Bounded Rationality, Reinforcement Learning, and Market Efficiency

Hyun Soo Doh, Jeonggyu Huh, Byung Hwa Lim*

October, 2024

Abstract

We consider an information-based asset market model where uninformed investors lack perfect knowledge of the economic environment. In the model, uninformed investors employ a reinforcement learning algorithm to find optimal investment strategies without directly inferring the relationship between prices and asset payoffs. Applying this approach to the settings considered by Grossman and Stiglitz (1980) and Breon-Drish (2015), which are analytically tractable, we show that the model outcomes tend to converge to the true rational expectations equilibrium. We extend our approach to more complex settings that are not analytically solvable. The model outcomes still converge to a limit, providing approximate solutions to these analytically intractable models.

Keywords: rational expectations; deep reinforcement learning; price informativeness; stability

JEL Classifications: G14, G11, G12

*Hyun Soo Doh is from College of Business and Economics, Hanyang University, Ansan, South Korea. Email: hsdoh@hanyang.ac.kr. Jeonggyu Huh is from Department of Mathematics, Sungkyunkwan University, Suwon, South Korea. Email: jghuh@skku.edu. Byung Hwa Lim is from Business School, Sungkyunkwan University, Seoul, South Korea. Email: limbh@skku.edu.

1 Introduction

A longstanding issue in finance concerns whether asset prices reflect all available information in financial markets (Fama, 1970; Jegadeesh and Titman, 1993; Welch and Goyal, 2008; Asness et al., 2013; McLean and Pontiff, 2016). To address this central question, Grossman and Stiglitz (1980), Hellwig (1980), and Verrecchia (1982) develop information-based asset market models that explain the specific mechanisms through which asset prices incorporate both public and private information. While the main message of these papers is that financial markets may not be perfectly efficient because investors would otherwise have no incentives to acquire costly information, another important legacy of these seminal papers is the conceptual framework that these papers have established for analyzing numerous issues related to market efficiency.

A key assumption imposed in these papers, however, is that all investors have the perfect knowledge of the economic environment, including payoff distributions, the information technologies of other investors, the populations of distinct types of investors, among other factors. Due to this so-called rational expectations assumption, uninformed investors in these models can accurately guess the true relationship between prices and asset payoffs. Relaxing this assumption is important for two reasons: (i) the economic outcomes of models without the rational expectations assumption may differ from those predicted by the above-mentioned papers and (ii) without a thorough understanding of an economy that is not bound by this assumption, we may not be able to rigorously assess the validity of investment strategies widely used in practice.

Recognizing the limitations of models with the rational expectations assumption, Bray (1982) and Routledge (1999) extend the model of Grossman and Stiglitz (1980) by assuming uninformed investors do not have full rationality. While these studies have significantly improved our understanding of this matter, the assumptions made in these papers are still restrictive. Specifically, Bray (1982) assumes that uninformed investors are still aware of the basic structure of the model, although they do not know the precise values of the model parameters. As such, these investors can still infer that prices and asset payoffs are linearly related to each other, as in Grossman and Stiglitz (1980), and thus naturally choose to run

a linear regression on past data to estimate the precise form of this relationship. Meanwhile, Routledge (1999) introduces a specific behavioral rule, such as mimicking the past successful behavior of other investors, to formulate how uninformed investors update their investment rules without providing solid microfoundations.

In our paper, we consider a setting in which uninformed investors have no prior knowledge of the economic environment. We then assume that the uninformed investors use a reinforcement learning algorithm, a machine learning tool which we will elaborate on later, to learn optimal investment strategies from past data. While reinforcement learning is not an entirely new concept to the field of machine learning, this learning algorithm has only recently gained popularity after several pioneering approaches that integrate reinforcement learning with deep learning models have made significant success in challenging tasks in video games and board games such as Atari and Go; see Mnih et al. (2015) and Silver et al. (2017). More importantly, a reinforcement learning algorithm is particularly suited to our research question because this algorithm was originally developed to tackle dynamic optimization problems where agents do not have any prior knowledge of the model environment; see Watkins (1989), Watkins and Dayan (1992), and Williams (1992).

In our model, uninformed investors do not attempt to directly infer the relationship between prices and asset payoffs from past data, unlike in Bray (1982). Instead, the uninformed investors merely calculate the average utility from realized historical data and adjust their beliefs about the optimal investment strategy in a direction that marginally increases their average utility. This particular type of reinforcement learning algorithm is known as a policy-gradient method in the machine learning literature; see, for instance, Sutton et al. (1999). Moreover, in line with the recent trends, we use a deep neural network to represent the investment strategies of uninformed investors because an optimal investment strategy can be potentially a nonlinear function of a price.

Using this approach, we first examine whether the outcomes in our model tend to converge to the true rational expectations equilibrium in the canonical model by Grossman and Stiglitz (1980). Interestingly, our numerical results show that our model outcomes, such the investment strategies of uninformed investors and the coefficients of a price-and-payoff relationship, indeed converge to their corresponding outcomes in Grossman and Stiglitz (1980)

under almost all reasonable parameter values we have tested. This result is not trivial because while the theoretical convergence results for reinforcement learning algorithms are generally available only for individual Markov decision problems (Watkins and Dayan, 1992), the price-and-payoff relationship in our model varies endogenously over time due to interactions among investors who keep updating their beliefs about optimal investment strategies based on data. Nonetheless, our numerical results strongly suggest that the long-run limit of our model coincides with the true rational expectations equilibrium, indicating that financial markets eventually become informationally efficient—or, to the same extent that is achieved in the model with the rational expectations assumption—if investors with no prior domain knowledge employ a reinforcement learning algorithm.

To justify the validity of our approach, we apply this reinforcement learning-based approach to an extension of Grossman and Stiglitz (1980) that is still analytically tractable. In the literature, Breon-Drish (2015) relaxes the restrictive assumptions regarding payoff distributions, asset supply, and signal structures, imposed in Grossman and Stiglitz (1980), and provides an analytic solution technique for the extended model. While this new solution method can be applied to a broad range of settings, we consider one particular setting that is highly empirically motivated. Specifically, we consider a model in which asset payoffs follow a mixture of normal distributions exhibiting negative skewness and heavy tails—two widely accepted stylized facts in the literature (Cont, 2001; Gabaix, 2009; Kelly and Jiang, 2014). To be more specific, in this model, an unobservable economic state is first realized and then the payoff distribution is determined, depending on the realized state. As advocated by Ang and Timmermann (2012), this state-dependent framework is not only intuitive to understand but also effectively captures many stylized facts, including the two aforementioned patterns.

Notably, our numerical results show that the outcomes in this extended model continue to converge to the true rational expectations equilibrium calculated via the solution method of Breon-Drish (2015). This result thus strengthens our experiment-based assertion that financial markets tend to eventually become informationally efficient, even when investors lack perfect knowledge of the economic environment. While this phenomenon is intriguing on its own, this result also suggests that practitioners, who may not have a complete understanding of the entire market structure and thus rely on a recent machine learning tool, that is,

the reinforcement learning algorithm, are actually on the right track as these investors will ultimately uncover optimal investment strategies in the long run, according to our results.

The fact that our approach successfully works for both the settings considered by Grossman and Stiglitz (1980) and Breon-Drish (2015) gives strong confidence that our approach can be further applied to more complex settings that do not permit analytic solutions. This additional work is important because given that extending the basic model of Grossman and Stiglitz (1980) while keeping analytical tractability has been notoriously challenging, our approach provides a valuable means of obtaining approximate solutions for a number of extended versions of Grossman and Stiglitz (1980). Specifically, we consider two additional extensions. First, we consider a more precise signal structure for the above-mentioned state-dependent economy. Second, we incorporate borrowing and short-sale constraints into the canonical model of Grossman and Stiglitz (1980).

In the first extension, we assume that informed investors in the above state-dependent model receive two signals about the current economic state and the asset payoffs separately. To the best of our knowledge, analytic solutions for this state-dependent model with separate signals are not available. Our result, however, shows that the outcomes of this model still converge to some limit, which can be considered the true rational expectations equilibrium of this model. Moreover, the qualitative results of this model are consistent with our intuition.

Specifically, we find that the long-run limit of the demand function of uninformed investors in this model is lower than that arising in the model with a single signal, especially when the price is at an intermediate level. This result makes sense because the additional signal on the current economic state contains valuable contents, primarily when the price falls in an intermediate region. More specifically, when the price is relatively high, uninformed investors believe that the current economy is highly likely to be in a good state and thus, their demand function would be almost indistinguishable from the demand function obtained in the model with a single signal. However, when the price falls, the degree of information asymmetry between informed investors and uninformed investors increases for the reason mentioned above and therefore, uninformed investors reduce their demand for the asset, compared to the case with a single signal. But when the price falls further, uninformed investors believe that the current economy is highly likely to be in a bad state. Thus, the

demand function of uninformed investors in this case again becomes almost the same as the demand function obtained in the model with a single signal.

Next, we consider an economy with financial constraints. Specifically, we consider a setting where investors are subject to borrowing and short-sale constraints. In the literature, Yuan (2005) extends the model of Grossman and Stiglitz (1980) by introducing borrowing constraints. However, Yuan (2005) formulates borrowing constraints in a reduced form to maintain tractability, which may not encompass a wide range of settings. In our paper, we consider borrowing and short-sale constraints in their most natural forms. That is, we assume that investors cannot take a position in a risky asset above a specific upper limit or below a certain lower limit. that exceeds a certain upper limit or falls below a certain lower limit. Again, while this model does not seem to allow for analytic solutions, our reinforcement learning-based approach can be readily applied.

Our results first show that a long-run limit of the model outcomes still exists, as in other cases. Moreover, in the long run, we find that when the price is low (high), the demand for the risky asset by uninformed investors is higher (low) than that obtained in the benchmark model without financial constraints. To understand this result, note that borrowing constraints tend to be binding when the price is low and short-sale constraints tend to be binding when the price is high. As such, compared to the outcomes in the benchmark economy without financial constraints, the asset in our model tends to be underpriced when the price is low and tends to be overpriced when the price is high. Accordingly, uninformed investors believe that the asset payoff would be higher than that inferred from the benchmark economy when the price is low and the opposite case must hold when the price is high. As such, the demand function of uninformed investors in the model with financial constraints must cross the demand function obtained in the benchmark model from above to below.

This paper contributes to the finance literature by investigating whether financial markets can eventually become informationally efficient, even when investors lack perfect knowledge of the economic environment. We explore this issue by positing that uninformed investors with no prior domain knowledge employ a reinforcement learning algorithm. In the previous studies in this literature, investors are assumed to still possess some partial knowledge of the economic environment (Bray, 1982) or rely on behavioral rules to update their

investment strategies (Routledge, 1999). Our paper extends the approaches of these papers by considering settings where investors have no prior knowledge and use a novel machine learning tool. In the literature, Bray and Kreps (1987) also develop a model similar to Bray (1982), in which uninformed investors use Bayesian learning to update their beliefs about the relationship between prices and asset payoffs. However, this model is not considered the model with bounded rationality because Bayesian learning is generally regarded as a form of rational learning.

Our paper also contributes to the literature by providing a useful means of calculating approximate solutions for extended models of Grossman and Stiglitz (1980) that are analytically intractable. In the literature, there are only a few papers that have successfully extended the model of Grossman and Stiglitz (1980). As mentioned, Breon-Drish (2015) relaxes the restrictive assumptions regarding payoff distributions, asset supply, and signal structures. However, his approach is still not fully far-reaching. For instance, his solution method typically allows for only a single signal and cannot deal with financial constraints. Regarding financial constraints, Yuan (2005) has made noticeable progress, but this model specifies borrowing constraints in a stylized form, which may not capture a wide range of situations, as mentioned. In our paper, we overcome these difficulties by utilizing the flexibility of a reinforcement learning algorithm that does not require investors to have the full knowledge of the economic environment. In the literature, Bernardo and Judd (2000) incorporate log-normal distributions for asset payoffs and constant-relative-risk-aversion (CRRA) utility functions into Grossman and Stiglitz (1980). But this paper specifies the market-clearing condition in an unconventional form for tractability. Also, we do not consider the log-normal and CRRA setting in our paper because we believe that this setting is more suitable for dynamic environments in which investors can transfer their wealth to the next periods, which is beyond the scope of this paper; see Peress (2004) for related work. Albagli et al. (2024) also relax the assumptions regarding payoff distributions, signal structures, and preferences in a similar spirit to the above-mentioned papers. However, this paper considers a setting with a continuum of investors receiving idiosyncratic signals, as in Hellwig (1980). While applying the reinforcement learning-based approach to this framework is possible, this task will be computationally more intensive because we need to introduce a large number of

heterogeneous agents, which is not the main focus of this paper.

Our paper also contributes to the rapidly growing interdisciplinary field spanning economics, finance, and machine learning. As an example, Calvano et al. (2020) show that computer programs powered by artificial intelligence persistently learn to charge supra-competitive prices in an oligopoly market without communicating with each other. In the context of financial markets, another seminal paper by Dou et al. (2024) develops an information-based asset market model with imperfect competitions and shows how informed traders equipped with artificial intelligence technologies autonomously learn to collude and achieve supra-competitive profits. As demonstrated by these pioneering papers, the merge between machine learning and economics or finance is expected to create numerous research questions and opportunities for researchers. In our paper, we take competitive markets as our main test-bed environment and focus on the classical issues regarding market efficiency in the absence of the rational expectations assumption, considering various extensions of Grossman and Stiglitz (1980).

The paper is organized as follows. In Section 2, for the purpose of a self-contained exposition, we present the model of Grossman and Stiglitz (1980) and the traditional learning approach developed by Bray (1982). In Section 3, we build our main model with reinforcement learning and apply our approach to a number of different settings. Section 4 concludes.

2 The Baseline Model and Traditional Approach

In this section, we first present the repeated version of the static model of Grossman and Stiglitz (1980), which will serve as the baseline model throughout the paper, and then discuss the traditional learning approach used by Bray (1982) to analyze an economy with bounded rationality. The second part can be skipped if readers are more interested in our approach that integrates reinforcement learning into Grossman and Stiglitz (1980).

2.1 Model with Full Rationality

As in Bray (1982) and Routledge (1999), we consider a repeated version of Grossman and Stiglitz (1980). Time is discrete, indexed by $t \in \{0, 1, 2, ...\}$. There are two types of securities:

a risky asset and a risk-free bond. The risky asset yields a payoff x_t at the end of each period t. The time-t payoff x_t, which is independently and identically distributed (i.i.d.) over time, follows a normal distribution with mean μ_x and standard deviation σ_x, that is, $x_t \sim \mathcal{N}(\mu_x, \sigma_x^2)$. Let $\tau_x = 1/\sigma_x^2$. The risk-free bonds are infinitely elastically supplied at a rate normalized to 0.

The supply of the risky asset, denoted by z_t, is randomly determined due to the presence of noise traders who trade the asset for unexpected liquidity reasons. Specifically, the supply of the asset at time t is determined by an i.i.d. normal random variable with mean μ_z and variance σ_z, that is, $z_t \sim \mathcal{N}(\mu_z, \sigma_z^2)$. Let $\tau_z = 1/\sigma_z^2$. The supply of the asset is not publicly observable.

The economy is populated with informed investors and uninformed investors, in addition to noise traders. The measure of informed investors is m_I and the measure of uninformed investors is m_U. As in Bray (1982) and Routledge (1999), we assume that investors cannot carry over their wealth to the next periods. Thus, every investor makes a static decision at each period. All investors have constant-absolute-risk-aversion (CARA) utility functions with a risk-aversion level η.

All informed investors observe the same signal s_t about a future payoff x_t. The signal is given by

$$s_t = x_t + \sigma_s \epsilon_t^s,$$

where $\tau_s = 1/\sigma_s^2$ denotes the precision of the signal and $\epsilon_t^s \sim \mathcal{N}(0,1)$. The error term ϵ_t^s is independent of all other random variables. After observing the signal, each informed investor updates her beliefs about the asset payoff as follows:

$$x_t | s_t \sim \mathcal{N}\left(\frac{\tau_x \mu_x + \tau_s s_t}{\tau_x + \tau_s}, \frac{1}{\tau_x + \tau_s}\right).$$

Let p_t denote the market clearing price of the asset, which is taken as given by all investors. Then each informed investor chooses to buy $q^I(s_t, p_t) \in (-\infty, \infty)$ units of the asset, where

$$q^I(s_t, p_t) = \frac{E[x_t|s_t] - p_t}{\eta Var(x_t|s_t)} = \frac{\tau_x \mu_x + \tau_s s_t - (\tau_x + \tau_s) p_t}{\eta}, \tag{1}$$

which is a standard result in the CARA-Normal setup. We will often just use q_t^I to denote the demand for the risky asset by each informed investor at time t.

Uninformed investors cannot observe the signal s_t. However, the uninformed investors with full rationality can extract information about the asset from its price. Specifically, uninformed investors postulate that the asset price is given by

$$p_t = As_t + B(z_t - \mu_z) + C \tag{2}$$

for some constants A, B, and C that are endogenously determined. Then we define

$$\tilde{p}_t := \frac{p_t - C}{A} = s_t + \rho(z_t - \mu_z) = x_t + \sigma_s \epsilon_t^s + \rho(z_t - \mu_z),$$

where $\rho = B/A$. Here, note that $\tilde{p}_t$ is solely determined by the price p_t. Then, since ϵ_t^s and z_t are independent, after observing the asset price p_t, uninformed investors update their beliefs about asset payoffs as follows:

$$x_t | p_t \sim \mathcal{N}\left(\frac{\tau_x \mu_x + \tau_s \kappa \tilde{p}_t}{\tau_x + \tau_s \kappa}, \frac{1}{\tau_x + \tau_s \kappa} \right), \tag{3}$$

where $\kappa = \frac{\tau_z}{\tau_z + \rho^2 \tau_s}$. As a result, the demand for the risky asset by each uninformed investor is given by

$$q^U(p_t) = \frac{E[x_t | p_t] - p_t}{\eta Var(x_t | p_t)} = \frac{1}{\eta} \left[\tau_x \mu_x + \tau_s \kappa \tilde{p}_t - (\tau_x + \tau_s \kappa) p_t \right], \tag{4}$$

which again comes from the standard result in the CARA-Normal setup. For uninformed investors, a price change has two effects: the substitution effect and the information effect. The substitution effect means that when a price falls, the asset becomes more attractive relative to a risk-free bond. On the other hand, the information effect indicates that a price drop may convey negative information about asset payoffs, making investors more reluctant to purchase the asset.

The total demand for the risky asset is then equal to $m_I q_t^I + m_t^U q_t^U$. Hence, the market-clearing price must satisfy

$$m_I q_t^I + m_U q_t^U = z_t, \tag{5}$$

which allows us to represent the price p_t in terms of s_t and z_t from (1) and (4). In equilibrium, the relationship between p_t and (s_t, z_t) derived from this market-clearing condition must coincide with the relationship initially guessed by uninformed investors, represented in terms A, B, and C in (2). The solutions for this fixed-point problem can be written explicitly and are provided in Appendix A.1.

In equilibrium, prices cannot perfectly reveal the information due to the presence of noise traders. Intuitively, if prices fully reveal all available information, investors would have no incentives to acquire costly information beforehand, leading to a contradiction. While Grossman and Stiglitz (1980) also consider the information acquisition game among investors, we do not address this issue in this paper. In our model, the composition of informed and uninformed investors is exogenously given, as in Bray (1982) and Routledge (1999).

2.2 Traditional Learning Approach

In this section, we provide a brief overview of the approach of Bray (1982), which relaxes the rational expectations assumption in Grossman and Stiglitz (1980) to some extent. The discussions in this section will be useful for comparing the traditional approach proposed by Bray (1982) to our reinforcement learning-based approach.

To begin, we assume that uninformed investors continue to know the basic environment of the economy. However, the uninformed investors are not aware of the precise values of the model parameters, such as the mean of asset payoffs, the precision of a signal, the population of informed investors, and so forth. Nonetheless, since uninformed investors at least know the basic structure of the economy, these investors can reasonably postulate that the price, p_t, and the asset payoff, x_t, are still linearly related to each other as in Grossman and Stiglitz (1980). Specifically, uninformed investors postulate that the price-and-payoff relationship is given in the following linear form:

$$x_t = a + bp_t + c\epsilon_t, \tag{6}$$

where a, b, and c are some constants to be estimated and $\epsilon_t \sim \mathcal{N}(0, 1)$ is an error term. In

the full rational expectations model, the true values of these constants are given by

$$a = \frac{\tau_x \mu_x - \tau_s \kappa \times \frac{C}{A}}{\tau_x + \tau_s \kappa}, \quad b = \frac{\tau_s \kappa}{A(\tau_x + \tau_s \kappa)}, \quad c = \sqrt{\frac{1}{\tau_x + \tau_s \kappa}},$$

which are derived from the property in (3) and the fact that all random variables follow joint normal distributions. Accordingly, we can naturally assume that uninformed investors run a linear regression on past data of prices and asset payoffs to estimate these coefficients, assuming the realized past asset payoffs are publicly observable.

More specifically, we first assume that all uninformed investors initially have the same beliefs about the coefficients a, b, and c, which substantially simplifies our analysis. The initial estimates of these coefficients are denoted by $\hat{a}_0$, $\hat{b}_0$, and $\hat{c}_0$. Then, at the beginning of each time $t \geq 2$, uninformed investors collect the data on prices and asset payoffs from the most recent M periods, denoted by $D_t^M := \{(p_{t-M}, x_{t-M}), ..., (p_{t-1}, x_{t-1})\}$. The idea of using only relatively recent data is reasonable because the data observed in the distant past was generated based on the outdated beliefs that uninformed investors held about the price-and-payoff relationship at that time. The parameter M is hereafter called the memory size.

Uninformed investors then run the ordinary least square regression on this data, using the price as an explanatory variable and the asset payoff as a dependent variable. From this regression, these investors can update their estimates of the coefficients a, b, and c at each time t, denoted by $\hat{a}_t$, $\hat{b}_t$, and $\hat{c}_t$, respectively. For clarification, at time $t = 1$, investors do not perform a regression on single-point data.

Based on these estimates, each uninformed investor with CARA utility sets her demand for the risky asset to

$$q_t^U = \frac{\hat{E}[x_t|p_t] - p_t}{\eta \hat{\sigma}^2(x_t|p_t)} = \frac{\hat{a}_t + \hat{b}_t p_t - p_t}{\eta \hat{c}_t^2}, \tag{7}$$

which is consistent with the formula in (4), where $\hat{E}[\cdot|\cdot]$ and $\hat{\sigma}(\cdot|\cdot)$ indicate the estimated conditional mean and the standard error of the regression, respectively. Here, note that depending on the realized outcomes, the estimated coefficient $\hat{b}_t$ can be potentially larger than 1, while we can easily verify that this outcome does not occur in the true rational expectations

equilibrium of Grossman and Stiglitz (1980). This result means that the positive information effect of a price can dominate the substitution effect in our model, especially when the market is substantially deviated from the true rational expectations equilibrium.

Regarding informed investors, as in Bray (1982), we assume that informed investors are at least aware of the exact forms of asset payoff distributions and signal structures, including all relevant parameter values. Accordingly, the demand for the asset by each informed investor, q_t^I, is still given by the expression in (1) obtained in the baseline model.

The market-clearing price p_t is then determined by the following condition:

$$m_I q_t^I + m_U q_t^U = z_t,$$

which implies

$$p_t = \frac{m_I(\tau_x \mu_x + \tau_s s_t)\hat{c}_t^2 + m_U \hat{a}_t - \eta \hat{c}_t^2 z_t}{m_I(\tau_x + \tau_s)\hat{c}_t^2 + m_U(1 - \hat{b}_t)}. \tag{8}$$

Here, following Bray (1982), we say that the market is in the temporary equilibrium if the price satisfies this market-clearing condition. This equilibrium is called "temporary" because uninformed investors do not accurately infer the relationship between prices and asset payoffs and therefore, the relationship derived in (8) holds only temporarily. For clarification, in the expression in (8), we ignore a measure-zero set of events in which $m_I(\tau_x + \tau_s)\hat{c}_t^2 + m_U(1 - \hat{b}_t)$ collapses to zero.

We conduct numerical simulations to examine whether the outcomes of this economy converge to the true rational expectations equilibrium, pinned down in Section 2. While Bray (1982) provides a sufficient condition for convergence by imposing some additional restrictive assumptions on the prior knowledge of uninformed investors, we are not aware of a more general proof for convergence in our context, which we leave for future research.

Figure 1 depicts the convergence results of this model by choosing reasonable parameter values. The top three panels in the figure plot the trajectories of estimated values of a, b, and c, respectively, over time. Recall that these coefficients represent the uninformed investors' beliefs about the price-and-payoff relationship. The results show that all these coefficients converge to their corresponding values arising in the true rational expectations equilibrium. Moreover, the graphs indicate that the uninformed investors' beliefs about the price-and-

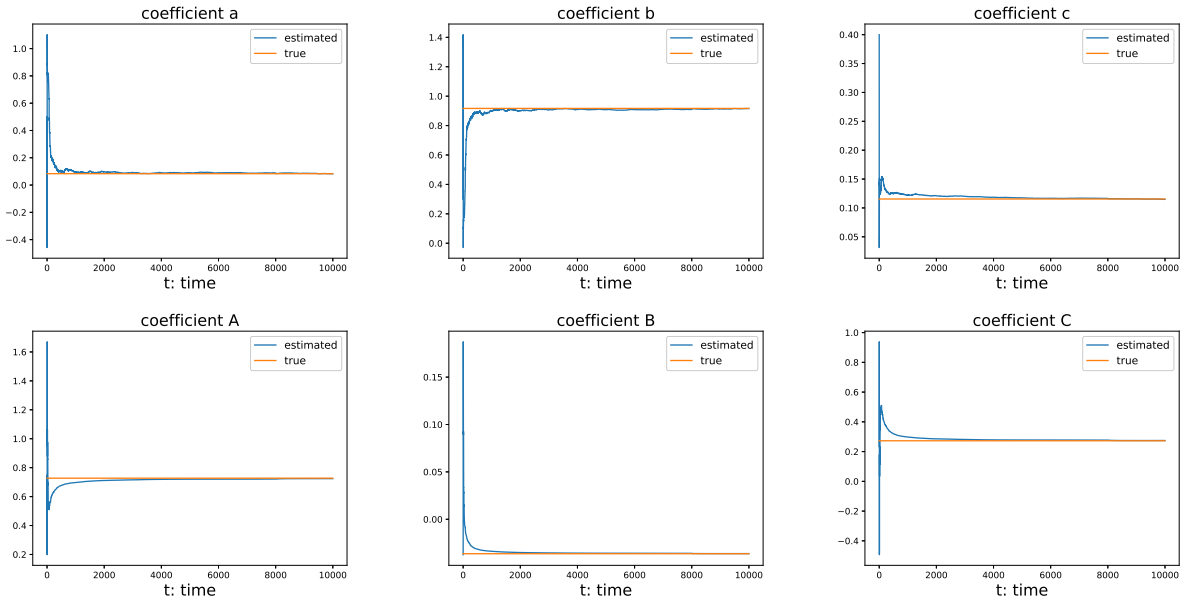

Figure 1: This figure illustrates the long-run behavior of our model outcomes. The top three panels plot whether the estimated coefficients of a, b, and c converge to their corresponding coefficients arising in the true rational expectations equilibrium. In these panels, the blue lines represent the estimated coefficients and the orange lines indicate the true coefficients. The bottom three panels plot whether the estimated coefficients of A, B, and C converge to their corresponding coefficients emerging in the true rational expectations equilibrium. Here, these coefficients are estimated by econometricians, not by uninformed investors directly. In these panels, the blue lines plot the estimated coefficients and the orange lines represent the true coefficients. The parameter values chosen in this example are $\mu_x = 1$, $\sigma_x = 0.2$, $\sigma_s = 0.1$, $\mu_z = 0$, $\sigma_z = 2$, $\eta = 5$, $m_I = 1$, and $m_U = 2$. The initial estimates of a, b, and c are $\hat{a}_0 = 0.5$, $\hat{b}_0 = 0.3$, and $\hat{c}_0 = 0.4$. The memory size is set to $M = 8000$. The simulation ends at $T = 10000$.

payoff relationship exhibit big swings in early stages. Nonetheless, the result shows that the market eventually identifies the true relationship between prices and asset payoffs in the long run.

The bottom three panels in Figure 1 plot the estimated coefficients A, B, and C, which determines the relationship between p_t (price) and the pair of s_t (signal) and z_t (supply), as expressed in (2). Specifically, in this model, uninformed investors do not directly estimate these coefficients by running this form of regression. In fact, this estimation will not be useful for uninformed investors because the current private signal and supply are not publicly observable, even if we assume that the past data on these variables is publicly observable. Nonetheless, as econometricians, we can estimate the values of A, B, and C by running this form of regression. This task enables us to examine whether this more refined relationship

between prices and the pair of asset payoffs and supplies also converges to its corresponding counterpart arising in the true rational expectations equilibrium. Our numerical results again show that all these coefficients converge to their corresponding values obtained in the true rational expectations equilibrium. This result strengthens our assertion that while uninformed investors do not have any prior knowledge of the economic environment, the market eventually become informationally efficient—to the same extent that is achieved in the economy with full rationality.

3 Bounded Rationality and Reinforcement Learning

In this section, we assume that uninformed investors have no prior knowledge of the economic environment, while informed investors are still aware of the exact form of payoff distributions and signal structures as in the previous section. In this setting, due to the lack of domain knowledge, uninformed investors cannot correctly infer the functional form of the price-and-payoff relationship. To overcome this limitation, uninformed investors use a novel machine learning tool, that is, a reinforcement learning algorithm, to learn optimal investment strategies. The key feature of this approach is that uninformed investors directly adjust their beliefs about optimal investment strategies based on realized data instead of estimating the precise relationship between prices and asset payoffs as done in Bray (1982).

The two widely used reinforcement learning algorithms are a Q-learning algorithm (Watkins, 1989; Watkins and Dayan, 1992; Mnih et al., 2015; Silver et al., 2017) and a policy-gradient algorithm (Williams, 1992; Sutton et al., 1999; Schulman et al., 2015). See also Sutton and Barto (2018) for a textbook-style treatment of these two distinct algorithms. Briefly put, a Q-learning algorithm attempts to learn the value function for all state variables and choice variables through trial and error and then exploits the accumulated knowledge to take an optimal action that maximizes the value function. On the other hand, a policy-gradient algorithm directly attempts to learn an optimal policy function by adjusting the beliefs about an optimal policy function in a direction that marginally increases the average utility calculated from realized data ex post. In the machine learning field, the Q-learning algorithm is more widely used in settings where state and choice variables are discrete and

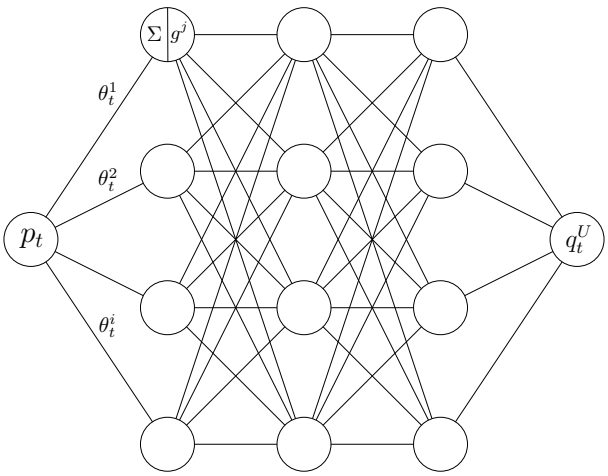

Figure 2: This figure plots a deep neural network that represents the investment policy function of uninformed investors. The input variable is a price p_t. The output variable, $q^U(p_t)$, is the demand for the risky asset. The weights on the edges are denoted by $\boldsymbol{\theta}_t = \{\theta_t^i\}_i$. The nodes representing bias terms are omitted in this figure. The symbol Σ denotes the weighted sum used in deep learning models and a function g^j denotes the activation function applied at the j-th layer.

the policy-gradient algorithm is considered more suitable for settings with continuous state and action spaces. As such, in this paper, we will use the policy-gradient algorithm. In the literature, Calvano et al. (2020) and Dou et al. (2024) use Q-learning by discretizing state and action spaces, while Buehler et al. (2019) employ a policy-gradient algorithm to study optimal hedging strategies for options.

In the subsequent sections, we will describe the policy-gradient algorithm in more detail in our specific context instead of describing this algorithm in generic settings. Since our model is basically a repeated version of a static model and does not consider mixed strategies, the policy-gradient algorithm can be described in a very simple manner.

3.1 Canonical Environment

In this section, we first apply our reinforcement learning-based approach to the canonical environment of Grossman and Stiglitz (1980). As in Section 2.2, this task enables us to examine whether our model outcomes continue to converge to the true rational expectations equilibrium. We will also consider several extensions of this canonical model in later sections.

3.1.1 Model with Reinforcement Learning

As mentioned above, the demand for the risky asset by each informed investor, q_t^I, is still given by the expression in (1). Meanwhile, each uninformed investor posits that an optimal investment strategy is a function of a price and some unknown parameters:

$$q_t^U = q^U(p_t; \boldsymbol{\theta}_t),$$

where $\boldsymbol{\theta}_t = \{\theta_t^i\}_i$ denotes the set of the unknown parameters that must be estimated over time. Moreover, following recent trends in machine learning, each uninformed investor uses a deep neural network to represent the unknown policy function $q^U(p; \boldsymbol{\theta})$.

As shown in Figure 2, a neural network consists of nodes (or neurons), arranged in distinct layers, and edges connecting the nodes in adjacent layers. Moreover, a certain weight is assigned to each edge of the network and a specific activation function is applied at each intermediate layer. The above-mentioned parameters, $\boldsymbol{\theta}_t = \{\theta_t^i\}_i$, indicate these weights on the edges. We choose activation functions appropriately. Then, for each input value, the weights and activation functions of the network produce the final output of the network, following the standard feedforward neural network operations; see, for instance, Goodfellow et al. (2016) for details. In our model, the input variable is a price p_t and the output variable, $q^U(p_t; \boldsymbol{\theta}_t)$, is the demand for the risky asset by each uninformed investor as mentioned before.

To proceed further, we assume that all uninformed investors set the initial values of the weights, that is, $\boldsymbol{\theta}_0 = \{\theta_0^i\}_i$, to be identical. This assumption allows us to treat all uninformed investors as a single representative investor. We typically set the initial values of the weights arbitrarily, following the convention in machine learning. We will soon discuss how the parameters, $\{\theta_t^i\}_i$, are updated over time.

Given that uninformed investors make investment decisions in this manner, the market-clearing price p_t is the price that satisfies the following condition:

$$m_I q^I(p_t) + m_U q^U(p_t; \boldsymbol{\theta}_t) = z_t,$$

where $z_t \sim \mathcal{N}(\mu_z, \sigma_z^2)$ denotes the asset supply as in the baseline model. We solve for

the market-clearing price numerically, using a divide-and-conquer method. One technical concern is that the above equation may not have a solution because the demand function of uninformed investors can be highly nonlinear, especially in early stages. If such a case occurs at a certain date, following the convention in the market microstructure literature, we assume that no trades occur at that date and move forward to the next date; see, for instance, Rostek and Yoon (2023). In other words, what happens at this particular date will be ignored by uninformed investors in their learning procedures. In fact, although we formally adopt this ad-hoc rule to deal with unusual cases, we have not encountered such cases in our numerical simulations under almost all reasonable parameter values.

Policy-gradient algorithm: We now describe the policy-gradient algorithm that is used to update the weights, $\boldsymbol{\theta}_t = \{\theta_t^i\}_i$, associated with the investment policy function q^U. We first assume that uninformed investors update the weights at every N periods, where N is a fixed integer. That is, uninformed investors update their investment policy functions only at the dates equal to $t = lN$, where $l \in \{1, 2, 3, ...\}$. We call N the update interval length. We adopt this infrequent learning procedure because updating the investment policy function at every period would be computationally inefficient as this procedure will repeatedly use largely overlapped data at every period.

Next, as in Section 2.2, uninformed investors update their investment policy functions using the data from only the most recent M periods. Let $D_t^M := \{(p_s, x_s)\}_{s=t-M}^{t-1}$ denote this data set as before. While the memory size M can be chosen to be larger than the update interval length N, we simply set the memory size equal to the update interval length. This choice is reasonable, particularly in our model, because the data generated before the previous updating date were produced under the outdated beliefs about optimal investment strategies. Therefore, using such old data may not improve the quality of learning.

Now, at the beginning of each updating date $t = lN$, each uninformed investor calculates the average partial derivative (that is, gradient) of her realized utility over the latest data set D_t^M with respect to each unknown parameter θ^i:

$$\Delta_t^i := \frac{1}{M} \sum_{s=t-M}^{t-1} \frac{\partial}{\partial \theta^i} u\left((x_s - p_s) q^U(p_s; \boldsymbol{\theta_{t-1}})\right). \tag{9}$$

Here, denoting the investment policy function used in this expression by $\boldsymbol{\theta_s}$ is unnecessary because the investment policy function is updated only at the dates equal to $t = lN$. Intuitively, the gradient Δ_t^i approximately measures the rate of changes in the investor's expected utility, if (i) she adjusts the parameter θ_{t-1}^i by an infinitesimally small amount while fixing all other parameters and (ii) all other investors keep their investment strategies. Accordingly, each uninformed investor reasonably (or heuristically) updates each parameter θ_{t-1}^i using the following update rule:

$$\theta_t^i \leftarrow \theta_{t-1}^i + \alpha_t \Delta_t^i, \tag{10}$$

where α_t is a so-called learning rate, which we will explain in more detail later. In other words, the investor adjusts the parameter θ_{t-1}^i in a direction that marginally increases her average utility calculated from the past data. After updating the parameters θ_{t-1}^i for each i at time $t = lN$, investors keep using these new parameters until the next updating date.

While this procedure clearly describes the key idea of the policy-gradient algorithm, we add one more procedure to this algorithm, following the conventions. Specifically, we divide the data set D_t^M into multiple batches of equal size and update the weights of the policy network, $\boldsymbol{\theta_{t-1}} = \{\theta_{t-1}^i\}_i$, applying the procedures in (9) and (10) over different batches separately. We then repeat this one cycle of learning procedure over the same data set D_t^M multiple times. The number of repetitions of this procedure is called the number of epochs. Both the batch size and the number of epochs are chosen appropriately.

Lastly, the learning rate α_t controls the speed at which investors update the parameters $\boldsymbol{\theta_{t-1}} = \{\theta_{t-1}^i\}_i$. If we choose a high learning rate, the learning process will be accelerated, but we may also increase the risk that the updated policy function would overshoot the true policy function. Conversely, if we choose a low learning rate, we can reduce the chance of the policy function diverging, but we may need to wait a long time or even fail to find an optimal policy. In general, we initially set the learning rate to an appropriately high level and then decrease this rate gradually over time. This specification allows for faster learning in early stages, followed by more delicate adjustments as the policy function converges toward the true optimal solution. This type of a parameter is commonly called a hyperparameter in the machine learning field.

3.1.2 Results

In this section, we discuss the model results. Figure 3 presents the results of this model, using the same parameters as those used in Figure 1. In this simulation, the number of hidden layers in the network is 5 and each hidden layer has 100 nodes. All other parameter values, such as the learning rate, the update interval length, and so forth, are specified in the figure.

The top four panels in Figure 3 depict how the investment policy function, $q_t^U(p) = q^U(p; \boldsymbol{\theta}_t)$, viewed as a function of price p, evolves over time. In each panel, the blue line represents the investment policy function trained in our model, while the orange line indicates the true investment policy function arising in the rational expectations model, which is obtained in (4). The result shows that the investment policy function in our model converges to the true solution in the long run, even though the initial investment policy function is chosen arbitrarily. The left and right boundaries of the price interval in the graphs are chosen to ensure that 99% of the realized data points are included within this interval. Also, as shown in the top-middle panel, the investment policy function converges to the true solution relatively faster in the region of intermediate price levels. This result is intuitive because investors would struggle to learn the optimal investment policy over the region of extreme price levels, as these investors are less likely to encounter data that falls in such an extreme region.

The bottom three panels in Figure 3 illustrate how the estimated values of coefficients A, B, and C evolve over time. Specifically, in this task, as in Section 2.2, we run a linear regression to estimate these coefficients, using the past data on prices, private signals, and asset supplies, from the perspective of econometricians. Then, as the graphs show, these estimated coefficients also converge to their corresponding values in the true rational expectations equilibrium in the long run. This result strengthens our assertion that financial markets tend to eventually achieve the maximum possible level of informational efficiency even if uninformed investors posses no prior knowledge of the economic environment. Before proceeding further, notice that the trajectories of estimated A, B, and C exhibit more oscillations around their corresponding true values, compared to the regression-based results in Figure 1. Although this finding is intriguing, we currently do not have a satisfactory explanation for this result.

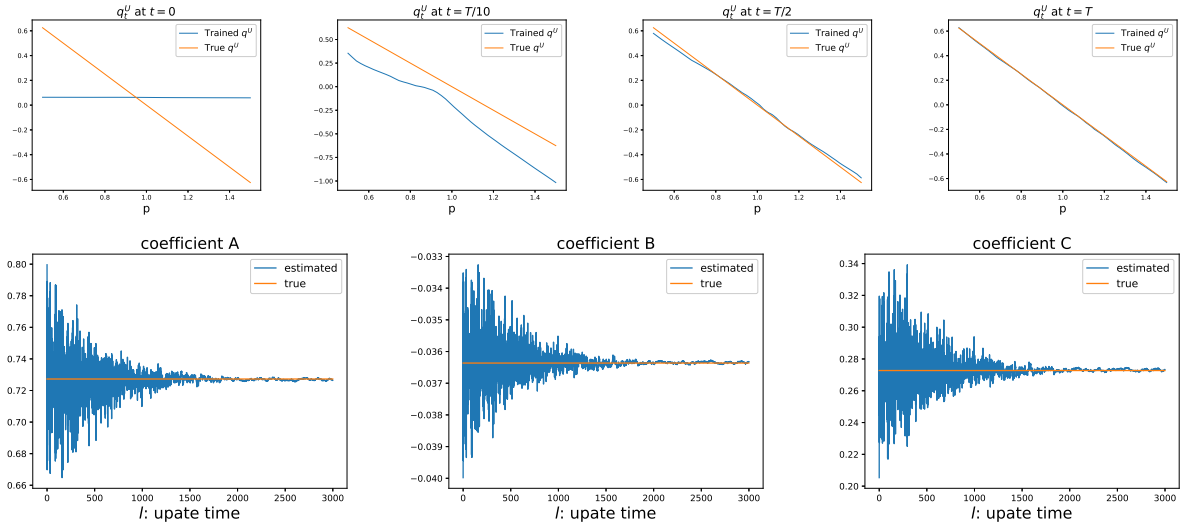

Figure 3: This figure depicts the results of our model when we apply the reinforcement learning algorithm to the canonical environment of Grossman and Stiglitz (1980). The top four panels show how the investment policy function $q_t^U = q^U(p; \boldsymbol{\theta}_t)$, as a function of price p, evolves over time, where T denotes the time at which the simulation ends. In each panel, the blue curve represents the trained investment policy function and the orange line represents the true policy function in the rational expectations model. The left and right boundaries of the graphs are chosen to ensure that 99% of the realized data points are included within that interval. The bottom three panels depict how the estimated values of coefficients A, B, and C evolve over time. Here, we plot these trajectories only at the update dates, that is, at the dates equal to $t = lN$, where $l \in \{1, 2, 3, ...\}$. The model parameter values are $\mu_x = 1$, $\sigma_x = 0.2$, $\sigma_s = 0.1$, $\mu_z = 0$, $\sigma_z = 2$, $\eta = 5$, $m_I = 1$, and $m_U = 2$. The update interval is $N = 2000$. The number of hidden layers is 5. The number of nodes in each hidden is 100. The batch size is 64. The number of epochs is 10. The learning rate is initially set to 10^{-4} and then is reduced by a factor 0.999 every 100 epochs. The ReLU function is used as an activation function. The simulation ends after 3000 updates in this example.

Here, note that we obtain this convergence result only when the hyperparameters (or learning-related parameters), such as the learning rate, the number of hidden layers, the number of nodes in each layer, among others, are appropriately chosen through trial and error. For instance, if we set the learning rate to 0, the model will never find the true rational expectations equilibrium. On the other hand, if we set the learning rate to 1, the model outcomes will never reach a stable state. Unfortunately, we are not aware of a systematic method to identify the hyperparameters that ensure convergence. In the literature, a general convergence result is available only for individual Market decision problems without any interactions among agents Watkins and Dayan (1992). Nonetheless, according to our numerical simulations, under most reasonable parameter values, our model tends to converge

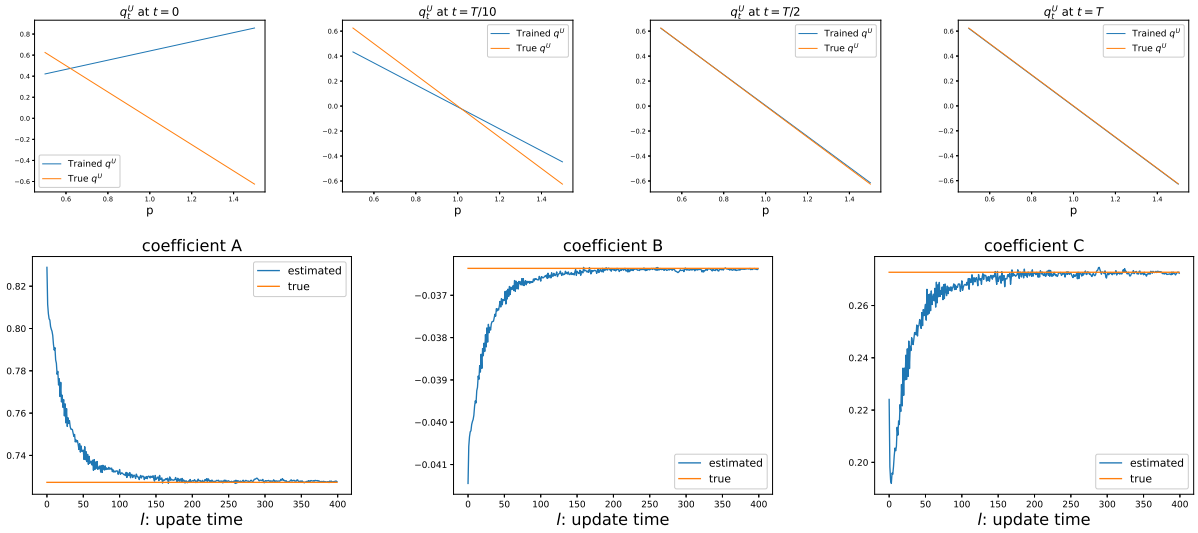

Figure 4: This figure depicts the results when we use a neural network without any hidden layers, leveraging the fact that the true investment strategy is a linear function of a price in this canonical environment. In the top four panels, the left and right boundaries of the graphs are chosen to ensure that 99% of the realized data points are included within that interval. In this simulation, the learning rate is initially set to 0.001 and then is reduced by a factor 0.999 every 50 epochs. All other parameter values are the same as those in Figure 3. The simulation ends after 600 updates.

to the true rational expectations as long as we choose the hyperparameters appropriately. Further results can be provided upon request.

In addition, in the above work, one may also wonder what would happen if we use a neural network without any hidden layers because this neural network of no hidden layers essentially represents a linear function and we know that the true investment policy function is a linear function of a price in Grossman and Stiglitz (1980). Although uninformed investors in the model are not aware of this fact, which is the crucial assumption of our setting, it is worth examining whether the model outcomes would converge to the true rational expectations equilibrium if uninformed investors use this simple neural network in their reinforcement learning process. Alternatively speaking, this setting is closer to Bray (1982), who also assumes that uninformed investors are aware of the linear structure of the model. However, in our setting, these investors still use reinforcement learning to learn optimal investment strategies, while uninformed investors in Bray (1982) use a regression approach to estimate the price-and-payoff relationship.

Figure 4 plots the results of this work. As expected, the model outcomes again converge

to the outcomes that emerge in the true rational expectations equilibrium. Moreover, in this setting, once uninformed investors have learned the optimal investment policy function precisely enough in the region of intermediate price levels, these investors do not need additional data to learn the optimal investment policy function in the region of extreme price levels due to the linearity assumption. Accordingly, as shown in the bottom three panels, the model outcomes tend to converge to the true rational expectations equilibrium relatively faster than in the previous case where uninformed investors are unaware of the linear structure of the model and thus use a highly complex deep neural network to represent their investment policy functions.

3.2 Non-Standard Environment

In what follows, we consider several extensions of Grossman and Stiglitz (1980), in which a linear relationship between prices and asset payoffs no longer holds. This task has two goals. First, we apply our approach to an extended model that is still analytically tractable and examine whether our model outcomes still converge to the true rational expectations equilibrium in the absence of a linear structure. Second, we extend our approach to more general settings that are not analytically tractable and provide the approximate solutions for the true rational expectations equilibrium of these models by obtaining the long-run limit of our model without rational expectations.

In the literature, Breon-Drish (2015) extends the model of Grossman and Stiglitz (1980) by relaxing restrictive assumptions on payoff distributions, asset supply, and signal structures to some extent. While the new solution developed by Breon-Drish (2015) covers a broad range of settings, in this paper, we consider one particular setting that is strongly motivated from empirical findings. Specifically, we consider a model in which asset payoffs follow a mixture of normal distributions, exhibiting negative skewness and heavy tails. These two patterns are the widely supported stylized facts (Cont, 2001; Gabaix, 2009; Kelly and Jiang, 2014). Also, Ang and Timmermann (2012) show that a regime-based Markov chain model, in which asset payoffs follow a mixture of normal distributions, explain many additional stylized facts besides negative skewness and heavy tails. Moreover, the solution technique invented by Breon-Drish

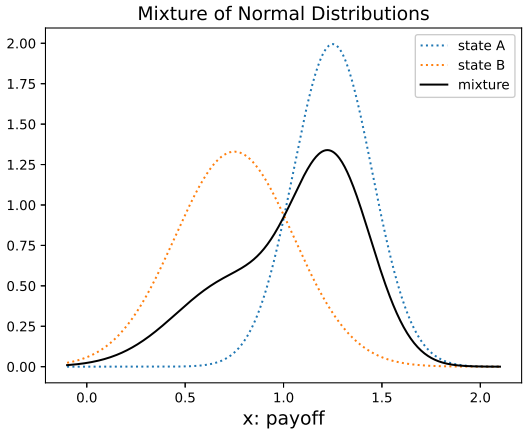

Figure 5: The black curve plots a typical pattern of the distribution of a mixture of normal random variables. The blue dotted curve plots the distribution of a normal random variable under state A. The orange dotted curve plots the distribution of a normal random variable under state B.

(2015) is applicable to this extension. In this regard, we take this setting as another test bed and examine the market efficiency issue in the absence of rational expectations. Regarding the second goal, we consider models with multiple signals and financial constraints.

3.2.1 Asset Payoffs with Negative Skewness and Heavy Tails

Model Setup: As mentioned, we now assume that asset payoffs follow a mixture of normal distributions. Specifically, in this model, the asset payoff x_t is drawn randomly according to the following law:

$$x_t = \begin{cases} x_t^A & \text{with probability } \lambda \\ x_t^B & \text{with probability } 1-\lambda, \end{cases}$$

where $x_t^A \sim \mathcal{N}(\mu_x^A, \sigma_x^A)$ and $x_t^B \sim \mathcal{N}(\mu_x^B, \sigma_x^B)$ for some constants μ_x^A, σ_x^A, μ_x^B, and σ_x^B. That is, in this model, a state $\pi \in \{A, B\}$ is realized first and then the asset payoff x_t is determined, depending on the state. State A occurs with a probability λ and state B occurs with the remaining probability. At each time, the state is realized independently of any other events. We also use $\pi_t \in \{A, B\}$ to denote the realized state at time t. For each state $\pi \in \{A, B\}$, let $\tau_x^\pi = 1/(\sigma_x^\pi)^2$.

Throughout, we interpret state A as a good state and state B as a bad state by choosing the parameters (μ_x^A, σ_x^A) and (μ_x^B, σ_x^B) such that $\mu_x^A > \mu_x^B$ and $\sigma_x^A < \sigma_x^B$. That is, the good

state represents the state with a high mean payoff and low variance, while the bad state represents the state with a low mean payoff and high variance. A typical pattern of the distribution of a mixture of normal random variables is shown in Figure 5. As indicated by this figure, a mixture of normal distributions is known to exhibit negative skewness and heavy tails under the above-mentioned parameter condition (Ang and Timmermann, 2012).

In the model, informed investors still receive a single signal s_t about the asset payoff in the following form:

$$s_t = x_t + \sigma_s \epsilon_t^s,$$

as in the canonical model. In the next section, we consider a setting in which informed investors receive two signals about the prevailing state and the asset payoff separately.

Interestingly, we can pin down the solution of this model analytically, following the method invented by Breon-Drish (2015) as mentioned above, provided that we impose the rational expectations assumption. For self-contained exposition, we briefly present how to solve this model analytically.

To begin, note that the cumulative distribution function of a payoff x, denoted by $F(x)$, is given by

$$F(x) = \lambda \Phi\left(\frac{x - \mu_x^A}{\sigma_x^A}\right) + (1-\lambda)\Phi\left(\frac{x - \mu_x^B}{\sigma_x^B}\right), \tag{11}$$

because the payoff follows a mixture of normal distributions, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. Then Breon-Drish (2015) shows that the optimal demand for the asset by each informed investor is equal to

$$q_t^I(s_t, p_t) = \frac{\tau_s s_t - G_I(p_t)}{\eta}, \tag{12}$$

where

$$G_I(p) = (g_I')^{-1}(p) \quad \text{and} \quad g_I(v) = \log\left(\int_{-\infty}^{\infty} \exp\left(vy - \frac{1}{2}\tau_s y^2\right) dF(y)\right). \tag{13}$$

We can compute the integral in the above expression using the Gauss-Hermite quadrature method, given that the distribution function F is represented by a sum of two normal distribution functions as shown above.

Regarding the investment problems of uninformed investors, we define $\tilde{p}_t$ as follows:

$$\tilde{p}_t = s_t + \rho(z_t - \mu_z),$$

where $\rho = -\frac{\eta}{m_I \tau_s}$, as done in Grossman and Stiglitz (1980). We will later see that $\tilde{p}_t$ serves as a sufficient statistic in this model. We also let $\sigma_U^2 = \sigma_s^2 + \rho^2 \sigma_z^2$ and $\tau_U = 1/\sigma_U^2$. Then, Breon-Drish (2015) shows that the optimal demand of each uninformed investor is given by

$$q_t^U(\tilde{p}_t, p_t) = \frac{\tau_U \tilde{p}_t - G_U(p_t)}{\eta}, \tag{14}$$

where

$$G_U(p) = (g_U')^{-1}(p) \quad \text{and} \quad g_U(v) = \log\left(\int_{-\infty}^{\infty} \exp\left(vy - \frac{1}{2}\tau_U y^2\right) dF(y)\right). \tag{15}$$

Here, note that uninformed investors cannot construct $\tilde{p}_t$ directly from the signal s_t and asset supply z_t, because these variables are not publicly observable. However, we will soon see that $\tilde{p}_t$ can be constructed solely from the price p_t without having to know the realized values of the signal and asset supply.

The market-clearing condition is then described as

$$m_I q^I(s_t, p_t) + m_U q^U(\tilde{p}_t, p_t) = z_t,$$

which can be rewritten as

$$\frac{1}{\eta}(m_I \tau_s + m_U \tau_U)\tilde{p}_t - \mu_z = \frac{1}{\eta}[m_I G_I(p_t) + m_U G_U(p_t)],$$

due to the fact that $\rho = -\frac{\eta}{m_I \tau_s}$ and $\tilde{p}_t = s_t + \rho(z_t - \mu_z)$. From this relationship, we can pin down the market-clearing price in terms of the signal s_t and the asset supply z_t, using the inverse function of $G(p) := \frac{1}{\eta}[m_I G_I(p) + m_U G_U(p)]$. Moreover, using this relationship, we can also express $\tilde{p}_t$ in terms of only the price p_t. This observation finally confirms that $\tilde{p}_t$ serves as a sufficient statistic in this model by summarizing the payoff-relevant information embedded

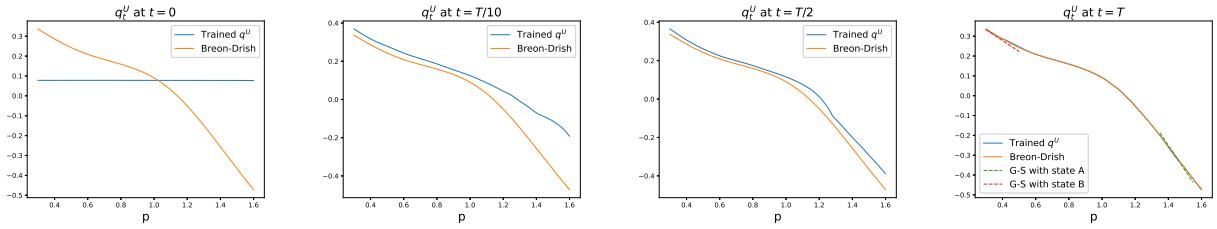

Figure 6: This figure depicts the results of the model in which asset payoffs follow a mixture of normal distributions. Specifically, the four panels show how the investment policy function of uninformed investors, $q^U(p; \boldsymbol{\theta}_t)$, evolves over time. In each panel, the orange curve plots the investment policy function arising in Breon-Drish (2015), in which full rationality is imposed, and the blue curve plots the investment policy function in our model without the rationality assumption. In the last panel, the green (red) dotted line plots the investment policy function in the canonical model of Grossman and Stiglitz (1980) with a fixed state set to A (B). The left and right boundaries of the graphs are chosen to ensure that 99% of the realized data points are included within that interval. The parameter values are $(\mu_x^A, \sigma_x^A) = (1.2, 0.2)$, $(\mu_x^B, \sigma_x^B) = (0.9, 0.3)$, $\lambda = 0.6$, $\sigma_s = 0.1$, $\mu_z = 0$, $\sigma_z = 2$, $\eta = 5$, $m_I = 1$, and $m_U = 2$. The update interval is set to $N = 20000$. The batch size is 2000. The number of epochs is 20. The learning rate is initially set to 10^{-4} and then is reduced by a factor 0.99 every 100 epochs. The number of hidden layers is set to 5. The number of nodes on each hidden layer is set to 100. The ReLU function is used as the activation function. The simulation ends after 400 updates in this example.

in the price. As a result, uninformed investors can determine their demand for the asset solely from the price, following the formula in (14), completing equilibrium characterization.

Next, notice that when uninformed investors have no prior knowledge of the economic environment, we need not make any changes in their learning procedures from those described in Section 3.1.1. This statement is easy to understand because the learning algorithm designed by agents who are not aware of the economic environment cannot be designed depending on the environment. Moreover, informed investors behave the same way as in the model with the rational expectations. As such, since we have already solved the individual problems of informed investors above, we can readily incorporate the reinforcement learning into the model with a mixture of normal distributions.

Results: Figure 6 plots our model results. Specifically, the orange curve in each panel denotes the true investment policy function of uninformed investors in the rational expectations equilibrium. As expected, the demand function of uninformed investors is no longer linear in the price. Notice that the investment strategy exhibits a tilted hump-shaped pattern. We can understand this pattern as follows. When the price is high, say, around 1.45, which is close to μ_x^A, uninformed investors would believe that the current economic state is

highly likely to be the good state. As such, the demand function of uninformed investors is close to the demand function obtained in the canonical model of Grossman and Stiglitz (1980) with a fixed state set to A. Meanwhile, when the price starts to fall, the demand function of uninformed investors becomes less steeper. This result is intuitive because when the price is at an intermediate level, uninformed investors struggle more to deduce the current economic state. Due to this elevated uncertainty about the current state, uninformed investors respond less elastically when the price falls.

However, when the price is sufficiently low, say, around 0.4, uninformed investors would believe that the current economic state is highly likely to be the bad state. Thus, the demand function of uninformed investors is close to the demand function arising in the canonical model of Grossman and Stiglitz (1980). Interestingly, when the price is at an intermediate level, say, around 0.7, the slope of the demand function can even less steeper than the demand function obtained when the price has a sufficiently low level. This result can be understood using the same reasoning discussed above. That is, when the price falls substantially, the uncertainty level about the current economic state is reduced and therefore, uninformed investors increase their demand relatively more elastically when the price falls.

Moreover, for clarification, note that when the price is extremely high, uninformed investors believe that the realized state is more likely to be the bad state because the payoff variance in the bad state is larger than that in the good state. However, such a region does not show up in our graph because realized data points rarely fall in that region.

Regarding the convergence result, the four panels in the figure show that the outcomes of the model without the rationality assumption converge to the true rational expectations equilibrium. These results give us strong confidence that financial markets eventually become informationally efficient even if we consider more complex environments and abandon the rational expectations assumption at the same time. In other words, these results suggest that uninformed investors with no prior domain knowledge can effectively capture the nonlinear relationship between prices and payoffs by utilizing reinforcement learning algorithms, although these investors do not directly attempt to infer this relationship from data. While these results are obtained from simulation experiments, we believe that these findings significantly improve of our understanding of the role of reinforcement learning in driving financial

markets to achieve informational efficiency.

In the subsequent sections, we will consider several further extensions that are not analytically tractable. As these models do not permit analytic solutions, we are not able to examine whether the market outcomes continue to converge to the true rational expectations equilibrium. Nonetheless, relying on the previous results, we can expect that the market outcomes will still converge to a limit, if exists, which can be regarded as the true rational expectations equilibrium. In this regard, our approach can be used not only for investigating the market efficiency issue but also for obtaining the approximate solutions to those models that are not analytically solvable.

3.2.2 State-Dependent Model with Separate Signals

An extension we can readily apply our approach to is a state-dependent economy, which we have considered in the previous section, in which informed investors receive two signals about the unobservable state and the asset payoff separately. Specifically, we assume that each informed investor first receives a signal h_t, which takes a value of either A or B. The probability of the signal being correct is assumed to be γ:

$$\gamma = \Pr(h_t = A|\pi_t = A) = \Pr(h_t = B|\pi_t = B),$$

which is assumed to be larger than $\frac{1}{2}$. We further assume that the signal h_t is independent of any other events and independently drawn over time.

After receiving the signal about the state, each informed investor receives an additional signal s_t about the asset payoff x_t in the same form as before:

$$s_t = x_t + \sigma_s \epsilon_t^s.$$

The informed investors then make the investment decisions using these two signals h_t and s_t together. The other parts of the model remain unchanged from the previous setting with a single signal. In fact, we may consider a more general state-dependent economy, in which the supply of the asset, the population of informed investors, and the risk-aversion level also

change, depending on the state. But we do not seek to consider these more general settings for the sake of brevity.

An analytical solution for the extended model with separate signals is not available at the moment of writing this paper. In particular, the approach of Breon-Drish (2015) accommodates only a single signal. However, we can still solve this model using the reinforcement learning approach because (i) the individual problem of informed investors is still tractable and (ii) the learning procedure of uninformed investors does depend on a particular environment.

To proceed, note first that the individual problem of informed investors can be similarly solved as before. Let $\tilde{\lambda}_t(h_t)$ denote the posterior beliefs of informed investors regarding the probability of the state being A, conditional on the realized signal h_t. Specifically, depending on the realized signal $h_t \in \{A, B\}$, the posterior beliefs are respectively given by

$$\tilde{\lambda}_t(A) = \frac{\lambda\gamma}{\lambda\gamma + (1-\lambda)(1-\gamma)} \quad \text{and} \quad \tilde{\lambda}_t(B) = \frac{\lambda(1-\gamma)}{\lambda(1-\gamma) + (1-\lambda)\gamma}.$$

Given that $\gamma > \frac{1}{2}$, we can easily see that $\tilde{\lambda}_t(B) < \lambda < \tilde{\lambda}_t(A)$.

Then, before receiving the additional signal s_t about a payoff x_t, each informed investor believes that a payoff x_t will be drawn from the distribution $F(x; \tilde{\lambda}_t)$ that is given by

$$F(x; \tilde{\lambda}_t) = \tilde{\lambda}_t \Phi\left(\frac{x - \mu_x^A}{\sigma_x^A}\right) + (1 - \tilde{\lambda}_t)\Phi\left(\frac{x - \mu_x^B}{\sigma_x^B}\right),$$

which is the same as the expression in (11) except that the prior λ is replaced by the posterior beliefs $\tilde{\lambda}_t$. Accordingly, the demand function of each informed investor, that is, q_t^I, is still given by (12) with the distribution function F in (13) being replaced by $F(x; \tilde{\lambda}_t)$.

Uninformed investors behave the same way as in the canonical model without rational expectations, as mentioned multiple times before. For clarification, in this model, uninformed investors do not need to collect the data on the past realized states even if the past states become publicly observable ex post, because the current state does not directly affect the utility of investors.

Moreover, note that we are not able to calculate the true rational expectations equi-

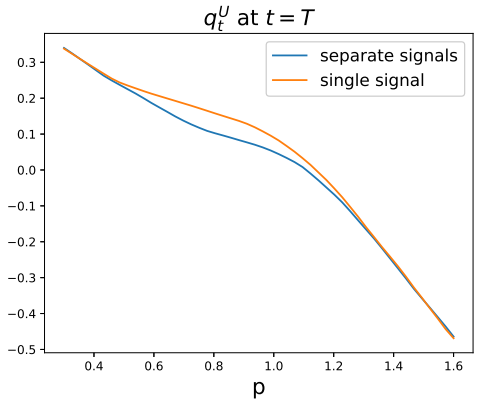

Figure 7: This figure depicts the results of the model with separate signals. The blue curve plots the long-run limit of the investment policy functions of uninformed investors. The orange curve plots the investment policy function of uninformed investors in the previous economy with a single signal. The left and right boundaries are chosen to ensure that 99% of realized data points lie within this interval. The parameter values are $(\mu_x^A, \sigma_x^A) = (1.2, 0.2)$, $(\mu_x^B, \sigma_x^B) = (0.9, 0.3)$, $\lambda = 0.6$, $\sigma_s = 0.2$, $\mu_z = 0$, $\sigma_z = 5$, $\lambda = 0.9$, $\eta = 5$, $m_I = 1$, and $m_U = 2$. The update interval is set to $N = 20000$. The batch size is set to 2000. The number of epochs is set to 20. The learning rate is initially set to 10^{-4} and then is reduced by a factor 0.99 every 50 epochs. The number of hidden layers is set to 5. The number of nodes of each hidden layer is set to 100. The activation function is set to the ReLU function. The simulation ends when the moving average of Δ_l^θ, which measures the changes in the policy network weights, over 20 consecutive update intervals falls below 10^{-5}.

librium in this model. Thus, we cannot use the error between the true solution and our model outcome as a stopping criterion for our simulation. Instead, following conventions in machine learning, we calculate the difference between the weights associated with the investment policy function of uninformed investors across two consecutive update intervals:

$$\Delta_l^\theta = \sum_i |\theta_{lN}^i - \theta_{(l-1)N}^i|.$$

We then stop the simulation when the moving average of this difference falls below a certain level. A similar stopping rule is used in Calvano et al. (2020) and Dou et al. (2024).

Figure 7 shows the results of this model under an appropriate parameter choice. Our numerical experiment first shows that the model outcomes converge to some limit under these parameter values, where we have used the changes in the policy network weights over time as the criterion for convergence as mentioned above. In the figure, the blue curve plots the long-run limit of the investment policy functions of uninformed investors, which can be

regarded as an approximate solution to the true investment policy function that must arise in the rational expectations equilibrium. The orange curve in the same figure plots the true investment policy function arising in the rational expectations equilibrium in the previous model with a single signal.

The demand functions of uninformed investors in these two different economies exhibit several interesting features. First, when the price is either low (say, around 0.4) or high (say, around 1.4), these two demand functions are nearly indistinguishable from each other. We have already discussed the underlying intuition behind this result. That is, when the price is high (low), uninformed investors firmly believe that the current state is highly likely to be good (bad) state. As such, in these regions, the additional signal received by informed investors would be less informative and therefore, the demand function arising in this economy should be almost the same as the demand function obtained in the economy with a single signal.

However, when the price is at an intermediate level, the demand function arising in the economy with separate signals is lower that that obtained in the economy with a single signal. To understand these results, note that when prices fall in an intermediate region, the additional signal received by informed investors becomes more valuable for these investors. This result leads to a higher degree of information asymmetry between informed investors and uninformed investors. As a result, the negative information effect of a price drop is enlarged, driving uninformed investors to reduce their demand for the asset, compared to the case where informed investors have access to only a single signal.

3.2.3 Model with Financial Constraints

In this section, we introduce financial constraints such as borrowing and short-sale constraints into the model of Grossman and Stiglitz (1980). We then attempt to approximate the solution to this extension by calculating the long-run limit of our model outcomes without rational expectations.

In the literature, Yuan (2005) extends the model of Grossman and Stiglitz (1980) by incorporating borrowing constraints and solves the extended model numerically. However, this model specifies borrowing constraints in a reduced form which may not encompass a wide

range of environments. In our model, we will introduce borrowing and short-sale constraints in their most natural forms. Also, Yuan (2005) assumes that uninformed investors have a mean-variance preference, although the expected utility of these investors is not equivalently reduced to a mean-variance preference, given that the linear relationship between prices and payoffs no longer holds in the presence of borrowing constraints. In our model, we do not adopt this simplifying specification regarding utility functions. In fact, this approach does not fit the reinforcement learning algorithm because this algorithm uses the realized utilities rather than the expected utility to update a policy function after observing realized data.

To begin, we assume that all investors are endowed with an initial wealth of $\bar{w}$ at the beginning of each time, where $\bar{w}$ is a positive constant. Also, following Yuan (2005), we assume that only a subset of informed investors are subject to financial constraints, say, borrowing and short-sale constraints. The remaining informed investors and all other uninformed investors do not face the financial constraints. The reason why uninformed investors are not assumed to be financially constrained is that, as pointed out by Yuan (2005), the main interest of this extension lies in examining how the presence of financial constraints affects the inference problem of uninformed investors. As such, imposing financial constraints on uninformed investors would unnecessarily make the model more complicated without providing additional insights. In other words, we can compare the demand function of uninformed investors in an economy to financial constraints with that in the original economy of Grossman and Stiglitz (1980) more effectively by assuming only some informed investors are financially constrained.

Let m_I^c denote the measure of informed investors who are subject to borrowing and short-sale constraints and let m_I^{uc} denote the measure of informed investors who are not subject to any financial constraints. We continue to denote the measure of uninformed investors by m_U as in the previous sections.

Regarding financial constraints, we assume that demand of each financially informed investor, denoted by $q_t^{I,c}$, must satisfy the following condition:

$$\begin{cases} -\xi \leq q_t^{I,c} \leq \frac{\bar{w}}{p_t}, & \text{if } p_t > 0 \\ \frac{\bar{w}}{p_t} \leq q_t^{I,c} \leq \xi, & \text{if } p_t < 0 \end{cases} \tag{16}$$

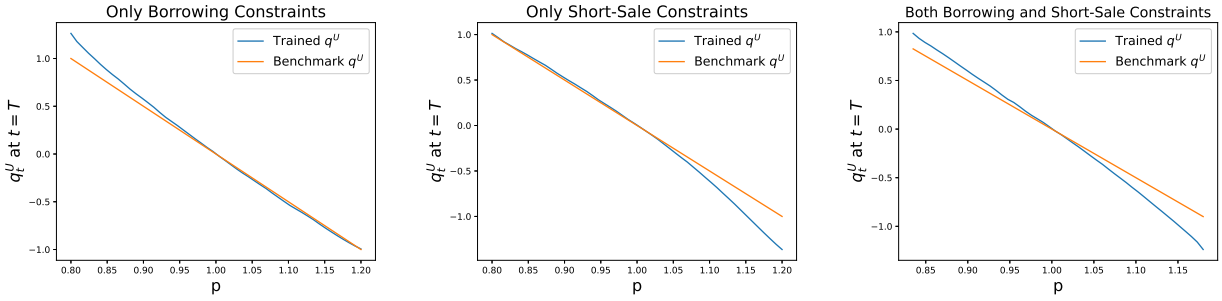

Figure 8: This figure depicts the results of the model with financial constraints. The left panel describes the case with only borrowing constraints. The middle panel describes the case with only short-sale constraints. The right panel plots the case with both borrowing and short-sale constraints. In each panel, the blue curve plots the demand function of uninformed investors in this extension, while the orange curve plots the demand function of uninformed investors in the canonical model without financial constraints. The parameter values chosen for the third case are $\mu_x = 1$, $\sigma_x = 0.2$, $\sigma_s = 0.1$, $\mu_z = 3$, $\sigma_z = 2$, $\eta = 5$, $\bar{w} = 2.3$, $\xi = 2.1$, $m_I^u = 0.7$, $m_I^c = 0.3$, and $m_U = 2$. For the first case, ξ is set to ∞. For the second case, $\bar{w}$ is set to ∞. In all three cases, the update interval length is set to $N = 3000$. The batch size is set to 64. The number of epochs is set to 10. The learning rate is initially set to 10^{-4} and then is reduced by a factor 0.99 every 30 epochs. The number of hidden layers is set to 5. The number of nodes of each hidden layer is set to 100. The activation function is set to the ReLu function.

for some constant $\xi \geq 0$. In the case of $p_t > 0$, the left inequality indicates a short-sale constraint and the right inequality represents a borrowing constraint. That is, financially constrained investors cannot make a negative position in the risky asset, which exceeds a certain limit, and cannot invest more than their initial wealth in the risky asset. In the case of $p_t < 0$, we can formally specify the borrowing and short-sale constraints in the above form. But since this case is unrealistic, when we simulate our model, we will choose parameter values to ensure that the second case would rarely occur. Also, note that the case of $\xi = \infty$ corresponds to the case with only borrowing constraints and the case of $\bar{w} = \infty$ indicates the case with only short-sale constraints. When we present the results later, we will consider these two special cases first to better illustrate the results.

The demand of each financially unconstrained informed investor, denoted by $q_t^{I,uc}$, is still given by the formula in (1), which depends on both the price p_t and private signal s_t.

Then, the demand of each financially constrained informed investor is simply given by

$$\begin{cases} q_t^{I,c} = \max\left\{\min\left\{q_t^{I,uc}, \frac{\bar{w}}{p_t}\right\}, -\xi\right\}, & \text{if } p_t > 0 \\ q_t^{I,c} = \min\left\{\max\left\{q_t^{I,uc}, \frac{\bar{w}}{p_t}\right\}, \xi\right\}, & \text{if } p_t < 0, \end{cases} \tag{17}$$

due to the constraints in (16). Next, regarding the decisions of uninformed investors, we need not modify their learning procedure, as expected.

Figure 8 presents the results of this model with borrowing and short-sale constraints. Specifically, the left two panels presents the results in case with only borrowing constraints. The top-left panel first shows that the model outcomes converge to some limit in the long run, which can be regarded as the approximate solution for the true rational expectations equilibrium of this model.

In the top-left panel, the orange line indicates the demand function of uninformed investors in the canonical model of Grossman and Stiglitz (1980) and the blue curve plots the long-run limit of the demand function of uninformed investors in our model. The result shows that (i) when the price is low, the demand by uninformed investors is higher in the economy with borrowing constraints than that in the canonical model and (ii) when the price is high, the two demand functions are close to each other.

To understand this result intuitively, note first that borrowing constraints tend to be binding when the price is low and therefore, assets in the economy with borrowing constraints would be underpriced when prices are low, compared to the canonical economy. As a result, uninformed investors in the economy with borrowing constraints would infer the quality of the asset to be higher than that inferred by uninformed investors, who observe the same level of a price, in the economy without borrowing constraints. Therefore, the demand by uninformed investors in the economy with borrowing constraints should be higher than than in the economy without borrowing constraints. Meanwhile, when prices are high, the demand functions by uninformed investors in the two different economies should be close to each other.

The middle two panels present the results for the case with only short-sale constraints. We can understand these results in the opposite way to the case with only borrowing constraints. That is, since short-sale constraints tend to binding when prices are high, assets

in the economy with short-sale constraints would be overpriced, compared to the economy without short-sale constraints. Hence, for the similar reason mentioned above, when prices are high, the demand by uninformed investors in the economy with short-sale constraints should be lower than that in the economy without short-sale constraints. When prices are low, the uninformed investors in the two different economies make almost the same decision.

The right two panels present the results in the economy with both borrowing and short-sale constraints. In this case, the outcomes that are observed in the above two cases, in which we have examined the effects of borrowing constraints and short-sale constraints separately, occur at the same time. That is, as shown in the top-right panel, the demand function of uninformed investors facing both the constraints cross the demand function arising in the economy with no financial constraints from the above to below. This result is consistent with our intuition for the reasons mentioned above. To the best of our knowledge, this paper is the first to provide an approximate solution to the extension of Grossman and Stiglitz (1980) that incorporates borrowing and short-sale constraints in their most natural forms.

4 Conclusions

In this paper, we investigate whether financial markets eventually achieve market efficiency, even if investors have no prior knowledge of the economic environment. To explore this issue, we assume that uninformed investors use a reinforcement learning algorithm. Using this framework, we show that our model outcomes tend to converge to the true rational expectations equilibriums in the settings considered by Grossman and Stiglitz (1980) and Breon-Drish (2015), both of which are analytically tractable. The fact that financial markets eventually become informationally efficient even in the absence of the rational expectations assumption has important implications for practitioners. That is, this result means that even if those practitioners in the investment industry design their investment strategies without having the perfect knowledge of the economic environment, they will eventually find an optimal investment strategy, especially when those investors use a reinforcement learning algorithm, which is a relatively new technology invented in the field of machine learning. We then apply our approach to more general settings that are not analytically tractable. For

these models, our approach provides credible approximate solutions to the true rational expectations equilibrium by calculating the long-run limit of our model. Our work contributes to the growing intersection of finance and machine learning, illustrating how adaptive algorithms can help investors refine their strategies in uncertain environments. This framework holds promise for exploring diverse, complex financial models and deepening insights into market behavior when traditional approaches are not feasible.

A Appendix

A.1 Omitted Solutions for Section 2.1

The solutions for the coefficients in relationship (2) are given by

$$A = \frac{m_I\tau_s + m_U\tau_s\kappa}{m_I(\tau_x+\tau_s) + m_U\left(\tau_x+\tau_s\kappa\right)}, \quad B = \frac{\rho\left(m_I\tau_s + m_U\tau_s\kappa\right)}{m_I(\tau_x+\tau_s) + m_U\left(\tau_x+\frac{\tau_s\tau_z}{\tau_z+\rho^2\tau_s}\right)},$$

$$C = \frac{(m_I+m_U)\tau_x\mu_x - \eta\mu_z}{m_I(\tau_x+\tau_s) + m_U\left(\tau_x+\tau_s\kappa\right)}, \tag{18}$$

where $\rho = -\frac{\eta}{m_I\tau_s}$ and $\kappa = \frac{\tau_z}{\tau_z+\rho^2\tau_s}$ as stated before.

References

Albagli, E., Hellwig, C., and Tsyvinski, A. (2024). Information aggregation with asymmetric asset payoffs. *The Journal of Finance*.

Ang, A. and Timmermann, A. (2012). Regime changes and financial markets. *Annual Review of Financial Economics*, 4(1):313–337.

Asness, C. S., Moskowitz, T. J., and Pedersen, L. H. (2013). Value and momentum everywhere. *The journal of finance*, 68(3):929–985.

Bernardo, A. E. and Judd, K. L. (2000). Asset market equilibrium with general tastes, returns, and informational asymmetries. *Journal of Financial Markets*, 3(1):17–43.

Bray, M. (1982). Learning, estimation, and the stability of rational expectations. *Journal of economic theory*, 26(2):318–339.

Bray, M. and Kreps, D. M. (1987). Rational learning and rational expectations. In *Arrow and the ascent of modern economic theory*, pages 597–625. Springer.

Breon-Drish, B. (2015). On existence and uniqueness of equilibrium in a class of noisy rational expectations models. *The Review of Economic Studies*, 82(3):868–921.

Buehler, H., Gonon, L., Teichmann, J., and Wood, B. (2019). Deep hedging. *Quantitative Finance*, 19(8):1271–1291.

Calvano, E., Calzolari, G., Denicolo, V., and Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative finance*, 1(2):223–236.

Dou, W. W., Goldstein, I., and Ji, Y. (2024). Ai-powered trading, algorithmic collusion, and price efficiency. *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper*.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of finance*, 25(2):383–417.

Gabaix, X. (2009). Power laws in economics and finance. *Annu. Rev. Econ.*, 1(1):255–294.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep learning. Cambridge, Massachusetts: The MIT Press.

Grossman, S. J. and Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American economic review*, 70(3):393–408.

Hellwig, M. F. (1980). On the aggregation of information in competitive markets. *Journal of economic theory*, 22(3):477–498.

Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance*, 48(1):65–91.

Kelly, B. and Jiang, H. (2014). Tail risk and asset prices. *The Review of Financial Studies*, 27(10):2841–2871.

McLean, R. D. and Pontiff, J. (2016). Does academic research destroy stock return predictability? *The Journal of Finance*, 71(1):5–32.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Peress, J. (2004). Wealth, information acquisition, and portfolio choice. *The Review of Financial Studies*, 17(3):879–914.

Rostek, M. J. and Yoon, J. H. (2023). Imperfect competition in financial markets: Recent developments. Technical report, SSRN Working Paper, 2023. https://doi. org/10. 2139/ssrn. 3710206.

Routledge, B. R. (1999). Adaptive learning in financial markets. *The Review of Financial Studies*, 12(5):1165–1202.

Schulman, J., Levine, S., Moritz, P., Jordan, M. I., and Abbeel, P. (2015). Trust region policy optimization. *In Proceedings of the 32nd International Conference on Machine Learning*, page 1889–1897.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.

Verrecchia, R. E. (1982). Information acquisition in a noisy rational expectations economy. *Econometrica*, pages 1415–1430.

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

Watkins, C. J. C. H. (1989). Learning from delayed rewards. *PhD Dissertation*.

Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256.

Yuan, K. (2005). Asymmetric price movements and borrowing constraints: A rational expectations equilibrium model of crises, contagion, and confusion. *The Journal of Finance*, 60(1):379–411.