A Long and a Short Leg Make For a Wobbly Equilibrium

Nicolae Gârleanu*[1], Stavros Panageas†[2], and Geoffery Zheng‡[3]

[1]WashU, Olin Business School and NBER
[2]UCLA, Anderson School of Management and NBER
[3]NYU Shanghai

March 2025

Abstract

We provide a model to explain how the interaction between the spot and lending markets for stocks can lead to abrupt changes in short selling activity. Furthermore, rational short sellers may choose to abandon the market even as mispricing widens. We document empirically that the dynamics of short selling are fat-tailed and subject to abrupt changes, especially for the stocks that the model identifies as susceptible to such dynamics.

Keywords: Asset pricing with frictions, Short selling, Runs, Limits to arbitrage

JEL Classification: G11, G12, G14

*garleanu@wustl.edu
†stavros.panageas@anderson.ucla.edu
‡geoff.zheng@nyu.edu. We would like to thank Adrien d'Avernas, Itamar Dreschler, Sergei Glebkin, Dan Greenwald, Christian Heyerdahl-Larsen, Paymon Khorrami, Alberto Teguia, the editor, and three anonymous referees, as well as seminar participants at the 2021 NBER Summer Institute Asset Pricing Program, the 2022 BI-SHoF Conference, the 2022 SFS Cavalcade, the 2022 WFA, the 16th Annual Cowles Conference on General Equilibrium, Boston University Questrom School of Business, BYU Marriott School of Business, Copenhagen Business School, CKGSB, CU Boulder, CUHK, Duke Fuqua School of Business, Florida International University, Harvard Business School, Oxford Saïd Business School, Shanghai Advanced Institute of Finance, Tsinghua SEM, Warwick Business School, University of Maryland Smith School of Business, UW Foster School of Business, UNC Kenan-Flagler Business School, and Victoria University of Wellington for their useful comments on the paper.

Short interest in individual stocks is unstable, exhibiting sudden and large changes. We propose a theoretical explanation for this instability, relying on a feedback loop between the spot and the lending markets. Our theory can also help explain why short sellers sometimes abandon their positions despite an increased profitability of shorting.

The novel aspect of our theory is that it does not rely on portfolio constraints, limitations to arbitrage capital, agency issues, etc. Instead, our mechanism is built on a detailed modeling of stock-lending income and its implications for spot-market clearing. We show that lending income gives rise to a feedback mechanism between a stock's expected return and short interest that can generate a "backward-bending demand," and accordingly sudden equilibrium shifts in short interest and expected returns.

We document that large and sudden changes in shorting activity are a broad phenomenon. Viewed as a time series, the short interest process of many stocks exhibits jump-like features. These features appear linked to the backward-bending demand channel that we highlight: the stocks that satisfy empirically the specific condition for a backward-bending demand curve implied by the model are also the ones that show the highest incidence of large and sudden changes in short interest.

The model features investors with heterogeneous beliefs about the expected return of a positive-supply risky stock: one group is optimistic, while the other holds rational beliefs.[1] This difference of opinion between investors prompts them to trade with each other, with the rational investors having an incentive to short the stock whenever the expected excess return becomes negative. Shorting a stock requires borrowing it, for a fee determined endogenously in the lending market as a result of bargaining.

We first discuss the model's implications for the Sharpe ratio; later, we turn to the price-dividend ratio. The presence of lending fees modifies the returns experienced by both long and short investors. The equilibrium risk compensation (the ratio of excess return to volatility, or "Sharpe ratio") is impacted both by the magnitude of the lending fee and the fraction of a representative lender's shares that are shorted. Following common terminology, we refer to the ratio of shorted-to-lendable shares as the "utilization" ratio.

1. Motivated by the empirical fact that stocks with high short interest tend to have low subsequent returns, we assume that the comparatively pessimistic investors are actually rational, but this is not an essential assumption for our results.

All else equal, a higher utilization ratio acts as an increased subsidy for long positions, since a larger fraction of the representative long position is lent out to short sellers. This subsidy can generate equilibrium multiplicity, since the increased incentive to purchase the stock ends up reducing its Sharpe ratio and consequently inducing more shorting, thus supporting the high utilization ratio as an equilibrium outcome. In addition, other equilibria may exist, including one equilibrium featuring no shorting at all. These equilibria feature lower utilization ratios, thus lower subsidies for long positions, and consequently higher Sharpe ratios and smaller short positions. The model also provides a necessary and sufficient condition for multiple equilibria (Proposition 4). The condition is in terms of an observable quantity, namely the gap between the lending fee paid by the short seller and the lending income received by the representative long investor.

While the main focus of the paper is theoretical, we also test the above model predictions empirically. Specifically, we argue that our theory may help explain some salient time-series properties of the utilization ratio. According to the model, short-run changes in the utilization ratio are small and (locally) normally distributed most of the time; but when there are changes in equilibrium, this ratio jumps. Therefore, the distribution of the changes in the utilization ratio should be fat-tailed. Empirically, the changes in the utilization rate are remarkably fat tailed. Further, we confirm that the stocks that satisfy our necessary and sufficient condition for multiplicity (Proposition 4) are the ones that are the most likely to exhibit jumps in utilization.

The baseline model targets the study of shorting activity and the Sharpe ratio, but log utility in a one-positive-supply-asset setting generates a constant price-dividend ratio. To discuss the model implications for the price-dividend ratio we develop two model extensions. These extensions highlight two distinct mechanisms, acting through the dynamics of the investor wealth shares, that cause the price-dividend ratio to increase when short-selling declines. We are particularly interested in illustrating this outcome, because one would think that a higher stock price is associated with more, not less, short selling.

In the first extension, we generalize the baseline model to allow for recursive preferences with an IES below one. We show that the wealth dynamics imply that the price-dividend ratio is lower when investors coordinate on the high- rather than the no-shorting equilibrium.

The intuition is that all investors believe that their wealth growth is higher when the shorting market is more active, since the active trading will help them vindicate their views. The higher anticipated wealth growth encourages more consumption today and pushes down the relative price of the positive supply asset to consumption (i.e., the price-dividend ratio).

The second extension is designed to address the fact that shorting frictions and belief disagreement usually pertain only to a few small stocks, rather than the market as a whole. To this end, we retain the unit IES assumption and instead extend the baseline model to allow for a large and a small stock.[2] In the spirit of realism, we also assume that only a small fraction of investors pay attention to the small stock and incur a small participation cost in doing so. We show that the shift to a low shorting equilibrium causes rational investors to exit the market for the small stock, since remaining in a market without a trading opportunity is not worth paying that participation cost. The exit of rational short-sellers increases the wealth share of irrational investors, which can lead to a rise in the price of the stock.

The paper concludes with a "case study" on the fickle behavior of short sellers. We document that the period between November 2020 and January 2021 saw an abrupt decline in short interest across hundreds of highly shorted stocks, and was also the worst period for a "betting against the short sellers" strategy, i.e., a strategy that goes long the top decile of most shorted stocks and shorts the market portfolio.

It is tempting to attribute this episode to the highly mediatized events involving the company GameStop, which saw online-forum-coordinated retail purchases resulting in a short squeeze of its stock. However, the broad-based short-seller retreat that we focus on started eight weeks before the GameStop episode and impacted stocks that were not particularly discussed online by retail traders and did not experience an appreciable change in retail purchase volume.

There are three aspects of this episode that are pertinent for our model. First, the episode helps illustrate how abruptly and dramatically short selling can decline. Second, the retreat of the short sellers coincides with a rise in prices, a puzzling phenomenon[3] that our model

2. With this assumption, the interest rate becomes essentially fixed and therefore fluctuations in the Sharpe ratio are mirrored in the price-dividend ratio of the small stock. By contrast, in the baseline model the assumption of log utility and i.i.d. dividend growth imply that fluctuations in the Sharpe ratio are exactly offset by fluctuations in the interest rate, leaving the price-dividend ratio unaffected.

3. In a static model, lower short-seller demand would only be consistent with a lower price and higher

can account for (see Section 6). Finally, the retreat of the short sellers predated the spike in online discussion. One possible explanation for this retreat was the fear of an impending change in retail-investor behavior. Absent the backward-bending demand feature of our model, however, an impending rise in irrationality would raise the profitability of short selling and increase short interest, which is the opposite of what happened in the data.

The paper is organized as follows. After a brief literature review, Section 1 lays out the baseline version of the model and Section 2 presents the main analytical results. Section 3 discusses the dynamics of the investor wealth shares. Section 4 generalizes the results of Section 2 and provides necessary and sufficient conditions for equilibrium multiplicity. Section 5 tests the model's main empirical implications. Section 6 presents extensions to non-unit IES and multiple stocks. Section 7 discusses the curious patterns of short selling between November 2020 and January 2021. Section 8 concludes. Proofs, detailed descriptions of the data, and additional results are contained in the appendix.

Related Literature

Our work relates to several strands of the asset-pricing literature. The most closely related one considers the joint determination of lending fees, short interest, and returns. In particular, D'Avolio (2002), Duffie, Gârleanu, and Pedersen (2002), Vayanos and Weill (2008), Banerjee and Graveline (2013), Evgeniou, Hugonnier, and Prieto (2022), and Atmaz, Basak, and Ruan (2023) consider explicit frictions to lending and borrowing shares, which translate into non-zero lending fees that in turn impact expected returns.[4] Similar to D'Avolio (2002),[5] Banerjee and Graveline (2013), and Atmaz, Basak, and Ruan (2023), the lending and spot markets clear simultaneously in our paper, but we use a different micro-foundation to obtain a positive lending fee. Specifically, we don't impose any hard constraint on the shares that a long investor can lend.[6] Instead, we obtain a positive lending fee by assuming that the

expected return.

4. Such frictions also motivated the empirical studies of Geczy, Musto, and Reed (2002), Lamont (2012), Jones and Lamont (2002), Kaplan, Moskowitz, and Sensoy (2013), Porras Prado, Saffi, and Sturgess (2016), and Asquith, Pathak, and Ritter (2005) among others.

5. More precisely, to a working-paper version of this study, which contains a theoretical model that did not appear in the published article.

6. Evgeniou, Hugonnier, and Prieto (2022) also does not impose a hard constraint on the quantity of lendable shares. Instead, it assumes that the supply of lendable shares is adjusted by a monopolistic entity to

process of matching share lenders and borrowers is a time-consuming activity, which requires compensation, similar in spirit to Duffie, Gârleanu, and Pedersen (2002). By taking that route, our model allows for a more general specification of the supply curve of lendable shares, which is not confined to being vertical.[7] This specification of the supply curve for lendable shares leads to a feedback loop between the Sharpe ratio and short interest that is not present in the aforementioned papers (which feature unique equilibria). In addition, our model allows us to explore the dynamic effects of an equilibrium shift, driven by the endogenous fluctuations in the wealth shares of the different types of agents.[8]

An even larger number of papers assume that shorting is prohibited and analyze implications for returns. Prominent examples here include Harrison and Kreps (1978), Miller (1977), Diamond and Verrecchia (1987), Detemple and Murthy (1997), Hong and Stein (2003), and Scheinkman and Xiong (2003). As in Harrison and Kreps (1978) and Miller (1977), we model the motive for trade in our paper in the convenient form of (dogmatic) differences of opinions among agents.

A large body of work studies the empirical relation between short interest and stock returns. Seneca (1967), Senchack and Starks (1993), Desai et al. (2002), Diether, Lee, and Werner (2009), Asquith, Pathak, and Ritter (2005), Blocher, Reed, and Van Wesep (2013), Beneish, Lee, and Nichols (2015), and Dechow et al. (2001) study the cross-sectional relation and find that stocks with higher short interest under-perform those with lower short interest. Cohen, Diether, and Malloy (2007) and Boehmer, Jones, and Zhang (2008) use proprietary data on quantities lent as well as shorting fees and find consistent results. Duong et al. (2017) studies the empirical relation between lending fees and stock returns and finds that high lending fees predict lower future returns. Drechsler and Drechsler (2014) documents that asset pricing anomalies concentrate in stocks with high shorting fees. Lamont and Stein

maximize lending revenue. In our paper, investors face search frictions in the lending market that make it costly to locate lendable shares.

7. An exception is Atmaz, Basak, and Ruan (2023). In their model, individual agents' supply curves are vertical, but the aggregate supply curve has finite elasticity due to composition effects when aggregating across agents.

8. The fact that shorting requires borrowing shares and is subject to natural collateral requirements has several interesting general equilibrium implications, as explored by Fostel and Geanakoplos (2008), Simsek (2013), and Biais, Hombert, and Weill (2021). In contrast, our model focuses on the general equilibrium implications of the associated lending fees.

(2004) studies the information content in aggregate short interest and finds that short interest declined as stock market valuations rose in the late 90's. Rapach, Ringgenberg, and Zhou (2016) shows that the predictive power of aggregate short interest stems predominantly from a cash-flow channel.

Our paper also relates to a sizable theoretical literature analyzing multiple equilibria in asset pricing and macroeconomics. Multiple equilibria can arise through a number of mechanisms, chief among them a) bubbles (or money) in OLG economies, b) increasing returns to scale and production externalities, and c) portfolio constraints.[9] The mechanism that gives rise to multiple equilibria in our paper is different, since it relies on the interaction between the lending and the spot markets. We also note in this context that, while Vayanos and Weill (2008) features multiple equilibria in the presence of shorting frictions and fees, the multiplicity of equilibria pertains to agents' choice of market to join, which renders one asset more liquid (that is, easier to find) and thus increases its attractiveness to future entrants. In addition, in our setup the spot market is not a search market, but is Walrasian.[10]

Finally, several recent papers target specifically the set of events involving GameStop. See, for instance, Pedersen (2022) and Allen et al. (2021).

1 Model

1.1 Agents: life-cycle and preferences

Time is continuous and infinite for tractability. To obtain a stationary wealth distribution, we follow Gârleanu and Panageas (2015) and assume that investors continuously arrive ("births") and depart ("deaths") from the economy. Per unit of time a mass π of investors arrives, and a mass π departs. Therefore, the population of agents born at time $s \leq t$ and still remaining at time t is $\pi e^{-\pi(t-s)}$. The total population is constant and equal to $\int_{-\infty}^{t} \pi e^{-\pi(t-s)} ds = 1$.

9. We refer the reader to the survey by Benhabib and Farmer (1999), which lists and discusses the different mechanisms that lead to multiple equilibria and indeterminacies. Recent examples of papers using multiple-equilibrium models in asset pricing include Gârleanu and Panageas (2021), Khorrami and Zentefis (2020), Khorrami and Mendo (2021), Zentefis (2022), and Farmer and Bouchaud (2020).

10. Coordination issues are central in economies admitting multiple equilibria, but can also be of first-order importance in unique-equilbrium settings, as highlighted by Abreu and Brunnermeier (2002) in a model featuring binding portfolio constraints and a non-Walrasian price protocol.

"Births" and "deaths" should be understood as arrivals and departures of market participants, a point that will become clearer in Section 6.2, where we introduce multiple stocks.

To introduce trade in equities, we assume that investors have heterogeneous beliefs. For simplicity, a fraction $\nu \in (0,1)$ of investors perceive the correct data-generating process. We refer to them as rational investors (R investors). The remaining fraction are overly optimistic (we model this optimism shortly), and we refer to these investors as I investors.

For tractability, both investors have logarithmic utilities and their expected discounted utility from consumption is

$$V_t^i \equiv E_t^i \int_t^\infty e^{-(\rho+\pi)(u-t)} \log\left(c_{u,t}^i\right) du \tag{1}$$

for $i \in \{I, R\}$, with ρ a discount rate and $c_{u,t}^i$ the time-u consumption of an agent of type i born at time $t \leq u$. The notation E_t^i reflects the different investor beliefs. Because of death, the effective discount rate is $\rho + \pi$.

Before proceeding, we note that, while we require heterogeneous beliefs to introduce a motivation for trading, the assumption that one group has correct beliefs helps mostly to save notation and can be easily relaxed. The same applies to the assumption that there are only two groups of investors, which can be relaxed to allow for multiple investor types, including a continuum (Section 4). Similarly, the overlapping-generations structure is just a technical device to ensure that no investor type disappears in the long run.[11] Finally, in setting up the model we make the (conventional) assumption that agents maximize over both their consumption and portfolio choices, which we introduce shortly. Our model is, however, equivalent to one in which agents delegate their portfolio decisions to professional managers and managers maximize their clients' expected portfolio (logarithmic) growth according to the managers' beliefs (R or I). The investors in our model can therefore be equivalently thought of as institutional investors.

11. In particular, the lack of inter-generational risk sharing, which is a feature of some of these models, is not driving any of the results in this paper.

1.2 Endowments

In order to support their consumption over their lives, we assume that the arriving investors at time t are equally endowed with shares of new "trees," which arrive at time t.[12] Letting $s \leq t$ denote the time of arrival of a tree, we specify its time-t dividends as

$$D_{t,s} = \delta e^{-\delta(t-s)} D_t, \tag{2}$$

where $\delta > 0$ captures depreciation and D_t follows a geometric Brownian motion with mean μ_D and volatility $\sigma_D > 0$,

$$\frac{dD_t}{D_t} = \mu_D dt + \sigma_D dB_t, \tag{3}$$

with B_t a standard Brownian motion. Accordingly, the time-t total endowment of this economy is the sum of the endowment produced by all trees born up to to time t,

$$\int_{-\infty}^{t} D_{t,s} ds = \left(\int_{-\infty}^{t} \delta e^{-\delta(t-s)} ds \right) \times D_t = D_t.$$

The arriving investors sell their shares, which become part of the market portfolio. An implication of assumption (2) is that the dividend growth, $\frac{dD_{t,s}}{D_{t,s}} = (\mu_D - \delta)dt + \sigma_D dB_t$, is the same for any vintage s, and equals the dividend growth of the market portfolio. In turn, the return of the market portfolio, dR_t, can be written as

$$dR_t = \mu_t \, dt + \sigma_t \, dB_t, \tag{4}$$

where μ_t and σ_t are stochastic processes to be determined in equilibrium.

In the real world, shorting frictions are more relevant for a small fraction of stocks rather than the broad stock market. In Section 6.2 we extend the model to allow for multiple stocks and study the special case in which the shorting frictions are relevant for small stocks only.

12. The assumption that investors are endowed with shares of newly arriving trees follows Gârleanu, Kogan, and Panageas (2012) and Panageas (2020). This assumption is just a convenient way to endow new cohorts as compared to introducing labor income (as in Gârleanu and Panageas (2015) or Gârleanu and Panageas (2023)). Since the goal of the overlapping generations structure in this paper is merely to ensure stationarity, we adopt this more convenient shortcut.

1.3 Beliefs

The irrational investors are optimistic and believe that the aggregate endowment grows at the rate $\mu^I > \mu_D$. Irrational investors hold this optimistic view over their lifetime and do not learn ("dogmatic beliefs"). Learning would be a distraction for the purposes of this paper and therefore we omit it.

1.4 Dynamic budget constraint and short-selling frictions

The main departure from a frictionless market is that selling the stock short requires paying a lending fee, f_t. Specifically, letting $W^i_{t,s}$ denote the time-t wealth of an investor of type i who was born at time $s \leq t$ and $w^i_{t,s}$ denote the fraction of wealth invested in the stock, the dynamic budget constraint is

$$dW^i_{t,s} = W^i_{t,s}\left(r_t + \pi + n_t + w^i_{t,s}\left(\mu_t - r_t + \lambda^i_{t,s}\right) - \frac{c^i_{t,s}}{W^i_{t,s}}\right)dt + w^i_{t,s}W^i_{t,s}\sigma_t dB_t, \tag{5}$$

where r_t is the equilibrium interest rate and $\pi W^i_{t,s}$ is the income per unit of time earned from annuitizing her entire wealth, since she has no bequest motives. (We follow Blanchard (1985) in assuming the existence of a competitive insurance company. Investors pledge their wealth upon death in exchange for receiving an income stream while alive. This income stream is equal to the hazard rate of death, π, per unit of pledged wealth, so that the insurance company breaks even.) The non-standard terms in equation (5) are the $\lambda^i_{t,s}$ and n_t, which we describe next.

The term $\lambda^i_{t,s}$ captures the presence of lending fees. It is defined as

$$\lambda^i_{t,s} \equiv \lambda_t(w^i_{t,s}) \equiv f_t \times \left(1_{\left\{w^i_{t,s}<0\right\}} + \tau y_t 1_{\left\{w^i_{t,s}\geq 0\right\}}\right), \tag{6}$$

where $1_{\{.\}}$ is an indicator function, y_t is the fraction of a long portfolio that is lent out by the representative "brokerage house," and τ is the fraction of the lending fees that accrues to the investor. (We discuss the determination of y_t, τ, and f_t shortly.) Equation (6) reflects that an investor with a short position $w^i_{t,s} < 0$ has to pay a proportion f_t of the value of her entire short position, $|w^i_{t,s}|W^i_{t,s}$, so that the net-of-fee excess rate of return per dollar shorted

is $-(\mu_t - r_t + f_t)\,dt - \sigma_t dB_t$. Similarly, an investor holding a positive position, $w^i_{t,s} > 0$, obtains an excess rate of return equal to $(\mu_t - r_t + \tau y_t f_t)\,dt + \sigma_t dB_t$ on her stock investments.

Market clearing for share lending requires that the fraction of the representative long position that is lent out, y_t, times the aggregate long position, W^+_t, equal the value of the aggregate short position, W^-_t :

$$y_t W^+_t = W^-_t, \tag{7}$$

where

$$W^-_t \equiv \sum_{i\in\{I,R\}} \int_{-\infty}^{t} |w^i_{t,s}| W^i_{t,s} 1_{\left\{w^i_{t,s}<0\right\}} ds \tag{8}$$

$$W^+_t \equiv \sum_{i\in\{I,R\}} \int_{-\infty}^{t} w^i_{t,s} W^i_{t,s} 1_{\left\{w^i_{t,s}>0\right\}} ds. \tag{9}$$

Following industry terminology, we henceforth refer to the quantity y_t as the utilization ratio (or utilization for short), since it captures the fraction of lendable shares that are utilized by shorters.

To close the model, we must specify the lending frictions and derive the fee. In the text, we specify f_t through a supply curve $f_t = f(y_t)$ given by a non-decreasing function f. In Appendix A, however, we model explicitly a search-and-bargaining friction yielding such a supply curve. Specifically, we introduce competitive firms specializing in servicing either borrowers ("brokers") or lenders ("security lenders"). Brokers are faced with a demand from would-be short sellers, while security lenders obtain investors' long portfolios. Brokers and security lenders are matched pairwise subject to a "labor cost" and engage in bilateral negotiations that result in a lending fee f_t. In equilibrium, the fee is the same for all shares that are lent, and therefore the total revenue from lending shares equals the fee multiplied by the value of all shares lent. This revenue is shared between the stock owners (a fraction τ of the lending revenue) and the households as compensation for their labor (the remaining $1-\tau$ fraction). These shares are driven by the relative bargaining powers of stock borrowers and lenders.

The term n_t in equation (5) captures the compensation for the labor cost in operating the matching technology. Denoting aggregate wealth at time t by W_t, we have $n_t = \frac{(1-\tau)f_t W_t^-}{W_t}$. We note that aggregate share-lending fees, $f_t W_t^-$, accrue back to the households as the sum of lending income to long portfolios, $\tau f_t y_t W_t^+ = \tau f_t W_t^-$, and compensation for operating the matching technology, $n_t W_t = (1-\tau) f_t W_t^-$.

1.5 Equilibrium definition

Equilibrium in the lending market requires that the supply of lendable shares $y_t W_t^+$ is equal to the demanded short interest, W_t^- (equation (7)).

The rest of the equilibrium definition is standard. We require that investors I and R maximize (1) over $c_{t,s}^i$ and $w_{t,s}^i$ subject to the budget constraint (5), and μ_t, r_t, and σ_t are such that the bond market clears, $\sum_{i\in\{I,R\}} \int_{-\infty}^t \nu^i \left(1 - w_{t,s}^i\right) W_{t,s}^i ds = 0$, the stock market clears, $\sum_{i\in\{I,R\}} \int_{-\infty}^t \nu^i w_{t,s}^i W_{t,s}^i ds = P_t$, and the goods market clears, $\sum_{i\in\{I,R\}} \int_{-\infty}^t \nu^i c_{t,s}^i ds = D_t$. By Walras' Law, market clearing of the bond market implies stock market clearing and vice versa, and accordingly the asset-market clearing requirements can be written equivalently as $W_t = \sum_{i\in\{I,R\}} \int_{-\infty}^t \nu^i W_{t,s}^i ds = P_t$.

For future reference, we note that stock market clearing implies $y_t = \frac{W_t^-}{W_t^+} = \frac{W_t^-}{P_t + W_t^-} < 1$. It also implies that there is a simple, monotone relation between the utilization ratio, y_t, and short interest, $\frac{W_t^-}{P_t}$, given by $y_t = \left(1 + \frac{W_t^-}{P_t}\right)^{-1} \frac{W_t^-}{P_t}$.

2 Analysis

We analyze the model in two steps. First, we consider a special parametric case that allows us to characterize all equilibrium quantities in closed form. The special case we analyze is the "elastic supply" case, that is, the limiting case where the supply of lendable shares is horizontal at some level $f(y_t) = \varphi$. (As we explain in Appendix A, this special case corresponds to a particular specification for the cost of lending out shares.) Section 4 extends the analysis to allow for an increasing function $f(y_t)$.

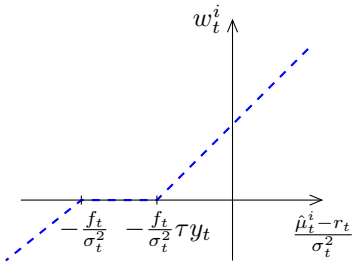


Figure 1: The optimal portfolio weight of investor i as a function of that investor's perceived value of $\frac{\hat{\mu}_t^i - r_t}{\sigma_t^2}$.

2.1 Optimal portfolio and consumption

For a log investor the wealth-to-consumption ratio is constant and equal to $\frac{c_{t,s}^i}{W_{t,s}^i} = \rho + \pi$. Given homothetic preferences, all agents of a given type choose the same portfolio independently of their cohort, s; therefore we may write w_t^i (rather than $w_{t,s}^i$). Additionally, a convenient property of logarithmic utility is that the portfolio is myopic and maximizes the logarithmic growth rate of an investor's wealth under the investor's beliefs,

$$w_t^i = \arg\max_w \left\{ r_t - \rho + w\left(\mu_t + \eta\sigma_t 1_{\{i=I\}} - r_t + \lambda_t(w)\right) - \frac{1}{2}\left(w\sigma_t\right)^2 \right\}, \tag{10}$$

where η is defined as

$$\eta \equiv \frac{\mu^I - \mu_D}{\sigma_D}. \tag{11}$$

Letting $\hat{\mu}_t^i \equiv \mu_t + \eta\sigma_t 1_{\{i=I\}}$ denote the expected return on the stock as perceived by investor $i \in \{I, R\}$, the optimal portfolio is

$$w_t^i = \begin{cases} \frac{\hat{\mu}_t^i - r_t + f_t}{\sigma_t^2} & \text{if } \hat{\mu}_t^i - r_t + f_t < 0 \\ \frac{\hat{\mu}_t^i - r_t + \tau f_t y_t}{\sigma_t^2} & \text{if } \hat{\mu}_t^i - r_t + \tau f_t y_t > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Figure 1 depicts equation (12), the optimal portfolio of investor i as a function of $\frac{\hat{\mu}_t^i - r_t}{\sigma_t^2}$. The figure shows the presence of an "inaction" region: for values of $\frac{\hat{\mu}_t^i - r_t}{\sigma_t^2}$ between $-\frac{f_t}{\sigma_t^2}$ and $-\frac{f_t}{\sigma_t^2}\tau y_t$, the investor optimally chooses a portfolio weight of zero.

One straightforward implication of equation (12) is that if investor R is actively shorting ($w_t^R < 0$) then the expected excess rate of return per dollar shorted is positive even after netting out the fee f_t.[13]

2.2 Equilibrium

It is useful to start by defining the wealth-weight ω_t^i of investors of type $i \in \{I, R\}$,

$$\omega_t^i \equiv \frac{\nu^i \int_{-\infty}^t \pi e^{-\pi(t-s)} W_{t,s}^i ds}{W_t}. \tag{13}$$

To save notation, henceforth we refer to ω_t^R simply as ω_t, and therefore $\omega_t^I = 1 - \omega_t$. Using $\frac{c_{t,s}^i}{W_{t,s}^i} = \rho + \pi$, the goods-market and stock-market clearing requirements imply

$$\begin{aligned} D_t &= \sum_{i \in \{I,R\}} \int_{-\infty}^t \nu^i \pi e^{-\pi(t-s)} c_{t,s}^i ds = (\rho + \pi) \sum_{i \in \{I,R\}} \int_{-\infty}^t \nu^i \pi e^{-\pi(t-s)} W_{t,s}^i ds \\ &= (\rho + \pi) W_t = (\rho + \pi) P_t. \end{aligned} \tag{14}$$

Taking logarithms gives $d \log D_t = d \log P_t$ and therefore the stock market volatility equals $\sigma_t = \sigma_D$. The implication of a constant stock volatility is convenient for obtaining closed-form solutions. In Section 6.2 we discuss extensions of the model that allow for a time-varying price-dividend ratio and volatility by introducing multiple stocks.

In an effort to obtain a closed-form solution we assume that the supply of lendable shares is perfectly elastic at the rate φ:

Assumption 1 $f(y) = \varphi > 0$.

We maintain this assumption until Section 4.

A remarkable feature of the model is its potential for multiple equilibria. Before stating formal conditions and results, it is instructive to sketch the argument of equilibrium multi-

13. This statement uses the assumption that agent R has the correct beliefs, and is a direct consequence of the agent's risk aversion. For a precise calculation, evaluate (12) with $i = R$, impose $w_t^R < 0$, and re-arrange to obtain $-(\mu_t - r - f_t) = -(\hat{\mu}_t^R - r - f_t) = -w_t^R \sigma_t^2 > 0$. The term $-w_t^R \sigma_t^2$, which equals the absolute value of the covariance of the stock's return with the short seller's portfolio, is the risk compensation to the agent for taking a short position.

plicity, by focusing first on an equilibrium that involves active shorting ($w_t^R < 0$). In such an equilibrium, the optimal portfolio holdings can be expressed as

$$w_t^R = \frac{\kappa_t + \frac{\varphi}{\sigma_D}}{\sigma_D} \tag{15}$$

$$w_t^I = \frac{\kappa_t + \eta + \frac{\varphi}{\sigma_D}\tau y_t}{\sigma_D}, \tag{16}$$

while asset-market clearing requires

$$\omega_t w_t^R + (1-\omega_t)\, w_t^I = 1. \tag{17}$$

Combining equations (15)–(17) leads to

$$\kappa_t = \sigma_D - (1-\omega_t)\,\eta - \frac{\varphi}{\sigma_D}\left(\omega_t + \tau y_t(1-\omega_t)\right). \tag{18}$$

Equation (18) shows that the Sharpe ratio, κ_t, is a declining function of utilization, y_t. To compute the value of y_t that clears the lending market, we note that in any equilibrium involving $w_t^R < 0$ and $w_t^I > 0$ we must have

$$y_t = \frac{W_t^-}{W_t^+} = \frac{-w_t^R W_t^R}{w_t^I W_t^I} = -\frac{w_t^R}{w_t^I} \times \frac{\omega_t}{1-\omega_t}. \tag{19}$$

Using (15) to compute the ratio $\frac{w_t^R}{w_t^I}$ gives

$$\begin{aligned} y_t &= -\frac{\kappa_t + \frac{\varphi}{\sigma_D}}{\kappa_t + \eta + \frac{\varphi}{\sigma_D}\tau y_t} \times \frac{\omega_t}{1-\omega_t} \\ &= \frac{\eta - \frac{\sigma_D}{1-\omega_t} - \frac{\varphi}{\sigma_D}(1-\tau y_t)}{\eta + \frac{\sigma_D}{\omega_t} - \frac{\varphi}{\sigma_D}(1-\tau y_t)}, \end{aligned} \tag{20}$$

where the last line follows from (18) after collecting terms and simplifying. Equation (20) is quadratic in y_t, and it may admit up to two solutions in the economically meaningful range $(0, 1)$. Proposition 2 below is devoted to studying this quadratic equation and confirming that its roots correspond to valid equilibria with non-zero shorting (under some additional assumptions on the parameters). The proposition further shows that when two such equilibria

exist, there also exists one with zero shorting.

The intuition for the equilibrium multiplicity can be explained starting with equation (18), which shows that the Sharpe ratio, κ_t, depends on utilization, y_t. This dependence gives rise to a feedback loop between κ_t and y_t: A higher value of y_t increases the rate of return on a long position and strengthens investor I's demand for the asset (equation (16)). This increased demand lowers the Sharpe ratio to clear the market. The lower Sharpe ratio strengthens the short-sellers' appetite to borrow the stock and short it. In turn, the increased shorting demand raises the utilization ratio, y_t, increasing the effective return to I investors, which further reduces the Sharpe ratio, etc.

We next provide a formal analysis of the full set of equilibria as a function of the wealth share of rational investors, ω_t. We start with a definition.

Definition 1 *Define the constant ω_1^* and the function $F : \mathbb{R} \to \mathbb{R}$ by*

$$\omega_1^* \equiv 1 - \frac{\sigma_D}{\eta - \frac{\varphi}{\sigma_D}}, \tag{21}$$

$$F(\omega) \equiv \left(\sigma_D - \omega\left((1+\tau)\frac{\varphi}{\sigma_D} - \eta\right)\right)^2 - 4\tau\frac{\omega^2}{1-\omega}\frac{\varphi}{\sigma_D}\left(\sigma_D + (1-\omega)\left(\frac{\varphi}{\sigma_D} - \eta\right)\right). \tag{22}$$

The next assumption provides a sufficient condition for the existence of multiple equilibria.

Assumption 2 *Assume that η, φ, σ_D, and τ are such that*

$$(1+\tau)\frac{\varphi}{\sigma_D} > \eta > \frac{\varphi}{\sigma_D}, \tag{23}$$

$$\omega_1^* > \frac{\sigma_D}{(1+\tau)\frac{\varphi}{\sigma_D} - \eta} > 0, \tag{24}$$

and F has a unique root in the interval $(0,1)$, denoted by ω_2^*.

The following proposition guarantees that Assumption 2 can be satisfied.

Proposition 1 *There exists an open set of positive values η, φ, σ_D, and τ that satisfy Assumption 2.*

The next proposition describes the equilibria in our economy.

Proposition 2 *Suppose that Assumptions 1 and 2 hold. Then* $\omega_2^* > \omega_1^*$ *and the equilibria in this economy are as follows.*

i) If $\omega_t \in (\omega_2^*, 1]$ *there is no short-selling in equilibrium. The equilibrium is unique and the Sharpe ratio* $\kappa_t \equiv \frac{\mu_t - r_t}{\sigma_D}$ *is given by*

$$\kappa_t = \begin{cases} \sigma_D - (1-\omega_t)\,\eta & \text{if } \omega_t > 1 - \frac{\sigma_D}{\eta} \\ \frac{\sigma_D}{1-\omega_t} - \eta & \text{if } \omega_t \in (\omega_2^*, 1 - \frac{\sigma_D}{\eta}] \end{cases}. \tag{25}$$

ii) If $\omega_t \in [\omega_1^*, \omega_2^*]$*, then there are three equilibria. The first equilibrium continues to be given by (25) and involves no short-selling. The second and third equilibria involve shorting and utilization,* y_t*, corresponds to the two roots* y^+ *and* y^- *of the quadratic equation*

$$y\left(\eta + \frac{\sigma_D}{\omega_t} - \frac{\varphi}{\sigma_D}(1-\tau y)\right) - \left(\eta - \frac{\sigma_D}{1-\omega_t} - \frac{\varphi}{\sigma_D}(1-\tau y)\right) = 0, \tag{26}$$

which has two real roots y^+ *and* y^- *in* $(0,1)$*. The Sharpe ratios in these equilibria are*

$$\kappa_t^{\pm} = \sigma_D - (1-\omega_t)\,\eta - \frac{\varphi}{\sigma_D}\left(\omega_t + \tau y^{\pm}(1-\omega_t)\right). \tag{27}$$

iii) If $\omega_t \in [0, \omega_1^*)$*, then the equilibrium is unique and involves shorting. In this case only the larger of the two roots* (y^+) *of equation (26) lies in the interval* $(0,1)$*, and the unique equilibrium Sharpe ratio is given by* κ^+*.*

In all three cases the interest rate is given by

$$r_t = \rho + \pi + \mu_D - \delta - \kappa_t \sigma_D. \tag{28}$$

Additionally, because κ_t*,* r_t*, and* y_t *are functions of* ω_t*, so is* w_t^R*, and the stochastic process for* ω_t*,* $d\omega_t = \mu_{\omega,t}dt + \sigma_{\omega,t}dB_t$*, is Markovian with volatility* $\sigma_{\omega,t} = \sigma_\omega(\omega_t)$ *and drift* $\mu_{\omega,t} = \mu_\omega(\omega_t)$ *given by*

$$\sigma_\omega(\omega_t) = \omega_t\left(w_t^R - 1\right)\sigma_D, \tag{29}$$

$$\mu_\omega(\omega_t) = \omega_t\left(-\mu_D + \sigma_D^2 - \pi + r_t - \rho + w_t^R\left(\mu_t - r_t + \lambda_t(w_t^R)\right) - w_t^R\sigma_D^2\right) + \nu^R\delta. \tag{30}$$

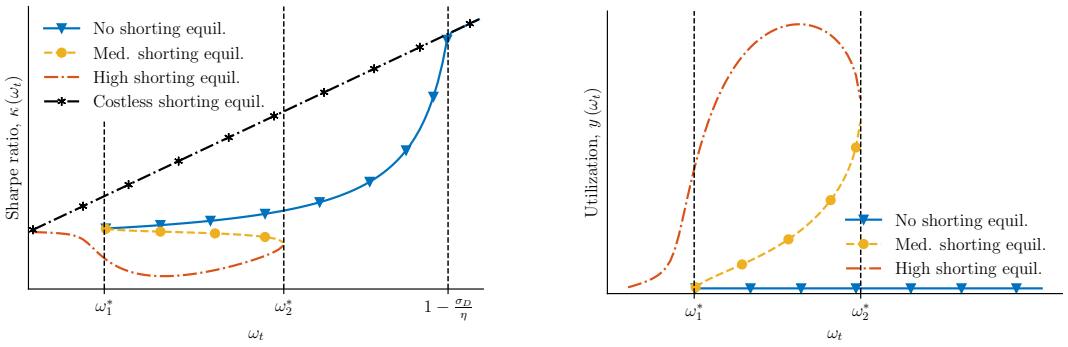


Figure 2: Left: All possible equilibrium values of the Sharpe ratio, as a function of ω_t. Right: The utilization ratio, $y(\omega_t)$, in all of the equilibria as a function of ω_t.

Figure 2 illustrates Proposition 2. The left graph plots $\kappa(\omega_t)$, the Sharpe ratio, as a function of the wealth share of rational agents, ω_t. As a benchmark, the line labeled "Costless shorting equil." depicts $\sigma_D - (1-\omega_t)\eta$, i.e., the Sharpe ratio that would obtain in this economy in the absence of any shorting frictions ($\varphi = 0$). The curve "No shorting equil." depicts the Sharpe ratio in the equilibrium that involves no shorting for the values of ω_t that such an equilibrium exists. Similarly for the curves "Med. shorting equil." and "High shorting equil.", which depict equilibria with shorting for the values of ω_t that permit such equilibria. To expedite the exposition of the results, we postpone a discussion of the quantitative implications of the model until Section 6.2. The graphs in the current section are meant to illustrate qualitative properties of the model.

The figure shows that, when ω_t is larger than $1-\frac{\sigma_D}{\eta}$, the lines "Costless shorting equil." and "No shorting equil." coincide, reflecting that all investors invest strictly positive amounts in the stock market in this region of ω_t.

When ω_t becomes smaller than $1-\frac{\sigma_D}{\eta}$ (but larger than ω_2^*), the rational investor puts zero weight on stocks, but the shorting fee φ deters her from actively short-selling. Since only the irrational investor is marginal in financial markets, the lines "Costless shorting equil." and "No shorting equil." deviate from each other when $\omega_t < 1-\frac{\sigma_D}{\eta}$. In this region the magnitude of the lending fee, φ, does not impact the Sharpe ratio directly (only by deterring the R investors from shorting).

If ω_t becomes smaller than ω_2^* (but larger than ω_1^*) the economy exhibits three equilibria. In the first equilibrium, there is still no shorting. In the second and third, there is active

shorting by the rational investor. Across these three equilibria, the higher the extent of shorting, the lower the Sharpe ratio. This is illustrated in the left graph of Figure 2.

Finally, if ω_t becomes smaller than ω_1^*, then the equilibrium becomes unique and involves shorting.[14]

Remark 1 *The fact that there are three equilibria, one of which features no shorting, is an implication of there being only two types of agents in the model. With more than two types of agents, more than three equilibria can obtain. Also, in the case of multiple equilibria, all of the equilibria can involve strictly positive short interest, as we show in Appendix B.*

Remark 2 *The presence of multiple equilibria implies that the aggregate demand curve for the stock,* $D(\kappa) \equiv W_t^+(\kappa, y(\kappa)) - W_t^-(\kappa, y(\kappa))$ *is a backward-bending function of* κ *(where* $y(\kappa)$ *is implicitly defined by the first line of equation* (20)*). The market-clearing requirement,* $D(\kappa) = 1$*, along with the fact that there are multiple values of* κ *such that* $D(\kappa) = 1$*, implies that* $D(\kappa)$ *is not monotonically declining, but instead is backward bending. As observed by Gennotte and Leland (1990), a backward bending demand curve gives rise to discontinuous changes in equilibrium. This necessary instability is illustrated in Figure 2: when the value of the continuous-path wealth-share process* ω_t *increases from below* ω_1^* *to above* ω_2^**, the processes* κ_t *and* y_t *experience discontinuous changes on the interval* $[\omega_1^*, \omega_2^*]$ *irrespective of how market participants select between high, medium, and no shorting equilibria.*[15]

2.2.1 Multiplicity and amplification

In our model multiplicity is a convenient way to illustrate a mutually reinforcing feedback loop between the Sharpe ratio, κ_t, and utilization, y_t. Before presenting general conditions

14. To see why there can be no equilibrium without shorting when $\omega_t < \omega_1^*$, assume otherwise. Indeed assume that the R investor holds zero stocks and is not marginal in the stock market $\left(w_t^R = 0\right)$. The market clearing requirement, $\omega_t w_t^R + (1-\omega_t) w_t^I = 1$, along with $w_t^I = \frac{\kappa_t + \eta}{\sigma_D}$ implies that the Sharpe ratio would be $\kappa_t = \frac{\sigma_D}{1-\omega_t} - \eta$. Under this supposition, it would therefore be the case that $\mu_t - r + \varphi = \sigma_D \left(\kappa_t + \frac{\varphi}{\sigma^D}\right) = \sigma_D \left(\frac{\sigma_D}{1-\omega_t} - \eta + \frac{\varphi}{\sigma^D}\right) < 0$, where the inequality follows from $\omega_t < \omega_1^*$. Because $\mu_t - r + \varphi < 0$, equation (12) implies that the R investor would want to short the market, contradicting the assumption that she is optimally holding zero stocks.

15. For instance, if the market participants always coordinate on the high shorting equilibrium, the jump will occur when $\omega_t = \omega_2^*$, and if market participants coordinate on the no shorting equilibrium, the jump will occur at ω_1^*.

that can lead to equilibrium multiplicity, in this section we confine attention to situations where the shorting market is active, but the equilibrium is unique. We show that even when the equilibrium is unique, the feedback loop between κ_t and y_t is still present and becomes the source of an "amplification" mechanism.

Specifically, assume that $\omega_t < \omega_1^*$, so that the shorting market is active and the equilibrium is unique. In this region, consider the impact of a change in the parameter η, which governs the optimism of irrational investors, on the Sharpe ratio, κ. Next, note that the market-clearing condition for lending (equation (20)) can be written compactly as $G(y, \kappa; \eta) = 0$, where $G(y, \kappa; \eta) \equiv y\left(\kappa + \eta + \frac{\varphi}{\sigma_D}\tau y\right) + \frac{\omega_t}{1-\omega_t}\left(\kappa + \frac{\varphi}{\sigma_D}\right)$. By the implicit function theorem, $dy_t = -\frac{G_\kappa}{G_y}d\kappa - \frac{G_\eta}{G_y}d\eta$. In turn, totally differentiating equation (18) yields $d\kappa_t = -(1-\omega_t)d\eta - \frac{\varphi}{\sigma_D}\tau(1-\omega_t)\,dy_t$. Combining these two equations yields

$$d\kappa_t = \Lambda d\eta + \Phi d\kappa_t, \tag{31}$$

where $\Phi = \frac{\varphi}{\sigma_D}\tau(1-\omega_t)\frac{G_\kappa}{G_y}$ and $\Lambda = -(1-\omega_t) + \frac{\varphi}{\sigma_D}\tau(1-\omega_t)\frac{G_\eta}{G_y}$.

The quantity Λ captures the "direct" effect of a change in η on κ_t. The presence of the term Φ on the right-hand side of equation (31) illustrates the presence of an "amplification" effect. Indeed, iterated substitution yields

$$\begin{aligned} d\kappa_t &= \Lambda d\eta + \Phi d\kappa_t = \Lambda d\eta + \Phi\left(\Lambda d\eta + \Phi d\kappa_t\right) \\ &= \Lambda(1+\Phi)\,d\eta + \Phi^2 d\kappa_t \\ &= \Lambda\left(1 + \Phi + \Phi^2 + \ldots\right)d\eta = \frac{\Lambda}{1-\Phi}d\eta. \end{aligned} \tag{32}$$

Lemma 2 in the Appendix shows that in the region where the equilibrium is unique, Φ lies between 0 and 1. Equation (32) captures a "multiplier" effect. Specifically, an increase in η has the direct effect of lowering the Sharpe ratio, since the optimists become more optimistic. In addition, this direct effect starts a "spiral" by increasing utilization, y_t, leading to a further reduction in the Sharpe ratio by a fraction $\Phi < 1$ of the original increase, further increasing y, lowering κ by a further Φ^2 of the original effect, etc. The expression for the fraction Φ is given by the product of a) the impact of a change in utilization on the Sharpe ratio,

$-\frac{\varphi}{\sigma_D}\tau\left(1-\omega_t\right)$, and b) the impact of a change in the Sharpe ratio on utilization, $-\frac{G_\kappa}{G_y}$.

The main difference between the region of multiplicity, $\omega_t \in (\omega_1^*, \omega_2^*)$, and that of uniqueness (with non-zero shorting), $\omega_t < \omega_1^*$, is that in the multiplicity region the feedback loop between κ_t and y_t becomes so strong that $\Phi > 1$ for some values of y.[16]

3 Wealth-Share Dynamics

When multiple equilibria are possible, both the drift rate $\mu_\omega\left(\omega_t\right)$ of the wealth share of type R investors and the expected logarithmic growth rate of their wealth are higher in equilibria that feature higher y_t, as the next proposition shows.

Proposition 3 *For a fixed wealth share of the R-agents, ω_t, consider two equilibria A and B with $y_{t,B} > y_{t,A}$ (and accordingly $\kappa_{t,B} < \kappa_{t,A}$). Then, the drift of investor R's wealth share in equilibrium $i \in \{A, B\}$, $\mu_{\omega,i}\left(\omega_t\right)$, satisfies $\mu_{\omega,B}\left(\omega_t\right) > \mu_{\omega,A}\left(\omega_t\right)$. In addition, the drift of the logarithmic growth rate of investor R's wealth, given by*

$$g(\omega_t) \equiv \frac{1}{dt} E\left[d\log(W_t^R)\right] = r_t - \rho + \max_{w \leq 0}\left\{w\left(\kappa_t\sigma_D + \varphi\right) - \frac{1}{2}\left(w\sigma_D\right)^2\right\}, \tag{33}$$

is higher in equilibrium B than in equilibrium A, i.e., $g_B\left(\omega_t\right) > g_A\left(\omega_t\right)$.

Figure 3 provides an illustration of Proposition 3. The figure shows the stationary distribution of ω_t in the equilibrium associated with no shorting for values $\omega_t \in (\omega_1^*, \omega_2^*)$ and in the equilibrium associated with the highest shorting, $y^+(\omega_t)$, for $\omega_t \in (\omega_1^*, \omega_2^*)$. The figure shows that the stationary distribution of ω_t has a higher mean in the high-shorting equilibrium than in the no-shorting equilibrium. This is consistent with Proposition 3, which asserts a higher (logarithmic) growth rate for the wealth of R investors in the second equilibrium.

When comparing a higher-shorting to a lower-shorting equilibrium, therefore, one must account for two competing effects on the stationary mean of the Sharpe ratio κ_t. On the one hand, for a fixed ω_t the Sharpe ratio is lower in the high-shorting equilibrium. On the other

16. For instance, one can show that $\Phi > 1$ for values of y in a neighborhood of y^-, while $\Phi < 1$ in a neighborhood of y^+. An implication is that — using the common definition of stability — the equilibria corresponding to y^+ and $y = 0$ are stable, while the equilibrium associated with y^- is unstable.

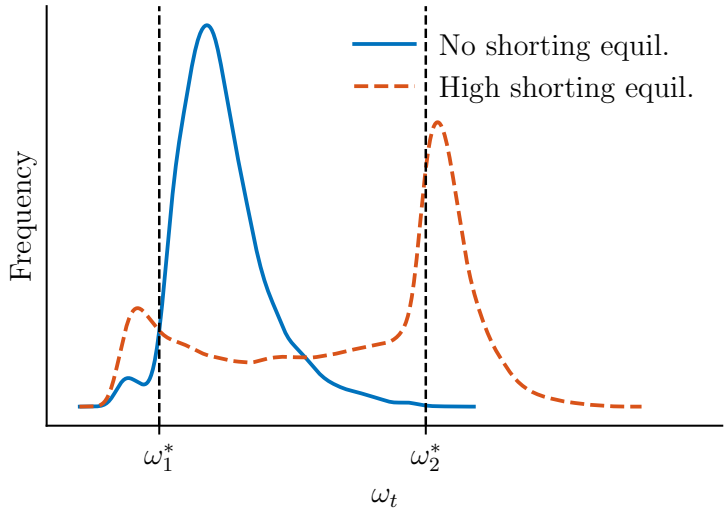


Figure 3: An illustration of Proposition 3. Simulating the model for the cases in which market participants coordinate on the high-shorting and, respectively, the no-shorting equilibrium, the figure depicts the stationary distribution of the wealth share of the rational investor, ω_t, for the economy of Figure 2.

hand, low values of ω_t become infrequent in the high-shorting equilibrium. The first channel makes the stationary mean of the Sharpe ratio lower in the high-shorting equilibrium, but the second channel has the opposite effect. The overall effect on the stationary value of the Sharpe ratio is ambiguous. This observation will become important in Section 6.2, when we discuss the impact of an equilibrium shift on the price-dividend ratio of a small stock.

4 Arbitrary Supply Curve for Lendable Shares

In Section 2 we assumed a perfectly elastic supply curve for lending shares ($f(y) = \varphi$), which allowed us to solve the model in terms of a simple, quadratic equation. Here we revisit our main result, namely the existence of multiple equilibria, for an arbitrary (differentiable) non-decreasing supply curve $f_t = f(y_t)$. In addition, to allay possible fears that our results are special to the discrete nature of the two-type distribution we considered so far,[17] the following proposition allows for a continuous distribution of beliefs (with connected support).

Proposition 4 *Let $h(y) \equiv f(y)(1 - \tau y)$. If there exists a value $y \in [0, 1)$ with (a) $h'(y) < 0$*

17. Note also that Appendix B illustrates multiple equilibria obtaining with three agent types.

and (b) $\sigma_D^2 < \frac{1}{4}(1-y)^2|h'(y)|$*, then there exist wealth distributions over beliefs for which multiple equilibrium values of* y_t *(and* κ_t*) obtain.*

A key role in Proposition 4 is played by the function $h(y)$. This function captures the difference between the (proportional) fee paid by a short seller, $f(y)$, and the (proportional) lending income received by a long investor, $\tau f(y)\, y$. To understand why the condition $h'(y) < 0$ for some $y \in [0, 1)$ is necessary for multiple equilibria, suppose that there are (at least) two equilibria $y_{t,1} < y_{t,2}$. From market clearing in the spot market, in the second equilibrium both long and short investors must choose a portfolio of larger absolute value than in the first equilibrium: $y_{t,1} < y_{t,2} \implies |w_{t,1}^R| < |w_{t,2}^R|$ and $w_{t,1}^I < w_{t,2}^I$.[18]

An immediate consequence is that $w_{t,1}^I - w_{t,1}^R < w_{t,2}^I - w_{t,2}^R$. From the expressions for the investors' portfolio weights, (15) and (16), this inequality is equivalent to $f(y_{t,1})\,(1-\tau y_{t,1}) > f(y_{t,2})\,(1-\tau y_{t,2})$, that is, $h(y_{t,1}) > h(y_{t,2})$. Since f is assumed differentiable, there exists $y \in (y_{t,1}, y_{t,2})$ such that $h'(y) < 0$.

Proposition 4 shows that condition (a) — when combined with condition (b) — is not just necessary, but also sufficient for the existence of multiple equilibria, in the sense that there exist (an open set of) wealth distributions over beliefs that ensure the existence of multiple equilibria. We show that condition (b) is satisfied as long as investor disagreement pertains to the idiosyncratic risk of a small stock relative to the market (see Remark 3 in Appendix E). Henceforth, we always implicitly refer to condition (a) when we write the "condition of Proposition 4."

5 Empirical Evidence

5.1 Overview

The novel intuition of our model is the presence of a feedback effect between utilization (y) and the Sharpe ratio (κ). This feedback loop can lead to equilibrium shifts that empirically

18. The fact that $w_t^R < 0$ along with equations (17) and (19) imply that $y_t = \frac{\omega_t |w_t^R|}{1+\omega_t |w_t^R|}$, which is an increasing function of $|w_t^R|$. Accordingly, $y_{t,1} < y_{t,2}$ implies that $|w_{t,1}^R| < |w_{t,2}^R|$. In turn, by market clearing, $w_t^I = \frac{1+\omega_t |w_t^R|}{1-\omega_t}$, and therefore $|w_{t,1}^R| < |w_{t,2}^R|$ implies $w_{t,1}^I < w_{t,2}^I$.

manifest themselves as jumps in utilization, irrespective of how agents coordinate on one equilibrium or another (Remark 2). Moreover, Proposition 4 states a key condition for the possibility of such jumps to occur, namely that the function $h(y) = f(y)(1 - \tau y)$, which reflects the difference between the fee paid by the short investor and the income received by the long investor, be declining for some y.

The next section shows that the assumption of a (locally) declining $h(y)$ is empirically plausible. In addition, we present evidence that the time series for utilization does exhibit jump-like behavior. Moreover, as predicted by the model, the incidence of utilization jumps for a given stock is significantly correlated with whether the estimated $h(y)$ (for that stock) has a declining segment.

5.2 Data description

Daily returns and market capitalization data are from the Center for Research in Security Prices (CRSP). Our source for stock lending fees and short interest is IHS Markit. These data start in January 2006.[19] Markit collects self-reported data on actual rates on security loans from active participants in the securities lending market. The data set covers roughly 30,000 securities, and contains 16 million unique stock-day observations.

We match the Markit data to the CRSP database and retain only common stocks of domestic companies. Furthermore, to ensure that our results are not driven by micro-cap stocks, for our main empirical results we only retain observations that correspond to stocks that are Russell 3000 constituents (on the day of observation), which we identify using the Datastream Monthly Index Constituents file. This reduces our number of observations to 10 million.

We follow Daniel, Klos, and Rottke (2022) and use the quantity "Indicative Fee" as our measure of the marginal cost of borrowing, which is the expected borrowing cost (expressed in percentage points per year) on a given day.[20] In addition to these data on fees, we use

19. The Markit data of other studies (Daniel, Klos, and Rottke (2022) and Drechsler and Drechsler (2014)) starts in 2004 and contains observations at a weekly frequency. The data set that was provided to us by Markit contains daily observations that start in 2006. Markit confirmed in an email that the pre-2006 data are no longer available.

20. Markit uses both borrowing costs between lenders and prime brokers as well as rates from hedge funds to produce this estimate of the current market rate. As discussed in Daniel, Klos, and Rottke (2022), Indicative

two additional data variables from Markit: a) "Daily Cost of Borrow Score" (DCBS) and b) daily utilization. The DCBS takes integer values between one and ten and is a "bucketed score (1-10) that reflects the cost to borrow the stock charged by the lenders from the Prime Brokers in the wholesale market, where 1 reflects a cheap or a GC ("general collateral") stock and 10 reflects an expensive or a special stock."[21] The literature has used this score as a way of identifying stocks that are on special. In the data, DCBS values equal to one are by far the most prevalent ones (74% of our sample) and tend to exhibit a high degree of persistence.[22] For some of our empirical results, we focus on stocks that are hard to borrow and we drop observations with DCBS equal to 1, since our model applies predominantly to stocks where lending frictions are non-trivial. Markit's "Utilization by Quantity" is computed as the fraction of assets on loan from lenders divided by the total lendable quantity. This variable takes values between 0 and 1 and corresponds to the variable y_t in our model.

5.3 Empirical results

Tables I.1 and I.2 in Appendix I provide some summary statistics on the lending fees. To expedite the presentation of the results that pertain to our paper, here we simply summarize our main findings from these tables as follows. The median lending fee ranges from 0.35% per annum (for stocks in the largest market-capitalization quintile) to 0.41% per annum (for stocks in the smallest market-capitalization quintile). However, lending fees exhibit substantial cross-sectional and time-series variation. The 99^{th} percentile of all lending-fee observations exceeds 7% for stocks in four out of the five market-capitalization quintiles. When we examine lending fees at the individual stock level, we find that 31% of Russell 3000 constituents exhibit a lending fee in excess of 1% for 5 out of 100 trading days, while 18% of Russell constituents exhibit lending fees in excess of 3% for 5 out out of 100 trading days. In addition, 45% of Russell constituents exhibit a fee in excess of 5% at some point in

Fee can be interpreted as a proxy for the marginal cost of short selling. Markit also reports "Simple Average Fee," which is the average fee over all outstanding contracts for a particular security. Following Daniel, Klos, and Rottke (2022), on each stock-day, we take the Indicative Fee as our measure of the stock's lending fee and (in the very rare instances) where the Indicative Fee is not reported, we use Simple Average Fee. This substitution applies to only 676 observations out of the roughly 10 million observations.

21. IHS Markit Securities Finance Quant Summary, July 2020 edition. Available on WRDS.

22. If a stock has a DCBS value of one on any given day, the probability that it has a DCBS value of one the next day is 98.83%

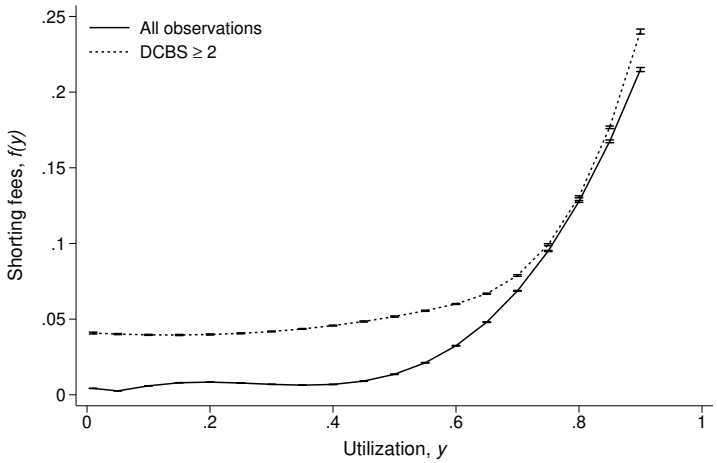


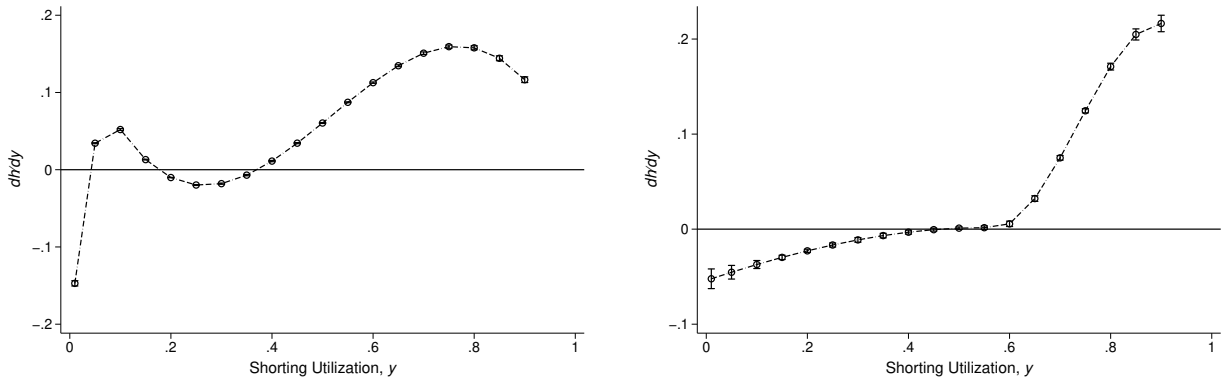
Figure 4: Relationship between shorting fees and utilization. Non-parametric series regression of daily shorting fees on utilization, pooled across Russell 3000 constituents. Daily shorting fees from 2006 to 2021 are from Markit and are reported as annualized percentage rates. (For instance, 0.05 on the y axis means 5% per annum.) Error bars denote 95% confidence intervals. Because this estimation utilizes several millions of observations, the standard errors of the estimates are negligible.

the sample. This is consistent with results reported in Engelberg, Reed, and Ringgenberg (2018), who write "loan fees can increase to levels that significantly decrease the profitability of nearly any trade."

We start the presentation of our main empirical findings with Figure 4, which examines the relationship between utilization and shorting fees. Specifically, the solid line pools all daily observations across all Russell-3000 stocks and depicts the estimates of a non-parametric regression of daily shorting fees (expressed in annual percentage terms) on utilization.[23] The dashed line depicts results from the subsample that only includes observations of stocks with a DCBS code of 2 or above, stocks to which we refer as being on special. As both plots show, the relationship between shorting fees and utilization is nonlinear, with a region that is approximately constant for low and intermediate values of utilization and a steeply increasing region for high values of utilization.

We next use the estimates from the non-parametric regression analysis to compute

23. We estimate a third order basis spline of fees on utilization and depict the point estimates and standard errors at 0.05 increments of utilization, along with standard errors. For the computations we use the command `npregress series` in Stata. Standard errors are produced using the command `margins` and the Δ method.



(a) Full sample

(b) $\text{DCBS}_{i,t} \geq 2$

Figure 5: $h'(y)$, pooled non parametric estimates. Estimated marginal effects are derived from a non-parametric series regression of daily shorting fees f on utilization, y. Marginal effects are computed using the formula $h'(y) = f'(y)\,(1 - \tau\, y) - \tau f(y)$. Short interest is based on utilization data from Markit. Sample consists of daily observations of shorting fees and utilization, pooled across Russell 3000 constituents from Markit for the period 2006 to 2021. Standard errors are derived from the standard-error estimates of $(1 - \tau y)\hat{f}'(y)$ and $\tau \hat{f}(y)$, while assuming a worst-case correlation of -1 between these two quantities. The parameter τ is set to 0.8.

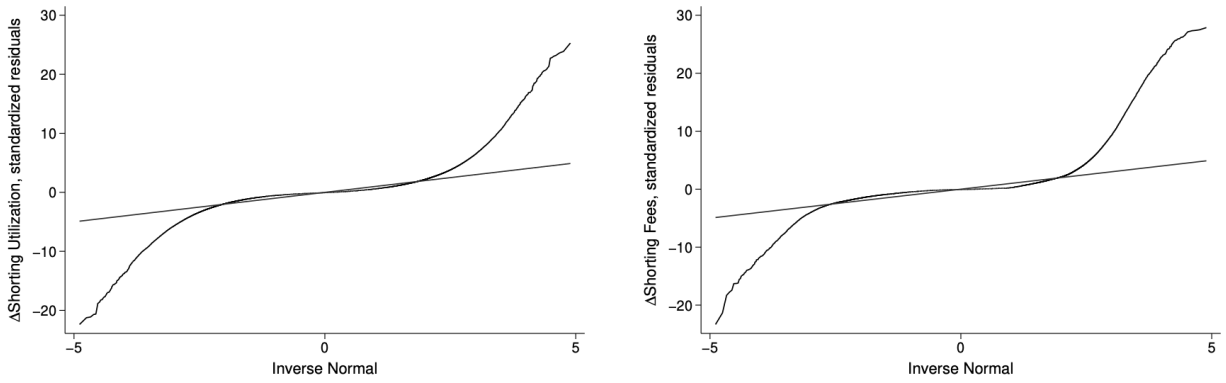
estimates of $h'(y)$ via $h'(y) = f'(y)\,(1 - \tau\, y) - f(y)\tau$. We calibrate τ to a value of 0.8 based on the industry practice of rebating about 80% to the mutual funds that provide their shares for lending.[24] Figure 5 depicts these estimates, along with an upper-bound estimate of the 95% confidence interval.[25] The figure shows that $h'(y)$ is statistically significantly negative for several y values between 0 and 0.4. Therefore, when we pool all observations, we can statistically reject the null hypothesis that $h'(y)$ is always positive.

The two plots of Figure 5 illustrate the results of a single non-parametric regression on pooled data to obtain more precise estimates. As a robustness check, in Appendix I we estimate a separate non-parametric regression of fees on utilization for each stock and compute a stock-specific $h'(y)$. Appendix I shows that the (cross-sectional) average of the estimated $h'(y)$ is negative (and statistically significant) for low values of y.[26] We also note that the

24. Source: "Unlocking the potential of your portfolios: iShares Security Lending." Blackrock, 2021. Available at `https://www.ishares.com`.

25. The standard errors are computed using Stata's estimates for the variance of the estimates $f(y)$ and $f'(y)$ and a worst-case assumption that the correlation between the estimates of $\hat{f}$ and $\hat{f}'$ is -1 to provide an upper bound on the variance of the estimate.

26. As can be expected, the estimation of a separate function $h'(y)$ for each stock increases the estimation-error bounds on $h'(y)$ and therefore the range of (statistically significant) negative y-values becomes smaller than in Figure 5.



(a) Utilization

(b) Lending fee

Figure 6: QQ-plots of weekly changes in utilization and lending fees. Left panel: An AR(1) of utilization is estimated at the stock level at a weekly frequency. The residuals of each time series are divided by their standard deviation and then pooled across all stocks. The quantiles of the standardized-residual distribution are then plotted against the quantiles of the standard normal distribution. Right panel: Same as the left panel, but for lending fees rather than utilization. Both utilization and lending fee data are provided by Markit and cover the period 2006 to 2021.

analysis so far is based on the model assumption that fees are a deterministic function of utilization, $f_t = f(y_t)$, and any residuals (in the data) are the result of orthogonal sampling and surveying errors in the measurement of the indicative lending fee. At the end of this section we revisit this orthogonality assumption.

Our next set of results pertains to the implications rather than the assumptions of our model. We start with the most basic empirical implication: due to the possibility of equilibrium shifts, utilization may exhibit jumps.

Figure 6 provides an informal way to visualize abrupt shifts in utilization in the data. For each Russell 3000 stock, we estimate a separate AR(1) process for weekly utilization, so that both the long-run mean and the persistence of utilization can vary at the stock level. We then normalize the innovations (i.e., the residuals of the AR(1) estimation) by their standard deviation. Assuming that utilization (at the stock level) follows an AR(1) process with normal, homoskedastic increments, these normalized residuals follow a standard normal distribution. The left panel of Figure 6 shows that this is not the case. The quantile-quantile plot of the standardized residuals clearly shows that the innovations to utilization exhibit remarkably fat tails (the excess kurtosis is 26) with a non-trivial mass of the residuals in the

Table 1: Determinants of Utilization Jump Rate

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$1_{\text{Fail} - \text{Satisfy}}$	-5.14***	-4.78***	-3.62***	-5.28***	-5.00***	-3.85***	-2.96***
	(-5.63)	(-5.67)	(-4.30)	(-5.79)	(-5.51)	(-4.37)	(-3.82)
Fee Ctrl	No	Yes	No	No	No	No	Yes
Size Ctrl	No	No	Yes	No	No	No	Yes
Equity Market Ctrls	No	No	No	Yes	No	No	Yes
Financial Ratios	No	No	No	No	Yes	No	Yes
Exchange Ctrls	No	No	No	No	No	Yes	Yes
N	1975	1975	1975	1975	1975	1975	1975

t statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

The variable $1_{\text{Fail} - \text{Satisfy}}$ is the indicator function that takes the value one when a stock belongs in the fifth quintile sorted on the t-statistic of $h'(y)$, as described in the text. In the regressions we include a constant and an additional indicator variable taking the value one if the stock belongs to in quintiles 2 through 4. "Fee Ctrls" indicates whether the time-series average shorting fee is included as a control. "Size Ctrls" indicates whether the stock's market capitalization is included as a control. "Equity Ctrls" indicates whether the stock's variance of returns and time-series average turnover are included as controls. "Fin Ratio Ctrls" indicates whether the stock's book-to-market ratio and leverage are included as controls. "Exchg Ctrls" indicates whether an indicator for NASDAQ traded stocks and an indicator for stocks with traded options are included as controls. Standard errors are heteroskedasticity-robust.

range of -20 standard deviations.[27,28] We note that utilization is not the only fat-tailed time series. The right subplot of Figure 6 shows that the standardized residuals of the lending-fee time-series are similarly fat tailed (the excess kurtosis is 75).

We next turn to the cross-section of stocks and examine the connection between the incidence of jumps and the condition that $h'(y) < 0$ for some y (Proposition 4). To test this connection, we start by estimating a stock-specific jump rate as follows. We fix a cutoff U above which we consider the weekly AR(1) residual in a stock's utilization as economically large. We refer to a week during which the absolute value of the change in utilization exceeds U as a jump event. (Results are quite similar if we consider the raw weekly changes in

27. The Jarque-Bera test rejects normality with a p-value essentially equal to zero.

28. The test proposed by Aït-Sahalia and Jacod (2009), which tests whether the discretely observed utilization data emanated from a continuous-sample-path diffusion process using daily data, rejects the null hypothesis of continuous sample paths for 85% of Russell 3000 constituents.

Table 2: Summary statistics, averaged within t-statistic quintiles.

	Quintile				
	1	2	3	4	5
Shorting Fee	0.063 (0.097)	0.041 (0.075)	0.044 (0.070)	0.043 (0.080)	0.057 (0.087)
Size Quintile					
1	270 (68.4%)	275 (69.6%)	262 (66.3%)	257 (65.1%)	220 (55.7%)
2	91 (23.0%)	80 (20.3%)	83 (21.0%)	76 (19.2%)	86 (21.8%)
3	19 (4.8%)	29 (7.3%)	31 (7.8%)	32 (8.1%)	48 (12.2%)
4	14 (3.5%)	7 (1.8%)	14 (3.5%)	19 (4.8%)	27 (6.8%)
5	1 (0.3%)	4 (1.0%)	5 (1.3%)	11 (2.8%)	14 (3.5%)
Var of Returns	0.003 (0.008)	0.004 (0.035)	0.003 (0.004)	0.002 (0.003)	0.003 (0.016)
Turnover	0.012 (0.013)	0.013 (0.016)	0.014 (0.019)	0.014 (0.020)	0.016 (0.017)
Debt/Total Assets	0.206 (0.228)	0.226 (0.232)	0.218 (0.220)	0.232 (0.236)	0.236 (0.240)
Log Book/Market	-1.089 (1.035)	-0.995 (0.926)	-1.070 (0.973)	-1.017 (0.949)	-1.087 (0.991)
$\mathbf{1}_{\text{Option}}$	0.113 (0.299)	0.136 (0.320)	0.128 (0.319)	0.179 (0.367)	0.239 (0.411)
$\mathbf{1}_{\text{NASDAQ}}$	0.709 (0.455)	0.701 (0.458)	0.742 (0.438)	0.676 (0.469)	0.615 (0.487)

Stocks are sorted into five quintiles based on the t-statistic of $h'(y)$, as described in the text. The table presents within-quintile averages of the control variables included in Table 1. For the variable "Size Quintile", the number in parentheses indicates the percentage of the stocks in that quintile. For all other variables, the number in parentheses indicates the standard error.

utilization rather than the AR(1) residuals.) The cutoff U corresponds to an unusually large (above two standard deviations) weekly change in utilization. We define the jump rate as the ratio of the number of jump events to the total number of weeks over which we observe the stock, which we then annualize for ease of interpretation. Given our interest in stocks that are on special, we confine attention to Russell 3000 stocks that have a DCBS score larger than one for at least 50 trading days. To gauge whether a stock is likely to satisfy the condition of Proposition 4 or not, we obtain a stock-specific non-parametric estimate of $h'(y)$, its standard error, and the associated t-statistic at various levels of utilization.[29] We then compute for each stock the lowest value of the t-statistic across utilization levels and sort stocks into five quintiles based on this quantity.[30] To mitigate the risk of misclassifying stocks as satisfying (failing) the condition of Proposition 4, when in fact they fail (satisfy) it, in all of our regressions we compare stocks in the first and fifth quintiles. (The mean (median) value of t for stocks in the first quintile is -11.6 (-9.4) and the mean (median)

29. For the estimation of $h'(y)$ at the stock level, we use the kernel estimator described in Appendix I. For each stock, we estimate $h'(y)$ at 11 values of shorting utilization (corresponding to each of the nine deciles, as well as the 5^{th} and 95^{th} percentiles of each stock's y).

30. We focus on the minimum value of the t-statistic because the null hypothesis is that $h'(y) \geq 0$ for all y.

Table 3: Alternative Specifications of the Utilization Jump Rate

	Dependent variable: Utilization Jump Rate		
	(1)	(2)	(3)
Panel A: Baseline Specification			
$1_{\text{Fail − Satisfy}}$	-2.96^{***}	-2.25^{***}	-2.03^{**}
	(-3.82)	(-2.76)	(-2.47)
N	1975	1975	1975
Panel B: Demand-Driven Jumps in Utilization			
$1_{\text{Fail − Satisfy}}$	-1.62^{***}	-1.27^{**}	-1.11^{**}
	(-3.32)	(-2.51)	(-2.18)
N	1975	1975	1975

Notes: t-statistics in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

The table re-runs the specification (7) of Table 1 for alternative definitions of the jump rate. In Column (1), jumps are identified as trading weeks during which the absolute value of the unanticipated change in shorting utilization exceeds the full-sample 95$^{\text{th}}$ percentile. In Columns (2) and (3), the cutoff is set at the 99$^{\text{th}}$ and 99.5$^{\text{th}}$ percentiles, respectively. In Panel A, all trading weeks exceeding the column-specific cutoff are considered jumps. In Panel B, only jumps in utilization that are demand-driven are considered. We label a jump as demand-driven if the absolute value of the percentage change in shorted shares exceeds the absolute value of the percentage change in lendable shares. Standard errors are heteroskedasticity-robust.

value of t for stocks in the fifth quintile is 7.8 (6.2).)

Table 1 provides a first test of our theory. We regress the utilization jump rate on an indicator variable taking the value one if the stock belongs in quintile 5, an indicator variable taking the value one if the stock belongs to quintiles 2 through 4, and a constant. (As a result, quintile 1 is the base quintile.) We report the coefficient and standard error of the first indicator variable as $1_{\text{Fail − Satisfy}}$, reflecting the notion that stocks in the fifth quintile are quite likely to fail the condition of Proposition 4, while stocks in the base quintile (first quintile) are likely to satisfy it. Column (1) of the Table reports the results of this regression without any additional controls. Columns (2)–(6) add different groups of control variables to this regression.[31] To save space, in the text we simply list the controls, while Table I.3 in the appendix contains the estimated coefficients and t-stats of these controls. Specifically, Column (2) includes the stock's average shorting fee as a control. Column (3) adds four

31. The controls are motivated by the empirical literature. (See, e.g., Blocher, Reed, and Van Wesep (2013).)

dummy variables corresponding to the quintile of the market-capitalization distribution that the stock belongs to. Column (4) includes the stock's weekly-return variance and average turnover.[32] Column (5) includes a stock's average log-book-to-market ratio and leverage ratio as additional controls. Column (6) includes two dummy variables, namely whether the stock is traded on the NASDAQ and whether the stock has traded options. Column (7) includes all controls together. Irrespective of the specification, the coefficient on the indicator variable $1_{\text{Fail - Satisfy}}$ is always negative and significant at the 1% significance level, indicating that the stocks that are quite likely to fail the condition of Proposition 4 (quintile 5) exhibit a lower jump rate compared to stocks that are likely to satisfy the condition (quintile 1).[33] The fact that the inclusion of controls does not impact significantly the coefficient on the indicator variable $1_{\text{Fail - Satisfy}}$ suggests that the various stock characteristics are more or less unrelated to the quintile to which a stock belongs. Table 2 confirms this conclusion by showing that stocks in different quintiles have roughly similar characteristics.

Table 3 reports robustness results when we use alternative definitions of the jump rate. The top panel of Table 3 repeats the regression of Column (7) of Table 1 but for progressively higher jump cutoffs U. The bottom panel of Table 3 performs a similar exercise to the top panel of the table, except that in order to define a jump in utilization, we impose an additional property of the model: we confine attention to jumps in utilization whereby the absolute value of the percentage change in shorted shares (the numerator of utilization) is larger than the absolute value of the percentage change in lendable shares (the denominator of utilization).[34] Taken together, Table 3 confirms that the results of Table 1 are not sensitive to the precise definition of the jump rate. Finally, to allow for potential time variation of utilization volatility at the stock level, we also used the rolling, jump-robust estimator of the

32. Turnover is defined as daily volume divided by market capitalization. Source: CRSP.

33. Results are qualitatively unchanged if we regress the utilization jump rate on the t-statistic that we use to form the quintiles instead of forming quintiles. Results are also unchanged if we form groups based on cutoffs of the t-statistics and form three groups based on whether the value of t is below -3, between -3 and 3, and above 3. Similarly, if we include four indicator variables, one corresponding to each quintile of the t value (rather than grouping stocks in quintiles 2–4 together) the results are essentially identical.

34. Inside the model, utilization is equal to the ratio of shorted shares to lendable shares; in turn lendable shares are equal to one plus the absolute value of shorted shares (this is the market clearing condition). Therefore, the absolute value of the percentage change in shorted shares must exceed the absolute value of the percentage change in lendable shares. By restricting attention to jumps that satisfy this requirement, we exclude changes in utilization due, for example, to some institutional client giving permission to short shares.

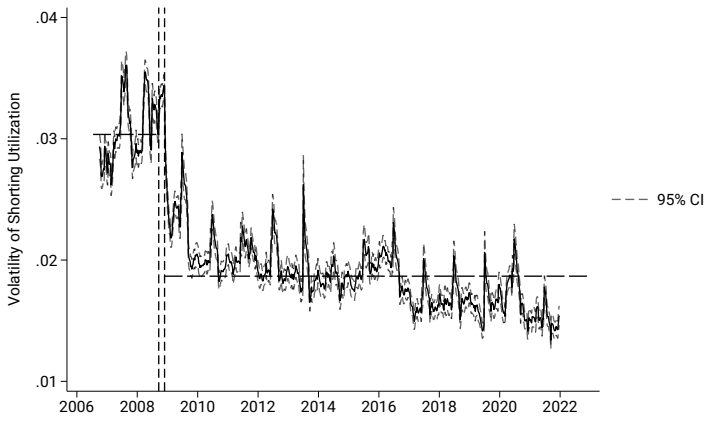


Figure 7: Volatility of Shorting Utilization. For each stock, we fit an AR(1) model for weekly shorting utilization, extract the residuals, and calculate the 10-week rolling standard deviation of residuals. We plot the cross-sectional average of this standard deviation of shorting utilization, along with a 95% confidence interval. The two vertical lines denote the beginning and end of the ten week period following the SEC announcement on September 17, 2008.

local volatility of utilization from Wang and Zheng (2022) to identify jump events as trading days where the absolute value of the change in utilization exceeds four times the estimated local utilization volatility of each stock. Using this alternative definition of the jump rate, the coefficient on $1_{\text{Fail} - \text{Satisfy}}$ in Column (7) of Table 1 remains significant (t-stat: -3.65).

The evidence in Tables 1 and 3 is cross-sectional in nature. This leaves open the possibility that the cross-sectional differences in jump intensity are driven by some omitted factor. To address this issue, we next report results from regressions that exploit a structural break in our sample. This structural break, which we describe shortly, allows us to show that the condition of Proposition 4 is useful at predicting within-stock changes in utilization jump rates (i.e., it effectively allows us to control for stock fixed effects).

The structural break in our sample is associated with the 2008 modification in SEC's regulation SHO that strengthened the regulatory delivery requirements on short selling.[35] Figure 7 provides a visual impression of how this regulatory change altered the time series of the volatility of shorting utilization. Specifically, at each point in time we compute a

35. On September 17, 2008, the SEC released Order No. 34-58572, which imposed enhanced delivery requirements as well as penalties for having a fail-to-deliver position on any equity security. Specifically, Rule 204T requires that: (i) fail-to-deliver positions must be closed on the next settlement day, and (ii) "the participant and any broker or dealer from which it receives trades for clearance and settlement [...] may not accept a short sale order in the equity security from another person, or effect a short sale in the equity security for its own account [...] until the participant closes out the fail to deliver position".

ten-week rolling standard deviation of utilization innovations for each stock and report the cross-sectional mean along with a 95% confidence interval. The figure shows that between July 2006 and September 2008 (the portion of our sample before the enactment of the new regulation) the volatility of the shorting utilization is markedly higher than in the 14 years that follow. The drop in utilization volatility occurs concurrently with the enactment of the regulation, and the volatility remains at these low levels long after the end of the great financial crisis (GFC), which suggests that the drop is not a temporary reaction to the financial turmoil of the GFC. (To interpret this drop in volatility through the lens of the model, in Appendix Section G we extend the baseline model to allow for an additional non-pecuniary cost of short selling, which we interpret as a regulatory cost; we show that a rise in this cost is consistent with a drop in the volatility of utilization.)

Table 4 examines the different behavior of stocks that likely satisfy (quintile 1) and fail (quintile 5) the condition of Proposition 4 around the regulatory change. The dependent variable in Column (1) is the difference in a stock's utilization volatility between the pre-regulation and the post-regulation sample. The regressors are the same indicator variables as in Column (1) of Table 1 and a constant. Column (1) of Table 4 shows that the constant is negative and significant, indicating that utilization volatility dropped significantly for the base quintile (quintile 1). By contrast, the indicator variable $1_{\text{Fail} - \text{Satisfy}}$ is insignificant, indicating that the utilization volatility dropped by similar amounts for stocks that are likely to fail and satisfy the condition of Proposition 4. In other words, the condition of Proposition 4 has no power for explaining within-stock changes in the volatility of utilization. The situation is different for the utilization jump rate. In Column (2), we use the same regressors, but the dependent variable is the difference in a stock's utilization jump rate between the pre- and the post-regulation sample. (We keep the jump cutoff, U, unchanged in the two subsamples.) The pattern in Column (2) is the opposite from Column (1). The constant is insignificant; the indicator variable $1_{\text{Fail} - \text{Satisfy}}$, however, has a negative and significant coefficient. The insignificant constant indicates that stocks in the base quintile (stocks that likely satisfy the condition of Proposition 4) experienced an insignificant change in the utilization jump rate, while the significant indicator variable implies that the stocks that likely fail the condition of Proposition 4 experienced a significant drop in the jump rate. Columns (3)–(5) show that the

conclusions of Column (2) for the jump rate continue to hold when we include controls for the change in the level of utilization (Column (4)) and the change in the volatility of utilization (Column (5)), as well as these two controls together (Column (6)). That is, comparing two stocks that experienced similar changes in their level and volatility of utilization, the stocks that likely satisfy the condition of Proposition 4 (quintile 1) experienced a significantly smaller (and insignificant) drop in their utilization jump rate compared to stocks that likely fail the condition.

It is intriguing that stocks in the base quintile (stocks that likely satisfy the condition of Proposition 4) experience a significant drop in utilization volatility (and similar in magnitude to the stocks that likely fail the condition), but experience only a small and insignificant change in the utilization jump rate. To interpret this finding, we start by noting that in our model utilization is given by the sum of two components: (a) a continuous diffusion (for all stocks) and (b) a pure-jump process (only for stocks that satisfy the condition of Proposition 4). Because our observations are in discrete time, we identify jumps as utilization changes that exceed a large cutoff, U, in absolute value. This identification implies the possibility of both type-I errors (unusually large realizations of the diffusive component that are not jumps) and type-II errors (jumps that are missed because of an offsetting realization of the diffusive component). According to our model, a rise in the regulatory cost reduces the volatility of the diffusive component across all stocks (Appendix G), and by implication reduces both type-I and type-II errors of identifying a jump. The reduction in type-I errors implies a reduction in the measured jump rate for all stocks. However, for stocks that satisfy the condition of Proposition 4 there is an opposing effect, namely the reduction in type-II errors (a larger number of correctly identified jumps), thus resulting in a smaller, indeed insignificant, effect on the measured jump rate. We elaborate on this point further in Appendix C, where we also provide a graphical illustration of the argument.

We conclude this section with a robustness check pertaining to the estimation of $h'(y)$. In the model the only shocks are exogenous dividend shocks, which cause shifts in the wealth distribution, the Sharpe ratio, and the demand for shorting shares. Further, the relation between the fee and the utilization is deterministic: $f_t = f(y_t)$. In the data, there is a residual ε_t in that relation, $f_t = f(y_t) + \varepsilon_t$, which throughout this section we have treated as

Table 4: Evidence from the regulatory change

	Dependent variable				
	Δ Volatility	Δ Jump Rate			
	(1)	(2)	(3)	(4)	(5)
$1_{\text{Fail − Satisfy}}$	−0.02	−4.24***	−4.87***	−3.59**	−3.00**
	(−0.81)	(−2.61)	(−2.92)	(−2.40)	(−2.04)
Constant	−0.09***	−2.07	−0.92	0.63	−0.28
	(−3.86)	(−1.57)	(−0.62)	(0.51)	(−0.22)
N	717	717	717	717	717

Notes: t-statistics in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

The dependent variable in Column (1) is the change in each stock's volatility of utilization (from before to after the regulatory change). In Columns (2)–(5) the dependent variable is the change in each stock's utilization jump rate (using the same jump-cutoff, U, before and after the regulatory change). The regressors in Columns (1) and (2) are the same indicator variables as in Column (1) of Table 1. Column (3) adds the change in the level of shorting utilization as an additional control to the specification of Column (2). Column (4) adds the change in the volatility of shorting utilization to the specification of Column (2). Column (5) adds changes in both the level and the volatility of shorting utilization to Column (2). Standard errors are heteroskedasticity-robust.

(orthogonal) measurement error caused by sampling the indicative lending fee. However, if one were to think of this ε_t as a non-orthogonal supply shock, the empirical estimates of $f'(y)$ could be biased upwards or downwards. The model offers a relatively simple approach to consistently estimate $\frac{\Delta f_t}{\Delta y_t}$ even in the presence of such supply shocks. The idea is to exploit the discontinuities that occur around equilibrium shifts:[36] Assuming that dividend shocks and the shock to the lending fee, ε_t are continuous processes, jumps in y_t can only be the result of an equilibrium shift. We can therefore measure $\frac{\Delta f_t}{\Delta y_t} = \frac{f_{t^+} - f_{t^-}}{y_{t^+} - y_{t^-}} = \frac{f(y_{t^+}) - f(y_{t^-}) + \varepsilon_{t^+} - \varepsilon_{t^-}}{y_{t^+} - y_{t^-}} \approx f'(y_{t^-})$, where we used the continuity assumption $\varepsilon_{t^+} = \varepsilon_{t^-}$. In other words, around jump events in utilization one is able to identify $f'(y_t)$. Figure I.2 in the Appendix repeats the exercise of Figures 4 and 5 except that the changes in the numerator and denominator of $\frac{\Delta f_t}{\Delta y_t}$ are evaluated on the weeks where y_t exhibits jump events, as identified earlier. The figure shows that the derived function $h'(y)$ is small and negative for most values of y. This suggests that the conclusion of Figures 4 and 5 is not the result of a bias due to joint-determination issues.

36. The idea is reminiscent of how Sweeting (2006) uses multiple equilibria to resolve identification issues.

6 Extensions: The Price-Dividend Ratio

In the baseline model, the price-dividend ratio and the volatility of the stock market are both constant and independent of the equilibrium on which the investors coordinate. This is an implication of a) logarithmic utility over intermediate consumption, which implies a unitary intertemporal elasticity of substitution (IES) and b) a single asset in positive net supply. In this section we consider two extensions of the baseline model that relax each of these two assumptions. The goal is to show how the dynamics of the wealth share, ω_t, can lead to situations where the price-dividend ratio is comparatively higher when agents coordinate on not shorting (whenever multiple equilibria are possible) rather than maximal shorting. This result obtains even though the Sharpe ratio is typically higher in the no-shorting equilibrium for any fixed ω_t.

6.1 Non-unitary elasticity of substitution

In Appendix D we generalize the baseline model by retaining the assumption that investors have unit risk aversion, but employing an Epstein-Zin-Weil specification to allow the IES to differ from one.[37] The unit risk aversion assumption allows us to preserve the "myopic" mean-variance portfolio equations (12). As a result, equations (18) and (20), which give the Sharpe ratio (κ_t) and utilization (y_t), respectively, remain unchanged, along with the possibility of multiple solutions.[38] However, by removing the unit-IES assumption, the consumption-to-wealth ratios of investors R and I are no longer constant; instead they depend on each investor's perceived evolution of their future wealth-growth rates. In turn, the dividend-to-price ratio is time varying, since it is a wealth-weighted average of the two consumption-to-wealth ratios.[39]

Figure D.1 in the Appendix compares the price-dividend ratio where investors coordinate

37. Specifically, we use the continuous-time version of Epstein-Zin-Weil utilities, proposed by Duffie and Epstein (1992) and Schroder and Skiadas (1999).

38. While the structure of the equations for κ_t and y_t remains unchanged, one complication is that the volatility of the stock market is no longer σ_D, but rather is an endogenous function of ω_t. To obtain this function, one needs to solve a system of ordinary differential equations that are provided in the appendix.

39. This statement follows from a standard market-clearing argument. To see this, let C_t^R and C_t^I denote the total consumption of the two investor types, W_t^R and W_t^I their aggregate wealth, and g_t^R and g_t^I their respective consumption-to-wealth ratios. Dividing the consumption-market clearing requirement, $C^I + C^R = D_t$, by the

on the no-shorting equilibrium with the price-dividend ratio where investors coordinate on the highest shorting equilibrium (whenever ω_t is in the range where multiple equilibria are possible). The figure shows that, if IES < 1, the price-dividend ratio is higher when investors coordinate on the no-shorting outcome, which we explain as follows. In the equilibrium where the shorting market is active whenever possible both agent types (rational and irrational) believe that their own wealth-growth rate will be higher compared to the equilibrium where the shorting market is inactive: An active shorting market allows investors an effective way to trade on their views. Since each agent believes that the other type of agent is wrong, she anticipates that her own wealth-growth rate will be higher when she can trade with the other type in an active shorting market.[40] When the income effect dominates the substitution effect (IES < 1), the perception of high future wealth-growth rates in the high-shorting equilibrium raises the consumption-to-wealth ratio of both investors. As a result, for any fixed wealth weights ω_t and $1-\omega_t$, the dividend-to-price ratio is higher — equivalently, the price-dividend ratio is lower — in the equilibrium in which coordination on the shorting equilibrium is more prevalent.[41]

The above discussion provides a first illustration of the role of anticipated wealth dynamics for the determination of the price-dividend ratio.[42] The discussion also illustrates that the price-dividend ratio may be higher in the no-shorting equilibrium, even though the Sharpe ratio is also higher. Admittedly, this outcome relies on general equilibrium effects, and in

stock-market clearing requirement $W_t^R + W_t^I = P_t$, implies

$$\frac{D_t}{P_t} = \frac{C_t^I}{P_t} + \frac{C_t^R}{P_t} = \frac{C_t^I}{W_t^I}\frac{W_t^I}{W_t^R + W_t^I} + \frac{C_t^R}{W^R}\frac{W_t^R}{W_t^R + W_t^I} = (1-\omega_t)g_t^I + \omega_t g_t^R.$$

40. In Section 3 we proved this result for agent R, and the argument holds mutatis mutandis through the lens on agent I.

41. This conclusion is reversed if the IES is larger than one. The literature is not conclusive on whether the IES is above or below one. See, e.g., Beeler and Campbell (2012) and the response by Bansal, Kiku, and Yaron (2012).

42. As an additional illustration of the role of wealth dynamics, in Appendix D we present a version of the model where the equilibrium switches frequently between high- and no-shorting on a partition of $\omega_t \in [\omega_1^*, \omega_2^*]$ into equal-sized intervals. The key takeaway from that exercise is that because ω_t is more volatile on the sub-intervals where the market coordinates on high shorting, market participants expect that ω_t will end up spending more time on sub-intervals where the market coordinates on no shorting. As a result the price-dividend ratio in the economy of frequent equilibrium switches resembles more closely the economy where the market always coordinates on no shorting rather than the economy where the market coordinates on high shorting.

particular the fact that the interest rate is lower in the no-shorting than in the high-shorting equilibrium for any given value of the wealth share, ω_t.[43] Such general equilibrium effects may not apply in situations where the shorting frictions affect only a small set of stocks. We address this issue next.

6.2 A limiting economy with a small and a large stock

In Appendix E we use logarithmic preferences (as in the baseline version of the model) and analyze a two-stock version, where the first stock is subject to the same shorting frictions as in the baseline model, while the second one isn't. Here we specialize this two-stock setup to a situation where (a) the first stock produces a dividend that is small compared to the dividend of the second stock, and (b) only a small fraction of the population expresses an active demand ("participates") in the market for the small stock. We proceed to briefly sketch the setup of the model and show how the dynamics of the wealth shares in the high- and no-shorting equilibria affect the price-dividend ratio of the small stock. We provide a complete analysis in Appendix F.

Specifically, we assume that there are two types of trees, namely "small" trees (type-1 trees) and "large" trees (type-2 trees). Type-2 trees have dividends similar to the baseline model, namely $D_{2,t,s} = \phi_2 \delta_2 D_{2,t} e^{-\delta_2 (t-s)}$, where $\phi_2 > 0$, $\delta_2 > 0$, s is the vintage of the tree, and $D_{2,t}$ follows a geometric Brownian motion, $\frac{dD_{2,t}}{D_{2,t}} = \mu_{2,D} dt + \sigma_{2,D} dB_{2,t}$, with drift $\mu_{2,D} > 0$. Type-1 trees produce dividends $D_{1,t,s} = \phi_1 \delta_1 D_{2,s} e^{-\delta_1 (t-s) + \sigma_{1,D}(B_{1,t} - B_{1,s})}$, with $\phi_1 > 0$ and $\delta_1 > 0$. The innovations $dB_{1,t}$ can be thought of as dividend innovations that are specific to stock 1 ("stock-1-specific risk"). With the above dividend specifications, the ratio of all type-1 trees' dividends to all type-2 trees' dividends, $\frac{D_{1,t}}{D_{2,t}}$, is stationary and given by

$$\frac{D_{1,t}}{D_{2,t}} = \frac{\int_{-\infty}^{t} D_{1,t,s} ds}{\int_{-\infty}^{t} D_{2,t,s} ds} = \frac{\phi_1}{\phi_2} \int_{-\infty}^{t} \frac{D_{2,s}}{D_{2,t}} \delta_1 e^{-\delta_1 (t-s) + \sigma_{1,D}(B_{1,t} - B_{1,s})} ds. \tag{34}$$

When type-1 trees are small compared to type-2 trees $\left(\frac{\phi_1}{\phi_2} \approx 0 \right)$, aggregate consumption $D_{1,t} + D_{2,t}$ is approximately equal to the aggregate dividends of the large, type-2 trees, and

43. Figure D.2 in Appendix D depicts the interest rate as a function of ω_t across the two equilibria.

therefore aggregate consumption follows a geometric Brownian motion. The implication is that the interest rate and the risk premium for type-2 trees both converge to constants as the ratio $\frac{\phi_1}{\phi_2}$ goes to zero.

In the baseline model, entry and exit of investors into the single stock market is tied to the arrival and departure of agents in the economy and is exogenous. The extension to two risky stocks requires that we model the entry and exit into the market for stock 1. For reasons of realism, we assume that only a (small) fraction of investors, $\widehat{\omega}$, participate in the market for stock 1 at any given point in time:[44] Participants in the market for stock 1 optimize their holdings of stocks 1 and 2 and the bond. By contrast, non-participants express a zero demand for stock 1 and only optimize their holdings of stock 2 and the bond. As we show in greater detail in the appendix, when the wealth share of participants, $\widehat{\omega}$, is sufficiently small (i.e., proportional to the market-capitalization share of stock 1), the risk premium for bearing the stock-1-specific risk does not converge to zero (as $\frac{\phi_1}{\phi_2} \approx 0$): Intuitively, while the stock-specific risk of stock 1 is small from an aggregate perspective (in the sense that aggregate consumption is unaffected by it), the fraction of the population that bears the stock 1-specific risk is also small, and therefore this risk carries a risk premium.

Per unit of time, a measure of investors with wealth share $\theta\widehat{\omega}$ are drawn from the general population of all investors at random and become participants in the market for stock 1.[45] Of the arriving investors, a fraction ν are of type R and $1-\nu$ are of type I, as in the baseline model. Specifically, while all investors agree on the dynamics of the Brownian motion $B_{2,t}$, R investors (correctly) believe that the stock-1-specific Brownian motion $B_{1,t}$ has no drift, while the irrational investors, I, believe that $B_{1,t}$ has a drift equal to η.

To remain active participants in stock 1, investors must incur a small, non-pecuniary, disutility flow ε capturing an attention cost. This cost may result in endogenous exit. Specifically, an investor of type $i \in (I, R)$ chooses to keep paying attention to stock 1 if and only if her expected net utility from remaining attentive to stock 1 an optimally chosen period of time strictly exceeds the attention cost over that period.

44. This assumption is in the spirit of the "limited recognition hypothesis" of Merton (1987).

45. The assumption that the hazard rate of entry "scales" with $\widehat{\omega}$ ensures that the ratio of the wealth of market-1-participants to the market-capitalization of stock 1 approaches a finite limit, and therefore the stock-1 specific risk does not disappear. See Appendix F.

Assuming that ε is sufficiently close to zero, the attention cost is irrelevant for investors of type I: these investors' perceived benefit from being able to invest in stock 1 is bounded away from zero. For investors of type R, however, there are regions of ω_t where the optimal holding of stock 1 is zero, and even a small disutility can lead them to exit the market.

Formally, the net value that an investor of type R derives from being in the market for stock 1 equals

$$V^R(\omega_t) \equiv \mathrm{E}_t \left[\max_{w_u, T} \int_t^T e^{-\rho(u-t)} \left(w_u \left(\mu_{1,u} - r_u + \lambda_u(w_u) \right) - \frac{1}{2} \left(w_u \sigma_{1,u} \right)^2 - \varepsilon \right) du \right], \quad (35)$$

where $T \geq t$ is the stochastic time of exit from the market for stock 1 (which can be endogenous, or occur exogenously with a hazard rate of θ). Equation (35) uses the assumption of logarithmic preferences to express the net expected utility gain from continued presence in market 1 as the increase the investor's logarithmic growth rate of wealth, $w_{1,u}^R \left(\mu_u - r_u + \lambda_u^R \right) - \frac{1}{2} \left(w_{1,u}^R \sigma_u \right)^2$, net of the flow disutility ε.[46] For given equilibrium functions $\kappa(\omega_t)$ and $y(\omega_t)$, there is a critical boundary $\overline{\omega}$, with the property $V^R(\overline{\omega}) = 0$, typically lying in the region of ω_t where $w_{1,u}^R(\omega_t) = 0$, that acts as a "reflecting barrier" for ω_t. Specifically, if the process ω_t were to ever exceed $\overline{\omega}$, there would be enough exit to restore ω_t to $\overline{\omega}$.[47,48]

Appendix F contains further technical details on entry and exit,[49] as well as a derivation of the differential equation obeyed by the price-dividend ratio, which depends on (a) whether investors coordinate on the high- or the low-shorting equilibrium and (b) on the fraction of market-1 participants who are rational, ω_t. Here we briefly illustrate the solution and discuss its properties. We are interested in situations where the disagreement is large ($\eta = 0.9$), and the speed of investor churn in market 1 is quite large ($\theta = 1$), to capture short-termism. The idiosyncratic dividend volatility is not too large, $\sigma_{1,D} = 7\%$, and the shorting fee is at the

46. To arrive at equation (35), we also use the fact that in the limit where stock 1 becomes small, the independence of the brownian motions $B_{1,t}$ and $B_{2,t}$ also implies that the returns of the two stocks are independent (Appendix E).

47. This behavior is reminiscent of models of industry equilibrium with endogenous entry and exit (e.g., Leahy (1993) and Baldursson and Karatzas (1996).)

48. Once an investor exits the stock for market 1, she could in principle re-enter the market for stock 1 at a future time. However, the rate of entry from the general population of stock-1-non-participants is proportional to $\widehat{\omega}$, which in turn approaches zero in the small-stock limit. Therefore, the likelihood of re-entry is negligible.

49. In a nutshell, the entry and exit assumptions ensure that the wealth share of the general population of investors who participate in market 1, $\widehat{\omega}$, is constant over time, while the fraction of rational investors who participate in market 1, ω_t, is time-varying and stationary.

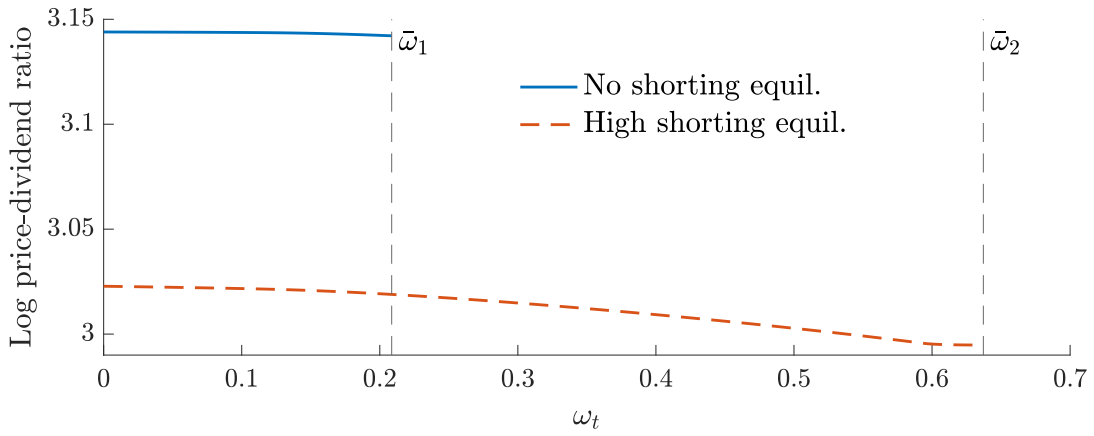


Figure 8: The line labeled "No shorting equil." (resp. "High shorting equil.") depicts the log-price-dividend ratio if investors coordinate on no (resp. highest) shorting whenever the wealth share of rational investors in market 1, ω_t, is in a region allowing for multiple equilibria (values of ω_t between ω_1^*=0.21 and ω_2^*=0.59 in this specific calibration). The dotted line $\overline{\omega}_1$ depicts the exit boundary at which $V^R(\overline{\omega}_1) = 0$ in the "no-shorting" case and $\overline{\omega}_2$ depicts the exit boundary in the "high-shorting" case. The exit boundaries satisfy the inequalities $\omega_1^* < \overline{\omega}_1$ and $\omega_2^* < \overline{\omega}_2$.

high levels that one encounters for stocks that are "on special" ($\varphi = 5.7\%$). A proportion $\nu = 0.7$ of new investors are of type R. In equilibrium, this value of ν ensures that the endogenous exit decision is meaningful, that is, under any equilibrium there is a possibility that ω_t "spends time" in a region where a zero holding of asset 1 is optimal for investor R. We assume that the sum of the interest rate and depreciation rate for stock 1, $r + \delta_1$, is 0.1. We set $\tau = 0.8$, as in Section 5.3. Finally, for the disutility ε we intentionally choose a very small amount (0.1 basis points on an annual basis).[50]

Figure 8 shows the log price-dividend ratio under two different assumptions on the equilibrium on which investors coordinate. Specifically, the line "zero shorting" assumes that investors always coordinate on the equilibrium with zero shorting, if one exists. By contrast, the line "high shorting" assumes that investors always coordinate on the equilibrium with the highest possible shorting, if one exists. Note that both lines extend only until the levels $\overline{\omega}_1$ and $\overline{\omega}_2$, respectively, which are the levels of ω_t at which R investors exit in the equilibrium with zero, respectively high, shorting.

There are several noteworthy features of Figure 8. First, the price-dividend ratio for

50. Given the large value of θ, any reasonable choice of ρ in (35) is inconsequential, since the effective discount rate is $\rho + \theta$. For simplicity, we set $\rho = 0$.

the zero shorting equilibrium is *higher* than the price-dividend ratio for the high shorting equilibrium. This may seem counterintuitive, since the high shorting equilibrium implies a lower Sharpe ratio *for a fixed* ω_t. To understand this feature, it is important to recall that in this small-stock/large-stock setup both the expected dividend growth rate of stock 1 and the interest rate are constant in the limit where $\frac{\phi_1}{\phi_2} \approx 0$. Accordingly, by a standard Campbell-Shiller decomposition argument, the log-price dividend ratio of stock 1 can be viewed as a geometrically-weighted average of the future risk premiums (from t to ∞) for stock 1, which are impacted by the expected future Sharpe ratios of stock 1. The reason why the price-dividend ratio is higher in the zero-shorting equilibrium is that the dynamics of the wealth shares of I and R investors differ depending on whether the economy coordinates on the high- or zero-shorting equilibrium, and as a result so do the dynamics of future Sharpe ratios. We already showed in Section 3 that, when the economy coordinates on the high-shorting equilibrium, the wealth dynamics favor R investors. As a result, their future wealth shares are higher, which in turn tends to raise the path of future Sharpe ratios.

Participation costs accelerate these wealth dynamics: To see this, suppose that the market coordinates on the high shorting equilibrium and that $\omega_t > \overline{\omega}_1$. If a coordination shock shifts the economy to the zero-shorting equilibrium, short sellers exit instantaneously until $\omega_{t^+} = \overline{\omega}_1$, where ω_{t^+} is the value of ω_t after the equilibrium shift. At the new value $\omega_{t^+} = \overline{\omega}_1$, the shorting market is still inactive ($\overline{\omega}_1 > \omega_1^*$) and the remaining short sellers pay the (flow) participation cost despite holding a zero position in stock 1. However, the exit of a sufficient number of short sellers has increased the probability that future values of ω_t will fall below ω_1^*, and thus the shorting market will become active again. (Recall that, for values of ω_t below ω_1^*, the equilibrium is unique and involves shorting.) For this reason, the instantaneous Sharpe ratio may well move up, while the expected, geometrically weighted average of the Sharpe ratio from t to ∞ may well move down, thus raising the price-dividend ratio. In Appendix F we provide a graphical illustration of why the Sharpe ratio can move up in the short run and down in the long run. (See Figure F.1 and the discussion surrounding it.)

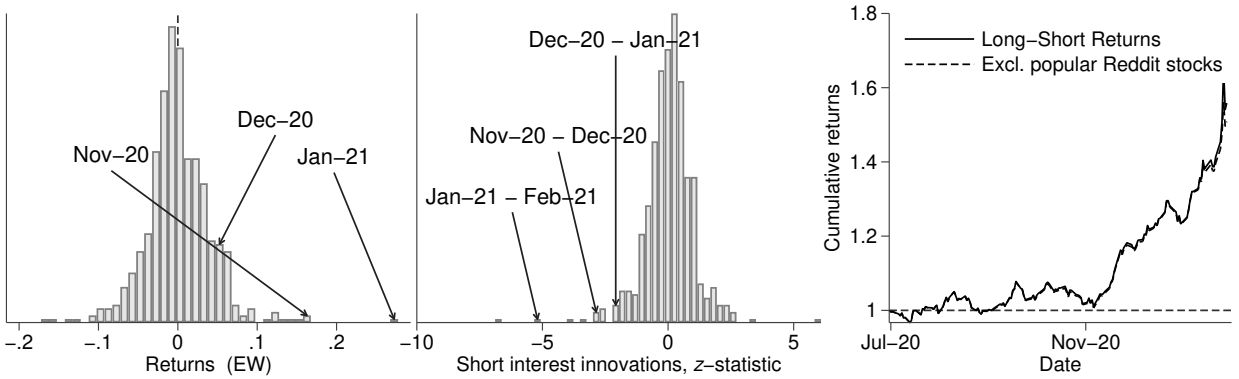


Figure 9: Left: Histogram of monthly returns (1973–2021). Equal-weighted, monthly returns on a portfolio long stocks in the top decile of short interest and short the market index. Center: Histogram of monthly AR(1) process innovations in short interest (1973–2021). Right: Cumulative returns to an equal-weighted long-short portfolio are shown by the solid black line. Arrows indicate observations for the months November 2020 to January 2021. The dashed line in the right panel excludes the six most-discussed tickers on the WallstreetBets subreddit (AMC, BBBY, GME, SPCE, TLRY, and TSLA).

7 The Retreat of Short Sellers between November 2020 and January 2021: A Case Study

In this final section we use the insights of our model to provide a potential explanation for the remarkable, and arguably puzzling, set of events that unfurled in the stock market between November 2020 and January 2021 — in particular, the unprecedented drop in short interest across a large number of stocks and the simultaneous appreciation of these stocks' price.

We document that (a) the short-seller retreat that started in November 2020 was followed by historically large losses for short-selling strategies, (b) this retreat preceded (by approximately two months) the dramatic and heavily publicized events surrounding meme stocks, and (c) the retreat was quite broad (across hundreds of stocks), and occurred among stocks that neither experienced a significant change in retail trading volume, nor were the topic of intense online discussion (as was the case with other meme stocks, predominately GameStop). We conclude, therefore, that the very poor performance of shorting strategies was the result of an abrupt shift in the behavior of short sellers, rather than that of a coordinated short squeeze by retail investors.

To start, in Figure 9 we plot the monthly returns to an equal-weighted portfolio that

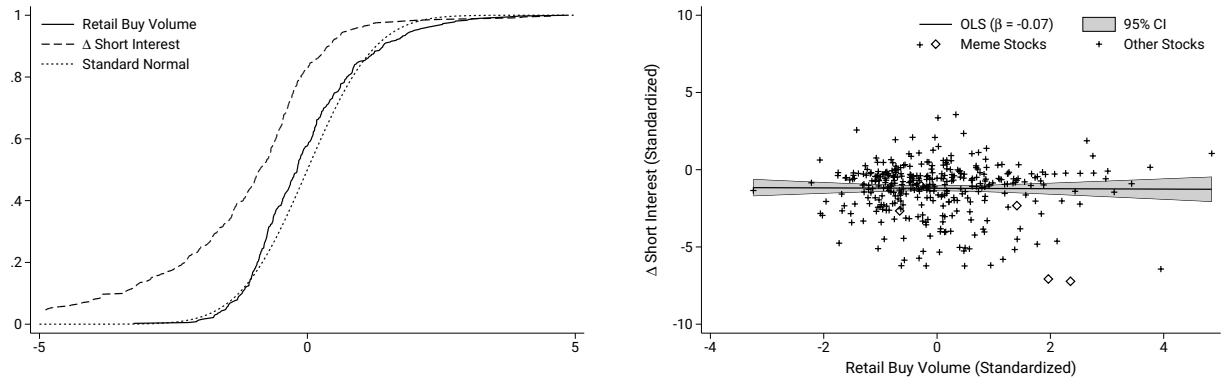
bets against the shorts. The portfolio is long the top decile of Russell 3000 stocks, ranked by short interest, and short the broad market. The left panel of Figure 9 shows that the November 2020 and January 2021 returns are the highest and second-highest (respectively) since the beginning of the sample (1973), while December 2020 is also in the top decile of the historically observed returns.[51] (The right panel of Figure 9 depicts the cumulative returns of the betting-against-the-shorts strategy to illustrate that November 2020 marks the start of the ascent.) The center panel of Figure 9 shows that the reduction in short interest was equally dramatic by historical standards and began prior to the meme stock events of January 2021. To further underscore that the short-seller retreat preceded these events, in Figure J.3 (Appendix I) we plot the daily submissions to the WSB subreddit (which was the online forum where users posted their opinions on Gamestop and other meme stocks) on a logarithmic scale. The graph shows that the explosive growth of online submissions occurred in early January 2021; November 2020 does not stand out.

Furthermore, while for GameStop there was a clear spike in retail purchase volume,[52] a remarkable feature of the data is that short sellers retreated across hundreds of stocks even though these stocks did not experience any unusual patterns in retail trading volume. Figure 10 plots the univariate distributions and scatter plots of (i) changes in short interest and (ii) retail purchase volume as a fraction of total volume for the most shorted stocks (top decile of stocks) ranked by short interest as of January 15, 2021. (All quantities are reported as standardized z-scores.) The distribution of the retail-purchase volume to total volume is centered around zero, with most values in a [–2, 2] range. By contrast, the distribution of changes in short interest is overwhelmingly negative, with most values in the [–5, 0] range. In addition, the relation between short interest and retail purchase volume is flat, as the scatter plot in Figure 10 illustrates.

To summarize, the short-seller retreat appears to have started in November 2020, a time when online discussion had not "picked up" yet. Moreover, it pertained to hundreds of stocks

51. Figure J.2 in Appendix I further shows that November 2020 and January 2021 remain outliers if we exclude tickers that were heavily discussed on the WallStreetBets subreddit (MC, BBBY, GME, SPCE, TLRY, and TSLA.), if we only include S&P500 constituents (i.e., larger stocks) in the formation of the long leg of the portfolio, and if we value-weight rather than equal-weight returns. See also Table J.1 in Appendix I for formal statistical tests.

52. Appendix Figure J.4 shows that the online discussion was highly correlated with retail trading volume in the case of GameStop.



(a) Marginal Distributions

(b) Joint Distribution

Figure 10: Retail purchase volume (as a fraction of total volume) and change in short interest, January 2021. Both quantities are reported as standardized z scores using TAQ and SEC data (respectively) from January 2015 through January 2021 to compute means and standard deviations. Panel (a) plots the empirical cumulative distribution of the two quantities, alongside a standard normal for reference. Panel (b) plots their joint distribution, along with the line of best fit. Tickers that were popular discussion topics on WSB and that are also in the top decile of short interest are indicated with "◊", while all other tickers are indicated with "+".

that saw no unusual change in retail trading volume. This suggests that for many stocks (other than the few meme stocks), the retreat in short selling happened even though the wealth composition across investor types (ω_t) did not change. Our model offers the possibility that some early signs of retail purchase instability triggered a coordination shock that led to a shift in equilibrium. An attractive aspect of this explanation for this particular episode is that there was no other obvious reason (such as a large loss by short sellers in other markets) that preceded the short selling retreat. We should also note that the seemingly simpler explanation that the broad declines in short interest were due to a correlated increase in irrationality (η) across stocks, but in an economy featuring a unique equilibrium, would run into the problem that $\frac{dy_t}{d\eta}$ is positive rather than negative.[53] Increased irrationality would therefore lead to a higher rather than lower level of short interest.[54]

53. See the proof of Lemma 2.

54. We also note that the decline in short interest was not just a short-run phenomenon. Figure J.5 in Appendix I shows that short interest didn't revert in the six months that followed January 2021.

8 Conclusion

Shorting can exhibit run-type patterns. The demand for shorting can be backward-bending and consequently for the same fundamentals there can be multiple equilibria featuring different degrees of shorting activity. We identify a general condition on the relation between fees and utilization that is necessary for the existence of multiple equilibria. We document that utilization and lending fees exhibit abrupt shifts. We also provide evidence that the stocks that satisfy the condition for the existence of multiple equilibria are more likely to experience jump-like behavior in utilization. In extensions of the baseline model we show that short sellers may choose to leave the market for the stock despite a rising stock price.

References

Abreu, Dilip, and Markus K Brunnermeier. 2002. "Synchronization risk and delayed arbitrage." *Journal of Financial Economics* 66 (2): 341–360.

Aït-Sahalia, Yacine, and Jean Jacod. 2009. "Testing for jumps in a discretely observed process." *The Annals of Statistics* 37 (1).

Allen, Franklin, Marlene Haas, Eric Nowak, Matteo Pirovano, and Angel Tengulov. 2021. "Squeezing Shorts Through Social Media Platforms." *History of Finance eJournal.*

Asquith, Paul, Parag A. Pathak, and Jay R. Ritter. 2005. "Short interest, institutional ownership, and stock returns." *Journal of Financial Economics* 78 (2): 243–276.

Atmaz, Adem, Suleyman Basak, and Fangcheng Ruan. 2023. "Dynamic Equilibrium with Costly Short-Selling and Lending Market." *The Review of Financial Studies* 37 (2): 444–506.

Baldursson, Fridrik M, and Ioannis Karatzas. 1996. "Irreversible investment and industry equilibrium." *Finance and Stochastics* 1 (1): 69–89.

Banerjee, Snehal, and Jeremy J. Graveline. 2013. "The Cost of Short-Selling Liquid Securities." *Journal of Finance* 68 (2): 637–664.

Bansal, Ravi, Dana Kiku, and Amir Yaron. 2012. "An Empirical Evaluation of the Long-Run Risks Model for Asset Prices." *Critical Finance Review* 1 (1): 183–221.

Beeler, Jason, and John Campbell. 2012. "The Long-Run Risks Model and Aggregate Asset Prices: An Empirical Assessment." *Critical Finance Review* 1 (1): 141–182.

Beneish, Messod Daniel, Charles MC Lee, and D Craig Nichols. 2015. "In short supply: Short-sellers and stock returns." *Journal of Accounting and Economics* 60 (2-3): 33–57.

Benhabib, Jess, and Roger Farmer. 1999. "Indeterminacy and sunspots in macroeconomics." Chap. 06 in *Handbook of Macroeconomics,* edited by J. B. Taylor and M. Woodford, vol. 1, Part A, 387–448.

Biais, Bruno, Johan Hombert, and Pierre-Olivier Weill. 2021. "Incentive Constrained Risk Sharing, Segmentation, and Asset Pricing." *American Economic Review, forthcoming.*

Blanchard, Olivier J. 1985. "Debt, Deficits, and Finite Horizons." *Journal of Political Economy* 93 (2): 223–247.

Blocher, Jesse, Adam V Reed, and Edward D Van Wesep. 2013. "Connecting two markets: An equilibrium framework for shorts, longs, and stock loans." *Journal of Financial Economics* 108 (2): 302–322.

Boehmer, Ekkehart, Charles M Jones, and Xiaoyan Zhang. 2008. "Which shorts are informed?" *Journal of Finance* 63 (2): 491–527.

Cohen, Lauren, Karl B Diether, and Christopher J Malloy. 2007. "Supply and demand shifts in the shorting market." *Journal of Finance* 62 (5): 2061–2096.

D'Avolio, Gene. 2002. "The market for borrowing stock." *Journal of Financial Economics* 66 (2): 271–306.

Daniel, Kent, Alexander Klos, and Simon Rottke. 2022. "The Dynamics of Disagreement." *The Review of Financial Studies* 36 (6): 2431–2467.

Dechow, Patricia M, Amy P Hutton, Lisa Meulbroek, and Richard G Sloan. 2001. "Short-sellers, fundamental analysis, and stock returns." *Journal of Financial Economics* 61 (1): 77–106.

Desai, Hemang, Kevin Ramesh, S Ramu Thiagarajan, and Bala V Balachandran. 2002. "An investigation of the informational role of short interest in the Nasdaq market." *Journal of Finance* 57 (5): 2263–2287.

Detemple, Jerome, and Shashidhar Murthy. 1997. "Equilibrium asset prices and no-arbitrage with portfolio constraints." *The Review of Financial Studies* 10 (4): 1133–1174.

Diamond, Douglas W, and Robert E Verrecchia. 1987. "Constraints on short-selling and asset price adjustment to private information." *Journal of Financial Economics* 18 (2): 277–311.

Diether, Karl B., Kuan-Hui Lee, and Ingrid M Werner. 2009. "Short-sale strategies and return predictability." *The Review of Financial Studies* 22 (2): 575–607.

Drechsler, Itamar, and Qingyi Freda Drechsler. 2014. *The Shorting Premium and Asset Pricing Anomalies.*

Duffie, Darrell, and Larry G. Epstein. 1992. "Stochastic Differential Utility." *Econometrica* 60 (2): 353–394.

Duffie, Darrell, Nicolae Gârleanu, and Lasse Heje Pedersen. 2002. "Securities lending, shorting, and pricing." *Journal of Financial Economics* 66 (2-3): 307–339.

Duong, Truong X, Zsuzsa R Huszár, Ruth S K Tan, and Weina Zhang. 2017. "The Information Value of Stock Lending Fees: Are Lenders Price Takers?" *Review of Finance* 21 (6): 2353–2377.

Engelberg, Joseph E., Adam V. Reed, and Matthew C. Ringgenberg. 2018. "Short-Selling Risk." *The Journal of Finance* 73 (2): 755–786.

Evgeniou, Theodoros, Julien Hugonnier, and Rodolfo Prieto. 2022. "Asset Pricing with Costly Short Sales." Working paper.

Farmer, Roger, and Jean-Philippe Bouchaud. 2020. *Self-Fulfilling Prophecies, Quasi Non-Ergodicity & Wealth Inequality.* Working Paper, Working Paper Series 28261. National Bureau of Economic Research.

Fostel, Ana, and John Geanakoplos. 2008. "Leverage cycles and the anxious economy." *American Economic Review* 98 (4): 1211–44.

Gârleanu, Nicolae, Leonid Kogan, and Stavros Panageas. 2012. "Displacement risk and asset returns." *Journal of Financial Economics* 105 (3): 491–510.

Gârleanu, Nicolae, and Stavros Panageas. 2015. "Young, old, conservative, and bold: The implications of heterogeneity and finite lives for asset pricing." *Journal of Political Economy* 123 (3): 670–685.

———. 2021. "What to expect when everyone is expecting: Self-fulfilling expectations and asset-pricing puzzles." *Journal of Financial Economics* 140 (1): 54–73.

———. 2023. "Heterogeneity and Asset Prices: An Intergenerational Approach." *Journal of Political Economy* 131 (4): 839–876.

Geczy, Christopher C., David K. Musto, and Adam V. Reed. 2002. "Stocks are special too: an analysis of the equity lending market." *Journal of Financial Economics* 66 (2–3): 241–269.

Gennotte, Gerard, and Hayne Leland. 1990. "Market Liquidity, Hedging, and Crashes." *American Economic Review* 80 (5): 999–1021.

Harrison, J Michael, and David M Kreps. 1978. "Speculative investor behavior in a stock market with heterogeneous expectations." *Quarterly Journal of Economics* 92 (2): 323–336.

Hong, Harrison, and Jeremy C Stein. 2003. "Differences of opinion, short-sales constraints, and market crashes." *The Review of Financial Studies* 16 (2): 487–525.

Jones, Charles M, and Owen A Lamont. 2002. "Short-sale constraints and stock returns." *Journal of Financial Economics* 66 (2-3): 207–239.

Kaplan, Steven N, Tobias J Moskowitz, and Berk A Sensoy. 2013. "The effects of stock lending on security prices: An experiment." *Journal of Finance* 68 (5): 1891–1936.

Khorrami, Paymon, and Fernando Mendo. 2021. "Rational Sentiments and Financial Frictions." Working paper.

Khorrami, Paymon, and Alexander Zentefis. 2020. "Arbitrage and Beliefs." Working paper.

Lamont, Owen A. 2012. "Go down fighting: Short sellers vs. firms." *The Review of Asset Pricing Studies* 2 (1): 1–30.

Lamont, Owen A, and Jeremy C Stein. 2004. "Aggregate short interest and market valuations." *American Economic Review* 94 (2): 29–32.

Leahy, John V. 1993. "Investment in competitive equilibrium: The optimality of myopic behavior." *Quarterly Journal of Economics* 108 (4): 1105–1133.

Merton, Robert. 1987. "A Simple Model of Capital Market Equilibrium with Incomplete Information." *Journal of Finance* 42 (3): 483–510.

Miller, Edward M. 1977. "Risk, uncertainty, and divergence of opinion." *Journal of Finance* 32 (4): 1151–1168.

Panageas, Stavros. 2020. "The Implications of Heterogeneity and Inequality for Asset Pricing." *Foundations and Trends® in Finance* 12 (3): 199–275.

Pedersen, Lasse Heje. 2022. "Game on: Social networks and markets." *Journal of Financial Economics* 146 (3): 1097–1119.

Porras Prado, Melissa, Pedro A. C. Saffi, and Jason Sturgess. 2016. "Ownership Structure, Limits to Arbitrage, and Stock Returns: Evidence from Equity Lending Markets." *The Review of Financial Studies* 29 (12): 3211–3244.

Rapach, David E., Matthew C. Ringgenberg, and Guofu Zhou. 2016. "Short interest and aggregate stock returns." *Journal of Financial Economics* 121 (1): 46–65.

Scheinkman, Jose A, and Wei Xiong. 2003. "Overconfidence and speculative bubbles." *Journal of Political Economy* 111 (6): 1183–1220.

Schroder, Mark, and Costis Skiadas. 1999. "Optimal Consumption and Portfolio Selection with Stochastic Differential Utility." *Journal of Economic Theory* 89 (1): 68–126.

Senchack, Andrew J, and Laura T Starks. 1993. "Short-sale restrictions and market reaction to short-interest announcements." *Journal of Financial and Quantitative Analysis* 28 (2): 177–194.

Seneca, Joseph J. 1967. "Short interest: bearish or bullish?" *Journal of Finance* 22 (1): 67–70.

Simsek, Alp. 2013. "Belief disagreements and collateral constraints." *Econometrica* 81 (1): 1–53.

Sweeting, Andrew. 2006. "Coordination, Differentiation, and the Timing of Radio Commercials." *Journal of Economics & Management Strategy* 15 (4): 909–942.

Vayanos, Dimitri, and Pierre-Olivier Weill. 2008. "A Search-Based Theory of the On-the-Run Phenomenon." *Journal of Finance* 63 (3): 1361–1398.

Wang, Bin, and Xu Zheng. 2022. "Testing for the presence of jump components in jump diffusion models." *Journal of Econometrics* 230 (2): 483–509.

Zentefis, Alexander K. 2022. "Self-Fulfilling Asset Prices." *The Review of Asset Pricing Studies* 12 (4): 886–917.

For Online Publication – Appendix

A The Determination of the Lending Fee

In the text we assume a "flat" supply curve for lending shares. That is, we assume $f_t = f(y_t) = \varphi$. We provide here the simplest model that supports this assumption. We also discuss how to extend the model to allow for an increasing $f(\cdot)$.

All interactions considered in this section happen anew every period, where the length of the period is idealized to be "dt," that is, infinitesimal. (We could formalize this assumption by considering a discrete-time model where the length Δ of a period is taken to go to zero, and focusing on the limit of resultant equilibria.)

We start by considering the long investors, who wish to lend their shares. Each investor lends all her shares to any one of a competitive fringe of profit-maximizing "security lenders" in exchange for an income stream that is proportional to the dollar value of shares the investor lends. This income stream is determined as follows. In equilibrium, each security lender lends a proportion y_t of the shares it borrows from investors and receives a fee f_l per dollar of shares it lends out. (We omit time subscripts from now on.) Competition between the security lenders drives the income stream paid to long investors to yf_l per dollar of stock owned by the investors.[55]

At the other end of the lending transaction, desirous short sellers interact with a competitive set of "borrower's brokers." Specifically, for every borrowing fee f_b the would-be short sellers provide the dollar amount that they would like to short, and the brokers take the value f_b as a given when they attempt to fill the investors' borrowing orders.

All of the frictions in this model pertain to the interaction between security lenders and borrower's brokers. Specifically, to initiate a stock loan the representative broker must pay a cost ξ per dollar value of share "located" with a security lender, per unit of time. This cost is construed as labor cost that compensates brokers for their disutility of labor.

The interaction between the broker and the security lender takes the form of bilateral Nash bargaining in which the broker has bargaining power $1/(1+z)$ for a parameter $z \in (0, \infty)$. Given our assumption that all interactions (between investors and brokers or security lenders and between brokers and security lenders) happen anew every period, the outside option for both brokers and security lenders is the failure to transact during the period. This means that the gains from trade to the security lender equal the lending fee f_l, while to the broker the borrowing fee net of the lending one $f_b - f_l$ — the searching and matching cost ξ has been sunk at this point. The total gains from trade equal f_b, the foregone revenue from the would-be short seller. Given the bargaining protocol, it follows that

$$f_l = \frac{z}{1+z}(f_l + f_b - f_l) = \frac{z}{1+z} f_b. \tag{A.1}$$

55. In that sense, the security lenders resemble the "insurance companies" in Blanchard (1985). Similar to how insurance companies collect payments from the fraction of agents who die and rebate them to the surviving population, the security lenders collect lending fees from the proportion of a long portfolio that gets loaned out and rebate it in the form of an income stream to the representative long investor.

Brokers break even on net, meaning that

$$f_b = f_l + \xi, \tag{A.2}$$

so that

$$f_l = z\xi, \tag{A.3}$$
$$f_b = (1+z)\xi. \tag{A.4}$$

To keep the model transparent and tractable, assume that all brokers are members of the representative household, and therefore the fees that compensate them for their effort are rebated to each households as an income stream proportional to the household's wealth and independent of the composition of the household's portfolio.

Setting $\varphi = (1+z)\xi$ and $\tau = z/(1+z)$, this extended model is equivalent to the model we assumed in the text. To generalize to upward-sloping supply curves, one would simply assume an increasing cost $\xi(y)$.

B Multiple Agent Types

We illustrate here that the multiplicity of equilibria may expand with the number of agent types. In particular, adding a third group of agents can result in a third equilibrium featuring non-zero shorting; such a model may admit, in fact, up to five equilibria.

Specifically, let us assume a third group of investors characterized by beliefs that are summarized by the quantity η^P. We think of these investors as pessimists, which implies $\eta^P < 0$. The intuition we wish to capture is that, in addition to the "high-shorting" and "medium-shorting" equilibria in the base-line model, low-shorting equilibria may exist in which investor R is inactive, while investor P shorts actively.

To make the point theoretically, one may argue by continuity. Specifically, consider the zero-shorting equilibrium in the baseline model, and perturb the setting by adding a small mass of sufficiently pessimistic investors ($|\eta^P|$ large enough). These investors will want to short, but will not be sufficiently numerous to move the Sharpe ratio or lending income to a point where investors R and I are no longer in equilibrium.

It is helpful to write down the equilibrium conditions in the augmented model — both to allow for a formal argument and in the interest of a numerical illustration. We repeat the analysis in the text — letting ω^P denote the wealth share of agents P — to obtain the market clearing condition

$$\begin{aligned} 1 = \frac{1}{\sigma_D}\Bigg[&\omega^P\left(\kappa+\eta^P+\frac{\varphi}{\sigma_D}\right)1_{\left\{\kappa+\eta^P+\frac{\varphi}{\sigma_D}<0\right\}} + \omega^R\left(\kappa+\frac{\varphi}{\sigma_D}\right)1_{\left\{\kappa+\frac{\varphi}{\sigma_D}<0\right\}} + \\ &\omega^I\left(\kappa+\eta^I+\frac{\varphi}{\sigma_D}\tau y\right)\Bigg], \end{aligned} \tag{B.1}$$

where the left-hand side is the proportion of aggregate wealth represented by the supply of the stock, while the right-hand side equals the proportion of aggregate wealth invested in the stock. We restricted attention to cases in which R agents do not take a long position in the

stock.

We solve for the Sharpe ratio κ:

$$\kappa = \sigma_D - \left(\omega^P \eta^P + \omega^I \eta^I\right) - \frac{\varphi}{\sigma_D}\left(\omega^P + \omega^R + \omega^I \tau y\right) \tag{B.2}$$

if $\kappa + \varphi/\sigma_D < 0$, respectively

$$\kappa = \frac{\sigma_D}{\omega^P + \omega^I} - \frac{\omega^P \eta^P + \omega^I \eta^I}{\omega^P + \omega^I} - \frac{\varphi}{\sigma_D}\frac{\omega^P + \omega^I \tau y}{\omega^P + \omega^I} \tag{B.3}$$

if $\kappa + \varphi/\sigma_D \geq 0 > \kappa + \eta^P + \varphi/\sigma_D$.

The other equilibrium condition concerns the determination of the value of y:

$$y = -\frac{\omega^P \left(\kappa + \eta^P + \frac{\varphi}{\sigma_D}\right) 1_{\left\{\kappa + \eta^P + \frac{\varphi}{\sigma_D} < 0\right\}} + \omega^R \left(\kappa + \frac{\varphi}{\sigma_D}\right) 1_{\left\{\kappa + \frac{\varphi}{\sigma_D} < 0\right\}}}{\omega^I \left(\kappa + \eta^I + \frac{\varphi}{\sigma_D}\tau y\right)}. \tag{B.4}$$

Depending on whether κ is determined according to (B.2) or (B.3) we obtain a different quadratic equation. For appropriate parameter choices all but one combinations are possible in terms of how many solutions in the interval $(0, 1)$ each of them admits. We are particularly interested in situations in which (B.2) applies and results in two admissible solutions, in addition to which at least one solution obtains when (B.3) applies.

We illustrate such outcomes in Figure B.1. The two panels differ in terms of parameters, but depict the same objects. Specifically, the x-axis records candidate values of y that agents anticipate. Agents form demands taking such a value y and a Sharpe ratio κ as given, and clearing in the asset market determines the Sharpe ratio. With the Sharpe ratio now specified for each candidate y, we can compute the actual resulting y — the value of the right-hand side of equation (B.4). This quantity is recorded on the y-axis. An equilibrium requires that the x and y coordinates are equal.

The line "R and P short" plots y as if both R and P shorted, that is, their portfolio weights are calculated by adding the return φ to their perceived intrinsic expected return from the asset; in that case, the Sharpe ratio is given by (B.2). The line "Only P shorts" is produced similarly, except that the demand of agent R is set to zero; equation (B.3) applies. The actual resulting y is depicted by the thick continuous line, labeled "Actual response." Finally, the line "Diagonal" depicts the equilibrium condition. Equilibria are therefore represented by points of intersection between the two continuous lines. The left panel presents a situation in which four equilibria with positive amounts of shorting and one with zero shorting obtain. The right panel presents a situation with three equilibria, all of which feature positive y.

We also flesh out the theoretical argument for the existence of a third equilibrium when ω^P is close to zero and two equilibria with $y > 0$ exist with $\omega^P = 0$ — i.e., the baseline model. By assumption, with $\omega^P = 0$ and $y = 0$ equation (B.3) applies and $\kappa + \frac{\varphi}{\sigma_D} > 0$. Choosing η^P

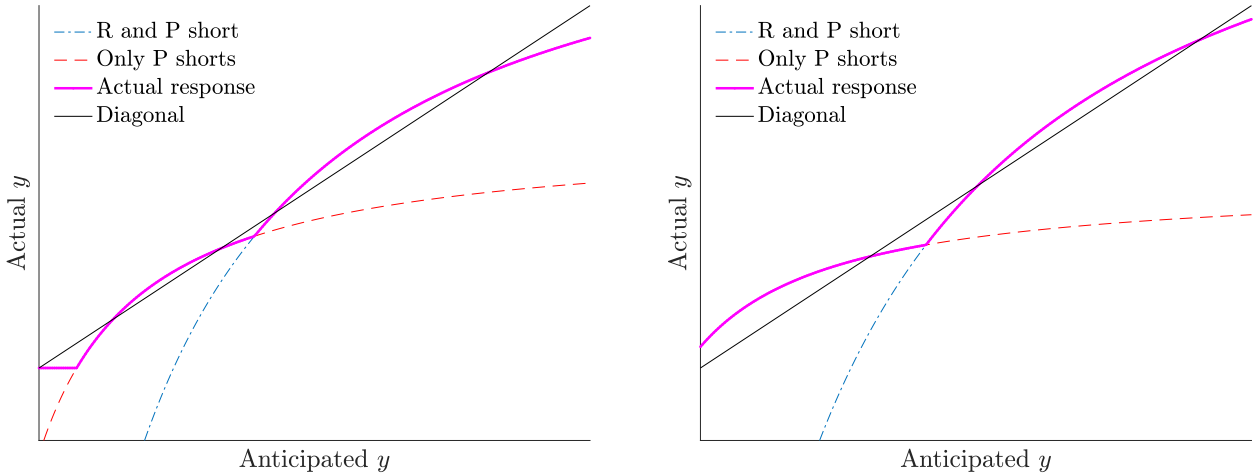


Figure B.1: The figure plots, in each panel, four lines pertaining to the model extension developed in this section. Equilibria are characterized by the satisfaction of equation (B.4), whose right-hand side is represented here by the line "Actual response" and the left-hand side by the line "Diagonal." Further details are provided in the text.

so that $\kappa + \eta^P + \frac{\varphi}{\sigma_D} < 0$, we wish to conclude that equations

$$y = -\frac{\omega^P \left(\kappa + \eta^P + \frac{\varphi}{\sigma_D}\right)}{\omega^I \left(\kappa + \eta^I + \frac{\varphi}{\sigma_D}\tau y\right)} \tag{B.5}$$

and (B.3) admit a solution that satisfies $\kappa + \frac{\varphi}{\sigma_D} > 0$ even for $\omega^P > 0$, at least when it is small enough. For simplicity, we keep ω^I constant as we increase ω^P from zero. Plugging (B.3) in (B.5) we obtain a quadratic that can be written as

$$y = \frac{\omega^P}{\omega^I} \frac{\left(\eta^I - \eta^P\right)\omega^I - \frac{\varphi}{\sigma_D}\omega^I \tau\left(1 - y\right) - \sigma_D}{\left(\eta^I - \eta^P\right)\omega^P - \frac{\varphi}{\sigma_D}\omega^P \tau\left(1 - y\right) + \sigma_D} \equiv H(\omega^P, y). \tag{B.6}$$

Our choice of η^P is such that the numerator of the second fraction on the right-hand side is positive at $y = 0$, which implies $\frac{\partial H}{\partial \omega^P} > 0$ evaluated at $\omega^P = 0$, as well as $\frac{\partial H}{\partial y} = 0$ at $\omega^P = 0$. We therefore have

$$\frac{dy}{d\omega^P} = \left(1 - \frac{\partial H}{\partial y}\right)^{-1} \frac{\partial H}{\partial \omega^P} > 0, \tag{B.7}$$

confirming that an equilibrium with positive y exists for small ω^P. (The condition $\kappa + \frac{\varphi}{\sigma_D} > 0$ is satisfied by continuity.)

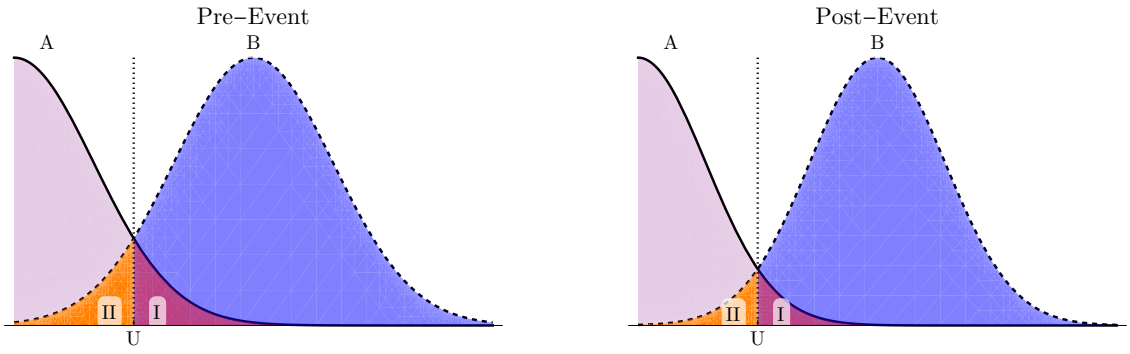


Figure C.1: Illustration of the effect of a decrease in volatility on the probability of Type-I and Type-II errors.

C A Graphical Illustration of the Findings of Table 4.

In this section we provide a more detailed explanation for why a drop in volatility lowers the measured jump intensity of stocks that are unlikely to satisfy the condition of Proposition 4, but not for those that are likely to satisfy it.

For stocks that don't satisfy the condition of Proposition 4, the innovations to utilization are diffusive and the absolute value of the innovations to utilization (over discrete time-intervals) follows a distribution that resembles the half-normal distribution (since we are considering absolute values of utilization-changes.) This half-normal is depicted by distribution A in Figure C.1. Of course, since a jump is defined as a realization above U, there is the possibility of declaring a "false positive" jump with a probability that is depicted as Region I in the figure. Stocks that satisfy the condition of Proposition 4 could exhibit either positive or negative jumps in utilization. The relevant distribution of utilization-changes looks like the absolute value of a mixture distribution: If no jump occurs, the relevant distribution is A. If a jump occurs, then the relevant distribution is B. Distribution B is the same as A, but shifted to the right because of the jump. (Note that because we take absolute values, the shift is always to the right.) Since a "jump" is an observation above the cutoff U, there are possibilities of both false positive jumps (Region I), but also false negative jumps (Region II). These false negative jumps are instances where the jump has occurred, but the absolute value of the discrete-time change in utilization is still below the cutoff U.

The right plot of figure C.1 depicts a decline in the (diffusive) volatility of utilization. When the volatility falls, the area associated with Regions I and II shrinks.

The shrinkage of area I implies that the possibility of a false positive jump declines, thus lowering the (measured) jump rate of stocks that don't satisfy the condition of Proposition 4.

For stocks that satisfy the condition of Proposition 4 there are two opposing effects. On the one hand side, the shrinkage of area I lowers the instances of false positive jumps, which lowers the measured jump rate; on the other hand, the shrinkage of area II lowers the instances of false negatives, which raises the measured jump rate. In short, a decline in volatility should unambiguously lower the measured jump rate for stocks that cannot exhibit jumps (fifth quintile stocks), but should have a lower (and in principle ambiguous effect) on the measured jump rate of stocks that are likely to satisfy the condition of Proposition 4 (first quintile stocks).

D Recursive Preferences with Unitary RRA

In this section we provide the details underlying the discussion in Section 6.1 in the text. For clarity, we develop the argument linearly, culminating in a formal result (Proposition 5).

We adopt the Epstein-Zin preference specification with the risk-aversion coefficient constrained to take a value of one:

$$V_t = \mathrm{E}_t \int_t^\infty e^{-(\rho+\pi)(s-t)} \left(\frac{c^\alpha}{\alpha}\, ds - \frac{1}{2V_s} d[V_s] \right). \tag{D.1}$$

This specification leads to V homogeneous of degree α in the consumption stream $\{c_t\}_t$, so that we can conjecture

$$V(W,x) = \frac{W^\alpha}{\alpha} g(x)^{\alpha-1}, \tag{D.2}$$

where the function $g(\cdot)$, to be computed, will be shown below to equal the agent's consumption-to-wealth ratio.

Finally, we maintain the conjecture that an appropriate — and sufficient — state variable is the wealth share of agent R, denoted by ω. Because of unit risk aversion, the optimal portfolio coincides with the one obtained under log utility. Treating the volatility of returns, σ_t, as known, we make use of the known (myopic) portfolio policy to write

$$w^R = \frac{\kappa_t + \frac{f(y_t)}{\sigma_t}}{\sigma_t} 1_{\kappa_t + \frac{f(y_t)}{\sigma_t} < 0} + \frac{\kappa_t + \tau y_t \frac{f(y_t)}{\sigma_t}}{\sigma_t} 1_{\kappa_t + \tau y_t \frac{f(y_t)}{\sigma_t} > 0} \tag{D.3}$$

$$w^I = \frac{\kappa_t + \eta + \frac{f(y_t)}{\sigma_t}}{\sigma_t} 1_{\kappa_t + \eta + \frac{f(y_t)}{\sigma_t} < 0} + \frac{\kappa_t + \eta + \tau y_t \frac{f(y_t)}{\sigma_t}}{\sigma_t} 1_{\kappa_t + \eta + \tau y_t \frac{f(y_t)}{\sigma_t} > 0}, \tag{D.4}$$

noting further that the first term on the right-hand side of (D.4) equals zero in equilibrium, since I cannot short in equilibrium. We have the equilibrium restrictions

$$y_t = \frac{\omega_t}{1-\omega_t} \frac{w^{R^-}}{w^I} \tag{D.5}$$

$$1 = w^R \omega_t + w^I (1-\omega_t). \tag{D.6}$$

A key determinant of the portfolio choices is the market return volatility σ_t, i.e., the diffusion parameter in the equation

$$\begin{aligned} dR_t &= \frac{dS_t}{S_t} - \delta dt + \frac{D_t}{S_t} dt \\ &= \mu_t dt + \sigma_t dB_t. \end{aligned} \tag{D.7}$$

We define the price-to-dividend ratio, $p_t = S_t/D_t$, an conjecture it to be a function of ω. The restriction that aggregate wealth equals the stock price and aggregate consumption the

stock dividend gives

$$p(\omega) := \frac{S_t}{D_t} = \left(\sum g_i(\omega)\omega^i\right)^{-1}, \tag{D.8}$$

so that we obtain

$$\sigma_t = \sigma_D + \frac{p'(\omega)}{p(\omega)}\sigma_\omega \tag{D.9}$$

$$\mu_t = \mu_D - \delta + \frac{1}{p(\omega)} + \frac{p'(\omega)}{p(\omega)}\mu_\omega + \frac{1}{2}\frac{p''(\omega)}{p(\omega)}\sigma_\omega^2 + \frac{p'(\omega)}{p(\omega)}\sigma_\omega\sigma_D. \tag{D.10}$$

We derive the dynamics of the wealth-share ω_t by remarking that $W_{R,t} = \omega_t S_t$ and thus

$$\frac{d(\omega S)}{\omega S} = (r + \pi + n)dt + w^R\left(dR_t + \left(f(y)1_{w^R<0} - r\right)dt\right) - g_R(\omega)\,dt - \pi dt + \frac{\nu\delta}{\omega}dt. \tag{D.11}$$

Writing

$$d\omega_t = \mu_\omega dt + \sigma_\omega dB_t, \tag{D.12}$$

we have the dynamics

$$\sigma_\omega + \omega\sigma_t = w^R\omega\sigma_t \tag{D.13}$$

$$\mu_\omega + \omega(\mu_t + \delta - p(\omega)^{-1}) + \sigma_\omega\sigma_t = \omega\left(r_t + n + w^R\left(\mu_t + f(y)1_{w^R<0} - r_t\right) - g_R\right) + \nu\delta. \tag{D.14}$$

Note that the drifts are subject to the agents' probability distortions. Specifically, agent I perceives $\mu_t^I = \mu_t + \eta\sigma_t$, which we already recognize in the portfolio choice problem. This must also be recognized in (D.14), i.e., agent I perceives

$$\mu_\omega^I = w^R\mu_t^I = \mu_\omega + \omega(w^R - 1)\eta\sigma_t = \mu_\omega + \eta\sigma_\omega. \tag{D.15}$$

Equations (D.9) and (D.13) lead to

$$\sigma_t - \omega\frac{p'(\omega)}{p(\omega)}(w^R - 1)\sigma_t = \sigma_D. \tag{D.16}$$

Given that the interest rate equals

$$r_t = \mu_t - \kappa_t\sigma_t, \tag{D.17}$$

the system of equations (D.3)–(D.6) comprises four equations and has four unknowns — w^R, w^I, y, and κ — treating $g_i(\cdot)$ as known.

To solve this system together with (D.16), note that (D.16) gives σ_t^{-1} as linear in w^R:

$$\frac{1}{\sigma_t} = \frac{1}{\sigma_D}\left(1 + \frac{xp'(x)}{p(x)}(1 - w^R)\right). \tag{D.18}$$

We can go through three cases, depending on the sign of w^R:
Case I: $w^R > 0$. We have

$$\kappa = \sigma_t - (1-x)\eta \tag{D.19}$$

$$w^R = \frac{\kappa}{\sigma_t} = 1 - \frac{(1-x)\eta}{\sigma_t}. \tag{D.20}$$

We can combine this equation with (D.18) to get

$$1 - w^R = (1-x)\frac{\eta}{\sigma_D}\left(1 + x\log(p)'(1-w^R)\right) \tag{D.21}$$

$$1 - w^R = \left(1 - (1-x)x\frac{\eta}{\sigma_D}\log(p)'\right)^{-1}(1-x)\frac{\eta}{\sigma_D} \tag{D.22}$$

$$w^R = 1 - \left(1 - (1-x)x\frac{\eta}{\sigma_D}\log(p)'\right)^{-1}(1-x)\frac{\eta}{\sigma_D},, \tag{D.23}$$

which must be positive if it is to be equilibrium value for w^R.

If (D.23) gives w^R negative, then we may be in either Case II, in which the type R investor stays out of the market, or III, in which the type R investor shorts the stock.

Case II: $w^R = 0$. Here,

$$\kappa = \frac{\sigma_t}{1-x} - \eta \tag{D.24}$$

and we know from the violation of Case I that, with $w^R = 0$, the condition $\kappa < 0$ is satisfied. In order for this case to provide an equilibrium, it must also be the case that shorting, with $y = 0$, is unattractive:

$$\frac{1}{1-x} - \frac{\eta}{\sigma_D}\left(1 + \frac{p'}{p}\right) + \frac{f(y)}{\sigma_D^2}\left(1 + \frac{p'}{p}\right)^2 \geq 0. \tag{D.25}$$

The left-hand side of this inequality, too, uses the form σ_t takes when $w^R = 0$.

If this inequality is not satisfied, then we are (verified below) in

Case III: $w^R < 0$. We get, from the market-clearing condition,

$$\kappa = \sigma_t - \eta(1-x) - \frac{f(y)}{\sigma_t}\left(x + (1-x)\tau y\right), \tag{D.26}$$

which we plug back in (D.3) to obtain

$$w^R = 1 - \frac{\eta(1-x)}{\sigma_t} + \frac{f(y)}{\sigma_t^2}(1-x)(1-\tau y) \tag{D.27}$$
$$= 1 - \frac{\eta(1-x)}{\sigma_D}\left(1 + (1-w^R)\frac{xp'}{p}\right) + \frac{f(y)}{\sigma_D^2}(1-x)(1-\tau y)\left(1 + (1-w^R)\frac{xp'}{p}\right)^2.$$

Finally, y is also expressed in terms of w^R:

$$y = -\frac{w^R x}{w^I(1-x)} = \frac{-w^R x}{1 - w^R x}, \tag{D.28}$$

leading to

$$w^R = 1 - \frac{\eta(1-x)}{\sigma_D}\left(1 + (1-w^R)\frac{xp'}{p}\right) + \tag{D.29}$$
$$\frac{f(y)}{\sigma_D^2}(1-x)\left(1 + \tau\frac{w^R x}{1-w^R x}\right)\left(1 + (1-w^R)\frac{xp'}{p}\right)^2.$$

This is a cubic equation, whose solutions are available explicitly. For the sake of completeness, we write the cubic in full, as

$$0 = \sum_{i=0}^{3} A_i^w w^i, \tag{D.30}$$

with

$$A_3^w = -\frac{f(y)}{\sigma_D^2}(1-x)x^3(1-\tau)\left(\frac{p'}{p}\right)^2 \tag{D.31}$$

$$A_2^w = x - \frac{\eta}{\sigma_D}(1-x)x^2\frac{p'}{p} + \frac{f(y)}{\sigma_D^2}(1-x)\left(2x(1-\tau)\left(1+\frac{xp'}{p}\right)\frac{xp'}{p} + \left(\frac{xp'}{p}\right)^2\right) \tag{D.32}$$

$$A_1^w = -1 - x + \frac{\eta}{\sigma_D}(1-x)\left(\frac{xp'}{p} + x\left(1+\frac{xp'}{p}\right)\right) - \tag{D.33}$$
$$\frac{f(y)}{\sigma_D^2}(1-x)\left(x(1-\tau)\left(1+\frac{xp'}{p}\right)^2 + 2\left(1+\frac{xp'}{p}\right)\frac{xp'}{p}\right)$$

$$A_0^w = 1 - \frac{\eta}{\sigma_D}(1-x)\left(1+\frac{xp'}{p}\right) + \frac{f(y)}{\sigma_D^2}(1-x)\left(1+\frac{xp'}{p}\right)^2. \tag{D.34}$$

We also note explicitly the expression for κ that holds in all of the cases:

$$\kappa = \left(\frac{\sigma_t}{1-x} - \eta\right)\left(1_{w^R=0} + (1-x)1_{w^R>0}\right) + \left(\sigma_t - \eta(1-x) - \frac{f(y)}{\sigma_t}\left(x + (1-x)\tau y\right)\right)1_{w^R<0}. \tag{D.35}$$

We are ready to derive the ODE obeyed by the functions g_i. We write the Hamilton-Jacobi-Bellman (HJB) equation:

$$0 = \sup_{c,w} \frac{(cW)^\alpha}{\alpha} - \frac{1}{2\alpha}\frac{d[W^\alpha g^{\alpha-1}]_t}{W^\alpha g^{\alpha-1}} + \mathrm{E}\left[d\frac{e^{-(\rho+\pi)t}W^\alpha g^{\alpha-1}}{\alpha}\right]. \tag{D.36}$$

This equation holds for each of the two agents, under their respective beliefs. Using the dynamics of W^i and $g(\omega_t)$, we expand (D.36) as

$$\begin{aligned} 0 = \sup_{c,w}\ & \frac{(cW)^\alpha}{\alpha} - \frac{1}{2\alpha}\frac{d[W^\alpha g^{\alpha-1}]_t}{W^\alpha g^{\alpha-1}} + W^\alpha g^{\alpha-1}\left(r + n + \pi + w(\hat{\kappa}\sigma_t) - c - \frac{\rho+\pi}{\alpha}\right) + \\ & \frac{\alpha-1}{\alpha}\frac{g'}{g}\mu_x W^\alpha g^{\alpha-1} + \frac{\alpha-1}{2}\left(w^2\sigma_t^2 + 2w\sigma_t\frac{g'}{g}\sigma_x + \frac{(\alpha-2)}{\alpha}\left(\frac{g'}{g}\right)^2\sigma_x^2\right)W^\alpha g^{\alpha-1} + \\ & \frac{\alpha-1}{2\alpha}\frac{g''}{g}\sigma_x^2 W^\alpha g^{\alpha-1}, \end{aligned} \tag{D.37}$$

where the second term equals

$$\begin{aligned} \frac{1}{2\alpha}\frac{d[W^\alpha g^{\alpha-1}]_t}{W^\alpha g^{\alpha-1}} &= \frac{\alpha}{2}\left(w\sigma_t + \frac{\alpha-1}{\alpha}\frac{g'}{g}\sigma_x\right)^2 W^\alpha g^{\alpha-1} \\ &= \left(\frac{\alpha}{2}w^2\sigma_t^2 + (\alpha-1)w\sigma_t\frac{g'}{g}\sigma_x + \frac{(\alpha-1)^2}{2\alpha}\left(\frac{g'}{g}\right)^2\sigma_x^2\right)W^\alpha g^{\alpha-1} \end{aligned} \tag{D.38}$$

and we used the notation $\hat{\kappa}$ for $\hat{\kappa}^i$ defined as

$$\hat{\kappa}^i = \kappa + \eta^i + \frac{f(y)}{\sigma_t}1_{w<0} + \frac{\tau y f(y)}{\sigma_t}1_{w>0}. \tag{D.39}$$

Plugging in the HJB, we have

$$\begin{aligned} 0 = \sup_{c,w}\ & \frac{c^\alpha}{\alpha} + g^{\alpha-1}\left(r + n + \pi + w(\hat{\kappa}\sigma_t) - c - \frac{\rho+\pi}{\alpha}\right) - \frac{1}{2}w^2\sigma_t^2 g^{\alpha-1} + \\ & \frac{\alpha-1}{\alpha}\frac{g'}{g}\mu_x g^{\alpha-1} + \frac{\alpha-1}{2\alpha}\frac{g''}{g}\sigma_x^2 g^{\alpha-1} - \frac{\alpha-1}{2\alpha}\left(\frac{g'}{g}\right)^2\sigma_x^2 g^{\alpha-1}, \end{aligned} \tag{D.40}$$

which implies $c = g$ and $w = \sigma_t^{-1}\hat{\kappa}$ when $w \neq 0$ (interior solution).

Finally, we plug these values back in the HJB to obtain the ODE

$$0 = r + n + \pi + w^i \hat{\kappa}^i \sigma_t - \frac{1}{2}(w^i)^2 \sigma_t^2 - \frac{\rho + \pi}{\alpha} - \frac{\alpha - 1}{\alpha} g_i + \frac{\alpha - 1}{\alpha} \frac{g_i'}{g_i} \mu_\omega^i - \tag{D.41}$$
$$\frac{\alpha - 1}{2\alpha} \left(\frac{g_i'}{g_i} \right)^2 \sigma_\omega^2 + \frac{\alpha - 1}{2\alpha} \frac{g_i''}{g_i} \sigma_\omega^2.$$

In the above equation, the interest rate r is a function of the unknown functions g_i and their derivatives, as follows.

We start by computing

$$\frac{p'}{p} = -\left(\sum g_i(x) x_i \right)^{-1} \left(\sum g_i'(x) x_i + g_R - g_I \right) \tag{D.42}$$

$$\frac{p''}{p} = 2 \left(\sum g_i(x) x_i \right)^{-2} \left(\sum g_i'(x) x_i + g_R - g_I \right)^2 \tag{D.43}$$
$$- \left(\sum g_i(x) x_i \right)^{-1} \left(\sum g_i''(x) x_i + 2 g_R' - 2 g_I' \right)$$

and rewrite (D.10) as

$$\begin{aligned} \mu_t &= B_{x0} + B_{x1} \mu_x - \frac{1}{2} \left(\sum g_i(x) x_i \right)^{-1} \left(\sum g_i''(x) x_i \right) \sigma_x^2 \\ &= B_{x0} + B_{x1} \mu_x - B_{x2} \left(\sum g_i''(x) x_i \right) \end{aligned} \tag{D.44}$$

for functions $B_{xk} = B_{xk}(g, g', x)$ and (D.14), based on the (myopic) optimal choice w^R, as

$$\begin{aligned} \mu_x + x(\mu_t + \delta - p(x)^{-1}) + \sigma_x \sigma_t &= x \left(\mu_t - \kappa \sigma_t + n + (\kappa + f(y) 1_{w^R < 0} / \sigma_t)^2 - g_R \right) + \nu \delta \\ \mu_x + x(\delta - p(x)^{-1}) + \sigma_x \sigma_t &= x \left(n - \kappa \sigma_t + (\kappa + f(y) 1_{w^R < 0} / \sigma_t)^2 - g_R \right) + \nu \delta. \end{aligned} \tag{D.45}$$

Equation (D.45) gives μ_x explicitly. Then we have μ_t from (D.44), which we use in (D.17) to compute r explicitly. We have

$$r = \mu_D - \delta + \frac{1}{p(x)} + \frac{p'(x)}{p(x)} \mu_x + \frac{1}{2} \frac{p''(x)}{p(x)} \sigma_x^2 + \frac{p'(x)}{p(x)} \sigma_x \sigma_D - \kappa \sigma_t. \tag{D.46}$$

A last element is the income flow due to the aggregate fees:

$$n = -(1 - \tau) f(y) w^R x 1_{w^R < 0}. \tag{D.47}$$

We summarize our analysis in the following proposition.

Proposition 5 *A stationary equilibrium is characterized by functions g_i that obey the second-order system of coupled ODE* (D.41) *whose coefficients are defined as follows: p by* (D.8)*, w^R by* (D.23) *or as a solution to* (D.29)*, w^I by* (D.6)*, y by* (D.28)*, σ_t by* (D.18)*, κ by* (D.35)*, n by* (D.47)*, the dynamics of ω by* (D.14) *and* (D.13)*, with μ_ω^I from* (D.15)*, and r by* (D.46)*.*

An explicit computation of the equilibrium price requires a selection rule when multiple

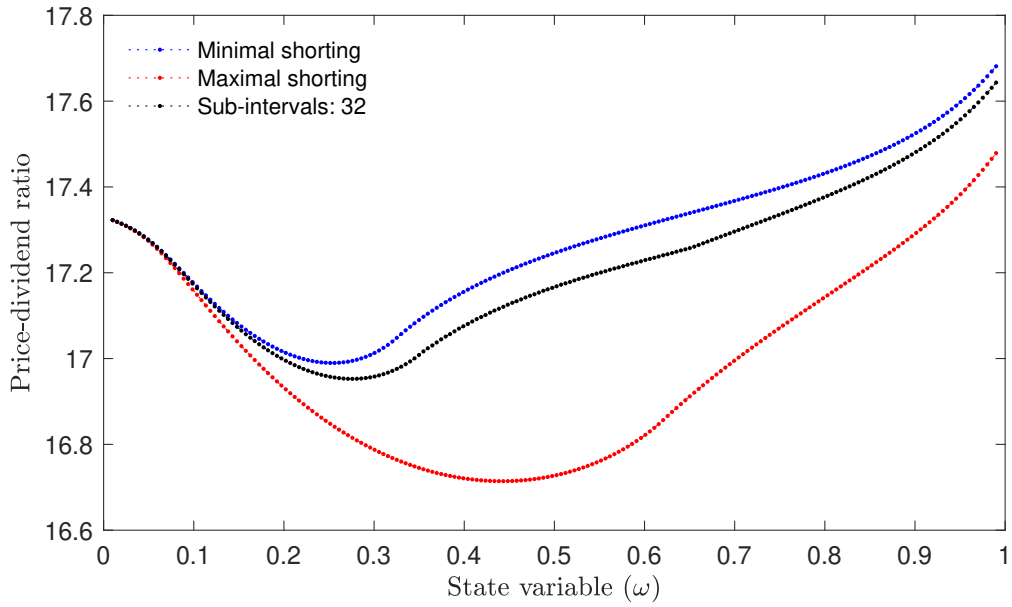


Figure D.1: Price-dividend ratios under different coordination assumptions in the multiplicity region $[\omega_1^*, \omega_2^*]$. The line "minimal shorting" assumes that investors always coordinate on the zero-shorting equilibrium whenever $\omega_t \in [\omega_1^*, \omega_2^*]$. The line "maximal shorting" assumes that investors coordinate on the highest shorting equilibrium whenever $\omega_t \in [\omega_1^*, \omega_2^*]$. The line "sub-intervals: 32" corresponds to splitting the interval $[\omega_1^*, \omega_2^*]$ into 32 sub-intervals and assumes high shorting on even-numbered sub-intervals and zero shorting on odd-numbered sub-intervals. The parameters are $\mu_D = 0.02$, $\sigma_D = 0.04$, $\pi = 0.05$, $\nu = 0.05$, $\delta = 0.05$, $\rho = 0$, $IES = 0.85$, and $\phi = 0.021$.

equilibria are possible. The lines "Maximal Shorting" and "Minimal Shorting" in Figure D.1 depict the price-dividend ratios obtaining when investors always coordinate on the maximal-shorting (resp. zero-shorting) equilibrium. The graph illustrates the general intuition we outlined in Section 6.1, explaining that an equilibrium selection featuring less shorting results in a higher price.

This model specification further enables us to address another question of interest, pertaining to the effect of the frequency of equilibrium switches. Specifically, one might fear that the expectation of switching back to a high-shorting equilibrium in the future reduces substantially the price response to the current switch to a low-shorting one.

To gain some insight in a tractable way, we implement a scheme of frequent equilibrium switching in the following way. We start by dividing the region of multiplicity (the interval $[\omega_1^*, \omega_2^*]$) into a relatively large number of equal-sized sub-intervals. On these sub-intervals, the coordination protocol prescribes high shorting on even-numbered sub-intervals and low shorting on odd-numbered sub-intervals. This specification allows us to capture the notion of frequent equilibrium switching, albeit in a stylized fashion.

The resulting price-dividend ratio is illustrated in Figure D.1 for the case of 32 sub-intervals. The figure shows that the lines labeled "Minimal shorting" and "Sub-intervals:32" are close, indicating that the price-dividend ratio with frequent equilibrium switches is close to the price-dividend ratio where investors always coordinate on the zero-shorting outcome. The reason is that, when the shorting market is active, the wealth share ω_t is more volatile.

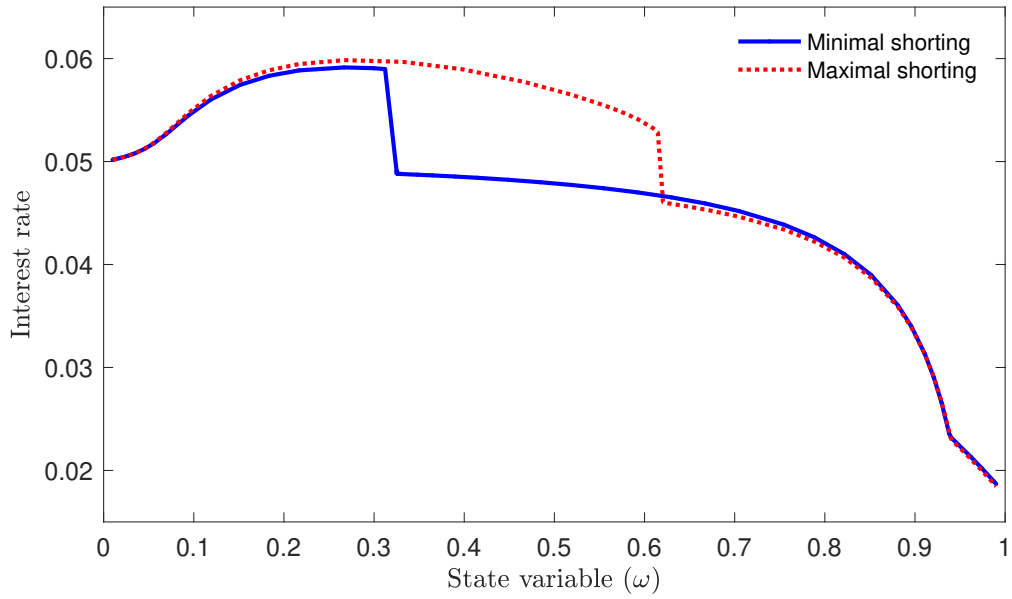


Figure D.2: Interest rate under different coordination assumptions in the multiplicity region $[\omega_1^*, \omega_2^*]$. The line "minimal shorting" assumes that investors always coordinate on the zero-shorting equilibrium whenever $\omega_t \in [\omega_1^*, \omega_2^*]$. The line "maximal shorting" assumes that investors coordinate on the highest shorting equilibrium whenever $\omega_t \in [\omega_1^*, \omega_2^*]$. The parameters are as for Figure D.1.

Accordingly, ω_t "travels fast" inside the regions where the shorting market is active and thus tends to exit these regions quickly. By contrast, when the shorting market is inactive, ω_t fluctuates less, moves more slowly, and as a result spends more time in these regions. We note that this is also the reason why the mass of the stationary distribution in Figure 3 of the paper is more concentrated in the region where the shorting market is inactive.

E A Two-Stock Economy with Limited Participation

In this section we introduce an additional security (stock 2) to our baseline model, which is not subject to any trading frictions. We continue to assume that borrowing stock 1 requires lending fees, as in the baseline model.

We allow one more feature, in the spirit of the "limited recognition hypothesis" of Merton (1987). Specifically, while all investors participate in the markets for stock 2 and the risk-free asset, only a fraction of investors pays any "attention" to stock 1. The remaining fraction of investors simply optimize their portfolio over the risk-free asset and stock 2 and assign zero weight to stock 1.

To ease the comparison of the results of this section with Proposition 2, we maintain the assumption that the lending supply curve is horizontal, and the lending fee is constant and equal to φ.

In this section, we take the equilibrium as given and derive several key restrictions that any equilibrium must obey; we solve fully for the equilibrium in Appendix F, in a limiting economy where stock 1 is small compared to stock 2. We assume that in equilibrium the

returns on stocks 1 and 2 follow a vector diffusion process of the form

$$dR_{1,t} = \mu_{1,t}dt + \sigma_{1,t}dB_{1,t} + b_t\sigma_{2,t}dB_{2,t} \tag{E.1}$$
$$dR_{2,t} = \mu_{2,t}dt + \sigma_{2,t}dB_{2,t}, \tag{E.2}$$

where $B_{1,t}$ and $B_{2,t}$ are independent Brownian motions, and $\mu_{1,t}$, $\mu_{2,t}$, $\sigma_{1,t}$, $\sigma_{2,t}$, and b_t are determined in equilibrium.We assume that investors I believe that Brownian motion 1 follows the dynamics[56] $dB_{1,t}+\eta dt$, while no investor has any belief distortions pertaining to Brownian motion 2.

To facilitate the statement of equilibrium returns, we define $\widetilde{m}_{1,t} \equiv \frac{m_{1,t}}{\widehat{\omega}_t}$ as the ratio of the stock-1 market capitalization share, denoted by $m_{1,t}$, to the wealth share of all investors participating in the market for stock 1, denoted by $\widehat{\omega}_t$. We also define $\kappa_{1,t} \equiv \frac{(\mu_{1,t}-r)-b_t(\mu_{2,t}-r)}{\sigma_{1,t}}$ as the Sharpe ratio of a portfolio long 1 unit of asset 1 and short b_t units of asset 2.

Proposition 6 *In an equilibrium with shorting in asset 1 ($y_t > 0$), y_t is given by the root(s) of the quadratic equation*

$$0 = y\left(\eta + \frac{\widetilde{m}_{1,t}}{\omega_t}\sigma_{1,t} - \frac{\varphi}{\sigma_{1,t}}(1-\tau y)\right) - \left(\eta - \frac{\widetilde{m}_{1,t}}{1-\omega_t}\sigma_{1,t} - \frac{\varphi}{\sigma_{1,t}}(1-\tau y)\right) \tag{E.3}$$

that lie(s) in the interval $(0,1)$, and the Sharpe ratio is given by

$$\kappa_{1,t} = \widetilde{m}_{1,t}\sigma_{1,t} - (1-\omega_t)\eta - \frac{\varphi}{\sigma_{1,t}}(\omega_t + (1-\omega_t)\tau y_t). \tag{E.4}$$

Similarly, in an equilibrium without shorting in asset 1 we have $\kappa_{1,t} = \sigma_{1,t}\widetilde{m}_{1,t} - (1-\omega_t)\eta$ if investor R holds an interior position in asset 1 and $\kappa_{1,t} = \frac{\sigma_{1,t}\widetilde{m}_{1,t}}{1-\omega_t} - \eta$ otherwise.

The excess return to asset 1 is given by

$$\mu_{1,t} - r_t = \kappa_{1,t}\sigma_{1,t} + b_t(\mu_{2,t} - r_t), \tag{E.5}$$

where $\mu_{2,t} - r_t$ is the excess return of asset 2, given by the conventional CAPM relation

$$\mu_{2,t} - r_t = b_t\sigma_{2,t}^2 m_{1,t} + \sigma_{2,t}^2 m_{2,t}. \tag{E.6}$$

Equations (E.3) and (E.4) are the same as (26) and (27), respectively, except that the volatility, σ_D, is replaced by $\widetilde{m}_{1,t}\sigma_{1,t}$. The reason for this replacement is intuitive: In the case of a single stock, the risk of that stock, σ_D, is aggregate (by construction) and commands a risk premium. When there are multiple stocks, the risk compensation for bearing the idiosyncratic risk[57] of stock 1, $\sigma_{1,t}$, is multiplied by $\widetilde{m}_{1,t}$, i.e., the stock market capitalization of stock 1 as a fraction of the wealth share of investors actively participating in the stock. An implication is that when $\widetilde{m}_{1,t}$ approaches zero, the idiosyncratic risk becomes diversifiable,

56. More formally, the Radon-Nikodym derivative of the true probability measure with respect to the subjective one is given by $Z_t^I \equiv e^{-\frac{\eta^2}{2}t+\eta B_{1,t}}$.

57. Recall that the Sharpe ratio in Proposition 6 pertains to a portfolio that invests one dollar in asset 1 and shorts b_t units of asset 2, hedging out the exposure of the portfolio to the second Brownian shock.

and there is no compensation for bearing that risk (the first term on the right-hand side of (E.4) disappears).

Remark 3 *Since the equations determining $\kappa_{1,t}$ and y_t are essentially the same as (27) and (26), Proposition 4 remains unchanged when there are multiple stocks, except that now condition (b) becomes $(\widetilde{m}_{1,t}\sigma_{1,t})^2 < \frac{1}{4}(1-y)^2|h'(y)|$. This condition therefore does not require that the total volatility of stock 1, or even its idiosyncratic part $\sigma_{1,t}$, be small, but rather that the risk of stock 1 be diversifiable by the agents trading it (small $\widetilde{m}_{1,t}$).*

F The Price-Dividend Ratio of a Small Stock

This section provides the details of the entry-and-exit process for the model of Section 6.2.

We start with a few definitions. We let $\overrightarrow{m}_t$ denote the vector of market-capitalization weights of the two stocks, and $m_{j,t}$, $j \in \{1,2\}$, its entries. Since the analysis of interest pertains to asset 1, from now on we use W_t^i to denote the wealth of all agents of type i that participate in the market for stock 1; the relevant state variable is $\omega_t^i \equiv W_t^i/(W_t^R + W_t^I)$, and to save notation we maintain the convention $\omega_t \equiv \omega_t^R$. As stated in the text, $\widehat{\omega}_t$ denotes the wealth share of the investors who actively participate in the market for stock 1. We also let $\widehat{w}_{2,t} = \frac{\mu_{2,t}-r_t}{\sigma_{2,t}^2}$ denote the optimal portfolio holding of stock 2 by investors who don't participate in stock 1, and $\overrightarrow{w}_t^i$ is the (row) vector of portfolio holdings of an investor $i \in \{I,R\}$ that is active in the market for stock 1. Finally, $\overrightarrow{B}_t \equiv (B_{1,t}, B_{2,t})^\top$.

We further define

$$\overrightarrow{B}_t \equiv \begin{bmatrix} B_{1,t} \\ B_{2,t} \end{bmatrix}, \ \sigma_t = \begin{bmatrix} \sigma_{1,t} & b_t\sigma_{2,t} \\ 0 & \sigma_{2,t} \end{bmatrix}, \ \overrightarrow{\varphi} = \begin{bmatrix} \varphi \\ 0 \end{bmatrix}, \ \overrightarrow{\eta} = \begin{bmatrix} \eta \\ 0 \end{bmatrix}. \tag{F.1}$$

The entry and exit into market 1 happens either for endogenous or exogenous reasons. By "endogenous" we mean that investors conduct a cost-benefit analysis before deciding whether to keep paying attention to the market for stock 1. In addition to this optimizing choice, we assume that investors enter and exit the market for exogenous reasons. This exogenous flux of investors is modeled with the sole purpose of making the model solution more tractable and transparent.

Specifically, with W_t^i the (aggregate) wealth of type-i investors that participate in market 1, we assume

$$dW_t^i = dW_t^{i,\text{part}} + \theta\left(\nu^i(W_t^I + W_t^R) - W_t^i\right)dt - 1_{i=R} \times \frac{W_t^I + W_t^R}{1-\omega_t} dF_t + \omega_t^i\left(dL_t - dN_t\right), \tag{F.2}$$

where $dW_t^{i,\text{part}}$ is the wealth growth of all investors of type $i \in \{I,R)$ who already participate in the market for stock 1.[58] The term $\theta\left(\nu_i(W_t^I + W_t^R) - W_t^i\right)dt$ reflects entirely exogenous,

58. For completeness, $dW_t^{i,\text{part}} = W_t^{i,\text{part}}\mu_W^i dt + W_t^{i,\text{part}}(\overrightarrow{w}^i)^\top \sigma_t d\overrightarrow{B}_t$ where

$$\mu_W^i = r_t + \pi + n_t + (\overrightarrow{w}_{t,s}^i)^\top\left(\overrightarrow{\mu}_t - r_t 1_{2\times 1} + \lambda_{t,s}^i \begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) - \frac{c_{t,s}^i}{W_{t,s}^i}.$$

non-optimizing entry, which happens at some rate θ.

As in the baseline model, we are assuming that this exogenous entry-and-exit process affects the composition, but not the sum, of $W_t^I + W_t^R$, since

$$\sum_{i\in\{I,R\}} \theta\left(\nu^i\left(W_t^I + W_t^R\right) - W_t^i\right) = 0.$$

The term $-1_{i=R}\times\frac{W_t^I+W_t^R}{1-\omega_t}dF_t$ captures the endogenous exit of R investors. As we described in the text, the (singular) process dF_t is constructed so that ω_t stays below the critical value $\overline{\omega}$ of ω_t (see (F.3) below) that ensures $V^R(\omega_t) > 0$ for $\omega_t < \overline{\omega}$.

Mostly for technical tractability reasons, we assume another source of exogenous entry and exit, which is reflected in the term $\omega_t^i\left(dL_t - dN_t\right)$ on the right-hand side of (F.2). This entry and exit process leaves the composition of wealth in the market (between R and I investors) unaffected, but ensures that the wealth of the investors who pay attention to the market 1 stays proportional to the "size" of market 1. Specifically, we define dL_t and dN_t as the two singular, increasing processes that control $W_t^I + W_t^R$ so that the ratio of stock market capitalization of asset 1 to the total wealth of investors participating in market 1, $\widetilde{m}_t = \frac{M_{1,t}}{W_t^I+W_t^R}$, stays constant across time ($\widetilde{m}_t = \widetilde{m}$).[59] Because $(dL_t - dN_t)$ is multiplied by ω_t^i, this exogenous entry-and-exit process does not impact the composition of wealth between R and I investors. The purpose of this exogenous entry-and-exit term is transparency and tractability: By ensuring a constant $\widetilde{m}_t$, if there were no differences of opinion ($\eta = 0$), the excess return, the price-dividend ratio, and the volatility of stock 1 would all be constant. Thus, we can eliminate a state variable from the problem, namely the ratio of market capitalization to the total wealth of investors in market 1. Economically, this means that we can abstract from the effects of limited participation (that have been studied extensively in the literature) and isolate the impact of shorting frictions. It is also worth highlighting that the term $\omega_t^i\left(dL_t - dN_t\right)$ endogenously approaches zero as δ_1 and θ approach infinity.[60] Thus, our computations would be approximately valid if we eliminated the term $\omega_t^i\left(dL_t - dN_t\right)$, as long as the analysis focuses on cases where investors are short-termist (θ is large) and the ratio of the dividends of a typical tree 1 to tree 2 mean reverts fast.

Having described the entry and exit of investors into the market for stock 1, we are ready to state a formal result describing the determination of equilibrium in this economy. For simplicity, we assume that the Brownian motions $B_{1,t}$ and $B_{2,t}$ are independent.

Proposition 7 *Using the expressions for w_t^i, $\kappa_{1,t}$ (with $b = 0$), and y_t from Proposition 6, the wealth share ω_t follows the diffusion process*

$$d\omega_t = \mu_\omega(\omega_t)dt + \sigma_\omega(\omega_t)dB_{1,t} - dF_t, \tag{F.3}$$

where F_t is an increasing (singular) process that reflects ω_t to remain below the value $\overline{\omega}$ that

59. These processes can be uniquely constructed from the running maximum and minimum of the difference between $(W_t^R + W_t^I) - M_{1,t}$. For details see Karatzas and Shreve (2012, p. 210) on the Skorohod equation.

60. The reason is that the price-dividend ratio and the ratio of the dividend processes for the two trees (given in (34)) approach constants, thus implying that $\widetilde{m}_t$ approaches a constant ($\widetilde{m}$).

is the lowest value for which $V^R(\omega_t) = 0$*, and* $\mu_\omega(\omega_t)$ *and* $\sigma_\omega(\omega_t)$ *are given by*

$$\mu_\omega(\omega_t) = \omega_t \left(\left(w_{1,t}^R - \widetilde{m} \right) \sigma_{1,t} \left(\kappa_t - \sigma_{1,t} \widetilde{m} \right) + 1_{w_{1,t}^R < 0} w_{1,t}^R \varphi + \frac{y_t \widetilde{m}}{1 - y_t} \varphi \left(1 - \tau \right) \right) + \tag{F.4}$$

$$\theta \left(\nu - \omega_t \right),$$

$$\sigma_\omega(\omega_t) = \omega_t \left(w_{1,t}^R - \widetilde{m} \right) \sigma_{1,t}, \tag{F.5}$$

where $\sigma_{1,t} = \frac{p'(\omega_t)}{p(\omega_t)} \sigma_\omega(\omega_t) + \sigma_{1,D}$ *is the volatility of stock* 1 *and the price-dividend ratio* $p_t = p(\omega_t)$ *solves the ordinary differential equation*

$$\frac{1}{2} \frac{\partial^2 p}{\partial \omega_t^2} \left(\sigma_\omega(\omega_t) \right)^2 + \frac{\partial p}{\partial \omega_t} \left(\mu_\omega(\omega_t) + \left(\sigma_{1,D} - \kappa_{1,t} \right) \sigma_\omega(\omega_t) \right) - p \left(r + \delta_1 + \kappa_{1,t} \sigma_{1,D} \right) + 1 = 0 \tag{F.6}$$

in the region $0 \leq \omega_t \leq \overline{\omega}$*.*

Remark 4 *Since there are multiple equilibrium values for* w_t^i*,* $\kappa_{1,t}$*, and* y_t *in Proposition 7, there exist a large set of solutions for* $p(\cdot)$ *and* $\overline{\omega}$*, depending on the equilibrium on which agents coordinate at each value of* ω_t*.*

The expressions for μ_ω and σ_ω in Proposition 7 coincide with (30) and (29) when $\widetilde{m} = 1$, $\theta = \pi$, and $\sigma_{1,t} = \sigma_D$.[61] Moreover, with the dividend growths of stocks 1 and 2 being independent, so are their stock-price processes (in the limit where stock 1 becomes small) and the expressions for y_t, $w_{1,t}^i$, and $\kappa_{1,t}$ in Proposition 7 (with $\widetilde{m} = 1$ and $\sigma_{1,t} = \sigma_D$) coincide with the respective expressions in the baseline model. Finally, if $\varepsilon = 0$, then $\overline{\omega} = 1$, as in the baseline model. In short, if one dropped the goods-market clearing requirement from the baseline model, the resulting expression for the price-to-dividend ratio would be given by (F.6) (with $\widetilde{m} = 1$ and $\varepsilon = 0$).

The main complications with solving (F.6) are that a) it is a non-linear ODE[62] and b) for $\varepsilon > 0$, this ODE is to be solved over a domain of values of ω_t on which $V^R(\omega_t) > 0$, with $V^R(\overline{\omega}) = 0$ and $\frac{dV^R(\overline{\omega})}{d\omega} = 0$ as boundary conditions.

We solve (F.6) with iterated use of MATLAB's ODE boundary value problem solver BVP5c. We start with the initial guess $\sigma_{1,t} = \sigma_{1,D}$. With that initial guess for $\sigma_{1,t}$ we solve the ODE for $V^R(\omega_t)$ using BVP5c for various boundaries $\overline{\omega}$ until we find the value of $\overline{\omega}$ that satisfies the boundary conditions $V^R(\overline{\omega}) = 0$ and $\frac{dV^R(\overline{\omega})}{d\omega} = 0$. With this $\overline{\omega}$ we compute the solution of the ordinary differential equation (F.6) on the interval $(0, \overline{\omega}]$ using the BVP5c solver and utilizing the (reflecting) boundary condition $p'(\overline{\omega}) = 0$. After obtaining the price-dividend ratio, $p(\omega_t)$ and its derivative, we evaluate $\frac{p'(\omega_t)}{p(\omega_t)}$, and compute $\sigma_{1,t} = \frac{p'(\omega_t)}{p(\omega_t)} \sigma_\omega(\omega_t) + \sigma_{1,D}$. Using this new guess for $\sigma_{1,t}$ we repeat the above procedure until convergence. For our numerical exercise, we assume that $\widetilde{m} = 1$, to make the expressions for the Sharpe ratio, utilization, etc. directly comparable with the baseline model. The rest of the parameters are described in the text.

61. To see this, substitute the expression for the equilibrium interest rate (28) into (30).

62. Equation (F.6) is non-linear because μ_t^i and σ_t^i depend on $\sigma_{1,t}$, which in turn depends on $p(\cdot)$ and $p'(\cdot)$.

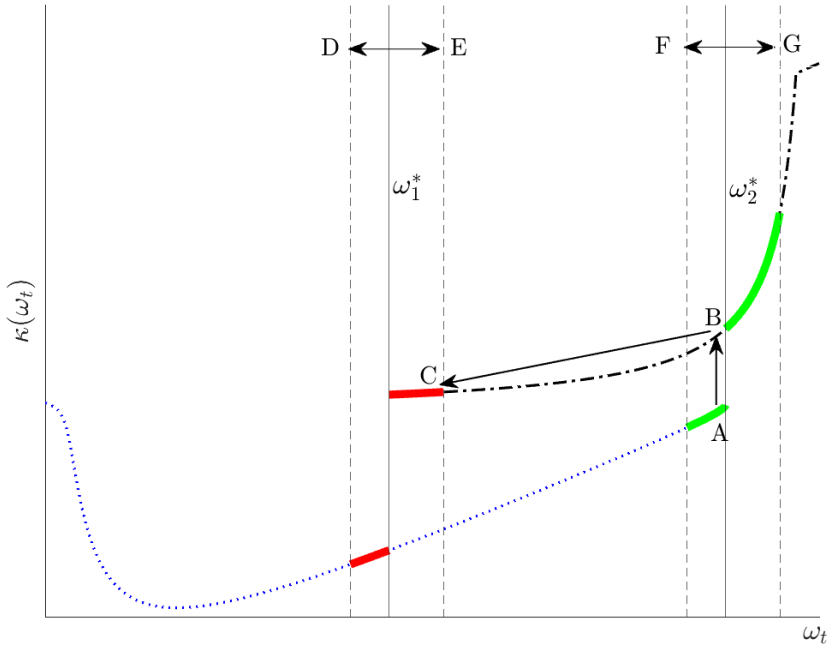


Figure F.1: An illustration of the short-run and long-run effects of an equilibrium shift.

F.1 The dependence of the price-dividend ratio on equilibrium coordination

To better explain how an equilibrium shift impacts the dynamics of the wealth shares and the Sharpe ratio, it is useful to refer to Figure F.1. The dash-dotted black line plots the Sharpe ratio values associated with the no shorting equilibrium. The blue dotted line shows the Sharpe ratios associated with the high shorting equilibrium. To the left of ω_1^* the Sharpe ratio is depicted by the blue line and is unique. Similarly, to the right of ω_2^* the Sharpe ratio is unique and depicted by the dash-dotted black line. Conditional on being in the high shorting equilibrium (point A), a shift from the high to the no shorting equilibrium can be decomposed into an "immediate" and an "eventual" effect. Figure F.1 illustrates the two effects. The first effect captures the immediate upward jump of the Sharpe ratio, and is depicted as a movement from point A to point B. We use the term "immediate" for this effect because the wealth shares are kept fixed. But the equilibrium shift also changes the dynamics of the wealth shares. As a result, the new typical stationary range of values of ω changes, and this is depicted as a move from point B to point C. In Figure F.1 the typical values of ω in the no shorting equilibrium are in the range D–E, while in the high shorting equilibrium, they are in the range F–G. When the market coordinates on the high shorting equilibrium, the associated typical values of the Sharpe ratio are the green points; similarly, when the market coordinates on the no-shorting equilibrium, the associated typical values of the Sharpe ratio are the red points. According to a Campbell-Shiller decomposition, the

log price-dividend ratio depends on the (geometrically-weighted) values of the risk premium over the infinite horizon. This average across the green points exceeds the respective average across the red points, and therefore the price-dividend ratio when agents coordinate on the high-shorting equilibrium is lower. (Intuitively, in the high shorting equilibrium the wealth shares places a significant mass to the right of ω_2^*, where the shorting market is inactive and the Sharpe ratio is particularly high.)

The small attention costs accelerate the transition from point A to point C. Upon an equilibrium shift to the no shorting equilibrium, a mass of short sellers endogenously chooses to exit the market for the small stock, thus making the transition from A to C essentially instantaneous.

G Adding a Non-Pecuniary Cost to Short Selling

In this section we make one change to the baseline model to allow for an additional — non-pecuniary — cost to shorting, meant to capture the increased burden of regulatory compliance brought about by the overhaul of Regulation SHO in 2008.

Specifically, suppose that the utility is given by

$$V_t^i = \int_t^\infty e^{-(\rho+\pi)(u-t)} \left(\log c_u^i + \frac{\chi}{\rho+\pi} w_u \times 1_{w_u<0} \right) du, \tag{G.1}$$

where $\frac{\chi}{\rho+\pi} w_u \times 1_{w_u<0}$ captures this added regulatory burden.

As is intuitive, the higher cost to shorting depresses the demand for shorting, which in turn reduces portfolio heterogeneity across agents and therefore the variability of their relative wealths, as measured by $|\sigma_\omega|$. It follows that the volatility of the utilization ratio y also drops, too, at least for values of ω_t that are small enough.

We can state the following formal result.

Proposition 8 *(i) For ω_t such that the shorting market is active, the stable utilization ratio $y^+ > 0$ decreases with χ. (ii) The diffusive variance σ_ω^2 of ω_t decreases with χ. (iii) For values ω_t sufficiently low, the diffusive variance of y, denoted by σ_y^2, decreases with χ.*

H Proofs

Proof of Proposition 1. Fix parameters $\eta > 0$ and $\psi > 1$ and define φ according to

$$\varphi = \sigma_D \left(\eta - \psi \sigma_D \right) \tag{H.1}$$

for any value of σ_D. Note that when σ_D is sufficiently small, φ is guaranteed to be positive.

We show next that, as σ_D gets close to zero, Assumption 2 is satisfied. Rearranging (H.1) gives

$$\frac{\eta}{\frac{\varphi}{\sigma_D}} = \frac{1}{1 - \psi \frac{\sigma_D}{\eta}}. \tag{H.2}$$

For sufficiently small σ_D we obtain

$$1+\tau > \frac{1}{1-\psi\frac{\sigma_D}{\eta}} > 1. \tag{H.3}$$

Combining (H.2) and (H.3) yields (23).

Turning to (24), we note that the definition of ω_1^* along with (H.1) implies

$$\omega_1^* = 1 - \frac{\sigma_D}{\psi\sigma_D} = \frac{\psi-1}{\psi} > 0,$$

while also

$$\lim_{\sigma_D\to 0} \frac{\sigma_D}{(1+\tau)\frac{\varphi}{\sigma_D}-\eta} = \lim_{\sigma_D\to 0} \frac{\sigma_D}{(1+\tau)(\eta-\psi\sigma_D)-\eta} = 0.$$

Therefore, for sufficiently small σ_D, the left-hand side of (24) converges to $\frac{\psi-1}{\psi} > 0$, while the right-hand side converges to zero, and therefore the inequality holds.

We conclude the proof by showing that $F(\omega)$ has a unique root in the interval $(\omega_1^*, 1)$. To this end, it is useful to introduce the definitions

$$A(\omega) \equiv \tau\frac{\omega}{\sigma_D}\varphi, \tag{H.4}$$

$$B(\omega) \equiv \sigma_D - \omega\left((1+\tau)\frac{\varphi}{\sigma_D}-\eta\right), \tag{H.5}$$

$$C(\omega) \equiv \frac{\omega}{1-\omega}\left(\sigma_D + (1-\omega)\left(\frac{\varphi}{\sigma_D}-\eta\right)\right). \tag{H.6}$$

With these definitions, $F(\omega)$ can be written as $F(\omega) = B^2(\omega) - 4A(\omega)C(\omega)$. We start by observing that $C(\omega_1^*) = 0$ for any parametric choice (since the definition of ω_1^* in equation (21) implies $\sigma_D + (1-\omega_1^*)\left(\frac{\varphi}{\sigma_D}-\eta\right) = 0$). Also, Inequality (24) implies that $B(\omega_1^*) \neq 0$, and thus $B^2(\omega_1^*) > 0$. Accordingly, $F(\omega_1^*) > 0$. Also $B(1) < \infty$, while $C(1) = \infty$. By continuity, there exists at least one value $\omega_2^* \in (\omega_1^*, 1)$ such that $F(\omega_2^*) = 0$.

To show that this value is unique, consider any value $\omega_2^* \in (\omega_1^*, 1)$ such that $F(\omega_2^*) = 0$. We next show that $F'(\omega_2^*) < 0$.

To this end, note that

$$\begin{aligned} F'(\omega) &= 2B(\omega)B'(\omega) - 4\left[A'(\omega)C(\omega) + A(\omega)C'(\omega)\right] \\ &= 2B^2(\omega)\frac{B'(\omega)}{B(\omega)} - 4A(\omega)C(\omega)\left(\frac{A'(\omega)}{A(\omega)} + \frac{C'(\omega)}{C(\omega)}\right). \end{aligned}$$

Since ω_2^* is a root of $F(\omega)$ it follows that $B^2(\omega_2^*) = 4A(\omega_2^*)C(\omega_2^*)$. Therefore,

$$F'(\omega_2^*) = B^2(\omega_2^*)\left(2\frac{B'(\omega_2^*)}{B(\omega_2^*)} - \frac{A'(\omega_2^*)}{A(\omega_2^*)} - \frac{C'(\omega_2^*)}{C(\omega_2^*)}\right). \tag{H.7}$$

We have

$$\frac{A'(\omega_2^*)}{A(\omega_2^*)} = \frac{1}{\omega_2^*}$$
$$\frac{B'(\omega_2^*)}{B(\omega_2^*)} = -\frac{(1+\tau)\frac{\varphi}{\sigma_D} - \eta}{\sigma_D - \omega_2^*\left((1+\tau)\frac{\varphi}{\sigma_D} - \eta\right)}$$

and

$$\frac{C'(\omega_2^*)}{C(\omega_2^*)} = \frac{1}{\omega_2^*(1-\omega_2^*)} + \frac{\eta - \frac{\varphi}{\sigma_D}}{\sigma_D + (1-\omega_2^*)\left(\frac{\varphi}{\sigma_D} - \eta\right)}.$$

Combining terms gives

$$\begin{aligned} &2\frac{B'(\omega_2^*)}{B(\omega_2^*)} - \frac{A'(\omega_2^*)}{A(\omega_2^*)} - \frac{C'(\omega_2^*)}{C(\omega_2^*)} \\ &= -\frac{2\left((1+\tau)\frac{\varphi}{\sigma_D} - \eta\right)}{\sigma_D - \omega_2^*\left((1+\tau)\frac{\varphi}{\sigma_D} - \eta\right)} - \frac{1}{\omega_2^*} - \frac{1}{\omega_2^*(1-\omega_2^*)} - \frac{\eta - \frac{\varphi}{\sigma_D}}{\sigma_D + (1-\omega_2^*)\left(\frac{\varphi}{\sigma_D} - \eta\right)}. \end{aligned} \tag{H.8}$$

For future reference, we note that using $\omega_2^* > \omega_1^*$ along with (23) and the definition of ω_1^* implies that

$$\sigma_D + (1-\omega_2^*)\left(\frac{\varphi}{\sigma_D} - \eta\right) > \sigma_D + (1-\omega_1^*)\left(\frac{\varphi}{\sigma_D} - \eta\right) = 0. \tag{H.9}$$

Using (H.1) we can write the right-hand side of (H.8) as

$$-\frac{2\left((1+\tau)(\eta - \psi\sigma_D) - \eta\right)}{\sigma_D - \omega_2^*\left((1+\tau)(\eta - \psi\sigma_D) - \eta\right)} - \frac{1}{\omega_2^*} - \frac{1}{\omega_2^*(1-\omega_2^*)} - \frac{\psi}{1 - \psi(1-\omega_2^*)}. \tag{H.10}$$

Taking the limit as σ_D approaches zero, the expression (H.10) converges to

$$-\frac{1}{1-\omega_2^*} - \frac{\psi}{1-\psi(1-\omega_2^*)} < 0,$$

where the inequality follows from (H.9) along with (H.1).[63]

The fact that the derivative $F'(\omega_2^*) < 0$ for any root of the equation $F(\omega_2^*) = 0$ in the interval $(\omega_1^*, 1)$ implies that the root ω_2^* must be unique. ■

Proof of Proposition 2. In preparation for the proof, we state and prove an auxiliary result.

63. Equation (H.1) implies $\frac{\varphi}{\sigma_D} - \eta = -\psi\sigma_D$, and therefore $0 < \sigma_D + (1-\omega_2^*)\left(\frac{\varphi}{\sigma_D} - \eta\right) = \sigma_D(1 - (1-\omega_2^*)\psi)$, where the inequality follows from (H.9).

Lemma 1 *The following statements hold for the quadratic equation* (26)*.*

1. $\omega_1^* < \omega_2^*$ *and the discriminant of (26) is non-negative for all* $\omega_t \leq \omega_2^*$*.*
2. *When* $\omega_1^* \leq \omega_t \leq \omega_2^*$*, the two roots of the equation are both in the interval* $[0,1)$*.*
3. *For* $\omega_t \in [0, \omega_1^*)$*, only the larger root of (26) is in the interval* $(0,1)$*.*
4. *If* y *is a root of (26), then* $(1-\omega_t)\eta - \sigma_D - \frac{1-\omega_t}{\sigma_D}\varphi(1-\tau y) > 0$*.*

Proof of Lemma 1. We start with part 1. Using the definitions (H.4)–(H.6), equation (26) can be written in the familiar form

$$A(\omega_t) y^2 + B(\omega_t) y + C(\omega_t) = 0,$$

and the discriminant of this quadratic equation is given by $F(\omega_t)$ as defined in equation (22).

For $\omega_t \leq \omega_1^*$, $C(\omega_t) < 0$ and the discriminant, $B^2(\omega_t) - 4A(\omega_t)C(\omega_t)$, is positive. The assumption that ω_2^* is the unique root of $F(\omega)$ along with the facts that $F(\omega_1^*) = B^2(\omega_1^*) > 0$ and $F(1) = -\infty$ imply that $\omega_1^* < \omega_2^*$.[64] The uniqueness of the root ω_2^* also implies that $F(\omega_t) = B^2(\omega_t) - 4A(\omega_t)C(\omega_t) \geq 0$ for all $\omega_t \leq \omega_2^*$.

We now turn to part 2. To economize on notation we write A rather $A(\omega_t)$ and similarly for B and C. Fix a given ω_t and let $g(y) = Ay^2 + By + C$. We have $g(1) = A+B+C = \frac{\sigma_D}{1-\omega_t} > 0$ and $g'(1) = 2A + B = \sigma_D + \omega_t\left(\eta - (1-\tau)\frac{\varphi}{\sigma_D}\right) > 0$, where the inequality follows from (23). Since $A > 0$, it follows that all roots of $g(y)$ must be smaller than one. Also, the fact that $\omega_t \geq \omega_1^*$ implies that $g(0) = C > 0$, while assumptions (23) and (24) together with the fact that $\omega_t \geq \omega_1^*$ imply that $g'(0) = B < 0$.

The facts that i) $g(y)$ is a convex, quadratic function of y, ii) $g(1) > 0$, $g(0) > 0$, $g'(1) > 0$, and $g'(0) < 0$ and iii) $B^2 - 4AC > 0$ for $\omega_t \in [\omega_1^*, \omega_2^*)$ imply that there are two roots in $(0,1)$.

For part 3, we note that, when $\omega_t < \omega_1^*$, $g(0) = C < 0$, while $g(1) = A+B+C = \frac{\sigma_D}{1-\omega_t} > 0$. Therefore there exists one and only one root in $(0,1)$.

Finally, let $y \in (0,1)$ denote a root of the quadratic equation (26). Accordingly,

$$\begin{aligned}(1-\omega_t)\eta - \sigma_D - (1-\omega_t)\frac{\varphi}{\sigma_D}(1-\tau y) &= \frac{1-\omega_t}{\omega_t} y\left(\sigma_D + \omega_t\eta - \omega_t\frac{\varphi}{\sigma_D}(1-\tau y)\right) \\ &= \frac{1-\omega_t}{\omega_t} y\left(\sigma_D + \omega_t\left(\eta - \frac{\varphi}{\sigma_D}\right) + \omega_t\frac{\varphi}{\sigma_D}\tau y\right) \\ &> 0,\end{aligned}$$

where the last inequality follows from (23). This proves property 4. ■

We now continue with the proof of the proposition. We provide expressions for r_t and κ_t that apply in any equilibrium in which $w_t^R \neq 0$. Since $\sum_i \omega_t^i = 1$, it follows that $\sum_i \sigma_t^i = 0$

64. Assumption (24) implies that $B(\omega_1^*) \neq 0$ and therefore $B^2(\omega_1^*) > 0$.

and $\sum_i \mu_t^i = 0$. Using (29) and $\sum_i \sigma_t^i = 0$ implies that $\sum_i \omega_t^i w_t^i = 1$. Combining $\sum_i \omega_t^i w_t^i = 1$ with (12) along with the definition $y_t = \frac{W_t^-}{W_t^+}$ gives

$$\kappa_t + (1-\omega_t)\,\eta + \left(\omega_t \frac{1}{\sigma_D}\varphi + (1-\omega_t)\,\tau y_t \frac{1}{\sigma_D}\varphi\right) 1_{\left\{w_t^R<0\right\}} = \sigma^D. \tag{H.11}$$

Similarly, using (30) along with $\sum_i \mu_t^i = 0$ and $\sum_i \omega_t^i \left(n_t + w_t^i s_t^i\right) = 0$ gives (28).

We next describe the equilibria for the three intervals of ω_t described in the statement of the proposition.

i) In this case, $\omega_t > \omega_2^*$. The equilibrium prescribes non-negative portfolios for both investors. If $\omega_t > 1 - \frac{\sigma_D}{\eta}$, equation (H.11) implies that $\kappa_t > 0$ and (12) implies that both investors hold positive portfolios and the shorting market is inactive. If $\omega_t \in [\omega_1^*, 1 - \frac{\sigma_D}{\eta})$, then there exists an equilibrium that involves no shorting and a zero portfolio for investor R. We check this assertion by observing that the associated market clearing requirement becomes $(1-\omega_t)\, w_t^I = 1$, which together with $y_t = 0$ leads to (25). We then note that

$$\begin{aligned}\kappa_t + \frac{\varphi}{\sigma_D} =& \frac{\sigma_D}{1-\omega_t} - \eta + \frac{\varphi}{\sigma_D} \\ >& \frac{\sigma_D}{1-\omega_1^*} - \eta + \frac{\varphi}{\sigma_D} \\ =& 0.\end{aligned} \tag{H.12}$$

The first line follows from (25), the second line follows from $\omega_t > \omega_1^*$ and the third line follows from the definition of ω_1^*. Since $\kappa_t + \frac{\varphi}{\sigma_D} > 0$, investor R does not choose a negative portfolio. And since $\kappa_t < 0$ for $\omega_t \in [\omega_1^*, 1 - \frac{\sigma_D}{\eta})$, the investor chooses a zero portfolio.

ii) In this case, $\omega_1^* < \omega_t < \omega_2^*$. Since $\omega_t > \omega_1^*$, equation (H.12) implies that the no-shorting equilibrium continues to be an equilibrium. There exist, however, two more equilibria. To compute them, we guess (and verify shortly) that $w_t^R < 0$. Using (12) and (H.11) gives

$$\begin{aligned} y_t &= \frac{W_t^-}{W_t^+} = \frac{-\omega_t w_{t,s}^R}{(1-\omega_t)\, w_{t,s}^I} = \frac{\omega_t}{1-\omega_t} \frac{-\left(\kappa_t + \frac{1}{\sigma_D}\varphi\right)}{\kappa_t + \eta_t + \frac{1}{\sigma_D}\varphi\tau y_t} \\ &= \frac{\omega_t}{1-\omega_t} \frac{(1-\omega_t)\,\eta - \sigma_D - \frac{1-\omega_t}{\sigma_D}\varphi\,(1-\tau y_t)}{\sigma_D + \omega_t \eta - \frac{\omega_t}{\sigma_D}\varphi\,(1-\tau y_t)}. \end{aligned}$$

Rearranging leads to (26). Statement 1 of Lemma 1 implies that, when $\omega_t \in (\omega_1^*, \omega_2^*)$, equation (26) has two roots in $(0,1)$. Under the supposition that $w_t^R < 0$, equation (H.11) leads to (27). In turn

$$\begin{aligned} \kappa_t^{\pm} + \frac{\varphi}{\sigma_D} &= \sigma_D - (1-\omega_t)\,\eta - \frac{\omega_t}{\sigma_D}\varphi\left(1 + \tau y^{\pm}\frac{1-\omega_t}{\omega_t}\right) + \frac{\varphi}{\sigma_D} \\ &= \sigma_D - (1-\omega_t)\left(\eta + \frac{\varphi}{\sigma_D}\left(1 - \tau y_t^{\pm}\right)\right) < 0, \end{aligned} \tag{H.13}$$

where the last inequality follows from statement 4 of Lemma 1. Combining this observation

with (12) confirms that $w_t^R < 0$. Note that in the second and third equilibria we have that

$$\kappa_t^{\pm} + \eta_t + \frac{1}{\sigma_D}\varphi\tau y_t^{\pm} = \sigma_D + \omega_t\eta - \frac{\varphi\omega_t}{\sigma_D}\left(1 - \tau y_t^{\pm}\right) > 0,$$

where the last inequality follows from (H.13) along with the fact that $y^{\pm}$ satisfy the equation (26). This implies that $w_t^I > 0$.

iii) In this case, $\omega_t < \omega_1^*$. Statement 3 of Lemma 1 implies that the quadratic equation (26) has only one solution in $(0,1)$. This shows that there can only be one equilibrium with shorting. Moreover, this is the unique equilibrium. If w_t^R were zero and the Sharpe ratio were $\frac{\sigma_D}{1-\omega_t} - \eta$, then the inequality in (H.12) reverses, i.e., $\frac{\sigma_D}{1-\omega_t} - \eta + \frac{\varphi}{\sigma_D} < 0$ and investor R would want to deviate from the equilibrium prescription and choose a negative portfolio.

The dynamics of the wealth share follow from a straightforward application of Ito's lemma. ■

Lemma 2 *When the equilibrium is unique,* $0 < \Phi < 1$.

Proof of Lemma 2. We start by noting that an application of the implicit function theorem to (26) gives $\frac{dy}{d\eta} = \frac{1-y}{Z'(y)}$, where $Z(y) \equiv y\left(\eta + \frac{\sigma_D}{\omega_t} - \frac{\varphi}{\sigma_D}(1-\tau y)\right) - \left(\eta - \frac{\sigma_D}{1-\omega_t} - \frac{\varphi}{\sigma_D}(1-\tau y)\right)$. $Z(y)$ is a quadratic equation in y with positive leading coefficient, and satisfies $Z(0) < 0$ when $\omega_t < \omega_1^*$. There consequently exists a unique value $y > 0$ such that $Z(y) = 0$; for this value, $Z'(y) > 0$. Hence, $\frac{dy}{d\eta} > 0$.

Next note that $G_y = \kappa + \eta + 2\frac{\varphi}{\sigma_D}\tau y > 0$, $G_\kappa = y + \frac{\omega_t}{1-\omega_t} > 0$, and $G_\eta = y > 0$. This proves $\Phi > 0$.

Finally, note that $Z(y) = G(y, \kappa(y))$. Therefore, $Z'(y) = G_y + G_\kappa\frac{d\kappa}{dy} = G_y(1-\Phi)$. Since $Z'(y) > 0$ at the equilibrium value of y, it follows that $G_y(1-\Phi) > 0$. Since $G_y = y > 0$, it follows that $\Phi < 1$. ■

Proof of Proposition 3. We note first that, for $w \leq 0$, the function

$$\iota(w,\kappa) \equiv w(\kappa\sigma_D + \varphi) - \frac{1}{2}(w\sigma_D)^2 \tag{H.14}$$

is decreasing in κ, and therefore it attains a higher maximum for equilibrium B (since $\kappa^B < \kappa^A$).

It immediately follows that

$$g_t^B - g_t^A = -\left(\kappa_t^B - \kappa_t^A\right)\sigma_D + \max_{w\leq 0}\iota(w,\kappa_t^B) - \max_{w\leq 0}\iota(w,\kappa_t^A) \geq 0.$$

We further have, based on the expressions for g_t and μ_ω (equation (30)),

$$\begin{aligned}\mu_\omega^B(\omega_t) - \mu_\omega^A(\omega_t) &= \omega_t\left(g_t^B - g_t^A\right) + \frac{1}{2}\omega_t\left(w_t^B(w_t^B-2) - w_t^A(w_t^A-2)\right)\sigma_D^2 \\ &= \omega_t\left(g_t^B - g_t^A\right) + \frac{1}{2}\omega_t\left(w_t^B - w_t^A\right)(w_t^B + w_t^A - 2)\sigma_D^2 \\ &> 0,\end{aligned}$$

with both of the factors in parentheses in the second term on the right-hand side of the second-to-last line being negative (since $w_t^B < w_t^A < 0$). ■

Proof of Proposition 4. We start by describing the determination of the equilibrium in this case. Fix a time t and let E denote expectations with respect to the wealth distribution over types η at time t. (For notational simplicity, we remove time-subscripts throughout the proof.) For a given Sharpe ratio κ and anticipated utilization ration y, define the following two functions, giving the aggregate long and short positions, respectively.

$$L(y,\kappa) = \mathrm{E}\left[\sigma^{-1}\left(\eta+\kappa+\sigma^{-1}\tau y f(y)\right)^{+}\right] \tag{H.15}$$

$$S(y,\kappa) = \mathrm{E}\left[\sigma^{-1}\left(\eta+\kappa+\sigma^{-1} f(y)\right)^{-}\right]. \tag{H.16}$$

An equilibrium is defined through the two market-clearing conditions

$$1 = L(y,\kappa) - S(y,\kappa) \tag{H.17}$$

$$y = \frac{S(y,\kappa)}{L(y,\kappa)}. \tag{H.18}$$

Furthermore, (H.17) defines κ uniquely as a function of y, so that we can write $S(y) = S(y,\kappa(y))$ and $L(y) = L(y,\kappa(y))$, and the equilibrium determination comes down to

$$F(y) \equiv \frac{S(y)}{L(y)} = y. \tag{H.19}$$

The remainder of the proof is organized as follows. We start by showing that, given y_1 with $h'(y_1) < 0$, a continuous distribution with connected support (thus the density does not drop to zero on an intermediate range to then become positive again) exists for which $F'(y_1) > 1$. Using this property, we show that there exist multiple equilibria for this distribution. The continuity of the problem then ensures that, for any sequence of distributions converging to the one we construct,[65] a sequence of equilibrium utilization rates $y_1^{(n)}$ obtain that converges to y_1, and consequently $F'\left(y_1^{(n)}\right) > 1$ for n large enough. In this sense, the set of type distributions admitting multiple equilibria is not "knife-edge" or even sparse, but in fact has non-empty interior.

For convenience, we define $\bar{h}(y) = \frac{h(y)}{\sigma}$ and note that $\bar{h}'(y) < 0$ is equivalent to $h'(y) < 0$. Equation (H.17) implies that

$$\kappa(y) = \frac{\sigma - \omega^S\overline{\eta}^S - \omega^L\overline{\eta}^L - \left(\omega^S\frac{f(y)}{\sigma} + \omega^L\tau y\frac{f(y)}{\sigma}\right)}{\omega^S+\omega^L}, \tag{H.20}$$

65. Convergence in the space of distribution is defined in terms of convergence of expectations of any smooth function with compact support.

where we defined the quantities

$$\omega^L = \mathrm{E}\left[1_{\{\eta+\kappa+\sigma^{-1}\tau y f(y)\geq 0\}}\right] \tag{H.21}$$
$$\omega^S = \mathrm{E}\left[1_{\{\eta+\kappa+\sigma^{-1} f(y)\leq 0\}}\right] \tag{H.22}$$
$$\overline{\eta}^L = \mathrm{E}\left[\eta \,|\, \eta+\kappa+\sigma^{-1}\tau y f(y)\geq 0\right] \tag{H.23}$$
$$\overline{\eta}^S = \mathrm{E}\left[\eta \,|\, \eta+\kappa+\sigma^{-1} f(y)\leq 0\right]. \tag{H.24}$$

(These quantities depend on y, but we suppress that dependence in our notation.)

Furthermore, one can differentiate the same equation (H.17) with respect to y to obtain

$$\kappa'(y) = -\sigma^{-1}\frac{\omega^S f'(y) + \tau\omega^L\left(f(y) + y f'(y)\right)}{\omega^S+\omega^L}, \tag{H.25}$$

where we have made use of the fact that $\frac{d}{dx}\mathrm{E}[(g(x,\eta))^+] = \mathrm{E}\left[\frac{d}{dx}g(x,\eta)1_{\{g(x,\eta)\geq 0\}}\right]$ for an arbitrary differentiable function g, given that the distribution of η is absolutely continuous.

Using equations (H.16) and (H.20) and the definitions of $h(y)$ and $\bar{h}(y)$, we compute

$$S(y) = \sigma^{-1}\frac{\omega^L\omega^S}{\omega^L+\omega^S}\left(\overline{\eta}^L - \overline{\eta}^S - \frac{\sigma}{\omega^L} - \bar{h}(y)\right) = B^{-1}(A-\bar{h}(y)) \tag{H.26}$$
$$F(y) = \frac{\overline{\eta}^L - \overline{\eta}^S - \frac{\sigma}{\omega^L} - \bar{h}(y)}{\overline{\eta}^L - \overline{\eta}^S + \frac{\sigma}{\omega^S} - \bar{h}(y)} = \frac{A-\bar{h}(y)}{A+B-\bar{h}(y)}, \tag{H.27}$$

where we also defined

$$A \equiv \overline{\eta}^L - \overline{\eta}^S - \frac{\sigma}{\omega^L} \tag{H.28}$$
$$B \equiv \frac{\sigma}{\omega^S} + \frac{\sigma}{\omega^L}. \tag{H.29}$$

Noting now, using (H.16) and (H.25), that

$$S'(y) = -\sigma^{-1}\frac{\omega^L\omega^S}{\omega^S+\omega^L}\bar{h}'(y) = -B^{-1}\bar{h}'(y), \tag{H.30}$$

we use (H.26) and (H.30), as well as $F(y) = \frac{S(y)}{L(y)} = \frac{S(y)}{1+S(y)} = 1 - \frac{1}{1+S(y)}$, to write

$$F'(y) = \frac{S'(y)}{(1+S(y))^2} = \frac{-B\bar{h}'(y)}{\left(A+B-\bar{h}'(y)\right)^2}. \tag{H.31}$$

Our intermediate goal, therefore, is to show that, given $\bar{h}'(y_1) < 0$, values A and B exist

satisfying

$$\frac{A-\bar{h}(y_1)}{A+B-\bar{h}(y_1)} = y_1 \tag{H.32}$$

$$\frac{-B\bar{h}'(y_1)}{\left(A+B-\bar{h}(y_1)\right)^2} = 1+\varepsilon > 1 \tag{H.33}$$

for some $\varepsilon > 0$. In fact, for any $\varepsilon > 0$, solutions A and B to these equations are given by

$$B = (1-y_1)^2 \frac{\left|\bar{h}'(y_1)\right|}{1+\varepsilon} > 0 \tag{H.34}$$

$$A = \bar{h}(y_1) + B\frac{y_1}{1-y_1} = \bar{h}(y_1) + y_1(1-y_1)\frac{\left|\bar{h}'(y_1)\right|}{1+\varepsilon} > \bar{h}(y_1). \tag{H.35}$$

To show the existence of a distribution yielding these desired values of A and B, we first note that the right-hand side of (H.29) can be made arbitrarily close to 4σ while keeping $\omega^L + \omega^S < 1$, and therefore condition b) of the proposition ensures that such ω^L and ω^S exist delivering B for a small enough ε. Fixing ω^L and ω^S, $\overline{\eta}^L$ and $\overline{\eta}^S$ can be chosen arbitrarily subject to (H.28) delivering the desired value of A. We therefore now have the value of $\kappa(y_1)$, which determines the sets of types that go long, respectively short, the asset. Finally, the density of the distribution on each of these two sets can be chosen freely subject to the two integrals defining ω^L and $\overline{\eta}^L$, respectively ω^S and $\overline{\eta}^S$. In the complementary, intermediate type region in which agents are inactive, the density is only subject to a total mass condition.

Finally, with $Y = \min\{1, y \,|\, \bar{h}(y) = A\}$, either $Y < 1$ and $F(Y) = 0 < Y$ or $F(Y) = F(1) < 1 = Y$. Since $F(Y) < Y$ in either case, and $F'(y_1) > 1$, a value $y_2 \in (y_1, Y)$ exists such that $y_2 = F(y_2)$. Thus, a second equilibrium exists. ∎

Proof of Proposition 6. The proof essentially repeats the steps from the one-risky asset case, so we provide only a sketch, focusing on the elements that differ.

With these definitions, the market clearing condition is

$$\widehat{\omega}_t \sum_{i\in\{I,R\}} \omega_t^i \overrightarrow{w}_t^i + (1-\widehat{\omega}_t)\begin{bmatrix} 0 \\ \widehat{w}_{2,t} \end{bmatrix} = \overrightarrow{m}_t. \tag{H.36}$$

We consider first an equilibrium with $y_t > 0$. Investor R's and I's optimal portfolios are given by

$$\overrightarrow{w}_t^R = (\sigma_t\sigma_t')^{-1}(\overrightarrow{\mu}_t - r_t 1_{2\times 1} + \overrightarrow{\varphi}), \tag{H.37}$$

$$\overrightarrow{w}_t^I = (\sigma_t\sigma_t')^{-1}(\overrightarrow{\mu}_t - r_t 1_{2\times 1} + \sigma_{1,t}\overrightarrow{\eta} + \tau y_t \overrightarrow{\varphi}). \tag{H.38}$$

Using (H.37) inside (H.36) yields

$$(\sigma_t\sigma_t')\,\overrightarrow{m}_t = \widehat{\omega}_t\left(\omega_t\left(\overrightarrow{\mu}_t - r1_N + \overrightarrow{\varphi}\right) + (1-\omega_t)\left(\overrightarrow{\mu}_t - r1_N + \sigma_1\overrightarrow{\eta} + \tau y_t\overrightarrow{\varphi}\right)\right) + (1-\widehat{\omega}_t)\,(\sigma_t\sigma_t')\begin{bmatrix} 0 \\ \frac{\mu_{2,t}-r}{\sigma_{2,t}^2}\end{bmatrix}. \tag{H.39}$$

Next we use the row selection vector $[0,1]$ to pre-multiply both sides of (H.39). Noting that $[0,1]\,\overrightarrow{\varphi} = [0,1]\,\overrightarrow{\eta} = 0$, and also

$$(\sigma_t\sigma_t')\begin{bmatrix} 0 \\ \frac{\mu_{2,t}-r}{\sigma_{2,t}^2}\end{bmatrix} = \begin{bmatrix} b_t\,(\mu_{2,t}-r) \\ \mu_{2,t}-r\end{bmatrix}, \tag{H.40}$$

leads to (E.6). We next note that

$$[1,-b_t]\,\sigma_t\sigma_t'\begin{bmatrix} m_{1,t} \\ m_{2,t}\end{bmatrix} = [\sigma_{1,t},0]\begin{bmatrix} \sigma_{1,t} & 0 \\ b_t\sigma_{2,t} & \sigma_{2,t}\end{bmatrix}\begin{bmatrix} m_{1,t} \\ m_{2,t}\end{bmatrix} = \sigma_{1,t}^2 m_{1,t}. \tag{H.41}$$

Pre-multiplying both sides of (H.39) with the row vector $[1,-b_t]$, using (H.40), (H.41), and the definition of $\kappa_{1,t}$, and re-arranging yields

$$\kappa_{1,t} = \widetilde{m}_{1,t}\sigma_{1,t} - (1-\omega_t)\,\eta - \frac{\varphi}{\sigma_{1,t}}\left(\omega_t + (1-\omega_t)\tau y_t\right). \tag{H.42}$$

Using the definition of $\kappa_{1,t}$ inside (H.37) gives

$$w_{1,t}^R = \frac{\kappa_{1,t}}{\sigma_{1,t}} + \frac{\varphi}{\sigma_{1,t}^2} \tag{H.43}$$

$$w_{1,t}^I = \frac{\kappa_{1,t}+\eta}{\sigma_{1,t}} + \frac{\tau y_t\varphi}{\sigma_{1,t}^2}, \tag{H.44}$$

where we used the notation $w_{1,t}^i$, $i \in \{R,I\}$, to denote the first element of w_t^i.

Using the market clearing condition $y_t = -\frac{\omega_t^R w_{1,t}^R}{\omega_t^I w_{1,t}^I} = -\frac{\omega_t w_{1,t}^R}{(1-\omega_t)w_{1,t}^I}$ leads to (E.3).

If agent R chooses not to short then the market clearing condition becomes

$$\widehat{\omega}_t(1-\omega_t)\,\overrightarrow{w}_t^I + (1-\widehat{\omega}_t)\begin{bmatrix} 0 \\ \widehat{w}_{2,t}\end{bmatrix} = \overrightarrow{m}_t. \tag{H.45}$$

Substituting in $\overrightarrow{w}_t^I$ from (H.38) and pre-mutiplying by $\sigma_t\sigma_t'$ gives

$$(\sigma_t\sigma_t')\,\overrightarrow{m}_t = \widehat{\omega}_t(1-\omega_t)\left(\overrightarrow{\mu}_t - r1_N + \sigma_1\overrightarrow{\eta}\right) + (1-\widehat{\omega}_t)\,(\sigma_t\sigma_t')\begin{bmatrix} 0 \\ \frac{\mu_{2,t}-r}{\sigma_{2,t}^2}\end{bmatrix}. \tag{H.46}$$

Premultiplying (H.46) by the row $[1, -b_t]$ and using (H.40) and (H.41) gives

$$\sigma_{1,t}^2 \widetilde{m}_{1,t} = (1-\omega_t)\,\sigma_{1,t}\,(\kappa_{1,t}+\eta)\,,$$

and therefore

$$\kappa_{1,t} = \sigma_{1,t}\frac{\widetilde{m}_{1,t}}{1-\omega_t} - \eta. \tag{H.47}$$

Finally, when both agents hold positive portfolios, the optimal portfolios are $\overrightarrow{w}_t^R = (\sigma_t\sigma_t')^{-1}(\overrightarrow{\mu}_t - r_t 1_{2\times 1})$, $\overrightarrow{w}_t^I = (\sigma_t\sigma_t')^{-1}(\overrightarrow{\mu}_t - r_t 1_{2\times 1} + \sigma_{1,t}\overrightarrow{\eta})$. Repeating the arguments in equations (H.37)–(H.42), we obtain $\kappa_{1,t} = \widetilde{m}_{1,t}\sigma_{1,t} - (1-\omega_t)\,\eta$. ■

Proof of Proposition 7. It remains to derive the differential equation in Proposition 7. Using the market clearing condition $\sum_{i\in\{I,R\}} \omega_t^i w_{1,t}^i = \widetilde{m}$, and applying Ito's Lemma to $\omega_t^i = \frac{W_t^i}{W_t^I + W_t^R}$ leads to

$$d\omega_t^i = \mu_{\omega,t}^i dt + \sigma_{\omega,t}^i dB_{1,t} \tag{H.48}$$

with

$$\begin{aligned}
\mu_{\omega,t}^i &= \omega_t^i\left[\left(w_{1,t}^i - \widetilde{m}\right)\sigma_{1,t}\left(\kappa_t - \sigma_{1,t}\widetilde{m}\right) + w_{1,t}^i\lambda_t^i + \widetilde{n}_t\right] + \theta\left(\nu_t^i - \omega_t^i\right),\\
\sigma_{\omega,t}^i &= \omega_t^i\left(w_{1,t}^i - \widetilde{m}\right)\sigma_{1,t},
\end{aligned}$$

and[66]

$$\widetilde{n}_t \equiv -\sum_{i\in\{I,R\}} w_{1,t}^i {\omega_t}^i \lambda_t^i = \frac{y_t\widetilde{m}}{1-y_t} f_t\,(1-\tau)\,.$$

Since $\frac{\phi_1}{\phi_2} \approx 0$, the aggregate endowment follows a geometric Brownian motion in the limit, and the interest rate is constant $r_t = r$. Accordingly, the price of a stock of type 1 follows the dynamics

$$\frac{dP_{1,t,s} + D_{1,t,s}dt}{P_{1,t,s}} = (r + \kappa_{1,t}\sigma_{1,t})dt + \sigma_t dB_{1,t}. \tag{H.49}$$

Applying Ito's Lemma to the product $P_{1,t,s} = p\,(\omega_t)\,D_{1,t,s}$ also implies that

$$\frac{dP_{1,t,s}}{P_{1,t,s}} = \frac{dp_t}{p_t} + \frac{dD_{1,t,s}}{D_{1,t,s}} + \frac{p'\,(\omega_t)}{p\,(\omega_t)}\sigma_{\omega,t}^R\sigma_{1,D}dt. \tag{H.50}$$

66. Using $\sum_{i\in\{I,R\}} w_{1,t}^i\omega_t^i = \widetilde{m}_t$, the definition $y_t = -\frac{w_{1,t}^R\omega_t 1_{\{w_{1,t}^R<0\}}}{w_{1,t}^I\omega_t^I}$ and the definition of λ_t^i leads to

$$-\sum_{i\in\{I,R\}} w_{1,t}^i\omega_t^i\lambda_t^i = \frac{y_t\widetilde{m}}{1-y_t} f_t\,(1-\tau)\,.$$

Combining (H.49) with (H.50) and using $\sigma_{1,t} = \frac{p'(\omega_t)}{p(\omega_t)}\sigma^R_{\omega,t} + \sigma_{1,D}$ and Ito's Lemma to compute the drift of $\frac{dp_t}{p_t}$ leads to

$$\frac{1}{2}\frac{\partial^2 p}{\partial \omega_t^2}\left(\sigma^R_{\omega,t}\right)^2 + \frac{\partial p}{\partial \omega_t}\left(\mu^R_{\omega,t} + \sigma^R_{\omega,t}\sigma_{1,D}\right) - p \times (r + \delta_1 + \kappa_{1,t}\sigma_{1,t}) + 1 = 0, \tag{H.51}$$

which in turn leads to (F.6) after substituting $\sigma_{1,t} = \frac{p'(\omega_t)}{p(\omega_t)}\sigma^R_{\omega,t} + \sigma_{1,D}$. ■

Proof of Proposition 8. Using a standard "guess-and-verify" approach, the modified first-order conditions for portfolio choice of the two investors (conditional on the shorting market being active, i.e., $w^R < 0$) give

$$w^R = \frac{\mu - r + f + \chi}{\sigma^2} = \frac{\kappa}{\sigma} + \frac{f + \chi}{\sigma^2} \tag{H.52}$$

$$w^I = \frac{\mu - r + \tau f y}{\sigma^2} = \frac{\kappa + \eta}{\sigma} + \frac{\tau f y}{\sigma^2}. \tag{H.53}$$

The consumption-to-wealth ratio for all investors continues to be $\rho + \pi$ and therefore the price-dividend ratio is constant and given by $\frac{1}{\rho+\pi}$. As a result, the volatility σ is constant. From this point on, repeating the steps summarized by equations (15)–(20), we obtain

$$y = \frac{\eta - \frac{\chi}{\sigma} - \frac{\sigma}{1-\omega} - \frac{f}{\sigma}(1 - \tau y)}{\eta - \frac{\chi}{\sigma} + \frac{\sigma}{\omega} - \frac{f}{\sigma}(1 - \tau y)}. \tag{H.54}$$

We note that (H.54) is the same as (20), but with η replaced by $\eta - \frac{\chi}{\sigma}$. Applying the implicit function theorem to (H.54) and focusing on the maximum shorting equilibrium (which is the "stable" equilibrium) shows that $\frac{dy}{d\chi} < 0$. In addition, equation (H.52) together with $w^R_t < 0$ and (29) implies $\frac{d\sigma_\omega}{d\chi} > 0$.[67] These observations prove parts (i) and (ii).

For part (iii), consider the derivative of $\sigma_y^2 = \left(\frac{dy}{d\omega}\sigma_\omega\right)^2$:

$$\frac{d\sigma_y^2}{d\chi} = 2\left(\frac{dy}{d\omega}\right)^2 \sigma_\omega^2 \left(\frac{\frac{d\sigma_\omega}{d\chi}}{\sigma_\omega} + \frac{\frac{d^2 y}{d\omega d\chi}}{\frac{dy}{d\omega}}\right). \tag{H.55}$$

The term $\frac{1}{\sigma_\omega}\frac{d\sigma_\omega}{d\chi}$ is negative because $\frac{d\sigma_\omega}{d\chi} > 0$ and $\sigma_\omega < 0$. The remainder of the proof shows that also the second term inside the last parentheses, $\left(\frac{dy}{d\omega}\right)^{-1}\frac{d^2 y}{d\omega d\chi}$, is negative for small ω. To

67. Using the expression for the equilibrium Sharpe ratio gives

$$\begin{aligned} w^R - 1 &= \frac{\kappa}{\sigma} + \frac{f + \chi}{\sigma^2} - 1 = -(1-\omega)\left(\frac{\eta}{\sigma} - \frac{\chi}{\sigma^2}\right) + \frac{f}{\sigma^2}(1-\omega)(1-\tau y) \\ &= -(1-\omega)\left[\left(\frac{\eta}{\sigma} - \frac{\chi}{\sigma^2}\right) - \frac{f}{\sigma^2}(1 - \tau y)\right] < 0. \end{aligned}$$

Therefore $\frac{d\left(w^R - 1\right)}{d\chi} = (1-\omega)\left[\frac{1}{\sigma^2} - \frac{f}{\sigma^2}\tau\frac{dy}{d\chi}\right] > 0$, since $\frac{dy}{d\chi} < 0$.

start, the implicit function theorem gives the following expression[68] for $\frac{dy}{d\omega}$:

$$\frac{dy}{d\omega} = \sigma \frac{\frac{(1-\omega)^2 y - \omega^2}{(1-\omega)^2 \omega^2}}{\eta + \frac{\sigma}{\omega} - \frac{f}{\sigma}(1-\tau y) - \frac{\chi}{\sigma} - \tau(1-y)\frac{f}{\sigma}}. \tag{H.56}$$

We next show that $\frac{dy}{d\omega} > 0$ for small enough ω. To see this, note that the larger root of equation (H.54) satisfies $\lim_{\omega \to 0} y(\omega) = 0$. Moreover, re writing equation (H.54) as

$$y\left(\eta - \frac{\chi}{\sigma} + \frac{\sigma}{\omega} - \frac{f}{\sigma}(1-\tau y)\right) = \eta - \frac{\chi}{\sigma} - \frac{\sigma}{1-\omega} - \frac{f}{\sigma}(1-\tau y)$$

and taking the limit as $\omega \to 0$ on both sides implies that

$$\sigma \lim_{\omega \to 0} \frac{y}{\omega} = \lim_{\omega \to 0} y\left(\eta - \frac{\chi}{\sigma} + \frac{\sigma}{\omega} - \frac{f}{\sigma}(1-\tau y)\right) = \lim_{\omega \to 0} \eta - \frac{\chi}{\sigma} - \frac{\sigma}{1-\omega} - \frac{f}{\sigma} > 0. \tag{H.57}$$

Therefore, for small enough ω $\frac{y}{\omega^2}$ is arbitrarily large and therefore the numerator in (H.56) is positive. We also note that the denominator of $\frac{dy}{d\omega}$ is also positive for the (stable) equilibrium associated with y^+. Accordingly, $\frac{dy}{d\omega} > 0$ for small enough ω.

Next, we totally differentiate $\frac{dy}{d\omega}$ with respect to χ and observe that the sign of $\left(\frac{dy}{d\omega}\right)^{-1} \frac{d^2 y}{d\omega d\chi}$ is the same as the sign of

$$\left(\frac{(1-\omega)^2}{(1-\omega)^2 y - \omega^2} + \frac{2\tau\frac{f}{\sigma}}{\eta + \frac{\sigma}{\omega} - \frac{f}{\sigma}(1-\tau y) - \frac{\chi}{\sigma} - \tau(1-y)\frac{f}{\sigma}}\right)\frac{dy}{d\chi} - \frac{-\frac{1}{\sigma}}{\eta + \frac{\sigma}{\omega} - \frac{f}{\sigma}(1-\tau y) - \frac{\chi}{\sigma} - \tau(1-y)\frac{f}{\sigma}}. \tag{H.58}$$

As $\omega \to 0$, the term $\eta + \frac{\sigma}{\omega} - \frac{f}{\sigma}(1-\tau y) - \frac{\chi}{\sigma} - \tau(1-y)\frac{f}{\sigma}$ approaches infinity, and hence the sign of the expression (H.58) is the same as the sign of $\frac{(1-\omega)^2}{(1-\omega)^2 y - \omega^2}\frac{dy}{d\chi}$, which as we argued above has the same sign as $\frac{dy}{d\chi}$. An application of the implicit function theorem to (H.54) shows that $\frac{dy}{d\chi} < 0$ (for the stable equilibrium y^+). Accordingly, $\frac{d\sigma_y^2}{d\chi} < 0$ for sufficiently small ω. ■

68. Note that equation (H.54) can be written as

$$H(y) \equiv y\left(\eta + \frac{\sigma}{\omega} - \frac{f}{\sigma}(1-\tau y) - \frac{\chi}{\sigma}\right) - \left(\eta - \frac{\sigma}{1-\omega} - \frac{f}{\sigma}(1-\tau y) - \frac{\chi}{\sigma}\right) = 0$$

and therefore (H.56) follows from the implicit function theorem, $\frac{dy}{d\omega} = -\frac{H_\omega}{H_y}$.

Table I.1: Summary Statistics of Shorting fees.

	Percentile				
Size quintile	50^{th}	75^{th}	90^{th}	95^{th}	99^{th}
1	0.41%	0.84%	2.96%	7.43%	28.48%
2	0.38%	0.50%	1.44%	3.97%	19.38%
3	0.36%	0.43%	0.86%	2.04%	12.30%
4	0.34%	0.39%	0.53%	1.02%	7.39%
5	0.35%	0.38%	0.43%	0.50%	1.72%
Total	0.37%	0.51%	1.24%	2.99%	13.85%

Lending fees by stock market capitalization quintile. Each year, we form 5 portfolios of Russell 3000 constituents sorted into size quintiles based on end-of-prior-year market capitalization. Within each size quintile, we compute the p^{th} percentile, $p \in \{50, 75, 90, 95, 99\}$, of daily shorting fees over the following year. We then report the time-series average of these percentiles from 2006 to 2021. Daily shorting fees from 2006 to 2021 are from Markit and are reported as annualized percentage rates.

I Additional Empirical Results

I.1 Summary Statistics — IHS Markit

We start by reporting some summary statistics on lending fees. In Table I.1, we group Russell 3000 constituents based on their end-of-prior-year market capitalization into five quintiles. We then fix the set of stocks in each quintile over the subsequent year and compute various statistics (median, 75th percentile, etc.) of the daily lending fees for the stocks in each quintile. We then average across the years. The table shows that the median lending fee ranges between 0.35% and 0.41%. However, the table also shows that some of the observations on lending fees can become quite large. For instance, for stocks that are in the size portfolios 1, 2, and 3, the 95-th percentile of fees exceeds 2% and the 99-th percentile exceeds 7% for stocks in portfolios 1, 2, 3, and 4. This table suggests that sometimes even relatively large stocks (by market capitalization) can exhibit sizable lending fees.

Table I.2 helps to illustrate this last point in greater detail. Specifically, Table I.2 reports some stock-level statistics on lending fees, and in particular the fraction of Russell 3000 constituents for which a given percentile of shorting fees across time exceeds certain cutoffs. The table shows that 96% of Russell 3000 constituents exhibit a lending fee in excess of 1% at some point between 2006 and 2021, while 45% of stocks exhibit a fee in excess of 5% at some point over that same time period. But even if we leave these extreme observations aside, and focus on — say — the 95-th percentile of the distribution of lending fees at the stock level, the numbers are large: 31% of Russell 3000 constituents exhibit a lending fee in excess of one percent for 5 out of 100 trading days, while 18% of Russell constituents exhibit lending fees in excess of 3% for 5 out out of 100 trading days.

Table I.2: Stock-level distribution of shorting fees.

	Shorting fee cutoffs				
Percentile	≥1%	≥2%	≥3%	≥5%	≥10%
90^{th}	0.23	0.16	0.12	0.09	0.05
95^{th}	0.31	0.21	0.18	0.14	0.07
99^{th}	0.50	0.32	0.26	0.21	0.13
99.5^{th}	0.62	0.38	0.30	0.23	0.14
Maximum	0.96	0.79	0.66	0.45	0.27

Fraction of Russell 3000 constituents for which the indicated percentile (first column) of daily shorting fees exceeds the cutoff noted in the header row. For example, the bottom rightmost number (0.27) means that 27% of the stocks in the Russell 3000 had a maximum daily shorting fee in excess of 10%. Similarly, the number 0.12 in the top row/ middle column indicates that 12% of the stocks have a lending fee in excess of 3 percent for one out of the ten trading days. Daily shorting fees from 2006 to 2021 are from Markit and are reported as annualized percentage rates.

I.2 Heterogeneous $h'(y)$

For our baseline results we pooled observations across all stocks and estimated a single function $h'(y)$. Figure I.1 shows results for the case where we allow $h'(y)$ to differ for each stock. Specifically, we focus on observations that are on special (DCBS > 1) and estimate a separate $h'(y)$ for each Russell 3000 constituent. We then evaluate $h'(y)$ for different values of y for each stock separately. Subsequently, we pool all $h'(y)$ values across all stocks and present them as a bin-scatter diagram.[69] Since stock-level estimates of $h'(y)$ are noisy, we trim stock-level estimates of $h'(y)$ at the 5^{th} and 95^{th} percentile levels. (Results are similar if we don't trim and instead report medians by shorting-utilization bin.) The main conclusion from Figure I.1 is similar to our conclusion in the text: for low values of y, $h'(y)$ is negative.

69. Since the observations per stock are not in the millions (as they are for the pooled regressions in the text), it is computationally feasible to use a kernel regression estimator with automatic, cross-validated, bandwidth selection. We present the results for this alternative estimation method, as a check that our conclusions are not driven by whether we use kernels or splines to estimate the non-parametric regression.

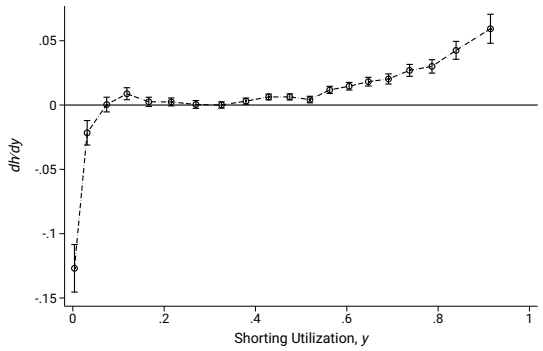
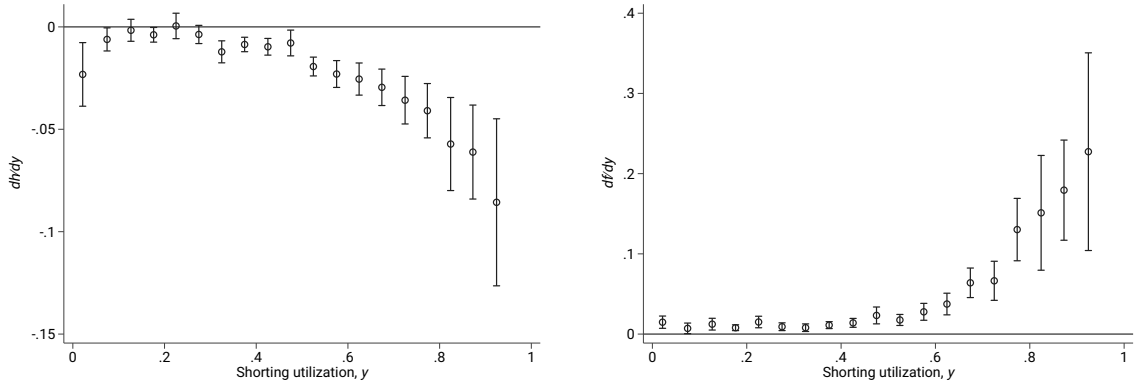


Figure I.1: $h'(y)$, binned-means from stock-level estimates using local-linear non-parametric kernel regressions. For each stock, we estimate the marginal effect $h'(y)$ at 11 points, corresponding to the stock-level deciles, as well as the 5^{th} and 95^{th} percentiles of utilization for observations exhibiting a Daily Cost of Borrowing Score (DCBS) over 1. Error bars represent 95% confidence intervals around bin means. Data on shorting fees and shorting utilization are from Markit over the period 2006 to 2021.



(a) $h'(y)$ (b) $f'(y)$

Figure I.2: $h'(y)$ and $f'(y)$, binned-means from jump discontinuities. Estimated marginal effects are calculated by dividing observed weekly changes in $h(y) = f(y)\,(1 - \tau\, y)$ by observed weekly changes in shorting utilization, restricting attention to weeks in which the magnitude of the change in shorting utilization magnitude exceeds 10.5%, four times the standard deviation of absolute changes in utilization. We calibrate τ to be 0.8 based on industry literature on the pass-through of shorting fees to institutional investors. Sample consists of daily observations of shorting fees and shorting utilization for Russell 3000 constituents. Error bars represent 95% confidence intervals around bin means. Data on shorting fees and shorting utilization are from Markit over the period 2006 to 2021.

Table I.3: Determinants of Utilization Jump Rate

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$1_{\text{Fail - Satisfy}}$	-5.14^{***}	-4.78^{***}	-3.62^{***}	-5.28^{***}	-5.00^{***}	-3.85^{***}	-2.96^{***}
	(-5.63)	(-5.67)	(-4.30)	(-5.79)	(-5.51)	(-4.37)	(-3.82)
Shorting Fee		59.68^{***}					52.64^{***}
		(11.03)					(9.65)
Size Quintile							
2			-8.86^{***}				-7.10^{***}
			(-14.79)				(-12.98)
3			-10.76^{***}				-8.17^{***}
			(-14.33)				(-11.38)
4			-13.17^{***}				-8.35^{***}
			(-14.62)				(-9.35)
5			-12.62^{***}				-8.88^{***}
			(-9.53)				(-6.60)
Var of Returns				63.37^{**}			24.73^{**}
				(1.99)			(2.31)
Turnover				36.80^{**}			-4.03
				(1.99)			(-0.24)
Debt/Total Assets					-4.61^{***}		-1.36
					(-3.36)		(-1.16)
Log Book/Market					0.86^{***}		1.22^{***}
					(2.72)		(3.86)
$\mathbf{1}_{\text{Option}}$						-9.29^{***}	-5.80^{***}
						(-11.82)	(-8.27)
$\mathbf{1}_{\text{NASDAQ}}$						1.36^{*}	-0.04
						(1.88)	(-0.05)
N	1975	1975	1975	1975	1975	1975	1975

t statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

This table reproduces the estimates of Table 1 and, in addition, reports the coefficients on the various stock-level characteristics. All columns also include dummies for the quintile that each stock belongs to based on the t-statistic of satisfying the assumption of Proposition 4 (see the description in the text). For ease of comparison with Table 1 we report (in the first row) only the coefficient on the dummy variable that compares stocks in the fifth quintile vs. the first (base) quintile under the heading $1_{\text{Fail-Satisfy}}$.

J Details and Additional Results for Section 7

J.1 Measuring ticker discussion on WallstreetBets

Our measure of ticker mentions on WallstreetBets is constructed as follows. We use the PushshiftAPI to collect all submissions posted on WallstreetBets subreddit from January 1, 2020 through February 7, 2021 (Baumgartner et al. 2020). For each submission, we observe the title text, the body of the submission, the author of the submission, and the time of the submission.

We then identify all cases in which these tickers are mentioned in submissions, irrespective of whether they are prefixed with a dollar sign. To address the possibility of falsely identifying tickers, we require that, if the ticker is a common word in the written English language, it must be prefaced by a dollar sign. For example, AT&T's ticker T is also a common word in written English, and thus we require that the text "$T" appear in a submission for it to be considered as mentioned AT&T. We consider a ticker as being mentioned in a submission if it appears in either the title or the body of the submission. We identify common word-stems based on the Google Trillion Word Corpus (Michel et al. 2011). In a robustness check, we account for the downward bias this restriction introduces by scaling common-word tickers by an in-sample estimated adjustment factor. This adjustment leaves the relative ranking of ticker mentions largely unchanged. We estimate the adjustment factor by comparing the frequency of tagged ticker mentions versus un-tagged ticker mentions for the set of tickers which do not commonly appear in written English.

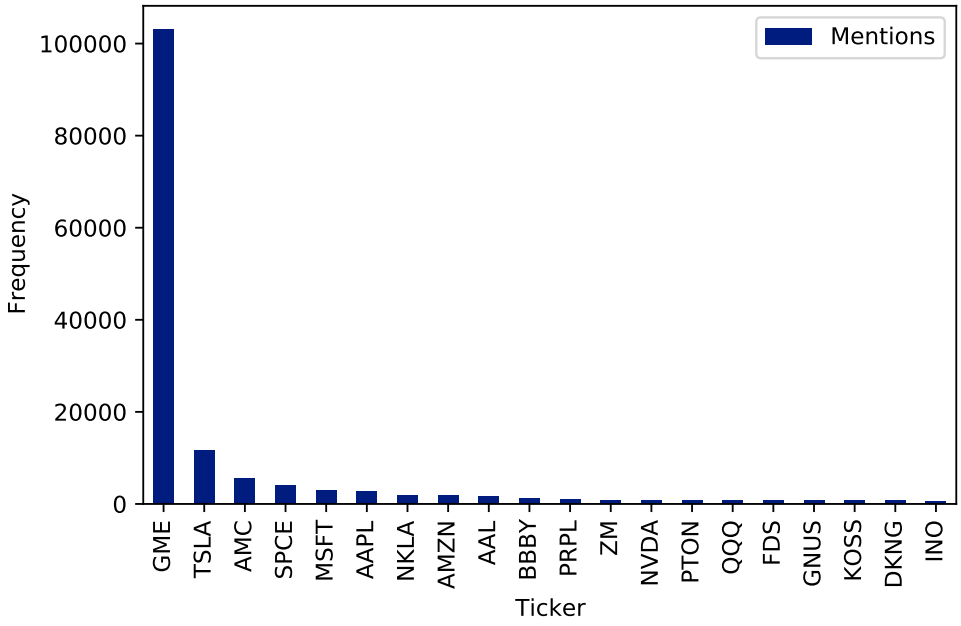
Revised submissions and comments. Authors of Reddit comments have the ability to edit their comments even after the comment has been posted. The PushshiftAPI records the comment text as of a certain day, and does not update to reflect potential revised comments. The same constraint applies to the content body of submissions. Titles of submissions cannot be revised and thus do not have this measurement problem.

Missed tickers Tickers that, for whatever reason, are never tagged with a leading dollar sign will be omitted from our dataset. Similarly, we under-count the occurrences of tickers that are common words, owing to requiring they appear with a leading "$" We attempt to correct for this by scaling the observed counts for common word tickers. For AAPL and GME, which are not common word tickers, the ticker appears with the leading "$" roughly 20% of the time. We can thus simply multiply our observed frequencies by a factor of five to adjust for the more stringent matching procedure. As can be seen in Figures J.1a and J.1b, the adjustment does not have a significant impact on the relative popularity of the top tickers.

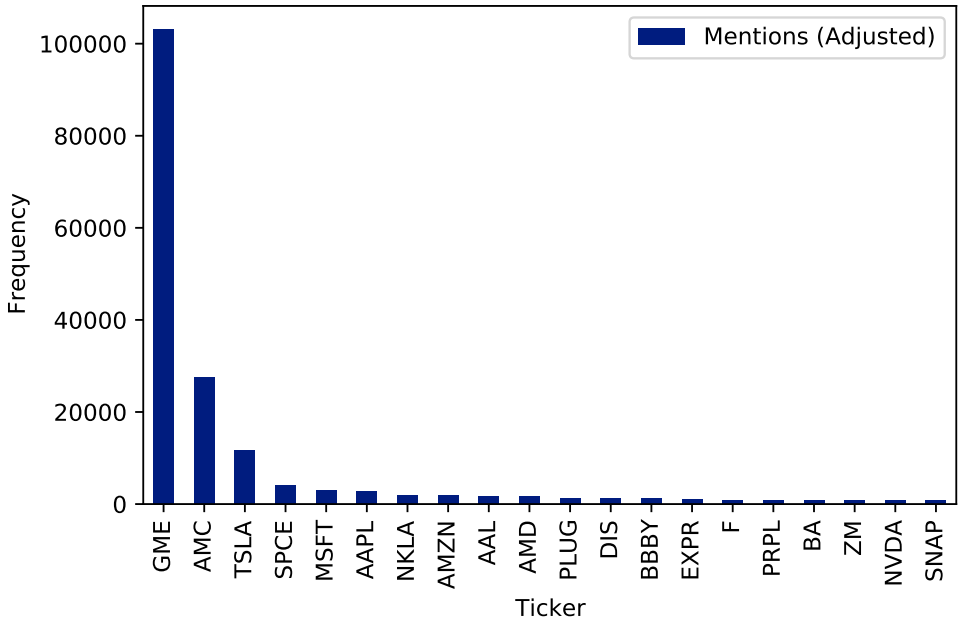
In some cases, users may choose to refer to the company by its name, rather than by its ticker. We do not attempt to identify mentions of companies by name.

J.2 Measuring retail trading

We adopt the methodology of Boehmer et al. (2021) to identify retail trades in the TAQ data. We briefly summarize the methodology here and refer readers to the paper for details.



(a) Submissions mentioning each Ticker



(b) Submissions mentioning each Ticker, adjusted for word-ticker overlap

Figure J.1: Popular Tickers on WallstreetBets (January 1, 2020 – February 7, 2021).

The intuition behind the methodology is the knowledge that retail trades are often executed by wholesalers or via broker internalization, rather than on the major trading exchanges. These trades appear in the TAQ consolidated tape data under the exchange code "D." These trades are given a small price improvement on the order of tenths of a penny as a means to induce brokers to route orders to the wholesaler. Similarly, brokers which internalize retail trades offer a subpenny price improvement in order to comply with Regulation 606T. Importantly, institutional trades are rarely, if ever, internalized or directed to wholesalers and their trades are usually in round penny prices, with the notable exception of midpoint trades.

The methodology of Boehmer et al. (2021) uses these institutional details to identify retail trades in the TAQ consolidated tape data. Trades flagged with exchange code "D" and with a subpenny amount in the set $(0, 0.40) \cup (0.60, 1.00)$ are identified as retail trades. Splitting these trades further, retail trades with subpenny amounts between zero- and forty-hundredths of a penny are labeled as "sell orders," whereas subpenny amounts between sixty- and one hundred-hundredths are considered "buy orders." The midpoint trades are excluded to avoid mis-classifying institutional trades executed at midpoints as retail trades.

J.2.1 Challenges

Derivatives The TAQ data only contains trades of equities. Options offer another way to benefit for investors to benefit from increases in the price of stock. As an added advantage for retail investors, options offer embedded leverage greater than what might otherwise be available through their broker. The Boehmer et al. (2021) methodology relies on institutional details to identify off-exchange retail trades, and thus cannot reliably identify replication trades by market makers. See Barber et al. (2024) for additional discussion of the limitations of the aforementioned methodology.

J.3 Betting against the shorts portfolio

As is standard in the literature, we restrict attention to common shares of COMPUSTAT firms which trade on the NYSE, AMEX, and NASDAQ exchanges. We further exclude companies for whom no share class has a price exceeding $1. The strategy equally weights each firm in the top decile, shorts the market index, and reconstitutes 8 trading days following the disclosure date, which is the first opportunity following the public dissemination of the short interest data.

	Highly Shorted Stocks	Excl. Popular Reddit Stocks	Excl. Small Stocks
Panel A: November 2020			
r^{EW}	0.164	0.161	0.226
	(4.330)	(4.249)	(5.077)
r^{VW}	0.093	0.092	0.132
	(3.208)	(3.186)	(3.446)
r^{EW}_{FF3}	0.086	0.083	0.157
	(3.641)	(3.512)	(4.373)
r^{VW}_{FF3}	0.044	0.042	0.080
	(1.871)	(1.819)	(2.461)
Panel B: December 2020			
r^{EW}	0.057	0.061	0.019
	(1.509)	(1.606)	(0.420)
r^{VW}	0.033	0.036	0.020
	(1.128)	(1.243)	(0.517)
r^{EW}_{FF3}	0.017	0.021	-0.000
	(0.727)	(0.882)	(-0.009)
r^{VW}_{FF3}	0.013	0.015	0.008
	(0.546)	(0.671)	(0.257)
Panel C: January 2021			
r^{EW}	0.271	0.233	0.156
	(6.835)	(5.865)	(3.576)
r^{VW}	0.194	0.160	0.182
	(6.658)	(5.576)	(4.762)
r^{EW}_{FF3}	0.205	0.167	0.113
	(8.685)	(7.059)	(3.135)
r^{VW}_{FF3}	0.163	0.128	0.155
	(6.895)	(5.598)	(4.741)

Table J.1: Portfolio returns (November 2020–January 2021). Test of whether the monthly return to the strategy of betting against the shorts is "abnormal" in November 2020 (Panel A), December 2020 (Panel B), and January 2021 (Panel C). The table reports the coefficient and the t-statistic of the month dummy variable that takes the value of one for the month listed in the title of the panel and zero otherwise from the regression:

$$r_{\text{Betting against the shorts}} = \text{const.} + \text{month dummy} + \beta' F_t + \varepsilon_t.$$

The first two rows of each panel do not control for any factor exposures and refer to equal-weighted (EW) and value-weighted (VW) returns, respectively. The last two rows of each panel control for Fama-French 3-factor exposures.

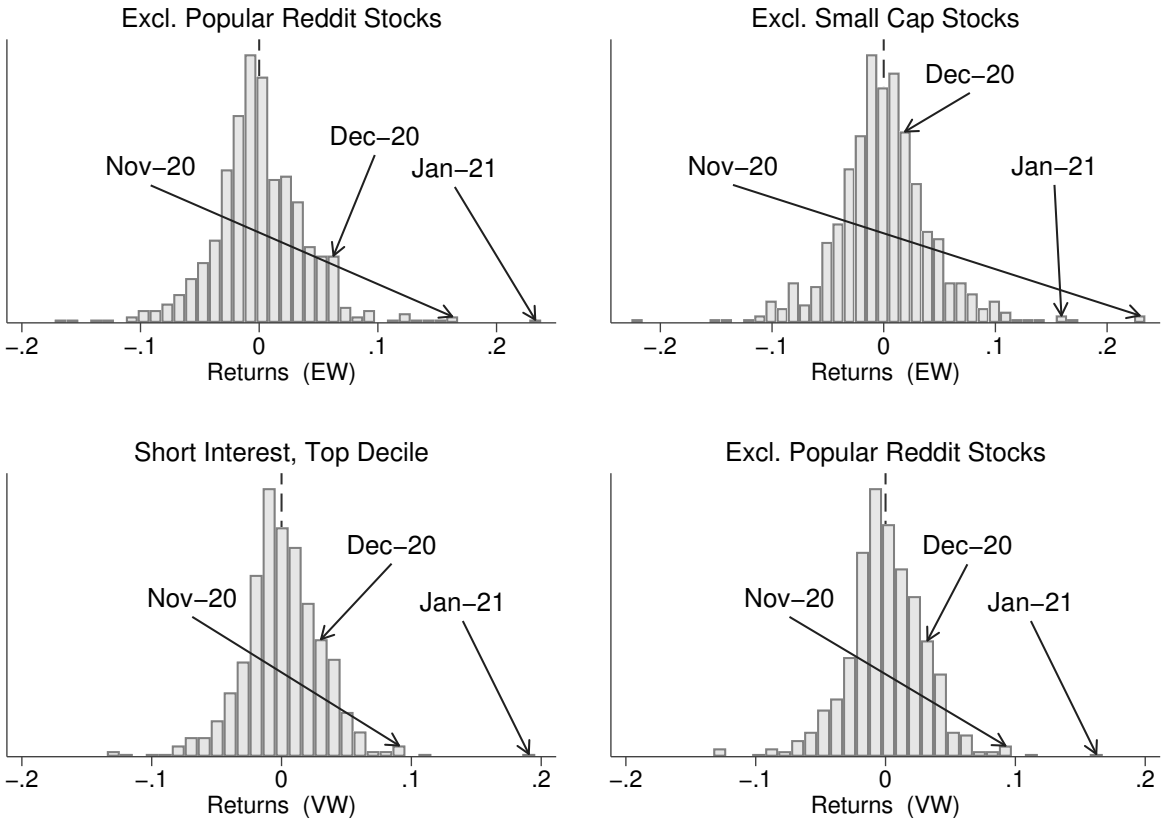


Figure J.2: Monthly returns (1973–2021). Histograms show monthly returns to a trading strategy long stocks in the top decile of short interest and short the market index. The top-left plot depicts equal-weighted returns, excluding the six most-popular stocks discussed on Reddit (AMC, BBBY, GME, SPCE, TLRY, and TSLA). The top-right plot depicts equal-weighted returns, further excluding small market capitalization stocks. The bottom-left plot depicts value-weighted returns. The bottom-right plot depicts value-weighted returns, excluding popular stocks discussed on Reddit. The arrows indicate the portfolio returns in the months of November and December 2020 and January 2021.

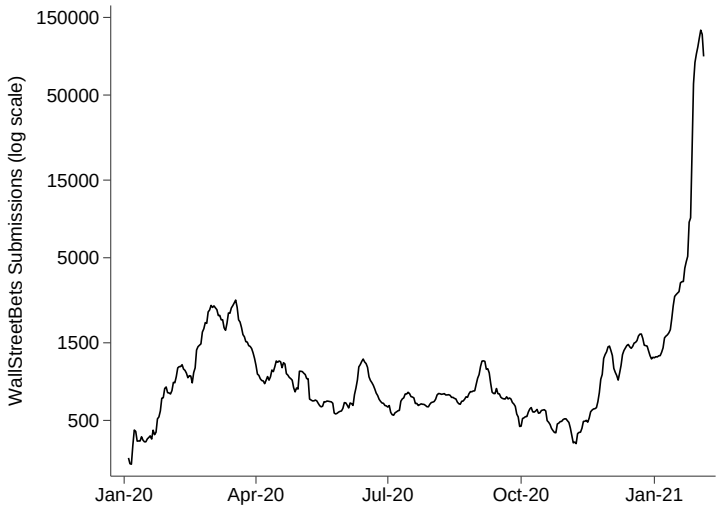


Figure J.3: Seven-day moving average of daily submissions to the WallstreetBets subreddit (January 1, 2020 – February 7, 2021). The vertical axis is on a logarithmic scale.

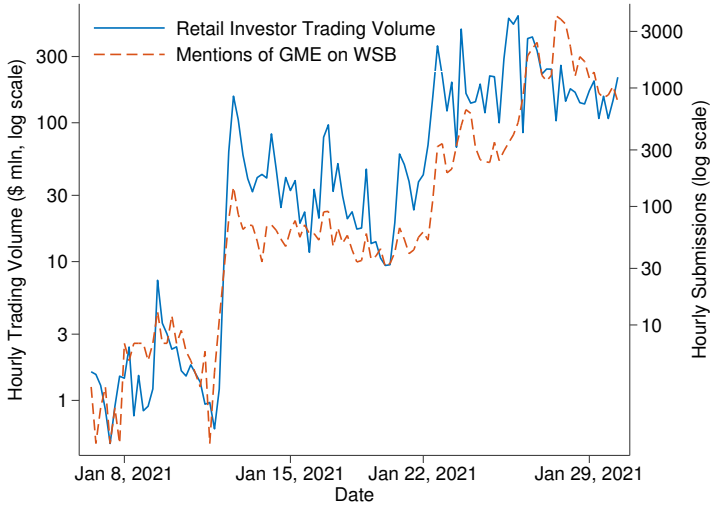


Figure J.4: Retail trading volume in GME (January 7 – January 29, 2021). Hourly trading volume in GME, measured using the methodology of Boehmer et al. (2021), plotted together with hourly mentions of the GME ticker on the WallStreetBets subreddit. Both vertical axes are on logarithmic scales.

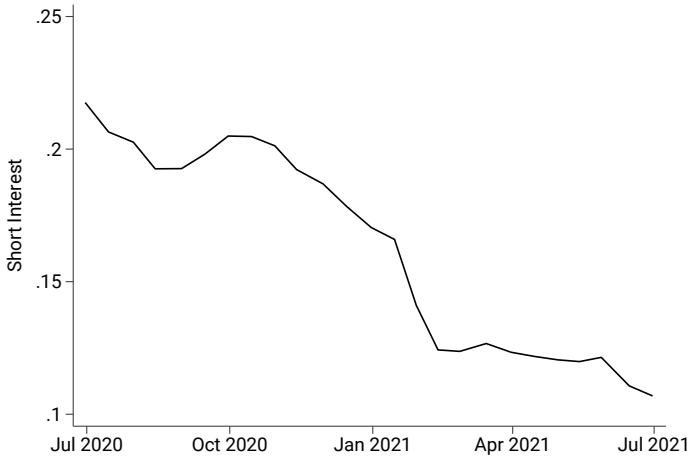


Figure J.5: Aggregate short interest (July 2020–June 2021). The figure plots value-weighted short interest for highly shorted stocks as of October 31, 2020. Highly shorted stocks are defined as the stocks in the top decile of the Russell 3000, ranked by short interest. The identities of these stocks is fixed and their short interest is plotted over the preceding four and subsequent eight months.

References - Online Appendix

Barber, Brad M., Xing Huang, Philippe Jorion, Terrance Odean, and Christopher Schwarz. 2024. "A (Sub)penny for Your Thoughts: Tracking Retail Investor Activity in TAQ." *The Journal of Finance.*

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. *The Pushshift Reddit Dataset.*

Boehmer, Ekkehart, Charles M. Jones, Xiaoyan Zhang, and Xinran Zhang. 2021. "Tracking Retail Investor Activity." *The Journal of Finance* 76 (5): 2249–2305.

Karatzas, Ioannis, and Steven Shreve. 2012. *Brownian motion and stochastic calculus.* Vol. 113. Springer Science & Business Media.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331 (6014): 176–182.