Granularity in Asset Markets*

Sergei Glebkin, Semyon Malamud, and Alberto Teguia

January 20, 2026

Abstract

We develop a tractable model of inelastic markets with heterogeneous, strategic investors who internalize their price impact. Investor granularity generates endogenous "granular wedges"—such as divergences between size- and equal-weighted holdings—that govern equilibrium outcomes. The model overturns classical predictions: non-competitive markets can be *more* liquid than competitive ones, and prices depend on the cross-sectional distribution of holdings. Moreover, capital flows toward risk-averse investors can paradoxically reduce aggregate risk aversion, explaining why safe-asset prices may decline during flight-to-safety episodes.

JEL Classification: D21, G31, G32, G35, L11

Keywords: Market Liquidity, Funding Liquidity, Price Impact, Strategic Trading, Wealth Effects

*Sergei Glebkin (glebkin@insead.edu) is at INSEAD, Semyon Malamud (semyon.malamud@epfl.ch) is at EPFL, and Alberto Teguia (alberto.mokakteguia@sauder.ubc.ca) is at UBC Sauder. For valuable feedback, we thank Viral Acharya, Ehsan Azarmsa, Frederico Belo, Bernard Dumas, Peter Kondor, Asaf Manela, Joel Peress, Anton Tsoy, Laura Veldkamp (discussant), Haoxiang Zhu, conference participants at the FTG meeting in UT Austin, the AFA meeting in Philadelphia, and seminar participants at INSEAD and the FTG Junior Research Group meeting.

1 Introduction

Modern financial markets are increasingly dominated by large institutional investors. A small number of firms—such as BlackRock, Vanguard, and Fidelity—now control a disproportionate share of total assets under management (AUM).[1] Consequently, financial markets have become *granular*: the actions of individual institutions can exert first-order effects on prices and liquidity.

This granularity shapes market outcomes through two primary channels. The first is *market power*: large investors have the ability to strategically influence prices (Rostek and Yoon, 2025). The second is *aggregation*: idiosyncratic shocks to large investors are not diversified away but instead drive aggregate market fluctuations (Gabaix, 2011; Gabaix and Koijen, 2021). Analyzing equilibria in this environment is challenging, as it requires a framework that simultaneously accounts for size (AUM) asymmetries and cross-sectional heterogeneity in other investor characteristics.

This paper develops a tractable model of inelastic markets populated by heterogeneous, strategic investors. The framework yields closed-form expressions for demand functions, equilibrium prices, and market liquidity. We leverage this tractability to address our central research question: how does investor granularity shape asset markets through the twin channels of aggregation and market power?

To isolate the effects of granularity, we begin with a single-period economy where size—interpreted as Assets Under Management (AUM)—is the sole source of heterogeneity among investors. All agents trade multiple risky assets and share identical Epstein-Zin preferences with a unit elasticity of intertemporal substitution (EIS); this specification preserves tractability while capturing the wealth effects central to our analysis. Trading occurs via a uniform-price double auction with symmetric information, where price impact arises solely from inventory risk as investors submit demand schedules to absorb a price-inelastic aggregate supply. We derive equilibrium quantities as functions of the realization of this supply, later endogenizing it for our welfare analysis as the hedging demand arising from investors' initial endowments.

We start with the competitive benchmark. Here, classic aggregation results imply that the distribution of AUM has no effect on returns, volatility, or liquidity, necessitating a non-competitive framework to address our research question.

Moving to the strategic equilibrium, we first analyze trading behavior. Analysis of the

[1] Since 1980, the share of U.S. equities held by the ten largest institutions has more than quadrupled, reaching 26.5% in 2016 (Ben-David, Franzoni, Moussawi, and Sedunov, 2021). Vanguard and Fidelity together represented roughly 30% of the U.S. mutual fund industry's market share in 2018 (Tjornehoj, 2018).

non-competitive equilibrium reveals that large investors do not merely scale up the behavior of small ones. While they trade larger volumes and provide more liquidity, they also face disproportionately higher price impact—a finding consistent with Koijen and Yogo (2019). Consequently, their turnover shares are *smaller* than their wealth shares, reversing the pattern of the competitive model. This matches the empirical patterns documented by Koijen and Yogo (2019) and arises endogenously in our model: higher price impact forces large investors to deviate from the competitive benchmark—where turnover shares align with wealth shares—more significantly than their smaller counterparts.

These distortions aggregate to shape asset prices. Relative to the competitive benchmark, the non-competitive equilibrium exhibits higher returns and volatility. Large investors exercise market power to tilt returns in their favor; this "scaling up" of returns naturally inflates volatility. Accordingly, greater AUM concentration amplifies this mechanism, further elevating both returns and volatility. These results match the empirical evidence on concentration and returns (Massa, Schumacher, and Wang, 2021) and volatility (Ben-David et al., 2021).

While the results on returns and volatility align with standard intuition, the implications for market liquidity are more surprising. We find that liquidity is *higher* in the non-competitive equilibrium and increases with AUM concentration. This counterintuitive result stems from the interaction between two levers that strategic investors use to exert market power: the *level*—shifting the price in their favor—and the *slope*—extracting additional discounts on incremental volume. Consider buyers, for concreteness. In concentrated markets, they exploit the first lever aggressively, pushing equilibrium prices toward their lower bound. Near this bound, little room remains to deploy the second lever; in effect, aggressive use of the level lever crowds out the slope lever.[2]

We extend the model to incorporate small initial holdings and heterogeneity in risk aversion, utilizing perturbation methods for the analysis.[3] We focus on how the joint distribution of preferences and holdings shapes aggregate outcomes in the presence of fund flows, modeled as exogenous reallocations of cash across funds.

In the competitive benchmark, outcomes follow standard intuition: (i) prices depend only on the aggregate supply of assets, not on their initial distribution across investors; (ii) flows between investors with identical risk aversion do not affect prices; and (iii) flows from less to more risk-averse investors increase the market's effective risk aversion.

Non-competitive behavior overturns these predictions. We show that: (a) The cross-

[2]This contrasts with traditional linear settings, where the two levers do not interact: the slope of aggregate demand is constant and independent of the price level.

[3]See Kogan and Uppal (2001) for similar techniques.

sectional distribution of holdings is pivotal. Prices and aggregate demand depend on a *granular holdings wedge*—the difference between size-weighted and equal-weighted holdings—consistent with empirical evidence. (b) This wedge predicts the direction of aggregate portfolio rebalancing. (c) Flows from less to more risk-averse investors can paradoxically *reduce* aggregate risk aversion, reversing the standard representative-agent logic.

Several implications follow. First, the model provides equilibrium microfoundations for the granular instrumental variables (GIV) methodology (Gabaix and Koijen, 2024, 2021). In the competitive benchmark, the hedging demands of investors of different sizes have the same sensitivity to their initial holdings. Consequently, aggregate demand depends solely on aggregate holdings. A redistribution of shocks that leaves aggregate holdings unaffected alters the GIV—which is constructed from the cross-sectional distribution of these shocks—but leaves aggregate demand invariant. Because the instrument varies while the endogenous variable (aggregate demand) does not shift, the relevance condition fails. In the non-competitive equilibrium, however, this neutrality breaks down. Because strategic investors hedge holding shocks with intensity related to their size, the size-weighted distribution of holdings drives shifts in aggregate demand. This restores the necessary link between the instrument and the endogenous variable, implying that market power is a prerequisite for GIV relevance. Intuitively, the strength of the instrument depends on the severity of the market power friction, providing a criterion for selecting markets where the GIV methodology is most applicable.

Second, our model implies a *predictive* relationship between the granular holdings wedge and aggregate portfolio rebalancing. Because large investors face higher price impact, their holding shocks receive a smaller weight in aggregate demand compared to those of smaller investors. This discrepancy creates a positive predictive relationship between the current granular holdings wedge and the direction of future aggregate portfolio rebalancing. This prediction is directly testable using standard holdings data.

Finally, flows from less risk-averse to more risk-averse investors can paradoxically depress safe-asset prices. In our model, this occurs because inflows enlarge the AUM of risk-averse funds, directly amplifying their price impact. To mitigate this impact, these funds must shade their demand more aggressively, effectively limiting their ability to express strong demand for safe assets. Funds specializing in safe assets serve as natural empirical proxies for these highly risk-averse agents. While flight-to-safety episodes typically raise safe-asset prices and generate convenience yields (Krishnamurthy and Vissing-Jorgensen, 2012, 2013), recent evidence documents episodes where safe-asset prices fall, creating an "inconvenience yield" (He, Nagel, and Song, 2022). Our model provides a new mechanism for this phenomenon, rooted in size-induced equilibrium inelasticity.

2 Literature Review

Our paper contributes to the extensive literature on strategic trading and price impact. In our model, information is symmetric, and price impact arises due to traders' limited risk-bearing capacity. We model trade using the classic double auction protocol, where traders submit price-contingent demand schedules. See, for example, Wilson (1979), Klemperer and Meyer (1989), Kyle (1989), Vayanos and Vila (1999), Vives (2011), Rostek and Weretka (2012), Kyle, Obizhaeva, and Wang (2017), Ausubel, Cramton, Pycia, Rostek, and Weretka (2014), Bergemann, Heumann, and Morris (2015), and Du and Zhu (2017) for the single-asset case, as well as Rostek and Weretka (2015) and Malamud and Rostek (2017) for the multi-asset case. Our key contribution to this literature is to provide a tractable framework that allows for wealth effects.

Our work is related to Kacperczyk, Nosal, and Sundaresan (2025), who also study strategic traders internalizing price impact in a multi-asset economy. The focus and mechanisms, however, differ. Kacperczyk et al. (2025) emphasize endogenous information acquisition and cross-asset learning, whereas we abstract from information frictions to isolate the effects of investor granularity and wealth-induced demand inelasticity. This distinction yields two advantages. First, we obtain closed-form equilibrium characterizations. Second, we demonstrate explicitly how the cross-sectional distribution of holdings and flows enters prices, liquidity, and aggregate risk aversion through a small set of granular wedges. Our model thus provides a transparent, microfounded mapping from concentration and flows to asset prices and liquidity—a channel that complements the information-based mechanism studied in Kacperczyk et al. (2025).

The papers discussed above typically rely on the standard CARA-Normal assumption to derive linear equilibria, where the slopes of demand schedules remain independent of price levels, and equilibrium price impact (given by the inverse slope of residual supply) is constant, independent of trade size. The linearity of equilibrium critically depends on the CARA-Normal assumption, which ensures that the marginal value of asset holdings is linear in holdings size, thereby guaranteeing the existence of linear equilibria.[4]

[4]The only exception is the two-agent case, where linear equilibria fail to exist, but as shown by Du and Zhu (2017), non-linear equilibria often arise. There is also a vast literature on competitive noisy rational expectations equilibria (REE) that extends beyond the CARA-Normal framework while assuming a continuum of non-strategic traders. For instance, some papers relax the assumption of normal payoff distributions while maintaining the CARA assumption or assuming risk neutrality—see Gennotte and Leland (1990), Ausubel (1990a), Ausubel (1990b), Barlevy and Veronesi (2003), Bagnoli, Viswanathan, and Holden (2001), Yuan (2005), Breon-Drish (2015), Pálvölgyi and Venter (2015), and Chabakauri, Yuan, and Zachariadis (2017). These studies assume CARA utilities and do not incorporate wealth effects. Glebkin, Gondhi, and Kuong (2021), however, introduce wealth effects in a CARA framework by considering margin constraints whose tightness depends on wealth levels. Meanwhile, Peress (2003), Malamud (2015), and Avdis and Glebkin (2023) analyze competitive

We offer a tractable alternative to the CARA-Normal framework. Whereas CARA-Normal models achieve tractability through the linearity of equilibria—resulting in constant price impacts—we achieve tractability by generating homogeneous equilibrium demands. This approach enables the study of wealth effects while accommodating general wealth distributions, all while preserving analytical tractability.

The double auction model in our paper enables the study of strategic liquidity provision while accounting for wealth effects. For the first time in the literature, we solve a fully micro-founded model that explicitly links market liquidity (price impact) with funding liquidity (the capital of strategic traders). Our model provides a fresh perspective on the classical results of Brunnermeier and Pedersen (2009), revealing subtle and unexpected interactions between the two forms of liquidity. In particular, we show that a higher concentration of funding liquidity can *improve* market liquidity.

There is a growing literature highlighting the significance of institutional investors in modern financial markets. Allen (2001) argue that financial crises are associated with liquidity shortages and that liquidity's effect on asset prices should be endogenous. Basak and Pavlova (2013) examine how institutional investors' trading impacts asset prices when their performance is measured relative to an index, leading to excess correlation among index stocks, heightened index stock volatility, and increased aggregate volatility. Brunnermeier and Pedersen (2009) show that institutional investors' aggregate capital (funding liquidity) influences risk premiums; see also Adrian, Etula, and Muir (2014) and He, Kelly, and Manela (2017). Micro-level evidence on individual institutional trades (Çötelioğlu, Franzoni, and Plazzi (2021), Ben-David et al. (2021)) suggests that aggregate measures overlook key market dynamics, and that investor granularity and strategic behavior—specifically, their internalization of price impact—are crucial for understanding the interplay between market and funding liquidity. We believe our model provides a tractable framework for analyzing this link and deepening our understanding of the precise role of institutional investor granularity in asset pricing.

Our paper is part of the broad literature on the effects of illiquidity in financial markets. Many papers in this literature take market frictions as exogenous, such as constant or random trading costs, portfolio constraints, and/or assets that cannot be traded (see Constantinides (1986), Longstaff (2009), Amihud and Mendelson (1986), Acharya and Pedersen (2005), Duffie, Gârleanu, and Pedersen (2005)). In our model, the only friction is the fact that there is a finite number of large traders who behave strategically. A trader is large simply because he owns a non-negligible fraction of the aggregate wealth. Wealth effects endogenously generate (1) portfolio constraints (due to nonnegativity of wealth), (2) illiquidity due to endogenous price

models with asymmetric information and non-CARA preferences.

impact, and (3) systemic liquidity that is priced in the cross-section of asset returns.[5]

The price impact of institutional trades has been extensively documented in the literature. See, for example, Chan and Lakonishok (1995), Griffin, Harris, and Topaloglu (2003), Chiyachantana, Jain, Jiang, and Wood (2004), Almgren, Thum, Hauptmann, and Li (2005), Coval and Stafford (2007), and Ben-David et al. (2021). Notably, Chung and Huh (2016) find that price impact is a priced factor and has a stronger effect on returns than adverse selection. Focusing on the non-informational component of price impact, our model provides a general framework for analyzing the relationship between the distribution of funding liquidity and market liquidity.

While our approach relies on supply function equilibria, as in Wilson (1979), Klemperer and Meyer (1989), and Kyle (1989), a related strand of literature models imperfect competition among traders in a *Cournot* fashion, where large traders are restricted to submitting market orders. See Gabszewicz and Vial (1972), Vives (1988), and, more recently, Neuhann, Sefidgaran, and Sockin (2021) and Neuhann and Sockin (2024). Through the lens of our model, this approach primarily captures imperfect competition among price-inelastic large traders, providing complementary insights to our focus on competition among price-elastic ones. Additionally, none of these studies examine the relationship between the distribution of wealth and market liquidity, which is central to our paper.

Finally, our paper relates to the literature on heterogeneity and asset prices (see Panageas (2020) for a review). These models are often difficult to analyze analytically, and several papers use perturbation techniques to study them. See, for example, Kogan and Uppal (2001), Kargar, Passadore, Silva, and Yang (2025), and Duarte, Kargar, Li, and Silva (2025).

3 The Model

There are two time periods $t \in \{0, 1\}$. A number $L > 1$ of strategic large investors trade assets with the rest of the market at $t = 0$. Large investors are indexed by $i \in \{1, 2, \ldots, L\}$. There are $N + 1$ assets, indexed by $k \in \{0, 1, 2, \cdots, N\}$. An asset k is a claim to a terminal dividend δ_k, where asset 0 is a risk-free asset with $\delta_0 = 1$. All payoffs are collected in the vector $\delta = (\delta_k)_k$.[6]

Each large investor i is endowed with $w_0^i = \alpha_i w_0$ units of the consumption good (henceforth, cash) at time 0, where w_0 represents the total wealth of all large investors, and α_i denotes

[5] Acharya and Bisin (2014) endogenize default risk using counterparty risk when positions are opaque. While there is no counterparty risk in our model, an agent's ability to borrow from other agents is effectively limited by the amount of liquid wealth that he can post as collateral.

[6] All vectors are assumed to be column vectors unless stated otherwise.

large investor i's share of the total wealth. By definition, the shares sum to 1, i.e., $\sum_i \alpha_i = 1$.

Large investors consume both at time 0 and time 1 and maximize the Epstein-Zin (1989) preferences with the elasticity of intertemporal substitution (EIS) equal to one and the relative risk aversion parameter (RRA) equal to γ:

$$U_i(c_0^i, c_1^i) = \log(c_0^i) + \log E\left[\left(c_1^i\right)^{1-\gamma}\right]^{1/(1-\gamma)}$$

with $\gamma > 0$. Log utility is a special case, with $\gamma = 1$.

Throughout, the expectation operator $E[\cdot]$ refers to the expectation taken with respect to the distribution of δ. The utility function $U_i(c_0^i, c_1^i)$ is well-defined for any $c_0^i > 0$ and $c_1^i > 0$. If either c_0^i or c_1^i is non-positive, we define $U_i(\cdot)$ to be equal to negative infinity. If investor i trades q_k units of asset k at time 0 at price P_k, his consumption satisfies $c_0^i = \alpha_i w_0 - q^\top P$, $c_1^i = \delta^\top q$, where $q = (q_k)_k$ and $P = (P_k)_k$.

The rest of the market supplies $Q \in \mathbb{R}^{N+1}$ units of the assets inelastically. We assume that Q is independent of δ.[7] We impose the following technical restrictions on the distributions of δ and Q. This assumption ensures that traders' optimization problems and the equilibrium outcomes are well-defined.

Assumption 1. *For every $y \in \text{supp}\{Q\}$, $E\left[\left(\delta^\top y\right)^{-\gamma} |\delta_k|\right] < \infty$, $E\left[\left(\delta^\top y\right)^{-(\gamma+1)} \delta_k^2\right] < \infty$ for all k; and (i) $\delta^\top y > 0$, and (ii) $E\left[\left(\delta^\top y\right)^{-\gamma} \delta\right] < \infty$. We restrict the admissible portfolios q to be of the form $q = ty$, for some $t \in \mathbb{R}_+$ and $y \in \text{supp}\{Q\}$. That is, q belong to the cone generated by Q.*

Trading is organized as a uniform-price double auction: each large investor i, $i \in \{1, 2, \cdots, L\}$ submits a demand function $D_i(P) : \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$ specifying the number of units of the assets they want to buy as a function of the prices of all assets. All trades are executed at prices P^* that clear the market, i.e., at a vector P^* such that $\sum_i D_i(P^*) = Q$. We examine Nash equilibria in demand schedules, where all large investors trade strategically, rationally anticipating the impact of their demand schedules on the market-clearing price.

[7] Independence of δ and Q implies that traders supplying Q are uninformed. Uncertainty about Q is needed to rule out the multiplicity of equilibria (cf. Klemperer and Meyer (1989) and Vayanos (1999)). As in Klemperer and Meyer, our assumptions imply that equilibrium quantities will depend on the realization of Q but not its distribution.

4 Competitive benchmark

We start by characterizing the benchmark equilibrium where large investors take prices as given. Large investor i solves the following optimization problem:

$$\sup_{q} \left\{ \log(\alpha_i w_0 - q^\top P) + \log\left(E\left[\left(q^\top \delta\right)^{1-\gamma} \right]^{\frac{1}{1-\gamma}} \right) \right\}.$$

The first-order (necessary and sufficient) condition can be written as

$$\left(\alpha_i w_0 - q^\top P\right) \frac{E\left[\left(q^\top \delta\right)^{-\gamma} \delta \right]}{E\left[\left(q^\top \delta\right)^{1-\gamma} \right]} = P. \tag{1}$$

Pre-multiplying (1) by $q^\top$, we obtain $q^\top P = \alpha_i w_0/2$. Substituting this back into (1) yields a closed-form expression for large investor i's inverse demand $I_i(q)$, specifying the prices they bid for a quantity vector q:

$$I_i^c(q) = \frac{\alpha_i w_0}{2} \frac{E\left[\left(q^\top \delta\right)^{-\gamma} \delta \right]}{E\left[\left(q^\top \delta\right)^{1-\gamma} \right]}. \tag{2}$$

It is important to note that this relationship holds for any joint distribution of asset payoffs δ, provided the relevant moments exist. We do not assume that asset payoffs are independent or normally distributed.

We note several key properties of the competitive inverse demand:

1. $I_i^c(q)$ exhibits *scale symmetry*. This means that $I_i^c(q) = I^c(q/\alpha_i)$ for all i, where $I^c(q) = 0.5 w_0 \, E\left[\left(q^\top \delta\right)^{-\gamma} \delta \right] / E\left[\left(q^\top \delta\right)^{1-\gamma} \right]$. This property implies that all investors hold the same portfolio weights in equilibrium—portfolios differ only in scale, not in composition.

2. The inverse demands are *homogeneous* in q. This means that there exists a constant k such that for any scalar $t \neq 0$ and any q, $I_i^c(tq) = t^k I_i^c(q)$. In our case, $k = -1$. This property reflects the absence of wealth effects on portfolio composition: the investor's marginal rate of substitution between assets depends only on the relative portfolio weights, not on the portfolio's overall size.

3. The inverse demands are *monotone* (strictly decreasing) functions of q. This means that $(I_i^c(q) - I_i^c(\hat{q}))^\top (q - \hat{q}) < 0$ for all $\hat{q} \neq q$.[8] This is the standard property that investors

[8] We prove that $I_i^c(q)$ is monotone in Lemma 1 in the Appendix.

require lower prices to absorb larger quantities. Additionally, $I_i^c(q)$ is continuously differentiable with a non-degenerate Jacobian, ensuring that demand functions are well-defined.

The properties above are linked to some of our modeling choices. In particular, our use of Epstein–Zin preferences with unit elasticity of intertemporal substitution (EIS= 1) is driven by tractability: the homogeneity of the inverse demand (Property 2) is essential for analyzing non-competitive equilibria, as we show in the next section. In general two-period settings, homogeneity requires that expenditure $q^\top I(q)$ be independent of the price level.[9] Logarithmic utility satisfies this requirement, as expenditure is a constant fraction of wealth. Standard CRRA preferences with $\gamma \neq 1$ do not: price scaling affects optimal consumption through wealth effects, breaking homogeneity and rendering the analysis intractable. Epstein–Zin preferences with EIS= 1 generalize log utility and allow us to accommodate general risk aversion while preserving the tractability that homogeneity provides.

We now turn to characterizing equilibrium allocations, prices, returns, volatility, and liquidity. Property 3 implies that the demand function, $D_i(P)$, which is the inverse of $I_i(q)$, is well-defined. Additionally, Properties 1 and 2 imply that the demand also exhibits scale symmetry and can be written as $D_i(P) = \alpha_i D(P)$. Given that supply is Q, the equilibrium allocation for trader i is $q_i = \alpha_i Q$. The equilibrium price is determined by $P(Q) = I_i(q_i) = I_i(\alpha_i Q)$, yielding the expression

$$P^c(Q) = \frac{w_0}{2} \frac{E\left[\left(Q^\top \delta\right)^{-\gamma} \delta\right]}{E\left[\left(Q^\top \delta\right)^{1-\gamma}\right]} \tag{3}$$

for the competitive equilibrium price $P^c(Q)$. Everywhere in the sequel, we use the superscript c to indicate the equilibrium outcomes in the competitive benchmark. The expected return on asset k is given by

$$\mu_k^c \equiv \frac{E[\delta_k]}{P_k^c(Q)} = \frac{2E[\delta_k] E\left[\left(Q^\top \delta\right)^{1-\gamma}\right]}{w_0 E\left[\left(Q^\top \delta\right)^{-\gamma} \delta_k\right]}, \tag{4}$$

while the return volatility is given by

$$\sigma_k^c \equiv \mathrm{Var}[\delta_k / P_k^c(Q)]^{1/2} = \frac{2\mathrm{Var}[\delta_k]^{1/2} E\left[\left(Q^\top \delta\right)^{1-\gamma}\right]}{w_0 E\left[\left(Q^\top \delta\right)^{-\gamma} \delta_k\right]}. \tag{5}$$

[9] Consider a price-taking trader solving $\sup_q \left\{u_0\left(w_0 - q^\top P\right) + u_1(q)\right\}$. The inverse demand is $I(q) = \nabla_q u_1(q) / u_0'(w_0 - q^\top I(q))$. For $I(q)$ to be homogeneous, the denominator $u_0'(w_0 - q^\top I(q))$ must be homogeneous, implying that time-zero consumption $w_0 - q^\top I(q)$ must be homogeneous as well. Since the first term w_0 is constant, the second term $q^\top I(q)$ can only be homogeneous of degree zero.

Our measure of illiquidity is the sensitivity of equilibrium prices to supply shocks, a standard measure in the literature (see Vayanos and Wang (2012)). When there are multiple assets, the illiquidity is characterized by a matrix Λ whose (k,l)-th element measures the marginal effect of a supply shock in asset l on the price of asset k,

$$\Lambda_{kl} = -\frac{\partial P_k}{\partial Q_l}.$$

Differentiating (3), we obtain

$$\Lambda^c(Q) = -\nabla P^c(Q) = \frac{w_0}{2}\gamma \frac{E\left[\left(Q^\top \delta\right)^{-\gamma-1} \delta\delta^\top\right]}{E\left[\left(Q^\top \delta\right)^{1-\gamma}\right]} + \frac{2}{w_0}(1-\gamma)P^c(Q)P^c(Q)^\top. \tag{6}$$

We summarize the properties of the unique competitive equilibrium in the following proposition.

Proposition 1. *The equilibrium prices, expected returns, return volatility, and illiquidity in the competitive case are given by equations (3), (4), (5) and (6). These quantities are invariant to changes in the wealth distribution* $\{\alpha_i,\ i \in \{1, 2, \cdots, L\}\}$.

In our model, agents have proportional endowments and homothetic preferences. Classic aggregation results (see, e.g., Varian (1992)) imply that the economy features a representative agent and, hence, wealth distribution does not affect equilibrium prices. As we show in the next section, this invariance breaks down when large investors act strategically.

Before we turn to the non-competitive equilibrium, we introduce the consumption-numeraire measure that allows to write equilibrium objects more compactly.

4.1 Prices, Liquidity, and the Consumption-Numeraire Measure

To express equilibrium objects more compactly, we introduce the consumption-numeraire measure. Define aggregate time-1 consumption as

$$C_{\text{agg}} \equiv \delta^\top Q,$$

and let $\hat{\delta} \equiv \delta / C_{\text{agg}}$ denote the vector of payoff shares—asset payoffs normalized by aggregate consumption. By construction, $\hat{\delta}^\top Q = 1$.

We define the consumption-numeraire measure P^* by the Radon–Nikodym derivative

$$\frac{dP^*}{dP} = \frac{C_{\text{agg}}^{1-\gamma}}{E\left[C_{\text{agg}}^{1-\gamma}\right]} = \frac{(\delta^\top Q)^{1-\gamma}}{E\left[(\delta^\top Q)^{1-\gamma}\right]}. \tag{7}$$

This measure tilts the physical probability P toward states in which aggregate consumption is high when $\gamma < 1$ (low risk aversion) and toward states in which aggregate consumption is low when $\gamma > 1$ (high risk aversion). The measure P^* is well-defined and equivalent to P because $C_{\text{agg}} > 0$ almost surely.

Using the consumption-numeraire measure, the competitive equilibrium price (3) can be written as

$$P^c(Q) = \frac{w_0}{2}\, E^*[\hat{\delta}], \tag{8}$$

where $E^*[\cdot]$ denotes expectation under P^*. The competitive price vector is thus proportional to the expected payoff shares under the consumption-numeraire measure.

Similarly, the illiquidity matrix (6) admits a compact representation:

$$\Lambda^c(Q) = \frac{w_0}{2}\left(\gamma\, \text{Var}^*[\hat{\delta}] + E^*[\hat{\delta}]\, E^*[\hat{\delta}]^\top\right), \tag{9}$$

where $\text{Var}^*[\hat{\delta}] \equiv E^*[\hat{\delta}\,\hat{\delta}^\top] - E^*[\hat{\delta}]\, E^*[\hat{\delta}]^\top$ is the variance-covariance matrix of payoff shares under P^*. This decomposition separates illiquidity into two components: a term proportional to the (risk-aversion-weighted) covariance of payoff shares and a rank-one term reflecting the outer product of expected payoff shares.

5 Non-competitive equilibrium

In this section, we derive an equilibrium where large investors act strategically. Following the classical approach introduced by Wilson (1979), we model their strategic interactions as competition in demand schedules, $D_i(P)$. We adopt a guess-and-verify approach. We hypothesize that the strategic demands exhibit the three key properties discussed in the previous section: scale symmetry, homogeneity, and monotonicity. From this point forward, we refer to the Nash equilibrium in demand schedules that satisfy these properties simply as the equilibrium. Therefore, our objective is to identify a Nash equilibrium where large investors' demands satisfy:

$$D_i(P) = \beta_i D(P) \quad \text{for all } i,$$

where $D(P)$ is a strictly decreasing, continuously differentiable, homogeneous function, and $\beta_i > 0$ are constants. Without loss of generality, we normalize $\sum_i \beta_i = 1$.

Definition 1. *A tuple $(D(P), \beta)$, consisting of a function $D(P)\colon \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$ and a vector $\beta \in \mathbb{R}^L_+$ with $\sum_i \beta_i = 1$, is an equilibrium if the following conditions hold:*

- *For any $i = 1, 2, \ldots, L$, if all other traders $j \neq i$ submit demands $D_j(P) = \beta_j D(P)$, then it is optimal for trader i to submit the demand $D_i(P) = \beta_i D(P)$.*
- *The function $D(P)$ is strictly decreasing, meaning $(D(P) - D(\hat{P}))^\top (P - \hat{P}) < 0$ for all $P \neq \hat{P}$. Additionally, $D(P)$ is continuously differentiable and has a non-degenerate Jacobian.*

We restrict our attention to scale-symmetric equilibria for three main reasons. First, this restriction is essential for tractability. In a general setting without scale symmetry, finding a Nash equilibrium in demand schedules requires solving a system of coupled, highly non-linear partial differential equations (PDEs). Such systems are generally intractable. By imposing scale symmetry, we reduce this complex problem to a manageable system of algebraic equations (as shown in Section 5.1).

Second, this restriction selects equilibria that preserve the intuition of the competitive benchmark. In the competitive case, we established that investors' demands are identical up to a scaling factor determined by their wealth. By looking for scale-symmetric strategic equilibria, we seek outcomes where this natural relationship between investor size and trading intensity is preserved, even when price impact is internalized.

Finally, this methodological choice parallels the standard approach in the CARA-Normal literature. Just as that literature typically restricts attention to linear equilibria to ensure solvability—despite the potential existence of non-linear equilibria—we focus on scale-symmetric equilibria to achieve a closed-form characterization while accommodating wealth effects.

Denoting the inverse of $D(P)$ by $I(q)$ (so that $I(D(P)) = P$), we can reformulate our Ansatz in terms of inverse demands as follows:

$$I_i(q) = I(q/\beta_i) \quad \text{for all } i.$$

By Definition 1, we seek an equilibrium where $I(q)$ is a monotone, continuously differentiable, and homogeneous function of q with a non-degenerate Jacobian. An important insight from the literature on supply function competition (see, e.g., Kyle (1989) and Klemperer and Meyer (1989)) is that the equilibrium can be reformulated in terms of inverse residual demand curves

faced by each trader. For a trader i, we denote the inverse residual demand by $P_i(q_i)$, which gives the vector of prices when agent i trades the quantity vector q_i. In this reformulation, the agent's objective is expressed as:

$$\sup_q \left\{ \log(\alpha_i w_0 - q^\top P_i(q)) + \log\left(E\left[\left(q^\top \delta\right)^{1-\gamma} \right]^{\frac{1}{1-\gamma}} \right) \right\}. \tag{10}$$

Importantly, the functions $P_i(q)$ are agent-specific, reflecting the heterogeneity assumed in the model.

5.1 Derivation of equilibrium

Recall the first-order condition (1) for a price-taking large investor i trading a quantity vector q_i at prices $P_i(q_i)$:

$$P_i(q_i) = \left(\alpha_i w_0 - q_i^\top P_i(q_i)\right) \frac{E\left[\left(q_i^\top \delta\right)^{-\gamma} \delta\right]}{E\left[\left(q_i^\top \delta\right)^{1-\gamma}\right]}.$$

A strategic trader, however, accounts for the fact that she can influence prices. The solution to the problem (10) can be formulated entirely in terms of the *price impact matrix*, $(\Lambda_i(q_i))_{kl} = \partial (P_i)_k / \partial (q_i)_l$, where the (k,l)-th element represents the sensitivity of the price of asset k to changes in large investor i's demand for asset l. In vector notation, $\Lambda_i(q_i) = \nabla P_i(q_i) \in \mathbb{R}^{(N+1)\times(N+1)}$. With this definition, the first-order condition for the strategic problem (10) becomes:

$$P_i(q_i) + \Lambda_i(q_i) q_i = \left(\alpha_i w_0 - q_i^\top P_i(q_i)\right) \frac{E\left[\left(q_i^\top \delta\right)^{-\gamma} \delta\right]}{E\left[\left(q_i^\top \delta\right)^{1-\gamma}\right]}. \tag{11}$$

To derive the equilibrium price impact, we determine the residual demand curve faced by investor i. Suppose large investor i modifies her quantity to q_i while the aggregate supply is Q. The remaining quantity, $Q - q_i$, must be absorbed by the other traders $j \neq i$. Here, we explicitly utilize the *scale symmetry* of the equilibrium. Since every other trader j submits an inverse demand of the form $I_j(q) = I(q/\beta_j)$, their demand schedules are identical up to the scalar β_j. Consequently, market clearing implies that the residual quantity is allocated across traders $j \neq i$ in proportion to their size β_j. Specifically, trader j's allocation is:

$$q_j = \frac{\beta_j}{\sum_{k \neq i} \beta_k} (Q - q_i) = \frac{\beta_j}{1 - \beta_i} (Q - q_i).$$

The market-clearing price is determined by the inverse demand of any trader j. Substituting q_j into $I_j(\cdot)$, we obtain the residual inverse demand $P_i(q_i)$ faced by trader i:

$$P_i(q_i) = I_j(q_j) = I\left(\frac{q_j}{\beta_j}\right) = I\left(\frac{Q - q_i}{1 - \beta_i}\right).$$

This derivation highlights the role of scale symmetry: it allows us to aggregate the heterogeneous demands of all other traders into a single representative residual demand curve scaled by $1-\beta_i$.

In equilibrium, trader i chooses $q_i = \beta_i Q$. Differentiating $P_i(q_i)$ with respect to q_i and evaluating at the equilibrium quantity yields the price impact matrix:

$$\Lambda_i(q_i) = \nabla P_i(q_i) = \frac{-1}{1-\beta_i} \nabla I\left(\frac{Q - q_i}{1 - \beta_i}\right)\Bigg|_{q_i = \beta_i Q} = \frac{-1}{1-\beta_i} \nabla I(Q). \tag{12}$$

We now substitute the expressions for $q_i = \beta_i Q$, the scale-symmetric inverse demand $I_i(q_i) = I(Q)$, and the derived price impact Λ_i into the strategic first-order condition (11). This yields the following system of partial differential equations (PDEs) governing the function $I(Q)$:

$$I(Q) - \frac{\beta_i}{1-\beta_i} \nabla I(Q) Q = \left(\alpha_i w_0 - \beta_i Q^\top I(Q)\right) \frac{E\left[\left(Q^\top \delta\right)^{-\gamma} \delta\right]}{\beta_i E\left[\left(Q^\top \delta\right)^{1-\gamma}\right]}. \tag{13}$$

The *scale symmetry* assumption has thus reduced a system of coupled equations for L unknown functions $I_i(q)$ into a condition for a single unknown function $I(Q)$.

The final simplification arises from *homogeneity*. If $I(Q)$ is a homogeneous function of degree k, Euler's homogeneous function theorem implies $\nabla I(Q) Q = k I(Q)$.[10] Substituting this identity into (13) reduces the system of PDEs to a system of linear algebraic equations:

$$\left(1 - k\frac{\beta_i}{1-\beta_i}\right) I(Q) = \left(\alpha_i w_0 - \beta_i Q^\top I(Q)\right) \frac{E\left[\left(Q^\top \delta\right)^{-\gamma} \delta\right]}{\beta_i E\left[\left(Q^\top \delta\right)^{1-\gamma}\right]}. \tag{14}$$

This system is solved using the same approach as in the competitive case: (1) premultiply (14) by $Q^\top$ to solve for the scalar expenditure $Q^\top I(Q)$; (2) substitute this scalar back into (14) to obtain the explicit form of $I(Q)$. We complete the derivation in the appendix, and the non-competitive equilibrium is summarized in the theorem below.

Theorem 1. *There exists a unique scale-symmetric equilibrium with a homogeneous $I(Q)$. The*

[10] To see this, differentiate the definition $I(tQ) = t^k I(Q)$ with respect to t, yielding $\nabla I(tQ) Q = k t^{k-1} I(Q)$, and set $t = 1$.

inverse demands are given by $I_i(q) = I(q/\beta_i)$. *The function* $I(q)$ *is given by*

$$I(q) = \frac{w_0}{2\phi} \frac{E\left[\left(\delta^\top q\right)^{-\gamma} \delta\right]}{E\left[\left(\delta^\top q\right)^{1-\gamma}\right]}. \tag{15}$$

The scaling constants are given by

$$\beta_i = \alpha_i \phi + 1 - \sqrt{(\alpha_i \phi)^2 + 1}.$$

The constant ϕ *is the unique positive solution to*

$$\sum_{i=1}^{L} \left(\alpha_i \phi + 1 - \sqrt{(\alpha_i \phi)^2 + 1}\right) = 1. \tag{16}$$

We note several properties of our equilibrium. First, unlike in traditional linear models, our equilibrium exists even when $L = 2$. The non-existence of linear equilibrium with $L = 2$ limits the applicability of the uniform-price double auction framework in modeling financial networks, where some network regions may naturally consist of only two players.[11] Given these challenges, we believe extending our approach to financial networks could provide valuable insights.

Second, we note the tractability of our model. A key challenge in solving models with market power is the price impact term $\Lambda_i(q_i)q_i$ in (11). Since price impact depends on the slopes of investors' demand functions in equilibrium, its presence in the first-order conditions transforms the problem into a system of partial differential equations (PDEs), as the FOCs relate inverse demands to the derivatives of other traders' inverse demands. In general, this system of PDEs is difficult to solve.[12]

Most of the literature circumvents this complexity by assuming that price impact is constant, as in the CARA-Normal framework, where the system of FOCs reduces to a system of algebraic equations. We propose an alternative that retains the tractability of the CARA-

[11] Malamud and Rostek (2017) and Babus and Kondor (2018) are two prominent examples of applying the uniform-price double auction to networks. The first paper effectively considers only networks with $L > 2$, while the second assumes that nodes with $L = 2$ (i.e., those with two dealers in their model) also have a mass of price-taking customers. However, not all real-world networks fit these restrictions. Du and Zhu (2017) demonstrate the existence of *non-linear* equilibria in a model with linear marginal utility and $L = 2$. While this non-linear equilibrium exists, it is significantly less tractable than the linear case, which has hindered its application to network models.

[12] In some special cases, this system of PDEs can be reduced to a single ordinary differential equation (ODE) that is more tractable. See Glebkin, Malamud, and Teguia (2023a) and Glebkin, Malamud, and Teguia (2023b).

Normal framework while allowing for wealth effects. Specifically, we focus on settings where demands are homogeneous. By Euler's homogeneous function theorem, the price impact term $\Lambda_i(q_i)q_i$ is then proportional to the inverse demand $I_i(q_i)$ itself, which again converts the system of FOCs into a system of algebraic equations, making the model solvable while capturing wealth effects.[13]

Third, our equilibrium permits a transparent analysis of how risk aversion affects demands. Figure 1 illustrates these patterns for a two-asset economy: a risk-free asset ($\delta_0 = 1$) and a risky asset with log-normal payoff. The left panel plots level curves of the risk-free inverse demand $I_0(q_0, q_1)$; the right panel plots level curves of the risky inverse demand $I_1(q_0, q_1)$. Dashed curves correspond to low risk aversion ($\gamma = 0.1$), while solid curves correspond to high risk aversion ($\gamma = 2$). The contour labels display the price levels.

The contours reveal how risk aversion shapes substitution patterns between assets. The dashed contours ($\gamma = 0.1$) are nearly linear, reflecting that a near-risk-neutral investor treats the two assets as close substitutes: the marginal rate of substitution between q_0 and q_1 is approximately constant. By contrast, the solid contours ($\gamma = 2$) display pronounced convexity. A more risk-averse investor values diversification: starting from a portfolio tilted toward the risky asset, she willingly sacrifices substantial risky holdings for a small increase in the safe asset, but this willingness diminishes as her portfolio becomes safer. Comparing the two panels, the curvature is more pronounced for the risky asset (right) than for the safe asset (left). Interestingly, even the safe asset exhibits convex substitution patterns, unlike in models with quasilinear preferences common in the strategic trading literature. The reason is that the marginal value of the safe asset depends on total time-1 consumption $\delta^\top q$: when consumption is low, marginal utility is high, making additional safe-asset holdings particularly valuable; as consumption grows, marginal utility declines, flattening the investor's willingness to pay—consistent with traditional consumption-based asset pricing logic.

5.2 The cross-section of investor demands

Recall that in a scale-symmetric equilibrium, individual demands $D_i(P)$ represent a fraction β_i of the aggregate demand $D(P)$, i.e., $D_i(P) = \beta_i D(P)$. Consequently, the slope of the individual demand $D_i(P)$, which corresponds to the amount of liquidity provided by large investor i, is also a fraction β_i of the slope of the aggregate demand (representing aggregate liquidity). Thus, the coefficient β_i has a dual interpretation: it represents large investor i's share of total turnover and

[13]This simplification relies crucially on the scale-symmetric structure of the equilibrium. Without scale symmetry, the system of PDEs remains intractable. Accordingly, we establish existence and uniqueness only within the class of scale-symmetric equilibria.

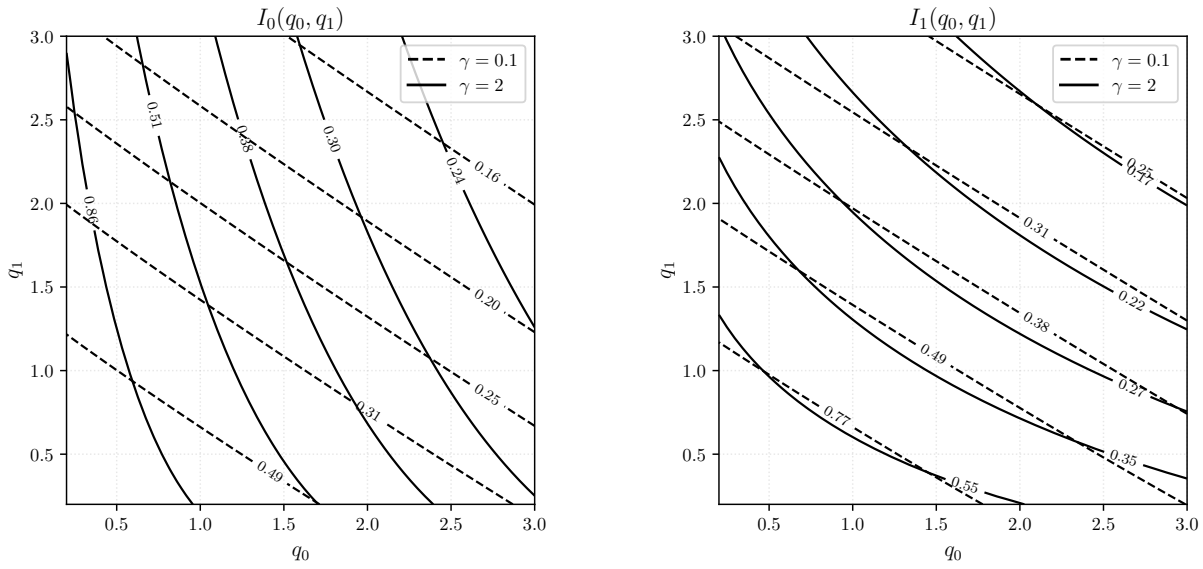

Figure 1: Level curves of the inverse demand functions $I_0(q_0, q_1)$ (left) and $I_1(q_0, q_1)$ (right) in a two-asset economy with a risk-free asset ($\delta_0 = 1$) and a risky asset ($\log \delta_1$ is standard normal). Dashed curves: $\gamma = 0.1$; solid curves: $\gamma = 2$. Contour labels indicate price levels. Higher risk aversion shifts contours outward, requiring smaller holdings for the same valuation.

its fraction of the total liquidity provided. In the next proposition, we examine the cross-section of the coefficients β_i and the individual price impacts Λ_i.[14]

Proposition 2. *Larger investors have a larger share of aggregate turnover and provide more liquidity, but have a higher price impact. Formally, for any i and j such that $\alpha_i > \alpha_j$, it holds that $\beta_i > \beta_j$, while $\Lambda_i > \Lambda_j$ (in the sense of positive semi-definite order).*[15] *Additionally, the share of the smallest (respectively, largest) investors in aggregate turnover exceeds (respectively, is less than) their share of aggregate wealth. Formally, ranking investors such that α_i increases with i, there exists a threshold i^* such that for any $i \geq i^*$ (respectively, $i < i^*$), $\beta_i < \alpha_i$ (respectively, $\beta_i > \alpha_i$).*

Our equilibrium exhibits several intuitive properties. First, larger investors engage in larger trades and provide more liquidity. This reflects their larger balance sheets, which generate greater trading needs and a higher capacity to supply liquidity.

Second, consistent with Koijen and Yogo (2019), larger investors experience a greater price impact. This occurs because their price impact is determined by the liquidity supplied by other investors, who, on average, provide less liquidity.

[14]Note the distinction between Λ, our measure of illiquidity, which is the slope of the aggregate inverse demand, and individual price impact Λ_i, the slope of the inverse residual demand faced by large investor i.

[15]That is, $\Lambda_i - \Lambda_j$ is positive definite.

Finally, unlike in the competitive model, where an investor's share of turnover equals their share of wealth, in the non-competitive setting, the largest investors have turnover shares that are smaller than their wealth shares, while the smallest investors exhibit the opposite pattern. This aligns with the empirical findings of Koijen and Yogo (2019), who show that the largest investors, managing one-third of total wealth, account for only 4% (less than one-third) of the cross-sectional variance of stock returns. In contrast, the smallest investors, who also manage one-third of total wealth, account for 47% (more than one-third) of the cross-sectional variance.

In our model, this pattern arises because larger investors face a higher price impact, which induces them to adjust their demands more aggressively away from the competitive benchmark, where turnover shares coincide with wealth shares, than smaller investors.

5.3 Contrasting with competitive benchmark

In this section, we compare aggregate quantities such as expected returns, return volatility, and liquidity across competitive and non-competitive equilibria. Quantities in the competitive equilibrium are denoted by the superscript c, while quantities in the non-competitive equilibrium are written without a superscript. The following proposition summarizes the comparison.

Proposition 3. *The non-competitive equilibrium is characterized by higher returns, higher return volatility, and lower illiquidity. Formally, for any k and $l \in \{1, 2, \ldots, N\}$, we have*

$$\frac{\mu_k}{\mu_k^c} = \frac{\sigma_k}{\sigma_k^c} = \frac{\Lambda_{kl}^c}{\Lambda_{kl}} = \phi > 1.$$

Here, the constant ϕ is determined by (16).

As highlighted in the proposition above, when large investors exercise market power, they tilt returns—both realized and expected—in their favor, effectively scaling them up. This amplification results in higher return volatility.[16]

By contrast, a more surprising implication concerns market liquidity: when traders exercise market power, the market becomes *more* liquid.[17] To build intuition, consider the case of a single risky asset ($N = 1$). As shown in Appendix IA.3, the logic below applies generally

[16]The finding that greater market power is associated with higher volatility is consistent with the empirical evidence in Ben-David et al. (2021). We revisit this connection in the next section, where we study the comparative statics of changes in wealth inequality, providing a more direct link to the empirical results in that paper.

[17]This implication is consistent with evidence in Pugachev (2024), who shows that hedge fund closures—which increase AUM concentration—lead to improved liquidity.

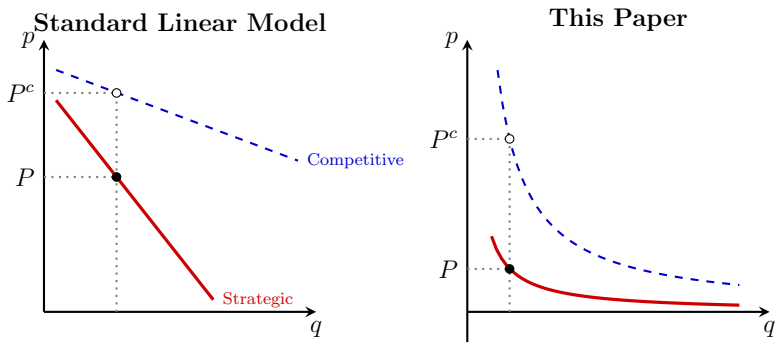

Figure 2: Demand Reduction and Price Impact: The Role of Bounded Demands.

to any preferences that produce bounded demands. Strategic traders "shade" their bids to manipulate prices, meaning their inverse demand curve always lies strictly below the competitive benchmark (demand reduction). However, as quantities become large, both strategic and competitive demands converge to the same limit (in our setting, zero, as implied by (2) and (15)).

This convergence forces a specific geometric relationship between the slopes, as illustrated in Figure 2. For the lower curve (strategic demand) to converge with the upper curve (competitive demand), the upper curve must descend more steeply to close the gap. While in the most general setting this condition need only hold over a subset of quantities, in our setting it holds globally for all quantities. Consequently, the strategic demand curve is *flatter* than the competitive one. In financial terms, a flatter inverse demand curve implies that prices are less sensitive to quantity changes—precisely the definition of higher liquidity.

Another important implication of Proposition 3 is that a single scalar wedge, ϕ, fully characterizes the departure of equilibrium outcomes from the competitive benchmark. Expected returns and return volatility are scaled up by ϕ, while illiquidity is scaled down by the same factor. This parsimonious structure greatly simplifies comparative statics: any change in the economic environment that affects ϕ proportionally shifts all aggregate quantities.

This wedge is tightly linked to concentration in the distribution of assets under management. The following proposition establishes that ϕ is linearly related to the Herfindahl–Hirschman Index (HHI), the standard measure of market concentration, when heterogeneity is not too large. Moreover, the linear relationship holds with high precision: the approximation is accurate up to third-order terms in the heterogeneity parameter.

Proposition 4 (HHI expansion for ϕ)**.** *Let market shares satisfy* $\alpha_i = \alpha + \epsilon a_i$*, with* $\sum_{i=1}^{L} a_i = 0$*,*

and define the Herfindahl–Hirschman Index as $\text{HHI} = \sum_{i=1}^{L} \alpha_i^2$. *Then,*

$$\phi = C_0 + C_1 \cdot \text{HHI} + O(\epsilon^3),$$

where

$$C_1 = \frac{L^3(2L-1)^2}{2(L-1)(2L^2-2L+1)^2},$$
$$C_0 = \frac{2L-1}{2(L-1)} - \frac{1}{L}C_1.$$

The proposition provides a direct quantitative link between market structure and the magnitude of strategic distortions: higher concentration, as measured by the HHI, implies a larger wedge ϕ and thus greater departures from competitive pricing.

5.4 Implications of changes in the distribution of wealth

In this section, we examine how changes in the wealth distribution, $\{\alpha_i\}_i$, influence expected returns, return volatility, and liquidity. Unlike the competitive benchmark—where variations in wealth distribution have no impact on these quantities—strategic interactions lead to a more nuanced relationship.

We focus on changes in wealth distribution that lead to an increase or decrease in inequality. When interpreting large investors as funds, an increase in inequality can result from two scenarios: (i) the merger of two funds or (ii) the flow of funds from a smaller fund to a larger one. In line with these scenarios, we define an increase and a decrease in inequality as follows.

Definition 2. *An* ***increase in inequality*** *corresponds to the following types of changes in the wealth distribution from* α *to* $\hat{\alpha}$*:*

1. *Flow of funds from a smaller large investor* i *to a larger large investor* j*:* $\alpha = \{\alpha_1, \ldots, \alpha_i, \ldots, \alpha_j, \ldots, \alpha_L\}$ *and* $\hat{\alpha} = \{\alpha_1, \ldots, \alpha_i - y, \ldots, \alpha_j + y, \ldots, \alpha_L\}$*, where* $y \leq \alpha_i \leq \alpha_j$*.*
2. *Merger of large investor* i *and large investor* j*:* $\alpha = \{\alpha_1, \ldots, \alpha_i, \ldots \alpha_{j-1}, \alpha_j, \alpha_{j+1}, \ldots, \alpha_L\}$ *and* $\hat{\alpha} = \{\alpha_1, \ldots, \alpha_i + \alpha_j, \ldots \alpha_{j-1}, \alpha_{j+1}, \ldots, \alpha_L\}$

A ***decrease in inequality*** *corresponds to changes in the wealth distribution described in 1 and 2, but in reverse, from* $\hat{\alpha}$ *to* α*. These changes represent a flow of funds from larger to smaller large investors or the split of a single large investor into two smaller entities.*

The following proposition summarizes how changes in wealth inequality influence expected returns, return volatility, and liquidity.

Proposition 5. *Consider a change in wealth distribution from α to $\hat{\alpha}$. Denote the equilibrium quantities corresponding to the distribution $\hat{\alpha}$ with a hat. We have:*

$$\frac{\hat{\mu}_k}{\mu_k} = \frac{\hat{\sigma}_k}{\sigma_k} = \frac{\Lambda_{kl}}{\hat{\Lambda}_{kl}} = \frac{\hat{\phi}}{\phi}.$$

When the change from α to $\hat{\alpha}$ corresponds to an increase (respectively, decrease) in inequality, we have $\hat{\phi} > \phi$ (respectively, $\hat{\phi} < \phi$). An increase in inequality leads to higher returns, higher return volatility, and lower illiquidity. Conversely, a decrease in inequality results in lower returns, lower return volatility, and higher illiquidity.

The results of the proposition above are intuitive: an increase in inequality amplifies the market power of large investors, producing an outcome qualitatively similar to the shift from a competitive equilibrium to a strategic one. The intuition outlined after Proposition 3 applies directly.

Our result, which shows that increased inequality leads to greater volatility, aligns with the evidence presented in Ben-David et al. (2021). They examine two scenarios involving changes in inequality: the merger of two funds (BlackRock and BGI) and an increase in the share of wealth managed by top institutions based on assets under management (AUM). In both cases, they find a positive relationship between inequality and volatility. These scenarios correspond directly to those described in Definition 2.

6 Heterogeneous Risk Aversion and Holding Shocks

The inelastic markets hypothesis (Gabaix and Koijen, 2021) and the literature on granular fluctuations (Gabaix, 2011) emphasize that the asset-pricing impact of investor attributes—such as preferences and endowments—depends critically on the distribution of investor size. Consequently, the relevant sufficient statistics are size-weighted aggregates rather than simple averages. Building on this insight, we extend our analysis to an environment with heterogeneity in risk aversion and holding shocks. We show that, consistent with the intuition in Gabaix and Koijen (2021), equilibrium outcomes are summarized by granular wedges: size-weighted cross-sectional moments of investor characteristics. These wedges encapsulate the effects of heterogeneity and yield novel empirical predictions regarding how market concentration shapes equilibria in inelastic markets.

Allowing for heterogeneity in risk aversion γ_i and initial holdings $q_{0,i}$ substantially complicates equilibrium characterization. In general, the problem reduces to a system of nonlinear partial differential equations governing investors' demand functions (see Appendix IA.3). This system generically admits multiple solutions and therefore gives rise to multiple equilibrium candidates. This makes the problem intractable both for analytical and even numerical methods (as numerical algorithms will have multiple equilibria to converge to). Glebkin et al. (2023b) address this issue in a model with homogeneous investors by imposing economically motivated boundary conditions that select a unique equilibrium. In our heterogeneous setting, however, it is not obvious how to impose analogous boundary conditions in a coherent and tractable way.

Instead, we adopt a perturbative approach and derive an approximate equilibrium around a tractable benchmark studied in the previous section. This method allows us to have a tractable characterization of the equilibrium while retaining the key economic forces introduced by heterogeneity and granularity.[18]

Specifically, we parameterize risk aversion and initial holdings as

$$\gamma_i = \gamma + \tilde{\gamma}_i \qquad \text{and} \qquad q_{0,i} = \tilde{q}_{0,i},$$

where $\tilde{\gamma}_i$ and $\tilde{q}_{0,i}$ denote small perturbations around the homogeneous benchmark. We let $\epsilon = \|(\tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{q}}_0)\|$ measure the magnitude of heterogeneity, where $\tilde{\boldsymbol{\gamma}} = (\tilde{\gamma}_i)_i$ and $\tilde{\boldsymbol{q}}_0 = (\tilde{q}_{0,i})_i$. We impose the normalization[19]

$$\sum_i \tilde{\gamma}_i = \sum_i \tilde{q}_{0,i} = 0. \tag{17}$$

We begin by introducing a key piece of notation that will be used throughout this section.

Definition 3 (Granular Wedge)**.** *The granular wedge for a variable X_i with weights w_i, $\sum_i w_i = 1$, is defined as:*

$$\Gamma_w[X] \equiv \underbrace{\sum_i w_i X_i}_{w\text{-weighted}} - \underbrace{\frac{1}{L}\sum_i X_i}_{\text{equal-weighted}} .$$

The granular wedge captures the difference between a size-weighted and an equal-weighted average (in our applications, the weights w_i are related to sizes). Under the nor-

[18] Perturbative and approximation-based methods are standard in economics; see, among others, Judd (1998), Uhlig (1999). In asset pricing, the most closely related approaches are that of Kogan and Uppal (2001), and more recently, Kogan and Mitra (2025), Kargar et al. (2025), and Duarte et al. (2025).

[19] For risk aversion, this normalization is without loss of generality. For initial holdings, we impose $\sum_i \tilde{q}_{0,i} = 0$ to isolate the effects of granular redistribution from standard aggregate endowment effects. Although it is straightforward to adjust the derivations to introduce a common component $\tilde{h} = \sum_i \tilde{q}_{0,i}$, we omit this aggregate shock as it yields no novel insights beyond standard effects.

malization $\sum_i X_i = 0$ (which holds for $\tilde{\gamma}_i$ and $\tilde{q}_{0,i}$ by (17)), the granular wedge simplifies to $\Gamma_w[X] = \sum_i w_i X_i$, the size-weighted sum. This operator naturally arises in granular markets: when investors differ in size, aggregate outcomes depend on how characteristics are distributed across sizes, not merely on their simple averages.

6.1 Competitive Equilibrium

In the competitive case, a standard implicit function theorem argument yields the following result.

Proposition 6 (Approximate Competitive Demand)**.** *The competitive demand of large investor i admits the expansion:*

$$D_i^c(p) = D_i^{*,c}(p) + \tilde{D}_i^c(p) + O(\epsilon^2),$$

where $D_i^{*,c}(p) = \alpha_i D^{*,c}(p,\gamma)$ is the demand in the unperturbed economy (with inverse demand given by (2)), and the first-order correction is

$$\tilde{D}_i^c(p) = -\tilde{q}_{0,i} + \frac{p^\top \tilde{q}_{0,i}}{w_0} D^{*,c}(p) + \alpha_i \tilde{\gamma}_i D_\gamma^{*,c}(p), \tag{18}$$

with $D_\gamma^{*,c}(p) \equiv \frac{\partial}{\partial \gamma} D^{*,c}(p,\gamma)$.

The first-order correction $\tilde{D}_i^c(p)$ consists of three economically intuitive terms: (i) a hedging component, $-\tilde{q}_{0,i}$, reflecting the investor's desire to offset her initial position; (ii) a wealth effect, $\frac{p^\top \tilde{q}_{0,i}}{w_0} D^{*,c}(p)$, capturing the additional demand arising from the value of initial holdings; and (iii) a risk-aversion effect, $\alpha_i \tilde{\gamma}_i D_\gamma^{*,c}(p)$, reflecting the tilt in demand due to the perturbation in risk aversion.

We next proceed to characterize the aggregate quantities, demand, and prices in equilibrium. For the sake of notational simplicity, we use notation $\Gamma^c[\tilde{\gamma}] = \Gamma_\alpha[\tilde{\gamma}]$ for competitive granular risk aversion wedge.

Proposition 7 (Aggregate Quantities in Competitive Equilibrium)**.** *The aggregate demand and equilibrium price in the competitive equilibrium admit the expansions:*

$$D^c(p) = D^{*,c}(p) + \tilde{D}^c(p) + O(\epsilon^2) \quad \textit{and} \quad p^c = p^{*,c} + \tilde{p}^c + O(\epsilon^2).$$

Here, $D^{*,c}(p)$ is the aggregate demand in the unperturbed economy, with the inverse $I^c(q)$ given

by $I^c(q) = 0.5 w_0 \, \mathbb{E}\left[\left(q^\top \delta\right)^{-\gamma} \delta\right] / \mathbb{E}\left[\left(q^\top \delta\right)^{1-\gamma}\right]$. *The aggregate demand perturbation is*

$$\tilde{D}^c(p) = \Gamma^c[\tilde{\gamma}] \, D^{*,c}_\gamma(p),$$

where $D^{*,c}_\gamma(p) \equiv \frac{\partial}{\partial \gamma} D^{*,c}(p, \gamma)$. *The baseline equilibrium price and its perturbation are:*

$$p^{*,c} = \frac{w_0}{2} E^*[\hat{\delta}], \tag{19}$$

$$\tilde{p}^c = -\frac{w_0}{2} \Gamma^c[\tilde{\gamma}] \, \mathrm{Cov}^*\left(\hat{\delta}, \ln C_{\mathrm{agg}}\right), \tag{20}$$

where E^* *and* Cov^* *denote expectation and covariance under the consumption-numeraire measure* P^*, $\hat{\delta} = \delta / C_{\mathrm{agg}}$ *is the vector of payoff shares, and* $C_{\mathrm{agg}} = \delta^\top Q$ *is aggregate consumption.*

Proposition 7 establishes that outcomes in the competitive benchmark follow standard intuition. First, prices depend only on the aggregate supply of assets Q, not on the initial distribution of holdings $\{\tilde{q}_{0,i}\}_i$ across investors—the purely redistributional endowment shocks do not affect the aggregate resource constraint faced by the representative investor. Second, flows between investors with identical risk aversion do not affect prices: since only the granular wedge $\Gamma^c[\tilde{\gamma}]$ enters the price perturbation, reallocations of wealth among investors sharing the same $\tilde{\gamma}_i$ leave $\Gamma^c[\tilde{\gamma}]$ unchanged. Third, flows from less to more risk-averse investors increase the market's effective risk aversion: transferring wealth from low-$\tilde{\gamma}$ to high-$\tilde{\gamma}$ investors raises $\Gamma^c[\tilde{\gamma}]$, which in turn decreases the price. As we will see in the next section, all of these predictions are overturned in the non-competitive equilibrium.

The covariance term $\mathrm{Cov}^*(\hat{\delta}, \ln C_{\mathrm{agg}})$ captures the exposure of each asset's payoff share to aggregate consumption risk. Importantly, the consumption-numeraire measure P^* already incorporates a complete adjustment for consumption risk, but only for an investor with risk aversion exactly equal to γ. Under heterogeneity, individual risk aversions are perturbed around γ, so the baseline adjustment is imperfect: investors who are more (less) risk-averse than γ require additional discounting (premium) for consumption risk exposure. The covariance term quantifies this residual adjustment.

Specifically, an asset with $\mathrm{Cov}^*(\hat{\delta}_k, \ln C_{\mathrm{agg}}) > 0$ delivers a larger fraction of aggregate consumption in good states (high C_{agg}) and a smaller fraction in bad states. When the effective risk aversion increases ($\Gamma^c[\tilde{\gamma}] > 0$), the marginal investor values an extra dollar in low-consumption states relatively more than P^* reflects. This tilts the stochastic discount factor further toward bad states, reducing the value of assets that pay relatively more in good times. Conversely, assets that provide insurance—delivering relatively more when aggregate consumption is low—experience price increases when the market becomes more risk-averse.

6.2 Perturbed Non-Competitive Equilibrium

We now turn to the non-competitive equilibrium, where strategic interactions fundamentally alter how heterogeneity affects prices. In contrast to the competitive case of Proposition 6, where the demand of agent i is independent of the actions of other agents, in the non-competitive case individual demands depend on other agents' behavior. In line with the perturbation logic, where demands are determined up to an $O(\epsilon^2)$ error, we assume that traders form their best responses to the demands of other traders known up to an $O(\epsilon^2)$ error.

Recall that the first-order condition determining the best response for agent i is given by:

$$p + \Lambda_i(p) D_i(p) = \left(\alpha_i w_0 - D_i(p)^\top p\right) \frac{E\left[\left((q_{0,i} + D_i(p))^\top \delta\right)^{-\gamma_i} \delta\right]}{E\left[\left((q_{0,i} + D_i(p))^\top \delta\right)^{1-\gamma_i}\right]}. \tag{21}$$

where $\Lambda_i(p)$ is the equilibrium price impact matrix of agent i that summarizes the behavior of other agents. It is given by the slope of the inverse residual demand:

$$\Lambda_i(p) = -\left(\nabla_p \sum_{j \neq i} D_j(p)\right)^{-1}. \tag{22}$$

Our approximate equilibrium definition requires that demands are mutual best responses up to an $O(\epsilon^2)$ error.

Definition 4 (Perturbed Non-Competitive Equilibrium)**.** *We say that demand schedules*

$$D_i(p) = D_i^*(p) + \tilde{D}_i(p) + O(\epsilon^2)$$

form a perturbed non-competitive equilibrium if $D_i(p)$ satisfy (21) and (22) up to an $O(\epsilon^2)$ error.

Using a guess-and-verify approach, we show the existence of a unique perturbed equilibrium, where demands have the structure (23) that follows the corresponding structure in the competitive equilibrium.

Theorem 2. *The first-order demand perturbation exhibits the following structure:*

$$\tilde{D}_i(p) = \eta_{0,i} \tilde{q}_{0,i} + \eta_{1,i}\, D_\gamma^*(p) + \eta_{2,i}\, \tilde{q}_{0,i}^\top p\, D^*(p). \tag{23}$$

Here, $D^*(p)$ denotes the baseline aggregate demand, and $D_\gamma^*(p)$ denotes its sensitivity to risk aversion:

$$D_\gamma^*(p) \equiv \frac{\partial}{\partial \gamma} D^*(p, \gamma).$$

The baseline demand satisfies $p = w_0/(2\phi)g(D^*(p), \gamma)$*, where*

$$g(x, \gamma) = \frac{\mathbb{E}\left[\left(x^\top \delta\right)^{-\gamma} \delta\right]}{\mathbb{E}\left[\left(x^\top \delta\right)^{1-\gamma}\right]}.$$

The coefficients η *are given by:*

$$\begin{aligned}
\eta_{0,i} &= -\frac{1}{1+\beta_i}, \\
\eta_{1,i} &= \beta_i(1-\beta_i)\tilde{\gamma}_i + \beta_i^2 \Gamma_{\omega^\gamma}[\tilde{\gamma}], \\
\eta_{2,i} &= \frac{\phi(1-\beta_i)}{w_0(1+\beta_i)},
\end{aligned}$$

where the granular risk aversion wedge uses weights $\omega_j^\gamma \equiv \frac{\beta_j(1-\beta_j)}{\sum_k \beta_k(1-\beta_k)}$ *:*

$$\Gamma_{\omega^\gamma}[\tilde{\gamma}] = \frac{\sum_j \beta_j(1-\beta_j)\tilde{\gamma}_j}{\sum_j \beta_j(1-\beta_j)}. \tag{24}$$

Moreover, the demand perturbation satisfies the orthogonality condition $p^\top D_\gamma^*(p) = 0$*.*

Theorem 2 characterizes the impact of holding shocks and heterogeneity on equilibrium demands. The first-order demand perturbation, $\tilde{D}_i(p)$, consists of three terms that we now discuss.

Hedging demand $\eta_{0,i}\tilde{q}_{0,i}$**.** Since $\eta_{0,i} = -1/(1+\beta_i)$ and the coefficients β_i are larger for bigger investors (i.e., those with larger α_i), larger traders hedge less aggressively than smaller ones. This reflects the classical bid-shading mechanism: agents with higher price impact cannot fully "undo" the holding shock and, as a result, these idiosyncratic shocks enter the aggregate demand with smaller weight.

Scaling of $D^*(p)$ **to reflect higher effective wealth.** The initial holdings $\tilde{q}_{0,i}$ represent an additional source of wealth. At price p, these holdings are equivalent to an increase in wealth (AUM) of $\tilde{q}_{0,i}^\top p$. Correspondingly, the demand scales up by $\eta_{2,i}(\tilde{q}_{0,i}^\top p)D^*(p)$, where $\eta_{2,i} > 0$.

Self-financing portfolio tilt $\eta_{1,i}\, D_\gamma^*(p)$ **to reflect heterogeneity.** The term $D_\gamma^*(p)$ captures the marginal change in baseline demand $D^*(p;\gamma)$ due to a shift in risk aversion. Notably, these demand tilts are self-financing, satisfying the zero-cost condition $p^\top D_\gamma^*(p) = 0$. Additionally,

this effect depends both on individual risk aversion $\tilde{\gamma}_i$ and on the granular risk aversion wedge $\Gamma_{\omega^\gamma}[\tilde{\gamma}]$ that we discuss in more detail in the next section.

We now summarize the aggregate implications of the perturbed equilibrium. Recall formula (19) for the competitive price $p^{*,c}$. We will also use the following two granular wedges

$$\begin{aligned}\Gamma_{\omega^\gamma}[\tilde{\gamma}] &= \sum_j \omega_j^\gamma \tilde{\gamma}_j \\ \Gamma_{\omega^q}[\tilde{q}_0] &= \sum_j \omega_j^q \tilde{q}_{0,j}\,,\end{aligned}$$

defined using the weights $\omega_j^q = \frac{\frac{1}{1+\beta_j}}{\sum_k (1+\beta_k)^{-1}}$ and $\omega_j^\gamma = \frac{\beta_j(1-\beta_j)}{\sum_k \beta_k(1-\beta_k)}$. As above, for simplicity, we use the notation $\Gamma[\tilde{\gamma}] = \Gamma_{\omega^\gamma}[\tilde{\gamma}]$ and $\Gamma[\tilde{q}_0] = \Gamma_{\omega^q}[\tilde{q}_0]$. We also define $S_q \equiv \sum_k (1+\beta_k)^{-1}$ and $\tilde{\phi} = \phi / S_q$.

Proposition 8 (Prices in Non-Competitive Equilibrium)**.** *Define the aggregate wealth effect* $\mathcal{W} \equiv \sum_i \eta_{2,i}(\tilde{q}_{0,i}^\top p^*)$. *The equilibrium price in the non-competitive equilibrium admits the expansion:*

$$p = \frac{1}{\phi} p^{*,c} + \tilde{p} + O(\epsilon^2) \tag{25}$$

with

$$\begin{aligned}\tilde{p} = &- \underbrace{\frac{w_0}{2\phi}\Gamma[\tilde{\gamma}]\,\mathrm{Cov}^*\left(\hat{\delta}, \ln C_{\mathrm{agg}}\right)}_{\textit{Granular consumption risk premium}} \\ &- \underbrace{\frac{w_0}{2\tilde{\phi}}\gamma \mathrm{Cov}^*\left(\hat{\delta}, \hat{\delta}^\top \Gamma[\tilde{q}_0]\right)}_{\textit{Granular hedging risk premium}} \\ &+ \underbrace{\bar{\mathcal{W}}\, p^*}_{\textit{Wealth effect}}\,,\end{aligned} \tag{26}$$

where $\bar{\mathcal{W}} \equiv \mathcal{W} - \frac{2\tilde{\phi}}{w_0}\Gamma[\tilde{q}_0]^\top p^*$ *captures the net wealth effect,* $\hat{\delta} = \delta / C_{\mathrm{agg}}$ *is the vector of payoff shares, and* $C_{\mathrm{agg}} = \delta^\top Q$ *is aggregate consumption.*

We now discuss the implications of Proposition 8.

6.2.1 Implications for Aggregate Portfolio Rebalancing

In the competitive benchmark, the distribution of holdings across investors does not affect equilibrium. As discussed above, standard aggregation results imply that only aggregate initial holdings—the resource constraint of the representative investor—matter. In contrast, the

aggregate demand in the non-competitive case admits the expansion

$$D(p) = D^*(p) + \tilde{D}(p) + O(\epsilon^2)\,,$$

with the aggregate demand perturbation $\tilde{D}(p)$ given by

$$\tilde{D}(p) = -S_q\,\Gamma[\tilde{q}_0] + \Gamma[\tilde{\gamma}]\,D^*_\gamma(p) + \mathcal{W}\,D^*(p), \tag{27}$$

As a result, the non-competitive equilibrium features a granular holdings wedge $\Gamma[\tilde{q}_0]$ that enters both aggregate demand (27) and equilibrium prices (26). This wedge arises because larger investors—those with higher β_i—hedge less aggressively due to bid shading, causing the cross-sectional distribution of holdings across investor sizes to matter for aggregate outcomes.

The granular holdings wedge represents a price-inelastic component of aggregate portfolio rebalancing. Since computing it requires only initial holdings and investor size, this component is predictable. To understand its structure, consider the case when size heterogeneity is moderate. Specifically, let $\alpha_i = \bar{\alpha} + \tilde{\alpha}_i$, where $\tilde{\alpha}_i$ is small. Lemma 3 in the appendix establishes that $\Gamma[\tilde{q}_0] \approx -\mathcal{K}\,\Gamma_\alpha[\tilde{q}_0]$, where $\mathcal{K} > 0$. Thus, the granular holdings wedge is negatively related to the size-weighted granular holdings wedge, the difference between size-weighted and equal-weighted holdings. This observation yields the following prediction.

Corollary 1. *When market share heterogeneity is small, the size-weighted granular holdings wedge* $\Gamma_\alpha[\tilde{q}_0]$ *positively predicts the aggregate change in holdings,* $\sum_i \tilde{D}_i$.

The advantage of replacing $\Gamma[\tilde{q}_0]$ with $\Gamma_\alpha[\tilde{q}_0]$ under small size heterogeneity is that the latter is directly computable from holdings data, making our prediction readily testable.

6.2.2 Implications for Granular Instrumental Variables

Gabaix and Koijen (2024) propose Granular Instrumental Variables (GIV) as instruments for endogenous demand changes. The GIV is constructed as the difference between size-weighted and equal-weighted price-inelastic demand shocks—shocks that are contemporaneous with the demand being instrumented. In our model, the price-inelastic demand shock of investor i is $\tilde{q}_{0,i}/(1+\beta_i)$, and the size-weighted shock is $\alpha_i\tilde{q}_{0,i}/(1+\beta_i)$. Accordingly, the GIV is given by:

$$\text{GIV} = \sum_i \frac{\alpha_i \tilde{q}_{0,i}}{1+\beta_i} - \frac{1}{L}\sum_i \frac{\tilde{q}_{0,i}}{1+\beta_i}.$$

Under moderate size heterogeneity, Lemma 4 in the appendix establishes that the GIV and the

granular holdings wedge are negatively related. This implies a positive relationship between the GIV and aggregate portfolio rebalancing, consistent with the first-stage regression results in Gabaix and Koijen (2021).

We analyze the relevance of the GIV as a comparative statics exercise with respect to the holding shocks, $\tilde{q}_{0,i}$. In the competitive benchmark, a redistribution of these shocks alters the GIV—which is constructed from the cross-sectional distribution of these shocks—but leaves aggregate demand invariant. Because the instrument varies while the endogenous variable (aggregate demand) does not shift, the relevance condition fails. In the non-competitive equilibrium, however, this neutrality breaks down. Because strategic investors hedge holding shocks with intensity inversely related to their size, the size-weighted distribution of $\tilde{q}_{0,i}$ drives shifts in aggregate demand. This restores the necessary link between the instrument and the endogenous variable, implying that market power is a prerequisite for GIV relevance.

6.2.3 Asset pricing implications

We now turn to the asset pricing implications of the model, juxtaposing our results against the competitive benchmark. Comparing the baseline price expressions (19) and (25) reveals that, unlike in the competitive case, flows between investors with identical risk aversion affect prices through their impact on ϕ. We have discussed these effects extensively in the previous sections.

The comparison of risk premium loadings reveals that granular markets aggregate risk aversion in a fundamentally different manner than competitive markets. In the competitive economy, the consumption risk premium is governed by the size-weighted risk aversion $\Gamma^c[\tilde{\gamma}]$. In granular markets, by contrast, it is governed by the granular risk aversion wedge $\Gamma[\tilde{\gamma}]$. This distinction has potentially far-reaching implications for how risk premia are formed and aggregated in heterogeneous economies.

An example: inconvenience yields of safe assets. To illustrate, consider a flight-to-safety episode in which capital flows from low-γ funds (e.g., equity-focused funds) to high-γ funds (e.g., funds investing in government or AAA-rated corporate bonds). In the competitive economy, such flows mechanically increase aggregate (representative-agent) risk aversion and therefore raise the consumption risk premium in (20). In a non-competitive equilibrium, however, the effect is ambiguous due to the opposing bid-shading force. As a fund grows larger, it shades its bids more aggressively, expressing its preferences less strongly in equilibrium prices. As a result, the consumption risk premium may remain unchanged or even decline following these flows.

We formalize this intuition with an example. Suppose there are two possible fund sizes, $\alpha_B > \alpha_S$, with $L_{i,B}$ ($L_{i,S}$) funds of size B (S) having risk aversion γ_i, $i = 1, 2$. Let $L_B = \sum_i L_{i,B}$ and $L_S = \sum_i L_{i,S}$. Then,

$$\Gamma[\tilde{\gamma}] \;=\; \frac{(L_{1,B}\beta_B(1-\beta_B) + L_{1,S}\beta_S(1-\beta_S))\tilde{\gamma}_1 + (L_{2,B}\beta_B(1-\beta_B) + L_{2,S}\beta_S(1-\beta_S))\tilde{\gamma}_2}{L_B\beta_B(1-\beta_B) + L_S\beta_S(1-\beta_S)}\,.$$

Each fund type is characterized by its size and risk aversion. We denote by $(1, B)$ a large fund with risk aversion γ_1, and analogously for other fund types.

Now consider a flow from a $(1, B)$ fund to a $(2, S)$ fund such that the $(2, S)$ fund becomes a $(2, B)$ fund. Since L_B and L_S do not change under such flows, neither do β_B and β_S. Nevertheless, the effect on $\Gamma[\tilde{\gamma}]$ is non-trivial:

$$\Gamma[\tilde{\gamma}]_{new} - \Gamma[\tilde{\gamma}] \;=\; (\beta_S(1-\beta_S) - \beta_B(1-\beta_B))\frac{\tilde{\gamma}_1 - \tilde{\gamma}_2}{L_B\beta_B(1-\beta_B) + L_S\beta_S(1-\beta_S)}\,.$$

Suppose $\tilde{\gamma}_1 < \tilde{\gamma}_2$, so that flight-to-safety flows increase the size-weighted risk aversion. However, if the size differential $\alpha_B - \alpha_S$ is sufficiently large, then $\beta_S(1-\beta_S) - \beta_B(1-\beta_B) > 0$, implying that the change in $\Gamma[\tilde{\gamma}]$ is negative. In other words, the effective (representative) risk aversion of the economy declines—in stark contrast to the competitive economy of Proposition 6.

This example demonstrates that in granular markets flows from less risk-averse to more risk-averse investors can paradoxically depress safe-asset prices. In our model, this occurs because inflows enlarge the AUM of risk-averse funds, directly amplifying their price impact. To mitigate this impact, these funds shade their demand more aggressively, effectively limiting their ability to express strong demand for safe assets. Funds specializing in safe assets serve as natural empirical proxies for these highly risk-averse agents. While flight-to-safety episodes typically raise safe-asset prices and generate convenience yields (Krishnamurthy and Vissing-Jorgensen, 2012, 2013), recent evidence documents episodes where safe-asset prices fall, creating an "inconvenience yield" (He et al., 2022). Our model provides a new mechanism for this phenomenon, rooted in non-competitive behavior of large investors.

Hedging risk premium. Juxtaposing (26) and (20) reveals that, unlike in the competitive economy, predictable, price-inelastic hedging demands $\Gamma[\tilde{q}_0]$ tend to be priced, as reflected in the term $-\frac{w_0}{2\bar{\phi}}\gamma\text{Cov}^*\left(\hat{\delta}, \hat{\delta}^\top\Gamma[\tilde{q}_0]\right)$. This implies that exploiting the predictability of aggregate rebalancing discussed in Section 6.2.1 to provide liquidity to large funds should earn excess returns. This implication is readily testable in the data.

Wealth effect. The last term in (26), $\bar{\mathcal{W}}\, p^*$, reflects a rescaling of the baseline price p^* due to an additional channel of AUM inequality arising from wealth generated by initial holdings. These wealth effects affect all asset prices uniformly and therefore have no cross-sectional implications, similar to the effect of size heterogeneity through ϕ discussed above.

7 Conclusion

This paper develops a tractable general-equilibrium model of asset markets with non-competitive trading, wealth effects, and heterogeneous investor size. Motivated by the growing concentration of assets under management among a small set of large institutions, we analyze how investor size shapes equilibrium prices, liquidity provision, risk sharing, and welfare. Our framework nests the competitive benchmark but highlights that, once investors internalize their price impact, the distribution of wealth—not merely aggregate wealth—becomes a first-order determinant of market outcomes.

The model delivers three central insights. First, non-competitive trading fundamentally alters the mapping between investor size and trading behavior. Large investors supply more liquidity but face disproportionately greater price impact, leading them to rebalance less aggressively than under perfect competition. These cross-sectional distortions generate predictable deviations in turnover shares, liquidity supply, and return exposures—patterns that align with recent empirical findings on institutional trading.

Second, market concentration amplifies aggregate outcomes. Compared to competitive markets, non-competitive markets exhibit higher returns and higher volatility, and increases in AUM concentration further strengthen these effects. Liquidity behaves in a striking way: despite stronger market power, liquidity increases in more concentrated markets due to a wealth-effect "cushion" that flattens demand curves when prices approach their lower bound. This mechanism offers a novel explanation for the empirical observation that concentration and liquidity need not move in opposite directions.

Third, the model provides a clear metric for the degree of non-competitiveness in asset markets. We show that the wedge between competitive and non-competitive equilibria is well approximated by the Herfindahl–Hirschman Index (HHI) of AUM. Large markets are competitive only if HHI vanishes; if HHI remains positive, so does market power. This result offers a theoretical foundation for the regulatory use of concentration measures in evaluating fund mergers and assessing market competitiveness.

Extending the model to incorporate small initial holdings and small heterogeneity in risk

aversion reveals new granular wedges—based on holdings, risk aversion, and wealth effects—that shape equilibrium prices and risk premia. These wedges generate testable predictions about the cross-sectional distribution of exposures, liquidity supply, and consumption-based pricing relationships.

Overall, our results underscore that market structure is central to asset pricing. Investor size, concentration, and strategic trading interact to generate rich implications for returns, volatility, and liquidity. The theory provides new tools for interpreting empirical patterns in increasingly granular financial markets and offers guidance for policymakers evaluating consolidation in the asset-management industry.

Appendices

A A Summary of Notation

Notation	Explanation
	General mathematical notation
$q^\top$	Transpose of a vector q
$\nabla f(q)$, where $f:\ \mathbb{R}^N \to \mathbb{R}$	Gradient of f, $(\nabla f)_l = \frac{\partial f}{\partial q_l}$
$\nabla^2 f(q)$, where $f:\ \mathbb{R}^N \to \mathbb{R}$	Hessian of f, $(\nabla^2 f)_{kl} = \frac{\partial^2 f}{\partial q_k \partial q_l}$
$\nabla I(q)$, where $I:\ \mathbb{R}^N \to \mathbb{R}^N$	Jacobian of I, $(\nabla I)_{ik} = \frac{\partial I^i}{\partial q_k}$
A_{ij}	ij-th element of a matrix A.
a_i	i-th element of a vector a.
	Model variables
$I^i(q)$	Trader i's inverse demand. $I^i_k(q)$ is a price that a trader i bids for asset k, given that he gets allocation q.
$P_i(q_i)$	Inverse residual demand faced by a trader i.
$\Lambda_i(q_i) = \nabla P_i(q_i)$	Price impact matrix of a trader i.
β_i	Scaling constants. We have $I_i(q) = I(q/\beta_i)$ in a scale-symmetric equilibrium
α_i	Investor i's share of the total wealth.

Notation	Explanation
$\text{HHI} = \sum \alpha_i^2$	Herfindahl-Hirschman Index (HHI) of the wealth distribution.
	Granular wedge notation
$\Gamma_w[X]$	Granular wedge operator: $\Gamma_w[X] \equiv \sum_i w_i X_i - \frac{1}{L}\sum_i X_i$.
$\Gamma^c[\tilde{\gamma}] = \Gamma_\alpha[\tilde{\gamma}]$,	Granular risk aversion wedge (competitive case), with wealth weights α_i.
$\Gamma_{\omega^\gamma}[\tilde{\gamma}]$	Granular risk aversion wedge (non-competitive case), with $\omega_j^\gamma = \frac{\beta_j(1-\beta_j)}{\sum_k \beta_k(1-\beta_k)}$.
$\Gamma_{\omega^q}[\tilde{q}_0]$	Granular holdings wedge, with $\omega_i^q = \frac{(1+\beta_i)^{-1}}{\sum_k (1+\beta_k)^{-1}}$.
S_q	Normalization factor for holdings weights: $S_q \equiv \sum_k (1+\beta_k)^{-1}$.
$\tilde{\phi}$	Adjusted liquidity parameter: $\tilde{\phi} = \phi / S_q$.

B Proofs

B.1 Proof of Theorem 1

Lemma 1. *The function* $f(q) = \frac{E\left[\left(q^\top \delta\right)^{-\gamma}\delta\right]}{E\left[\left(q^\top \delta\right)^{1-\gamma}\right]}$ *is strictly decreasing in* q.

Proof. We compute the Jacobian $\nabla f(q)$ as follows:

$$\nabla f(q) = -\gamma \frac{E\left[\left(q^\top \delta\right)^{-\gamma-1}\delta\delta^\top\right]}{E\left[\left(q^\top \delta\right)^{1-\gamma}\right]} - (1-\gamma)\frac{E\left[\left(q^\top \delta\right)^{-\gamma}\delta\right] E\left[\left(q^\top \delta\right)^{-\gamma}\delta^\top\right]}{E\left[\left(q^\top \delta\right)^{1-\gamma}\right]^2}.$$

To rewrite this in a more interpretable form, define the probability measure $\mathcal{Q}$ via its Radon-Nikodym derivative:

$$\frac{d\mathcal{Q}}{d\mathbb{P}} = \frac{(q^\top \delta)^{1-\gamma}}{E\left[(q^\top \delta)^{1-\gamma}\right]}.$$

Under $\mathcal{Q}$, we can write

$$\nabla f(q) = -\gamma \left(\mathrm{Var}_{\mathcal{Q}} \left[\frac{\delta}{q^\top \delta} \right] \right) - E_{\mathcal{Q}} \left[\frac{\delta}{q^\top \delta} \right] E_{\mathcal{Q}} \left[\frac{\delta}{q^\top \delta} \right]^\top .$$

The first term is a positive-definite matrix (a covariance), and the second is an outer product of a vector with itself, which is positive semi-definite. Thus, the entire expression is negative definite, implying that $f(q)$ is strictly decreasing in q. ■

Proof of Theorem 1. As derived in Section 5.1, the inverse residual demand that trader i faces is given by

$$I\left(\frac{Q - q^i}{1 - \beta_i} \right),$$

where q^i represents the portfolio trader i intends to trade, and Q denotes a specific realization of supply. Therefore, trader i's ex-post optimization problem can be written as

$$\sup_q \left\{ \log \left(\alpha_i w_0 - q^\top I \left(\frac{Q - q}{1 - \beta_i} \right) \right) + \log \left(E \left[\left(q^\top \delta \right)^{1-\gamma} \right]^{\frac{1}{1-\gamma}} \right) \right\}. \tag{28}$$

The first-order condition yields

$$I\left(\frac{Q - q}{1 - \beta_i} \right) - \frac{1}{1 - \beta_i} \nabla I \left(\frac{Q - q}{1 - \beta_i} \right) q = \left(\alpha_i w_0 - q^\top I \left(\frac{Q - q}{1 - \beta_i} \right) \right) \frac{E\left[\left(q^\top \delta \right)^{-\gamma} \delta \right]}{E\left[\left(q^\top \delta \right)^{1-\gamma} \right]}. \tag{29}$$

Lemma 2 below establishes that the first-order condition (29) is both necessary and sufficient. In the scale-symmetric equilibrium, $q = \beta_i Q$ must be optimal for any admissible Q. Substituting $q = \beta_i Q$ into the expression above yields the system of PDEs (13). Applying homogeneity then reduces this system to the algebraic equations (14). For convenience, we restate (14) below:

$$\left(1 - k \frac{\beta_i}{1 - \beta_i} \right) I(Q) = \left(\alpha_i w_0 - \beta_i Q^\top I(Q) \right) \frac{E\left[\left(Q^\top \delta \right)^{-\gamma} \delta \right]}{\beta_i E\left[\left(Q^\top \delta \right)^{1-\gamma} \right]}. \tag{30}$$

Multiply by $Q^\top$ to find expenditure $E = Q^\top I(Q)$ that solves

$$E = \left(\alpha_i w_0 - \beta_i E \right) 1/\beta_i + \frac{k \beta_i}{1 - \beta_i} E. \tag{31}$$

Then it follows from (30) that $I_i(q) \propto \frac{E\left[\left(\delta^\top q \right)^{-\gamma} \delta \right]}{E\left[\left(\delta^\top q \right)^{1-\gamma} \right]}$, which implies that the inverse demand function is homogeneous of degree -1. Thus, $k = -1$. Substituting $k = -1$ back into (31), we

obtain

$$E = \frac{\alpha_i(1-\beta_i)w_0}{(2-\beta_i)\beta_i}.$$

And, from (30) we obtain

$$I(q) = \frac{\alpha_i(1-\beta_i)w_0}{(2-\beta_i)\beta_i} \frac{E\left[\left(\delta^\top q\right)^{-\gamma}\delta\right]}{E\left[\left(\delta^\top q\right)^{1-\gamma}\right]}.$$

For the scale-symmetric equilibrium to exist, we must have that

$$\frac{\alpha_i(1-\beta_i)}{(2-\beta_i)\beta_i} = \frac{1}{2\phi}, \tag{32}$$

for some constant ϕ. There is a unique solution to (32) which is between 0 and 1, given by

$$\beta_i = \alpha_i\phi + 1 - \sqrt{(\alpha_i\phi)^2 + 1}.$$

The constant ϕ is pinned down by the condition $\sum_i \beta_i = 1$:

$$\sum_i \left(\alpha_i\phi + 1 - \sqrt{(\alpha_i\phi)^2 + 1}\right) = 1.$$

The solution to the equation above exists, as the function on the left-hand side is continuous, attains 0 as $\phi \to 0$, and goes to $L > 1$ as $\phi \to \infty$. The solution is unique as the function is monotone. ■

Lemma 2. *For the equilibrium inverse demand $I(\cdot)$ given by (15), the solution $q = \beta_i Q$ attains the global maximum of the optimization problem (28).*

Proof. Define the admissible set

$$\mathcal{A}_i = \left\{ q \;\middle|\; \delta^\top q > 0,\ \alpha_i w_0 - q^\top I\left(\frac{Q-q}{1-\beta_i}\right) > 0 \right\}.$$

By assumption, the utility function U_i evaluates to negative infinity for any $q \notin \mathcal{A}_i$. Therefore, the optimum must lie either in the interior or on the boundary of $\mathcal{A}_i$. We proceed in three steps. First, we show that the first-order condition (29) admits a unique solution given by $q = \beta_i Q$. Second, we verify that the second-order conditions are satisfied at this point, confirming that $q = \beta_i Q$ is a unique interior local maximizer of problem (28). Finally, we demonstrate that the objective in (28) cannot attain its maximum at the boundary of $\mathcal{A}_i$ or for $q \to \infty$.

We begin with **the first step.** Define

$$\xi = \frac{Q - q}{1 - \beta_i}, \tag{33}$$

and let

$$f(q) = \frac{E\left[\left(q^\top \delta\right)^{-\gamma} \delta\right]}{E\left[\left(q^\top \delta\right)^{1-\gamma}\right]}.$$

The first-order condition (FOC) can be rewritten as

$$I(\xi) = \left(\alpha_i w_0 - q^\top I(\xi)\right) f(q) + \frac{1}{1 - \beta_i} \nabla I(\xi) q. \tag{34}$$

The equation (29) is thus equivalent to the pair of equations (34) and (33). We show that the unique solution to equation (34) is $q = \beta_i \xi$. The fact that $q = \beta_i \xi$ is a solution can be verified by direct substitution of $q = \beta_i \xi$, along with equation (15), into (34).

Now suppose, for the sake of contradiction, that there exists another solution $\hat{q} \neq \beta_i \xi$, $\hat{q} \in \mathcal{A}_i$. Substitute $q = \hat{q}$ and $q = \beta_i \xi$ into equation (34):

$$I(\xi) = \left(\alpha_i w_0 - \beta_i \xi^\top I(\xi)\right) f(\beta_i \xi) + \frac{1}{1 - \beta_i} \nabla I(\xi) \beta_i \xi,$$
$$I(\xi) = \left(\alpha_i w_0 - \hat{q}^\top I(\xi)\right) f(\hat{q}) + \frac{1}{1 - \beta_i} \nabla I(\xi) \hat{q}.$$

Subtracting these two expressions and premultiplying both sides by $(\hat{q} - \beta_i \xi)^\top$, we obtain:

$$\begin{aligned} 0 = &\frac{1}{1 - \beta_i} (\hat{q} - \beta_i \xi)^\top \nabla I(\xi) (\hat{q} - \beta_i \xi) \\ &+ (\hat{q} - \beta_i \xi)^\top \left(g(\hat{q}) - g(\beta_i \xi)\right), \end{aligned}$$

where we define, for the purpose of this proof only,

$$g(q) \equiv \left(\alpha_i w_0 - q^\top I(\xi)\right) f(q).$$

Note that $\nabla I(\xi)$ is negative definite. This follows from the fact that it is proportional (with a positive coefficient) to $\nabla f(\xi)$, which we have shown to be negative definite in Lemma 1. Therefore,

$$(\hat{q} - \beta_i \xi)^\top \nabla I(\xi) (\hat{q} - \beta_i \xi) < 0.$$

It remains to show that

$$(\hat{q} - \beta_i \xi)^\top (g(\hat{q}) - g(\beta_i \xi)) < 0, \tag{35}$$

which will lead to a contradiction and thus establish the uniqueness of the solution. To establish (35), consider:

$$\begin{aligned}
(g(\hat{q}) - g(\beta_i \xi))^\top (\hat{q} - \beta_i \xi) &= \left[\left(\alpha_i w_0 - \hat{q}^\top I(\xi) \right) f(\hat{q}) - \left(\alpha_i w_0 - \beta_i \xi^\top I(\xi) \right) f(\beta_i \xi) \right]^\top (\hat{q} - \beta_i \xi) \\
&= -(\hat{q} - \beta_i \xi)^\top I(\xi)\, f(\beta_i \xi)^\top (\hat{q} - \beta_i \xi) \qquad (36) \\
&\quad + \underbrace{\left(\alpha_i w_0 - \hat{q}^\top I(\xi) \right)}_{>0 \text{ since } \hat{q} \in \mathcal{A}_i} \underbrace{\left(f(\hat{q}) - f(\beta_i \xi) \right)^\top (\hat{q} - \beta_i \xi)}_{<0 \text{ since } f \text{ is strictly decreasing}}. \qquad (37)
\end{aligned}$$

The second equality above follows by adding and subtracting

$$(\alpha_i w_0 - \hat{q}^\top I(\xi)) f(\beta_i \xi)^\top (\hat{q} - \beta_i \xi).$$

Moreover, since $I(\xi)$ is proportional to $f(\beta_i \xi)$ with a positive coefficient, the term in (36) is negative. The term in (37) is also negative, establishing (35). We have therefore shown that the pair of equations (34) and (33) reduce to $q = \beta_i \xi$ and $\xi = (Q - q)/(1 - \beta_i)$, which imply that $q = \beta_i Q$. Thus, the first-order condition (29) admits a unique solution given by $q = \beta_i Q$.

We proceed to the **second step**, verifying that the second-order conditions are satisfied at $q = \beta_i Q$. Specifically, we show that both

$$\log \left(\alpha_i w_0 - q^\top I \left(\frac{Q - q}{1 - \beta_i} \right) \right) \quad \text{and} \quad \log \left(E \left[\left(q^\top \delta \right)^{1-\gamma} \right]^{\frac{1}{1-\gamma}} \right)$$

are locally concave at $q = \beta_i Q$. The concavity of the second expression follows from the fact that its Hessian is proportional to $\nabla f(q)$, which is negative definite by Lemma 1.

To establish the concavity of the first term, we first exploit the homogeneity of the inverse demand function. Since $I(\cdot)$ is homogeneous of degree -1, we have

$$I \left(\frac{Q - q}{1 - \beta_i} \right) = (1 - \beta_i) I(Q - q).$$

Consequently, the term inside the logarithm simplifies to $\alpha_i w_0 - (1 - \beta_i) q^\top I(Q - q)$. Since $1 - \beta_i > 0$, it suffices to show that the expenditure function $q \mapsto q^\top I(Q - q)$ is convex at

$q = \beta_i Q$. This, in turn, follows from the convexity of the auxiliary function

$$h(y) = \frac{E\left[\left(\delta^\top y\right)^{-\gamma} \delta^\top Q\right]}{E\left[\left(\delta^\top y\right)^{1-\gamma}\right]},$$

evaluated at $y = (1-\beta_i)Q$. Indeed, by decomposing $Q = (Q-q)+q$ in the numerator of $h(Q-q)$ and applying the identity $y^\top I(y) = w_0/(2\phi)$ implied by (15), we obtain the relationship:

$$h(Q-q) - \frac{2\phi}{w_0} q^\top I(Q-q) = 1.$$

Thus, the convexity of $h(\cdot)$ directly implies the convexity of the function $q \mapsto q^\top I(Q-q)$, completing the argument. We compute the Hessian of $h(y)$:

$$\begin{aligned}
\nabla^2 h(y) = \gamma(\gamma+1)&\frac{E\left[\left(\delta^\top y\right)^{-\gamma-2}\left(\delta^\top Q\right)\delta\delta^\top\right]}{E\left[\left(\delta^\top y\right)^{1-\gamma}\right]} \\
&+ 2\gamma(1-\gamma)\frac{E\left[\left(\delta^\top y\right)^{-\gamma-1}\left(\delta^\top Q\right)\delta\right] E\left[\left(\delta^\top y\right)^{-\gamma}\delta^\top\right]}{E\left[\left(\delta^\top y\right)^{1-\gamma}\right]^2} \\
&+ \gamma(1-\gamma)\frac{E\left[\left(\delta^\top y\right)^{-\gamma}\delta^\top Q\right] E\left[\left(\delta^\top y\right)^{-\gamma-1}\delta\delta^\top\right]}{E\left[\left(\delta^\top y\right)^{1-\gamma}\right]^2} \\
&+ 2(1-\gamma)^2\frac{E\left[\left(\delta^\top y\right)^{-\gamma}\delta^\top Q\right] E\left[\left(\delta^\top y\right)^{-\gamma}\delta\right] E\left[\left(\delta^\top y\right)^{-\gamma}\delta^\top\right]}{E\left[\left(\delta^\top y\right)^{1-\gamma}\right]^3}.
\end{aligned}$$

Substituting $y = (1-\beta_i)Q$, we obtain after simplification:

$$\frac{1}{2}(1-\beta_i)^3 \nabla^2 h((1-\beta_i)Q) = \gamma \cdot \mathrm{Var}_{\mathcal{Q}}\left[\frac{\delta}{Q^\top\delta}\right] + E_{\mathcal{Q}}\left[\frac{\delta}{Q^\top\delta}\right] E_{\mathcal{Q}}\left[\frac{\delta}{Q^\top\delta}\right]^\top,$$

which is clearly positive definite. Here, $\mathcal{Q}$ is a probability measure defined via its Radon–Nikodym derivative with respect to the original measure $\mathbb{P}$:

$$\frac{d\mathcal{Q}}{d\mathbb{P}} = \frac{(Q^\top\delta)^{1-\gamma}}{E\left[(Q^\top\delta)^{1-\gamma}\right]}.$$

Third step. To conclude, we show that the maximum of the objective function in (28)

cannot be attained at the boundary of $\mathcal{A}_i$ or at infinity. First, observe that for any q such that $\delta^\top q > 0$, the term $(Q-q)^\top \delta$ becomes negative for sufficiently large q. In such cases, the expression $I\left(\frac{Q-q}{1-\beta_i}\right)$ is not well-defined. We resolve this by defining the residual supply such that $q^\top(\text{Residual Supply}(q))$ is infinite whenever $(Q-q)^\top \delta \leq 0$. Hence, such q do not belong to $\mathcal{A}_i$, and the objective cannot attain a maximum as $\|q\| \to \infty$. Second, on the boundary of $\mathcal{A}_i$, either $\delta^\top q \to 0$ or the budget constraint becomes tight, which again leads to utility approaching $-\infty$. Therefore, the maximum cannot be attained on the boundary of $\mathcal{A}_i$ either. We conclude that the unique maximizer lies in the interior of $\mathcal{A}_i$ and is given by $q = \beta_i Q$. ■

B.2 Proof of Proposition 2

Proof of Proposition 2.

The function

$$g(x;\, a) \;=\; ax + 1 - \sqrt{(ax)^2 + 1}$$

is strictly increasing in x, for any arbitrary constant $a > 1$. Thus, for any i and j such that $\alpha_i > \alpha_j$, we have

$$\beta_i \;=\; g(\alpha_i;\, \phi) \;> g(\alpha_j;\, \phi) \;=\; \beta_j.$$

.

From the relationship $\Lambda_i \propto 1/(1-\beta_i)$ (cf. (12)), it immediately follows that for any i and j such that $\alpha_i > \alpha_j$, we also have $\Lambda_i > \Lambda_j$ in the positive-definite order.

For the final part, consider $\beta_i = \alpha_i \phi + 1 - \sqrt{(\alpha_i \phi)^2 + 1}$ as a function of $\alpha_i \in [0, 1]$ for a given ϕ. It can be shown that β_i is concave, starts at zero, and crosses the 45-degree line ($\beta_i = \alpha_i$) exactly once for $\alpha_i > 0$. Consequently, there can be at most one threshold i^*.

Such a threshold must exist because the largest β_i is smaller than the largest α_i. If this were not the case, then given the single crossing property established earlier, we would have $\beta_i > \alpha_i$ for all i, violating the condition to pin down ϕ in (16). Furthermore, there must exist some i for which $\beta_i > \alpha_i$, as otherwise (16) would again be violated. ■

B.3 Proof of Proposition 3

Proof of Proposition 3. It suffices to prove that, in equilibrium, $\phi > 1$. Consider (16). The left-hand side of this equation is a continuously increasing function of ϕ that approaches $L > 1$ as $\phi \to \infty$. Therefore, it is enough to show that the left-hand side of (16) is strictly less than

1 when $\phi = 1$.

Indeed, by multiplying and dividing by $\alpha_i\phi + 1 + \sqrt{(\alpha_i\phi)^2 + 1}$, we can rewrite the left-hand side of (16) as

$$\sum_i \frac{2\alpha_i\phi}{\alpha_i\phi + 1 + \sqrt{(\alpha_i\phi)^2 + 1}}\bigg|_{\phi=1} = \sum_i \frac{2\alpha_i}{\alpha_i + 1 + \sqrt{\alpha_i^2 + 1}}.$$

This sum satisfies

$$\sum_i \frac{2\alpha_i}{\alpha_i + 1 + \sqrt{\alpha_i^2 + 1}} < \sum_i \alpha_i = 1.$$

■

B.4 Proof of Proposition 5

Proof of Proposition 5. The relationship $\frac{\hat{\mu}_k}{\mu_k} = \frac{\hat{\sigma}_k}{\sigma_k} = \frac{\hat{\Lambda}_{kl}}{\Lambda_{kl}} = \frac{\hat{\phi}}{\phi}$ follows directly from Proposition 3. For instance, for expected returns, we can derive:

$$\frac{\hat{\mu}_k}{\mu_k} = \frac{\hat{\mu}_k}{\hat{\mu}_k^c} \cdot \frac{\mu_k^c}{\mu_k} = \frac{\hat{\phi}}{\phi}.$$

To show that an increase in inequality leads to an increase in ϕ, we analyze the impact of wealth redistribution. The case of a decrease in inequality is analogous and omitted for brevity.

Consider a change in the wealth distribution from α to $\hat{\alpha}$, corresponding to a transfer of funds from a smaller large investor i to a larger large investor j. Specifically,

$$\alpha = \{\alpha_1, \ldots, \alpha_i, \ldots, \alpha_j, \ldots, \alpha_L\}, \quad \hat{\alpha} = \{\alpha_1, \ldots, \alpha_i - y, \ldots, \alpha_j + y, \ldots, \alpha_L\},$$

where $y \leq \alpha_i \leq \alpha_j$.

Define $b(\alpha, \phi) = \alpha\phi + 1 - \sqrt{(\alpha\phi)^2 + 1}$. The equation (16), which determines ϕ, can then be expressed as:

$$\sum_i b(\alpha_i, \phi) = 1.$$

Observe that for a given ϕ, the function $b(\alpha, \phi)$ is concave and increasing in α. Consequently,

$$|b(\alpha_i - y, \phi) - b(\alpha_i, \phi)| > b(\alpha_j + y, \phi) - b(\alpha_j, \phi).$$

Thus, for a given ϕ,

$$\sum_i b(\alpha_i, \phi) > \sum_i b(\hat{\alpha}_i, \phi).$$

Since $b(\alpha, \phi)$ is increasing in ϕ, for the equation $\sum_i b(\hat{\alpha}_i, \hat{\phi}) = 1$ to hold, we must have $\hat{\phi} > \phi$.

The case of a merger is equivalent (in terms of pinning down ϕ) to the case considered above with $y = \alpha_i$. ■

C Proof of Proposition 4

Proof of Proposition 4. We proceed in three main steps: expanding the equilibrium condition, solving order-by-order, and substituting the HHI definition.

1. Expanding the equilibrium condition

Let $x_i = \alpha_i \phi$ and define the function $f(x) = x + 1 - \sqrt{x^2 + 1}$. The equilibrium condition is $\sum_{i=1}^{L} f(x_i) = 1$. We assume an expansion for ϕ of the form:

$$\phi = \phi_0 + \epsilon\phi_1 + \epsilon^2\phi_2 + O(\epsilon^3)$$

Substituting the expansions for α_i and ϕ into x_i:

$$x_i = (\alpha + \epsilon a_i)(\phi_0 + \epsilon\phi_1 + \epsilon^2\phi_2)$$

We expand x_i around the base value $x_0 = \alpha\phi_0$, grouping terms by powers of ϵ. Let $x_i = x_0 + \Delta_i$:

$$\Delta_i = \epsilon(\alpha\phi_1 + a_i\phi_0) + \epsilon^2(\alpha\phi_2 + a_i\phi_1) + O(\epsilon^3)$$

We Taylor expand the sum condition $\sum f(x_i) = 1$ around x_0:

$$\sum_{i=1}^{L} \left[f(x_0) + f'(x_0)\Delta_i + \frac{1}{2} f''(x_0)\Delta_i^2 \right] + O(\epsilon^2) = 1 \tag{38}$$

Let $f_0 \equiv f(x_0)$, $f_1 \equiv f'(x_0)$, and $f_2 \equiv f''(x_0)$.

2. Solving Order by Order

Step 2a: Zeroth Order (ϵ^0)

Matching terms of order ϵ^0 in Eq. (38):

$$\sum_{i=1}^{L} f_0 = Lf(x_0) = 1 \implies f(x_0) = \frac{1}{L}$$

Solving $x_0 + 1 - \sqrt{x_0^2 + 1} = 1/L$ yields:

$$x_0 = \frac{2L-1}{2L(L-1)} \implies \phi_0 = \frac{x_0}{\alpha} = \frac{2L-1}{2(L-1)}$$

Here we substituted $\alpha = 1/L$.

Step 2b: First Order (ϵ^1)

Matching terms of order ϵ^1:

$$f_1 \sum_{i=1}^{L} (\alpha\phi_1 + a_i\phi_0) = 0$$

Since α and ϕ_1 are constants inside the sum, and $\sum a_i = 0$:

$$f_1(L\alpha\phi_1 + \phi_0 \underbrace{\sum a_i}_{0}) = 0 \implies \phi_1 = 0$$

Step 2c: Second Order (ϵ^2)

With $\phi_1 = 0$, the perturbation simplifies to $\Delta_i = \epsilon a_i\phi_0 + \epsilon^2\alpha\phi_2 + O(\epsilon^3)$. We collect ϵ^2 terms from Eq. (38), which come from the linear term ($f_1\Delta_i$) and the quadratic term ($\frac{1}{2}f_2\Delta_i^2$):

$$\sum_{i=1}^{L} \left[f_1\alpha\phi_2 + \frac{1}{2} f_2 a_i^2 \phi_0^2 \right] = 0$$

Summing over i (noting $\sum a_i^2$ remains):

$$L\alpha f_1\phi_2 + \frac{1}{2} f_2\phi_0^2 \sum_{i=1}^{L} a_i^2 = 0$$

Solving for ϕ_2:

$$\phi_2 = -\frac{f_2\phi_0^2}{2L\alpha f_1} \sum_{i=1}^{L} a_i^2$$

Substituting derivatives $f_1 = \mu/S$ and $f_2 = -1/S^3$ (with $S = \sqrt{x_0^2 + 1}, \mu = 1 - 1/L, \alpha = 1/L$):

$$\phi_2 = \left[\frac{L^3(2L-1)^2}{2(L-1)(2L^2-2L+1)^2}\right] \sum_{i=1}^{L} a_i^2 \tag{39}$$

3. Relating to HHI

The HHI is defined as $HHI = \sum \alpha_i^2$. Substituting $\alpha_i = 1/L + \epsilon a_i$:

$$HHI = \sum_{i=1}^{L} \left(\frac{1}{L^2} + \frac{2\epsilon a_i}{L} + \epsilon^2 a_i^2\right) = \frac{1}{L} + \epsilon^2 \sum_{i=1}^{L} a_i^2$$

Solving for the sum of squared perturbations:

$$\epsilon^2 \sum_{i=1}^{L} a_i^2 = HHI - \frac{1}{L} \tag{40}$$

4. Final Form

Let K be the bracketed coefficient in Eq. (39). Substituting (40) into the expansion $\phi \approx \phi_0 + \epsilon^2 \phi_2$:

$$\phi = \phi_0 + K\left(HHI - \frac{1}{L}\right) + O(\epsilon^3)$$

Grouping terms yields $\phi = C_0 + C_1 \cdot HHI$, where $C_1 \equiv K$ and $C_0 \equiv \phi_0 - K/L$. ■

D Proof of Proposition 6

Proof of Proposition 6. We first derive the demand function for general parameters $\theta_i \equiv (w_{0,i}, \gamma_i, q_{0,i})$, where $w_{0,i} = \alpha_i w_0$. The first-order condition (FOC) for the investor's optimization is:

$$p = c_{0,i}\, g(q_{0,i} + D_i^c(p), \gamma_i),$$

where $c_{0,i} = w_{0,i} - p^\top D_i^c(p)$ is time-0 consumption and the function g is defined as:

$$g(q, \gamma) \equiv \frac{\mathbb{E}\left[\left(q^\top \delta\right)^{-\gamma} \delta\right]}{\mathbb{E}\left[\left(q^\top \delta\right)^{1-\gamma}\right]}.$$

The function $g(\cdot, \gamma)$ is homogeneous of degree -1 in its first argument, implying the identity $q^\top g(q, \gamma) = 1$. Multiplying the FOC by $(q_{0,i} + D_i^c(p))^\top$ yields:

$$p^\top (q_{0,i} + D_i^c(p)) = c_{0,i} \underbrace{(q_{0,i} + D_i^c(p))^\top g(q_{0,i} + D_i^c(p), \gamma_i)}_{=1} = c_{0,i}.$$

Substituting the budget constraint $c_{0,i} = w_{0,i} - p^\top D_i^c(p)$ into the equation above, we solve for the expenditure:

$$p^\top D_i^c(p) = \frac{1}{2}(w_{0,i} - p^\top q_{0,i}) \implies c_{0,i} = \frac{1}{2}(w_{0,i} + p^\top q_{0,i}).$$

We define the baseline aggregate inverse demand function implicitly by $p \equiv \frac{w_0}{2} g(D^{*,c}(p, \gamma), \gamma)$. Utilizing the homogeneity of g, the FOC can be inverted to yield:

$$q_{0,i} + D_i^c(p) = \frac{2c_{0,i}}{w_0} D^{*,c}(p, \gamma_i). \tag{41}$$

Substituting $c_{0,i}$ into (41) provides the exact demand function:

$$D_i^c(p) = \frac{\alpha_i w_0 + p^\top q_{0,i}}{w_0} D^{*,c}(p, \gamma_i) - q_{0,i}.$$

We expand this expression around the baseline parameters $q_{0,i} = 0$ and $\gamma_i = \gamma$. Let $\tilde{q}_{0,i}$ and $\tilde{\gamma}_i$ denote the perturbations. The zeroth-order term is $D_i^{*,c}(p) = \alpha_i D^{*,c}(p, \gamma)$. The first-order term is obtained by differentiating with respect to $q_{0,i}$ and γ_i:

$$\begin{aligned}
\tilde{D}_i^c(p) &= \nabla_{q_{0,i}} D_i^c \cdot \tilde{q}_{0,i} + \frac{\partial D_i^c}{\partial \gamma_i} \tilde{\gamma}_i \\
&= \left(\frac{p^\top \tilde{q}_{0,i}}{w_0} D^{*,c}(p, \gamma) - \tilde{q}_{0,i} \right) + \left(\frac{\alpha_i w_0}{w_0} \frac{\partial D^{*,c}(p, \gamma)}{\partial \gamma} \tilde{\gamma}_i \right).
\end{aligned}$$

Rearranging terms yields (18). ■

E Proof of Proposition 7

Proof of Proposition 7. Aggregating the individual demand expansions from Proposition 6 yields the aggregate demand perturbation $\tilde{D}^c(p) \equiv \sum_i \tilde{D}_i^c(p)$:

$$\tilde{D}^c(p) = -\sum_i \tilde{q}_{0,i} + \frac{p^\top (\sum_i \tilde{q}_{0,i})}{w_0} D^{*,c}(p) + \Gamma_\alpha[\tilde{\gamma}]\, D_\gamma^{*,c}(p).$$

The normalization $\sum_i \tilde{q}_{0,i} = 0$ (and hence $\sum_i \tilde{\gamma}_i = 0$) implies that the first two terms vanish, and that the granular wedge $\Gamma_\alpha[\tilde{\gamma}] = \sum_i \alpha_i \tilde{\gamma}_i$. The aggregate demand perturbation therefore simplifies to:

$$\tilde{D}^c(p) = \Gamma_\alpha[\tilde{\gamma}]\, D^{*,c}_\gamma(p). \tag{42}$$

Thus, aggregate demand is unaffected by the purely redistributional endowment shocks $\tilde{q}_{0,i}$ and depends only on the granular wedge of risk aversion shocks.

To derive the price perturbation, consider the market clearing condition $D^c(p^c) = Q$. Substituting the expansions $p^c = p^{*,c} + \tilde{p}^c$ and $D^c(p) = D^{*,c}(p) + \tilde{D}^c(p)$ around the baseline price $p^{*,c}$, we obtain:

$$\underbrace{D^{*,c}(p^{*,c})}_{=Q} + \nabla_p D^{*,c}(p^{*,c})\tilde{p}^c + \tilde{D}^c(p^{*,c}) = Q.$$

The zero-order terms cancel. Substituting (42) into the linearized market clearing condition yields:

$$\nabla_p D^{*,c}(p^{*,c})\tilde{p}^c + \Gamma_\alpha[\tilde{\gamma}]\, D^{*,c}_\gamma(p^{*,c}) = 0.$$

Solving for $\tilde{p}^c$, we obtain:

$$\tilde{p}^c = -\left(\nabla_p D^{*,c}(p^{*,c})\right)^{-1} D^{*,c}_\gamma(p^{*,c})\, \Gamma_\alpha[\tilde{\gamma}] = \Gamma_\alpha[\tilde{\gamma}]\, p^{*,c}_\gamma, \tag{43}$$

where the second equality follows from differentiating the baseline identity $D^{*,c}(p^{*,c}(Q,\gamma),\gamma) = Q$ with respect to γ.

To express $\tilde{p}^c$ in terms of the consumption-numeraire measure, recall from (8) that $p^{*,c} = \frac{w_0}{2} E^*[\hat{\delta}]$, where E^* denotes expectation under the measure P^* defined in (7). Since P^* depends on γ through its Radon–Nikodym derivative, we compute:

$$E^*[\hat{\delta}] = \frac{E\left[\hat{\delta}\, C_{\text{agg}}^{1-\gamma}\right]}{E\left[C_{\text{agg}}^{1-\gamma}\right]}.$$

Differentiating with respect to γ:

$$\begin{aligned}
\frac{\partial}{\partial\gamma} E^*[\hat{\delta}] &= \frac{-E\left[\hat{\delta}\, C_{\text{agg}}^{1-\gamma} \ln C_{\text{agg}}\right] \cdot E\left[C_{\text{agg}}^{1-\gamma}\right] + E\left[\hat{\delta}\, C_{\text{agg}}^{1-\gamma}\right] \cdot E\left[C_{\text{agg}}^{1-\gamma} \ln C_{\text{agg}}\right]}{\left(E\left[C_{\text{agg}}^{1-\gamma}\right]\right)^2} \\
&= -E^*\left[\hat{\delta} \ln C_{\text{agg}}\right] + E^*[\hat{\delta}]\, E^*\left[\ln C_{\text{agg}}\right] \\
&= -\text{Cov}^*\left(\hat{\delta}, \ln C_{\text{agg}}\right).
\end{aligned}$$

Therefore, the price sensitivity to risk aversion is:

$$p_{\gamma}^{*,c} = \frac{w_0}{2}\frac{\partial}{\partial \gamma}E^*[\hat{\delta}] = -\frac{w_0}{2}\mathrm{Cov}^*\left(\hat{\delta}, \ln C_{\mathrm{agg}}\right).$$

Substituting into (43) yields:

$$\tilde{p}^c = -\frac{w_0}{2}\Gamma_\alpha[\tilde{\gamma}]\,\mathrm{Cov}^*\left(\hat{\delta}, \ln C_{\mathrm{agg}}\right).$$

■

F Proof of Proposition 8

Proof of Proposition 8. Aggregating the individual demand perturbations from Theorem 2 yields:

$$\tilde{D}(p) \equiv \sum_i \tilde{D}_i(p) = \sum_i \eta_{0,i}\tilde{q}_{0,i} + \left(\sum_i \eta_{1,i}\right) D_\gamma^*(p) + \left(\sum_i \eta_{2,i}\tilde{q}_{0,i}^\top p\right) D^*(p).$$

Hedging term: Let $S_q \equiv \sum_k (1+\beta_k)^{-1}$. Since $\eta_{0,i} = -\frac{1}{1+\beta_i}$ and $\omega_i^q = \frac{(1+\beta_i)^{-1}}{S_q}$, we have $\eta_{0,i} = -S_q\,\omega_i^q$. The first term is:

$$\sum_i \eta_{0,i}\tilde{q}_{0,i} = -\sum_i \frac{\tilde{q}_{0,i}}{1+\beta_i} = -S_q \sum_i \omega_i^q \tilde{q}_{0,i} = -S_q\,\Gamma_{\omega^q}[\tilde{q}_0].$$

Risk aversion term: Using $\eta_{1,i} = \beta_i(1-\beta_i)\tilde{\gamma}_i + \beta_i^2\Gamma_{\omega^\gamma}[\tilde{\gamma}]$ and the granular wedge definition $\Gamma_{\omega^\gamma}[\tilde{\gamma}] = \frac{\sum_j \beta_j(1-\beta_j)\tilde{\gamma}_j}{\sum_j \beta_j(1-\beta_j)}$, we compute:

$$\sum_i \eta_{1,i} = \sum_i \beta_i(1-\beta_i)\tilde{\gamma}_i + \Gamma_{\omega^\gamma}[\tilde{\gamma}]\sum_i \beta_i^2.$$

The first sum equals $\Gamma_{\omega^\gamma}[\tilde{\gamma}]\sum_i \beta_i(1-\beta_i)$ by definition of $\Gamma_{\omega^\gamma}[\tilde{\gamma}]$. Since $\sum_i \beta_i = 1$, we have $\sum_i \beta_i(1-\beta_i) = 1 - \sum_i \beta_i^2$. Thus:

$$\sum_i \eta_{1,i} = \Gamma_{\omega^\gamma}[\tilde{\gamma}](1 - \sum_i \beta_i^2) + \Gamma_{\omega^\gamma}[\tilde{\gamma}]\sum_i \beta_i^2 = \Gamma_{\omega^\gamma}[\tilde{\gamma}].$$

Wealth effect term: The third term is $\mathcal{W}D^*(p)$ where $\mathcal{W} = \sum_i \eta_{2,i}(\tilde{q}_{0,i}^\top p)$.

Combining these results yields (27).

Price perturbation: The market clearing condition $D(p) = Q$ implies:

$$D^*(p^*) + \nabla_p D^*(p^*)\tilde{p} + \tilde{D}(p^*) + O(\epsilon^2) = Q.$$

Since $D^*(p^*) = Q$, the zeroth-order terms cancel. Solving for $\tilde{p}$:

$$\tilde{p} = -(\nabla_p D^*(p^*))^{-1}\tilde{D}(p^*) = \Lambda^*(p^*)\tilde{D}(p^*).$$

We use the following identities. First, from Euler's theorem for homogeneous functions (since D^* is homogeneous of degree -1 in p), we have $\nabla_p D^*(p) \cdot p = -D^*(p)$, which implies:

$$\Lambda^*(p^*)D^*(p^*) = p^*.$$

Second, differentiating the baseline equilibrium condition $D^*(p^*(Q,\gamma),\gamma) = Q$ with respect to γ yields:

$$\nabla_p D^* \cdot p^*_\gamma + D^*_\gamma = 0 \implies p^*_\gamma = \Lambda^*(p^*)D^*_\gamma(p^*).$$

Substituting the aggregate demand perturbation (27) and applying these identities:

$$\begin{aligned}\tilde{p} &= \Lambda^*(p^*)\left(-S_q\,\Gamma_{\omega^q}[\tilde{q}_0] + \Gamma_{\omega^\gamma}[\tilde{\gamma}]D^*_\gamma(p^*) + \mathcal{W}D^*(p^*)\right)\\ &= -S_q\,\Lambda^*(p^*)\Gamma_{\omega^q}[\tilde{q}_0] + \Gamma_{\omega^\gamma}[\tilde{\gamma}]p^*_\gamma + \mathcal{W}p^*,\end{aligned} \tag{44}$$

which is the stated expression.

Representation under P^*: To express prices in terms of the consumption-numeraire measure, recall from (15) that the baseline inverse demand is $I^*(Q) = \frac{w_0}{2\phi}g(Q,\gamma)$, where $g(Q,\gamma) = \frac{E[(Q^\top\delta)^{-\gamma}\delta]}{E[(Q^\top\delta)^{1-\gamma}]}$. Using $\delta = \hat{\delta}C_{\text{agg}}$ with $C_{\text{agg}} = \delta^\top Q$, we obtain:

$$g(Q,\gamma) = \frac{E[\hat{\delta}\,C_{\text{agg}}^{1-\gamma}]}{E[C_{\text{agg}}^{1-\gamma}]} = E^*[\hat{\delta}],$$

where $E^*[\cdot]$ denotes expectation under the measure P^* defined in (7). Thus, the baseline price is:

$$p^* = I^*(Q) = \frac{w_0}{2\phi}E^*[\hat{\delta}],$$

which establishes the stated result.

For the price sensitivity to risk aversion, differentiating $E^*[\hat{\delta}]$ with respect to γ yields

(following the same calculation as in the proof of Proposition 7):

$$\frac{\partial}{\partial\gamma}E^*[\hat{\delta}] = -\mathrm{Cov}^*\left(\hat{\delta}, \ln C_{\mathrm{agg}}\right).$$

Therefore:

$$p^*_\gamma = \frac{w_0}{2\phi}\frac{\partial}{\partial\gamma}E^*[\hat{\delta}] = -\frac{w_0}{2\phi}\mathrm{Cov}^*\left(\hat{\delta}, \ln C_{\mathrm{agg}}\right). \tag{45}$$

For the illiquidity matrix, differentiating the baseline inverse demand $I^*(Q) = \frac{w_0}{2\phi}g(Q,\gamma)$ yields (by the same calculation as in (9)):

$$\Lambda^*(Q) = \nabla_Q I^*(Q) = \frac{w_0}{2\phi}\left(\gamma\,\mathrm{Var}^*[\hat{\delta}] + E^*[\hat{\delta}]\,E^*[\hat{\delta}]^\top\right).$$

Applying $S_q\,\Lambda^*(Q)$ to the granular holdings wedge $\Gamma_{\omega^q}[\tilde{q}_0]$:

$$\begin{aligned} S_q\,\Lambda^*(p^*)\Gamma_{\omega^q}[\tilde{q}_0] &= \frac{S_q\,w_0}{2\phi}\left(\gamma\,\mathrm{Var}^*[\hat{\delta}]\,\Gamma_{\omega^q}[\tilde{q}_0] + (E^*[\hat{\delta}]^\top\Gamma_{\omega^q}[\tilde{q}_0])\,E^*[\hat{\delta}]\right) \\ &= \frac{\gamma w_0}{2\tilde{\phi}}\mathrm{Cov}^*\left(\hat{\delta}, \hat{\delta}^\top\Gamma_{\omega^q}[\tilde{q}_0]\right) + \frac{2\tilde{\phi}}{w_0}\Gamma_{\omega^q}[\tilde{q}_0]^\top p^*\cdot p^*, \end{aligned}$$

where we used $S_q = \phi/\tilde{\phi}$, $\mathrm{Var}^*[\hat{\delta}]\,\Gamma_{\omega^q}[\tilde{q}_0] = \mathrm{Cov}^*(\hat{\delta}, \hat{\delta}^\top\Gamma_{\omega^q}[\tilde{q}_0])$, and $E^*[\hat{\delta}] = \frac{2\phi}{w_0}p^*$.

Substituting (45) and the expression for $S_q\,\Lambda^*(p^*)\Gamma_{\omega^q}[\tilde{q}_0]$ into (44):

$$\begin{aligned} \tilde{p} &= -\frac{w_0}{2\phi}\Gamma_{\omega^\gamma}[\tilde{\gamma}]\,\mathrm{Cov}^*\left(\hat{\delta}, \ln C_{\mathrm{agg}}\right) - \frac{\gamma w_0}{2\tilde{\phi}}\mathrm{Cov}^*\left(\hat{\delta}, \hat{\delta}^\top\Gamma_{\omega^q}[\tilde{q}_0]\right) \\ &\quad - \frac{2\tilde{\phi}}{w_0}\Gamma_{\omega^q}[\tilde{q}_0]^\top p^*\cdot p^* + \mathcal{W}p^* \\ &= -\frac{w_0}{2\phi}\Gamma_{\omega^\gamma}[\tilde{\gamma}]\,\mathrm{Cov}^*\left(\hat{\delta}, \ln C_{\mathrm{agg}}\right) - \frac{\gamma w_0}{2\tilde{\phi}}\mathrm{Cov}^*\left(\hat{\delta}, \hat{\delta}^\top\Gamma_{\omega^q}[\tilde{q}_0]\right) + \bar{\mathcal{W}}\,p^*, \end{aligned}$$

where $\bar{\mathcal{W}} \equiv \mathcal{W} - \frac{2\tilde{\phi}}{w_0}\Gamma_{\omega^q}[\tilde{q}_0]^\top p^*$. This establishes (26). ■

G Proof of Theorem 2

Proof of Theorem 2. We derive the perturbed equilibrium by linearizing the first-order conditions around the symmetric baseline equilibrium where $q_{0,i} = 0$ and $\gamma_i = \gamma$.

Step 1: Linearization of the First-Order Condition. The optimality condition for

agent i is:

$$p + \Lambda_i(p) D_i(p) = \left(\alpha_i w_0 - D_i(p)^\top p\right) g(q_{0,i} + D_i(p), \gamma_i), \tag{46}$$

where $g(x, \gamma) \equiv \mathbb{E}[(x^\top \delta)^{-\gamma} \delta] / \mathbb{E}[(x^\top \delta)^{1-\gamma}]$. We expand all terms to first order in ϵ. Let $D_i(p) = D_i^*(p) + \tilde{D}_i(p)$, where the baseline demand is $D_i^*(p) = \beta_i D^*(p)$.

Price Impact Expansion: The price impact is $\Lambda_i(p) = -(\sum_{j \neq i} \nabla_p D_j(p))^{-1}$. Using the matrix inversion formula $(A + \epsilon B)^{-1} = A^{-1} - \epsilon A^{-1} B A^{-1} + O(\epsilon^2)$, the expansion of the price impact matrix is:

$$\Lambda_i(p) = \Lambda_i^*(p) + \Lambda_i^*(p) \left(\nabla_p \tilde{D}_{-i}(p)\right) \Lambda_i^*(p) + O(\epsilon^2),$$

where $\Lambda_i^*(p) = \frac{1}{1-\beta_i} \Lambda^*(p)$ and $\Lambda^*(p) = -(\nabla_p D^*(p))^{-1}$.

LHS Expansion: Using the baseline relationship $\Lambda_i^* D_i^* = \frac{\beta_i}{1-\beta_i} p$, the LHS of (46) expands to:

$$\text{LHS} = p \left(1 + \frac{\beta_i}{1-\beta_i}\right) + \frac{1}{1-\beta_i} \Lambda^* \tilde{D}_i + \frac{\beta_i}{(1-\beta_i)^2} \Lambda^* (\nabla_p \tilde{D}_{-i}) p + O(\epsilon^2).$$

RHS Expansion: We exploit the homogeneity of g. Specifically, $g(D_i^*, \gamma) = \frac{2\phi}{w_0 \beta_i} p$, $\nabla_q g(D_i^*, \gamma) = -\frac{2\phi}{w_0 \beta_i^2} \Lambda^*$, and $g_\gamma(D_i^*, \gamma) = \frac{2\phi}{w_0 \beta_i} p_\gamma^*$, where p_γ^* is the baseline price sensitivity. The baseline expenditure is $p^\top D_i^* = \frac{\beta_i w_0}{2\phi}$. Expanding the RHS of (46) and keeping first-order terms yields:

$$\begin{aligned} \text{RHS} = \frac{p}{1-\beta_i} + w_0 \left(\alpha_i - \frac{\beta_i}{2\phi}\right) \left[-\frac{2\phi}{w_0 \beta_i^2} \Lambda^* (\tilde{D}_i + \tilde{q}_{0,i}) + \frac{2\phi}{w_0 \beta_i} p_\gamma^* \tilde{\gamma}_i\right] \\ - (\tilde{D}_i^\top p) \frac{2\phi}{w_0 \beta_i} p + O(\epsilon^2). \end{aligned}$$

Equating LHS and RHS, canceling the zeroth-order terms, premultiplying by $\nabla_p D^* = -(\Lambda^*)^{-1}$, and using the identity $\frac{\alpha_i(1-\beta_i)}{(2-\beta_i)\beta_i} = \frac{1}{2\phi}$ (from the baseline ϕ condition (32)) simplifies the system to:

$$\frac{\tilde{D}_i + \tilde{q}_{0,i}}{\beta_i} - D_\gamma^* \tilde{\gamma}_i + \frac{2 - \beta_i}{\alpha_i} \frac{\tilde{D}_i^\top p}{w_0} D^* = -\tilde{D}_i - \frac{\beta_i}{1-\beta_i} (\nabla_p \tilde{D}_{-i}) p, \tag{47}$$

where we used $D_\gamma^*(p) = -\nabla_p D^* p_\gamma^*$. Note that $p^\top D_\gamma^* = 0$ due to the zero-degree homogeneity of the budget constraint with respect to γ.

Step 2: Ansatz and Verification. We conjecture the solution form (23). Note that $\nabla_p (D^*) p = -D^*$ (homogeneity of degree -1) and $\nabla_p (D_\gamma^*) p = -D_\gamma^*$. Applying the gradient to the ansatz for agents $j \neq i$, we find:

$$(\nabla_p \tilde{D}_{-i}) p = -\left(\sum_{j \neq i} \eta_{1,j}\right) D_\gamma^*.$$

The expenditure term is $p^\top \tilde{D}_i = (\eta_{0,i} + \eta_{2,i}\frac{w_0}{2\phi})(p^\top \tilde{q}_{0,i})$, since $p^\top D^*_\gamma = 0$.

Step 3: Coefficient Matching. We substitute the ansatz into (47) and match terms for the basis vectors $\tilde{q}_{0,i}$, D^*_γ, and D^*.

Matching $\tilde{q}_{0,i}$: The terms proportional to $\tilde{q}_{0,i}$ require:

$$\frac{\eta_{0,i}+1}{\beta_i} = -\eta_{0,i} \implies \eta_{0,i} = -\frac{1}{1+\beta_i}.$$

Matching D^*_γ: Let $K = \sum_j \eta_{1,j}$. Then $\sum_{j\neq i}\eta_{1,j} = K - \eta_{1,i}$. The condition becomes:

$$\frac{\eta_{1,i}}{\beta_i} - \tilde{\gamma}_i = -\eta_{1,i} + \frac{\beta_i}{1-\beta_i}(K - \eta_{1,i}).$$

Rearranging to isolate $\eta_{1,i}$:

$$\eta_{1,i}\underbrace{\left[\frac{1}{\beta_i} + 1 + \frac{\beta_i}{1-\beta_i}\right]}_{=\frac{1}{\beta_i(1-\beta_i)}} = \frac{\beta_i}{1-\beta_i}K + \tilde{\gamma}_i.$$

Solving for $\eta_{1,i}$ gives $\eta_{1,i} = \beta_i^2 K + \beta_i(1-\beta_i)\tilde{\gamma}_i$. Summing over all i to determine K:

$$K = K\sum \beta_i^2 + \sum \beta_i(1-\beta_i)\tilde{\gamma}_i \implies K(1-\sum\beta_i^2) = \sum \beta_i(1-\beta_i)\tilde{\gamma}_i.$$

Since $\sum \beta_i = 1$, we have $1 - \sum\beta_i^2 = \sum\beta_i(1-\beta_i)$. Thus, $K = \Gamma_{\omega^\gamma}[\tilde{\gamma}]$, where $\Gamma_{\omega^\gamma}[\tilde{\gamma}]$ is defined in (24). Substituting K back yields $\eta_{1,i} = \beta_i(1-\beta_i)\tilde{\gamma}_i + \beta_i^2\Gamma_{\omega^\gamma}[\tilde{\gamma}]$.

Matching $(p^\top\tilde{q}_{0,i})D^*$: The condition for the wealth effect term is:

$$\frac{\eta_{2,i}}{\beta_i} + \frac{2-\beta_i}{\alpha_i w_0}\left(\eta_{0,i} + \eta_{2,i}\frac{w_0}{2\phi}\right) = -\eta_{2,i}.$$

Substituting $\eta_{0,i} = -1/(1+\beta_i)$ and solving for $\eta_{2,i}$ yields the expression in the Theorem. ∎

H Proof of Lemma 3

Lemma 3 (Linear Approximation of Granular Wedges)**.** *Let heterogeneity be measured by $\epsilon = \|(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{q}}_0)\|$, where $\alpha_i = \bar{\alpha} + \tilde{\alpha}_i$ with $\bar{\alpha} = 1/L$ and $\sum_i \tilde{\alpha}_i = 0$, and $\sum_i \tilde{q}_{0,i} = 0$. The granular*

holdings wedge $\Gamma_{\omega^q}[\tilde{q}_0]$ *satisfies:*

$$\Gamma_{\omega^q}[\tilde{q}_0] = -\mathcal{K}(\bar{\alpha}, \phi)\, \Gamma_\alpha[\tilde{q}_0] + O(\epsilon^3),$$

where $\Gamma_\alpha[\tilde{q}_0] = \sum_i \alpha_i \tilde{q}_{0,i}$ *is the size-weighted holdings deviation. The coefficient* $\mathcal{K}(\bar{\alpha}, \phi)$ *is strictly positive and given by:*

$$\mathcal{K}(\bar{\alpha}, \phi) = \frac{\phi(1-\bar{\beta})}{L(1+\bar{\beta})(\bar{\alpha}\phi + 1 - \bar{\beta})},$$

where $\bar{\beta}$ *is the symmetric equilibrium scaling constant satisfying* $\bar{\beta} = \bar{\alpha}\phi + 1 - \sqrt{(\bar{\alpha}\phi)^2 + 1}$.

Proof. Let $S_q \equiv \sum_k (1+\beta_k)^{-1}$. With normalized weights $\omega_i^q = (1+\beta_i)^{-1}/S_q$, the granular holdings wedge is:

$$\Gamma_{\omega^q}[\tilde{q}_0] = \sum_{i=1}^{L} \omega_i^q \tilde{q}_{0,i} = \frac{1}{S_q} \sum_{i=1}^{L} \frac{\tilde{q}_{0,i}}{1+\beta(\alpha_i)},$$

where $\beta(\alpha)$ is the smooth function defined by the equilibrium condition. Define the weighting function $h(\alpha) \equiv (1+\beta(\alpha))^{-1}$ and perform a Taylor expansion around the symmetric share $\bar{\alpha}$. Since $\alpha_i = \bar{\alpha} + \tilde{\alpha}_i$, we have:

$$h(\alpha_i) = h(\bar{\alpha}) + h'(\bar{\alpha})\tilde{\alpha}_i + O(\epsilon^2).$$

For the normalization factor, $S_q = \sum_k h(\alpha_k) = L \cdot h(\bar{\alpha}) + O(\epsilon^2)$ since $\sum_k \tilde{\alpha}_k = 0$.

Consider the numerator $\sum_i h(\alpha_i)\tilde{q}_{0,i}$. Substituting the expansion yields:

$$\sum_{i=1}^{L} h(\alpha_i)\tilde{q}_{0,i} = \sum_{i=1}^{L} \left(h(\bar{\alpha}) + h'(\bar{\alpha})\tilde{\alpha}_i + O(\epsilon^2)\right) \tilde{q}_{0,i}.$$

Distributing the sum yields three terms. The first term, $h(\bar{\alpha}) \sum_i \tilde{q}_{0,i}$, vanishes by the constraint $\sum_i \tilde{q}_{0,i} = 0$. The second term is $h'(\bar{\alpha}) \sum_i \tilde{\alpha}_i \tilde{q}_{0,i}$. Note that since $\sum_i \tilde{q}_{0,i} = 0$, we have $\sum_i \alpha_i \tilde{q}_{0,i} = \sum_i (\bar{\alpha} + \tilde{\alpha}_i)\tilde{q}_{0,i} = \sum_i \tilde{\alpha}_i \tilde{q}_{0,i} = \Gamma_\alpha[\tilde{q}_0]$. The residual involves $O(\epsilon^2)$ terms multiplied by $O(\epsilon)$ perturbations, yielding $O(\epsilon^3)$.

Thus, $\sum_i h(\alpha_i)\tilde{q}_{0,i} = h'(\bar{\alpha})\Gamma_\alpha[\tilde{q}_0] + O(\epsilon^3)$. Dividing by $S_q = L \cdot h(\bar{\alpha}) + O(\epsilon^2)$:

$$\Gamma_{\omega^q}[\tilde{q}_0] = \frac{h'(\bar{\alpha})}{L \cdot h(\bar{\alpha})} \Gamma_\alpha[\tilde{q}_0] + O(\epsilon^3).$$

Computing the derivative $h'(\bar{\alpha}) = -(1+\bar{\beta})^{-2}\beta'(\bar{\alpha})$ and using $h(\bar{\alpha}) = (1+\bar{\beta})^{-1}$, the coefficient

becomes $-\mathcal{K}(\bar{\alpha}, \phi)$ where:

$$\mathcal{K}(\bar{\alpha}, \phi) = \frac{-h'(\bar{\alpha})}{L \cdot h(\bar{\alpha})} = \frac{\beta'(\bar{\alpha})}{L(1+\bar{\beta})}.$$

Substituting the explicit form of $\beta'(\bar{\alpha})$ derived in the main text yields the stated expression for $\mathcal{K}(\bar{\alpha}, \phi)$. ■

I Proof of Lemma 4

Lemma 4 (Linear Approximation of GIV)**.** *Under the same conditions as Lemma 3, the Granular Instrumental Variable satisfies:*

$$\text{GIV} = \mathcal{K}_{\text{GIV}}(\bar{\alpha}, \phi)\, \Gamma_{\alpha}[\tilde{q}_0] + O(\epsilon^3),$$

where $\mathcal{K}_{\text{GIV}}(\bar{\alpha}, \phi) > 0$*. Consequently, the GIV and the granular holdings wedge* $\Gamma_{\omega^q}[\tilde{q}_0]$ *are negatively related:*

$$\text{GIV} = -\frac{\mathcal{K}_{\text{GIV}}}{\mathcal{K}}\, \Gamma_{\omega^q}[\tilde{q}_0] + O(\epsilon^3).$$

Proof. The GIV is defined as:

$$\text{GIV} = \sum_{i=1}^{L} \frac{\alpha_i \tilde{q}_{0,i}}{1+\beta(\alpha_i)} - \frac{1}{L}\sum_{i=1}^{L} \frac{\tilde{q}_{0,i}}{1+\beta(\alpha_i)}.$$

Define $g(\alpha) \equiv \alpha/(1+\beta(\alpha))$ and $h(\alpha) \equiv 1/(1+\beta(\alpha))$. Note that $g(\alpha) = \alpha h(\alpha)$. Expanding around the symmetric share $\bar{\alpha} = 1/L$:

$$\begin{aligned} g(\alpha_i) &= g(\bar{\alpha}) + g'(\bar{\alpha})\tilde{\alpha}_i + O(\epsilon^2), \\ h(\alpha_i) &= h(\bar{\alpha}) + h'(\bar{\alpha})\tilde{\alpha}_i + O(\epsilon^2). \end{aligned}$$

Substituting these expansions into the GIV definition and using the condition $\sum_i \tilde{q}_{0,i} = 0$:

$$\sum_i g(\alpha_i)\tilde{q}_{0,i} = g(\bar{\alpha}) \underbrace{\sum_i \tilde{q}_{0,i}}_{0} + g'(\bar{\alpha}) \sum_i \tilde{\alpha}_i \tilde{q}_{0,i} + O(\epsilon^3) = g'(\bar{\alpha})\Gamma_{\alpha}[\tilde{q}_0] + O(\epsilon^3),$$

$$\frac{1}{L}\sum_i h(\alpha_i)\tilde{q}_{0,i} = \frac{h(\bar{\alpha})}{L} \underbrace{\sum_i \tilde{q}_{0,i}}_{0} + \frac{h'(\bar{\alpha})}{L} \sum_i \tilde{\alpha}_i \tilde{q}_{0,i} + O(\epsilon^3) = \frac{h'(\bar{\alpha})}{L}\Gamma_{\alpha}[\tilde{q}_0] + O(\epsilon^3).$$

Subtracting the two terms yields:

$$\text{GIV} = \left(g'(\bar{\alpha}) - \frac{h'(\bar{\alpha})}{L} \right) \Gamma_\alpha[\tilde{q}_0] + O(\epsilon^3).$$

Using the identity $g'(\alpha) = h(\alpha) + \alpha h'(\alpha)$ evaluated at $\bar{\alpha}$:

$$g'(\bar{\alpha}) - \frac{h'(\bar{\alpha})}{L} = h(\bar{\alpha}) + \bar{\alpha} h'(\bar{\alpha}) - \frac{h'(\bar{\alpha})}{L} = h(\bar{\alpha}) + h'(\bar{\alpha}) \left(\bar{\alpha} - \frac{1}{L} \right).$$

Since $\bar{\alpha} = 1/L$, the term involving $h'(\bar{\alpha})$ vanishes exactly. Thus:

$$\mathcal{K}_{\text{GIV}}(\bar{\alpha}, \phi) = h(\bar{\alpha}) = \frac{1}{1 + \beta(\bar{\alpha})} > 0.$$

The negative relationship with $\Gamma_{\omega^q}[\tilde{q}_0]$ follows from Lemma 3. ■

References

Viral Acharya and Alberto Bisin. Counterparty risk externality: Centralized versus over-the-counter markets. Journal of Economic Theory, 149:153–182, 2014.

Viral V. Acharya and Lasse Heje Pedersen. Asset pricing with liquidity risk. Journal of Financial Economics, 77(2):375–410, 2005.

Tobias Adrian, Erkko Etula, and Tyler Muir. Financial intermediaries and the cross-section of asset returns. The Journal of Finance, 69(6):2557–2596, 2014.

Franklin Allen. Do financial institutions matter? The Journal of Finance, 56(4):1165–1175, 2001.

Robert Almgren, Chee Thum, Emmanuel Hauptmann, and Hong Li. Direct estimation of equity market impact. Risk, 18(7):58–62, 2005.

Yakov Amihud and Haim Mendelson. Asset pricing and the bid-ask spread. Journal of financial Economics, 17(2):223–249, 1986.

Lawrence M Ausubel. Insider trading in a rational expectations economy. The American Economic Review, pages 1022–1041, 1990a.

Lawrence M Ausubel. Partially-revealing rational expectations equilibrium in a competitive economy. Journal of Economic Theory, 50(1):93–126, 1990b.

Lawrence M Ausubel, Peter Cramton, Marek Pycia, Marzena Rostek, and Marek Weretka. Demand reduction and inefficiency in multi-unit auctions. The Review of Economic Studies, 81(4):1366–1400, 2014.

Efstathios Avdis and Sergei Glebkin. Chile. Working Paper, 2023.

Ana Babus and Péter Kondor. Trading and information diffusion in over-the-counter markets. Econometrica, 86(5):1727–1769, 2018.

Mark Bagnoli, S Viswanathan, and Craig Holden. On the existence of linear equilibria in models of market making. Mathematical Finance, 11(1):1–31, 2001.

Gadi Barlevy and Pietro Veronesi. Rational panics and stock market crashes. Journal of Economic Theory, 110(2):234–263, 2003.

Suleyman Basak and Anna Pavlova. Asset prices and institutional investors. American Economic Review, 103(5):1728–1758, 2013.

Itzhak Ben-David, Francesco Franzoni, Rabih Moussawi, and John Sedunov. The granular nature of large institutional investors. Management Science, 67(11):6629–6659, 2021.

Dirk Bergemann, Tibor Heumann, and Stephen Morris. Information and volatility. Journal of Economic Theory, 158:427–465, 2015.

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

Markus K Brunnermeier and Lasse Heje Pedersen. Market liquidity and funding liquidity. The review of financial studies, 22(6):2201–2238, 2009.

Georgy Chabakauri, Kathy Yuan, and Konstantinos E Zachariadis. Multi-asset noisy rational expectations equilibrium with contingent claims. LSE working paper, 2017.

Louis KC Chan and Josef Lakonishok. The behavior of stock prices around institutional trades. The Journal of Finance, 50(4):1147–1174, 1995.

Chiraphol N Chiyachantana, Pankaj K Jain, Christine Jiang, and Robert A Wood. International evidence on institutional trading behavior and price impact. The Journal of Finance, 59(2): 869–898, 2004.

Kee H Chung and Sahn-Wook Huh. The noninformation cost of trading and its relative importance in asset pricing. The Review of Asset Pricing Studies, 6(2):261–302, 2016.

George M Constantinides. Capital market equilibrium with transaction costs. Journal of Political Economy, 94(4):842–862, 1986.

Efe Çötelioğlu, Francesco Franzoni, and Alberto Plazzi. What constrains liquidity provision? evidence from institutional trades. Review of Finance, 25(2):485–517, 2021.

Joshua Coval and Erik Stafford. Asset fire sales (and purchases) in equity markets. Journal of Financial Economics, 86(2):479–512, 2007.

Eduardo Dávila and Anton Korinek. Pecuniary externalities in economies with financial frictions. The Review of Economic Studies, 85(1):352–395, 2018.

Songzi Du and Haoxiang Zhu. Bilateral trading in divisible double auctions. Journal of Economic Theory, 167:285–311, 2017.

Victor Duarte, Mahyar Kargar, Jiacui Li, and Dejanir Silva. Dissecting the aggregate market elasticity. Technical report, Working Paper, 2025.

Darrell Duffie, Nicolae Gârleanu, and Lasse Heje Pedersen. Over-the-counter markets. Econometrica, 73(6):1815–1847, 2005.

Larry G Epstein and Stanley E Zin. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: A theoretical framework. Econometrica (1986-1998), 57(4): 937, 1989.

Andreas Fagereng, Matthieu Gomez, Emilien Gouin-Bonenfant, Martin Holm, Benjamin Moll, and Gisle Natvik. Asset-price redistribution. Journal of Political Economy, 133(11):3494–3549, 2025.

Xavier Gabaix. The granular origins of aggregate fluctuations. Econometrica, 79(3):733–772, 2011.

Xavier Gabaix and Ralph SJ Koijen. In search of the origins of financial fluctuations: The inelastic markets hypothesis. Technical report, National Bureau of Economic Research, 2021.

Xavier Gabaix and Ralph SJ Koijen. Granular instrumental variables. Journal of Political Economy, 132(7):2274–2303, 2024.

Jean Jaskold Gabszewicz and Jean-Philippe Vial. Oligopoly "a la cournot" in a general equilibrium analysis. Journal of economic theory, 4(3):381–400, 1972.

Gerard Gennotte and Hayne Leland. Market liquidity, hedging, and crashes. The American Economic Review, pages 999–1021, 1990.

Sergei Glebkin, Naveen Gondhi, and John Chi-Fong Kuong. Funding constraints and informational efficiency. The Review of Financial Studies, 34(9):4269–4322, 2021.

Sergei Glebkin, Semyon Malamud, and Alberto Teguia. Illiquidity and higher cumulants. The Review of Financial Studies, 36(5):2131–2173, 2023a.

Sergei Glebkin, Semyon Malamud, and Alberto Teguia. Strategic trading with wealth effects. 2023b.

John M Griffin, Jeffrey H Harris, and Selim Topaloglu. The dynamics of institutional and individual trading. The Journal of Finance, 58(6):2285–2320, 2003.

Zhiguo He, Bryan Kelly, and Asaf Manela. Intermediary asset pricing: New evidence from many asset classes. Journal of Financial Economics, 126(1):1–35, 2017.

Zhiguo He, Stefan Nagel, and Zhaogang Song. Treasury inconvenience yields during the covid-19 crisis. Journal of Financial Economics, 143(1):57–79, 2022.

Kenneth L Judd. Numerical methods in economics. MIT press, 1998.

Marcin Kacperczyk, Jaromir Nosal, and Savitar Sundaresan. Market power and price informativeness. Review of Economic Studies, 92(3):1955–1986, 2025.

Mahyar Kargar, Juan Passadore, Dejanir Silva, and Yucheng Yang. Liquidity and risk in otc markets: A theory of asset pricing and portfolio flows. Technical report, SSRN Working Paper No. 3731019, 2025.

Paul D Klemperer and Margaret A Meyer. Supply function equilibria in oligopoly under uncertainty. Econometrica: Journal of the Econometric Society, pages 1243–1277, 1989.

Leonid Kogan and Indrajit Mitra. Near-rational equilibria in heterogeneous-agent models: A verification method. The Review of Financial Studies, page hhaf030, 2025.

Leonid Kogan and Raman Uppal. Risk aversion and optimal portfolio policies in partial and general equilibrium economies, 2001.

Ralph S. J. Koijen and Motohiro Yogo. A demand system approach to asset pricing. Journal of Political Economy, 127(4):1475–1515, 2019. doi: 10.1086/701683. URL https://doi.org/10.1086/701683.

Arvind Krishnamurthy and Annette Vissing-Jorgensen. The aggregate demand for treasury debt. Journal of Political Economy, 120(2):233–267, 2012.

Arvind Krishnamurthy and Annette Vissing-Jorgensen. Short-term debt and financial crises: What we can learn from us treasury supply. Annual Review of Financial Economics, 5: 311–340, 2013.

Albert S Kyle. Informed speculation with imperfect competition. The Review of Economic Studies, 56(3):317–355, 1989.

Albert S Kyle, Anna A Obizhaeva, and Yajun Wang. Smooth trading with overconfidence and market power. The Review of Economic Studies, 85(1):611–662, 2017.

Jeongmin Lee and Albert S Kyle. When are financial markets perfectly competitive? 2018.

Francis A Longstaff. Portfolio claustrophobia: Asset pricing in markets with illiquid assets. American Economic Review, 99(4):1119–1144, 2009.

Semyon Malamud. Noisy arrow-debreu equilibria. Available at SSRN 2572881, 2015.

Semyon Malamud and Marzena Rostek. Decentralized exchange. American Economic Review, 107(11):3320–62, 2017.

Massimo Massa, David Schumacher, and Yan Wang. Who is afraid of blackrock? The Review of Financial Studies, 34(4):1987–2044, 2021.

Daniel Neuhann and Michael Sockin. Financial market concentration and misallocation. Journal of Financial Economics, 159:103875, 2024.

Daniel Neuhann, Mahyar Sefidgaran, and Michael Sockin. Portfolio regulation of large financial institutions. Available at SSRN 3971043, 2021.

Dömötör Pálvölgyi and Gyuri Venter. Multiple equilibria in noisy rational expectations economies. Available at SSRN 2524105, 2015.

Stavros Panageas. The implications of heterogeneity and inequality for asset pricing. Foundations and TrendsŴ in Accounting, 12(3):199–275, 2020.

Joel Peress. Wealth, information acquisition, and portfolio choice. The Review of Financial Studies, 17(3):879–914, 2003.

Dmitrii Pugachev. How do hedge funds affect stock market quality? evidence from hedge fund terminations. Technical report, INSEAD Working Paper, 2024.

Marzena Rostek and Marek Weretka. Price inference in small markets. Econometrica, 80(2): 687–711, 2012.

Marzena Rostek and Marek Weretka. Information and strategic behavior. Journal of Economic Theory, 158:536–557, 2015.

Marzena Rostek and Ji Hee Yoon. Equilibrium theory of financial markets: Recent developments. Journal of Economic Literature, 1, 2025.

Jeff Tjornehoj. Exploring fund industry concentration: The good, the bad, and the unknown. Technical report, Broadridge Financial Solutions, 2018.

Harald Uhlig. A toolkit for analyzing nonlinear dynamic stochastic models easily. In John B. Taylor and Michael Woodford, editors, Handbook of Macroeconomics, volume 1, pages 1175–1236. Elsevier, 1999.

Hal R Varian. Microeconomic analysis. WW Norton & Company, 1992.

Dimitri Vayanos. Strategic trading and welfare in a dynamic market. The Review of Economic Studies, 66(2):219–254, 1999.

Dimitri Vayanos and Jean-Luc Vila. Equilibrium interest rate and liquidity premium with transaction costs. Economic theory, 13(3):509–539, 1999.

Dimitri Vayanos and Jiang Wang. Theories of liquidity. Foundations and Trends® in Finance, 6(4):221–317, 2012.

Xavier Vives. Aggregation of information in large cournot markets. Econometrica: Journal of the Econometric Society, pages 851–876, 1988.

Xavier Vives. Strategic supply function competition with private information. Econometrica, 79(6):1919–1966, 2011.

Robert Wilson. Auctions of shares. The Quarterly Journal of Economics, 93(4):675–689, 1979.

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

Internet Appendix

for

"Granularity in Asset Markets"

IA.1 Is a large market competitive?

In this section, we study equilibrium outcomes in an economy with a large number of large investors, focusing on the non-competitive equilibrium in the limit as $L \to \infty$. Our objective is to determine whether aggregate quantities such as expected returns, return volatility, and illiquidity converge to their competitive counterparts, or whether market power persists even in very large markets.

Proposition 3 implies that aggregate outcomes in a large non-competitive economy differ from those in the competitive benchmark whenever

$$\phi(\infty) \equiv \lim_{L\to\infty} \phi(L) > 1,$$

where $\phi(L)$ denotes the equilibrium wedge defined in (16) for a non-competitive economy with L large investors. The following proposition provides sufficient conditions under which this wedge remains strictly positive in the large-market limit, as well as conditions under which it vanishes.

Proposition 9. *Expected returns, return volatility, and illiquidity in a large non-competitive market differ from those in a competitive market if and only if the large market exhibits strictly positive concentration, as measured by the Herfindahl–Hirschman Index (HHI) of the wealth distribution.*

Formally, let $\mathrm{HHI}(L) \equiv \sum_{i=1}^{L} \alpha_i^2$ *denote the HHI of wealth shares, and define* $\mathrm{HHI}(\infty) \equiv \lim_{L\to\infty} \mathrm{HHI}(L)$. *Assume that both* $\mathrm{HHI}(\infty)$ *and* $\phi(\infty)$ *exist. If* $\mathrm{HHI}(\infty) > 0$, *then* $\phi(\infty) > 1$.

Conversely, if $\mathrm{HHI}(\infty) = 0$*, then* $\phi(\infty) = 1$*. Moreover,*

$$\phi(\infty) \geq 1 + \frac{\mathrm{HHI}(\infty)}{1+\sqrt{2}}.$$

Proposition 9 provides new insight into the classic question of whether markets become perfectly competitive as the number of traders grows large (see, e.g., Lee and Kyle (2018) and references therein). The key distinction relative to the existing literature is that we explicitly account for wealth heterogeneity and wealth effects. These features imply that market power need not vanish with market size if the wealth distribution remains sufficiently concentrated.

The proposition also establishes a tight link between the wedge separating competitive and non-competitive equilibria and a standard measure of market concentration. By Proposition 3, this wedge is measured by $\phi - 1$. Aggregate quantities in the competitive economy are independent of both L and the wealth distribution $\{\alpha_i\}_i$, whereas aggregate quantities in the non-competitive economy are scaled by ϕ: expected returns and return volatility are multiplied by ϕ, while illiquidity is divided by ϕ.

When $\mathrm{HHI}(\infty) > 0$, large investors retain market power even as their number becomes arbitrarily large, and the wedge $\phi(\infty) - 1$ remains strictly positive. Conversely, when $\mathrm{HHI}(\infty) = 0$, concentration vanishes, market power disappears, and the non-competitive equilibrium coincides with the competitive benchmark. Thus, the HHI, a measure commonly used by the FTC to evaluate mergers, directly captures the degree of non-competitive behavior in the market. The lower bound in Proposition 9 further shows that this wedge is quantitatively meaningful, scaling at least linearly with concentration.

We conclude with several illustrative examples.

Example 1 (Homogeneous large investors)**.** *Suppose all large investors have equal wealth shares,* $\alpha_i = 1/L$ *for all* i*. Then*

$$\mathrm{HHI}(L) = \sum_{i=1}^{L} \frac{1}{L^2} = \frac{1}{L} \to 0 \quad \text{as } L \to \infty.$$

Hence, the wedge between competitive and large non-competitive markets vanishes, highlighting the central role of wealth heterogeneity in sustaining market power.

Example 2 (Power-law wealth distribution)**.** *Suppose wealth shares follow a power law,*

$$\alpha(i) = \frac{1}{\zeta(\psi)} i^{-\psi}, \quad \psi > 1,$$

where $\zeta(\cdot)$ denotes the Riemann zeta function. Then

$$\mathrm{HHI}(\infty) = \frac{1}{\zeta(\psi)^2} \sum_{i=1}^{\infty} i^{-2\psi} = \frac{\zeta(2\psi)}{\zeta(\psi)^2} > 0.$$

The wedge between competitive and large non-competitive markets therefore remains strictly positive.

Example 3 (At least one granular large investor)**.** *Suppose that, in the limit $L \to \infty$, at least one large investor retains a strictly positive wealth share. Let this large investor be indexed by $i = 1$, with limiting share $\alpha_1 > 0$. Then*

$$\mathrm{HHI}(\infty) \geq \alpha_1^2 > 0,$$

and the wedge between competitive and large non-competitive markets remains strictly positive.

IA.1.1 Proof of Proposition 9

Proof of Proposition 9. Note that we can rewrite the left-hand side of (16) as

$$\phi + \sum_i \left(1 - \sqrt{(\alpha_i\phi)^2 + 1}\right) = \phi - \sum_i \frac{\alpha_i^2\phi^2}{1 + \sqrt{(\alpha_i\phi)^2 + 1}}. \tag{IA.1}$$

This follows from multiplying and dividing each term in the sum by $1 + \sqrt{(\alpha_i\phi)^2 + 1}$ and applying the identity $(a-b)(a+b) = a^2 - b^2$.

Since $\alpha_i < 1$ for all i, we have

$$\phi - \sum_i \frac{\alpha_i^2\phi^2}{1 + \sqrt{(\alpha_i\phi)^2 + 1}} \leq \phi - \sum_i \frac{\alpha_i^2\phi^2}{1 + \sqrt{\phi^2 + 1}} = \phi - \frac{\mathrm{HHI}\,\phi^2}{1 + \sqrt{\phi^2 + 1}}.$$

Moreover, since $\frac{\mathrm{HHI}\,\phi^2}{1+\sqrt{\phi^2+1}}$ is strictly increasing in ϕ, and in equilibrium $\phi \geq 1$ (see the proof of Proposition 3), we obtain

$$\phi - \frac{\mathrm{HHI}\,\phi^2}{1 + \sqrt{\phi^2 + 1}} \leq \phi - \frac{\mathrm{HHI}}{1 + \sqrt{2}}. \tag{IA.2}$$

Recall that $\phi(L)$ solves the equation

$$\phi - \sum_i \frac{\alpha_i^2\phi^2}{1 + \sqrt{(\alpha_i\phi)^2 + 1}} = 1.$$

From the inequality (IA.2), it follows that $\phi(L)$ is greater than the solution to $\phi - \frac{\text{HHI}}{1+\sqrt{2}} = 1$. Hence,

$$\phi(L) \geq 1 + \frac{\text{HHI}(L)}{1+\sqrt{2}}.$$

Taking the limit as $L \to \infty$, we obtain

$$\phi(\infty) \geq 1 + \frac{\text{HHI}(\infty)}{1+\sqrt{2}}.$$

Thus, if $\text{HHI}(\infty) > 0$, it follows that $\phi(\infty) > 1$.

We now turn to deriving an upper bound for ϕ. Consider the right-hand side of (IA.1), and note that since $\alpha_i \geq 0$ for all i, we can write

$$\phi - \sum_i \frac{\alpha_i^2 \phi^2}{1 + \sqrt{(\alpha_i \phi)^2 + 1}} \geq \phi - \sum_i \frac{\alpha_i^2 \phi^2}{2} = \phi - \phi^2 \frac{\text{HHI}}{2}.$$

From this inequality, it follows that $\phi(L)$ is less than the smaller solution to $\phi - \phi^2 \frac{\text{HHI}}{2} = 1$. Denote this solution by $\phi^*(L)$. Hence,

$$1 \leq \phi(\infty) \leq \phi^*(\infty).$$

The left-hand inequality follows because $\phi(L) \geq 1$ in equilibrium.

Since $\phi^*(\infty) = 1$ when $\text{HHI}(\infty) = 0$, we conclude that if $\text{HHI}(\infty) = 0$, then $\phi(\infty) = 1$.

■

IA.2 Welfare

Investigating welfare requires endogenizing the behavior of all market participants, including the rest of the market. In this section, we extend the model to incorporate price-elastic, competitive traders (henceforth, small traders). We assume that the representative small trader holds a fraction α_S of total wealth and is endowed with holdings q_0 of the traded assets.[20] The theorem below characterizes the equilibrium in this extended setup.

Theorem 3. *The equilibrium demand of the competitive trader is given by:*

$$D_S(p) = \frac{\alpha_S w_0 + p^\top q_0}{w_0} D^*(p) - q_0.$$

[20] Below, we show how such a representative trader can be "disaggregated."

Here, $D^*(\cdot)$ is defined implicitly by $I^*(D^*(p)) \equiv p$, where

$$I^*(x) = \frac{w_0}{2} \frac{E[(x^\top \delta)^{-\gamma} \delta]}{E[(x^\top \delta)^{1-\gamma}]}.$$

The equilibrium inverse demand of large investor i is given by:

$$I_i^*(q) = \frac{\beta_i}{\phi} I^*(q).$$

The coefficients β_i are given by:

$$\beta_i = \alpha_i \phi + 1 - \sqrt{(\alpha_i \phi)^2 + 1},$$

and the constant ϕ solves:

$$\sum_i \beta_i = 1 - \alpha_S \phi.$$

The result in Theorem 3 highlights a notable asymmetry in how initial endowments affect market participants. For the competitive trader, the demand function $D_S(p)$ is explicitly modified by the endowment q_0. The term $\alpha_S w_0 + p^\top q_0$ represents the trader's total wealth, comprising initial cash and the market value of their asset holdings. As price takers, they treat p as given, and their demand simply reflects the standard portfolio choice problem with this augmented wealth budget.

Note that the competitive trader's demand fits the scale-symmetric structure—where demand is strictly proportional to the representative demand $D^*(p)$—only in the special case where $q_0 = 0$. In the general case, the endowment introduces an additive deviation. Despite this, surprisingly, the large investors' equilibrium demands remain structurally unchanged compared to the baseline economy (where all traders hold no assets).

The intuition for the large investors' invariance stems from the *ex-post* nature of their optimization. Since large investors do not observe the realization of the aggregate endowment q_0 before submitting their demand schedules, they must optimize pointwise for every possible realization of the residual supply curve. In equilibrium, this strategic requirement causes the specific level of the price-taker's endowment to wash out of the large investors' first-order conditions. Consequently, their strategy is identical to the case where the competitive sector enters with zero initial assets ($q_0 = 0$).

IA.2.1 Welfare Analysis

To analyze welfare implications, we disaggregate the competitive trader sector. Suppose there are M classes of competitive traders indexed by j, each holding a fraction α_j of total wealth and initial holdings q_0^j. The aggregation constraints are:

$$\sum_j \alpha_j = \alpha_S, \quad \sum_j q_0^j = q_0.$$

To facilitate the discussion, we introduce the following definition regarding the net trade position.

Definition 5. *A trader j is a "net buyer" if the value of assets held after trading, $p^\top x_j$, exceeds the value of their initial holdings, $p^\top q_0^j$.*

We summarize the main result of this section in the proposition below.

Proposition 10. *An increase in market concentration (higher ϕ) benefits all large investors, i.e., $\partial U^i/\partial\phi > 0$ for all i. Among the small traders, it benefits net buyers but harms net sellers. The aggregate welfare of the competitive sector increases in ϕ if and only if:*

$$\sum_j \frac{\alpha_j\phi - k_j}{\phi\left(\alpha_j\phi + k_j\right)} > 0.$$

Here, k_j denotes trader j's share of the aggregate endowment value, $k_j = \frac{p^\top q_0^j}{p^\top q_0}$.

An increase in ϕ—which reflects greater concentration and is associated with lower equilibrium prices—affects welfare through two distinct channels:

1. **Redistribution Channel:** An increase in ϕ lowers asset prices. This effectively transfers wealth from asset-rich agents (net sellers) to cash-rich agents (net buyers). Due to the concavity of the log-utility function, the marginal utility gain to the "poor" (cash-rich but asset-poor) buyers can mathematically outweigh the utility loss to the "rich" (endowment-holding) sellers.

2. **Liquidity Channel:** As established in the previous section, non-competitive large investors generate a market that is more liquid. This implies that a marginal decrease in price results in a larger volume of assets being transferred from sellers to buyers. This "volume effect" amplifies the redistribution channel described above.

These results connect to the literature on pecuniary externalities (e.g., Dávila and Korinek (2018)) and the redistributive effects of asset prices (e.g., Fagereng, Gomez, Gouin-Bonenfant, Holm, Moll, and Natvik (2025)). The positive welfare effect on LDs operates through what Dávila and Korinek (2018) term *distributive externalities*. However, while in their framework the externality typically arises from financial constraints, in our context it arises endogenously from market power.

Furthermore, our results provide a counter-narrative to the "asset price redistribution" mechanism highlighted by Fagereng et al. (2025), where high asset prices typically exacerbate inequality. Here, increased market concentration depresses prices, acting as a mean-reverting force that can partly undo the negative inequality effects associated with high asset valuations.

Example: Welfare Increasing in ϕ

Suppose we have two groups, each of size $M/2$:

- **Group 1 (Pure Buyers):** Hold only cash ($\alpha_1 > 0, k_1 = 0$).

$$\frac{\partial U_1}{\partial \phi} = \frac{1}{\phi}$$

- **Group 2 (Predominant Sellers):** Hold all assets ($k_2 = 2/M$) and remaining cash.

$$\frac{\partial U_2}{\partial \phi} = \frac{1}{\phi}\left(\frac{\alpha_2\phi - k_2}{\alpha_2\phi + k_2}\right)$$

The aggregate welfare change is:

$$\sum_j \frac{\partial U_j}{\partial \phi} \propto 1 + \frac{\alpha_2\phi - k_2}{\alpha_2\phi + k_2} = \frac{2\alpha_2\phi}{\alpha_2\phi + k_2} > 0$$

Thus, aggregate welfare strictly increases in ϕ due to the gains from trade realized by the buyers.

IA.2.2 Proof of Theorem 3

The price taker's demand is given by:

$$D_S(p) = \frac{\alpha_S w_0 + p^\top q_0}{w_0} D^*(p) - q_0.$$

Here, $D^*(\cdot)$ is defined by $I^*(D^*(p)) \equiv p$, where

$$I^*(x) = \frac{w_0}{2} \frac{E[(x^\top \delta)^{-\gamma} \delta]}{E[(x^\top \delta)^{1-\gamma}]}$$

Computing the Jacobian of $D_S(p)$, we obtain:

$$\nabla_p D_S(p) = \left(\alpha_S + \frac{p^\top q_0}{w_0} \right) \nabla_p D^*(p) + \frac{1}{w_0} D^*(p) q_0^\top$$

We conjecture that for large investor i, the demand is given by:

$$D_i(p) = \frac{1}{\phi} \beta_i \, D^*(p)$$

The normalization constant ϕ is chosen such that

$$\alpha_S \phi + \sum_i \beta_i = 1.$$

(This corresponds to ϕ in the economy with a continuum of small investors without holdings.)

Large Investor Optimization

The endowment q_0 is not known to large investors. They optimize pointwise for every realization of q_0.

The slope of the residual demand for trader i is:

$$\nabla_p D_{-i}(p) = \left(\alpha_S + \frac{p^\top q_0}{w_0} + \frac{\beta_{-i}}{\phi} \right) \nabla_p D^*(p) + \frac{1}{w_0} D^*(p) q_0^\top$$

We invert this using the Sherman–Morrison formula:

$$[\nabla_p D_{-i}(p)]^{-1} = \frac{1}{\alpha_S + \frac{p^\top q_0}{w_0} + \frac{\beta_{-i}}{\phi}} \left(I + \frac{p q_0^\top}{w_0 (\alpha_S + \frac{\beta_{-i}}{\phi})} \right) [\nabla_p D^*(p)]^{-1}$$

The First Order Condition (FOC) for large investor i is given by:

$$p + \Lambda_i(p) D_i(p) = \left(\alpha_i w_0 - D_i(p)^\top p \right) \frac{\mathbb{E}\left[\left(D_i(p)^\top \delta \right)^{-\gamma} \delta \right]}{\mathbb{E}\left[\left(D_i(p)^\top \delta \right)^{1-\gamma} \right]}.$$

Substituting terms, we have:

$$\begin{aligned}
\Lambda_i(p)D_i(p) &= -\frac{\beta_i}{\phi}[\nabla_p D_{-i}(p)]^{-1}D^*(p) \\
&= -\frac{\beta_i}{\phi\alpha_S + \phi\frac{p^\top q_0}{w_0} + \beta_{-i}}\left(I + \frac{pq_0^\top}{w_0(\alpha_S + \frac{\beta_{-i}}{\phi})}\right)\underbrace{[\nabla_p D^*(p)]^{-1}D^*(p)}_{=-p \text{ (by Euler)}} \\
&= \frac{\beta_i}{\phi\alpha_S + \beta_{-i}}p \\
&= \frac{\beta_i}{1-\beta_i}p.
\end{aligned}$$

Note: A crucial simplification occurs here: the term $p^\top q_0/w_0$ cancels out. Consequently, the FOC of large traders is the same as in the economy where small traders have no endowments. This is intuitive, as large investors do not know the realization of q_0.

Back to the FOC, now written in terms of inverse demand:

$$\left(1 + \frac{\beta_i}{1-\beta_i}\right) I_i(q_i) = \left(\alpha_i w_0 - q_i^\top I_i(q_i)\right)\frac{2}{w_0}I^*(q_i)$$

Note that $D_i(p) = \frac{1}{\phi}\,\beta_i\,D^*(p)$ implies $I_i(q) = \frac{\beta_i}{\phi}\,I^*(q)$ and therefore $q^\top I_i(q) = \frac{\beta_i w_0}{2\phi}$. So:

$$I_i^\top q_i = \beta_i \frac{w_0}{2\phi}$$

Also noting that $I_i(q_i) = \frac{\beta_i}{\phi}I^*(q_i)$, the expression to pin down β_i is:

$$\frac{\alpha_i\,(1-\beta_i)}{(2-\beta_i)\,\beta_i} = \frac{1}{2\phi}$$

This is the same expression as in the main text. Hence, the unique $\beta_i < 1$ solving the above is given by:

$$\beta_i = \alpha_i\phi + 1 - \sqrt{(\alpha_i\phi)^2 + 1}$$

The constant ϕ is determined by

$$\sum_i \beta_i = 1 - \alpha_S\phi.$$

IA.2.3 Proof of Proposition 10

We start with the market clearing condition:

$$\alpha_S D^*(p) + \frac{\sum_i \beta_i}{\phi} D^*(p) + \frac{p^\top q_0}{w_0} D^*(p) = q_0$$

Multiplying by $p^\top$, and accounting for the fact that $p^\top D^*(p) = w_0/2$ and $\phi\alpha_S + \sum_i \beta_i = 1$, we get:

$$p^\top q_0 = w_0/\phi$$

Substituting this back implies:

$$D^* = \frac{\phi q_0}{2} \implies \text{Aggregate allocation to large investors} = \frac{q_0}{2}(1 - \alpha_S \phi)$$

Price Takers' Consumption

Let W_j denote the total wealth of trader j at time 0. This wealth consists of their initial cash holding (fraction of w_0) and the market value of their endowment q_0^j.

$$W_j = \underbrace{\alpha_j w_0}_{\text{Cash}} + \underbrace{p^\top q_0^j}_{\text{Endowment Value}}$$

From the market clearing derivation, $p^\top q_0 = w_0/\phi$. Using the definition $k_j = \frac{p^\top q_0^j}{p^\top q_0}$, we express the value of trader j's endowment as:

$$p^\top q_0^j = k_j (p^\top q_0) = k_j \frac{w_0}{\phi}$$

Thus, total wealth is:

$$W_j = \alpha_j w_0 + k_j \frac{w_0}{\phi} = w_0 \left(\alpha_j + \frac{k_j}{\phi} \right)$$

Consumption

By the homotheticity of preferences, agents consume half their wealth at $t = 0$ and invest the other half.

- **Time-0:** $c_0^j = \frac{1}{2} W_j = \frac{w_0}{2} \left(\alpha_j + \frac{k_j}{\phi} \right)$.

- **Time-1:** Time-1 consumption is the payoff from the optimal asset portfolio x_j. The gross demand is proportional to representative demand:

$$x_j = \frac{W_j}{w_0} D^*(p) = \left[\frac{w_0(\alpha_j + k_j/\phi)}{w_0}\right]\left(\frac{\phi}{2} q_0\right) = \frac{1}{2}(\alpha_j \phi + k_j) q_0$$

  Therefore, $c_1^j = \frac{1}{2}(\alpha_j \phi + k_j)(q_0^\top \delta)$.

Price Takers' Welfare (U^j)

Substituting the consumption expressions into the log-utility function:

$$U^j = \log(c_0^j) + \log(c_1^j) = \log\left(\alpha_j + \frac{k_j}{\phi}\right) + \log(\alpha_j \phi + k_j) + \dots$$

where $\dots$ denotes terms independent of the wealth distribution. Differentiating with respect to ϕ:

$$\frac{\partial U_j}{\partial \phi} = \frac{\alpha_j \phi - k_j}{\phi(\alpha_j \phi + k_j)}$$

The utility of trader j is **increasing in** ϕ if and only if:

$$\alpha_j \phi - k_j > 0 \iff \alpha_j \phi > k_j$$

Interpretation: Net Buyers vs. Net Sellers

This condition corresponds exactly to being a net buyer. A trader is a "net buyer" if the value of assets held after trading ($p^\top x_j$) exceeds initial holdings ($p^\top q_0^j$).

$$\frac{1}{2} W_j > p^\top q_0^j \iff \frac{1}{2}(\alpha_j w_0 + p^\top q_0^j) > p^\top q_0^j \iff \alpha_j w_0 > p^\top q_0^j$$

Substituting equilibrium prices ($p^\top q_0^j = k_j w_0/\phi$) yields $\alpha_j \phi > k_j$.

Result: An increase in ϕ (market illiquidity/price drop) benefits net buyers and harms net sellers.

Aggregate Welfare of Price Takers

The aggregate welfare increases in ϕ if and only if:

$$\sum_j \frac{\partial U_j}{\partial \phi} = \sum_j \frac{\alpha_j \phi - k_j}{\phi(\alpha_j \phi + k_j)} > 0$$

Large Investors Welfare

For the Large Investors, we have:

- Time-0: $c_0^i = \alpha_i w_0 - \beta_i \frac{w_0}{2\phi}$
- Time-1: $c_1^i = \beta_i \frac{q_0}{2}$

The utility is:

$$U^i = \log\left(\alpha_i - \frac{\beta_i}{2\phi}\right) + \log(\beta_i) + \dots$$

Using the derived β_i, the derivative is:

$$\frac{\partial U^i}{\partial \phi} = \frac{1}{\phi}\left(\frac{1}{\sqrt{(\alpha_i \phi)^2 + 1}}\right) > 0.$$

IA.3 When is the competitive market less liquid?

Consider a market with L identical traders trading 1 risky asset. The utility after trading q units at price p, starting with initial endowment q_0 and wealth w_0, is given by $U(q+q_0, w_0 - pq)$. We assume U is increasing in both arguments.

Let $f(q, x)$ denote the marginal rate of substitution given quantity q and expenditure x:

$$f(q, x) = -\frac{U_q(q, x)}{U_w(q, x)}$$

Remark: We use $U_q(q, x)$ to denote the partial derivative of U with respect to its first argument, evaluated at (q, x).

Glebkin et al. (2023a) show that the inverse demand functions are characterized by:

1. **Strategic Demand:**

$$I(q) = f(q, qI(q)) + \frac{qI'(q)}{L-1}$$

2. **Competitive Demand:**

$$I^c(q) = f(q, qI^c(q))$$

Assumption 2 (Boundedness & Monotonicity)**.** *The competitive inverse demand $I^c(q)$ is strictly decreasing and bounded as $q \to \infty$.*

Given that the competitive demands satisfy natural monotonicity and boundedness properties, we focus on equilibria where the strategic demand $I(q)$ is strictly decreasing and bounded as well.

Assumption 3 (Stability)**.** *For all relevant q and x, the wealth effect satisfies the condition:*

$$1 - qf_x(q, x) > \epsilon > 0,$$

and the limit

$$F_\infty(p) \equiv \lim_{q\to\infty} f(q, qp)$$

exists and is finite for any $p > 0$. Furthermore, the derivative of F_∞ can be computed by interchanging limits:

$$\frac{d}{dp}F_\infty(p) = \lim_{q\to\infty} \frac{\partial}{\partial p} f(q, qp).$$

Remark: Assumption 3, combined with the condition that the direct effect of quantity on valuation is negative ($f_q + \frac{x}{q}f_x < 0$), implies that the competitive inverse demand is downward sloping.

IA.3.1 Main Result

Our main result is that, given the assumptions above, the strategic inverse demand is more elastic than the competitive demand for large enough quantities. Our proof proceeds in several steps. First, we show that the two demands approach the same value as $q \to \infty$. Second, due to demand reduction, the strategic demand $I(q)$ is always strictly below the competitive demand $I^c(q)$. Consequently, the gap between the two, $\Delta(q) = I^c(q) - I(q)$, is positive but vanishes to zero asymptotically. This implies that $\Delta'(q)$ must be negative for large q. It follows that $(I^c)'(q) < I'(q)$. Since both slopes are negative, the competitive demand has a more negative slope (i.e., is steeper). Thus, the competitive demand is less elastic for large enough quantities. The rest of this section formalizes these ideas.

Lemma 5. *The competitive and strategic inverse demands converge to the same limit as $q \to \infty$:*

$$\lim_{q\to\infty} I(q) = \lim_{q\to\infty} I^c(q)$$

Proof of Lemma 5. We proceed in several steps.

Step 1: Integral Representation of Strategic Demand

We first transform the differential equation for strategic demand into an integral equation. The ODE is given by:

$$I'(q) - \frac{L-1}{q} I(q) = -\frac{L-1}{q} f(q, qI(q))$$

Multiplying by the integrating factor $\mu(q) = q^{-(L-1)}$ and integrating from q to ∞ yields the general solution:

$$I(q) = kq^{L-1} + (L-1)q^{L-1} \int_q^\infty t^{-L} f(t, tI(t))\, dt$$

By Assumption 2, $I(q)$ is bounded as $q \to \infty$. Since $L > 1$, the term kq^{L-1} grows without bound unless $k = 0$. Therefore, we must have $k = 0$.

Substituting variables $t = q\xi$ (where $dt = qd\xi$), the equation simplifies to:

$$I(q) = (L-1) \int_1^\infty \xi^{-L} f(q\xi, q\xi I(q\xi))\, d\xi$$

Step 2: Characterizing the Limits

Since both $I(q)$ and $I^c(q)$ are monotonic and bounded (Assumption 2), their limits as $q \to \infty$ exist. Let us define:

$$C = \lim_{q\to\infty} I^c(q) \quad \text{and} \quad L^* = \lim_{q\to\infty} I(q)$$

We also define the asymptotic function $F_\infty(p)$ as the limit of the marginal rate of substitution when quantity becomes large:

$$F_\infty(p) \equiv \lim_{q\to\infty} f(q, qp)$$

Step 3: The Fixed Point Equation

First, consider the competitive limit. Taking the limit $q \to \infty$ on both sides of the implicit definition $I^c(q) = f(q, qI^c(q))$, we obtain:

$$C = F_\infty(C)$$

Next, consider the strategic limit. We take the limit $q \to \infty$ of the integral representation from

Step 1:

$$L^* = \lim_{q\to\infty}\left[(L-1)\int_1^\infty \xi^{-L} f(q\xi, q\xi I(q\xi))\, d\xi\right]$$

By the Dominated Convergence Theorem (valid due to the boundedness of $I(q)$ and integrability of ξ^{-L}), we can exchange the limit and the integral. Inside the integral, for any fixed $\xi \geq 1$, we have $I(q\xi) \to L^*$ as $q \to \infty$. Thus:

$$\lim_{q\to\infty} f(q\xi, q\xi I(q\xi)) = F_\infty(L^*)$$

Substituting this limit back into the integral equation:

$$L^* = (L-1)\int_1^\infty \xi^{-L} F_\infty(L^*)\, d\xi$$

Since $F_\infty(L^*)$ is constant with respect to ξ, it factors out:

$$L^* = F_\infty(L^*)\underbrace{\left[(L-1)\int_1^\infty \xi^{-L}\, d\xi\right]}_{=1}$$

$$L^* = F_\infty(L^*)$$

Step 4: Uniqueness of the Limit

We have established that both C and L^* are roots of the equation $g(p) = 0$, where $g(p) = p - F_\infty(p)$.

To determine if the root is unique, we examine the monotonicity of $g(p)$. The derivative is:

$$g'(p) = 1 - \frac{\partial F_\infty(p)}{\partial p}$$

Recall that $F_\infty(p)$ is the limit of $f(q, qp)$. By the chain rule, the derivative with respect to p (which enters via the expenditure argument $x = qp$) corresponds to the wealth effect scaled by q. Specifically, $\frac{\partial}{\partial p} f(q, qp) = q f_x(q, x)$.

By Assumption 3, we have $1 - q f_x > 0$ for all finite q. Taking the limit $q \to \infty$, this inequality implies:

$$1 - \frac{\partial F_\infty(p)}{\partial p} \geq \epsilon > 0$$

Consequently, $g(p)$ is non-decreasing.

Thus, the fixed point is unique:

$$L^* = C$$

■

With the lemma above at hand, we can now establish the main result.

Proposition 11. *For all q, the strategic inverse demand is strictly lower than the competitive one ($I(q) < I^c(q)$). Furthermore, there exists an unbounded set $\Theta \subset \mathbb{R}_+$ such that the competitive demand is steeper than the strategic demand (i.e., $|(I^c)'(q)| > |I'(q)|$) for all $q \in \Theta$. If I' and $(I^c)'$ are continuous, then Θ is open.*

Proof of Proposition 11. We proceed in several steps.

Step 1: Establishing demand reduction

Let $\Delta(q) = I^c(q) - I(q)$. We start with the defining equations for strategic and competitive demands:

1. $I(q) = f(q, qI(q)) + \frac{qI'(q)}{L-1}$
2. $I^c(q) = f(q, qI^c(q))$

Subtracting the first from the second:

$$I^c(q) - I(q) = f(q, qI^c(q)) - f(q, qI(q)) - \frac{qI'(q)}{L-1}$$

Consider the function $h(p) = f(q, qp)$. By the Mean Value Theorem with respect to the price argument p, there exists a value $\tilde{p}(q)$ lying strictly between $I(q)$ and $I^c(q)$ such that:

$$f(q, qI^c(q)) - f(q, qI(q)) = h'(\tilde{p})(I^c(q) - I(q))$$

Calculating the derivative $h'(p) = \frac{\partial}{\partial p} f(q, qp) = q f_x(q, q\tilde{p})$, we substitute this back:

$$\Delta(q) = [q f_x(q, q\tilde{p})]\, \Delta(q) - \frac{qI'(q)}{L-1}$$

Rearranging to collect the $\Delta(q)$ terms:

$$\Delta(q)\,[1 - q f_x(q, q\tilde{p})] = -\frac{qI'(q)}{L-1}$$

We analyze the sign of each term in the equation derived above:

1. **Right Hand Side:** Since the strategic demand is strictly decreasing ($I'(q) < 0$) and $L > 1$, the term $-\frac{qI'(q)}{L-1}$ is strictly positive.

2. **Left Hand Side Coefficient:** The term in brackets is $1 - qf_x(q, q\tilde{p})$. By the Stability Assumption (Assumption 3), $1 - qf_x(q, x) > 0$ for all x. Thus, the coefficient is strictly positive.

For the equation to hold, $\Delta(q)$ must be positive:

$$\Delta(q) > 0 \implies I^c(q) > I(q)$$

Step 2: Comparison of Slopes

We invoke Lemma 5, which establishes that both demands converge to the same limit:

$$\lim_{q\to\infty} I(q) = \lim_{q\to\infty} I^c(q) = L^*$$

This implies that the gap vanishes at infinity:

$$\lim_{q\to\infty} \Delta(q) = 0$$

Thus, we can represent it as

$$\Delta(q) = -\int_q^\infty \Delta'(t)\, dt > 0.$$

Thus, the set

$$\Theta = \{q > 0 : \Delta'(q) < 0\}$$

is unbounded and open (the latter because Δ' is continuous). Substituting the definition $\Delta'(q) = (I^c)'(q) - I'(q)$:

$$(I^c)'(q) - I'(q) < 0$$

$$(I^c)'(q) < I'(q)$$

Since both demand curves are strictly decreasing, their slopes are negative numbers. The inequality $(I^c)'(q) < I'(q)$ implies that the competitive slope is larger in magnitude:

$$|(I^c)'(q)| > |I'(q)|$$

for $q \in \Theta$. ■