two AMs. Distribution of the price chosen by each AM in the last episode, evolution of average payoffs over episodes.



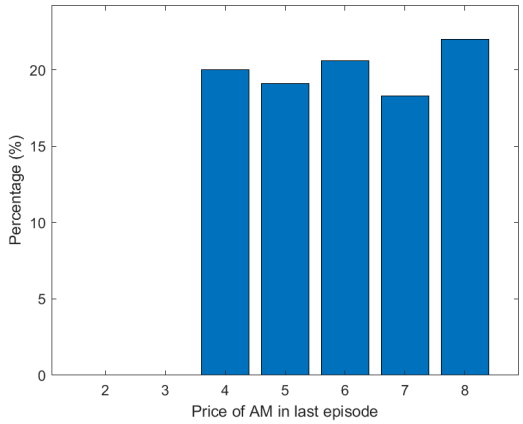
(a) Distribution of AMs' quotes. Almost all quotes are at 4, and the rest at 5.



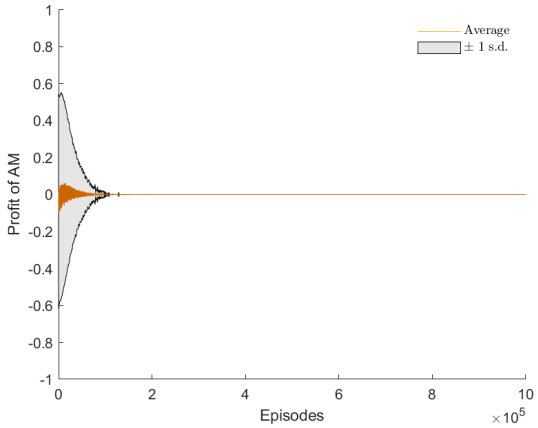
(b) Evolution of AMs' average profits: Average profits reach around 0.2 after 20,000 episodes and then remain at that level, with a standard deviation around 0.6.

[42] In the last episode, the AM plays 8 in 992 out of 1,000 experiments while the OM plays 7 (and gets the monopoly profit) in 988 out of 1,000 experiments.

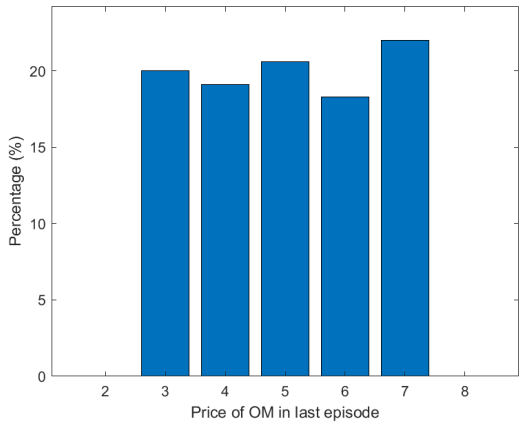
Figure OA.7: The OM plays "Exclusion". Distribution of the price chosen by each player in the last episode, evolution of average payoffs over episodes.



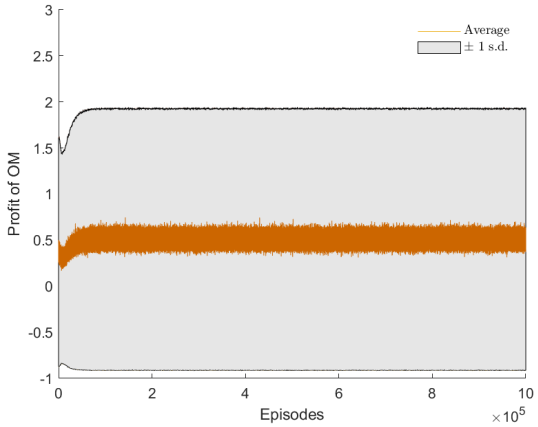
(a) Distribution of the AM's quotes: Quotes are approximately uniformly distributed between 4 and 8.



(b) Evolution of the AM's average profits: Average profits converge to around 0 after 100,000 episodes, with a standard deviation close to 0.

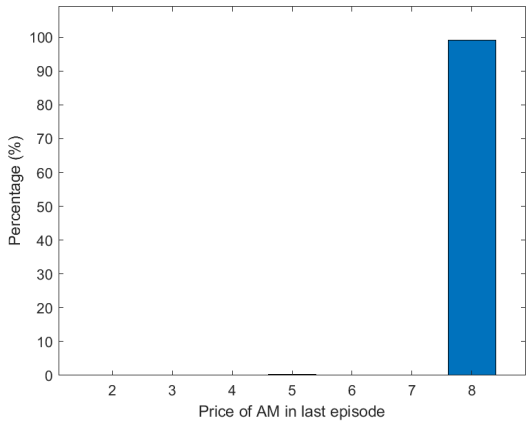


(c) Distribution of the OM's quotes: Quotes are approximately uniformly distributed between 3 and 7.

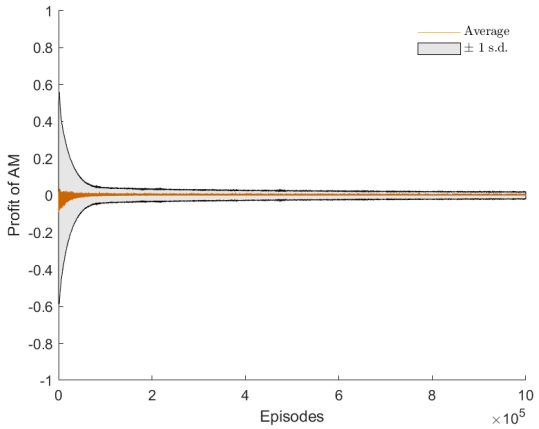


(d) Evolution of the OM's average profits: Average profits converge to around 0.5 after 100,000 episodes, with a standard deviation around 1.5.

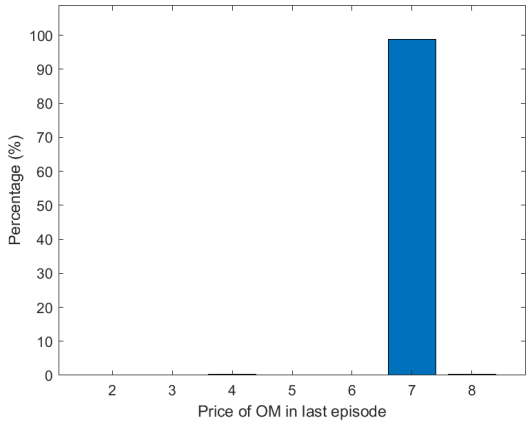
Figure OA.8: The OM plays "Accommodation". Distribution of the price chosen by each player in the last episode, evolution of average payoffs over episodes.



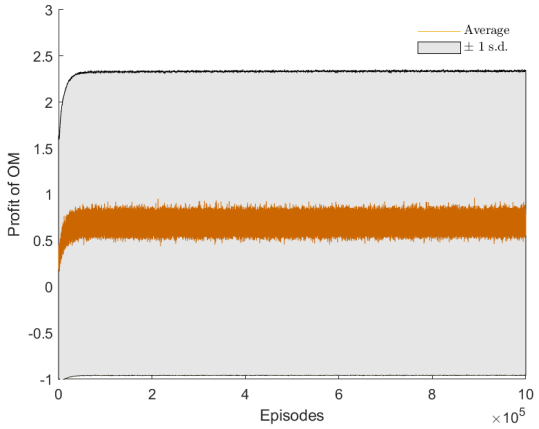
(a) Distribution of the AM's quotes: Almost all quotes are at 8, and the rest at 5.



(b) Evolution of the AM's average profits: Average profits converge to around 0 after 100,000 episodes, with a small but positive standard deviation.



(c) Distribution of the OM's quotes: Almost all quotes are at 7, and the rest at 4.



(d) Evolution of the OM's average profits: Average profits converge to around 0.6 after 100,000 episodes, with a standard deviation around 1.5.