# Trading against Algorithms: Price Dynamics and Risk-sharing in a Market with Q-learners

Snehal Banerjee and Martin Szydlowski*

First Draft: August 5, 2025
This Draft: June 12, 2026

**Abstract**

We study pricing dynamics and risk-sharing in a market with rational investors and Q-learning traders. The Q-learners' trading generates a feedback loop in prices: their demand for the risky security depends on their perceived benefit from trading, which in turn, depends on realized returns. We show that this loop generates state-dependent stochastic volatility, predictable returns, and novel price dynamics which depend on the mass and learning rate of the Q-learners. When rational investors have strong risk-sharing motives for trading, we show that Q-learners can (i) earn trading profits and (ii) improve average investor utility, even though they increase the volatility of prices.

**JEL Classification:** D83, G12, G14, G40

**Keywords:** Algorithmic trading, Reinforcement learning, Liquidity Predictability, Human-AI interaction

*Banerjee (snehalb@umich.edu) is at the University of Michigan - Ann Arbor and and Szydlowski (mszydlowski@ust.hk) is at the Hong Kong University of Science and Technology. We thank audiences at CKGSB, CUHK, NYU-Shanghai, the 2nd Asian FTG conference, and the FIRS 2026 Annual Meeting as well as Martino Banchio, Marc Dordal i Carreras, In-Koo Cho, Jean-Edouard Colliard, Winston Dou, Yunzhi Hu, John Kuong, Ali Lazrak, Tse-Chun Lin, Konstantin Milbradt, Mariann Ollar, Emiliano Pagnotta, Kevin Smith, Jiang Wang, Noah Williams, and Andrzej Skrzypacz for helpful comments.

# 1 Introduction

Algorithmic trading is transforming financial markets. An increasing share of trading volume in US stocks is generated by algorithms that are designed to learn and adapt to market conditions in real time.[1] Moreover, such strategies are no longer restricted to high frequency traders or quantitative hedge funds, but are increasingly popular among ETF providers and large asset managers.[2] As a result, regulators are increasingly concerned about algorithmic trading's effect on price volatility, market stability, shock amplification, and liquidity.[3] While a growing body of empirical and simulation-based work explores the implications of such trading, a theoretical framework for understanding their economic impact is largely missing from the literature.

We develop a stylized model to answer a fundamental question: How does the presence of algorithmic traders affect equilibrium asset prices, return dynamics, and the welfare of rational investors? Existing simulation-based studies are limited in their ability to address this question because they often restrict attention to economies that are exclusively populated by algorithmic agents.[4] In contrast, our paper provides a first step in understanding human-AI interaction in financial markets. By explicitly modeling how rational investors and algorithms trade against each other, we can characterize the impact of algorithmic trading not only on market outcomes, but also on investor welfare.

**Model.** We consider a stylized, continuous-time economy in which there is a continuum of rational investors with CARA utility and a continuum of algorithmic traders of non-negligible mass. The algorithmic traders use Q-learning, which is a foundational reinforcement learning algorithm known for its tractability, economic interpretability, and performance.[5] The market participants trade a risky security in fixed supply which pays dividends following an arithmetic Brownian motion. In the absence of Q-learners, the equilibrium is standard: the

[1] For example, Brogaard, Hendershott, and Riordan (2014) finds that algorithms make up 42% of the trading volume in stocks, and Chaboud, Chiquoine, Hjalmarsson, and Vega (2014) find that algorithms make up above 60% of trading volume in some currency markets.

[2] Blackrock, the largest asset manager in the world, introduced active ETFs that are managed by machine learning algorithms in 2018, and more recently announced the introduction of a virtual investment analyst, "Asimov", for use by the firm's fund managers. Traditional money managers, like AQR, are also quickly adopting machine learning-based strategies.

[3] See Regulatory approaches to Artificial Intelligence in finance, OECD Intelligence Papers, No. 24, Sept. 2024, and Financial Stability in Focus: Artificial intelligence in the financial system, Bank of England, Financial Policy Committee, April 2025, for instance.

[4] The focus of many of these papers, as we discuss in Section 2, is the microstructure impact of algorithms on market efficiency and liquidity. In contrast, our focus is on understanding the asset pricing and welfare implications of such traders, and as such, we require a model in which algorithms and human investors interact.

[5] See e.g. Wiering and Otterlo (2012), Ch. 1.7.1 or Sutton and Barto (2018), Ch. 6.5.

price reflects the present value of future dividends, adjusted for a constant risk premium, which depends on the supply of the asset. Expected returns are constant, and prices exhibit no predictability or stochastic volatility.

When Q-learners are introduced, the equilibrium changes fundamentally. Rational investors face a nontrivial forecasting problem, because their optimal portfolio choice depends on their expectation of how Q-learners will trade in the future. The Q-learners' future behavior, in turn, depends endogenously on realized returns. With standard Q-learning in discrete time, solving for this equilibrium is not analytically tractable as it involves tracking the evolution of an entire distribution of Q-values. However, by taking the appropriate continuous-time limit, we show that the Q-learners' trading strategies can be summarized by a single state variable, $Q_t$, which reflects their current estimate of the net benefit to buying a share, and is predictable given the price path.[6]

When $Q_t$ is positive (negative), Q-learners are more likely to buy (sell) shares of the security. Moreover, a positive return realization increases the estimated net benefit, while a negative return realization decreases it. Because realized returns affect the evolution of $Q_t$, and consequently, future trading decisions, there is a feedback loop between the security price and the Q-learners'trading. In particular, because rational investors are risk-averse, they require compensation for the induced variation in residual asset supply in the form of a (stochastic) risk premium. This feedback loop between the security price and the Q-learners' trading leads to an amplification of fundamental shocks and is potentially destabilizing. We characterize conditions on the Q-learning algorithm under which there exists an equilibrium.

**Return dynamics.** The feedback loop between returns and the Q-values gives rise to rich endogenous dynamics. We show that the equilibrium $Q_t$ process is stochastic and exhibits mean-reversion. A positive dividend shock increases the net benefit of a share, $Q_t$, and thus Q-learner demand. This pushes the price up, which makes it less attractive to buy the risky asset going forward, which leads to lower future $Q_t$. The volatility of the $Q_t$ process depends on how sensitive it is to innovations in security returns, which depends on the learning rate of the Q-learning algorithm. Moreover, the Q-learners' demand is most sensitive to $Q_t$ when the algorithm is most uncertain about the benefit of trading, i.e., when $Q_t$ is near zero.

As a result, in the presence of Q-learners, security prices exhibit stochastic volatility

[6]Specifically, we assume that the Q-learners can only buy or sell up to one share of the risky asset, but (i) use Boltzmann exploration when choosing their demand and (ii) engage in counterfactual learning. As we show in Section 3.3, this allows us to reduce the number of state variables that drive the Q-learners' demand to $Q_t$. A key technical contribution of the paper is to formally show that the evolution of the $Q_t$ process can be approximated by a stochastic differential equation in continuous time, where the (instantaneous) innovation in the $Q_t$ process is driven by instantaneous return process. Moreover, as we discuss in Section 4.3, our analysis is robust to relaxing the assumption that the rational investors know the hyper-parameters of the Q-learning algorithm.

and predictable returns, even though fundamentals evolve as a Brownian motion. Since contemporaneous Q-learner demand is positively correlated with dividend shocks, trading by Q-learners amplifies return volatility. In fact, we show that return volatility is a hump-shaped function of $Q_t$: it is highest for $Q_t$ near zero (but lower for extreme $Q_t$), since Q-learner demand is very sensitive to realized returns in this region. Moreover, return volatility increases in both the mass of Q-learning traders and the rate of learning of the Q-learning algorithm.

We show that expected returns can also depend non-monotonically on $Q_t$. This is because the expected return on the security depends on the product of the return volatility and (approximately) the residual supply of the security that rational investors have to bear in equilibrium. Since Q-learner demand affects the residual supply borne by rational investors, higher $Q_t$ typically lowers the expected return by lowering the quantity of risk investors have to bear. However, because the return volatility is a hump-shaped function of $Q_t$ which peaks at zero, expected returns can be increasing, then decreasing, and then increasing again in $Q_t$.

Finally, because rational investors are risk averse and trading by Q-learners induces stochastic shifts in the residual supply of the asset, returns exhibit negative auto-correlation at all horizons. However, the degree of reversals first increases and then decreases with horizon. This is because shocks to the return process induce moves in $Q_t$ and, consequently, Q-learner demand, that are persistent in the short run, but gradually unwind. We show that the learning rate and the mass of Q-learners differentially affect the strength and persistence of these return dynamics.

**Risk sharing and investor utility.** A common justification for regulating algorithmic trading is that it leads to heightened volatility and possible fragility and destabilization in markets. For instance, the Financial Policy Committee of the Bank of England emphasizes that[7]

> Advanced AI models could rationally exploit profit-making opportunities in a destabilising way or engage in other adverse behaviours. Under a scenario of advanced AI trading models being deployed to act with more autonomy, these models might identify and exploit weaknesses in the trading strategies of other firms in a way that triggers or amplifies price movements.

Similarly, the OECD report on "Regulatory approaches to Artificial Intelligence in Finance" finds that "Herding and Volatility," "Shock Amplification," and "Destabilising Events" were

[7]See Bank of England, Financial Policy Committee, April 2025.

among the top five financial stability risk areas in relation to the use of AI in finance identified by a survey of OECD countries.

To evaluate such concerns, we extend our main model to incorporate hedging needs for the rational investors. Specifically, we assume that rational investors are exposed to background risks that are correlated with the dividend process, and so have an intrinsic motive to trade. A fraction of the rational investors have an incentive to be long the risky asset, while the rest have an incentive to be short. As in our main model, trading by Q-learners leads to volatility amplification and non-trivial return dynamics. Yet, we show that rational investors can often *benefit* from the presence of Q-learners. Perhaps surprisingly, we show that the Q-learners can earn positive profits while improving aggregate investor utility at the same time.

The presence of Q-learners affects investor utility via two channels. First, rational investors are able to exploit the Q-learners' lack of sophistication to extract trading gains. This is because rational investors choose optimal consumption and portfolio allocations taking into account the endogenous demand dynamics induced by Q-learners, while Q-learners adjust their demand based on realized returns without internalizing equilibrium consequences.

Second, the Q-learners "learn" to provide liquidity to one side of the market or the other. For instance, suppose the net demand from rational investors is to short the risky security (e.g., if the hedging demand from the short side of the market exceeds the hedging demand from long investors). This excess demand pushes the security price down, and as a result, Q-learners learn that trading along long investors is profitable. This makes long investors worse off compared to there being no Q-learners, since they now compete with these Q-learners. Short investors, however, profit from the presence of Q-learners, as the Q-learners' trading leads to more favorable prices for them. On average, investor utility increases in the presence of Q-learners and Q-learners realize positive profits. This liquidity provision role of Q-learners is consistent with the evidence documented by Hendershott, Jones, and Menkveld (2011), Brogaard et al. (2014) and Boehmer, Fong, and Wu (2021) who show that algorithmic traders improve liquidity, especially for large stocks.

Our setting delivers clean comparative statics on the Q-learners' profits and the rational investors' utilities. When the magnitude or asymmetry in hedging needs increases, the Q-learners' profits increase, while rational investor are worse off. On the other hand, holding all else fixed, average investor utility tends to increase as the mass of Q-learners increases because of both more trading profits and higher liquidity provision. Finally, we show that an increase in fundamental volatility can lead to **both** higher Q-learner profits and an increase in average investor utility. Together, these findings suggest that blanket regulatory efforts to limit algorithmic trading may be misguided. Even if Q-learners increase volatility or amplify

fundamental shocks, their presence can improve allocative efficiency and investor welfare.

**Overview.** The rest of the paper is as follows. The next section provides a brief discussion of the related literature. Section 3 introduces the model and provides a discussion of the key assumptions. Section 4 provides the main analysis of the paper, by characterizing the equilibrium and describing its properties. Section 5 discusses the model's implications for volatility, liquidity and expected returns. Section 6 presents an extension of the model that incorporates risk-sharing motives for trading. Section 7 concludes. Unless mentioned otherwise, all proofs and additional analysis are in Appendix A and B, respectively.

# 2 Related literature

Our paper is most closely related to the small but growing literature that focuses on the impact of Q-learners in financial markets.[8] In a representative investor setting, Barberis and Jin (2023) show how a Q-learning algorithm can give rise to commonly observed patterns in investor behavior, including extrapolative demand, experience effects and the inertia in portfolio allocations. Colliard, Foucault, and Lovo (2022) show that a Q-learning market maker in a Glosten and Milgrom (1985)-setting adapts to adverse selection due to informed trading, but may charge a markup over the competitive price. They show that noisier environments slow down learning by algorithmic traders thereby leading to less competition and higher spreads. Guarino, Jehiel, and Symons-Hicks (2025) study a similar setting with market makers who use Q-learning. They find that prices vary depending on the setup of the algorithm, and range between competitive and loss-free prices which do not allow trade. Dou, Goldstein, and Ji (2023) introduce Q-learning agents in a multi-trader Kyle (1985)-setting and show how such agents can learn to use collusive trading strategies autonomously, by choosing to dampen the sensitivity of their trades to their private information.[9] In related work, Gufler, Sangiorgi, and Tarantino (2025) study how deep reinforcement learning algorithms operate in a financial market that is calibrated to empirically observed levels of return predictability and price impact. Using simulations, they show that algorithmic traders qualitatively match the behavior of rational investors (who have full knowledge of

[8] There is an earlier literature in finance that uses Q-learning algorithms to numerically solve for equilibria where analytical solutions are infeasible e.g., Goettler, Parlour, and Rajan (2005, 2009).

[9] The latter paper is more broadly related to models of algorithmic collusion in IO settings, including models that focus on Q-learning (e.g., Calvano, Calzolari, Denicoló, and Pastorello (2021), Calvano, Calzolari, Denicolò, and Pastorello (2020), Klein (2021), Banchio and Skrzypacz (2022), Johnson, Rhodes, and Wildenbeest (2023), Xu, Zhang, and Zhao (2024)), richer reinforcement learning algorithms (e.g., Asker, Fershtman, and Pakes (2022) and Cho and Williams (2024)) and large language models (e.g., Fish, Gonczarowski, and Shorrer (2024)).

the economy), but fall short quantitatively — the presence of other algorithmic traders slows down learning and leads to lower profits, lower liquidity and less market efficiency.

Relative to this literature, our work differs in two important ways. First, most of the existing analysis is based on simulations of economies with algorithmic agents. In contrast, we provide an explicit characterization of the financial market equilibrium with Q-learning traders, by establishing convergence properties of the Q-learning algorithm in our setting. We view our approach as complementary, as it makes modeling the market participation by Q-learners amenable to conventional asset pricing tools. This allows us to transparently describe the key pricing dynamics that result from trading by Q-learners along the entire equilibrium path, and compare our results to the existing literature.

Second, while existing work has largely focused on the interaction among multiple reinforcement learning agents, our analysis focuses on the interaction between Q-learners and rational investors. As such, our analysis provides a useful benchmark model of human-AI interaction in financial markets. Specifically, our model provides a basic framework to evaluate the impact of algorithmic trading on investor utility and the efficacy of regulatory policy. For instance, we show that the presence of Q-learners in the market can improve average investor utility, even though it leads to higher return volatility and induces non-fundamental dynamics in prices.

Tractably characterizing models with both rational agents and Q-learners is difficult, as rational agents need to predict the Q-learners' behavior, and, therefore, the evolution of Q-values. We render our setting tractable by (i) assuming that Q-learners use counterfactual updating[10] and (ii) by taking a continuous-time limit. The latter allows us to characterize the entire equilibrium dynamics via a system of ODEs. That is, we can characterize prices, demands, and trader utilities for any arbitrary history of dividends.

Banchio and Mantegazza (2023) use similar a continuous-time limit to study the emergence of collusive equilibria in a prisoner's dilemma setting. Specifically, they use a "fluid approximation" technique which renders the stochastic Q-value process deterministic in the limit, and which can be interpreted as the mean behavior of the Q-value process. In our continuous-time limit, by contrast, the Q-values follow a stochastic differential equation and are adapted to the filtration generated by prices, since the only relevant shocks are those to dividends. This allows us to capture the inherently stochastic nature of Q-learning while preserving tractability, and renders the rational traders' inference problem tractable.[11]

[10] Counterfactual updating is a common technique in reinforcement learning, e.g. see Wachter, Mittelstadt, and Russell (2017) and Foerster, Farquhar, Afouras, Nardelli, and Whiteson (2018). In our setting, it means that whenever a Q-learner buys and receives a certain (realized) return, she understands that selling would have yielded minus that return. With this information, the Q-learner can update the Q-values from both buying and selling simultaneously.

[11] See also Cho (2001) studies asymptotic convergence in an REE setting where agents use least-squares

Dou, Goldstein, Li, and Yang (2025) consider a setting in which better-informed, human traders who exhibit $k$-level thinking compete against an algorithmic trader. They establish conditions under which the algorithmic trader converges to playing the rational expectations strategy, and then characterize equilibrium profits, price volatility and information sensitivity for different levels of investor sophistication and information precision. We view our analysis as complementary. Instead of assuming a behavioral bias, we consider rational investors who have hedging motives for trade — this arguably makes investor welfare analysis more transparent.[12] Moreover, while the algorithmic trader converges to rational expectations in their setting, our analysis suggests that this need not always be the case. In fact, the presence of the Q-learner can endogenously lead to a risk-premium because of the feedback between their trading behavior and prices.

More generally, our paper is related to two other areas in the literature. First, it provides a theoretical complement to the empirical literature that has focused on the impact of algorithmic trading in financial markets (e.g., Brogaard et al. (2014), Chaboud et al. (2014), Weller (2018)). Our model is broadly consistent with the empirical evidence that implies algorithmic traders provide liquidity in markets. Second, it adds to the theoretical literature that focuses on trading settings with non-rational traders. These include models in which rational investors trade against noise traders whose demands are exogenous and uninformed (e.g., Campbell and Kyle (1993), Wang (1993), Wang (1994)) or investors with behavioral biases (e.g., overconfidence, dismissiveness).

We show that trading by Q-learners induces a novel source of state-dependent dynamics, because they systematically and endogenously update their strategies based on past return realizations.[13] As such, our model is closely related to models of extrapolative investors. For instance, Barberis, Greenwood, Jin, and Shleifer (2015) consider a related setting where rational investors trade a risky security with extrapolative investors, who incorrectly believe that the drift in the price process is driven by sentiment, which is a weighted average of past price changes. Because sentiment affects the demand of extrapolative investors linearly, prices exhibit excess but constant volatility and negative correlation which is monotonic in horizon. In our setting, the non-linear dependence of the Q-learners' demand on $Q_t$ implies that prices exhibit stochastic volatility, expected returns are state-dependent and can be non-monotonic in Q-learner demand, and serial correlation in returns exhibits non-monotonicity

learning.

[12] Welfare is difficult to define in settings with heterogeneous beliefs or behavioral biases (e.g., Brunnermeier, Simsek, and Xiong (2014)).

[13] In contrast, traditional models of noise trading assume their security demands are uncorrelated with fundamentals and insensitive to prices, while behavioral investors are usually assumed to maximize subjective utility under biased beliefs or non-standard preferences.

across horizons. As we discuss further in Section 5, these distinctive predictions are consistent with existing empirical evidence.

# 3 Model

In Section 3.1, we present the key assumptions of the model, Section 3.2 provides a discussion of the important assumptions, and Section 3.3 provides intuition for the process we assume for the Q-learners' trading strategy.

## 3.1 Setup

**Payoffs.** There are two securities: a risk free security in perfectly elastic supply with a constant interest rate $r$, and a risky security which pays a dividend stream:

$$dD_t = \mu dt + \sigma dB_t. \tag{1}$$

The supply of the risky asset is $S$ shares per capita, and the price at date $t$ is given by $P_t$. Denote the instantaneous return process by

$$dR_t = D_t dt + dP_t - rP_t dt.$$

**Investors.** There are two types of infinitely lived investors. There is a mass $1-\theta$ of identical, rational traders with discount factor $\delta$ and CARA utility with risk aversion $\phi$. Each investor takes prices as given and chooses consumption $C_t$ and risky security demand $N_t^R$ to maximize

$$\max_{\left\{C_t, N_t^R\right\}_{t\geq 0}} -\mathbb{E}\left[\frac{1}{\phi}\int_0^\infty e^{-\delta t}e^{-\phi C_t}dt\right]$$

subject to the wealth constraint

$$dW_t = \left(rW_t - C_t + N_t^R\left(D_t - rP_t\right)\right)dt + N_t^R dP_t. \tag{2}$$

There is also mass $\theta$ of Q-learners, with aggregate demand $N_t^Q$ for the risky asset, where[14]

$$N_t^Q = N^Q(Q_t) = \frac{1 - \exp\left(-\frac{1}{\beta}Q_t\right)}{1 + \exp\left(-\frac{1}{\beta}Q_t\right)} \in (-1, 1), \tag{3}$$

and $Q_t \equiv Q_t^B - Q_t^S$, where $Q_t^B$ and $Q_t^S$ denote the time $t$ Q-values from buying and selling, respectively. As we shall argue below, we can focus on the evolution of a single Q-value, $Q_t$, which can be expressed as

$$dQ_t = -\alpha Q_t dt + 2\alpha\, dR_t. \tag{4}$$

Here, the parameter $\alpha \in [0, 1]$ is the learning rate of the Q-learning algorithm and controls how quickly the algorithm updates $Q_t$, and the parameter $\beta \geq 0$ captures how responsive the algorithm's demand is to $Q_t$. Specifically, as $\beta$ increases the Q-learners' demand $N_t^Q$ responds more to changes in $Q_t$.

**Market clearing and equilibrium.** The market clearing condition implies that the aggregate demand for the risky asset equals its supply, i.e.,

$$(1 - \theta) N_t^R + \theta N_t^Q = S.$$

The following defines the notion of equilibrium in our setting.

**Definition 1.** A Rational Expectations Equilibrium with Q-Learning (**Q-REE**) is given by a risk premium $\Pi(Q_t)$, demands for the risky asset $N_t^R = N^R(W_t, P_t, D_t, Q_t)$ (for the human traders) and $N_t^Q = N^Q(Q_t)$ (for the Q-learners), and consumption $C_t = C(W_t, P_t, D_t, Q_t)$, such that

1. Prices: For any $(D_t, Q_t)$ and time $t \geq 0$, the price is given by

$$P_t = P(D_t, Q_t) = \frac{1}{r}\left(D_t + \frac{\mu}{r}\right) + \Pi(Q_t), \tag{5}$$

2. Q-value evolution: Q-values follow Equation (4) given $P(D, Q)$,

3. Utility Maximization: $\left(N^R(.), C(.)\right)$ maximizes a rational trader's expected utility

$$V(W_0, Q_0) = \sup_{\left\{C_t, N_t^R\right\}_{t\geq 0}} -E_t\left[\int_0^\infty e^{-\delta t} e^{-\phi C_t} dt\right] \tag{6}$$

[14] We provide a foundation for this functional form in Section 3.3 below. An alternative interpretation is that there is a single large Q-learner who has a misspecified model of price impact. In this case, the demand $N_t^Q$ is simply the Q-learner's demand instead of the aggregate demand.

subject to the dynamic budget constraint in Equation (2) and the transversality condition $\lim_{\tau\to\infty} E_t\left[V\left(W\left(t+\tau\right), Q\left(t+\tau\right)\right)\right] = 0$,

4. Q-learner demand: $N_t^Q = N^Q\left(Q_t\right)$ is given by the expression in Equation (3),

5. Market Clearing: For any $(D_t, Q_t)$ and $W_t$, $P\left(D, Q\right)$ is such that for all $t \geq 0$,

$$\left(1-\theta\right) N^R\left(W_t, P\left(D_t, Q_t\right), D_t, Q_t\right) + \theta N^Q\left(Q_t\right) = S, \tag{7}$$

6. Rational expectations: Rational traders have correct beliefs about the evolution of $\{D_t, P_t, Q_t\}$.

As we shall establish in Section 4.1, the risk premium component of the price $\Pi$ is constant in the absence of Q-learners, and pinned down by the aggregate supply, $S$, of the asset, the risk aversion of the rational traders $\phi$, and their perceived risk $\sigma$ from holding the asset. In the presence of Q-learners, however, we will show that it depends non-linearly on the evolution of the Q-value process, $Q_t$, which itself endogenously depends on the risk premium component $\Pi\left(Q\right)$. Moreover, in equilibrium, while Q-learners are not optimizing agents, the rational traders have correct beliefs about the joint distribution of $\{D_t, P_t, Q_t\}$ and optimize their trading behavior accordingly.

## 3.2 Discussion of assumptions

Our model is stylized for analytical tractability and ease of exposition. Below we discuss the relevance of some of the important simplifying assumptions.

**Hyperparameter knowledge.** We assume for simplicity that the Q-learning algorithm's hyperparameters $\alpha$ and $\beta$ as well as the starting Q-value $Q_0$ are common knowledge. In Section 4.3, we prove that this common knowledge assumption is without loss of generality. That is, if rational traders do not know $(\alpha, \beta, Q_0)$, they can infer them from observing dividends, prices, and their own demands for an arbitrarily short amount of time in equilibrium. In this sense, equilibrium prices reveal the hyperparameters (almost) instantaneously.

**Trading motives of rational investors.** We assume that rational investors speculate against Q-learners but have no additional motives. This simplifies the analysis and helps us isolate the impact of Q-learners on price dynamics. However, it implies that rational investors are always better off in the presence of Q-learners, since they are more sophisticated and can make additional trading gains. In Section 6, we extend the model to endow the rational investors with a hedging motive for trade: we assume that they are subject to endowment

shocks that are correlated with the dividend process. In this case, we show that the trading by Q-learners can improve or worsen outcomes for rational investors.

**Sophistication of algorithmic traders.** In practice, investors employing algorithmic trading use proprietary algorithms to guide their trading strategies. Since we do not know specifically which algorithms are used in practice, we assume that algorithmic traders use Q-learning as a simple and transparent stand in. Q-learning is a foundational reinforcement learning algorithm which is known for its tractability and performance.[15] Q-learning has the additional advantage of being analytically tractable: the updating equation resembles a Bellman equation and we can interpret the Q-value as an estimate for the actual value function. More generally, however, we expect that many of our results will be qualitatively similar for different reinforcement learning algorithms. Used in a live (or "online") context, such algorithms rely on backward looking data to update a decision rule at each time period. As a result, different algorithms will give rise to a similar feedback loop as described in the introduction: an increase in prices leads the algorithm to buy more, which can then lead to further price increases. We expect that such trading will amplify volatility and generate predictability in equity returns in a similar manner to Q-learning.[16]

**Q-learner profits.** In equilibrium, Q-learners make profits whenever $|S|$ is sufficiently large. Intuitively, for a large positive net supply, the stock trades at a discount to induce rational investors to buy, and Q-learners will learn that buying is profitable. For small $|S|$, Q-learners make negative profits in equilibrium. We abstract from Q-learners' decision whether to participate in trading for simplicity. In Section 6, we show that endowment shocks to rational traders endogenously generate a non-zero net supply which Q-learners can profitably exploit. In that case, Q-learners learn to provide liquidity to the short side of the market.

**Endogenous Hyperparameters.** Reinforcement learning models such as Q-learning depend on exogenously chosen hyperparameters, i.e., the learning rate $\alpha$ and the temperature parameter $\beta$. They are hence subject to the Lucas critique – in reality, firms who delegate trading to an algorithm choose the hyperparameters, so the optimal hyperparameters change depending on the environment.[17] In Appendix C.2, we endogenize the hyperparameters. There, we assume that a representative trader with CARA utility delegates trading to the Q-learning algorithm and, anticipating the resulting equilibrium, optimally chooses $\alpha$

[15] For textbook treatments of Q-learning, see Wiering and Otterlo (2012), Ch. 1.7.1 or Sutton and Barto (2018), Ch. 6.5. In single-agent dynamic decision problems, Q-learning has been proven to converge to the optimal policy under mild regularity assumptions, see Watkins and Dayan (1992).

[16] For example, a gradient ascent algorithm will increase the probability of buying following positive returns, which in equilibrium further drives up returns. Temporal-Difference (e.g. Sutton and Barto (2018), Ch. 6) or Actor-Critic (e.g. Sutton and Barto (2018), p. 322) algorithms will act similarly.

[17] We are grateful for In-Koo Cho for pointing this out.

and $\beta$ at the outset. The equilibrium takes the same form as in our main model.

### 3.3 Specification of Q-learning demand

In this subsection, we provide an intuitive argument for why the demand from Q-learners takes the form specified in Equations (3)-(4). We begin with a quick overview of Q-learning in discrete time — Sutton and Barto (2018), Ch. 6.5 provides a more detailed discussion.

Suppose $t = 0, 1, 2, \ldots$ and consider the problem of choosing a policy $\{a_t\}_{t\geq 0}$ given a stochastically evolving state $\{s_t\}_{t\geq 0}$, where $a_t \in A$ and $s_t \in S$ for some finite sets $A$ and $S$:

$$V(s_0) = \max_{\{a_t\}_{t\geq 1}} E\left[\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t)\right].$$

Here, $r(s_t, a_t)$ is the current payoff given $(s_t, a_t)$, $\gamma$ is the discount factor and $V(s)$ is the value function. The optimal policy and value function can be characterized using the Q-matrix $Q(s, a)$, which is the expected value of choosing action $a$ at state $s$, assuming optimal play in all future periods, i.e.

$$Q(s, a) = r(s, a) + \gamma E\left[V(s') \,|s, a\right].$$

We then have $V(s) = \max_{a\in A} Q(s, a)$ and we can exploit this fact to write the Q-matrix recursively as

$$Q(s, a) = r(s, a) + \gamma E\left[\max_{a\in A} Q(s', a) \,|s, a\right].$$

A Q-learner has no information about the payoff function $r(s, a)$ or the evolution of the state, but instead estimates the Q-matrix recursively: starting with an initial guess for the Q-matrix $\hat{Q}_0$, she updates the Q-matrix via

$$\hat{Q}_{t+1}(s_t, a_t) \leftarrow (1-\alpha)\hat{Q}_t(s_t, a_t) + \alpha\left(r_t + \gamma \max_{a'\in A} \hat{Q}_t(s_{t+1}, a')\right)$$

whenever action $a_t$ is chosen given state $s_t$. Here, $r_t$ is the realized payoff at time $t$, and $\alpha \in [0, 1]$ is the learning rate of the Q-learning algorithm, which is set exogenously. It governs how "quickly" the Q-matrix is updated from period to period. In each iteration, the Q-learner takes action $a$ with probability $p(s, a)$.[18]

Now consider a discretization of the model in Section 3.1 in which the risky security pays

[18] If a particular value $a \in A$ is not chosen in period $t$ at state $s_t$, then the Q-matrix is not updated for that value, i.e. $\hat{Q}_{t+1}(s, a) = \hat{Q}_t(s, a)$ whenever $(s, a) \neq (s_t, a_t)$.

dividends $D_t$ at time $t$, where

$$D_{t+1} - D_t = \mu + \sigma\varepsilon_{t+1},$$

and $\varepsilon_{t+1} \sim N(0,1)$ is independent over time.

There is a mass $\theta$ of Q-learners indexed by $i$, each with demand $N^Q_{i,t} \in \{-1, 1\}$ and Q-matrix $Q^a_{i,t}$ for $a \in \{B, S\}$. Here, $Q^B_{i,t}$ is Q-learner $i$'s Q-value from buying a share and $Q^S_{i,t}$ is the Q-value from selling a share.[19] We assume that the Q-learners are identical, i.e. they have the same hyperparameters $\alpha$ and $\beta$ and their Q-values are initialized at the same values.

The per period return on the risky security is given by

$$R_{t+1} = P_{t+1} - P_t + D_t - rP_t,$$

and so conditional on buying, each Q-learner updates the Q-value $Q^B_{i,t}$ as follows:

$$Q^B_{i,t+1} = Q^B_{i,t} + \alpha\left(R_{t+1} + \gamma \max_{a\in\{B,S\}} Q^a_{i,t} - Q^B_{i,t}\right), \tag{8}$$

and conditional on selling, she updates $Q^S_{i,t}$ as:

$$Q^S_{i,t+1} = Q^S_{i,t} + \alpha\left(-R_{t+1} + \gamma \max_{a\in\{B,S\}} Q^a_{i,t} - Q^S_{i,t}\right). \tag{9}$$

We make two assumptions on the Q-learning algorithm. First, we assume that the algorithm uses Boltzmann exploration (e.g. Sutton and Barto (2018), p. 37) when choosing demand, i.e.,

$$N^Q_{i,t} = \begin{cases} 1 & \text{with probability } p\left(Q^B_{i,t}, Q^S_{i,t}\right) \\ -1 & \text{with probability } 1 - p\left(Q^B_{i,t}, Q^S_{i,t}\right) \end{cases}$$

where the probability of buying a share $p\left(Q^B_{i,t}, Q^S_{i,t}\right)$ is given by

$$p\left(Q^B_{i,t}, Q^S_{i,t}\right) = \frac{\exp\left(\frac{1}{\beta}Q^B_{i,t}\right)}{\exp\left(\frac{1}{\beta}Q^B_{i,t}\right) + \exp\left(\frac{1}{\beta}Q^S_{i,t}\right)},$$

and where $\beta \geq 0$ is the temperature parameter. Intuitively, a Q-learner is more likely to buy

[19] We assume that the Q-learners do not treat the dividend as a state. This is natural, since in CARA-Brownian models, the dividend is priced in and returns are independent of realized dividends. Hence, treating the dividend as a state will not improve the Q-learners' payoffs. As we show in Proposition 2, the return is indeed independent of the dividend $D_t$ in equilibrium.

(sell) a share of the risky security when $Q^B_{i,t} > Q^S_{i,t}$ ($Q^B_{i,t} < Q^S_{i,t}$, respectively). The temperature parameter $\beta$ controls the tradeoff between exploitation and exploration. Specifically, when $\beta \to 0$, the demand function is "greedy" since $p \to \mathbf{1}\left\{Q^B_{i,t} > Q^S_{i,t}\right\}$. On the other hand, when $\beta \to \infty$, the Q-learner randomizes uniformly between buying and selling since $p \to 1/2$.

Second, we assume that Q-learners engage in counterfactual learning. If a Q-learner buys a share today and gets a payoff $R_{t+1}$, she not only updates $Q^B_{i,t}$ as in Equation (8), but also recognizes that selling would have yielded $-R_{t+1}$ and so simultaneously updates $Q^S_{i,t}$ as in Equation (9). This allows us to gain tractability when solving the model.[20] We can then summarize each Q-learner's behavior via a single Q-value $Q_{i,t} \equiv Q^B_{i,t} - Q^S_{i,t}$, and characterize the evolution of $Q_{i,t}$ as:

$$\begin{aligned} Q_{i,t+1} &= Q_{i,t} + \alpha\left(R_{t+1} + \gamma \max_{a\in\{B,S\}} Q^a_{i,t} - Q^B_{i,t}\right) - \alpha\left(-R_{t+1} + \gamma \max_{a\in\{B,S\}} Q^a_{i,t} - Q^S_{i,t}\right) \\ &= Q_{i,t}\left(1-\alpha\right) + 2\alpha R_{t+1}. \end{aligned}$$

Since each Q-learner faces the same returns $R_{t+1}$ and updates her Q-matrix the same way, their Q-values are identical, i.e., $Q_{i,t} = Q_t$ for all $t$, and $Q_t$ follows[21]

$$Q_{t+1} - Q_t = 2\alpha R_{t+1} - \alpha Q_t. \tag{10}$$

We will henceforth refer to $Q_t$ simply as *the Q-value*. With slight abuse of notation, we can express the probability of buying a share as:[22]

$$p\left(Q_t\right) = \frac{\exp\left(\frac{1}{\beta}Q_t\right)}{1+\exp\left(\frac{1}{\beta}Q_t\right)}.$$

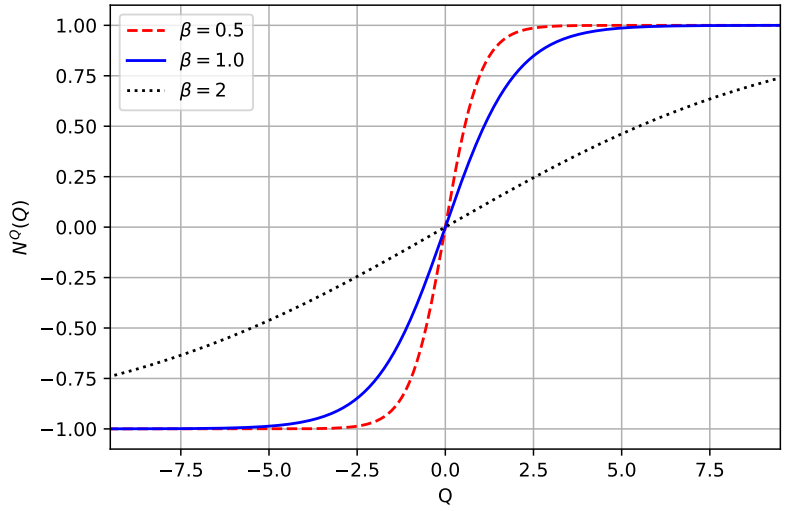
[20]For seminal work on the use of counterfactuals in reinforcement learning see Wachter et al. (2017) and Foerster et al. (2018). Our notion of counterfactuals is particularly simple. We only require that the algorithm recognizes that in each period buying and selling yield opposite returns, and that this information is incorporated into Q-values no matter if the algorithm currently buys or sells. In this sense, we assume that the Q-learner is a "price taker."

[21]Note that Equation (10) is equivalent to the law of motion for $Q^B_t - Q^S_t$ under an expected SARSA algorithm (e.g. Sutton and Barto (2018), Ch. 6.6). Hence, we can alternatively interpret our setting as the algorithmic traders using expected SARSA instead of Q-learning with counterfactual updating.

[22]Since the Q-learning algorithm uses counterfactual updating, the evolution equation (10) is unchanged if we change the algorithm's decision rule, e.g. if we use an $\varepsilon$-greedy rule (e.g. Sutton and Barto (2018), p. 27f) instead of Boltzmann exploration.

Figure 1: Q-learner demand $N^Q(Q_t)$

The figure plots $N^Q(Q)$ for different values of $\beta$: $\beta = 0.5$ (dashed), $\beta = 1$ (solid) and $\beta = 2$ (dotted).




Then, we can express the Q-learners' aggregate demand as

$$N_t^Q = \frac{1}{\theta}\int N_{i,t}^Q di = E[N_{i,t}^Q | Q_t] = p(Q_t) - (1 - p(Q_t)),$$

where we normalize by $\theta$ for convenience. Intuitively, because we have a continuum of Q-learners with the same Q-values, the aggregate demand equals the expected demand for each individual Q-learner.

Using the definition of $p(Q_t)$, we can then write the Q-learners' aggregate demand as

$$N_t^Q = N^Q(Q_t) = \frac{1 - \exp\left(-\frac{1}{\beta}Q_t\right)}{1 + \exp\left(-\frac{1}{\beta}Q_t\right)} \in (-1, 1). \tag{11}$$

Figure 1 provides an illustration of the Q-learners' demand as a function of their Q-value, for different levels of $\beta$.

Note that the specification for the Q-learners' demand in Equations (3)-(4) is just the continuous time analogue of the discrete-time specification in Equations (10) and (11). In Appendix (B), we establish this convergence formally under the conjectured equilibrium price function,

$$P(D_t, Q_t) = \frac{1}{r}\left(D_t + \frac{\mu}{r}\right) + \Pi(Q_t).$$

# 4 Analysis

In this section, we characterize the equilibrium in our model. We begin with the special case in which there are no Q-learners in the economy in Section 4.1. This provides a natural and intuitive benchmark, which will allow us to clarify the key distinguishing features that arise once we introduce Q-learners in Section 4.2.

## 4.1 Benchmark: No Q-learners

The following result characterizes the equilibrium when there are no Q-learners in the economy i.e., $\theta = 0$.[23]

**Proposition 1.** *If $\theta = 0$ (i.e. there are no Q-learners), there exists an equilibrium such that:*

*(i) The equilibrium price function is given by*

$$P(D_t) = \frac{1}{r}\left(D_t + \frac{\mu}{r}\right) - \phi\sigma_{P0}^2 S, \tag{12}$$

*(ii) the trader's value function is given by*

$$V(W_t) = -\frac{1}{r\phi}\exp\left(-r\phi W_t - \frac{\delta - r}{r} - \frac{1}{2r}\phi^2\sigma^2 S^2\right),$$

*(iii) the equilibrium consumption and investment in the risky asset are*

$$C_t = rW_t + \frac{\delta - r}{r\phi} + \frac{1}{2r}\phi\sigma^2 S^2 \text{ and } N_t^R = \frac{D_t + \mu/r - rP_t}{r\phi\sigma_{P0}^2},$$

*and where the price volatility is $\sigma_{P0} = \sigma/r$.*

The above equilibrium is intuitive. The equilibrium price reflects the present value of the future stream of dividends, $\frac{1}{r}\left(D_t + \frac{\mu}{r}\right)$, adjusted for a risk premium term $-\phi\sigma_{P0}^2 S$ which accounts for the discount that investors require for holding $S$ shares of the risky asset in equilibrium. Moreover, the assumption of CARA utility implies that the optimal demand, $N_t^R$, has the traditional "mean-variance" form and is independent of the agents wealth. Similarly, the each trader optimally chooses to consume the interest on her wealth, adjusted for her relative impatience (as reflected by the $\frac{\delta - r}{r\phi}$ term).

[23] In Appendix C.1, we consider the alternative benchmark in which Q-learners are replaced with a risk-neutral trader who, just as the Q-learners, is restricted to buying or selling at most one share.

Note that in the absence of Q-learners, the price is (weakly) lower than the present value of future dividends, i.e., the risk premium component is always negative. Moreover, price volatility is constant and expected returns do not exhibit predictability since

$$\sigma_{P0} = \frac{\sigma}{r}, \quad \text{and} \quad \mathbb{E}\left[dR\right] = r\phi\sigma_{P0}^2 S.$$

As we shall see in the next section, in the presence of Q-learners, price volatility is state dependent. This implies that the risk premium is state-dependent and expected returns are predictable.

## 4.2 Equilibrium

We provide a heuristic argument for characterizing the equilibrium — the formal proofs are in the Appendix. Conjecture that the price can be characterized as in Equation (5) and let

$$dP_t \equiv \mu_P\left(Q_t\right) dt + \sigma_P\left(Q_t\right) dB_t. \tag{13}$$

Since the flow reward for the Q-learners is $dR_t = \left(D_t - rP_t\right) dt + dP_t$, this implies we can express the evolution of $Q$ as

$$dQ_t \equiv \mu_Q\left(Q_t\right) dt + \sigma_Q\left(Q_t\right) dB_t, \tag{14}$$

where the price conjecture above implies:

$$\mu_Q\left(Q_t\right) = 2\alpha\left(\mu_P\left(Q_t\right) - \frac{\mu}{r} - r\Pi\left(Q_t\right) - \frac{1}{2}Q_t\right), \text{ and } \sigma_Q\left(Q_t\right) = 2\alpha\sigma_P\left(Q_t\right). \tag{15}$$

Using Ito's lemma, we can express

$$\mu_P\left(Q_t\right) = \frac{\mu}{r} + \Pi'\left(Q_t\right)\mu_Q(Q_t) + \frac{1}{2}\Pi''\left(Q_t\right)\sigma_Q^2\left(Q_t\right), \text{ and } \sigma_P\left(Q_t\right) = \frac{\sigma}{r} + \Pi'\left(Q_t\right)\sigma_Q\left(Q_t\right). \tag{16}$$

Next, conjecture that the value function for the rational trader is of the form:

$$V\left(W_t, Q_t\right) = -\frac{1}{r\phi}\exp\left(-\phi r W_t - \frac{\delta - r}{r} - G\left(Q_t\right)\right). \tag{17}$$

We refer to $G\left(Q_t\right)$ as the utility gain, because it captures the incremental value that accrues to rational traders because they can predict the demand from Q-learners based on the evolution of $Q_t$.

The HJB equation for the trader implies that

$$\delta V = \max_{C,N} \begin{array}{c} -\frac{1}{\phi}e^{-\phi C} + V_W \times (rW - C + N(D + \mu_P - rP)) + V_Q \mu_Q \\ +\frac{1}{2}\left(V_{WW}\sigma_P^2 N^2 + 2V_{WQ}N\sigma_Q\sigma_P + V_{QQ}\sigma_Q^2\right) \end{array},$$

where we have suppressed the arguments for expositional clarity. Given the conjecture for the value function, one can show that the first order condition for $C_t$ implies

$$C_t = rW_t + \frac{\delta - r}{r\phi} + \frac{1}{\phi}G(Q_t). \tag{19}$$

Since the traders have CARA utility, their optimal demand for the risky asset does not depend on the additional wealth generated by trading against the Q-learners. Instead, this additional value (reflected in $G(Q_t)$) increases consumption, relative to the benchmark without Q-learners.

Similarly, one can show that the first order condition for $N_t$ implies:

$$N_t^R = \underbrace{\frac{D_t + \mu_P(Q_t) - rP_t}{\phi r \sigma_P^2(Q_t)}}_{\text{speculative}} - \underbrace{2\alpha\frac{G'(Q_t)}{\phi r}}_{\text{hedging demand}}. \tag{20}$$

Intuitively, the optimal demand from rational traders has two components. The "speculative" component is of the the standard "mean-variance" form and is analogous to the demand in the benchmark with no Q-learners. The "hedging demand" component reflects how the demand from Q-learners affects the utility gain for rational investors.

One can then calculate the equilibrium price by imposing the market clearing condition, which implies:

$$P_t = \frac{1}{r}\left(D_t + \mu_P(Q_t) - \sigma_P^2(Q_t)\left(\frac{r\phi}{1-\theta}\left(S - \theta N^Q(Q_t)\right) + 2\alpha G'(Q_t)\right)\right). \tag{21}$$

The following result characterizes the existence of a Q-REE in our setting.

**Proposition 2.** *There exists a Q-REE where the rational trader's value function is given by (17), optimal consumption is given by (19), optimal demand is given by (20), and the equilibrium price is given by*

$$P_t = \frac{1}{r}\left(D_t + \frac{\mu}{r}\right) + \Pi(Q_t),$$

*where the risk premium* $\Pi(Q)$ *satisfies the nonlinear*[24] *second-order ODE*

$$r\Pi(Q) = \Pi'(Q)\,\mu_Q(Q) + \frac{1}{2}\Pi''(Q)\,\sigma_Q^2(Q) - G'(Q)\,\sigma_Q(Q)\,\sigma_P(Q) - \frac{r\phi\sigma_P^2(Q)}{1-\theta}\left(S - \theta N^Q(Q)\right) \tag{22}$$

*on* $\mathbb{R}$ *with boundary conditions* $\Pi'(-\infty) = \Pi'(\infty) = 0$*, and for all* $Q$*,* $\Pi'(Q) < \frac{1}{2\alpha}$*. The utility gain* $G(Q)$ *satisfies the nonlinear second-order ODE*

$$rG(Q) = \frac{1}{2}\sigma_P^2(Q)\left(\frac{r\phi}{1-\theta}\left(S - \theta N^Q(Q)\right)\right)^2 + G'(Q)\,\mu_Q(Q) + \left(G''(Q) - G'(Q)^2\right)\frac{1}{2}\sigma_Q^2(Q) \tag{23}$$

*on* $\mathbb{R}$ *with boundary conditions* $G'(-\infty) = G'(\infty) = 0$*. The evolution of* $Q$ *is driven by the drift and diffusion terms:*

$$\mu_Q(Q) = \frac{2\alpha}{1 - 2\alpha\Pi'(Q)}\left(\frac{1}{2}\sigma_Q^2(Q)\,\Pi''(Q) - r\Pi(Q) - \frac{1}{2}Q\right)$$

*and*

$$\sigma_Q(Q) = \frac{\sigma}{r}\frac{2\alpha}{1 - 2\alpha\Pi'(Q)}.$$

The expressions in (22) and (23) can be derived by (i) plugging in the expressions for $\mu_P$ and $\sigma_P$ from (16) into (21), and (ii) plugging in the optimal consumption and demand expressions and the market clearing condition into the trader's HJB equation (18) and simplifying. The key challenge in the proof is to establish existence of the ODE system (22) and (23). One cannot rely on standard Lipschitz conditions, since the volatility $\sigma_Q(Q)$ may potentially explode as $\Pi'(Q) \to 1/2\alpha$. Instead, we use sub- and supersolutions to construct bounded solutions on arbitrary finite domains, and then extend these solutions to infinity via the Arzela-Ascoli theorem. The details are in Appendix A.1.2.

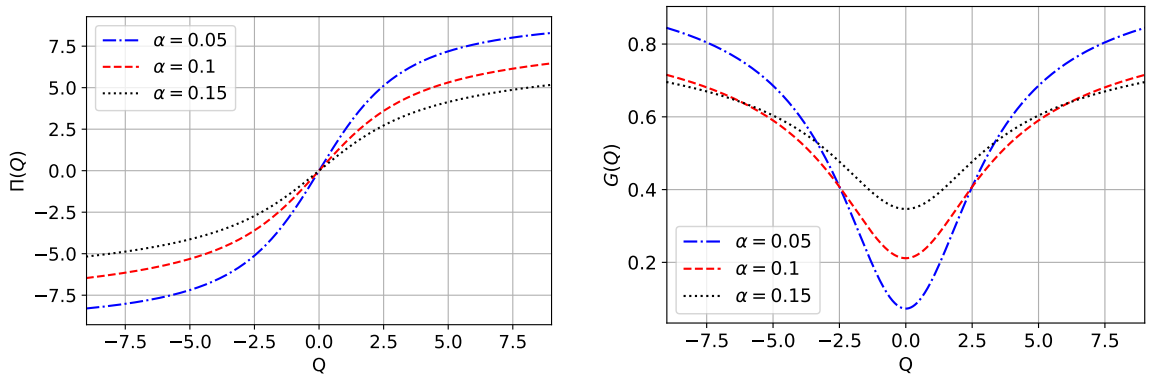
The boundary conditions at infinity are the standard conditions for non-explosion. We can translate these conditions into economically meaningful ones as follows. When $Q = \infty$, the Q-learners always buy, i.e. $N^Q(\infty) = 1$, and when $Q = -\infty$ the Q-learners always sell, i.e. $N^Q(-\infty) = -1$. Then, the risk premium $\Pi$ equals

$$\Pi(-\infty) = -\frac{\sigma^2}{r^2}\frac{\phi}{1-\theta}(S+\theta) \text{ and } \Pi(\infty) = -\frac{\sigma^2}{r^2}\frac{\phi}{1-\theta}(S-\theta). \tag{24}$$

[24]The coefficients $\mu_P(\cdot)$, $\mu_Q(\cdot)$, $\sigma_P(\cdot)$, and $\sigma_Q(\cdot)$ depend on the function $\Pi(\cdot)$, see Equations (16) and (15).

Figure 2: Risk premium $\Pi(Q)$ and utility gains $G(Q)$

The figure plots the risk-premium $\Pi(Q_t)$ and utility gains $G(Q_t)$ for different values of the learning rate $\alpha$. Other parameters are set to: $\phi = 10$, $\sigma = 0.075$, $r = 0.05$, $\beta = 1$, $\mu = 0.05$, $\theta = 0.30$ and $S = 0$.




(a) Risk premium $\Pi(Q)$ (b) Utility gain $G(Q)$

Intuitively, this is the risk premium in the benchmark model without Q-learners, where the asset supply is adjusted for the Q-learners' constant demand (see Equation (12)). Similarly, the gain equals

$$G(-\infty) = \frac{1}{2}\frac{\phi^2\sigma^2}{r}\left(\frac{1}{1-\theta}(S+\theta)\right)^2 \text{ and } G(\infty) = \frac{1}{2}\frac{\phi^2\sigma^2}{r}\left(\frac{1}{1-\theta}(S-\theta)\right)^2. \tag{25}$$

The boundary conditions in Proposition 2 are equivalent to the ones in Equations (24) and (25).

### 4.2.1 The impact of Q-learning on the risk premium and utility gain

Figure 2 provides a numerical illustration of $\Pi(Q)$ and $G(Q)$ for different learning rates $\alpha$. To gain some intuition for the equilibrium, first note that when $\alpha \to 0$, then the demand from the Q-learners is constant at $N(Q_0)$. In this case, the price volatility is again constant and given by $\sigma_P = \sigma/r$ as in the benchmark from Section 4.1, and the risk premium term collapses to

$$\Pi(Q) = -\frac{\phi\sigma_P^2}{1-\theta}\left(S - \theta N^Q(Q)\right).$$

This is the natural analogue to the benchmark model in Section 4.1, after accounting for the fact that now, the Q-learners' demand $N^Q(Q_0)$ shares and the rational traders have a mass of $1-\theta$ in equilibrium. Note that the risk premium $\Pi(Q_0)$ is negative (positive) when the rational traders have to bear a net positive (negative) supply of shares in equilibrium i.e.,

when $N_t^R = \frac{1}{1-\theta}\left(S - \theta N^Q\left(Q_t\right)\right)$ is positive (negative). As a result, the risk premium $\Pi\left(Q\right)$ is increasing in the net Q-value from buying $Q$.

Moreover, note that the utility gain for rational traders is given by

$$G\left(Q\right) = \frac{1}{2}\frac{\sigma_P^2}{r}\left(\frac{r\phi}{1-\theta}\left(S - \theta N^Q\left(Q\right)\right)\right)^2,$$

which reflects the incremental benefit that rational traders gain from the presence of Q-learners. Note that the utility gain is quadratic in $N_t^R$. Intuitively, this reflects the fact that when rational traders are net buyers (i.e., $N_t^R > 0$), the risk-premium implies that the price is lower than expected (discounted) cash flows, while when they are net sellers (i.e., $N_t^R < 0$), the price is higher. As a result, $G\left(Q_0\right)$ is U-shaped in $Q$.

When $\alpha > 0$, the Q-learners update their Q-value $Q_t$ and, consequently, their demand $N^Q\left(Q_t\right)$ for the risky asset, based on realized returns. This induces a feedback channel to prices, analogous to those in models of positive-feedback or extrapolative investors (e.g., De Long, Shleifer, Summers, and Waldmann (1990), Barberis et al. (2015)), where by larger return realizations lead Q-learners to increase their estimate of the net benefit from buying (versus selling), $Q_t$, which leads to higher demand $N^Q\left(Q_t\right)$, which then leads to higher expected returns. Moreover, this feedback effect cannot be too large — a necessary condition for the existence of an equilibrium is that the rate of learning by the Q-learners is bounded above i.e., $\alpha < \frac{1}{2\Pi'(Q_t)}$. To see why, note that we can express price volatility as

$$\sigma_P\left(Q_t\right) = \frac{\sigma}{r} + \frac{\sigma}{r}\frac{2\alpha\Pi'\left(Q_t\right)}{1 - 2\alpha\Pi'\left(Q_t\right)} = \frac{\sigma}{r}\frac{1}{1 - 2\alpha\Pi'\left(Q_t\right)}.$$

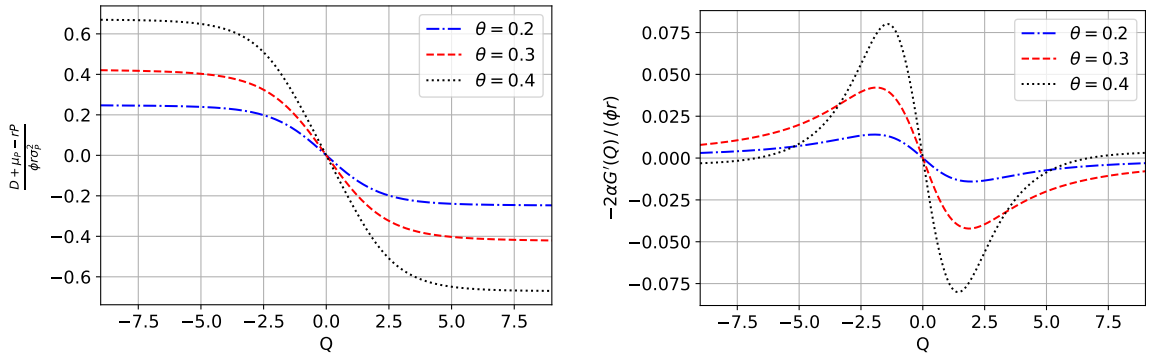
This implies that, as $\alpha \to \frac{1}{2\Pi'(Q_t)}$, $\sigma_P\left(Q_t\right) \to \infty$, and consequently, the rational traders' demand is insensitive to price (see equation (20)). As a result, there is no finite price $P_t$ which clears the market.[25] We establish that $\alpha < \frac{1}{2\Pi'(Q_t)}$ as part of the proof of Proposition 2.

Note that the optimal demand of rational investors depends *non-linearly* on the state-variable $Q_t$. Figure 3 provides an illustration by showing how the speculative and hedging demand components of rational demand, as decomposed in expression (20), vary with $Q_t$. The non-linearity in the speculative, or "mean-variance", component of demand (panel (a)) stems from the fact that the drift $\mu_P$ and volatility $\sigma_P$ of the price process are non-linear in $Q_t$. Similarly, the hedging demand component is non-linear because rational investors anticipate the dynamics of the Q-learners' demand and adjust their demand for the risky

[25] Moreover, when $\alpha > \frac{1}{2\Pi'(Q_t)}$, price volatility is negative.

Figure 3: Rational trader demand components versus $Q_t$

The figure plots the speculative and hedging component of rational trader demand (as in expression (20)) for different values of the mass of Q-learners $\theta$. Other parameters are set to: $\phi = 10$, $\sigma = 0.075$, $r = 0.05$, $\beta = 1$, $\mu = 0.05$, $\alpha = 0.10$ and $S = 0$.




(a) Speculative component

(b) Hedging demand component

asset accordingly. Importantly, this distinguishes our model from most standard models with CARA investors (e.g., Wang (1993), Barberis et al. (2015)), where demand is linear in the relevant state variables. As a result, our model gives rise to stochastic volatility in returns, while existing models exhibit constant or deterministic volatility (as we shall discuss below).

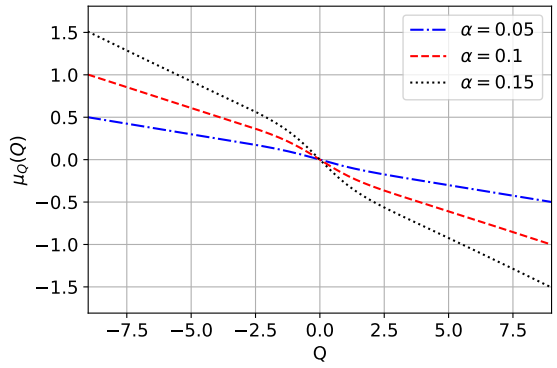
### 4.2.2 The equilibrium evolution of $Q_t$

The evolution of $Q_t$ is characterized by its drift $\mu_Q(Q)$ and volatility $\sigma_Q(Q)$. Figure 4 provides an illustration of how these vary with the rate of learning, $\alpha$, and the mass of Q-learners, $\theta$. As panels (a) and (b) suggest, the drift $\mu_Q(Q)$ is decreasing in $Q$, and is positive when $Q_t$ is negative and vice versa, which implies $Q_t$ is a mean-reverting process. To see why this is intuitive, suppose $Q_t$ is positive and large. This implies Q-learners are more likely to buy shares of the risky security, which pushes up its price. But this makes it less profitable to buy the risky asset going forward, which decreases the net value from buying shares $Q_t$. Moreover, consistent with this intuition, the response is larger when either the rate of learning $\alpha$ is higher or the mass of Q-learners $\theta$ is larger. A higher $\alpha$ implies that Q-learners update $Q_t$ more quickly in response to price changes induced by their past behavior, while a larger $\theta$ implies that changes in $Q_t$ have a larger impact on prices, which in turn, leads to more mean-reversion.

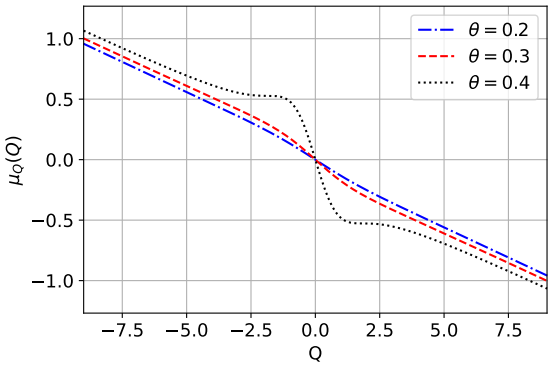
Panels (c) and (d) illustrate that $\sigma_Q(Q)$ is hump-shaped in $Q_t$ and highest when $Q_t = 0$. Intuitively, when $Q_t$ is zero, the Q-learners are indifferent between buying and selling, and so their beliefs are maximally sensitive to realized returns. In contrast, when $Q_t$ is extremely positive (negative), Q-learners perceive the net benefit from buying (selling, respectively)

Figure 4: Drift and volatility of the $Q_t$ process

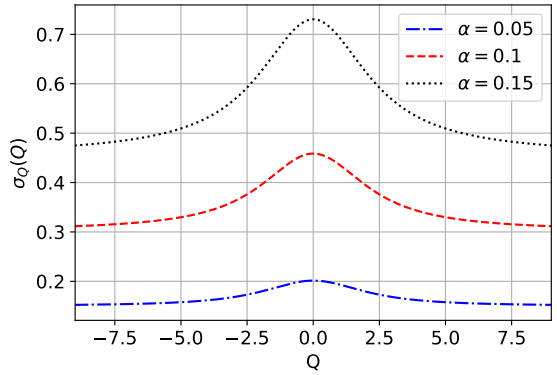
The figure plots the drift $\mu_Q(Q)$ and volatility $\sigma_Q(Q)$ for the $Q_t$ process, for different values of learning rate $\alpha$ and mass of Q-learners, $\theta$. Other parameters are set to: $\phi = 10$, $\sigma = 0.075$, $r = 0.05$, $\beta = 1$, $\mu = 0.05$, $\theta = 0.30$, $\alpha = 0.10$ and $S = 0$.



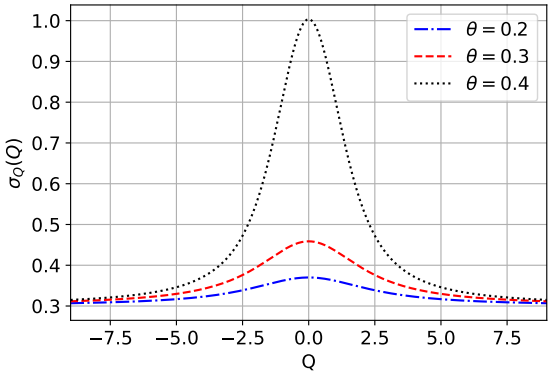

(a) Drift $\mu_Q(Q)$ versus $\alpha$




(b) Drift $\mu_Q(Q)$ versus $\theta$




(c) Volatility $\sigma_Q(Q)$ versus $\alpha$




(d) Volatility $\sigma_Q(Q)$ versus $\theta$

to be extremely high and so $Q_t$ is not very sensitive to realized returns, and consequently, $\sigma_Q(Q)$ is very low. Moreover, all else equal, $\sigma_Q(Q)$ increases as the rate of learning $\alpha$or the mass of Q-learners $\theta$ increases, since either of these changes make $Q_t$ more sensitive to realized returns.

The feedback effect induced by Q-learners implies that their demand for the risky security evolves in a stochastic but persistent manner. As we shall describe in the Section 5, this has novel implications for return dynamics in our setting.

## 4.3 Hyperparameter uncertainty

In baseline model, we assume that rational traders know the hyperparameters of the Q-learning algorithm. Specifically, traders know (1) that the algorithmic traders use Q-learning

with counterfactual updating, (2) the learning rate $\alpha$, (3) the initial Q-value $Q_0$, and (4) the parameter $\beta$ in the specification for demand $N^Q(Q)$.

However, the equilibrium in Proposition 2 survives even if traders do not know the hyperparameters $(\alpha, \beta, Q_0)$.[26] Intuitively, traders can infer all these parameters from observing price volatility, demand volatility, and realized returns over an arbitrarily small horizon, and from calculating excess volatility.

**Proposition 3.** *If traders do not know $(\alpha, \beta, Q_0)$, then the Q-REE of Proposition 2 remains an equilibrium.*

# 5 Return dynamics

The Q-learners' demand depends non-linearly on their net benefit $Q_t$ from buying the security. As the previous section demonstrates, this implies that the presence of Q-learners induces stochastic volatility and predictability in expected returns. We explore how these depend on the parameters of Q-learning in this section. Recall that **dollar** returns evolve according to

$$dR_t = (D_t - rP_t)\,dt + dP_t,$$

where $P_t$ satisfies Equation (5).[27] This implies that returns can be characterized as follows.

**Corollary 1.** *In equilibrium, returns evolve according to:*

$$dR_t = \sigma_P^2(Q_t)\left(\frac{r\phi}{1-\theta}\left(S - \theta N^Q(Q_t)\right) + 2\alpha G'(Q_t)\right)dt + \sigma_P(Q_t)\,dB_t.$$

*where*

$$\sigma_P(Q_t) = \frac{\sigma}{r}\frac{1}{1 - 2\alpha\Pi'(Q_t)}.$$

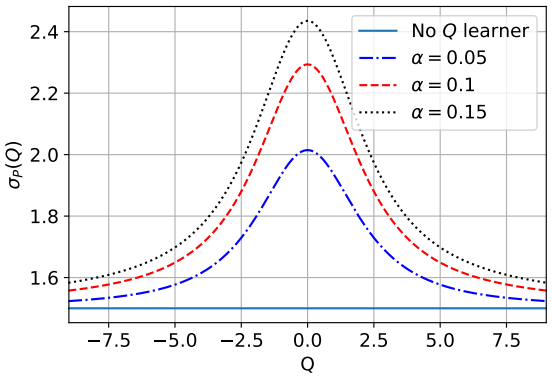
As we discuss below, the above implies that prices exhibit excess, stochastic volatility, expected returns are predictable and can depend non-monotonically in $Q_t$, and serial correlation in return is non-monotonic in horizon.

[26] That is, traders know that there are Q-learners using a Boltzman rule with some parameter $\beta$ and know that the Q-learners use counterfactual updating, but they do not know the actual values $(\alpha, \beta, Q_0)$.

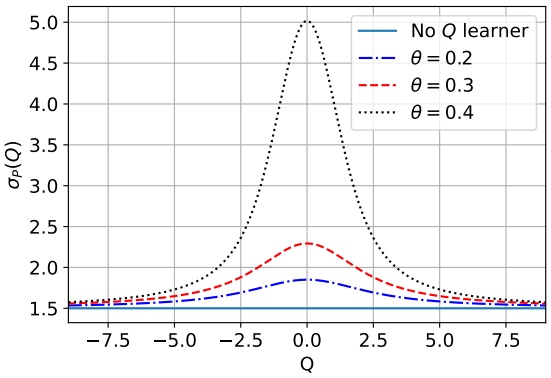
[27] It is worth emphasizing that this is the return that investors receive for holding one share of the risky asset, and therefore distinct from the rate of return one would receive from investing one dollar in the risky asset.

Figure 5: Return volatility with Q-learning

The figure plots return volatility $\sigma_P(Q)$ as a function of $Q$, for different values of learning rate $\alpha$ and mass of Q-learners, $\theta$. Other parameters are set to: $\phi = 10$, $\sigma = 0.075$, $r = 0.05$, $\beta = 1$, $\mu = 0.05$, $\theta = 0.30$, $\alpha = 0.10$ and $S = 0$.




(a) Volatility $\sigma_P(Q)$ versus $\alpha$




(b) Volatility $\sigma_P(Q)$ versus $\theta$

## 5.1 Stochastic volatility

Figure 5 provides an illustration of how stochastic volatility depends on parameters that capture the impact of Q-learners. In the absence of Q-learners, return volatility is constant and given by $\sigma/r$. However, in the presence of Q-learners, return volatility depends on the current state of $Q_t$, and is highest when $Q_t$ is at zero. This captures the fact that, when $Q_t = 0$, the Q-learners are indifferent between buying and selling, which introduces maximal uncertainty and volatility in prices going forward. As a result, Q-learner demand $N_t^Q(\cdot)$ is most sensitive to changes in $Q_t$ at this point. Moreover, the plots illustrate that return volatility is increasing in both the mass of Q-learners, $\theta$, and their learning rate, $\alpha$. These results are also intuitive and follow because trading by Q-learners induces an amplification effect. For instance, a positive innovation in dividends leads to higher prices directly, but also increases the net benefit $Q_t$ of being long. This leads to a higher demand $N(Q_t)$ from Q-learners, which in turn, pushes the price up further. When the mass of Q-learners is larger, the variation in their trading behavior (as driven by the evolution of $Q_t$) has a larger impact on prices, and so increases volatility more. Similarly, when the learning rate $\alpha$ is higher, the net Q-value from buying is more volatile. This leads to more volatility in trading by Q-learners, and consequently, higher return volatility.

The model's prediction of stochastic volatility is consistent with empirical evidence on algorithmic trading. Using the introduction of co-location services as an exogenous instrument for algorithmic trading intensity, Boehmer et al. (2021) show that higher intensity of such trading *causally* increases short-term volatility across a number of empirical proxies

and across 42 equity exchanges in 37 countries. Moreover, the prediction also distinguishes our setting from standard models of CARA investors with noise traders (e.g., Wang (1993)) or behavioral investors (e.g., Barberis et al. (2015)), which exhibit constant or deterministic return volatility.

## 5.2 Expected returns

Figure 6 illustrates the impact of Q-learning on the instantaneous expected return. Recall that in the absence of Q-learners, the instantaneous return is given by $\mathbb{E}[dR] = r\phi\sigma_P^2 S$. For the numerical illustration, we set the aggregate supply of the asset $S$ to zero, which implies that in the absence of Q-learners, the expected return and Sharpe ratio are zero. In contrast, the expected return and Sharpe ratio vary with $Q_t$ in the presence of Q-learners. It is worth noting that expected returns, and consequently, sharpe ratios can be non-monotonic in $Q$. To see why, note that the impact of $Q$ on the expected return can be decomposed as follows:

$$\mathbb{E}[dR_t] = \underbrace{\sigma_P^2(Q_t)}_{\text{scale effect}} \times \left( \underbrace{\frac{r\phi}{1-\theta}\left(S - \theta N^Q(Q_t)\right) + 2\alpha G'(Q_t)}_{\text{level effect}} \right).$$

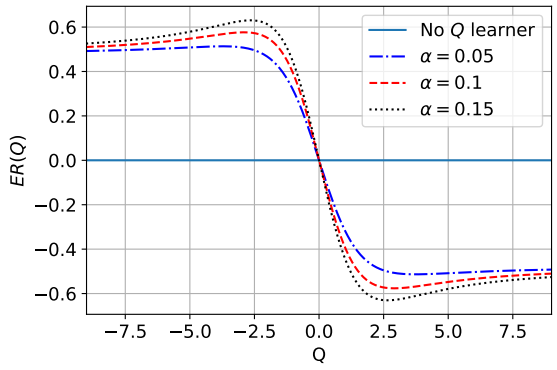
To gain some intuition, note that changes in $Q_t$ affect expected returns through two channels. First, an increase in $Q_t$ implies that Q-learners demand more of the asset. This implies that rational investors have to hold less of the asset in equilibrium, which leads to a decrease in the risk-premium they require, and consequently, lower expected returns. We refer to this as the **level effect**. As panels (c) and (d) of Figure 6 illustrate, the level effect implies that expected returns are positive when $Q_t$ is negative and negative when $Q_t$ is positive. Intuitively, when $Q_t$ is negative, the Q-learners are net sellers (recall $S = 0$) and so rational investors require a positive expected return for being long. In contrast, when $Q_t$ is positive, the Q-learners are net buyers and so rational investors earn a negative expected return since they are short in equilibrium.

Second, as panels (e) and (f) of Figure 6 illustrate, when $Q_t$ becomes closer to zero, stochastic volatility increases. This serves to amplify the impact of the level effect on expected returns by increasing the *per-share* risk of holding the asset — as such, we refer to this as the **scale effect**. This tends to make expected returns more positive (negative) when $Q_t$ is negative (positive) as $Q_t$ gets closer to zero.

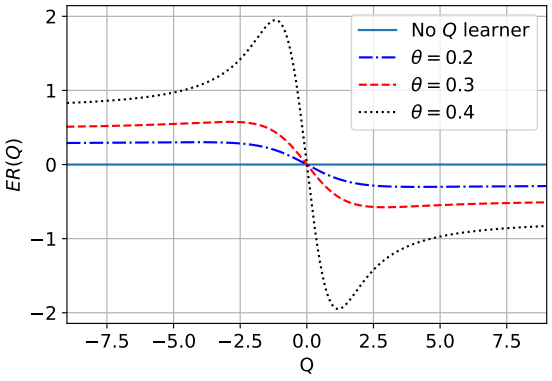
The overall impact of $Q_t$ on expected returns depends on the interaction of the level and scale effects, as panels (a) and (b) of Figure 6 illustrate. When the scale effect is muted (e.g., when the rate of learning $\alpha$ or the mass of Q-learners $\theta$ is low), the level effect dominates

Figure 6: Expected return with Q-learning

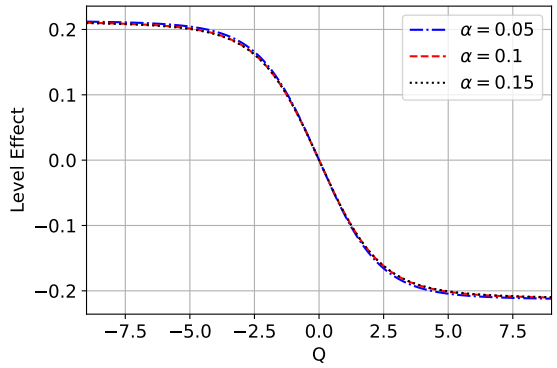
The figure plots the instantaneous expected return $\mathbb{E}[dR_t]$ and the corresponding level and scale effects as a function of $Q$, for different values of learning rate $\alpha$ and mass of Q-learners, $\theta$. Other parameters are set to: $\phi = 10$, $\sigma = 0.075$, $r = 0.05$, $\beta = 1$, $\mu = 0.05$, $\theta = 0.30$, $\alpha = 0.10$ and $S = 0$.



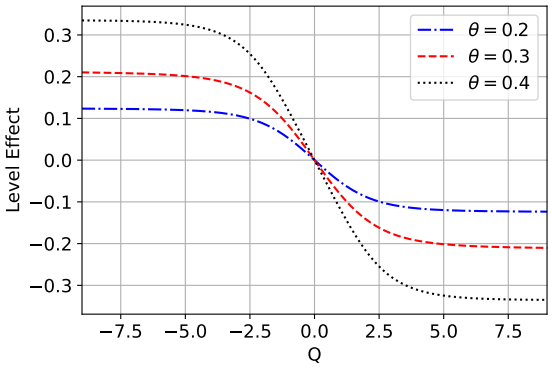

(a) Expected return $\mathbb{E}[dR_t]$ versus $\alpha$



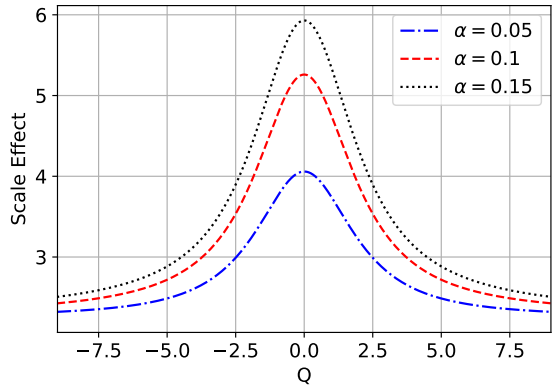

(b) Expected return $\mathbb{E}[dR_t]$ versus $\theta$



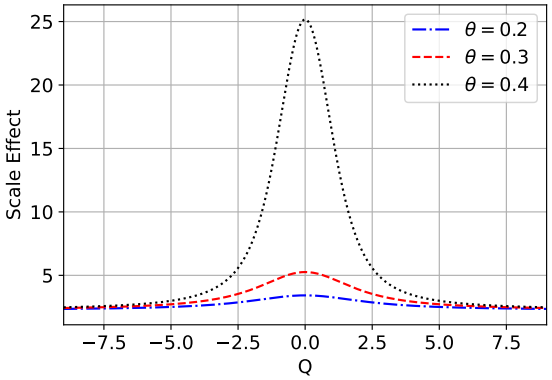

(c) Level effect versus $\alpha$




(d) Level effect versus $\theta$




(e) Scale effect versus $\alpha$




(f) Scale effect versus $\theta$

and to expected returns decrease with $Q_t$. However, when the scale effect is sufficiently large (e.g., for the dotted lines in either panel), the interaction with the scale effect implies that expected returns can be *increasing* in $Q_t$ when $Q_t$ is sufficiently away from zero in either direction. In other words, *higher* demand from Q-learners can lead to *higher* expected returns. This is because variation in demand from Q-learners affects not only the net supply of shares that the rational investors have to hold in equilibrium, but also affects the risk per share (through its effect on stochastic volatility).

This non-monotonicity distinguishes our predictions from traditional models in which variation in demand from non-rational participants (e.g., noise traders) affects the net supply of shares that rational investors have to bear, but does not generate stochastic variation in the volatility of returns (e.g., Wang (1993), Wang (1994), Barberis et al. (2015)). Moreover, it is broadly consistent with the evidence in Brogaard et al. (2014) about the impact of algorithmic trading on future returns. Brogaard et al. (2014) find that liquidity demanding (marketable) trades by high-frequency traders are positively correlated with future returns, while liquidity supplying (non-marketable) trades are negatively correlated with such returns. While our model does not distinguish between marketable and non-marketable orders, one could interpret the Q-learners' demand for extreme $Q_t$ as marketable, since they are less sensitive to $Q_t$, and consequently the price. In contrast, the demand for $Q_t$ near zero are very sensitive to $Q_t$, and thus the price, and could be classified as non-marketable. Under this interpretation, the model predicts that an increase in the marketable demand (driven by an increase in $Q_t$ in the extremes) is associated with an increase in expected returns, while an increase in non-marketable demand (driven by an increase in $Q_t$ near zero) is associated with a decrease in expected returns.

## 5.3 Autocorrelation in returns

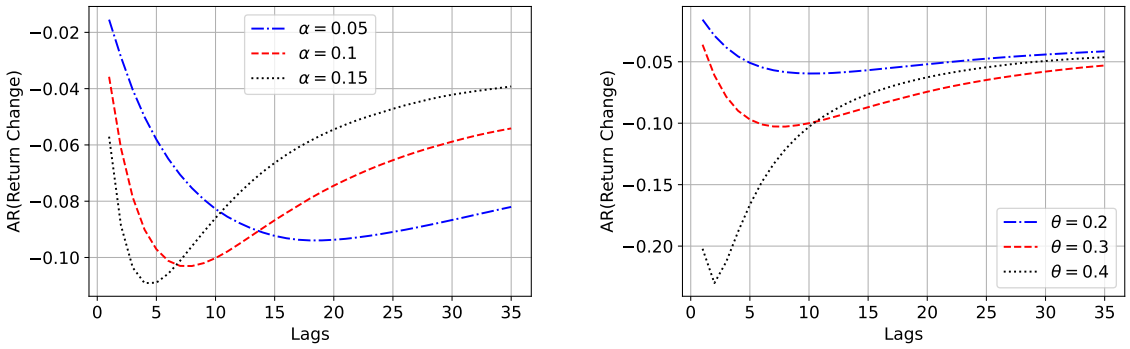
As Figure 7 illustrates, the mean-reverting properties of trading by Q-learners induces serial correlation in returns. Specifically, following Wang (1993), we plot the autocorrelation in returns at lag $\tau$ as:

$$AC(\tau) \equiv \frac{\mathbb{C}\left(R_{t+\tau} - R_t, R_t - R_{t-\tau}\right)}{\sqrt{\mathbb{V}\left(R_{t+\tau} - R_t\right)\mathbb{V}\left(R_t - R_{t-\tau}\right)}}.$$

The feedback into prices induced by Q-learners amplifies the negative correlation in the short-term, but mitigates it in the long-term. To gain some intuition, consider a period in which the dividend shock $D_t$ is higher than expected and there is a contemporaneous increase in the price, $P_t$. This leads to a decrease net benefit from buying, $Q_{t+dt}$, over the next instant, which leads Q-learners to sell and prices to decrease going forward. Since $Q_t$ is a persistent process, Q-learners continue to sell for a period, until the current price decreases

Figure 7: Autocorrelation in returns

The figure plots the autocorrelation in dollar returns $AC(\tau)$ at different lags $\tau$ for different values of learning rate $\alpha$ and mass of Q-learners, $\theta$. Other parameters are set to: $\phi = 10$, $\sigma = 0.075$, $r = 0.05$, $\beta = 1$, $\mu = 0.05$, $\theta = 0.30$, and $S = 0$.




(a) $AC(\tau)$ versus $\alpha$ (b) $AC(\tau)$ versus $\theta$

sufficiently and $Q_t$ starts increasing again.

All else equal, a higher learning rate $\alpha$ implies that Q-learners trading is more sensitive to the realized return process. As panel (a) illustrates, this implies that the short-term serial correlation in returns is more negative, but also that it mean-reverts more quickly. In contrast, a larger mass $\theta$ of Q-learners leads to a larger negative impact in the short-term since the price impact of a unit change in $Q_t$ is larger. However, the autocorrelation at longer lags are approximately the same for different levels of $\theta$.

The patterns in autocorrelation are broadly consistent with empirical evidence about variation in serial correlation across different horizons (e.g., Heston, Korajczyk, and Sadka (2010), Bogousslavsky (2016), Baule, Schlie, and Zhou (2025)). Using high-frequency data on US stocks, Baule et al. (2025) document a term-structure of return autocorrelation that is qualitatively similar to our model predictions. They show that average intraday return autocorrelations are negative across most horizons, are close to zero for sub-minute returns, decline to a minimum for intermediate horizons (e.g., around 15 minutes) and then gradually revert to zero for longer horizons.

Our model generates novel testable predictions that relate the term structure of return autocorrelation to the intensity and speed of algorithmic trading. To the extent that such traders have a larger impact (higher $\theta$) and learn more slowly (lower $\alpha$) for small stocks, our model predictions are consistent with the additional evidence from Baule et al. (2025) that suggests small stocks exhibit a sharper drop and slower reversion in return autocorrelation.

# 6 Risk sharing and investor utility

Policymakers are concerned about the growing importance of algorithmic trading because such behavior may increase price volatility, amplify shocks, and potentially destabilize markets. Our model provides a natural benchmark in which to evaluate such concerns.

As the results from the earlier sections show, the presence of Q-learners leads to higher volatility, amplification of shocks, and non-fundamental dynamics in prices. However, rational investors are always better off (in expectation) in the presence of Q-learners — as panel (b) of Figure 2 illustrates, the additional utility gain (due to the presence of Q-learners) is always positive. This is because rational investors are both better informed about the equilibrium and have no intrinsic need to trade, and so are able to exploit the relative lack of sophistication of Q-learners.

To consider a more realistic benchmark, in this section we extend the model to introduce an additional motive for trade for rational investors. Specifically, we assume that rational traders are exposed to background risks that are correlated with the dividend process of the risky asset. In this case, the presence of Q-learners can have more nuanced effects on the welfare of rational investors.

In equilibrium, Q-learners learn to provide liquidity to the short side of the market, which allows them to realize profits. These profits, however, do not necessarily come at the expense of investors. While investors on the same side as Q-learners lose, since they now have to compete with the Q-learners, investors on the opposite side gain as they benefit from more favorable prices. In aggregate, the presence of Q-learners can improve the average utility across rational investors, even though it leads to higher volatility and decreases utility for some fraction of the investors.

## 6.1 Setup

We consider an extension of our benchmark model in which Q-learners serve the role of liquidity provision. Recall that the mass of rational traders in the economy is $1-\theta$. Further, assume that there are two types of rational traders, indexed by $i \in \{L, S\}$, where fraction $m \in (0,1)$ are of type $i = S$. Suppose that investor $i$ has endowment shocks that are correlated with the dividend process of the risky asset – specifically, suppose

$$dW_{it} = \left(rW_{it} - C_{it} + N_{it}\left(D_t - rP_t\right)\right)dt + N_{it}dP_t + Y_i dB_t,$$

where $Y_S = Y > 0$ and $Y_L = -Y$. Moreover, the market clearing condition is now given by:

$$\theta N_t^Q + (1-\theta)\left(mN_{St} + (1-m)N_{Lt}\right) = S.$$

A non-zero $Y_i$ generates a risk-sharing or liquidity motive for trading.[28] Intuitively, $L$ investors are endowed with an exposure that is negatively correlated with the dividend shocks, and so have a motive to be long the risky asset, while $S$ investors have a positive exposure and so have a motive to short the asset. Note, however, that whether these groups are long or short in equilibrium will depend on the equilibrium price and the demand from Q-learners. Finally, let $\bar{Y} = mY - (1-m)Y$ denote the average exposure to endowment shocks across rational traders.

The following result characterizes the equilibrium in this setting.

**Proposition 4.** *There exists a Q-REE in which the equilibrium price is given by*

$$P_t = \frac{1}{r}\left(D_t + \frac{\mu}{r}\right) + \Pi\left(Q_t\right).$$

*The gains from trade for types $i = L, S$ are given by the ODEs*

$$\begin{array}{rcc} rG_i\left(Q\right) & = & \frac{1}{2}\left(\phi r\right)^2 N_i^2\left(Q\right)\sigma_P^2\left(Q\right) - \frac{1}{2}\left(\phi r Y\right)^2 - G_i'\left(Q\right)\phi r\sigma_Q\left(Q\right)Y_i \\ & & +G_i'\left(Q\right)\mu_Q\left(Q\right) + \left(G_i''\left(Q\right) - G_i'\left(Q\right)^2\right)\frac{1}{2}\sigma_Q^2\left(Q\right) \end{array}$$

*with boundary conditions $G_i'\left(-\infty\right) = G_i'\left(\infty\right) = 0$. The risk premium satisfies the ODE*

$$\begin{array}{rcc} r\Pi\left(Q\right) & = & \Pi'\left(Q\right)\mu_Q\left(Q\right) + \frac{1}{2}\Pi''\left(Q\right)\sigma_Q^2\left(Q\right) \\ & & -\frac{\phi r\sigma_P^2(Q)}{1-\theta}\left(S - \theta N^Q\left(Q\right)\right) - \bar{G}'\left(Q\right)\sigma_P\left(Q\right)\sigma_Q\left(Q\right) - \phi r\bar{Y}\sigma_P\left(Q\right) \end{array}$$

*on $\mathbb{R}$ with boundary conditions $\Pi'\left(-\infty\right) = \Pi'\left(\infty\right) = 0$, and where $\bar{G}'\left(Q\right) = mG_S'\left(Q\right) + (1-m)G_L'\left(Q\right)$. The evolution of $Q$ is driven by the drift and diffusion terms:*

$$\mu_Q\left(Q\right) = \frac{2\alpha}{1 - 2\alpha\Pi'\left(Q\right)}\left(\frac{1}{2}\sigma_Q^2\left(Q\right)\Pi''\left(Q\right) - r\Pi\left(Q\right) - \frac{1}{2}Q\right)$$

*and*

$$\sigma_Q\left(Q\right) = \frac{\sigma}{r}\frac{2\alpha}{1 - 2\alpha\Pi'\left(Q\right)}.$$

The above proposition is analogous to the one for the benchmark analysis, but accounts for the fact that rational investors have endowments which are correlated with the dividend

[28] We do not explicitly model the source of $Y_i$, but it could capture the net risk-exposure from the rest of investor $i$'s (non-tradable) portfolio, including due to shocks to income or liquidity.

process. Specifically, as the proof of the above proposition shows, the optimal demand for a rational investor $i$ can be expressed as:

$$N_i\left(Q_t\right) = \frac{D_t + \mu_P(Q_t) - rP_t}{\phi r \sigma_P^2\left(Q_t\right)} - 2\alpha \frac{G_i'\left(Q_t\right)}{\phi r} - \frac{Y_i}{\sigma_P\left(Q_t\right)}, \tag{28}$$

which extends the optimal demand expression in the main model (see equation (20)) with a term $-\frac{Y_i}{\sigma_P(Q)}$ that captures the impact of the endowment. Specifically, since $Y_L = -Y < 0$ and $Y_S = Y > 0$, $L$ investors buy more of the risky security relative to the no-endowment benchmark, while $S$ investors sell more. Market clearing then implies that the risk premium $\Pi\left(Q\right)$ depends on the average risk exposure $\bar{Y}$ (as captured by $\phi r \bar{Y} \sigma_P\left(Q\right)$ term in equation (27)). Since exposures for $L$ and $S$ investors are symmetric, the risk premium (and consequently, the price of the risky asset) is higher when the mass of $S$ investors is less than half (i.e., $m < 1/2$, and consequently, $\bar{Y} < 0$), and lower when the mass of $S$ investors is more than half.

## 6.2 Intuition

To gain some intuition, consider a case where $m > 1/2$. In this case, the average exposure is $\bar{Y} > 0$, and so the price of the risky asset is lower than under the no-endowment case. In the absence of Q-learners, this makes $L$ traders better off and $S$ traders worse off. A positive average exposure (i.e., $\bar{Y} > 0$) implies that the hedging demand to short the risky security from $S$ traders is larger than the demand from $L$ traders, and the price drop compensates $L$ traders for providing "risk-sharing" or "liquidity" to $S$ traders.

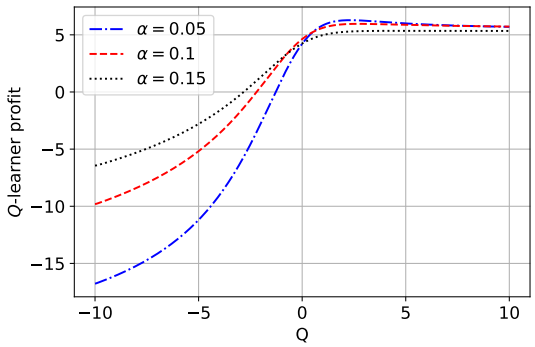
The presence of Q-learners interacts with this risk-sharing. On the one hand, since rational traders are better informed than Q-learners, the presence of the latter group tends to improve utility gains for both $L$ and $S$ traders (as in the no-endowment benchmark). However, the asymmetry in hedging demands across the two groups provides an opportunity for Q-learners to make profits. When the average risk exposure is positive (i.e., $\bar{Y} > 0$), Q-learners can earn positive profits by buying shares at a low price. Intuitively, they learn that, just like $L$ traders, they can earn rents for providing risk-sharing or liquidity to $S$ traders. This improves payoffs for $S$ investors, but reduces payoffs for the $L$ investors.

Figure 8 provides an illustration. Specifically, the figure plots the expected discounted profits $U\left(Q_t\right)$ for a Q-learner, where

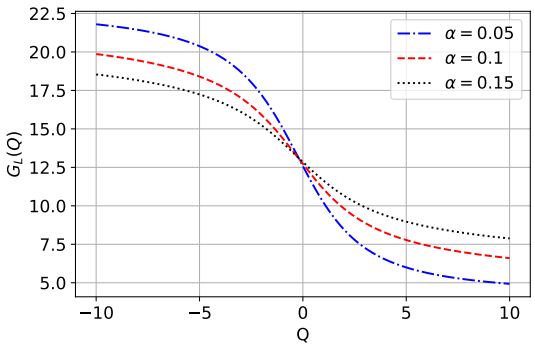
$$U\left(Q_t\right) = E\left[\int_t^{\infty} e^{-r(s-t)} N^Q\left(Q_s\right) dR_s\right],$$

Figure 8: Q-learner profits and rational investor utility gains

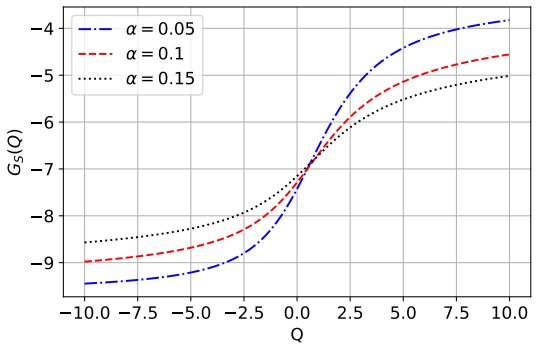
The figure plots the expected discounted profits for a Q-learner, $U(Q_t)$, and the utility gains for $L$ and $S$ investors, as functions of $Q$, for different values of learning rate $\alpha$. Other parameters are set to: $\phi = 10$, $\sigma = 0.075$, $r = 0.05$, $\beta = 1$, $\mu = 0.05$, $\theta = 0.30$, $\alpha = 0.10$, $S = 0$, $m = 0.75$ and $Y = 2$.




(a) Q-learner profits versus $\alpha$




(b) Long utility gain $G_L(Q)$ versus $\alpha$




(c) Short utility gain $G_S(Q)$ versus $\alpha$

and the utility gains $G_S(Q_t)$ and $G_L(Q_t)$ for $S$ and $L$ investors. The illustration sets $m = 0.75$ and $Y = 2$, which implies that the average exposure $\bar{Y} > 0$, and so the price of the risky asset is lower than under the no-endowment case. As a result, as panel (a) illustrates, the expected profit for a Q-learner can be positive when she is sufficiently long (i.e., when $Q_t$ is sufficiently above zero), since she is more likely to buy the security when its price is low. In contrast, when $Q_t$ is negative, the Q-learner is short the security and, consequently, her expected profit is negative. Moreover, the presence of Q-learners affects the utility gains for the rational traders. As $Q_t$ increases, Q-learners are more likely to be long (or less likely to be short). This reduces the utility gains that $L$ investors can generate from providing risk-sharing to $S$ investors, making them worse off (see panel (b)). On the other hand, increased buying demand from Q-learners makes $S$ investors better off (see panel (c)). Essentially,

the Q-learners learn that in equilibrium, it is profitable to "compete" with $L$ investors to provide risk-sharing to $S$ investors.

## 6.3 Comparative statics

Figure 9 characterizes how expected Q-learner profits and rational traders' utility gains depend on model parameters. Specifically, the figure plots expected discounted profits (i.e., $U$), utility gains for $L$ and $S$ traders (i.e., $G_L \equiv \mathbb{E}\left[G_L\left(Q_t\right)\right]$ and $G_S \equiv \mathbb{E}\left[G_S\left(Q_t\right)\right]$, respectively) and the average utility gain (i.e., $\bar{G} = mG_S + (1-m)\, G_L$), where the expectations are computed using the steady state distribution of $Q_t$.[29] Utility gains for rational investors depend on (i) profits generated by trading against Q-learners (as in the main model) and (ii) hedging needs due to risky endowments. As panel (a) illustrates, when $Y$ is small, trading profits dominate and so average $G_L$, $G_S$ and $\bar{G}$ are positive while average $U$ is negative. However, as $Y$ increases, hedging demands begin to dominate. As a result, Q-learner profits are increasing in $Y$ (all else equal) and become positive for sufficiently high $Y$. Moreover, since $m = 0.75 > 1/2$, aggregate hedging demand from $S$ investors are larger, and so $G_S$ decreases with $Y$ while $G_L$ increases with $Y$.

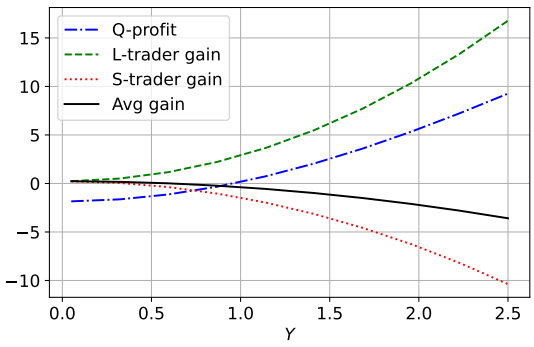
Panel (b) illustrates the impact of varying the mass $m$ of $S$ investors on trading profits and utility gains. When $m$ is low, hedging demand from rational traders is dominated by $L$ investors (i.e., $\bar{Y} < 0$), and consequently, trading gains for $S$ investors and Q-learners is high. Similarly, when $m$ is sufficiently high, hedging demand from $S$ investors dominates and so trading gains for $L$ investors and Q-learners is high. For intermediate $m$, the demand from $S$ and $L$ investors offset each other, so the net demand from rational investors is not very large. In this case, Q-learners do not earn large profits from providing risk-sharing, but instead are a source of trading profits for rational investors. As a result, Q-learner profits are negative, while average utility gains for rational investors are higher than for extreme $m$.

Intuitively, as the mass of Q-learners increases, average utility gains for rational investors increases while trading profits for Q-learners decrease — panel (c) provides an illustration of this. Recall that in this illustration, $m > 1/2$ and rational hedging demand is dominated by $S$ investors. An increase in $\theta$ implies there are more Q-learners to provide risk-sharing, which improves $G_S$ but reduces $G_L$. Moreover, an increase in $\theta$ leads to higher trading profits for rational investors, which tends to increase the average utility gain.

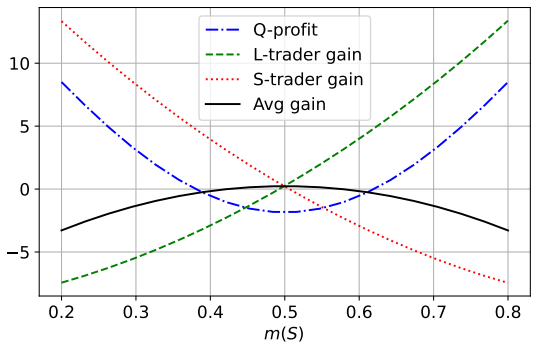
[29] It is worth nothing that $\bar{G}$ does not correspond to the utility gain of rational investors on average because of Jensen's inequality. In Figure 12 of Appendix C, we present the corresponding plots for the expected value function for $L$ and $S$ traders (i.e., $E[V_L(Q_t)]$ and $E[V_S(Q_t)]$) and the weighted average $mE[V_S(Q_t)] + (1-m)\, E[V_L(Q_t)]$, where the expectations are taken under the steady state distribution. While the comparative statics are qualitatively similar, it is clearer to see the impact of parameters when we plot expected utility gains $G(\cdot)$, and so we do so in the main text.

Figure 9: Expected trading profits and utility gains

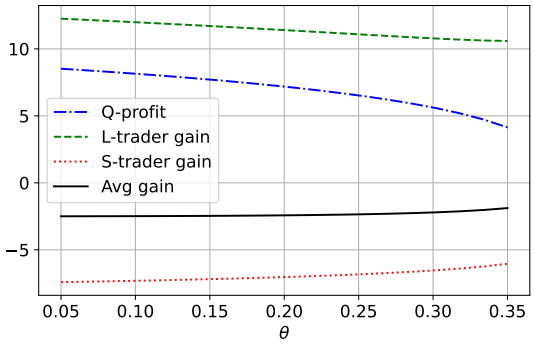
The figure plots the expected discounted profits for a Q-learner, $U(Q_t)$ (dot-dashed), and the utility gains for $L$ and $S$ investors, $G_L$ (dashed) and $G_S$ (dotted), and the average gain $\bar{G} = mG_S + (1-m)G_L$ (solid), versus various model parameters. Unless specified, parameters are set to: $\phi = 10$, $\sigma = 0.075$, $r = 0.05$, $\beta = 1$, $\mu = 0.05$, $\theta = 0.30$, $\alpha = 0.10$, $S = 0$, $m = 0.75$ and $Y = 2$.



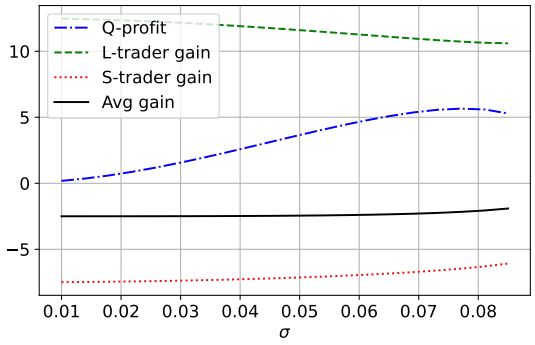

(a) Average Profits / utility gains versus $Y$




(b) Average Profits / utility gains versus $m$




(c) Average Profits / utility gains versus $\theta$

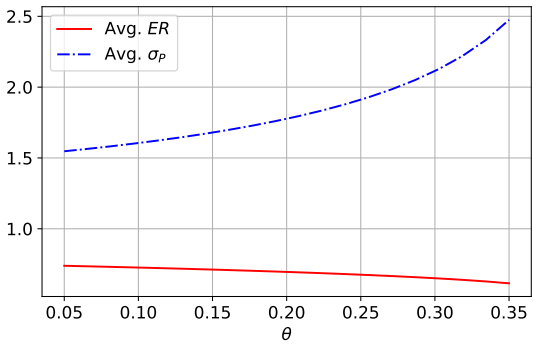



(d) Average Profits / utility gains versus $\sigma$

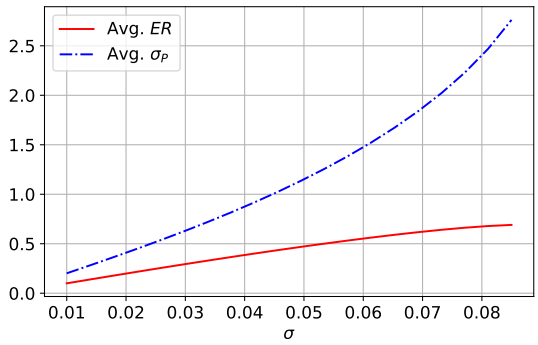
In contrast, panel (d) shows that an increase in fundamental volatility $\sigma$ can have a non-monotonic impact on Q-learner profits. This is because an increase volatility has two effects: (i) it increases the benefit from providing risk-sharing to rational investors, and (ii) it leads to larger expected losses from trading against better informed investors. When $\sigma$ is low, the first effect dominates and Q-learner profits increase with volatility. However, when $\sigma$ is sufficiently large, the second effect dominates and profits can decrease with volatility. Moreover, in this case, average utility gains for rational investors also increases.

Figure 10: Expected returns and average return volatility

The figure plots the expected return and average return volatility versus various model parameters. Unless specified, parameters are set to: $\phi = 10$, $\sigma = 0.075$, $r = 0.05$, $\beta = 1$, $\mu = 0.05$, $\theta = 0.30$, $\alpha = 0.10$, $S = 0$, $m = 0.75$ and $Y = 2$.




(a) Average $ER$ and $\sigma_P$ versus $\theta$




(b) Average $ER$ and $\sigma_P$ versus $\sigma$

## 6.4 Policy implications

A standard argument for regulating algorithmic trading is that they increase volatility and fragility in markets because the feedback induced by their trading behavior. The common wisdom suggests that increased market volatility should reduce welfare for risk averse market participants and that higher trading profits by Q-learners necessarily lead to lower gains for others.

Our results suggests that this intuition may not hold generally. Instead, investors' average utility may improve,[30] because the Q-learners learn to act as liquidity providers. When risk sharing is an important motivation for trading, the above results suggest that (i) Q-learners may be able to make positive trading profits and (ii) their presence may lead to improved average utility of other participants, although it may make some investors worse off. Interestingly, this improvement in average utility gains may happen even though observables like return volatility may suggest otherwise.

For instance, consider the numerical illustrations in Figure 10. For the same parameters considered in Figure 9, panel (a) implies that an increase in the fraction of Q-learners leads to higher return volatility due to the feedback in trading they generate. Similarly, panel (b) suggests that in the presence of Q-learners leads to a sharp amplification effect: return volatility increases non-linearly in cash-flow volatility. However, panels (c) and (d) of Figure 9 imply that average trading gains for rational investors increase with $\theta$ and $\sigma$, respectively,

[30] See specifically Figure 9, panel (c), which shows that average investor gains are increasing in the mass of Q-learners $\theta$.

for these parameter values. Intuitively, since rational investors are risk averse and have hedging needs, while Q-learners are risk-neutral, there are gains from trade. Moreover, since rational investors have an informational advantage relative to Q-learners, an increase in the presence of Q-learners can lead to higher welfare due to more trading profits, despite the increase in return volatility.

# 7 Conclusions

We develop of a model in which rational investors and Q-learners trade a risky security. The trading behavior of the Q-learners is driven by their perceived net benefit from buying a share, $Q_t$, which we show can be approximated in continuous-time by an endogenous SDE. This allows us to analytically characterize equilibrium prices and trading. The Q-learners' trading generates a feedback loop in equilibrium prices, which leads to stochastic volatility, state-dependence in expected returns, and novel patterns in return autocorrelation which depend on the mass and learning rate of Q-learners. Our model also allows us to characterize the impact of Q-learners on investor utility. We show that when risk-sharing is an important trading motive for investors, we show that Q-learners can earn positive profits and improve average investor utility, even though they increase the volatility of prices.

Our model is stylized for analytical tractability and expositional clarity. There are a number of natural directions for future work. First, it would be interesting to explore human-AI interactions in other market environments e.g., in a strategic trading setting like Kyle (1985). Second, we have assumed that the Q-learners do not have an informational advantage over the rational investors. However, in practice, one argument in favor of algorithmic trading is that such approaches allow the use of better information (e.g., in the form of big data). Incorporating both superior information and a lack of sophistication (e.g., about the structure of the economy) in Q-learning traders would be another interesting avenue.

# References


John Asker, Chaim Fershtman, and Ariel Pakes. Artificial intelligence, algorithm design, and pricing. In *AEA Papers and Proceedings*, volume 112, pages 452–456, 2022.

Martino Banchio and Giacomo Mantegazza. Artificial intelligence and spontaneous collusion. *Available at SSRN*, 2023.

Martino Banchio and Andrzej Skrzypacz. Artificial intelligence and auction design. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 30–31, 2022.

Nicholas Barberis, Robin Greenwood, Lawrence Jin, and Andrei Shleifer. X-capm: An extrapolative capital asset pricing model. *Journal of Financial Economics*, 115(1):1–24, 2015.

Nicholas C Barberis and Lawrence J Jin. Model-free and model-based learning as joint drivers of investor behavior. Technical report, National Bureau of Economic Research, 2023.

Rainer Baule, Sebastian Schlie, and Xiaozhou Zhou. The term structure of intraday return autocorrelations. *Available at SSRN 5227714*, 2025.

Ekkehart Boehmer, Kingsley Fong, and Juan Julie Wu. Algorithmic trading and market quality: International evidence. *Journal of Financial and Quantitative Analysis*, 56(8): 2659–2688, 2021.

Vincent Bogousslavsky. Infrequent rebalancing, return autocorrelation, and seasonality. *The Journal of Finance*, 71(6):2967–3006, 2016.

Jonathan Brogaard, Terrence Hendershott, and Ryan Riordan. High-frequency trading and price discovery. *The Review of Financial Studies*, 27(8):2267–2306, 2014.

Markus K Brunnermeier, Alp Simsek, and Wei Xiong. A welfare criterion for models with distorted beliefs. *The Quarterly Journal of Economics*, 129(4):1753–1797, 2014.

Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. Artificial Intelligence, Algorithmic Pricing and Collusion. *American Economic Review*, 110(10): 3267–3297, 2020.

Emilio Calvano, Giacomo Calzolari, Vincenzo Denicoló, and Sergio Pastorello. Algorithmic collusion with imperfect monitoring. *International journal of industrial organization*, 79: 102712, 2021. Publisher: Elsevier.

John Y Campbell and Albert S Kyle. Smart money, noise trading and stock price behaviour. *The Review of Economic Studies*, 60(1):1–34, 1993.

Alain P. Chaboud, Benjamin Chiquoine, Erik Hjalmarsson, and Clara Vega. Rise of the Machines: Algorithmic Trading in the Foreign Exchange Market. *The Journal of Finance*, 69(5):2045–2084, October 2014. ISSN 0022-1082, 1540-6261. doi: 10.1111/jofi.12186.

In-Koo Cho. Convergence of least squares learning in self-referential discontinuous stochastic models. *Journal of Economic Theory*, 101(1):78–114, 2001.

In-Koo Cho and Noah Williams. Collusive Outcomes Without Collusion: Algorithmic Pricing in a Duopoly Model. *Available at SSRN 4753617*, 2024.

Jean-Edouard Colliard, Thierry Foucault, and Stefano Lovo. Algorithmic pricing and liquidity in securities markets. *HEC Paris Research Paper No. FIN-2022-1459*, 2022.

Colette De Coster and Patrick Habets. *Two-Point Boundary Value Problems: Lower and Upper Solutions*, volume 205. Elsevier Science, 2006.

J Bradford De Long, Andrei Shleifer, Lawrence H Summers, and Robert J Waldmann. Positive feedback investment strategies and destabilizing rational speculation. *Journal of Finance*, 45(2):379–395, 1990.

Winston Wei Dou, Itay Goldstein, and Yan Ji. Ai-powered trading, algorithmic collusion, and price efficiency. *Available at SSRN 4452704*, 2023.

Winston Wei Dou, Itay Goldstein, Zigang Li, and Liyan Yang. Human edge, machine limits: Ai-human competition in financial markets. *Machine Limits: AI-Human Competition in Financial Markets (November 18, 2025)*, 2025.

Sara Fish, Yannai A. Gonczarowski, and Ran I. Shorrer. Algorithmic Collusion by Large Language Models, March 2024. arXiv:2404.00806 [cs, econ, q-fin].

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Lawrence R Glosten and Paul R Milgrom. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1): 71–100, 1985.

Ronald L Goettler, Christine A Parlour, and Uday Rajan. Equilibrium in a dynamic limit order market. *The Journal of Finance*, 60(5):2149–2192, 2005.

Ronald L Goettler, Christine A Parlour, and Uday Rajan. Informed traders and limit order markets. *Journal of Financial Economics*, 93(1):67–87, 2009.

Antonio Guarino, Philippe Jehiel, and James Symons-Hicks. Q-learning and algorithmic market making: Loss-free, collusive, or competitive prices? *CEPR Discussion Paper*, 2025.

Ivan Gufler, Francesco Sangiorgi, and Emanuele Tarantino. (deep) learning to trade: An experimental analysis of ai trading and market outcomes. *Available at SSRN 5375160*, 2025.

Philip Hartman. *Ordinary Differential Equations*. Classics in Applied Mathematics. SIAM, 2002.

Terrence Hendershott, Charles M Jones, and Albert J Menkveld. Does algorithmic trading improve liquidity? *The Journal of finance*, 66(1):1–33, 2011.

Steven L Heston, Robert A Korajczyk, and Ronnie Sadka. Intraday patterns in the cross-section of stock returns. *The Journal of Finance*, 65(4):1369–1407, 2010.

Justin P. Johnson, Andrew Rhodes, and Matthijs Wildenbeest. Platform Design When Sellers Use Pricing Algorithms. *Econometrica*, 91(5):1841–1879, 2023. ISSN 0012-9682. doi: 10.3982/ECTA19978.

Timo Klein. Autonomous algorithmic collusion: Q-learning under sequential pricing. *The RAND Journal of Economics*, 52(3):538–558, September 2021. ISSN 0741-6261, 1756-2171. doi: 10.1111/1756-2171.12383.

Thomas G. Kurtz. Equivalence of Stochastic Equations and Martingale Problems. In Dan Crisan, editor, *Stochastic Analysis*, pages 113–130. Springer Verlag, Berlin, Heidelberg, 2010.

Albert S Kyle. Continuous auctions and insider trading. *Econometrica*, pages 1315–1335, 1985.

Daniel W. Stroock and SR Srinivasa Varadhan. *Multidimensional diffusion processes*, volume 233. Springer Science & Business Media, 1997.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Jiang Wang. A model of intertemporal asset prices under asymmetric information. *The Review of Economic Studies*, 60(2):249–282, 1993.

Jiang Wang. A model of competitive stock trading volume. *Journal of Political Economy*, 102(1):127–168, 1994.

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

Brian M Weller. Does algorithmic trading reduce information acquisition? *The Review of Financial Studies*, 31(6):2184–2226, 2018.

Marco Wiering and Martijn Otterlo, editors. *Reinforcement Learning*. Springer Berlin, Heidelberg, 2012. ISBN 978-3-642-27644-6. doi: 10.1007/978-3-642-27645-3.

Zhang Xu, Mingsheng Zhang, and Wei Zhao. Algorithmic collusion and price discrimination: The over-usage of data. *arXiv preprint arXiv:2403.06150*, 2024.

Jiongmin Yong and Xun Yu Zhou. *Stochastic controls: Hamiltonian systems and HJB equations*, volume 43. Springer Verlag, 1999.

# A Proofs

## A.1 Proof of Proposition 2

### A.1.1 Traders' HJB Equation

Given the SDEs (2), (13), and (14), standard verification arguments imply that the solution to the trader's optimization problem $V(W,Q)$ in Equation (6) is characterized by the following Hamilton-Jacobi-Bellman (HJB) equation:

$$\begin{aligned}\delta V &= \max_{C,N} -\frac{1}{\phi}e^{-\phi C} + V_W\left(rW - C + N\left(\mu_P(Q) - rP(D,Q) + D\right)\right)\\ &\quad + \frac{1}{2}V_{WW}\sigma_P^2(Q)N^2\\ &\quad + V_Q\mu_Q(Q) + V_{QQ}\frac{1}{2}\sigma_Q^2(Q)\\ &\quad + V_{WQ}N\sigma_P(Q)\sigma_Q(Q),\end{aligned}$$

where we omit the dependence of $V$ on $(W,Q)$ to keep the notation simple. Conjecture that $V(W,Q) = -\exp\left(-\phi r W - G(Q) - K\right)$, where

$$K = \frac{\delta - r}{r} + \log(\phi r). \tag{30}$$

The first-order condition for $N$ is given by

$$N = -\frac{V_W}{V_{WW}}\frac{\mu_P(Q) - \frac{\mu}{r} - r\Pi(Q)}{\sigma_P^2(Q)} - \frac{V_{WQ}}{V_{WW}}\frac{\sigma_Q(Q)}{\sigma_P(Q)}. \tag{31}$$

The conjecture implies that $V_{WQ} = \phi r G'(Q)V$, $V_W = -\phi r V$, and $V_{WW} = (\phi r)^2 V$, so that we can simplify this expression to

$$N = \frac{1}{\phi r}\frac{\mu_P(Q) - \frac{\mu}{r} - r\Pi(Q)}{\sigma_P^2(Q)} - \frac{G'(Q)}{\phi r}\frac{\sigma_Q(Q)}{\sigma_P(Q)}.$$

Plugging in $\sigma_Q(Q) = 2\alpha\sigma_P(Q)$ yields Equation (20).

The market clearing condition (7) implies that

$$N = \frac{1}{1-\theta}\left(S - \theta N^Q(Q)\right)$$

or equivalently

$$\frac{1}{\phi r}\frac{\mu_P(Q) - \frac{\mu}{r} - r\Pi(Q)}{\sigma_P^2(Q)} - 2\alpha\frac{G'(Q)}{\phi r} = \frac{1}{1-\theta}\left(S - \theta N^Q(Q)\right).$$

After some algebra, this implies

$$\begin{aligned} r\Pi(Q) &= \Pi'(Q)\mu_Q(Q) + \frac{1}{2}\Pi''(Q)\sigma_Q^2(Q) \\ &\quad -G'(Q)\sigma_Q(Q)\sigma_P(Q) - \frac{r\phi\sigma_P^2(Q)}{1-\theta}\left(S - \theta N^Q(Q)\right), \end{aligned}$$

which is Equation (22).

The FOC for consumption is given by

$$C = -\frac{1}{\phi}\log(V_W),$$

which yields Equation (19). To derive $G(Q)$, plug in the FOC for $N$ in Equation (31) and the FOC for $C$ into the HJB Equation (29) and use $V_{QQ} = \left(G'(Q)^2 - G''(Q)\right)V$, which yields

$$\begin{aligned} \delta V &= rV - \phi rV\left(\frac{1}{\phi}\log(\phi r) - \frac{1}{\phi}G(Q) - \frac{1}{\phi}K\right) \\ &\quad +\frac{1}{2}(\phi r)^2 V\sigma_P^2(Q)N^2 \\ &\quad -G'(Q)V\mu_Q(Q) + \left(G''(Q) - G'(Q)^2\right)V\frac{1}{2}\sigma_Q^2(Q). \end{aligned}$$

Canceling $V$ and plugging in the market clearing Condition (7) for $N$ yields

$$\begin{aligned} \delta &= r - r\log(\phi r) + rG(Q) + rK \\ &\quad -\frac{1}{2}(\phi r)^2\sigma_P^2(Q)\left(\frac{1}{1-\theta}\left(S - \theta N^Q(Q)\right)\right)^2 \\ &\quad -G'(Q)\mu_Q(Q) + \left(G'(Q)^2 - G''(Q)\right)\frac{1}{2}\sigma_Q^2(Q). \end{aligned}$$

$$\begin{aligned} rG(Q) &= \frac{1}{2}\sigma_P^2(Q)\left(\frac{r\phi}{1-\theta}\left(S - \theta N^Q(Q)\right)\right)^2 + G'(Q)\mu_Q(Q) \\ &\quad + \left(G''(Q) - G'(Q)^2\right)\frac{1}{2}\sigma_Q^2(Q) \end{aligned}$$

Plugging in the expression for $K$ in Equation (30) and rearranging the expressions involving $G(Q)$ yields Equation (23).

### A.1.2 Existence of Solutions for Equations (22) and (23)

The proof proceeds as follows. We first fix $G'(Q)$ to be an arbitrary continuous function which is bounded on $\mathbb{R}$ and we show that, taking $G'(Q)$ as given, Equation (22) has a twice continuously differentiable and bounded solution on $\mathbb{R}$, which satisfies $\Pi'(Q) \in (-R, 1/(2\alpha))$ for some $R > 0$. Then, we show that for any continuous $\Pi'(Q)$ such that $\Pi'(Q) \in (-R, 1/(2\alpha))$, Equation (23) has a twice continuously differentiable and bounded solution on $\mathbb{R}$.

Taken together, these two results imply that we can define a modification of the ODE system (22) and (23), which is bounded in both the function values $\Pi(Q)$ and $G(Q)$ and the derivatives $\Pi'(Q)$ and $G'(Q)$. Any solution of that ODE system is also a solution of the original system (22) and (23). We then prove that the modified system has a solution on any bounded interval of $Q$, and then use the Arzela-Ascoli theorem to extend the solutions to the entire real line.

**Proposition 5.** *Fix $G'(Q)$ to be an arbitrary continuous function, which satisfies $\sup_{Q\in\mathbb{R}} |G'(Q)| < \infty$ and $\sup_{Q\in\mathbb{R}} |G''(Q)| < \infty$. Then, Equation (22) has at least one solution $\Pi(Q) \in \mathcal{C}^2(\mathbb{R})$.*[31] *Any solution satisfies $\sup_{Q\in\mathbb{R}} |\Pi(Q)| < \infty$ and $\Pi'(Q) \in (-R, 1/(2\alpha))$ for all $Q \in \mathbb{R}$ for some $R > 0$.*

*Proof.* Rewrite Equation (22) as

$$\Pi''(Q) = F_\Pi\left(Q, \Pi(Q), \Pi'(Q), G'(Q)\right) \tag{33}$$

where

$$F_\Pi(Q, u, v, G'(Q)) = (ru + \alpha Q v)\, 2\frac{(1-2\alpha v)^2}{4\alpha^2\left(\frac{\sigma}{r}\right)^2} + \frac{1}{2\alpha^2}(1-2\alpha v)\left(2\alpha G'(Q) + \frac{r\phi}{1-\theta}\left(S - \theta N^Q(Q)\right)\right).$$

Pick $M > 0$ sufficiently large and define for $(Q, u, v) \in \mathbb{R}^3$

$$\bar{F}_\Pi(Q, u, v, G'(Q)) = F_\Pi(Q, u, \max\{\min\{v, M\}, -M\}, G'(Q)).$$

Consider the ODE

$$\bar{\Pi}''(Q) = \bar{F}_\Pi\left(Q, \bar{\Pi}(Q), \bar{\Pi}'(Q), G'(Q)\right). \tag{34}$$

[31] Here, $\mathcal{C}^2(\mathbb{R})$ is the space of twice continuously differentiable functions on $\mathbb{R}$.

We now apply De Coster and Habets (2006), Th. 5.6, p. 122, to Equation (34), which is restated below for the convenience of the reader.

**Theorem 1** (De Coster and Habets (2006), Th. 5.6)**.** *Let $A(Q), B(Q) \in \mathcal{C}^2(\mathbb{R})$ such that $A(Q) \leq B(Q)$ for all $Q \in \mathbb{R}$, $E = \{(Q,u,v) \in \mathbb{R}^3 | A(Q) \leq u \leq B(Q)\}$ and let $f : E \to \mathbb{R}$ be continuous.*

*Assume that $A(Q)$ and $B(Q)$ are such that for all $Q \in \mathbb{R}$,*

$$A''(Q) \geq f(Q, A(Q), A'(Q)) \text{ and } B''(Q) \leq f(Q, B(Q), B'(Q)).$$

*Also, assume that for any bounded interval $I$, there exists a positive continuous function $\varphi_I : \mathbb{R}^+ \to \mathbb{R}$ that satisfies*

$$\int_0^\infty \frac{sds}{\varphi_I(s)} = \infty \tag{35}$$

*and for all $Q \in I$, $(u,v) \in \mathbb{R}^2$ with $A(Q) \leq u \leq B(Q)$,*

$$|f(Q,u,v)| \leq \varphi_I(|v|).$$

*Then, the ODE*

$$u''(Q) = f(Q, u(Q), u'(Q))$$

*has at least one solution $u \in \mathcal{C}^2(\mathbb{R})$ such that for all $Q \in \mathbb{R}$*

$$A(Q) \leq u(Q) \leq B(Q).$$

In particular, we show that there exist numbers $A < B$ such that the constant functions $A(Q) = A$ and $B(Q) = B$ are sub-solutions and super-solutions to Equation (34), respectively, on $\mathbb{R}$ and that $\bar{F}_\Pi(Q,u,v,G'(Q))$ satisfies all conditions of the theorem.

We have

$$F_\Pi(Q,A,0,G'(Q)) = \frac{rA}{2\alpha^2\left(\frac{\sigma}{r}\right)^2} + \frac{1}{2\alpha^2}\left(2\alpha G'(Q) + \frac{r\phi}{1-\theta}\left(S - \theta N^Q(Q)\right)\right).$$

Since $G'(Q)$ and $N^Q(Q)$ are bounded on $\mathbb{R}$, we have $F_\Pi(Q,A,0,G'(Q)) < 0$ all $Q \in \mathbb{R}$ whenever $A$ is sufficiently small. In particular, it holds that $A''(Q) = 0 \geq F_\Pi(Q, A(Q), A'(Q), G'(Q)) = F_\Pi(Q,A,0,G'(Q))$, so that $A(Q) = A$ is indeed a subsolution to Equation (33). Since $A'(Q) = 0 \in (-R, 1/(2\alpha))$, it is also a subsolution to Equation (34). Similarly, for $B$ sufficiently large, $F_\Pi(Q,B,0,G'(Q)) > 0$ for all $Q \in \mathbb{R}$ and by the same argument, $B(Q) = B$ is a supersolution to Equations (33) and (34). Since $F_\Pi(Q,u,0,G'(Q))$ is strictly increasing

in $Q$, and $F_\Pi(Q, A, 0, G'(Q)) < F_\Pi(Q, B, 0, G'(Q))$, it must be the case that $B > A$.

For any bounded interval $I \subset \mathbb{R}$, define $\varphi_I(v) = \max_{(Q,u,v)\in I\times[A,B]\times\mathbb{R}} \left|\bar{F}_\Pi(Q, u, v, G'(Q))\right|$ and note that $\varphi_I(v)$ satisfies the Nagumo condition in Equation (35). By construction of $\varphi_I(v)$, we have $\left|\bar{F}_\Pi(Q, u, v, G'(Q))\right| \leq \varphi_I(|v|)$ for any $(Q, u, v) \in I \times [A, B] \times \mathbb{R}$.

Hence, the theorem implies that Equation (34) has at least one solution $\bar{\Pi}(Q)$, which is twice continuously differentiable on $\mathbb{R}$ and which satisfies $\bar{\Pi}(Q) \in [A, B]$ for all $Q \in \mathbb{R}$. □

We next show that any solution of Equation (34) has a bounded derivative. This implies that for $M$ sufficiently large, any solution to Equation (34) is also a solution to Equation (33).

**Lemma 1.** *Any solution $\bar{\Pi}(Q)$ to Equation (34) satisfies $\bar{\Pi}'(Q) \in (-R, 1/(2\alpha))$ for some $R > 0$ sufficiently large.*

*Proof.* We first show that $\bar{\Pi}'(Q) < 1/(2\alpha)$. Pick $M > 1/(2\alpha)$. By way of contradiction, suppose that there exists a solution $\bar{\Pi}(Q)$ to Equation (34) such that $\bar{\Pi}'(Q) = 1/(2\alpha)$ for some $Q$. Then, since $\bar{\Pi}(Q)$ is bounded, Equation (34) implies that $\bar{\Pi}''(Q) = 0$. Hence, $\bar{\Pi}'(Q') = 1/(2\alpha)$ for $Q' > Q$. But this implies that $\bar{\Pi}(Q) > B$ for $Q$ sufficiently large, which is a contradiction. Similarly, if there exists a $Q$ such that $\bar{\Pi}'(Q) > 1/(2\alpha)$, it must be the case that for some $Q' > Q$, $\bar{\Pi}'(Q') = 1/(2\alpha)$, otherwise $\bar{\Pi}(Q) \to \infty$ as $Q \to \infty$, which contradicts the fact that $\bar{\Pi}(Q) \in [A, B]$. But then, the same contradiction hols at $Q'$. Thus, any solution to Equation (34) must satisfy $\bar{\Pi}'(Q) < 1/(2\alpha)$ for all $Q \in \mathbb{R}$.

Next, we show that $\bar{\Pi}'(Q) > -R$ for some $R > 0$ sufficiently large. Wlog pick $R$ so that

$$(1 + 2\alpha R)(r + \alpha) R \frac{1}{2\alpha^2 \left(\frac{\sigma}{r}\right)^2} > \frac{1}{2\alpha^2}\left(2\alpha G''(Q) + \frac{r\phi}{1-\theta}\left(S - \theta N^{Q\prime}(Q)\right)\right) \tag{36}$$

for all $Q$, which is possible because $G''(Q)$ and $N^{Q\prime}(Q)$ are bounded, and pick $M > R$. Suppose by way of contradiction that for some $Q$, $\bar{\Pi}'(Q) \leq -R$. Then, since $\bar{\Pi}(Q)$ is bounded, it must hold that $\lim_{|Q|\to\infty} \bar{\Pi}'(Q) = 0$ and there exists a $\hat{Q} = \inf\left\{Q : \bar{\Pi}'(Q) \leq -R\right\}$ for which

$$\begin{aligned}\bar{\Pi}''\left(\hat{Q}\right) = F_\Pi\left(\hat{Q}, \bar{\Pi}\left(\hat{Q}\right), R, G'\left(\hat{Q}\right)\right) &= (1 + 2\alpha R)\left(\left(r\bar{\Pi}\left(\hat{Q}\right) - \alpha\hat{Q}R\right)\frac{1}{2\alpha^2\left(\frac{\sigma}{r}\right)^2}(1 + \alpha R)\right. \\ &\quad \left. + \frac{1}{2\alpha^2}\left(2\alpha G'\left(\hat{Q}\right) + \frac{r\phi}{1-\theta}\left(S - \theta N^{Q}\left(\hat{Q}\right)\right)\right)\right) \\ &< 0.\end{aligned}$$

That is, $\hat{Q}$ is the first time $\bar{\Pi}'(Q)$ crosses $-R$ from above. Since we picked $R$ sufficiently large, Equation (36) implies that the RHS is strictly decreasing in $Q$ whenever $\bar{\Pi}'(Q) \leq -R$.

But this implies that $\bar{\Pi}''(Q') < 0$ for all $Q' > Q$ if $\bar{\Pi}'(Q') = -R$. Therefore $\bar{\Pi}'(Q') \leq -R$ for all $Q' > Q$, which contradicts the fact that $\bar{\Pi}'(\infty) = 0$. Hence, $\bar{\Pi}'(Q) > -R$ for all $Q$. □

Lemma 1 immediately implies the following.

**Corollary 2.** *For $R$ sufficiently large, any solution of Equation (34) also satisfies Equation (33).*

The corollary implies that Equation (33) has at least one solution $\Pi(Q)$ such that $\Pi(Q) \in [A, B]$ and $\Pi'(Q) \in (-R, 1/(2\alpha))$, which is what we set out to prove.

Next, we show that for any function $\Pi(Q)$ such that $\Pi'(Q)$ is continuous and $\Pi'(Q) \in (-R, 1/(2\alpha))$ for all $Q \in \mathbb{R}$, Equation (23) has a solution, such that $G'(Q)$ is continuous and bounded on $\mathbb{R}$.

**Proposition 6.** *Fix $\Pi'(Q)$ to be an arbitrary continuous function such that for some $R > 0$, $\Pi'(Q) \in (-R, 1/(2\alpha))$ for all $Q \in \mathbb{R}$. Then Equation (23) has a bounded solution $G(Q) \in \mathcal{C}^2(\mathbb{R})$ and $G'(Q)$ is bounded on any finite interval $I \subset \mathbb{R}$.*

*Proof.* For a given $\Pi'(Q)$, define

$$\sigma_P(Q) = \frac{\sigma}{r} \frac{1}{1 - 2\alpha\Pi'(Q)}$$

and notice that $\sigma_P(Q)$ continuous and uniformly bounded on $\mathbb{R}$,[32] and satisfies $\inf_{Q \in \mathbb{R}} \sigma_P(Q) \geq \underline{\sigma} > 0$ for some $\underline{\sigma} > 0$.[33] Proceeding similarly as for Equation (22), write Equation (23) as

$$G''(Q) = F_G(Q, G(Q), G'(Q), \sigma_P(Q)) \tag{37}$$

where

$$\begin{aligned} F_G(Q, u, v, \sigma_P(Q)) &= (ru + \alpha Q v) \frac{1}{2\alpha^2 \sigma_P^2(Q)} \\ &\quad - \frac{1}{4\alpha^2} \left( \frac{r\phi}{1-\theta} \left( S - \theta N^Q(Q) \right) + 2\alpha v \right)^2. \end{aligned}$$

We now apply the same theorem as above, De Coster and Habets (2006), Th. 5.7, p. 122, to Equation (37). Since $\Pi'(Q)$ and $N^Q(Q)$ are bounded and since $F_G(Q, u, v, \sigma_P(Q))$ is

[32] This follows because for any $Q$, $\Pi'(Q) < 1/(2\alpha)$ and $\Pi'(Q) \to 0$ as $|Q| \to \infty$, so $\Pi'(Q)$ cannot approach $1/(2\alpha)$ as $|Q|$ approaches infinity.

[33] This is because we proved that for all $Q$, $\Pi'(Q) > -R$ for some $R > 0$.

increasing in $u$, there exist $A < B$ such that the constant functions $A(Q)$ and $B(Q)$ are sub-solutions and super-solutions to Equation (37). For any bounded interval $I \subset \mathbb{R}$, pick

$$\varphi_I(v) = K_I\left(1 + v^2\right)$$

for some $K_I > 0$. Then, $\varphi_I(v)$ satisfies Equation (35) and

$$|F_G(Q, u, v, \sigma_P(Q))| \leq \varphi_I(|v|) \tag{38}$$

whenever $K_I$ is sufficiently large. Hence, Equation (37) has at least one solution $G(Q) \in \mathcal{C}^2(\mathbb{R})$ which satisfies $G(Q) \in [A, B]$ for all $Q \in \mathbb{R}$.

It remains to show that any solution of Equation (37) has a derivative $G'(Q)$ which is bounded on finite intervals. This follows from Hartman (2002), Lemma 5.1, p. 428. The lemma states that if $G$ is bounded on $\mathbb{R}$ and $G''(Q) \leq \varphi(|G'(Q)|)$ on any finite interval $I$ of length at least $l_0 > 0$, then $|G'(Q)| \leq M$. Picking $\varphi(v) = \varphi_I(v) = K_I(1 + v^2)$ and using Equation (38) then yields the result. □

Propositions 5 and 6 establish that we can pick constants $A, B, R_P, R_G$ and $R_\alpha$ and wlog consider the system of ODEs

$$\hat{\Pi}''(Q) = \hat{F}_\Pi\left(Q, \hat{\Pi}(Q), \hat{\Pi}'(Q), \hat{G}'(Q)\right) \text{ and } \hat{G}''(Q) = \hat{F}_G\left(Q, \hat{G}(Q), \hat{G}'(Q), \hat{\Pi}'(Q)\right) \tag{39}$$

on finite intervals of $\mathbb{R}$ where

$$\hat{F}_\Pi(Q, u, v, x) = F_\Pi(Q, \min\{A, \max\{B, u\}\}, \min\{-R_P, \max\{v, 1/(2\alpha) - R_\alpha\}\}, \min\{-R_G, \max\{x, R_G\}\}$$

and

$$\hat{F}_G(Q, u, v, x) = F_\Pi(Q, \min\{A, \max\{B, u\}\}, \min\{-R_G, \max\{v, R_G\}\}, \min\{-R_P, \max\{x, 1/(2\alpha) - R_\alpha\}\}$$

**Proposition 7.** *For any finite interval $[\underline{Q}, \bar{Q}] \subset \mathbb{R}$, the BVP (39) with boundary conditions*

$$\hat{\Pi}(\underline{Q}) = \underline{\Pi},\ \hat{\Pi}(\bar{Q}) = \bar{\Pi},\ \hat{G}(\underline{Q}) = \underline{G},\ \text{and}\ \hat{G}(\bar{Q}) = \bar{G}$$

*has a solution.*

*Proof.* The proof is an adaptation of the proof of Hartman (2002), Th. 4.2, p. 424. The proof relies on constructing an operator on the space of continuously differentiable functions on finite intervals and applying Schauder's Theorem (e.g. Hartman (2002), Th. 0.2, p. 405).

We can write the ODE system in vector form as

$$X''(Q) = \hat{F}(Q, X(Q), X'(Q))$$

where $X(Q) = (\Pi(Q), G(Q))$ and $\hat{F} = \left(\hat{F}_{\Pi}, \hat{F}_G\right)$. We now establish that this system has a solution on any finite interval $I \equiv [\underline{Q}, \bar{Q}] \subset \mathbb{R}$ with the above boundary conditions.

Let $l_I = \bar{Q} - \underline{Q}$ and denote with $\mathcal{D}_I$ the Banach space of continuously differentiable vector-valued functions $d(Q) : I \to \mathbb{R}^2$. Endow $\mathcal{D}_I$ with the norm

$$|d| = \max\left(\max_{Q \in I} \|d(Q)\|, \frac{l_I}{4} \max_{Q \in I} \|d'(Q)\|\right)$$

where $\|.\|$ is the Euclidean norm. Consider the subset $\mathcal{D}_I^R = \{d(Q) : \max_{Q \in I} |d(Q)| \leq R\}$ for some $R > 0$. Hartman (2002), Th. 3.1, p. 419 implies that the ODE system

$$X''(Q) = \hat{F}(Q, d(Q), d'(Q))$$

with boundary conditions $X(\underline{Q}) = \underline{X}$ and $X(\bar{Q}) = \bar{X}$, where $\underline{X} = (\underline{\Pi}, \underline{G})$ and $\bar{X} = (\bar{\Pi}, \bar{G})$, has a unique solution for any $d \in \mathcal{D}_I^R$, which is given by

$$X(Q) = -\int_{\underline{Q}}^{\bar{Q}} G(Q, s) \hat{F}(s, d(s), d'(s)) ds + \left(Q - \underline{Q}\right) \frac{1}{l_I} \left(\bar{X} - \underline{X}\right) + \underline{X} \tag{40}$$

where

$$G(Q, s) = \frac{1}{l_I} \left(\left(\bar{Q} - Q\right)\left(s - \underline{Q}\right) \mathbf{1}_{\{s \leq Q\}} + \left(\bar{Q} - s\right)\left(Q - \underline{Q}\right) \mathbf{1}_{\{s > Q\}}\right)$$

Now, define the operator $T : \mathcal{D}_I \to \mathcal{D}_I$ such that $T(d(Q)) = X(Q)$. To apply Schauder's fixed point theorem, we must establish that $T$ is a continuous and compact operator which maps $\mathcal{D}_I^R$ onto itself.

To establish that $T$ maps $\mathcal{D}_I^R$ onto itself, note that it holds that

$$\int_{\underline{Q}}^{\bar{Q}} G(Q, s) ds \leq \frac{l_I^2}{8} \text{ and } \int_{\underline{Q}}^{\bar{Q}} \frac{\partial G(Q, s)}{\partial Q} \leq \frac{l_I}{2}.$$

Therefore,

$$\|X(Q)\| \leq \frac{l_I^2}{8} \max_{Q \in [\underline{Q}, \bar{Q}]} \left\|\hat{F}(Q, d(Q), d'(Q))\right\| + \left\|\bar{X} - \underline{X}\right\|$$

and, by differentiating Equation (40) with respect to $Q$,

$$\|X'(Q)\| \leq \frac{l_I}{2} \max_{Q \in [\underline{Q}, \bar{Q}]} \left\| \hat{F}\left(Q, d(Q), d'(Q)\right) \right\| + \frac{1}{l_I} \left\| \bar{X} - \underline{X} \right\|.$$

By construction of $\hat{F}$, it holds that

$$\max_{Q \in [\underline{Q}, \bar{Q}]} \left\| \hat{F}\left(Q, d(Q), d'(Q)\right) \right\| \leq K_I$$

and therefore

$$\|X(Q)\| \leq \frac{l_I^2}{8} K_I + \left\| \bar{X} - \underline{X} \right\| \text{ and } \|X'(Q)\| \leq \frac{l_I}{2} K_I + \left\| \bar{X} - \underline{X} \right\|$$

so that $|X(Q)| \leq \frac{l_I^2}{8} K_I + \left\| \bar{X} - \underline{X} \right\|$. Picking $R > \frac{l_I^2}{8} K_I + \left\| \bar{X} - \underline{X} \right\|$ ensures that $T$ maps $\mathcal{D}_I^R$ into itself.

To establish continuity, pick $d_1, d_2 \in \mathcal{D}_I^R$ and let $X_1 = T(d_1)$ and $X_2 = T(d_2)$. Then,

$$|X_1 - X_2| \leq \frac{l_I}{4} \int_{\underline{Q}}^{\bar{Q}} \left\| \hat{F}\left(Q, d_1(Q), d_1'(Q)\right) - \hat{F}\left(Q, d_2(Q), d_2'(Q)\right) \right\| dQ,$$

so that $|d_1 - d_2| \to 0$ implies $|X_1 - X_2| \to 0$. Thus, $T$ is continuous.

We finally establish that $T$ is a compact operator, i.e. the range of $T$ has a compact closure. Equation (40) implies that

$$\|X(Q_1) - X(Q_2)\| \leq K_I l_I |Q_1 - Q_2|$$

and

$$\|X'(Q_1) - X'(Q_2)\| \leq 2K_I |Q_1 - Q_2|,$$

which follows after some algebra. Hence, functions in the range of $T$ are equicontinuous and the Arzela-Ascoli theorem implies that the range of $T$ has a compact closure.

We can hence apply Schauder's theorem to $T$, which concludes the proof. □

It remains to extend the existence result in Proposition (7) to the entire real line. To this end, consider a sequence of intervals $I_n = [\underline{Q}_n, \bar{Q}_n]$ for $n \in \mathbb{N}$ with $I_n \subset I_{n+1}$ and $\underline{Q}_n \to -\infty$ and $\bar{Q}_n \to \infty$ as $n \to \infty$. Consider the sequence of BVPs (39) on $I_n$ with boundary conditions $\hat{\Pi}(\underline{Q}) = \underline{\Pi}_n$, $\hat{\Pi}(\bar{Q}) = \bar{\Pi}_n$, $\hat{G}(\underline{Q}) = \underline{G}_n$, and $\hat{G}(\bar{Q}) = \bar{G}_n$, such that as $n \to \infty$,

$$\underline{\Pi}_n \to \Pi(-\infty) = -\frac{\sigma^2}{r^2} \frac{\phi}{1-\theta}(S + \theta) \text{ and } \bar{\Pi}_n \to \Pi(\infty) = -\frac{\sigma^2}{r^2} \frac{\phi}{1-\theta}(S - \theta),$$

and

$$\underline{G}_n \to G(-\infty) = \frac{1}{2}\frac{\phi^2\sigma^2}{r}\left(\frac{1}{1-\theta}(S+\theta)\right)^2 \text{ and } \bar{G}_n \to G(\infty) = \frac{1}{2}\frac{\phi^2\sigma^2}{r}\left(\frac{1}{1-\theta}(S-\theta)\right)^2.$$

On any interval $I_n$, Proposition (7) implies that there exists a solution $(\Pi_n(Q), G_n(Q))$. We will use the Arzela-Ascoli theorem to construct a solution on $\mathbb{R}$ as a limit of $(\Pi_n(Q), G_n(Q))$. To this end, we use Hartman (2002), Corollary 5.1, p. 431, which we reproduce below.

**Corollary 3.** *[Hartman (2002), Cor. 5.1] Suppose that a (vector-valued) function $X(Q)$ is twice continuously differentiable on the interval $[\underline{Q}, \bar{Q}]$ with $[\underline{Q}, \tilde{Q}] \subset [\underline{Q}, \bar{Q}]$ for some $\tilde{Q}$ and*

$$\|X\| \leq R \text{ and } \|X''\| \leq K_1 + K_2\|X'\|^2 \tag{41}$$

*for some denote strictly positive constants $R$, $K_1$, and $K_2$. Then, there exists a constant $M$ such that $\|X'\| \leq M$ for all $Q \in p = [\underline{Q}, \bar{Q}]$. The constant $M$ depends only on $R$, $K_1$, $K_2$, and $\tilde{Q}$.*

From our construction of $\hat{F}$ in Equation (39), it follows that the conditions in Equation (41) hold on any finite interval $[\underline{Q}, \bar{Q}]$. Fix the interval $[\underline{Q}_1, \bar{Q}_1]$. Propositions 5 and 6 imply that any solution $(\Pi_n(Q), G_n(Q))$ satisfies $\Pi_n(Q), G_n(Q) \in [A, B]$ for $Q \in [\underline{Q}_1, \bar{Q}_1]$, and Corollary 3 implies that $|\Pi_n'(Q)|, |G_n'(Q)| \leq M$ for $Q \in [\underline{Q}_1, \bar{Q}_1]$. Notably, $M$ depends on $[\underline{Q}_1, \bar{Q}_1]$ but not on $n$. Then, given $\Pi_n(Q), G_n(Q), \Pi_n'(Q)$, and $G_n'(Q)$ are bounded on $[\underline{Q}_1, \bar{Q}_1]$, Equation (39) implies that $\Pi_n''(Q)$ and $G_n''(Q)$ are bounded on $[\underline{Q}_1, \bar{Q}_1]$ as well. Since the bounds do not depend on $n$, the sequence of functions $(\Pi_n(Q), G_n(Q))$ is bounded (uniformly in $n$) and equicontinuous. Hence, the Arzela-Ascoli theorem implies that there is a subsequence which converges uniformly to a continuously differentiable function $\left(\tilde{\Pi}_1, \tilde{G}_1\right)$ on $[\underline{Q}_1, \bar{Q}_1]$. Since the second derivatives are also bounded on $[\underline{Q}_1, \bar{Q}_1]$, uniformly in $n$, $(\Pi_n'(Q), G_n'(Q))$ are also equicontinuous, so that $\left(\tilde{\Pi}, \tilde{G}\right)$ is twice continuously differentiable wlog and satisfies Equation (39).

Now, pick $[\underline{Q}_2, \bar{Q}_2]$. Repeating the argument above, the original subsequence has another subsequence that converges uniformly to a limit $\left(\tilde{\Pi}_2, \tilde{G}_2\right)$ on $[\underline{Q}_2, \bar{Q}_2]$, and the limit function satisfies Equation (39) on $[\underline{Q}_2, \bar{Q}_2]$. Proceeding iteratively, we can cover the entire real line $\mathbb{R}$, which concludes our proof.

## A.2 Proof of Proposition 3

We break the argument down into multiple steps.

1. *Inferring $\Pi(Q_t)$ and $\alpha\Pi'(Q_t)$ from dividends and prices.* Traders observe the dividend process $D_t$, and, therefore, the path of the Brownian Motion $B_t$, since $D_t = \mu t + \sigma B_t$. Since traders also observe the price path, they infer the price markup $\Pi(Q_t)$ (see Equation (5)). From observing the price path, traders also infer the price volatility $\sigma_P(Q_t)$,[34] which implies that they infer $\alpha\Pi'(Q_t)$.

2. *Inferring Q-learner demand $N^Q(Q_t)$, $Q_t/\beta$, and the ratio $\alpha/\beta$ from asset demands.* Since asset supply $S$ is fixed and traders know their own demand $N_t^R$, they know $N^Q(Q_t)$ at each time $t$, i.e. they know the path of the Q-learner demand up to any time $t$. Equation (11) implies that this is equivalent to knowing $Q_t/\beta$ for all $t$. From knowing the path of $N^Q(Q_t)$, they similarly infer the volatility of $N^Q(Q_t)$, which is given by

$$\frac{\alpha}{2\beta}\left(1+N^Q(Q_t)\right)\left(1-N^Q(Q_t)\right)\sigma_P(Q_t).$$

   Since both $N^Q(Q_t)$ and $\sigma_P(Q_t)$ are known, traders infer the ratio $\alpha/\beta$.

3. *Inferring $\alpha G'(Q_t)$ and $\alpha^2\Pi''(Q_t)$ from realized returns.* Knowing both $\sigma_P(Q_t)$ and the path of $B_t$, traders can infer $\alpha G'(Q_t)$ from observing the path of realized returns $R_t$. Since traders know the price volatility, they can also infer excess volatility. Excess volatility itself is follows a diffusion process, which is given by

$$d\sigma_P(Q_t) = \left(\frac{2\alpha\sigma\Pi''(Q_t)}{r(1-2\alpha\Pi'(Q_t))^2}\mu_Q(Q) + \frac{1}{2}\frac{2\alpha\sigma\left(4\alpha\Pi''(Q_t)^2+\Pi'''(Q_t)(1-2\alpha\Pi'(Q_t))\right)}{r(1-2\alpha\Pi'(Q_t))^3}\sigma_Q^2(Q_t)\right)dt + \frac{2\alpha\sigma\Pi''(Q_t)}{r(1-2\alpha\Pi'(Q_t))^2}\sigma_Q(Q_t)\,dB_t.$$

   From this process, traders can infer the "volatility of excess volatility,"

$$\alpha\sigma'_{\Delta P}(Q_t)\,\sigma_P(Q_t) = \frac{\sigma}{r}\frac{4\alpha^2\Pi''(Q_t)}{(1-2\alpha\Pi'(Q_t))^2}\sigma_P(Q_t).$$

   Hence, traders can infer $\alpha^2\Pi''(Q_t)$.

4. *Inferring $\alpha Q_t$ from the first-order constraint.* Traders know their own demand $N_t^R$ and understand that it must satisfy the first-order constraint (20), which implies that they know

$$\frac{\mu_P(Q_t) - \frac{\mu}{r} - r\Pi(Q_t) - \alpha G'(Q_t)\sigma_P^2(Q_t)}{\phi r\sigma_P^2(Q_t)}.$$

   Using Equation (16), and their knowledge of $\alpha\Pi'(Q_t)$, $\alpha^2\Pi''(Q_t)$, and $\alpha G'(Q_t)$, traders

[34] Intuitively, observing path of $P_t$ is equivalent to observing the path of $P_t^2$, and Ito's Lemma implies that $dP_t^2 - 2P_t dP_t = \sigma_P^2(Q_t)\,dt$, so traders can infer $\sigma_P(Q_t)$.

can then back out the value of $\alpha Q_t$.

5. *Inferring* $(\alpha, \beta, Q_0)$. We have now established that traders know $\alpha/\beta$ and the paths of $Q_t/\beta$ and $\alpha Q_t$. Hence, they can infer $\alpha$, $\beta$, and $Q_t$ for all $t$. Thus, the hyperparameters $(\alpha, \beta, Q_0)$ are common knowledge without loss of generality.

## A.3 Proof of Proposition 4

Given the price conjecture in Equation (44), the Q-value and risk premium evolve according to Equations (14) and (13). The HJB equation of trader type $i$ is given by

$$\begin{aligned}\delta V^i &= \max_{C_i, N_i} -\frac{1}{\phi} e^{-\phi C_i} + V_W^i \left(rW - C + N_i \left(\mu_P(Q) - rP(D,Q) + D\right)\right) \\ &\quad + \frac{1}{2} V_{WW}^i \left(N_i \sigma_P(Q) + Y_i\right)^2 \\ &\quad + V_Q^i \mu_Q(Q) + V_{QQ} \frac{1}{2} \sigma_Q^2(Q) \\ &\quad + V_{WQ}^i \left(N_i \sigma_P(Q) + Y_i\right) \sigma_Q(Q).\end{aligned}$$

Using the conjecture $V^i(W,Q) = -\frac{1}{\phi r} \exp\left(-\phi r W - G_i(Q) - \frac{\delta - r}{r}\right)$, the FOC for $N_i$ is given by

$$N_i(Q) = \frac{1}{\phi r \sigma_P^2(Q)} \left(\mu_P(Q) - rP(D,Q) + D\right) - \frac{2\alpha}{\phi r} G_i'(Q) - \frac{Y_i}{\sigma_P(Q)}.$$

Plugging the FOC for $N_i$ and the FOC for $C$ into the HJB equation, using $Y_i^2 = Y^2$, and simplifying yields

$$\begin{aligned}r G_i(Q) &= \frac{1}{2} (\phi r)^2 N_i^2(Q) \sigma_P^2(Q) - \frac{1}{2} (\phi r Y)^2 \\ &\quad + G_i'(Q) \mu_Q(Q) + \left(G_i''(Q) - G_i'(Q)^2\right) \frac{1}{2} \alpha^2 \sigma_P^2(Q) \\ &\quad - G_i'(Q) \phi r \sigma_Q(Q) Y_i.\end{aligned}$$

The market clearing condition is given by

$$m N_1(Q) + (1-m) N_2(Q) = \frac{1}{1-\theta} \left(S - \theta N^Q(Q)\right).$$

Plugging in the optimal demands and rearranging then yields Equation (27). We can rewrite that equation as

$$\begin{aligned}\mu_P(Q) - rP(D,Q) + D &= \Pi'(Q)\mu_Q(Q) + \frac{1}{2}\Pi''(Q)\sigma_Q^2(Q) - r\Pi(Q) \\ &= \frac{\phi r\sigma_P^2(Q)}{1-\theta}\left(S - \theta N^Q(Q)\right) + \bar{G}'(Q)\sigma_P(Q)\sigma_Q(Q) + \phi r\bar{Y}\sigma_P(Q)\end{aligned}$$

and then substitute into $N_i(Q)$, which yields Equation (28). Plugging that equation into Equation (43) then yields Equation (26).

Inspecting Equations (26) and (27), the only difference to Equations (22) and (23) are the terms $G_i'(Q)\phi r\sigma_Q(Q)Y_i - \frac{1}{2}(\phi rY)^2$ and $\bar{G}'(Q)\sigma_P(Q)\sigma_Q(Q) + \phi r\bar{Y}\sigma_P(Q)$, respectively. The existence proof of Proposition 2 then carries through with only minor modifications. In particular, fixing two functions $(G_1(Q), G_2(Q))$ with uniformly bounded derivatives implies that $\bar{G}'(Q)$ is uniformly bounded, so that we can apply Proposition 5 to Equation (27) replacing $G'(Q)$ with $\bar{G}'(Q)$. The argument in Lemma 1 applies with minor modifications, which establishes that $\sigma_P(Q) \geq \underline{\sigma} > 0$ for all $Q$ and that $\sigma_P(Q)$ is uniformly bounded. Then, we can apply the argument in Proposition 6 separately to the functions $G_1(Q)$ and $G_2(Q)$. Thus, Propositions 5 and 6 imply that we can a priori assume bounded and continuous solutions to the ODEs (26) and (27). This is the main part of the existence argument in Proposition 7. The remainder of the proof of Proposition 7 holds for an arbitrary dimensional ODE system, and thus continues to go apply.

# B Convergence of discrete time Q-learner demand

In single-agent environments, the Q-learning algorithm is guaranteed to converge to the true Q-matrix $Q(s,a)$ under mild conditions, which among others guarantee that each action is sampled infinitely many times in each state over an infinite horizon. However, in more general settings, such convergence is not guaranteed.

In this Appendix, we show that the Q-learner demand from the discrete time specification converges to the continuous time specification that we assume in Section (3.1). The discrete-time analog of Equation (1) is given by

$$D_{t+1} - D_t = \mu + \sigma\varepsilon_{t+1},$$

where $\varepsilon_t \sim N(0,1)$ is iid.

In what follows, we conjecture that the equilibrium price can be expressed as

$$P_t = \frac{1}{r}\left(D_t + \frac{\mu}{r}\right) + \Pi\left(Q_t\right), \tag{44}$$

where $\Pi\left(Q\right)$ is a twice continuously differentiable function. Then, a Q-learner's realized return is given by

$$R_{t+1} = P_{t+1} - P_t + D_t - rP_t = \frac{\sigma}{r}\varepsilon_{t+1} + \Pi\left(Q_{t+1}\right) - \left(1+r\right)\Pi\left(Q_t\right)$$

and the evolution of the Q-value is given by

$$Q_{t+1} - Q_t \;\; = \;\; 2\alpha\left(\frac{\sigma}{r}\varepsilon_{t+1} + \Pi\left(Q_{t+1}\right) - \left(1+r\right)\Pi\left(Q_t\right) - \frac{1}{2}Q_t\right).$$

Note that the evolution of Q-values exhibits a dynamic feedback effect. Specifically, next period's Q-value $Q_{t+1}$ is determined by future price markup $\Pi\left(Q_{t+1}\right)$, as the price markup affects the realized return. However, the future markup is determined by the future Q-value $Q_{t+1}$,and so on. As a result, the continuous-time limit is well defined only if $\Pi'\left(Q\right) < \frac{1}{2\alpha}$ for all $Q$. To see why heuristically, take a small change in $Q_t$, which implies that

$$\frac{Q_{t+1} - 2\alpha\Pi\left(Q_{t+1}\right) - \left(Q_t - 2\alpha\Pi\left(Q_t\right)\right)}{Q_{t+1} - Q_t} \approx 1 - 2\alpha\Pi'\left(Q_t\right),$$

and so we can express Equation (45) as

$$Q_{t+1} - Q_t \approx \frac{2\alpha}{1 - 2\alpha\Pi'\left(Q_t\right)}\left(\frac{\sigma}{r}\varepsilon_{t+1} - r\Pi\left(Q_t\right) - \frac{1}{2}Q_t\right).$$

This implies that as $\Pi'\left(Q_t\right) \to \frac{1}{2\alpha}$, the RHS explodes because the noise in the dividend process is amplified infinitely.

As such, in the following result, we impose $\Pi'\left(Q\right) < \frac{1}{2\alpha}$ as a condition. However, we verify this condition holds in equilibrium in Proposition 2.

**Proposition 8.** *Suppose that* $\Pi\left(Q\right)$ *is twice continuously differentiable and that* $\sup_{Q\in\mathbb{R}}\Pi'\left(Q\right) < \frac{1}{2\alpha}$. *Then, as* $h \to 0$, *the approximation*

$$Q_h^*\left(t\right) = \sqrt{h}\left(Q_{\lfloor t/h\rfloor} + \left(t - h\left\lfloor\frac{t}{h}\right\rfloor\right)\left(Q_{\lfloor t/h\rfloor+1} - Q_{\lfloor t/h\rfloor}\right)\right) \tag{46}$$

*converges in distribution to the unique weak solution of the stochastic differential equation*

*(SDE)*

$$Q_t = Q_0 + \int_0^t \mu_Q(Q_s)\, ds + \int_0^t \sigma_Q(Q_s)\, dB_s,$$

*where*

$$\mu_Q(Q_t) = \frac{2\alpha}{1 - 2\alpha\Pi'(Q_t)} \left( \frac{1}{2}\sigma_Q(Q_t)^2 \Pi''(Q_t) - r\Pi(Q_t) - \frac{1}{2}Q_t \right)$$

*and*

$$\sigma_Q(Q_t) = \frac{\sigma}{r} \frac{2\alpha}{1 - 2\alpha\Pi'(Q_t)}.$$

Similarly, the approximation of the price $P_t$ converges to Equation (13).

**Corollary 4.** *Under Conjecture (44), as $h \to 0$ the approximation*

$$P_h^*(t) = \sqrt{h} \left( P_{\lfloor t/h \rfloor} + \left( t - h \left\lfloor \frac{t}{h} \right\rfloor \right) \left( P_{\lfloor t/h \rfloor + 1} - P_{\lfloor t/h \rfloor} \right) \right)$$

*converges in distribution to the unique weak solution of the SDE*

$$P_t = \frac{\mu}{r^2} + \int_0^t \mu_P(Q_s)\, ds + \int_0^t \sigma_P(Q_s)\, dB(s)$$

*where*

$$\mu_P(Q_t) = \frac{\mu}{r} + \Pi'(Q_t)\, \mu_Q(Q_t) + \frac{1}{2}\Pi''(Q_t)\, \sigma_Q^2(Q_t)$$

*and*

$$\sigma_P(Q_t) = \frac{\sigma}{r} \frac{1}{1 - 2\alpha\Pi'(Q_t)}.$$

*In particular, for all $t \geq 0$,*

$$P_t = \frac{\mu}{r^2} + \frac{1}{r} D_t + \Pi(Q_t).$$

*Proof.* This follows from applying Ito's Lemma to Conjecture (44). □

## B.1 Proof of Proposition 8

Rewrite Equation (45) as

$$Q_{t+1} - 2\alpha\Pi(Q_{t+1}) = Q_t - 2\alpha\Pi(Q_t) + \frac{\alpha\sigma}{r}\varepsilon_{t+1} + \mu(Q_t),$$

where $\mu(Q) = -\alpha Q - 2\alpha r \Pi(Q)$. Define

$$\tilde{Q}_t = H(Q_t) = Q_t - \alpha\Pi(Q_t).$$

Given the assumption $\sup_{Q\in\mathbb{R}}\Pi'(Q) < 1/(2\alpha)$ , $H(Q)$ is invertible. Then, we can write

$$\tilde{Q}_{t+1} = \tilde{Q}_t + \frac{2\alpha\sigma}{r}\varepsilon_{t+1} + \tilde{\mu}\left(\tilde{Q}_t\right)$$

where $\tilde{\mu}\left(\tilde{Q}_t\right) = \mu\left(H^{-1}\left(\tilde{Q}_t\right)\right)$. Define a grid with step size $h$ between any points $t$ and $t'$. Indexing points on the grid with $k$, define the approximation

$$\tilde{Q}_{k+1,h} = \tilde{Q}_{k,h} + h\tilde{\mu}\left(\tilde{Q}_{k,h}\right) + \sqrt{h}\frac{2\alpha\sigma}{r}\varepsilon_{k+1,h},$$

where $\varepsilon_{k,h} \sim N(0,1)$ for all $k$ and $h$. Finally, define the linear interpolation as $\tilde{Q}^*_h(t)$, where

$$\tilde{Q}^*_h(t) = \sqrt{h}\left(\tilde{Q}_{\lfloor t/h\rfloor,h} + \left(t - h\left\lfloor\frac{t}{h}\right\rfloor\right)\left(\tilde{Q}_{\lfloor t/h\rfloor+1,h} - \tilde{Q}_{\lfloor t/h\rfloor,h}\right)\right).$$

The function $\tilde{Q}^*_h(t)$ defines a Markov process which is deterministic at all points $t \in (kh,(k+1)h)$ and for which the transition probabilities times $t \in \{kh\}_{i\in\mathbb{N}}$ are defined by

$$\Pr\left(\tilde{Q}^*_h((k+1)h) \in \Gamma | \tilde{Q}^*_h(kh) = x\right) = \Pi_h(x,\Gamma)$$

for any $\Gamma \in \mathcal{B}(\mathbb{R})$. Since $\varepsilon_{k,h}$ is normally distributed, $\tilde{Q}^*_h(t)$ admits a transition density, which is given by

$$\Pi_h(x,y) = \frac{1}{\sqrt{2\pi h}\tilde{\sigma}}\exp\left(-\frac{1}{2}\frac{(y-x-h\tilde{\mu}(x))^2}{h\tilde{\sigma}^2}\right),$$

where we wrote $\tilde{\sigma} = 2\alpha\sigma/r$ to save notation.

Let $C^\infty_0(\mathbb{R})$ denote the space of infinitely continuously differentiable functions $f:\mathbb{R}\to\mathbb{R}$ with compact support. For any $f \in C^\infty_0(\mathbb{R})$,define

$$A_h f = \frac{1}{h}\int (f(y) - f(x))\,\Pi_h(x,y)\,dy.$$

**Lemma 2.** *For any $f \in C^\infty_0(\mathbb{R})$*

$$A_h f \to \mathcal{L} f$$

*uniformly on compact subsets of $\mathbb{R}$ as $h \to 0$, where*

$$\mathcal{L} f(x) = \tilde{\mu}(x) f'(x) + \frac{1}{2}\tilde{\sigma} f''(x).$$

*Proof.* Since $\mathbb{R}$ is locally compact, to establish uniform convergence on compact subsets, it

is sufficient to establish locally uniform convergence, i.e. for all $x \in \mathbb{R}$, there exists an $\varepsilon > 0$ such that

$$\lim_{h \to 0} \sup_{x' \in B(x,\varepsilon)} \left| A_h f\left(x'\right) - \mathcal{L} f\left(x'\right) \right| = 0$$

for any $f \in C_0^\infty(\mathbb{R})$, where $B\left(x, \varepsilon\right)$ is the $\varepsilon$-ball around $x$.

Since we assumed that $\Pi\left(.\right)$ is continuously differentiable, $\tilde{\mu}\left(.\right)$ is locally Lipschitz continuous, i.e. there exists a $K > 0$ such that $\left|\tilde{\mu}\left(x'\right) - \tilde{\mu}\left(x\right)\right| \leq K \left|x' - x\right|$ for all $x' \in B\left(x, \varepsilon\right)$, and in particular, $\left|\tilde{\mu}\left(x'\right) - \tilde{\mu}\left(x\right)\right| \leq K\varepsilon$.

Since $f$ is infinitely differentiable, Taylor's theorem implies that

$$f\left(y\right) = \sum_{l=0}^{\infty} \frac{1}{l!} f^{(l)}\left(x\right) \left(y - x\right)^l.$$

Thus,

$$A_h f = \frac{1}{h} \left( f'\left(x\right) E_h\left[y - x\right] + \frac{1}{2} f''\left(x\right) E_h\left[\left(y - x\right)^2\right] + \sum_{l=3}^{\infty} \frac{1}{l!} f^{(l)}\left(x\right) E_h\left[\left(y - x\right)^l\right] \right)$$

where $E_h\left[.\right]$ is the expectations operator under $\Pi_h$. We have $E_h\left[y - x\right] = h\tilde{\mu}\left(x\right)$ for all $x$, which in particular implies that

$$\lim_{h \to 0} \sup_{x' \in B(x,\varepsilon)} \left| \frac{1}{h} E_h\left[y - x'\right] - \tilde{\mu}\left(x'\right) \right| = 0.$$

Similarly,

$$\begin{aligned} E_h\left[\left(y - x\right)^2\right] &= E_h\left[\left(y - x - h\tilde{\mu}\left(x\right) + h\tilde{\mu}\left(x\right)\right)^2\right] \\ &= E_h\left[\left(y - x - h\tilde{\mu}\left(x\right)\right)^2\right] + 2E_h\left[\left(y - x - h\tilde{\mu}\left(x\right)\right)\right] h\tilde{\mu}\left(x\right) \\ &\quad + h^2 \tilde{\mu}\left(x\right)^2 \end{aligned}$$

and

$$\begin{aligned} \lim_{h \to 0} \sup_{x' \in B(x,\varepsilon)} \left| \frac{1}{h} E_h\left[y - x'\right]^2 - \tilde{\sigma} \right| &= \lim_{h \to 0} \sup_{x' \in B(x,\varepsilon)} \left| h^2 \tilde{\mu}^2\left(x'\right) \right| \\ &\leq \lim_{h \to 0} h^2 K^2 \varepsilon^2 \\ &= 0. \end{aligned}$$

Proceeding similarly for the higher order terms, we have

$$\begin{aligned} E_h\left[(y-x)^3\right] &= E_h\left[(y-x-h\tilde{\mu}(x)+h\tilde{\mu}(x))^3\right] \\ &= E_h\left[(y-x-h\tilde{\mu}(x))^3\right]+3h^2\tilde{\mu}(x)\tilde{\sigma}^2 \\ &= +3h^2\tilde{\mu}(x)^2 E_h\left[(y-x-h\tilde{\mu}(x))\right]+h^3\tilde{\mu}(x)^3. \end{aligned}$$

Stein's Lemma[35] implies that

$$E_h\left[(y-x-h\tilde{\mu}(x))^3\right]=2h\tilde{\sigma}^2 E_h\left[y-x-h\tilde{\mu}(x)\right]=0$$

and thus

$$E_h\left[(y-x)^3\right]=3h^2\tilde{\mu}(x)\tilde{\sigma}^2+h^3\tilde{\mu}(x)^3$$

which implies that

$$\begin{aligned} \lim_{h\to 0}\sup_{x'\in B(x,\varepsilon)}\left|\frac{1}{h}E_h\left[(y-x)^3\right]\right| &\leq \lim_{h\to 0} 3h^2\tilde{\sigma}^2K\varepsilon+h^3K^3\varepsilon^3 \\ &= 0 \end{aligned}$$

as $h\to 0$. Similarly,

$$\begin{aligned} E_h\left[(y-x)^4\right] &= E_h\left[(y-x-h\tilde{\mu}(x)+h\tilde{\mu}(x))^4\right] \\ &= E_h\left[(y-x-h\tilde{\mu}(x))^4\right]+4h\tilde{\mu}(x)E_h\left[(y-x-h\tilde{\mu}(x))^3\right] \\ &\quad +6h^4\tilde{\mu}(x)^2\tilde{\sigma}^2+4h^3\tilde{\mu}(x)^3E_h\left[(y-x-h\tilde{\mu}(x))\right] \\ &\quad +h^4\tilde{\mu}(x)^4. \end{aligned}$$

Using Stein's Lemma again implies that

$$\begin{aligned} E_h\left[(y-x-h\tilde{\mu}(x))^4\right] &= 3h\tilde{\sigma}^2E_h\left[(y-x-h\tilde{\mu}(x))^2\right] \\ &= 3h^2\tilde{\sigma}^4, \end{aligned}$$

so that

$$\begin{aligned} \lim_{h\to 0}\sup_{x'\in B(x,\varepsilon)}\left|\frac{1}{h}E_h\left[(y-x)^4\right]\right| &\leq \lim_{h\to 0} 3h^2\tilde{\sigma}^4+6h^4K^2\varepsilon^2\tilde{\sigma}^2+h^4K^4\varepsilon^4 \\ &= 0. \end{aligned}$$

[35]For any differentiable function $g(X)$ and $X\sim N(\mu,\sigma^2)$, $E[g(X)(X-\mu)]=\sigma^2E[g'(X)]$.

Proceeding inductively, it follows that

$$\lim_{h\to 0}\sup_{x'\in B(x,\varepsilon)}\frac{1}{h}\sum_{l=3}^{\infty}\frac{1}{l!}f^{(l)}(x)\,E_h\left[(y-x)^l\right]=0$$

as $h\to 0$ and therefore

$$A_h f(x)\to f'(x)\,\tilde{\mu}(x)+\frac{1}{2}f''(x)\,\tilde{\sigma}^2=\mathcal{L}f$$

uniformly on compact subsets. □

We now apply Stroock and Varadhan (1997), Th. 11.2.3. Define with $P_h$ the probability measure induced by $\tilde{Q}_h^*(t)$ on $C([0,\infty),\mathbb{R})$, the space of continuous trajectories from $[0,\infty)$ into $\mathbb{R}$. Similarly, define with $P$ the probability measure on $C([0,\infty),\mathbb{R})$ which solves the martingale problem for $\mathcal{L}$.[36] The theorem states that if the martingale problem for $\mathcal{L}$ has a unique solution, and for all $f\in C_0^\infty(\mathbb{R})$, $A_h f\to\mathcal{L}f$ uniformly on compact subsets of $\mathbb{R}$, then $P_h\to P$ uniformly on compact subsets of $\mathbb{R}$.

The previous lemma establishes the necessary convergence. It only remains to show that the martingale problem for $\mathcal{L}$ has a unique solution. Kurtz (2010), Th. 1, states that the martingale problem for $\mathcal{L}$ has a unique solution if and only if the SDE associated with $\mathcal{L}$ has a unique weak solution. The SDE

$$\tilde{Q}_t=\int_0^t\tilde{\mu}\left(\tilde{Q}_s\right)ds+\tilde{\sigma}B_t \tag{47}$$

has a constant and positive volatility $\tilde{\sigma}$, and $\tilde{\mu}(\cdot)$ is continuous and therefore bounded on compact subsets of $\mathbb{R}$. Yong and Zhou (1999), Th. 6.1.2 then establishes that Equation (47) has a unique weak solution. Hence, Stroock and Varadhan (1997), Th. 11.2.3 implies that $P_h\to P$, or equivalently

$$\tilde{Q}_h^*(t)\to^d\tilde{Q}_t$$

on compact subsets of $\mathbb{R}$.

It remains to transform $\tilde{Q}_t$ back to $Q_t$. By the continuous mapping theorem, $Q_h^*(t)$, the linear interpolation of $Q_t$ in Equation (46), converges in distribution to

$$Q_t=H^{-1}\left(\tilde{Q}_t\right).$$

Since $\tilde{Q}_t$ is a diffusion process, we can use Ito's Lemma and the implicit function rule to

[36] See Stroock and Varadhan (1997), p. 138 for a definition of martingale problems.

obtain

$$\begin{aligned} dQ_t &= \left( H^{-1\prime}\left(\tilde{Q}_t\right)\tilde{\mu}\left(\tilde{Q}_t\right) + H^{-1\prime\prime}\left(\tilde{Q}_t\right)\frac{1}{2}\left(\frac{2\alpha\sigma}{r}\right)^2 \right) dt + H^{-1\prime}\left(\tilde{Q}_t\right) 2\alpha\frac{\sigma}{r} dB_t \\ &= \frac{2\alpha}{1-2\alpha\Pi'\left(Q_t\right)}\left( -r\Pi\left(Q_t\right) - \frac{1}{2}Q_t + \frac{\Pi''\left(Q_t\right)}{\left(1-2\alpha\Pi'\left(Q_t\right)\right)^2}\frac{1}{2}\left(\frac{2\alpha\sigma}{r}\right)^2 \right) dt \\ &\quad + \frac{1}{1-2\alpha\Pi'\left(Q_t\right)}\frac{2\alpha\sigma}{r} dB_t \\ &= \mu_Q\left(Q_t\right) dt + \sigma_Q\left(Q_t\right) dB_t. \end{aligned}$$

Given our assumption $\sup_{Q\in\mathbb{R}} \Pi'(Q) < 1/(2\alpha)$, it holds that $\sigma_Q(Q) \geq \underline{\sigma}_Q > 0$ for all $Q \in \mathbb{R}$ and $|\mu_Q(Q)| < \infty$ for all $Q \in \mathbb{R}$, i.e. the SDE for $Q_t$ is uniformly elliptic and has locally bounded drift. It is hence well-posed.

# C Supplementary Figures and Analysis

## C.1 Replacing Q-learners with a risk-neutral trader

We now consider a benchmark in which the Q-learners are replaced with a risk-neutral trader who chooses trades optimally and is a price taker.[37] All other assumptions in the model are unchanged, and, in particular, the risk-neutral trader is subject to choosing $N_t^N \in [-1, 1]$. This model variant features multiple equilibria and is prone to sunspots. For example, depending on $S$ and $\theta$, there may be an equilibrium where the risk-neutral trader always buys a share, but also another equilibrium where the risk-neutral trader sells a share.

Ignoring sunspots, however, any equilibrium is simple and resembles the equilibrium in Proposition 1. It features a constant risk premium $\Pi$, so that

$$P(D_t) = \frac{1}{r}D_t + \frac{\mu}{r^2} + \Pi.$$

Either $D_t + \mu/r - rP_t > 0$ and the risk-neutral trader buys one unit, i.e. $N_t^N = 1$, $D_t + \mu/r - rP_t < 0$, and the risk-neutral trader sells one unit, i.e. $N_t^N = -1$, or $D_t + \mu/r - rP_t = 0$ and the risk-neutral trader is indifferent.

**Proposition 9.** *If $S < \theta$, there exists an REE in which $\Pi < 0$ and $N_t^N = 1$ for all $t$. If $S > -\theta$, there exists an REE in which $\Pi > 0$ and $N_t^N = -1$ for all $t$. If $S \in [-\theta, \theta]$, there exists an REE in which $\Pi = 0$ for all $t$.*

[37] Equivalently, there exists a continuum of rational traders of mass $\theta$.

*Proof.* Conjecture that $\Pi < 0$. Then, $N_t^N = 1$ and the market clearing condition becomes $(1-\theta) N_t^R + \theta = S$. Together with the FOC of the traders with CARA utility, this implies that

$$\Pi = \frac{\phi \sigma_{P0}^2}{1-\theta} (S - \theta).$$

Thus, $\Pi < 0$ whenever $S < \theta$, otherwise, this cannot be an equilibrium. Next, conjecture that $\Pi > 0$. Then, $N_t^N = -1$ and the market clearing condition is $(1-\theta) N_t^R - \theta = S$, which implies that

$$\Pi = \frac{\phi \sigma_{P0}^2}{1-\theta} (S + \theta).$$

Thus, $\Pi > 0$ whenever $S > -\theta$. Finally, conjecture that $\Pi = 0$. Then, it must be the case that

$$\Pi = \frac{\phi \sigma_{P0}^2}{1-\theta} \left(S - \theta N^N\right)$$

so that $N^N = S/\theta$. □

In particular for $|S| > \theta$, there exists a unique REE in which the risk premium is constant, in line with Proposition 1. For $|S| < \theta$, there may be sunspots. With Q-learners, any variation in risk premia is driven by changes in the Q-value, whereas here, any potential change in the risk premium is due to sunspots. Thus, the two models generate price changes for fundamentally different reasons.

## C.2 Hyperparameter Choice

In the baseline model, we assumed that the parameters of the Q-learning algorithm are exogenous. In reality, these parameters are determined by traders who choose to delegate their trading to an algorithm. In this section, we assume that a there is a representative algorithmic trader with CARA utility, who delegates trading to the algorithm[38] and optimally chooses the learning rate $\alpha$ and the temperature parameter $\beta$. That is, the algorithmic trader determines how fast the algorithm learns and how aggressively it trades to maximize her utility, anticipating the resulting equilibrium. The proposition below shows that there exists a Q-REE in this setting. This equilibrium takes the same form as in Proposition 2, except that $\alpha$ and $\beta$ are now endogenous.

[38]This is optimal if trading "rationally" is costly. For example, hedge funds may need to hire quants and traders and share profits with them to incentivize them to gather information and trade. Trading via an algorithm may be cheaper, and it is optimal to delegate whenever $V^D(W_0 - I, Q_0) \geq V(W_0, Q_0)$, where $V^D$ is the trader's expected value form delegating and $I$ is the cost.

Specifically, the algorithmic trader's problem is now given by

$$V^D(W_0, Q_0) = \sup_{\{C_s\}_{s\geq 0}, \alpha\in[0,1], \beta\geq 0} -E_t\left[\int_0^\infty e^{-\delta(s-t)}\frac{1}{\phi}e^{-\phi C_s}ds\right]$$

subject to

$$dW_t = \left(rW_t - C_t + N^Q(Q_t)(D_t - rP_t)\right)dt + N^Q(Q_t)\,dP_t$$

and

$$dQ_t = \alpha dR_t - \frac{\alpha}{2}Q_t dt.$$

That is, the trader chooses $\alpha$ and $\beta$ at time zero and chooses how much to consume at each time, but lets the algorithm trade on his behalf. Here, note that $N^Q(Q)$ depends implicitly on $\beta$ (see Equation (11)).[39] All other traders behave as in the baseline model.

**Proposition 10.** *A Q-REE in which the large trader delegates trading to the algorithm and chooses $\alpha$ and $\beta$ optimally exists and takes the same form as the equilibrium in Proposition 2. The delegating trader's value function is given by*

$$V^D(W, Q) = -\frac{1}{\phi r}\exp\left(-\phi r W - G_D(Q) - \frac{\delta - r}{r}\right),$$

*where $G_D(Q)$ is determined by the ODE*

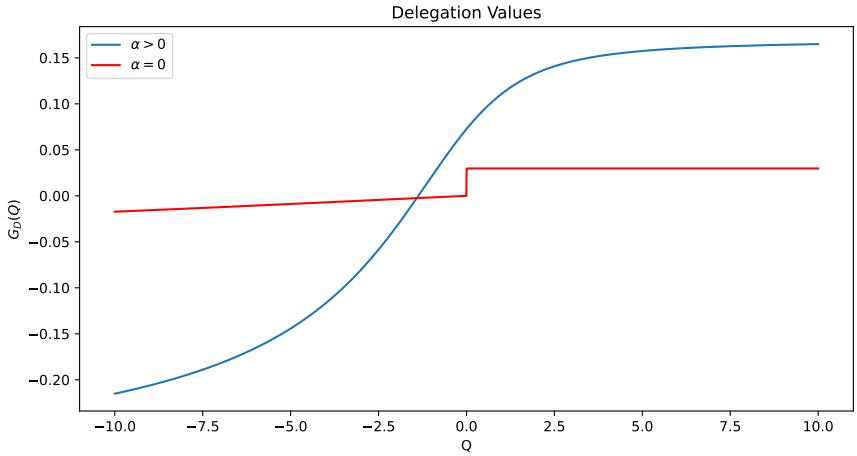
$$\begin{aligned} rG_D(Q) &= \phi r N^Q(Q)\left(\mu_P(Q) - \frac{\mu}{r} - rP(Q)\right) \\ &\quad -\frac{1}{2}(\phi r)^2\sigma_P^2(Q)N^Q(Q)^2 \\ &\quad +G_D'(Q)\left(\mu_Q(Q) - \alpha\phi r\sigma_P^2(Q)N^Q(Q)\right) \\ &\quad +\frac{1}{2}\left(G_D''(Q) - G_D'(Q)^2\right)\alpha^2\sigma_P^2(Q) \end{aligned}$$

*with boundary conditions $G_D'(-\infty) = G_D'(\infty) = 0$, and $P(Q)$ and $G(Q)$ satisfy Equations (22) and (23).*

[39] As we have shown in Section 4.3, traders learn the hyperparameters from observing equilibrium prices over an arbitrarily small time horizon. Thus, without loss of generality, we assume that the trader's choice of $\alpha$ and $\beta$ is public.

Figure 11: Delegation Values for $\alpha = 0$ and $\alpha > 0$

The figure plots the delegation values $G_D(Q)$ for the large trader for $\alpha = 0$ (red) and $\alpha > 0$ (blue). Whenever the blue line is above the red line, it is optimal to set $\alpha > 0$ for a given $Q_0$. Parameter values are set to: $r = 0.05$, $\phi = 1$, $\sigma = 0.075$, $\alpha = 0.1$ and $\beta = 1$ (for the blue line), $\theta = 0.3$, and $S = 5$.




*Proof.* The large trader's HJB equation is given by

$$\begin{aligned}\delta V^D &= \max_C -\frac{1}{\phi}e^{-\phi C} + V_W^D\left(rW - C + N^Q(Q)\left(\mu_P(Q) - \frac{\mu}{r} - rP(Q)\right)\right)\\ &\quad + \frac{1}{2}V_{WW}^D\sigma_P^2(Q)\,N^Q(Q)^2 + V_Q^D\mu_Q(Q) + \frac{1}{2}V_{QQ}^D\alpha^2\sigma_P^2(Q)\\ &\quad + V_{WQ}^D\alpha\sigma_P^2(Q)\,N^Q(Q).\end{aligned}$$

Conjecturing an equilibrium of the same form as in Proposition 2, plugging the FOC for consumption $e^{-\phi C} = V_W^D(Q)$, and rearranging yields Equation (48). Existence of the equilibrium follows from the same argument as for Proposition 2, which established existence for arbitrary values of $\alpha$ and $\beta$. □

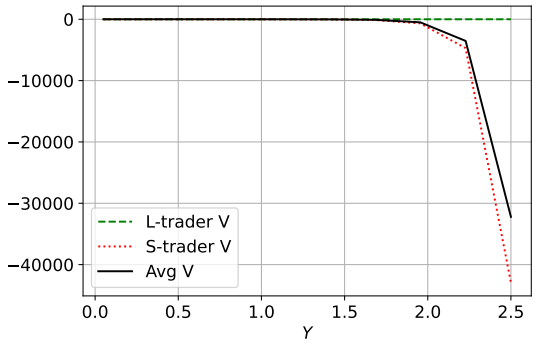
Whenever $|S|$ is sufficiently large, the Q-learner makes positive profits. In Figure 11, we provide a numerical example when it is optimal to set $\alpha > 0$ and $\beta > 0$, so that in equilibrium, the Q-value indeed reacts to price changes.[40]

[40]The kink in the red line occurs because for $S > 0$, buying is profitable, but when $Q_0 < 0$, $N^Q \leq 0$. It is then optimal to set $\beta = \infty$, so that $N^Q = 0$. By contrast, when $Q_0 > 0$, $N^Q \geq 0$ and it is optimal to set $\beta = 0$ which ensures that $N^Q = 1$, i.e. the policy is greedy and the algorithm buys as much as possible.

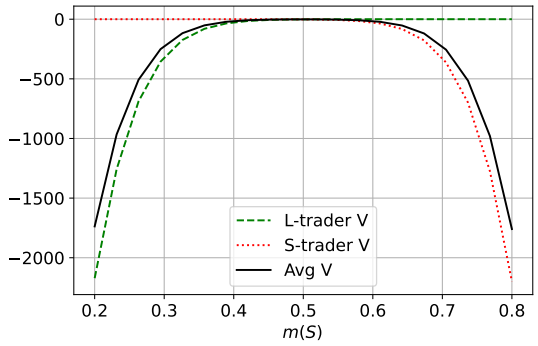
## C.3 Endowment risk: Expected value for rational investors

Figure 12: Expected Value for rational investors

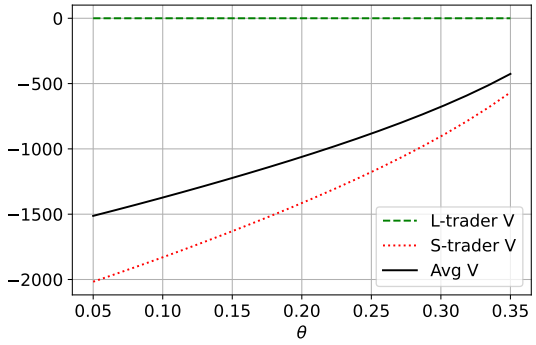
The figure plots the expected utility for $L$ and $S$ investors, $V_L$ (dashed) and $V_S$ (dotted), and the average gain $V = mV_S + (1-m)V_L$ (solid), versus various model parameters. Unless specified, parameters are set to: $\phi = 10$, $\sigma = 0.075$, $r = 0.05$, $\beta = 1$, $\mu = 0.05$, $\theta = 0.30$, $\alpha = 0.10$, $S = 0$, $m = 0.75$ and $Y = 2$.



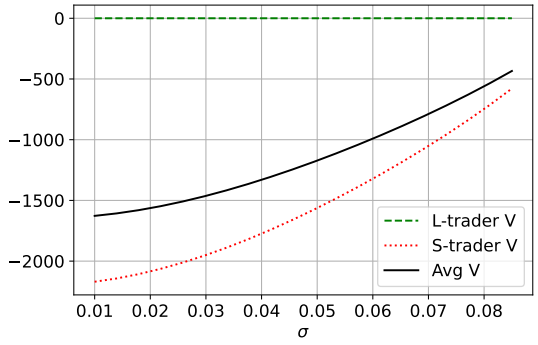

(a) Investor value versus $Y$




(b) Investor value versus $m$




(c) Investor value versus $\theta$




(d) Investor value versus $\sigma$