Will ETFs drive Mutual Funds extinct?∗

Anna Helmke†

May 31, 2023
Click here for the latest version

Abstract

I study trade-offs investors make between exchange traded funds (ETFs) and open-end mutual funds in the presence of idiosyncratic liquidity risk and aggregate uncertainty. Based on a portfolio choice model, I show that ETFs and mutual funds provide liquidity at different maturities. Mutual funds (ETFs) are preferred by investors facing high (low) idiosyncratic liquidity risk and shorter (longer) investment horizons. In equilibrium, the pooling of investors into fund types based on their expected investment horizon directly emerges from the differential frictions of ETFs and mutual funds. Over the long-term, payoff complementarities in mutual funds dilute investors fund holdings and generate underperformance vis-à-vis ETFs. Yet, in the short-run, ETFs can be mispriced due to intermediary arbitrage constraints. The optimal size of the mutual fund sector relative to ETFs decreases in the illiquidity of portfolio assets but increases in the proportion of mutual fund shares held via retirement accounts.

∗I thank Itay Goldstein, Luke Taylor and Yao Zeng for providing me with invaluable guidance throughout the preparation of this work. I am grateful to seminar participants at the 20th Macro Finance Society Workshop, Finance Theory Group conference and Wharton for helpful suggestions and conversations.

†University of Pennsylvania, Wharton School, Email: ahelmke@wharton.upenn.edu. I gratefully acknowledge financial support from the Jacobs Levy Equity Management Center, Mack Institute for Innovation Management, and the Rodney White Center for Financial Research. All errors are my own.
1 Introduction

It doesn't matter if you own the Vanguard Total Stock Market mutual fund (VTSMX) or the Vanguard Total Stock Market ETF (VTI). I am vehicle agnostic.


Open-end mutual funds (MFs) and exchange traded funds (ETFs) are both professionally managed investment funds that pool money from many investors to purchase securities. In the growing index investment segment featuring funds that mechanically track a specific benchmark index, ETFs and MFs hold virtually identical portfolios. Yet, capital flows into index ETFs have outpaced flows into index MFs over the past decade, leading index ETFs to surpass their MF peers in terms of assets under management (AuM) for the first time in 2021.\(^1\) The broader shift from active to passive investing alone is insufficient to explain the 'age of ETFs'. The fundamental economic difference between the two fund types is that ETFs are exchange traded, just like stocks, whereas MFs are purchased or redeemed directly from the fund sponsor at the end-of-day fund net asset value (NAV) (see figure ??). How do these differences in how ETFs and open-end MFs are priced and traded in financial markets affect funds’ relative liquidity provision? And, what role do ETFs’ and MFs’ distinct payoff structures play in rational investors’ portfolio allocation decision between both fund types? Are open-end mutual funds just like dinosaurs headed for extinction? Or is there still a role for index MFs now that ETFs have been "discovered"?

To answer these questions, I propose a simple Diamond and Dybvig (1983)-style portfolio choice model featuring investors with heterogeneous liquidity risks and aggregate uncertainty. I establish that the non-portfolio differences between ETFs and MFs do matter for fund returns. Yet, ETFs and MFs optimally co-exist in equilibrium because they provide investors with different liquidity provision services. In particular, both fund types provide liquidity at different maturities.\(^2\) ETFs provide more liquidity over the long-term, whereas MFs provide more liquidity over the short-term. This result directly emerges from the differential frictions built into ETFs and MFs. Outflows from ETFs only have temporary effects on investors’ payoffs, whereas MF flows have persistent effects on investors’ payoffs. Over the short-term, ETFs can be mispriced relative to their fund NAV when balance sheet capacity constraints prevent authorized participants (APs) from providing liquidity in secondary ETF markets via primary market creations or redemptions. Due to the countercyclical nature of financial intermediaries’ balance sheet capacity, ETFs tend to trade at a discount in "bad" states when aggregate market liquidity is low. In contrast, MFs guarantee investors the ability to trade at

---

\(^1\)See figures 2a and 2b.

\(^2\)In this paper, liquidity provision refers to the difference between the price at which fund shares can be liquidated and the direct liquidation value of the underlying security portfolio. Accordingly, the notion of liquidity provision is defined by the relative payoffs of equity claims issued by different financial intermediaries (here ETFs and MFs).
the end-of-day fund NAV, irrespective of the state of the economy. However, payoff complementarities between MF investors make them less attractive at longer horizons. Long-term MF investors bear the transaction costs caused by early redemptions of other investors, leading to share dilution. Meanwhile, ETF prices eventually tend to converge back to the fund NAV once AP balance sheet capacity constraints subside.

I show that the differences in ETFs’ and MFs’ relative liquidity provision over time affect index investors’ portfolio allocation. Naturally, investors with higher liquidity risks, for example labor income risk, or shorter horizons value the short-term liquidity insurance provided by MFs. They are willing to sacrifice long-term expected returns to avoid the potential short-term mispricing in ETFs. Conversely, ETFs are preferred by investors with lower liquidity risks or longer-term horizons. Overall, individual investors’ tradeoffs between ETFs and MFs have implications for the aggregate fund market composition. Holding constant investor characteristics, the equilibrium market share of MFs decreases in the illiquidity of fund portfolio assets and increases in the fraction of MF assets held by sleepy investors, such as in retirement accounts. MFs with more illiquid portfolio holdings tend to pay larger transaction costs and generate more price impact in asset markets, and thus experience larger share dilutions from early fund redemptions. At the same time, sleepy investors who never redeem their shares early effectively subsidize MFs’ liquidity provision to investors with shorter-term horizons. In equilibrium, these sleepy MF investors suffer from the pooling of investors into fund types based on their liquidity needs and investment horizons.

The fundamental source of financial frictions in ETFs and MFs is illiquidity. The mechanism works as follows. On one side, the key financial friction associated with ETFs is the potential for relative mispricing. ETFs are traded intra-day in secondary markets at the prevailing market price. Secondary market transactions of ETF shares occur between investors (via a market maker). Investors do not directly trade with the ETF sponsor. Therefore, ETF trades are not directly linked to transactions in portfolio securities. Instead, the ETF’s market price is indirectly linked to its NAV and asset markets via the law of one price and the arbitrage conducted by APs. APs are financial institutions, usually large broker-dealers, with the right, but not the obligation, to create and redeem ETF shares outright. For example, when the ETF price exceeds the fund NAV, they can deliver a basket of securities, the creation basket, to the fund sponsor in exchange for new ETF shares, keeping the price difference as an arbitrage profit. Generally, creation and redemption baskets resemble the portfolio securities held by the ETF. Through this process, APs increase the number of ETF shares outstanding. Importantly, creations and redemptions of ETF shares normally do not involve cash but occur in-kind. As a special case, in the U.S. such in-kind transfers of securities are

3The described mechanism refers to ETFs that physically replicate their benchmark index by holding the underlying securities. There also exist synthetic ETFs that track their benchmark index using derivatives, such as swaps. These ETFs represent the minority of funds. The large majority of ETFs pursue the physical index replication process which this paper builds upon.
also tax-exempt. This mechanism turns APs into the central suppliers of liquidity in ETF markets. At the same time, the dependence on AP arbitrage renders ETFs vulnerable to shocks to financial intermediaries’ balance sheet capacity. When APs temporarily retreat from ETF creations or redemptions due to regulatory capital constraints, ETF prices can start deviating substantially from the fund NAV, leading to relative mispricing (Pan & Zeng, 2019). This potential discrepancy between the price at which investors can liquidate ETF shares at short notice and the fund NAV is what I call relative mispricing. It constitutes the key friction in ETF markets.

On the other side, the key friction associated with MFs are payoff externalities among fund investors. In contrast to ETFs, MF shares are purchased or redeemed directly from the fund sponsor at the fund NAV. Unlike ETFs, all MF trades submitted during the trading day are executed at the same price, the end-of-day fund NAV. By construction, MFs have zero relative mispricing. The daily fund NAV is calculated based on the closing prices, or estimates thereof, of the funds’ portfolio holdings and typically posted by 6pm ET. Hence, the MF NAV is flexible in the sense that it reflects the same-day price impact in asset markets. However, the MF NAV is not fully forward looking, as it does not account for potential future price impact generated by funds’ flow-induced selling pressure (Zeng, 2020). The MF can satisfy net investor redemptions by temporarily depleting cash buffers, if available, but will eventually need to liquidate asset holdings when outflows are large. Thus, by construction, net MF redemption are directly linked to transactions in portfolio securities and costly for the fund. Flow induced transaction costs for MFs include commissions, bid-ask spreads, price impact and taxes on capital gains distributions. When capital gains are realized as a result of security sales after redemptions, these taxes are borne by the remaining fund investors. This implies that MF investors may be subject to an early realization of capital gains taxes even if they remain invested in the fund. Fund-level transaction costs also include indirect costs if flows force fund managers to deviate from their optimal portfolio. However, these indirect costs are less significant for index funds compared to active MFs. Overall, because of how MF shares are priced, transaction costs are caused by exiting investors but borne by the remaining MF shareholders. This externality among MF investors gives rise to run risks, whereby the anticipation of redemptions by other investors increases the incentives for an individual investor to prematurely redeem MF shares herself. Ultimately, payoff complementarities in MFs and the associated run risks can dilute long-term MF investors’ shareholdings and constitute the key friction in MFs. Previous studies have shown that these effects are larger for funds with illiquid portfolio holdings, lower turnover and a larger share of smaller (retail) investors (Chen, Goldstein, & Jiang, 2010; Goldstein, Jiang, & Ng, 2017). When trading ETF shares, each investor bears her own transaction costs. Consequently, ETFs are not subject to such strategic complementarities among investors. For a more extensive dis-

4In the U.S., orders for MFs must usually be submitted by 4pm ET to be executed at the same-day fund NAV. Any orders submitted thereafter will be executed at the next available NAV, that is on the next trading day.
cussion of the institutional details behind ETFs, including their differences from open-end MFs, I refer to Lettau and Madhavan (2018) and Ben-David, Franzoni, and Moussawi (2017).

The first U.S. mutual fund was launched in 1924 (Federal Reserve, 2000). For most of the time, mutual funds have been the only way for retail investors as well as many institutional investors to obtain cheap portfolio diversification and access illiquid market segments. MFs were a revolution because they brought the average person into the stock market. In contrast, ETFs were originally established in 1993 by stock exchanges and targeted at traders of futures contracts. ETFs simply represented a new way for investors to trade bundles of stocks. Initially, they were not intended to directly compete with MFs as a long-term investment vehicle for the average investor. This is consistent with the idea of ETFs being tailored towards short-term traders who require intra-day liquidity, a view that remains widespread in the academic literature. Yet, ever since the financial crisis in 2008, ETFs have become more than just a vehicle for high turnover trading strategies. ETFs now constitute a popular alternative investment vehicle for many different types of investors.\(^5\) In this paper, I provide an explanation reconciling the simultaneous popularity of ETFs among very short-term, high frequency traders as well as long-term investors. Intra-day trading of ETF shares implies instantaneous liquidity provision. However, it also implies a larger flexibility of ETF prices. ETF prices immediately adjust to reflect market demand and supply conditions, and therefore new information. Even though long-term investors may not require intra-day liquidity, they benefit from the forward looking nature of ETF market prices because it shields them from externalises associated with the trading activities of other fund investors.

1.1 Related literature

This paper contributes to the growing ETF literature by formalizing a unified framework to understand investors’ trade-offs between ETFs and MFs. I show that index ETFs and MFs are imperfect substitutes because of differences in their payoff structures, and the associated financial frictions. As a result, I demonstrate that MFs provide liquidity over the short-term, whereas ETFs generate higher returns over the long-term when underlying portfolio holdings are illiquid and fund flows are large. According to a conventional argument, ETFs primarily appeal to investors with a higher liquidity demand due to their ability to trade intraday. I challenge this common conception. On the contrary, I establish that in equilibrium investors facing relatively higher idiosyncratic liquidity risks and shorter investment horizon

\(^5\)Based on survey estimates, around 11% of U.S. households (≈ 13.9 mm) owned ETFs in 2021. For comparison, in the same year, around 52.3% of U.S. households held MFs. (see 2022 Investment Company Fact Book and www.ici.org/news-release/22-news-ownership). In response to investors’ increasing demand for ETFs, some asset managers have already converted existing MFs into ETFs (see www.ft.com/content/ce92665-b481-4c2e-877d-9a9abf4f7f8d5, www.ft.com/content/3eabe21a-ea76-486d-813bb657c1f9 and www.ft.com/content/9eb2fbb-5f76-4c43-982b-88844d1f1bc for recent examples of (planned) open-end mutual fund to ETF conversions).
prefer open-end mutual funds. Investors with relatively lower idiosyncratic liquidity risks and longer investment horizon, on the other hand, prefer open-end mutual funds ETFs.

The closest paper to mine is Huang and Guedj (2009) who develop an equilibrium portfolio choice model in which investors trade off the liquidity insurance services of MFs with their moral hazard costs. In a framework featuring ETFs that are frictionless, thus equivalent to their benchmark index, they show that ETFs are more suitable when investors have more correlated liquidity shocks or underlying securities markets are less liquid. I complement their results by accounting for the frictions in both, ETFs and MFs, and showing that MFs can be more suitable for investors with short-term liquidity needs due to the potential for short-term relative mispricing in ETFs.

A few other papers have previously explicitly considered investors’ trade-offs between ETFs and open-end mutual funds. Consistent with my results, these papers generally agree that ETFs and MFs can co-exist in equilibrium. However, they attribute the co-existence of both fund types to clientele effects resulting from different fee structures and ETFs’ intra-day liquidity provision (Agapova, 2011) or tax efficiency (Moussawi, Shen, & Velthius, 2020). Specifically, Agapova (2011) predicts that investors with higher liquidity needs or longer time horizons prefer ETFs because of their exchange traded nature and lower fees, while short-term traders prefer MFs because of their commission-free nature. In contrast, I argue for payoff structure differences as the fundamental economic feature distinguishing ETFs and open-end MFs. Fees and taxes certainly represent important factors in investors’ portfolio allocation decision, especially with respect to index funds. Yet, this paper takes the stance that different fee structures and tax rules do not constitute the key economic differences ETFs and MFs. Higher fee levels and lower tax efficiency are not an integral part of MF’s organizational structure. Instead, these parameters are at discretion of fund sponsors, brokerage firms or regulators. Fees can change over time and tax rules vary across countries. Depending on the share class, index MFs are now available at similar or even lower expense ratios than their ETF competitors and U.S. retail investors can largely trade ETFs commission free. At the same time, the ETF tax advantage is unique to the U.S. In many other jurisdictions where ETFs have experienced a similar growth as in the U.S., such as in Germany, MFs and ETFs are taxed identically. In contrast, ETFs’ and MFs’ payoff structure is universal and inherently defines the security design of both fund types. Yet, even though the channels driving market segmentation between ETFs and MFs in the existing literature differ from mine, the predictions largely point in a similar direction by suggesting that ETFs are preferred by longer-term investors with relatively higher liquidity needs and shorter investment horizons.

---

6 In-kind transfers of securities which form the basis of the ETF creation and redemption process are tax exempt. In contrast, MFs’ (flow-induced) transactions in portfolio securities constitute a taxable event and may result in annual capital gains tax distribution to fund investors.

7 This finding is consistent with Box, Davis, and Fuller (2019) who explain the co-existence of same-index ETFs by clientele effects according to which more price sensitive long-run investors prefer low cost funds, whereas short-term liquidity traders self-select into high turnover ETFs.
Therefore, I contribute to this literature by showing that even abstracting from fee and tax differences, ETFs provide relatively larger payoffs over the long-term due to the absence of share dilution risks. I also demonstrate that MFs, and not ETFs, may in fact be preferred by investors with short-term but no intra-day liquidity needs because MFs’ guaranteed redemption at the fund NAV allows these investors to avoid the potential short-term mispricing in ETFs.

In addition to contributing to the growing literature on ETFs, my paper adds to the literature on investment funds’ choice of their organizational structure. The previous literature largely focuses on the choice between the open- and closed-end fund structure. The consensus proposes both fund types co-exist because of complementarities in their liquidity provision. Deli and Varma (2002) empirically show that funds holding securities with lower liquidity or price transparency are more likely to be structured as closed-end funds (CEFs). Cherkes, Sagi, and Stanton (2009) theoretically investigate why CEFs exist despite their tendency to trade at a discount to their fund NAV, the ‘CEF puzzle’. Similar to ETFs in my model, they argue that CEFs exist because they facilitate investments in illiquid securities without the externality costs of open-end MFs. I continue this line of research by considering the tradeoff between the ETF and open-end MF structure. ETFs represent a hybrid organizational form between open- and closed-end MFs. Their shares are exchange traded like CEFs, but a selected group of intermediaries (APs) can create and redeem ETF shares as in open-end MFs. I focus on investors’ choice between ETFs and MFs, taking as given the existence of both. Yet, in line with previous results, my model predicts the relative size of the ETFs compared to the open-end MF sector to be increasing in market segment illiquidity.

Third, this paper complements the literature on liquidity provision by non-bank financial intermediaries (NBFIs) and the associated risks. Fundamentally, I build on the seminal work by Diamond and Dybvig (1983) on bank liquidity provision. Equity issuing intermediaries including ETFs and MFs also provide liquidity (Ma, Xiao, & Zeng, 2022a). Similar to banks, the liquidity mismatch between ETF or MF shares and their portfolio holdings is one of the key drivers of fragility in these markets.

For MFs, this literature has identified the combination of payoff complementarities among MF investors with the liquidity mismatch on MFs’ balance sheet as the central source of run risk in open-end MFs. Chen et al. (2010) formalize this mechanism and show, empirically and theoretically in a global games framework, that the payoff structure of open-end MFs gives rise to a first-mover advantage among investors which increases in the illiquidity of fund assets and decrease with institutional fund ownership. Thereby, outflows predict future declines in fund NAVs. These payoff complementarities arise because MF redemptions are associated with costs for the fund that are not fully reflected in the fund NAVs at which MF investors trade. On the one side, these flow-induced costs include direct transaction costs (e.g., com-
missions, bid-ask spreads or price impact) arising from trades of portfolio assets as well as indirect costs, such as capital gains distributions. Amongst others, Edelen (1999) quantifies the significant cost of liquidity-motivated trading for long-term MF investors. Coval and Stafford (2007) show that MFs tend to conduct costly and unprofitable trades ex-post large outflows at the cost of their remaining shareholders. They estimate that most flow induced MF trades occur with a lag of one day after redemption events. Dickson and Sialm (1999) provide evidence for the importance of capital gains taxes as a source of externalities among MF investors. On the other side, MF NAVs are not perfectly forward looking. As a result, flow induced transaction costs may be borne by the remaining instead of the redeeming fund investors. Most recently, the staleness in MF NAVs is documented by Choi, Kronlund, and Oh (2022) for fixed income funds. Dickson and Sialm (1999) provide evidence for the importance of capital gains taxes as a source of externalities among MF investors. On the other side, MF NAVs are not perfectly forward looking. As a result, flow induced transaction costs may be borne by the remaining instead of the redeeming fund investors. Most recently, the staleness in MF NAVs is documented by Choi, Kronlund, and Oh (2022) for fixed income funds. Tufano, Quinn, and Taliaferro (2012) highlight outdated fund portfolio weights as another source of staleness in MF NAVs. Fund liquidity management tools, such as swing pricing, that have been shown to mitigate payoff externalities among MF investors. Nonetheless, U.S.-based MFs have not yet adopted swing pricing policies to this date, in part due to operational challenges in its implementation (International Monetary Fund, 2022). Similarly, whereas cash buffers decrease portfolio illiquidity, Zeng (2017) argues that the predictable re-building of cash reserves ex-post outflows exacerbates run risks. I build on this literature by accounting for the potential short-term staleness in MF NAVs and incorporating transaction costs driven payoff complementarities among MF investors and the resulting run risk into my model.

The literature on risks in ETFs has focused on the spillover effects from ETFs to financial markets caused by the AP arbitrage channel. From investors’ perspective, the key friction in ETFs is the potential for relative mispricing between the ETF price at which investors can trade and the fund NAV. Liquidity in secondary ETF markets directly depends on APs’ engagement in the creation or redemption of ETF shares. Due to their dual role as dedicated ETF arbitrageurs and broker-dealers in underlying security markets, APs are subject to balance sheet capacity constraints. As the first to endogenously model the AP arbitrage process, Pan and Zeng (2019) show theoretically and empirically how the combination of a liquidity mismatch between ETF shares and portfolio securities with APs’ balance sheet constraints

---

8See also Feroli, Kashyap, Schoenholtz, and Shin (2014); Goldstein et al. (2017). Falato, Goldstein, and Hortaçsu (2021); Ma, Xiao, and Zeng (2022b) offer empirical evidence from the Covid-19 crisis. Kacperczyk and Schnabl (2013); Schmidt and Wermers (2016) document shareholder runs in money market funds. Previous attempts to quantify the costs of MF redemptions include (Chordia, 1996; Wermers, 2000; Greene & Hodges, 2002; Johnson, 2004).

9Previous evidence, focusing on the international and illiquid domestic equity MFs, include Goetzmann, Ivković, and Rouwenhorst (2001); Chalmers, Edelen, and Kadlec (2001); Boudoukh, Richardson, Subrahmanyan, and Withelaw (2002); Zitzewitz (2003).

10For example, previous studies have analyzed the effect of ETF ownership on non-fundamental asset price volatility Dannhauser and Hoseinzade (2017); Ben-David, Franzoni, and Moussawi (2018), asset price discovery Brown, Davies, and Ringgenberg (2020); Israeli, Lee, and Sridharan (2017); Glosten, Nallareddy, and Zou (2019); Madhavan and Sobczyk (2016); Bhattacharya and O’Hara (2018) and return co-movement Da and Shive (2018); Shim (2019); Madhavan and Morillo (2018).
gives rise to limits to arbitrage in ETF markets. The effects are especially pronounced in times of heightened market volatility. These frictions can explain persistent and countercyclical mispricing in ETFs holding fundamentally illiquid securities, such as corporate bond ETFs. Shim and Todorov (2022) propose the endogenous selection of creation and redemption baskets as another driver of relative mispricing in fixed income ETFs. Consistent with my model predictions, they suggest that ETFs may be more effective in managing illiquid assets compared to MFs. Haddad, Moreira, and Muir (2021) document large discounts in bond ETFs during the covid-19 sell-off. (Petajisto, 2017) show that the relative ETF mispricing remain statistically and economically significant even after accounting for potential staleness in the fund NAV. I contribute to this literature by showing that the frictions in ETFs and MFs affect fund liquidity provision and returns at different horizons and thus constitute key drivers of investors’ portfolio choice between both fund types.

The remainder of the paper is organized as follows. Section 2 introduces the model. Section 3 presents the equilibrium predictions. Section 4 concludes.

2 Theoretical framework

I propose a portfolio choice model, built upon Diamond and Dybvig (1983), to study on how rational investors with different liquidity needs and time horizons optimally choose to allocate their portfolio between index ETFs and open-end mutual funds. Thereby, I will also generate predictions regarding the role of MFs in the index investment industry going forward. The model is designed to capture the fundamental economic differences between ETFs and MFs, the mechanisms by which both funds’ shares are priced and traded in financial markets. I do not aim to capture all dimensions on which funds may differ. Most importantly, I abstract from variation in fund expense ratios, capital gains taxation as well as search and switching costs. However, I consider the effects of fee and tax differences between ETFs and MFs on investors’ portfolio allocation in an extension of the model. Accounting for the on average higher expense ratios across MF share classes and the U.S. specific tax efficiency of ETFs further strengthens my model predictions. I refer to Moussawi et al. (2020) and Agapova (2011) for a detailed empirical analysis of the implications of tax and fee differences between ETFs and MFs.

There are three periods, \( t = \{1, 2, 3\} \), time is discrete and financial markets are competitive. There is a single consumption good, dollars. The economy consists of a continuum of in-

\footnotesize
\begin{itemize}
  \item Other related studies include Todorov (2021); Gorbatikov and Sikorskaya (2022); Malamud (2016).
  \item ETFs’ relative tax advantage, that is the deferral of capital gains taxes until investors sell their ETF shares, effectively provides investors with an interest free loan from the IRS. This tax deferral is most valuable for long-term investors. Similarly, the effect of fee differences on fund returns increases with the investors’ horizon.
\end{itemize}
vestors $i$ and financial intermediaries (broker-dealers). Investors are heterogeneous in terms of their liquidity risk exposure. Some investors are more likely to be forced to liquidate their assets early than others. There also exists a group of "sleepy" or inattentive investors. I call them pension funds. Consistent with the menu of funds offered by many retirement plans in practice, pension funds can only invest in mutual funds. They follow a simple buy-and-hold strategy, making no portfolio re-allocation decisions between the initial and terminal period. Financial markets feature risky composite securities, index MFs and index ETFs. MFs and ETFs are both investment technologies that pool money from investors to invest in a diversified portfolio of potentially illiquid securities. The composite security represents a benchmark index and can only directly be traded by fund managers and intermediaries. Investors can only invest in financial securities via investment funds. They do not have direct access to trade the index. For any benchmark index $j$, there exists both a MF and an ETF tracking it by physically replicating the index portfolio. I abstract from individual securities markets. Hence, ETFs and MFs simply hold the composite security. Abstracting from portfolio differences allows me to focus on non-portfolio differences as the source of heterogeneity between ETFs and MFs. Besides, there also exists a risk-free asset which pays zero interest $R^f = 1$.

The key model frictions directly arise from the different payoff structures of ETFs and MFs. Financial frictions in ETFs emerge from the potential for ETF prices at which investors transact in secondary markets to diverge from the fund portfolio value (NAV). I refer to this difference between ETF market prices and fund NAV as relative mispricing (ETF premium / discount). In contrast, frictions in MFs are due to payoff externalities among fund investors. I abstract from agency frictions between fund sponsors, intermediaries and investors. MFs and ETFs solely serve as investment technologies that facilitate liquidity transformation and diversification for investors. The objective of fund sponsors is outside of the scope of my model.

The $j \times 1$ vector of index fundamentals, $\vec{x}$, serves as the state variable. Each component of vector of state variables, $x_1, ..., x_j$, reflects the fundamental value of a composite security $j$ and is observed at the beginning of period $t = 1$. The index fundamentals are independently distributed according to $x_j \sim N(\mu_j, \sigma_j^2)$ with $\mu_j > 1$. The remainder of the paper focuses on the case of one single index. Hence, $j = 1$ and $\vec{x}$ reduces to the scalar, $x_j$. I assume that all agents have identical information about $x$. Table 1 in Appendix C summarizes the model’s notation.

2.1 Financial markets

**Benchmark index.** Each benchmark index $j$ represents a diversified portfolio of securities (e.g., S&P 500 Index, Bloomberg Aggregate Bond Index, MSCI Emerging Markets Index). The remainder of the paper focuses on the case of one single index that is tracked by one

---

13Since $x_j$ represents the terminal index payoff, $\mu_j < 1$ would imply a negative expected return for index $j$. 
Investors allocate portfolio between ETF & MF $\theta_0^{E,j}, \theta_0^{M,j}$

$t = 0$

Early redemption decision $\theta_1^{E,j}, \theta_1^{M,j}$

$t = 1$

Terminal index & fund payoffs

$t = 2$

Index value (state) $x_j$ realized

Investor types = Patient, Impatient

Figure 1: Timeline

ETF and MF. I nevertheless keep the notation $j$ to illustrate the dependence of key parameters, and consequently the model predictions, on the return distribution and liquidity of the benchmark portfolio. For tractability, I disregard the dynamics of the individual security prices and directly model the price of the benchmark index as a composite security. In the following, I use the terms composite security and index interchangeably. The composite security pays a terminal dividend in $t = 2$. This composite security can be traded by mutual fund managers, financial intermediaries in their role as APs and market makers (broker-dealers) who are specialized to provide liquidity in these index markets. Composite security markets are segmented. This assumption is standard in the literature (e.g., Malamud (2016), Gromb and Vayanos (2002)) and allows for potential relative mispricing between markets. Accordingly, MFs and APs trade with distinct index market makers, potentially at different prices. The timing of when MFs and APs respectively trade in index markets also differs. APs respond to demand imbalances in secondary ETF markets in real time intra-day. In contrast, MFs adjust their portfolio holdings with a lag after observing their investors’ net redemptions. More fundamentally, the segmented index markets assumption allows me to distinguish ETFs’ and MFs’ respective price impact in underlying security markets. MFs and APs are both price takers and do not compete on the asset side in index markets. In an extension of the model I relax this assumption and show that the same overall predictions continue to hold when ETFs and MFs trade with a common index market maker. Investors cannot directly trade composite securities. In theory, it is equivalent to assuming that the transaction costs investors face under autarky when trying to replicate, and sequentially rebalance, the index themselves, for instance through direct indexing in separately managed accounts (SMAs), are always strictly larger than the cost of holding an ETF or MF.

The index price in the interim period is endogenous and given by $P^j_t$. It is determined by market clearing between market makers, MFs and APs. In the terminal period, the index pays a terminal dividend equal to its fundamental value, $v(x_j) = x_j$. Hence, the cum-dividend index price at $t = 2$ is given by

\[ P^j_2 = P^j_1 + v(x_j) \]

Similarly, the assumption that market makers are distinct from banks acting as APs in ETF markets is a simplification. In practice, the same broker-dealer may act in multiple roles as market maker in index market, as a counterparty to MFs, as well as an AP for ETFs. Studying the implications of the associated conflicts of interests is out of the scope of this paper but should be explored in future research.
Investment funds. ETFs and MFs passively track these benchmark indices by physically replicating their underlying securities portfolios. For every benchmark, \( j \), there exist both an open-end mutual fund as well as an ETF tracking it. Funds’ ex-ante choice which benchmark index to track is outside the scope of this model. For simplicity, neither ETFs nor MFs charge any fees. Net expense ratios are zero, \( f^\text{ETF} = f^\text{MF} = 0 \) \( \forall t \) and \( j \). I abstract from any differences in ETFs’ and MFs’ portfolio replicating methodologies. Funds do not hold any cash. The portfolio holdings of the ETF and MF tracking benchmark \( j \) are identical.

ETFs. ETFs are traded on competitive financial markets intra-day. The ETF market price, \( P^E \), arises endogenously in equilibrium from market clearing between ETF investors and APs. By definition, the ETF NAV is simply equal to the value of all its index holdings divided by the number of its shares outstanding. I abstract from discrepancies between creation or redemption baskets and the ETF portfolio holdings.

Assumption 1 ETF creation and redemption baskets are always identical to the ETF benchmark index.

Lemma 1 Following assumption 1, the ETF NAV is always equal to the value of the benchmark index, \( \text{NAV}^E \equiv P^j \). ETFs perfectly replicate their benchmark. There is no tracking difference.\(^{15}\)

Yet, what matters to ETF investors is not the ETF NAV but the ETF market price at which they trade. \( P^E \) is linked to the index price, \( P^j \), via arbitrage. There are limits to arbitrage. As a result, the ETF price can deviate from its NAV, resulting in relative mispricing vis-à-vis the benchmark index. Arbitrage constraints arise endogenously from intermediaries’ balance sheet capacity constraints. These balance sheet capacity constraints are increasing in the index segment specific illiquidity. Intuitively, due to their greater illiquidity, redemptions of corporate bond ETF shares consume more AP balance sheet capacity than redemptions of large cap equity ETFs. Pan and Zeng (2019) establish the microfoundations for this modeling choice and study the ETF arbitrage mechanism in an equilibrium model. In their model, APs face several costs associated with ETF arbitrage, including transaction costs in illiquid underlying securities markets, inventory costs and creation or redemption fees. They show that limits to arbitrage in ETF markets are time-varying and arise endogenously from optimal AP arbitrage in the face of balance sheet constraints. The AP arbitrage mechanism in my model builds on their framework. However, for tractability, I focus on inventory

\(^{15}\)Here, tracking difference refers to the difference between the fund net asset value per share and the value of its benchmark index at any time, \( \text{NAV}^E - P^j \). In practice, tracking difference often refers to the return difference between a fund and its benchmark instead. Given that the initial fund and index price is normalized to equal one and there are no dividend payments at \( t = 1 \), the ETF return and payoffs are the same in this model.
Definition 1 The relative ETF mispricing refers to the difference between the ETF’s net asset value, $NAV_{E,j}^t$, and its market price, $P_{E,j}^t$. At any given point in time, the relative mispricing is given by

$$\epsilon_{E,j}^t \equiv NAV_{E,j}^t - P_{E,j}^t. \quad (1)$$

Then, the ETF price is given by

$$P_{E,j}^t = NAV_{E,j}^t + \epsilon_{E,j}^t.$$

In equilibrium, $\epsilon_{E,j}^t$ is endogenously determined by the AP’s optimal net ETF share redemptions. Given $P_j^t$, the mispricing captures the wedge between the index and the ETF price. Because transaction costs for trading ETF shares are incurred at the investor level, investors’ effective (pre-tax) transaction price may deviate from $P_{E,j}^t$ by the amount of bid-ask spreads and trading commissions charged per ETF share. For simplicity, I assume ETF bid-ask spreads are zero. I also assume investors trade ETFs commission free.

Open-end mutual funds. In contrast to ETFs, all MF creations or redemption orders submitted throughout the trading day are executed at the end of each trading day at the fund NAV. The daily MF NAV is calculated based on the closing price of the funds’ portfolio holdings, or estimates thereof. By design, MFs have zero relative mispricing. The price at which investors trade MF shares is always equal to its estimated end-of-day portfolio value per share,

$$P_{M,j}^t \equiv NAV_{M,j}^t.$$

Rational investors have perfect foresight and correctly anticipate the fund NAV $P_{M,j}^t$ when submitting their MF orders throughout the trading day. This assumption is standard in the literature. The MF NAV is flexible to the extent that it accounts for new fundamental information as $x_j$ is revealed to everyone. However, the NAV is not fully forward looking. Total net fund flows and the price impact caused by MFs’ ensuing flow-induced trading are generally unknown in the U.S. until the next trading day (Ma et al., 2022b). While ETF prices are determined in equilibrium from the supply-demand conditions in secondary ETF markets, MF prices are set based on supply-demand conditions in underlying security markets. The imperfect flexibility of MF prices constitutes the key friction of mutual funds in this model. Formally, the potential stale pricing of MF shares is captured by the specification of the MF price at which investors can redeem shares in the interim period. The MF NAV at $t = 1$ is exogenous and given by
\[\text{NAV}_1^{M,j} = \psi E_1[x_j].\]

Since \(x_j\) becomes common knowledge at \(t = 1\), \(2\) reduces to \(\text{NAV}_1^{M,j} = \psi x_j\). I assume \(P_t^{M,j}\) is a function of the expected terminal index value and a parameter \(\psi\) with \(0 < \psi \leq 1\). \(\psi\) reflects a penalty for the early liquidation of MF shares. The larger \(\psi\), the higher the cost to the individual investor of liquidating their MF holdings early. This specification is consistent with the literature (Chen et al., 2010). The model predictions are robust to alternative specifications as long as the condition \(\text{NAV}_1^{M,j} < E_1[x_j]\) is satisfied. It ensures that the early liquidation of all MF holdings irrespective of the index fundamentals does not become the unique dominant strategy for risk-averse investors. If \(\psi = 1\), the expected return of MF shares between \(t = 1\) and \(t = 2\) is zero, \(E_1[\frac{P_2^{M,j} - P_1^{M,j}}{P_1^{M,j}}] = 0\). In this case, a MF run would be the unique equilibrium outcome at \(t = 1\).

All net MF redemptions are executed in cash at the end of trading day \(t = 1\). MFs cannot satisfy redemptions using in-kind transfers (RIK) of security baskets.\(^{16}\) In practice, MF managers actively manage their cash positions. They may first meet net redemptions by depleting their cash holdings before restoring liquidity buffers by trading portfolio securities (Zeng, 2020). For tractability I abstract from MFs’ liquidity management in this model. This simplification however introduces a gap between the time when MFs have to pay off redeeming investors in cash (at \(t = 1\)) and when MFs receive the cash proceeds from their trades in index markets (at \(t = 1^+\)). To overcome this technical inconsistency, MFs initially meet net redemptions at \(t = 1\) by drawing down overnight credit lines from exogenous banks. Implicitly, I assume MFs can borrow overnight at \(t = 1\) at the risk-free rate. They then fully repay any overnight credit lines at \(t = 1^+\) by liquidating portfolio assets. There is no default. Period \(t = 1^+\) represents the morning of the next trading day following \(t = 1\) when MFs learn about the volume of their investors’ net redemptions. The existence of these credit lines only constitutes a technical feature of the model resulting from the assumption that funds do not hold cash. It has no implications on the model predictions. The only purpose of bank credit lines is to bridge the gap between the time at which MFs have to satisfy net redemptions by investors in cash, at \(t = 1\), and when trading in index markets starts again at the beginning of the subsequent period \(t = 1^+\). Eventually, MFs must eventually liquidate portfolio securities to meet net fund redemptions. Because all portfolio trades are flow induced, MFs liquidate whatever amount of index shares necessary to pay off redeeming investors.

\(^{16}\)This is consistent with the data. RIKs are only feasible for large investors. MFs rarely make use of the option to meet redemptions in-kind even if they are legally permitted to do so. Based on data from U.S. funds’ shareholder reports, Agarwal, Ren, Shen, and Zhao (2022) find that only 13.1% of the funds that reserve the right to execute redemptions in kind actually engage in in-kind redemptions at least once during a sample from 1997 to 2017.
2.2 Agents

There is a continuum of individual investors with mass 1. The mass of pension funds, the sleepy investors, is $\eta$. Pension funds only invest in MFs, while all other investors can choose to allocate their portfolio between ETFs and MFs. This exogenous market segmentation is consistent with the data. In the U.S., 58% of defined contribution (DC) pension fund assets were invested in mutual funds as of 2019. ETFs currently only play a minor role in DC plans. In turn, 54% of AuM in U.S. equity, hybrid and bond MFs are held via retirement accounts.\textsuperscript{17} Only financial intermediaries, including ETF authorized participants, mutual funds and market makers can directly trade the index.

**Individual investors.** Individual investors are the agents of interest. This group is broadly defined to include retail and institutional investors, such as endowments or family offices, who may choose to invest in ETFs or MFs. In the following I will refer to these agents simply as the investors. In practice, retail investors also constitute the ultimate beneficiary of pension funds. Hence, their fund market investments can be interpreted as additional savings, beyond the capital that they may already have invested through their 401k or IRA. I abstract from capital gains taxes in this model. There is one important limitation. Investors, such as high frequency trading firms, who require intra-day liquidity are outside of the scope of this model. Since only ETFs provide intra-day liquidity, such agents would never invest in MFs in the first place. They do not face an interesting trade-off between different fund types and are hence not explicitly featured in my model.

The unit mass of investors are homogenous in terms of their initial wealth, $\theta_i^0 = \theta_0$. I normalize $\theta_0 = 1$. They invest long-term but face idiosyncratic liquidity risks as in Diamond and Dybvig (1983). Liquidity shocks can force investors to liquidate their portfolio early at $t = 1$. In contrast to classic bank run models in which all investors initially face the same probability of a liquidity shock, here the ex-ante probability of experiencing a liquidity shock, denoted by $\lambda_i$, differs across investors. Specifically, $\lambda_i$ is independently and identically distributed according to

$$\lambda_i \sim U[\lambda_0, \lambda_1],$$

where $0 \leq \lambda_0 \leq \lambda_1 \leq 1$. This idiosyncratic liquidity risk exposure is privately observed at $t = 0$ and causes ex-ante heterogeneity among investors. $\lambda_i$ is directly linked to investor $i$’s expected investment horizon, $T_i$, according to

$$E_0[T_i] = 2 - \lambda_i.$$  \hfill (3)

The larger $\lambda_i$, the shorter an investor’s expected investment horizon. Accordingly, variation in $\lambda_i$ can reflect differences in investors’ probabilities of being hit by an unexpected expense

\textsuperscript{17}See 2022 Investment Company Fact Book.
(e.g., health care bill, home repairs, lawsuit) or income shock (e.g., unemployment) as well as differences in their holding periods (e.g., age). I denote the expected mass of impatient investors, that is investors who receive a liquidity shock at \( t = 1 \) and are forced to liquidate their portfolios early, by

\[
\bar{\lambda} = \int_{i} \lambda_i d_i = \frac{\lambda_0 + \lambda_1}{2}.
\]  

(4)

Given \( \lambda_i \) and \( x_j \sim N(\mu_j, \sigma^2_j) \), at \( t = 0 \) investors choose to allocate their endowment between ETFs and MFs. They do not have access to the risk-free asset in the initial period, but must exchange their entire endowment into fund shares. After observing their idiosyncratic liquidity shocks, at \( t = 1 \) investors can decide to liquidate any amount of their fund holdings early and store the proceeds in the risk-free asset until the terminal period. Investors’ hit by a liquidity shock must always liquidate their entire fund portfolio. Investors cannot reallocate assets between MF and ETF markets at \( t = 1 \). Conditional on their initial allocation to MFs and ETFs, \( \{\theta_{i,M,j}, \theta_{i,E,j}\} \), they can only keep or liquidate fund shares. Investors do not receive any income from other sources than their financial investments. They do not have access to leverage and face short-selling constraints. These restrictions allow me to isolate the effect of different liquidity risk exposures on investors’ allocation across fund types.

All investors have identical time-separable preferences. The primitive utility function, \( u(c) \), is increasing, concave, and satisfies the Inada conditions. In the baseline case, I simply assume all agents are risk-neutral. Financial assets are agents’ only income source. There exists no income from labor or real capital investment. In the absence of storage technologies for the consumption good cash, other than investment funds, agents’ consumption always equals their wealth at the end of their lifespan, \( c_T = w_T \). Accordingly, investors choose their initial portfolio allocation between ETFs and MFs to maximize their expected utility over terminal wealth, \( E[u(w_T)] \), subject to their endowment 6, leverage 7 and short-selling constraints 8:

\[
\max_{\theta_{i,M,j}, \theta_{i,E,j}} E[u(w_T)] \]

(5)

\[
s.t. \quad \theta_{0,E}^i P_0^E + \theta_{0,M}^i P_0^M = \theta_{0}^i P_j^i, \]

(6)

\[
\theta_{0,E}^i - \theta_{1,E}^i > 0, \quad \theta_{0,M}^i - \theta_{1,M}^i > 0, \]

(7)

\[
\theta_{t,E}^i \geq 0, \quad \theta_{t,M}^i \geq 0 \quad \forall \ t = 0, 1. \]

(8)

\( \theta_{0,E}^i \) and \( \theta_{0,E}^i \) are the number of shares of the ETF and MF tracking index \( j \) bought by investor \( i \) at \( t = 0 \) respectively. I only model the allocation among funds tracking a given benchmark index \( j \) and do not consider investors’ choice between different index segments. Initial wealth is normalized to one unit of the index, \( \theta_0^i = 1 \), which serves as the numeraire in the economy.

The terminal wealth of patient or late investors, denoted by \( i = l \), is given by
\[ w^i_2 = \theta^{i,E,j}_1 P^{E,j}_2 + \theta^{i,M,j}_1 P^{M}_2 + (\theta^{i,E,j}_0 - \theta^{i,E,j}_1) P^{E,j}_1 R^f + (\theta^{i,M,j}_0 - \theta^{i,M,j}_1) P^{M,j}_1 R^f. \] (9)

Investors who are subject to a liquidity shock at \( t = 1 \) have to liquidate all portfolio holdings at \( t = 1 \). Ex-post, I denote these impatient or early investors by \( i = e \). They only derive utility from wealth in the interim period, \( w^i_1 \). Hence, the terminal wealth of impatient investors is given by

\[ w^i_1 = \theta^{i,E,j}_0 P^{E,j}_1 + \theta^{i,M,j}_0 P^{M,j}_1. \] (10)

Impatient investors’ terminal wealth only depends on their initial allocations, \( \theta^{i,E}_0 \) and \( \theta^{i,M}_0 \), as well as fund prices at \( t = 1 \).

**Pension funds.** The mass \( \eta \) of pension funds are sleepy investors. Their portfolio allocation is exogenous. Specifically, pension funds invest their entire endowment in MFs with the sole objective of meeting their long-term liabilities. This reflects the idea that the majority of 401(k) retirement plans in the U.S. does not offer ETFs but constrain participants’ choice set to MFs only. Formally, their allocation is given by

\[ \{ \theta^{P,M}, \theta^{P,E} \} = \{ \theta_0, 0 \}. \] (11)

Pension funds do not face any short-term liquidity risks. They do not rebalance their portfolio holdings in the interim period and never withdraw early. Instead, they hold their assets until maturity. From this follows the aggregate value of the portfolio holdings of the pension sector in the final period,

\[ W^P_2 = \eta \theta_0 P^M_2. \] (12)

Pension funds serve two purposes in my model. First, they stabilize the mutual fund markets in the interim period \( t = 1 \) by serving as sticky investors. Their existence prevents MFs from going bankrupt even amid runs by investors. This simplifies the model solution and is consistent with modeling conventions in the previous MF literature.\(^{18}\) Second, including pension funds facilitates counterfactual analyses on the effects of changes in pension funds capital allocation to MFs, \( \eta \), on other investors’ equilibrium allocation between ETFs and MFs. Because of the payoff complementarities among MF investors, MFs’ investor base composition matters for an individual fund investors’ payoff expectations (Chen et al., 2010). As a case in point, if DC retirement plans (e.g., 401-k plans) were to offer an increasing number of ETFs besides MFs going forward, this may translate into a reduction of "sleepy money" under management by MFs. The implications of such a policy change can be studied using this model.

---

\(^{18}\)See for example Chen et al. (2010).
2.3 Financial intermediaries

Financial intermediation in the economy occurs via index market makers, a representative authorized participant (AP) and a mutual fund. Index market makers supply liquidity in composite security markets. In this model framework, the AP provides liquidity in secondary ETF markets while requiring liquidity in index markets. Similarly, a MF provides liquidity to its investors while requiring liquidity in index markets. This directly follows from the liquidity risk exposure of fund investors. By assumption, investors face adverse liquidity shocks only. They can be impatient consumers, but not impatient savers. There are no positive cash flows shocks. Hence, in any period after \( t = 0 \), aggregate net flows into the fund sector are weakly negative. More generally, if investors were to also face positive cash flow shocks, investment funds may at times also turn into liquidity providers in index markets.

**Index market makers.** Risk neutral index market makers are the broker-dealers in composite security markets. They only appear in the interim period. Their sole purpose is to provide liquidity to investment funds at \( t = 1 \). These intermediaries are not frictionless. They face inventory quadratic costs for holding potentially illiquid securities on their balance sheet. In equilibrium, index market makers absorb the supply of index shares resulting from MFs’ flow induced portfolio liquidations and APs’ arbitrage activities. In the baseline model with index market segmentation, MFs and ETFs trade with distinct but identical index market makers. This convention keeps the model tractable and allows me to focus on competition between MFs and ETFs on the liability side. Each market maker submits a price schedule for index shares, \( P^j_t(\Theta^D,j_t) \), to maximize her expected trading profits given her inventory costs. Market makers’ optimization problem is given by

\[
\max_{\Theta^D,j_t} E_t[\Pi^j_{t+1}]
\]

s.t. \( \Pi^j_{t+1} = P^j_{t+1} - P^j_t - \frac{c_j}{2}(\Theta^D,j_t)^2 \).

For each representative market maker, 13 implies a downward sloping demand schedule for index shares:

\[
P^j_t = E_t[P^j_{t+1}] - c_j\Theta^D,j_t.
\]

\( P^j_t \) is the index price offered by dealer \( D \) for \( \Theta^D,j_t \) units of index shares at \( t \). The inventory cost parameter, \( c_j \), is increasing in the index segment \( j \) specific liquidity. Intuitively, \( c_{\text{Equity}} < c_{\text{Corporate Bond}} \), the inventory costs of holding corporate bonds is strictly larger than those of holding stocks. When \( c_j > 0 \), MFs and APs have price impact when trading in index markets. This represents the baseline specification. However, for \( c_j = 0 \), the model nests the case in which funds have zero security market price impact. This condition may be satisfied
for large cap domestic equities. Each representative market maker faces identical inventory costs. ETFs, via APs, and MFs therefore face the same net demand for index shares in composite security markets. The equilibrium market prices both funds trade at can nevertheless differ depending on the net redemptions by MF as compared to ETF shareholders as well as on APs endogenous arbitrage activities.

**Authorized participants.** APs are deep pocketed risk-neutral financial intermediaries with the right to create and redeem ETF shares outright through in-kind transactions with the ETF sponsor. This role turns APs into the central and only counterparty between ETF investors and fund sponsors. If idiosyncratic liquidity shocks or optimal portfolio reallocation decisions by investors at \( t = 1 \) generate excess demand (supply) for ETF shares, the AP can step in and create (redeem) ETF shares by buying (short selling) the underlying creation (redemption) basket in index markets and delivering (redeeming) the proceeds (corresponding security basket). APs trade in index and ETF markets with the purpose of generating short-term profits. In practice, there are other reasons for APs to engage in ETF creations or redemptions. For example, ETF arbitrage can help financial intermediaries manage their own liquidity risks or hedge balance sheet exposures. I abstract from such motives here. APs immediately offload any temporary composite security or ETF positions obtained as part of their arbitrage trades in the same period \( t = 1 \). They do not hold any assets over several periods. There exists a single representative AP for each ETF \( j \). Similar to index market makers, APs only operate in the interim period, \( t = 1 \).

There are costs associated with the AP’s ETF arbitrage trades. Due to the potential time lag between an AP’s transactions in ETF markets and its offsetting trade in composite security markets, ETF arbitrage is not entirely risk free. Specifically, regulatory capital requirements associated with temporary security or ETF positions on intermediaries’ balance sheets as part of ETF arbitrage trades give rise to limits to arbitrage. Pan and Zeng (2019) document that AP balance sheet capacity constraints distort AP arbitrage of corporate bond ETFs. In addition, when ETF track indices composed of illiquid securities that are traded in over-the-counter markets, such as corporate bonds, ETF arbitrage also entails search and matching costs (Koont, Ma, Pastor, & Zeng, 2023). Altogether, these costs increase in the size of ETF creations or redemptions and are thus modeled as variable costs. I assume these balance sheet costs are symmetric for ETF creations and redemptions.

The AP’s profit maximization problem at \( t = 1 \) is

---

19In an extension, I model \( c \) as a function of both, security market \( j \) and aggregate market liquidity conditions. This is consistent with Ma et al. (2022b) who, amongst others, find that MFs had a significant price impact even in generally liquid market segments during the Covid-19 related market sell-off.
\[
\max_{\Theta_{E,AP}^{j}} \Pi_{1}^{AP}(\Theta_{1}^{E,AP,j}) \tag{15}
\]
\[
s.t. \quad \Pi_{1}^{AP} = (P_{1}^{j} - P_{1}^{E,j})\Theta_{1}^{E,AP,j} - \frac{1}{2}\phi_{j}(x_{j})(\Theta_{1}^{E,AP,j})^{2}, \tag{16}
\]
\[
\Pi_{1}^{AP} > 0. \tag{17}
\]

16 is the AP’s profit function and 17 is the AP’s break-even condition. The AP never engages in any ETF arbitrage at a loss. \(P_{1}^{j}\) is the index price and \(P_{1}^{E}\) is the ETF price per share. The AP takes the index price \(P_{1}^{j}\) as given when deciding on its optimal arbitrage strategy. In equilibrium \(P_{1}^{j}\) is determined by market clearing between the AP and its index market maker at \(t = 1\). \(P_{1}^{E}\) will be the result of the AP’s constrained optimal arbitrage activity in ETF markets. \(\Theta_{1}^{E,AP}\) is the AP’s net redemptions of ETF shares. \(\Theta_{1}^{E,AP} < 0\) implies that the AP is creating \(\Theta_{1}^{E,AP}\) units of new ETF shares by delivering a basket of \(\Theta_{1}^{E,AP}\) units of the composite security to the exogenous ETF sponsor. For \(\Theta_{1}^{E,AP} > 0\), the AP is redeeming \(\Theta_{1}^{E,AP}\) existing ETF shares and receiving a basket of \(\Theta_{1}^{E,AP}\) units of the composite security from the ETF sponsor in exchange. \(\phi_{j}(x_{j}) \geq 0\) is the balance sheet capacity parameter that captures any variable transaction costs and balance sheet risks associated with ETF arbitrage. In equilibrium, \(c_{j}(x_{j}) > 0\) generates limits to arbitrage. \(\phi\) is a function of the unconditional index segment \(j\) specific liquidity as well as the contemporary index fundamentals, \(x_{j}\). Generally, creations or redemptions of corporate bond ETFs consume more balance sheet capacity than for large cap equities which can be offloaded from APs’ balance sheet to liquid securities market quickly without significant price impact. The dependence of \(\phi\) on \(j\) captures this relationship. \(\phi_{Corporate\ Bond} > \phi_{Equity}\). Besides, regulatory capital constraints are more binding when fundamentals, and accordingly index market liquidity, are low. The decrease in balance sheet space for ETF arbitrage in times of low fundamentals is reflected by \(\frac{\partial \phi_{j}}{\partial x_{j}} > 0\). Overall, the specification of \(\phi_{j}(x_{j})\) gives rise to countercyclical variation in AP arbitrage intensity within the overall ETF sector as well as cross-sectional variation in AP arbitrage across ETF market segments at any given point in time.

**Assumption 2** Over the long-term, that is at \(t = 2\), APs do not face any balance sheet capacity constraints. They resume ETF arbitrage until all relative mispricing has been eliminated and the ETF and index price have converged. Formally,

\[
c_{2}^{E} = 0. \tag{18}
\]

2.4 Model timeline

The sequence of events and actions in the model is as follows. At \(t = 0\), each investor \(i\) is born with an endowment of one unit of the index, \(\theta_{0}^{i} = 1\). Each investor observes her idiosyncratic liquidity risk exposure, \(\lambda_{i}\) as well as the unconditional distribution of the fundamental index value \(x_{j} \sim N(\mu_{j}, \sigma_{j}^{2})\). Investors then pool their index endowment to form ETFs and open-end
MFs in exchange for fund shares at a 1:1 exchange ratio. Since they cannot own the index directly, they must exchange all of their initial endowment for a portfolio of mutual fund and ETF shares. Neither MFs nor ETFs hold any cash, \( C_0^E = C_0^M = 0 \). They keep all shareholder capital fully invested in the benchmark index.

In the interim period, \( t = 1 \), the fundamental \( x_j \) is realized and observed by everyone. At the same time, investors privately learn about the realization of their idiosyncratic liquidity shock. All investors who receive a liquidity shock liquidate their entire fund portfolio immediately. The remaining, patient investors can choose to liquidate any amount of their fund holdings early and store the proceeds in the risk-free asset until the terminal period, or keep their fund portfolio unchanged until \( t = 2 \). Investors submit orders in MF and ETF markets, taking as given the MF NAV, \( P_{1}^{M,j} \), and ETF price \( P_{1}^{E,j} \).

Given the net supply of ETF shares from investors, the AP chooses how many ETF shares to redeem and submits offsetting orders for the composite security to index market makers. For APs, trading in index and ETF markets takes place simultaneously and intra-day at \( t = 1 \). In contrast, the MF can only start adjusting its portfolio holdings after the market closes and aggregate net MF redemptions are observed. I use \( t = 1^+ \) to denote the time when MFs trade in index markets to rebalance their portfolio and satisfy their obligations to redeeming shareholders at \( t = 1 \). Intuitively, if \( t = 1 \) reflects one trading day, \( t = 1^+ \) can be interpreted as the morning of the subsequent trading day. \( t = 2 \) is then the market close of another trading day. The model sub-periods do not necessary have the same duration. Instead, the time interval between \( t = 0 \) and \( t = 1 \) should be understood as the short-term (days or weeks) whereas the intervals between \( t = 1^+ \) and \( t = 2 \) represents the long-term (months or years).

Over the long-term, at \( t = 2 \) the index pays the terminal dividend, \( P_2^j = x_j \). Investors’ terminal payoff depends on their remaining share holdings of ETFs and MFs, and investment funds’ terminal portfolio value.

### 3 Optimal portfolio allocation between ETFs and MFs

Solving the model entails finding investors’ optimal portfolio allocation to ETFs and MFs, \( \theta_0^E \) and \( \theta_0^M \), and deriving the equilibrium sizes of the ETF and MF sectors. The two period game equilibrium is a pure strategy Nash equilibrium. In every period, \( t = 0 \) and \( t = 1 \), conditional on her endowment, (expected) liquidity needs and fund prices, each investor \( i \) chooses her optimal allocation to ETFs and MFs \( (\theta_i^{E,j}, \theta_i^{M,j}) \) to maximize her expected utility from terminal wealth taking as given fund prices as well as other investors’ portfolio strategy. Formally:

**Definition 2** Given \( \theta_0^i, \lambda_i, \mu_j, \sigma_j, c_j, \phi_j(x_j), \) and \( \eta \), a two-stage pure strategy symmetric Nash
equilibrium is defined as a sequence of portfolio allocations \( \{\theta^{E,j}_i, \theta^{M,j}_i, \theta^{E,j}_i(x), \theta^{M,j}_i(x)\} \in [0,1] \) and asset prices, \( \{P^j_i, P^{E,j}_i, P^{j}_i\} \), such that

(i) given other investors’ allocations at \( t = 1 \), patient investor \( i \)’s portfolio allocation strategy \( \{\theta^{E,j}_i(x_j), \theta^{M,j}_i(x_j)\} \) maximizes her utility function 5 in the interim period \( t = 1 \), and

(ii) given investor \( i \)’s expectation of other investors’ future allocation strategies, and thus the probability of a mutual fund run at \( t = 1 \), \( i \)’s allocation \( \{\theta^{E,i,j}_0, \theta^{M,i,j}_0\} \) maximizes her expected utility 5 at \( t = 0 \), and

(iii) all investors share a common belief about the probability of a mutual fund run at \( t = 1 \), and

(iv) ETF and index markets clear at all times.

I abstract from mixed strategy equilibria because such equilibria are not economically meaningful in this context. I solve for the equilibrium of the model using backward induction. First, I take as given the realization of the fundamental index value \( x_j \) at \( t = 1 \) as well as investors’ initial allocations, \( \theta^{E,j}_0 \) and \( \theta^{M,j}_0 \), and solve for the MF, ETF and index market equilibrium at \( t = 1 \). In the baselines model in which APs and MFs trade with separate index market makers, the ETF and MF market equilibria at \( t = 1 \) are not directly interlinked and can be solved separately. This follows from the assumption that investors can no longer switch between fund types in the interim period. Since MF prices are fixed at the fund NAV over the short-term, MF investors take prices as given. They do not internalize the impact of their trades on the future MF NAV. In contrast, ETF investors take into account their price impact in ETF markets as they trade intra-day at the fully flexible market clearing ETF price. On the other side, due to my focus on passive investment funds, MFs’ index market trades are flow-induced. MFs liquidate whatever amount of composite security shares necessary to meet their investors’ net redemptions. APs instead respond to investors’ ETF orders in real time. When deciding on their optimal arbitrage strategy, they take into account price impact in index and ETF markets as well as their balance sheet capacity. A complication emerges because patient MF investors’ optimal allocation at \( t = 1 \) depends on their beliefs regarding other MF investors’ redemption decisions. The interdependency of investors’ allocation decisions and the possibility of shareholder runs on MFs gives rise to multiple equilibria in MF markets. This challenge has received extensive treatment in the previous literature (e.g., (Chen et al., 2010)). Anticipating this possibility of multiple MF market equilibria, I resort to the sunspot equilibrium selection technique to coordinate MF investors’ behavior and choose the unique MF market equilibrium at \( t = 1 \) following Davila and Goldstein (2022). 20 Second, taking as given investors’ optimal investment policy and

\[ \text{footnote: An alternative approach to overcome the problem of multiple equilibria is the global games technique by Goldstein and Pauzner (2005). The global games solution however requires asymmetric information regarding fundamentals among investors which would compromise the tractability of the model solution. Since the} \]
equilibrium price functions at \( t = 1 \), I solve for investors’ equilibrium allocations at \( t = 0 \) as a function of their idiosyncratic liquidity risks, \( \lambda^i \).

3.1 Fund payoffs in the terminal period

Asset prices at \( t = 2 \) directly follow from the terminal index payoff and are given by

\[
P^j_2 = x_j, \quad (19)
\]
\[
P^{E,j}_2 = P^j_2, \quad (20)
\]
\[
P^{M,j}_2 = \frac{X^{M,j}_2 P^j_2}{\kappa^M + \eta - \nu^M}. \quad (21)
\]

20 follows from assumption 2. The terminal ETF price equals the fundamental value of its benchmark index. There is zero ETF mispricing in the terminal period. APs do not face any balance sheet capacity constraints over the long-term and resume ETF arbitrage until all relative mispricing has been eliminated.21 20 also requires zero ETF tracking error which in turn is a direct consequence of assumption 1 that ETF creation and redemption baskets are always identical to the ETF benchmark index.

By definition, the terminal MF payoff 21 is equal to the fund NAV as of \( t = 2 \). \( X^{M,j}_2 \) is the number index shares held by the MF at \( t = 2 \) after accounting for the fund’s trades in index markets at \( t = 1^+ \). It must satisfy \( 0 \leq X^{M,j}_2 \leq \kappa^M + \eta \). Amid short selling constraints, the least amount of index shares the MF can own is zero. Similarly, without the possibility of fund inflows at \( t = 1 \), the amount of index shares the MF can still own at \( t = 2 \) is bounded from above by the fund’s initial investment. \( \kappa^M \) is the number of MF shares outstanding at \( t = 0 \) after investors initial fund investments. \( \eta \) are the MF shares held by pension funds. \( \nu^M \leq \kappa^M \) is the number of MF shares redeemed early by MF investors in the interim period. If the MF experiences net outflows at \( t = 1 \), \( \nu^M > 0 \). Since there are no more fund flows at \( t = 2 \), all right-hand side variables in 21 are known as of \( t = 1^+ \). Overall, the long-term MF payoff can deviate from the terminal index value when premature liquidations of fund shares at \( t = 1 \) are non-zero, \( \nu^M \neq 0 \).

**Definition 3** The MF tracking difference refers to the difference between the mutual fund net asset value (NAV) and the benchmark index price at any time and is given by

\[
\Delta^{M,j}_t \equiv P^j_t - NAV^{M,j}_t. \quad (22)
\]

MF equilibrium is not a key contribution of this paper, I therefore resort to using the less elegant sunspot equilibrium selection technique.

21 The terminal ETF payoff can be endogenized by specifying a time-varying function for the AP’s balance cost parameter \( \phi_j x_j, t \) in a way that \( \phi_1 x_j, 1 > 0 \) and \( \phi_2 x_j, 2 = 0 \). Without loss of generality, I model the terminal ETF price as exogenous instead.
When $\Delta_{t}^{M,j} > 0$, the MF NAV is smaller than the value of its benchmark index.\textsuperscript{22}

$\Delta_{t}^{M}$ can be interpreted as wedge between the fund and index price. It arises from the cumulative transaction costs incurred at the fund level as a result of early redemptions by MF shareholders.

Following 1, for ETFs, the tracking difference is always zero, $\Delta_{t}^{E,j} = 0$ $\forall t$. ETF investors trading in secondary ETF markets bear their own transaction costs. Their actions do not directly affect the ETF’s net asset value but only its secondary market price. Outflows from ETFs only have temporary effects on its investors’ payoff, whereas outflows from MFs can have permanent effects on investors’ payoffs.

**Lemma 2**  
In the terminal period the MF tracking difference is given by

$$
\Delta_{t=2}^{M,j} = \frac{P_{t=2}^{j} \left( \nu_{t=1}^{M} (NAV_{t=1}^{M,j} - P_{t=1}^{j}) \right)}{P_{t=1}^{j} + \kappa_{t=1}^{M} + \eta_{t=1} - \nu_{t=1}^{M}}.
$$

$P_{t=1}^{j}$ is the index price at which the MF trades when liquidating portfolio holdings in response to MF investors’ redemptions in the interim period.

**Corollary 1**  
The MF tracking difference in the terminal period is zero, $\Delta_{t=2}^{M,j} = 0$, iff at least one of the following two conditions is satisfied

(i) Net mutual fund outflows at $t = 1$ are zero, $\nu_{t=1}^{M} = 0$.

(ii) The mutual fund NAV at $t = 1$ is perfectly forward looking and equal to the index price at which MFs trade in index markets, $NAV_{t=1}^{M,j} = P_{t=1}^{M,j}$.

If MFs have no price impact in index markets ($c_{j} = 0$), (ii) is always satisfied. Alternatively, (ii) would hold if the MF used swing pricing and swing factors are calibrated optimally to reflect all flow induced transaction costs.

**Corollary 2**  
When at least one of the conditions in corollary 1 are satisfied, the mutual fund NAV at $t = 2$ is equal to the terminal index payoff. In this case, the MF and ETF tracking benchmark index $j$ provide identical payoffs over the long-term, $P_{t=2}^{M,j} = P_{t=2}^{E,j} = P_{t=2}^{j}$.

Equation 23 illustrates how frictions in MFs arise from flow induced share dilution. The imperfect flexibility of the MF NAV relative to prices in underlying security markets, is the fundamental source of tracking error, $\Delta_{t=2}^{M,j} \neq 0$, over the long-term. When $NAV_{t=1}^{M,j} - P_{t=1}^{j} > 0$, short-term investors’ payoffs exceeds the current liquidation value of their share holdings. Accordingly, the MF provides short-term investors with liquidity insurance in the amount of

\textsuperscript{22}Here the tracking difference is defined based on fund and index prices. In practice, the tracking difference generally refers to the difference in the returns of a fund and its benchmark index.
NAV_1^{M,J} - P_1^{j+}. This liquidity provision to short-term investors is not free. The relatively higher payoff received by short-term investors comes at the cost of long-term investors. In this sense, Δ_2^{M,J} can also be interpreted as a negative risk premium or an insurance premium paid in exchange for the short-term liquidity insurance service provided by MFs. It represents a redistribution of assets from long-term to short-term investors. Consistent with classical insurance problems, mutual fund markets are characterized by adverse selection and moral hazard due to asymmetric information between shareholders and mutual funds. Ex-ante, shareholders can freely choose to allocate their portfolio between ETFs and MFs irrespective of their idiosyncratic liquidity risk \( \lambda_i \) which is private information. There is no way for funds to screen their investors. Ex-post, in contrast to typical insurance contracts, it is impossible for mutual funds to distinguish between impatient investors and patient investors who choose to redeem their shares early due to concerns about a MF run. As a result, if not only impatient but also patient investors choose to exploit the MFs’ short-term liquidity insurance, the ex-post cost of the MFs’ liquidity provision, Δ_2^{M,J}, will be inefficiently high.

### 3.2 Portfolio allocations in the interim period

At \( t = 1 \) the idiosyncratic liquidity shocks are realized. Each agent learns if she is the impatient (\( e \)) or patient (\( l \)) type. Besides, the fundamental index value \( x_j \) is revealed and becomes common knowledge.

**Lemma 3** Conditional on the realization of the idiosyncratic liquidity shocks at \( t = 1 \), there are four groups of ex-post identical investors: Patient- and impatient ETF investors, and patient- and impatient MF investors. These investors are characterized by the following conditions

(i) **Patient ETF investors**: All \( i = l \) who initially invested in ETFs, \( \theta_0^{l,E} = 1 \)

(ii) **Impatient ETF investors**: All \( i = e \) who initially invested in ETFs, \( \theta_0^{l,E} = 1 \).

(iii) **Patient MF investors**: All \( i = l \) who initially invested in MFs, \( \theta_0^{l,M} = 1 \).

(iv) **Impatient MF investors**: All \( i = e \) who initially invested in MFs, \( \theta_0^{l,M} = 1 \).

Solving for the fund market equilibrium at \( t = 1 \) entails separately deriving the optimal portfolio choice of each of these investor groups.

(i) **Patient ETF investors** are investors who invested in ETFs at \( t = 0 \) and received no idiosyncratic liquidity shock. At \( t = 1 \), they must decide if they want to liquidate any of their ETF shares early. Formally, their problem is given by
\[
\max_{\theta_1^{i,E}} E[w_2^j|x]
\]
\[
s.t. \quad w_2^j = \theta_1^{i,E} P_2^E + (\theta_0^{i,E} - \theta_1^{i,E}) P_1^E R_f,
\]
\[
\theta_0^{i,E} - \theta_1^{i,E} \geq 0,
\]
\[
\theta_1^{i,E} \geq 0.
\]

It follows directly from the assumption that investors are no longer able to switch between fund types in the interim period. Rational investors anticipate that the ETF mispricing will converge to zero in the terminal period, such that \( P_2^{E,j} = P_2^j \). Since the terminal index payoff is realized and observed at \( t = 1 \), \( E[P_2^j|x] = x_j \). There remains no uncertainty regarding the final ETF payoff. Hence, patient ETF investors optimal portfolio allocation at \( t = 1 \) only depends on the current ETF price, the fundamental index value and the risk-free rate. Correspondingly, in the absence of any residual uncertainty, patient ETF investors’ allocation at \( t = 1 \) is independent of their risk-aversion. However, investor risk-aversion plays a role in their initial allocation.

**Assumption 3** *If the payoff from holding the ETF until the terminal period is equal to the payoff from immediate liquidation at \( t = 1 \), patient investors always hold the investment fund until maturity.*

Using assumption 3, patient ETF investors’ optimal investment policy is given by

\[
\theta_1^{i,E} = \begin{cases} 
0 & \text{if } P_2^j - P_1^{E,j} R_f < 0, \\
1 & \text{if } P_2^j - P_1^{E,j} R_f \geq 0,
\end{cases}
\]

To illustrate the dependence of patient ETF investors portfolio allocation decision at \( t = 1 \) on the key friction in ETF markets, that is the ETF mispricing, \( \epsilon_{1}^{E,j} \), the optimal allocation in 25 can be rewritten as

\[
\theta_1^{i,E} = \begin{cases} 
0 & \text{if } P_2^j - (P_1^j - \epsilon_{1}^{E,j}) R_f < 0, \\
1 & \text{if } P_2^j - (P_1^j - \epsilon_{1}^{E,j}) R_f \geq 0,
\end{cases}
\]

The ETF discount \( \epsilon_{1}^{E,j} \) is the liquidity premium ETF investors pay in exchange for short-term liquidity provision. The larger the current ETF discount, the greater the payoff from waiting to liquidate ETF shares until the terminal period. If the expected payoff of the ETF in the terminal period is smaller than the current ETF share price reinvested at the risk-free rate, \( x_j < P_1^E R_f \), investors always liquidate all of their ETF shares prematurely. In the absence of ETF inflows, this condition is never satisfied in the competitive equilibrium.
Proposition 1 In equilibrium, patient investors never liquidate any ETF shares early,

\[ \theta_{i,E,j}^1 = \theta_{i,j}^0 \quad \forall i = l \text{ with } \theta_{i,E,j}^0 = 1. \]  

(ii) **Impatient ETF investors** are investors who invested in ETFs at \( t = 0 \) and received an idiosyncratic liquidity shock. Their terminal wealth (consumption) is given by

\[ w_2 = \theta_0 P_{1,E,j}^1. \]  

By definition, \( u(c_2^i) = u(w_1^i) \quad \forall i = e. \)

(iii) **Patient MF investors** are investors who invested in MFs at \( t = 0 \) and received no idiosyncratic liquidity shock. Like patient ETF investors, at \( t = 1 \) they can decide to liquidate any of their MF shares early. They solve the following problem:

\[
\begin{align*}
\max_{\theta_{i,M,j}^1} & \quad E[w_2^i|x] \\
\text{s.t.} & \quad w_2^i = \theta_{i,M,j}^1 P_{2,M,j}^1 + (\theta_{i,M,j}^0 - \theta_{i,M,j}^1) P_{1,M,j}^1 R_f, \\
& \quad \theta_{i,M,j}^0 - \theta_{i,M,j}^1 \geq 0, \\
& \quad \theta_{i,M,j}^1 \geq 0.
\end{align*}
\]  

Patient MF investors trade off the payoff from early redemption, \( P_{1,M,j}^1 R_f \), with the expected payoff from holding on to their MF shares until the final period. However, in contrast to the terminal ETF payoff, the terminal MF payoff is uncertain even at \( t = 1 \). Beyond \( P_{2}^j \), \( P_{2,M,j}^1 \) depends on the early redemption decision of other MF investors. This interdependency arises because the MF tracking difference, \( \Delta_{2,M,j} \), depends on past fund flows. By definition,

\[ E[P_{2,M,j}^1|x] = E[NAV_{2,M,j}^1|x_j] = E[P_{2}^j - \Delta_{2,M,j}^1|x_j] = x_j - E[\Delta_{2,M,j}^1|x_j]. \]  

The solution for \( \theta_{i,M,j}^1 \) can be a boundary or interior solution. Patient MF investors optimal MF allocation at \( t = 1 \) as a function of the expected MF tracking error is then given by

\[
\theta_{i,M,j}^1 = \begin{cases} 
0 & \text{if } P_{2}^j - E[\Delta_{2,M,j}^1|x] - P_{1,M,j}^1 R_f \leq 0, \\
1 & \text{if } P_{2}^j - E[\Delta_{2,M,j}^1|x] - P_{1,M,j}^1 R_f \geq 0.
\end{cases}
\]  

According to 31, at \( t = 1 \) patient MF investors trade off the payoff from early withdrawal, \( P_{1,M,j}^1 = \psi x_j \), and waiting until \( t = 2 \). In the latter case, they receive an uncertain payoff, \( E[P_{2,M,j}^1|x] = P_{2}^j - E[\Delta_{2,M,j}^1|x] \), which depends on the known fundamental index value as well as the uncertain redemption decisions by other MF investors. Hence, key to solving patient MF investors’ problem at \( t = 1 \) is solving for the expected MF tracking error per MF share conditional on the state variable, \( E[\Delta_{2,M,j}^1|x] \).
Impatient MF investors are investors who invested in MFs at \( t = 0 \) and received an idiosyncratic liquidity shock. The final wealth of these investors is given by

\[ W_2^i = W_1^i R^f = \psi x_j. \] (32)

### 3.3 Equilibrium prices in the interim period

Equilibrium prices in the interim period are determined by market clearing between the MF, APs and index market makers given ETF and MF investors’ fund liquidations. In the baseline model with segmented index markets, prices are set in the following sequence: First, at \( t = 1 \) the ETF price, \( P_1^{E,j} \), is determined relative to the index price from the AP’s optimal ETF arbitrage (redemption) policy and investors’ net demand for ETF shares. Simultaneously, the index price faced by the AP, \( P_1^{j*} \), follows from the other side of the AP’s arbitrage trade and the net the market maker’s net demand schedule for index shares. Second, given MF investors’ net redemptions at \( t = 1 \), at \( t = 1^+ \) the index price faced by MFs, \( P_{1^+}^{j*} \) is determined from market clearing between the MF’s flow induced index sales and its market maker’s net demand schedule for index shares. Thereby, \( P_{1^+}^{j*} \) is derived independently of \( P_1^{j*} \) in line with the timing differences between the AP’s and the MF’s index trading activity.

In an extension, I relax the assumption that index markets are segmented and allow APs and MFs to trade with the same index market maker. I show that the key model predictions continue to hold. Yet, the formal equilibrium results become substantially more complicated. For tractability the rest of the paper builds on the simplified baseline model with segmented index markets.

**ETF price.** \( P_1^{E,j} \) follows from market clearing between APs and ETF investors, taking as given the equilibrium index price \( P_1^{E,j_1} \). Since patient ETF investors never redeem early, the aggregate number of ETF shares sold by investors at \( t = 1 \), denoted by \( \nu^E \), is equal impatient investors’ liquidations,

\[ \nu^E = \int_i (\theta_0^{E,j_1} - \theta_1^{E,j_1}) \, di. \] (33)

From 15 follows AP’s net demand function for ETF shares at \( t = 1 \),

\[ \Theta_1^{E,j,AP} = \frac{(P_1^j - P_1^{E,j})}{\phi_j(x_j)}. \] (34)

\( \Theta_1^{E,j,AP} \) refers to the number of ETF shares redeemed by APs. ETF market clearing requires \( \Theta_1^{E,j,AP} = \nu^E \). Hence, given the index price and ETF outflows by impatient investors, the ETF price is

\[ P_1^{E,j} = P_1^j - \nu^E \phi_j. \] (35)
\( \nu^E \phi_j \) is the relative ETF mispricing, \( \epsilon_1^{E,j} = \nu^E \phi_j(x_j) \). If \( \nu^E > 0 \), the ETF trades at a discount to its benchmark index because \( \phi_j(x_j) > 0 \) \( \forall x_j \). Intuitively, \( \ref{eq:34} \) implies that when the ETF trades at a discount (premium), the AP will choose to redeem (create) ETF shares. As a result, net supply of ETF shares by secondary market investors, \( \nu^E > 0 \), is associated with ETF discounts while net demand for ETF shares, \( \nu^E < 0 \), will lead the ETF to trade at a premium relative to the fund NAV. The ETF discount also increases with the benchmark segment specific illiquidity. \( \phi_j \) is larger for more illiquid market segments. The latter reflects the empirical fact that ETFs tracking benchmark indices in more illiquid market segments, such as corporate bond ETFs, tend to have larger relative mispricings compared to ETFs tracking more liquid benchmarks (e.g., large cap domestic equity ETFs) across the cycle. Besides, with \( \frac{\partial \phi_j}{\partial x_j} < 0 \), AP balance sheet capacity constraints, and thus ETF mispricing are countercyclical across all index segments.

**Index price faced by APs.** The index price faced by APs at \( t = 1 \) then follows from market clearing between the index market maker’s net demand for index shares, \( \Theta_{1, D,j} = \frac{E[P^j_2|X] - P^E_{1j}}{c_j} \), and the AP’s net supply of index shares. The net demand for index shares by market makers is defined over \( P^j_1 \in [0, x_j] \). From the AP’s optimization problem and the ETF market clearing condition, the AP’s net supply of index shares follows as

\[
\Theta_{1, AP}^{j} = \frac{P^E_{1j} - (P^E_{1j} - \nu^E \phi_j)}{\phi_j}. \tag{36}
\]

\( \Theta_{1, AP}^{j} = \Theta_{1, E, AP}^{j} \) since the AP can always exchange one unit of the index for one ETF share. Then, index market clearing requires \( \Theta_{1, D,j} = \Theta_{1, AP}^{j} \).

In the case in which there is price impact in index markets (\( c_j > 0 \)) \( \frac{E[P^j_2|X] - P^j_1}{c_j} = \nu^E \). \( \ref{eq:24} \)

The \( t = 1 \) index price is

\[
P^j_1 = E[P^j_2|X] - \nu^E c_j. \tag{37}
\]

For the ETF price it implies

\[
P^{E,j}_1 = E[P^j_2|X] - \nu^E (c_j + \phi_j(x_j)). \tag{38}
\]

The relative ETF mispricing at \( t = 1 \) is strictly increasing in the AP’s balance sheet capacity constraint and the selling pressure from ETF investors

\[\text{In the alternative scenario with } c_j = 0 \text{, market makers net demand for index shares is perfectly inelastic. Index markets are very liquid and APs do not have price impact. This condition represents the benchmark for very liquid market segments, such as large cap domestic equities. Under this scenario, the unique equilibrium index price faced by APs as well as MFs at } t = 1 \text{ and } t = 1^+ \text{ respectively is given by } P^j_1 = P^j_{1+} = x_j.\]
ETF prices adjust in real time to supply and demand conditions in financial markets. ETF investors internalize the price impact in index markets caused by AP arbitrage. Yet, in the presence of AP balance sheet capacity constraints, \( \phi_j(x_j) > 0 \), ETF prices are too flexible over the short-term, \( P_{1}^{E,j} < P_{1}^{j} \).

**Mutual fund price.** The MF NAV at which investors transact at \( t = 1 \), is exogenously given by

\[
P_{1}^{M,j} \equiv \psi E[P_{2}^{j}|x],
\]

where \( 0 < \psi < 1 \) reflects the exogenous deadweight loss associated with premature fund liquidations. Formally, \( \psi < 1 \) ensures that, in the absence of payoff complementarities, early redemptions would never be optimal for patient investors. If \( P_{1}^{M,j} \geq E[P_{2}^{j}|x] \), patient MF investors would always optimally choose to redeem their shares early even if they do not bear any losses from the redemptions by impatient investors.

In practice, trades of most MFs in the U.S., with the exception of money market MFs, settle within one business day after the trade date \( T + 1 \). For example, in the case of MF redemptions, investors receive the cash proceeds from their sale with a delay of one business day. In my model investors receive the cash from their ETF and MF transactions right away on the same business day. I follow this convention for simplicity. This assumption does not affect the model predictions. Instead, accounting for the time lag between the settlement of MF trades at \( t = 1 \) and MF’s trading activity in index markets at \( t = 1^{+} \), I make another simplifying assumption. Mutual funds initially meet net fund redemptions by drawing down credit lines held at exogenous intermediaries. I assume they can borrow at the risk-free rate overnight. This is only a technical assumption to abstract from funds’ liquidity management and inconsequential for my key results.}

\[25\] See SEC settlement cycle recommendation. Generally, the settlement cycle for transactions of publicly traded securities, including equities, corporate bonds, municipal bonds, unit investment trusts and ETFs, in the U.S. is \( T + 2 \). Money market MFs tend to settle \( T + 0 \) or \( T + 1 \). On February 15, 2023, the Securities and Exchange Commission (SEC) adopted an amendment to an existing rule to further reduce the settlement cycle for standard securities transactions to \( T + 1 \), see SEC release Nos. 34-96930, IA-6239; File No. S7-05-22.

\[26\] In an extension, I consider the case in which MFs are allowed to hold cash and meet net redemptions by depleting cash reserves.
I denote the number shares redeemed from MFs at $t = 1$ by

$$\nu^M = \kappa^M - \int_0^1 \theta^M_{i+1} di. \quad (41)$$

$\kappa^M = 1 - \kappa^E$ is the fraction of investors who invested in MFs at $t = 0$. The remaining number of MF shares outstanding as of $t = 1^+$ is $\kappa^M + \eta - \nu^M$. Even in the case of a MF run in which all patient MF investors redeem their share early, the presence of the 'sticky' pension fund investors ensures that the MF never disappears, $\kappa^M \geq \eta > 0$. Since pension funds never sell any MF shares and there are no inflows at $t = 1$, the total number of MF share redemptions at $t = 1$ must satisfy $0 \leq \nu^M \leq 1$.

**Index price faced by MFs.** At $t = 1^+$ the MF liquidates index shares to repay its obligations following investor redemptions at $t = 1$. The index price at $t = 1^+$ follows from market clearing between the market maker’s net demand for index shares, $\Theta^{D,j}_{1^+} = \frac{E[P^j_{1^+} | \omega] - P^j_{1^+}}{c_j}$, and MFs flow-induced trading in index markets. The net demand for index shares by market makers is defined over $P^j_{1^+} \in [0, x]$. The supply of index shares by MFs is

$$\Theta^{M,j}_{1^+} = \frac{\nu^M P^M_{1^+}}{P^j_{1^+}}. \quad (42)$$

The maximum number of index shares which the MF can sell to meet investor redemptions is bounded by its portfolio holdings according to $\Theta^{M,j}_{1^+} \leq \kappa^M + \eta$. The MF cannot engage in short sales, $0 \leq \Theta^{M,j}_{1^+}$.

In the special case in which $c_j = 0$, the number of index shares sold by MFs is then given by $\Theta^{M,j}_{1^+} = \psi \nu^M$. When index markets are perfectly liquid, there are negative externalities among MF investors. Abstracting from potential taxable distributions of capital gains and trading commissions, there are no transaction costs associated with fund outflows. Accordingly, if $c_j = 0$, patient MF investors would never redeem MF shares early just like patient ETF investors.

**Lemma 4** If $c_j = 0$, patient MF investors never choose to redeem any MF shares early in equilibrium. There are no run risks in the index MF.

This result is consistent with prior studies documenting higher run risk in more illiquid fund market segments (e.g., Chen et al. (2010)).

In the more general case in which $0 < c_j \leq 1$, investment funds have price impact in index markets. The market clearing index price at $t = 1^+$, follows from $\Theta^{M,j}_{1^+} = \Theta^{D,j}_{1^+}$ and solves

$$\frac{E[P^j_{1^+} | \omega] - P^j_{1^+}}{c_j} = \frac{\nu^M P^M_{1^+}}{P^j_{1^+}}. \quad (43)$$
It exists as long as the fundamental index value $x$ is be large relative to market makers’ inventory costs, $c_j$.

**Lemma 5** In the case with $c_j > 0$, iff $x_j \geq 4c_j\psi$, $\exists P_{1+}^j$ that solves 43 over $0 \leq \Theta_{1+}^{M,j} \leq 1 + \eta$.

To establish the uniqueness of this equilibrium index price at $t = 1^+$, I impose an additional assumption.

**Assumption 4** If there exist a pair of candidate prices and quantities, $\{P_{1+}', \Theta_{1+}'\}$, that satisfy 43, the equilibrium index market price is given by the larger price candidate

$$P_{1+}^j = P_1' \text{ if } P_1' \geq P_1''$$

and $\Theta_{1+}^j = \Theta_1'$.

Assumption 4 is consistent with a model in which market makers and MF managers engage in price negotiations for their trade defined by $\{P_1', \Theta_1'\}$. They start at $P^j = x_j$ and adjust the index price downward until the market clears.

**Proposition 2** If $c_j > 0$ and $x_j - \frac{1}{2}(x_j - \sqrt{x_j^2 - 4c_j\psi x_j^3}) \leq 1 + \eta$, for any amount of MF redemptions $\nu^M \in [0, 1]$, the unique index market equilibrium at $t = 1^+$ is given by

$$P_{1+}^j = \frac{1}{2}(x_j + \sqrt{x_j^2 - 4c_j\psi x_j^3})$$

and

$$\Theta_{1+}^j = \frac{x_j + \frac{1}{2}(x_j - \sqrt{x_j^2 - 4c_j\psi x_j^3})}{c_j}$$

**Corollary 3** In the special case in which MF net redemptions at $t = 1$ are given by

$$\nu^M = \frac{(1 - \psi)x_j}{c_j},$$

the equilibrium index price at $t = 1^+$ is exactly equal to the MF net asset value at $t = 1$, $P_{1+}^j = \psi x_j$. Then, $\Theta_{1+}^j = \nu^M$, and the MF tracking error is zero, $\Delta_{2}^{M,j} = 0$. In this special case, ex-post the fund NAV at $t = 1$ is perfectly forward looking.

Ex-post, this equilibrium is comparable to the case in which the MF employed swing pricing to ensure that exiting investors at $t = 1$ bear the transaction costs associated with their redemptions. Ex-ante it differs from a swing pricing equilibrium due to the effect of swing pricing policies on investors’ expectations regarding others’ redemption decisions.

**Corollary 4** When $\nu^M \neq \frac{(1 - \psi)x_j}{c_j}$, the equilibrium index price at which the MF trades at $t = 1^+$ deviates from the MF NAV at which fund investors can redeem shares at $t = 1$.
i. If \( \nu^M < \frac{(1-\psi)x_j}{c_j} \), the equilibrium index price at \( t = 1^- \) is larger than the MF NAV at \( t = 1 \), \( P_{1+}^j > \psi x_j \). Then, \( \Theta_{1+}^{M,j} < \nu^M \), and the terminal MF tracking difference becomes negative, \( \Delta_2^{M,j} < 0 \). The remaining MF shareholders gain from other investors’ redemptions at \( P_{1}^{M,j} \) because the latter sell at a fund NAV that is too low given fundamentals.

ii. If \( \nu^M > \frac{(1-\psi)x_j}{c_j} \), the equilibrium index price at \( t = 1^- \) is lower than the MF NAV at \( t = 1 \), \( P_{1+}^j < \psi x_j \). Then, \( \Theta_{1+}^{M,j} > \nu^M \), and the terminal MF tracking difference becomes positive, \( \Delta_2^{M,j} > 0 \). The MF NAV at \( t = 1 \) is too high as it does not fully account for the full price impact associated with investor redemptions at \( t = 1 \).

### 3.4 Equilibrium allocations in the interim period

Investors’ equilibrium portfolio allocations in the interim period, \( \theta_{1,E,\cdot}^* \) and \( \theta_{1,M,\cdot}^* \), directly follow from market clearing in the ETF, MF and index markets at \( t = 1 \) and investors optimal investment strategies.

**ETF market equilibrium.** From proposition 1 follows

\[
\theta_{1,E,j}^* = \begin{cases} 
1 & \forall i = l \text{ (patient types)} \text{ with } \theta_{0,E,j}^* = 1, \\
0 & \forall i = e \text{ (impatient types)} \text{ with } \theta_{0,E,j}^* = 1 
\end{cases}
\]  

(48)

and

\[
P_{1,E,j}^* = x_j - \nu^*(c_j + \phi_j(x_j)).
\]  

(49)

\( \nu^* = \int \theta_{0,E,j}^* - \theta_{0,E,j} \, di \) is simply the fraction of ex-post impatient ETF investors. The higher the fraction of impatient ETF investors at \( t = 1 \), the larger the ETF mispricing, and therefore the lower the incentive for any remaining patient ETF investors to liquidate their shares early.

**MF market equilibrium.** Generally, patient MF investors’ early redemption incentive is increasing in the expected volume of redemptions by other MF investors, \( \uparrow E[\nu^M|x_j] \), index market price impact, \( \uparrow c_j \), and decreasing in the fundamental, \( \downarrow x_j \), as well as the share of sleepy investors \( \downarrow \eta \). Notwithstanding, the MF market at \( t = 1 \) is characterized by the possibility of multiple equilibria. Each MF investors’ optimal allocation strategy \( 31 \) is a function of the expected terminal MF tracking error which in turn depends on the allocation decisions of all MF investors at \( t = 1 \). To show this formally, I first analyze two regions of very bad and very good fundamentals, where each patient MF investor’s optimal MF allocation is independent of her beliefs of other patient MF investors’ actions. In line with the bank run literature, I refer to these regions of fundamental values as the lower and upper dominance regions. In these two regions, patient MF investors allocations at \( t = 1 \) are only based on the fundamental and general model parameters.
**Lower dominance region.** The lower dominance region encompasses values of the fundamental in the range \( x_j \in (0, \underline{x}_j] \). In this region, fundamentals are sufficiently bad, such that early redemption is the dominant strategy for any individual patient investor. After observing \( x_j \leq \underline{x}_j \), investors will redeem all of their MF shares early even if all other patient MF investors choose not to redeem their shares early. I denote the mass of impatient MF investors at \( t = 1 \) by \( e^M \). Then, the lower dominance region is characterized by the value \( \underline{x}_j \) that solves

\[
E[P_2^j - \Delta_2^{M,j} - P_1^{M,j} R^f | x_j = \underline{x}_j \cup \nu^M = e^M] = 0. \tag{50}
\]

**Upper dominance region.** The upper dominance region is defined by \( x_j \in [\overline{x}_j, \infty) \). In this region, fundamentals are extremely good, such that patient MF investors always choose to keep all of their MF shares until \( t = 2 \). After observing \( x_j \geq \overline{x}_j \), investors will never redeem any of their MF shares early even if all other patient MF investors are liquidating at \( t = 1 \). The mass of all patient and impatient MF investors at \( t = 1 \) is \( \kappa^M = e^M + l^M \). Then, the upper dominance region is characterized by the value \( \overline{x}_j \) that solves

\[
E[P_2^j - \Delta_2^{M,j} - P_1^{M,j} R^f | x_j = \overline{x}_j \cup \nu^M = \kappa^M] = 0 \tag{51}
\]

**Proposition 3** There exists a lower and upper dominance region with respect to the fundamental that is characterized by the boundary values \( \underline{x}_j \) and \( \overline{x}_j \), with \( \underline{x}_j \leq \overline{x}_j \), such that for any realization of the fundamental within these regions (\( x_j \leq \underline{x}_j \) or \( x_j \geq \overline{x}_j \)) the MF market at \( t = 1 \) has an unique equilibrium in which patient MF investors always redeem all of their shares early, or never redeem any shares early at all, irrespective of their beliefs of other patient MF investors portfolio allocations. Formally, \( \exists \underline{x}_j, \overline{x}_j \) s.t.

\[
\begin{align*}
\theta_1^{i,M,j}*(x_j \leq \underline{x}_j) &= 0, \tag{52} \\
\theta_1^{i,M,j}*(x_j \geq \overline{x}_j) &= 1. \tag{53}
\end{align*}
\]

If additionally \( \underline{x}_j < \overline{x}_j \), the range of \( x_j \) over which which multiple equilibria exist in the MF market, \( x_j \in (\underline{x}_j, \overline{x}_j) \) is non-empty.

The **lower dominance region** in which a MF run is the unique equilibrium outcome irrespective of patient MF investors beliefs of others’ actions is defined by

\[
\underline{x}_j = \frac{c_j \psi (\kappa^M + \eta + e^M \psi R^f - \kappa^M \psi R^f - \eta \psi R^f)^2}{(1 + \psi)(e^M(-1 + \psi(R^f - 1)) - (\kappa + \eta)(\psi R^f - 1))}. \tag{54}
\]
The size of the lower dominance region does not depend on investors’ risk aversion, $\gamma$. $x_j$ is the same for risk-neutral and risk-averse (mean-variance) investors.

The upper dominance region in which all patient MF investors choose not to sell any of their MF shares irrespective of their beliefs of other patient MF investors’ actions is defined by the condition

$$\bar{x}_j = \frac{c_j \psi (\kappa^M + (1 - \psi R^I)\eta)^2}{(1 + \psi)(\eta(1 - \psi R^I) - \kappa^M \psi)}.$$  \hspace{1cm} (55)

Intuitively, as the fundamental improves, the price impact of MF in index markets becomes negligible, whereby the penalty for early liquidation increases in absolute terms. Fund outflows at $t = 1$ do not impose a significant cost on the remaining investors when expected long-term returns are sufficiently high. The relative importance of payoff complementarities in patient MF investors’ allocation decision diminishes with $x_j$.

**Corollary 5** The size of the run (lower dominance) region $(0, \bar{x}_j]$ is increasing in $c_j$, the illiquidity of the index segment, as well as in $e^M$, the mass of impatient MF investors. The size of the run region is decreasing in $\eta$, the share of sleepy MF investors who never redeem early.

This result directly follows from the partial derivatives of 54. $\frac{\partial x}{\partial c_j} > 0$ implies that funds with more illiquid portfolios are more prone to runs by investors. In contrast, $\frac{\partial x}{\partial \eta} > 0$. the larger the base of sleepy or long-term investors, $\eta$, the smaller the run region. These investors stabilize the MF as other investors rationally never anticipate them to redeem early, thereby decreasing their own run incentives. These results are consistent with the previous literature on fragility in open-end mutual funds (Chen et al., 2010). Finally, $\frac{\partial x}{\partial e_j} > 0$ suggests that a larger base of investors with high short-term liquidity needs or shorter investment horizons makes the MF inherently more prone to run risk.

**Corollary 6** The size of the no-run (upper dominance) region $[\bar{x}_j, \infty)$ is decreasing in $c_j$, the illiquidity of the index segment, as well as in $\kappa^M$, the total mass of non-sleepy investors who initially invested in MFs and may potentially redeem their shares early at $t = 1$. The size of the no-run region is increasing in $\eta$, the share of sleepy MF investors who never redeem early.

**Corollary 7** In the special case in which $c_j = 0$ and MFs do not have price impact in index markets, $\bar{x}_j = \bar{x}_j = 0$. The region of $x_j$ over which a MF run is the unique or just one possible equilibrium outcome is empty. There are no externalities among MF investors and MFs never occur in equilibrium. This is true for the baseline case with risk-neutral investors as well as for for investors with CRRA utility and $\gamma > 0$.  

35
**Corollary 8** For \( x_j \in (\underline{x}_j, \overline{x}_j) \), there exist multiple equilibria in MF markets at \( t = 1 \) depending on investors beliefs regarding other patient investors’ redemption strategy, \( E[\nu^M | x_j] \).

I overcome the multiplicity of equilibria, I assume that over the region \((\underline{x}_j, \overline{x}_j)\) there is an i.i.d. sunspot for every realization of \( x_j \). Specifically, I assume that investors believe that the run (no-run) equilibrium will materialize with the probability \( \pi^{\text{Run}} (1 - \pi^{\text{Run}}) \) for every \( x_j \in (\underline{x}_j, \overline{x}_j) \). For simplicity, I assume that the run and no-run equilibria are perceived as equally likely, \( \pi^{\text{Run}} = 0.5 \).

**Assumption 5** For any \( x_j \in (\underline{x}_j, \overline{x}_j) \), investors beliefs are such that they expect all other patient MF investors to run versus not run with equal probabilities. Formally,

\[
\pi^{\text{Run}} \equiv \text{Prob}(\nu^M = e^M | x_j < x_j < \overline{x}_j) = 0.5 \tag{56}
\]

\[
\pi^{\text{No-Run}} \equiv \text{Prob}(\nu^M = e^M | x_j < x_j < \overline{x}_j) = 0.5 \tag{57}
\]

### 3.5 Equilibrium allocations in the initial period

Investors’ demand function for ETFs and MFs at \( t = 0 \) is a function of their idiosyncratic liquidity risk exposure, \( \lambda_i \), the unconditional distribution of the terminal index payoff \( x_j \), the mass of sleepy MF investors, \( \eta \), as well as the liquidity of the index market, \( c_j \) and \( \phi_j \). Any cross-sectional variation in investors’ initial portfolio allocation is due to heterogeneity in \( \lambda_i \) because all investors have identical preferences and there is no asymmetric information regarding the index fundamental. Accordingly, in the initial period, the model is characterized by a separating equilibrium. Investors self-select into different fund types based on their idiosyncratic liquidity needs. Formally, since investors endowment is indivisible and they can only invest in one fund at \( t = 0 \), they simply trade off their expected lifetime utility from investing in the ETF with their expected utility from investing in the MF. In the risk-neutral benchmark these are

\[
E_0[u(w^i)|\theta_0^{i,E,j} = 1] = \lambda_i E_0[P_{1}^{E,j} R^f] + (1 - \lambda_i) E_0[P_{2}^{E,j}]
\]

\[
= \lambda_i \left( \mu_j - E_0[e^E](c_j + \phi_j(\mu_j)) \right) R^f + (1 - \lambda_i) \mu_j \tag{58}
\]

where \( E_0[e^E] \) is the expected mass of impatient ETF investors, and

\[
E_0[u(w^i)|\theta_0^{i,M,j} = 1] = \lambda_i \psi \mu_j R^f + (1 - \lambda_i) \left( R^f \int_{\underline{x}_j}^{\overline{x}_j} \psi x_j dx + \frac{1}{2} R^f \int_{\underline{x}_j}^{\overline{x}_j} \psi x_j dx \right)
\]

\[
+ \frac{1}{2} \int_{\underline{x}_j}^{\overline{x}_j} (x_j - \Delta_2^{M,j}) dx + \int_{\overline{x}_j}^{\infty} (x_j - \Delta_2^{M,j}) dx \tag{59}
\]
Proposition 4 Investors with high liquidity risk or short-term investment horizon self-select into mutual funds, while investors with low liquidity risk or long-term investment horizon will self-select into ETFs:

\[
\exists \lambda' \text{ for which } \{\theta_0^{i,ETF}, \theta_0^{i,MF}\} = \{1, 0\} \forall i \text{ with } \lambda_i < \lambda'
\]

and \( \{\theta_0^{i,ETF}, \theta_0^{i,MF}\} = \{0, 1\} \forall i \text{ with } \lambda_i > \lambda' \).

Corollary 9 The investor at the boundary with \( \lambda_i = \lambda' \) is indifferent between allocating her entire portfolio to ETFs or MFs at \( t = 0 \).

At \( t = 0 \), the fund sponsors receive the index shares from investors and convert them into MF and ETF shares respectively at an exchange rate of one. Since fund shares are created outright at \( t = 0 \), initially, both funds and the index have the same price. ETF mispricing and MF tracking error are zero at \( t = 0 \):

\[
P_{0}^{j} = P_{0}^{M} = P_{0}^{E}.
\]

I normalize \( P_{0}^{j} = P_{0}^{M} = P_{0}^{E} = 1 \).

Let \( 0 \leq \kappa^E \leq 1 \) be the share of investors who initially choose to invest in ETFs. \( \kappa^M = 1 - \kappa^E \) is the mass of investors investing in MFs. Using 4, \( \kappa^{E^*} = \frac{\lambda'}{\lambda_1 - \lambda_0} \). The total MF and ETF shares outstanding at \( t = 0 \) respectively are

\[
X_0^M = \eta + 1 - \kappa^{E^*},
\]

\[
X_0^E = \kappa^{E^*}.
\]

Corollary 10 The threshold liquidity risk level, \( \lambda' \), which defines the self-selection of investors into ETFs versus MF markets at \( t = 0 \) is increasing in fund portfolio illiquidity. Fund tracking indices associated with larger \( c_j \), are subject to stronger payoff complementarities. This decreases the expected terminal fund payoff for patient fund investors making them less likely to invest in the first place.

Because \( \lambda' \) is decreasing in the liquidity of the fund’s underlying benchmark index, the most illiquid mutual funds attract the most liquidity sensitive investor base. Inherently more risky MFs become endogenously even more risky when investors have the choice to invest in an otherwise identical ETF instead because only investors with high idiosyncratic liquidity risk choose to remain invested in the MF.
4 Conclusion

This paper studies investors’ optimal portfolio allocation between ETFs and open-end mutual funds. I show that the non-portfolio differences between ETFs and mutual funds matter for fund returns at different horizons. ETFs are not universally "better" than mutual funds. Due to the guaranteed redemption of fund shares at the end-of-day fund NAV, MFs provide relatively more liquidity over the short-term. In particular in illiquid fund market segments, such as corporate bond funds, and during periods of market stress, MFs insure investors against the potentially significant price impact in illiquid market segments. However, this short-term liquidity insurance comes at a cost. The cost for MFs’ short-term liquidity provision is born by remaining long-term MF investors who bear the transaction costs caused by the early redemptions of other investors. On the contrary, ETFs provide relatively higher returns over the long-term. Unlike in MFs, there are no externalities among investors in ETFs. Each ETF investor bears their own transaction costs when trading ETF shares in secondary markets. This forces ETF investors to internalize their price impact in illiquid market segments and limits the potential for ETF share dilution. However, the pricing of ETF shares in secondary markets depends on the continuous arbitrage activities by financial intermediaries (APs). Over the short-term, when APs face balance sheet capacity constraints and temporarily retreat from ETF creations or redemptions, ETFs can be mispriced relative to their underlying portfolio NAV. In fact, relative mispricing in ETFs gives rise to "reverse run incentives", whereby increased selling pressure in ETF markets discourages the remaining ETF investors from liquidating their shares early, the opposite effect as in open-end MFs. Countercyclicality in intermediary balance sheet constraints can explain why ETFs tend to trade at a discount to their fund NAV when market volatility is high and aggregate liquidity is low. Over the long-term, once APs resume their ETF arbitrage activities, ETF prices tends to converge back to its fair value.

In this framework, MFs are naturally preferred by investors with relatively higher liquidity needs or shorter investment horizons. These investors are willing to sacrifice long-term expected returns in exchange for MF’s short-term liquidity insurance. ETFs in turn are preferred by investors with low liquidity needs and longer term horizons. This result challenges the popular view that ETFs may appeal more to investors with a higher liquidity demand. The equilibrium size of the ETF as compared to the MF sector is increasing in the illiquidity of a given security market segment. Abstracting from search and switching costs, as well as the demand from high frequency traders who require intra-day liquidity, ETFs are predicted to dominate the corporate bond index segment, whereas both funds are perfect substitutes in large cap equities.
References


Appendices

A Figures

Figure 2: Size of U.S. index fund market

(a) Net assets in U.S. index funds, bn USD
(b) Monthly net flows into U.S. index funds

Figure 3: Number of U.S. index ETFs and MFs (share classes)

B Proofs

Proof of Lemma 1. The only way for a non-zero tracking difference to occur in ETFs is for creation and redemption baskets to diverge from the underlying benchmark index. This is ruled out by assumption 1.

Proof of Lemma 2. By definition $\Delta_2^M \equiv P_2^j - NAV_2^M$. The terminal index price is given exogenously by $P_2^j = \nu^j(x) = x$. The fund NAV at the end of any given business day is defined by the accounting identity $NAV_2^M = \frac{X_2^M P_2^j}{\kappa^M + \eta - \nu^M}$, where $X_2^M$ is the number of index shares held by the fund, $P_2^j$ is the price per unit of index share and $\kappa^M + \eta - \nu^M$ is the
remaining number of MF shares outstanding. The number of index shares held by the MF at $t = 1$ and $t = 2$ are given by

$$X_1^M = (1 - \kappa + \eta)$$

(62)

$$X_2^M = (1 - \kappa + \eta) - \frac{\psi E[v^j(x)|s]\nu^M}{P_{2^-}^j}$$

(63)

Flow-induced index sales

It follows

$$\Delta_2^M = P_2^j - NAV_2^M$$

$$= P_2^j - \frac{P_2^j X_2^M}{\kappa^M + \eta - \nu^M}$$

$$= P_2^j - \frac{P_2^j (\kappa^M + \eta - \Theta_2^{j,MF})}{\kappa^M + \eta - \nu^M}$$

(64)

where $\Theta_2^{j,MF}$ is the number of index shares sold by the MF at $t = 2$ to meet net fund redemptions in the interim period. Using the expression for $\Theta_2^{j,MF}$ in 42,

$$\Delta_2^M = P_2^j - \frac{P_2^j (\kappa^M + \eta - \nu^M)}{\kappa^M + \eta - \nu^M} - \frac{P_2^j \nu^M}{\kappa^M + \eta - \nu^M}$$

(65)

Finally,

$$\Delta_2^M = \frac{P_2^j (\kappa^M + \eta - \nu^M)}{\kappa^M + \eta - \nu^M}$$

(66)

The MF tracking error represents the total cost of the share dilution experienced by remaining fund shareholders and can be decomposed into two components

$$\Delta_2^M = \frac{P_2^j (\kappa^M + \eta - \nu^M)}{\kappa^M + \eta - \nu^M} - \frac{P_2^j \nu^M}{\kappa^M + \eta - \nu^M}$$

(67)

Value of index shares sold to meet early redemptions

↓ # remaining shareholders

44
Proof of Corollary 1.

From equation 66 it directly follows that $\Delta_2^M = 0$ if $\nu^M = 0$ since

$$\Delta_2^M = \frac{P^j_2 \left( \frac{\nu^E [v^j(x)|s_1]}{P^j_2^M} - 0 \right)}{\kappa^M + \eta - 0} = \frac{P^j_2 \times 0}{\kappa^M + \eta - 0} = 0. \quad (68)$$

Similarly, if $\psi E[v^j(x)|s_1] = P^j_2^M$, $\Delta_2^M = 0$ because

$$\Delta_2^M = \frac{P^j_2 \left( \frac{\nu^E [v^j(x)|s_1]}{P^j_2^M} - \nu^M \right)}{\kappa^M + \eta - \nu^M} = \frac{P^j_2 \left( \nu^M - \nu^M \right)}{\kappa^M + \eta - \nu^M} = 0 \quad (69)$$

Finally, if $\Delta_2^M \neq 0$ and $\psi E[v^j(x)|s_1] \neq P^j_2^M$, then $\Delta_2^M \neq 0$. This follows directly from the fact that $\eta \leq \nu^M$. Due to short-selling constraints, the mass of investors redeeming shares, $\nu^M$, can never be larger than the total mass of investors who exist in the economy $\eta$. By assumption the mass of pension funds is non-zero $\kappa^M > 0$. Thus, $\kappa^M + \eta - \nu^M > 0$. By assumption $P^j_2 > 0$, the index price must be strictly positive. Hence $\Delta_2^M = 0$ requires

$$\left( \frac{\nu^M \psi E[v^j(x)|s_1]}{P^j_2^M} - \nu^M \right) = 0 \quad (70)$$

which is never satisfied for $\Delta_2^M \neq 0$ and $\psi E[v^j(x)|s_1] \neq P^j_2^M$.

Proof of Lemma 3. I start by proving the results (i) and (iii): Conditional on receiving no liquidity shock, $\Lambda^i = 0$, ETF (MF) investors are ex-post identical. This is because at $t = 0$ investors only had the choice between investing their entire endowment into MFs or ETFs. They did not have the option to invest parts of their endowment in the risk-free asset. All investors characterized by $\lambda^i < \lambda^i$ ($\lambda^i > \lambda^i$) at $t = 0$ and $\Lambda^i = 0$ at $t = 1$, that is patient ETF (MF) investors, have the initial allocation $\theta^i_0^E = \theta_0 = 1$ ($\theta^i_0^M = \theta_0 = 1$). This allocation implies that all patient ETF (MF) investors’ budget constraints at $t = 1$ must be identical. In addition, by assumption, all investors have identical preferences over terminal wealth, $\gamma \perp i$. Thus, all remaining patient ETF (MF) investors must have the same optimal investment policy at $t = 1$, that is

$$\theta^i_1^E^* = \theta^E^* \quad \forall \ i \text{ with } \theta^i_0^E = 1 \text{ and } \Lambda^i = 0 \quad (71)$$

$$\theta^i_1^M^* = \theta^M^* \quad \forall \ i \text{ with } \theta^i_0^M = 1 \text{ and } \Lambda^i = 0 \quad (72)$$

The proofs for (ii) and (iv) follow. Similarly, proposition ?? implies that all investors characterized by $\lambda^i < \lambda'$ ($\lambda^i > \lambda'$) at $t = 0$ and $\Lambda^i = 1$ at $t = 1$, that is impatient ETF (MF)
investors, have the initial allocation \( \theta_{0}^{i,E} = \theta_{0} = 1 \) (\( \theta_{0}^{i,M} = \theta_{0} = 1 \)). By assumption they must liquidate their entire portfolio upon receiving a liquidity shock at \( t = 1 \). Thus.

\[
\begin{align*}
\theta_{1}^{i,E} &= 0 \forall i \text{ with } \theta_{0}^{i,E} = 1 \text{ and } \Lambda^{i} = 1 \quad (73) \\
\theta_{1}^{i,M} &= 0 \forall i \text{ with } \theta_{0}^{i,M} = 1 \text{ and } \Lambda^{i} = 1 \quad (74)
\end{align*}
\]

Their terminal wealth is given by

\[
\begin{align*}
W_{2}^{i} &= P_{1}^{E} \forall i \text{ with } \theta_{0}^{i,E} = 1 \text{ and } \Lambda^{i} = 1 \quad (75) \\
W_{1}^{i} &= \text{NAV}_{1}^{M}^{\ast} \forall i \text{ with } \theta_{0}^{i,M} = 1 \text{ and } \Lambda^{i} = 1 \quad (76)
\end{align*}
\]

**Proof of Proposition 1.** Because all investors face non-zero liquidity risk, \( \lambda_{0} \leq \lambda_{i} \leq \lambda_{1} \) with \( 0 > 0 \), and liquidity shocks are iid across investors, the mass of impatient investors at \( t = 1 \) is non-zero. Given market makers downward sloping demand function for index shares 14, market clearing in equilibrium implies \( P_{1}^{j}R_{1}^{j} < E[P_{2}^{j}|s_{1}] \). At the same time, AP’s objective function 34 with \( \phi_{j} > 0 \) implies that \( P_{1}^{E} \leq P_{1}^{j} \) as long as net outflows from ETFs are weakly positive. Then, the result in proposition 1 follows directly from assumption ??.

**Proof of Proposition 2.** Market clearing in index markets between index market makers and MFs, \( \Theta_{2}^{M,j} = \Theta_{2}^{D,j,M} \), implies

\[
\begin{align*}
\frac{E[v^{j}(x)|s] - P_{2}^{j,M}}{c_{j}} &= \frac{\nu^{M}P_{1}^{M}}{P_{2}^{M}} \\
\iff \frac{s - P_{2}^{j,M}}{c_{j}} &= \frac{\nu^{M}\psi s}{P_{2}^{j,M}} \\
\iff 0 &= P_{2}^{j,M} - P_{2}^{j,M}s + c_{j}\psi s\nu^{M}.
\end{align*}
\]

Using assumption 4, it follows \( P_{2}^{j,M}^{\ast} = \frac{1}{2}(s + \sqrt{s^{2} - 4c_{j}\psi s\nu^{M}}) \).

**Proof of Corollary 3.** This result directly follows from 77 and proposition 2.

**Proof of Corollary 4.** These results directly follow from 77 and the fact that the inverse net index demand function of market makers, \( P_{2}^{j,D,M} \), and the inverse net index supply by MFs, \( P_{2}^{j,M} \), is strictly decreasing in the number of index shares

\[
\begin{align*}
\frac{dP_{2}^{j,D}}{d\Theta_{2}^{j,D}} &= -c_{j} \quad (78) \\
\frac{dP_{2}^{j,M}}{d\Theta_{2}^{j,M}} &= -\frac{\nu^{M}P_{1}^{M}}{(\Theta_{2}^{j,M})^{2}} \quad (79)
\end{align*}
\]

over the permissible range of \( \Theta_{2}^{j,M} \).

46
Proof of Proposition 3. \( x_j \) must be such that after observing the fundamental \( x_j < x_j \) patient MF investors always decide to run and redeem their entire portfolio of MF shares early no matter their beliefs of other patient MF shareholders’ actions. Patient MF investors will choose to liquidate their shares at \( t = 1 \) even if they believe only impatient MF shareholders redeem early. Conversely, for any \( x_j > x_j \), patient MF investors will never want to redeem any of their shares early. In this region runs never occur in equilibrium.

**Lower-dominance region** \((0, x_j)\). Using ?? and lemma 2, the lower dominance region with respect to the fundamental \( x_j \) is defined by the following condition:

1. Case with risk-neutral investors:

   \[
   0 = E[P_2^j - \Delta_2^M, j - P_1^M, j R^f | x_j = x_j \cup \nu^M = e^M] \tag{80}
   \]
   
   where \( e^M < \kappa^M \leq 1 \) is the mass of impatient (early) MF investors at \( t = 1 \).

2. Case with mean-variance preferences:

   \[
   0 = \frac{E[P_2^j - \Delta_2^M, j - P_1^M, j R^f | x_j = x_j \cup \nu^M = e^M]}{\gamma \text{Var}[\Delta_2^M, j | x_j = x_j \cup \nu^M = e^M]} \tag{81}
   \]
   
   Using the definition of the MF tracking difference \( 3 \) and the terminal value of the index, equation \( 80 \) [81] can be rewritten as

1. Case with risk-neutral investors:

   \[
   0 = x_j - E[\Delta_2^M, j | x_j = x_j \cup \nu^M = e^M] - \psi x_j R^f
   \]
   
   \[
   = x_j - E \left[ \left. \frac{\nu^M (NAV_1^M - P_{1+}^j)}{P_{1+}^j} \right| x_j = x_j \cup \nu^M = e^M \right] - \psi x_j R^f
   \]
   
   \[
   = x_j - x_j E \left[ \left. \frac{\nu^M (NAV_1^M - P_{1+}^j)}{P_{1+}^j} \right| x_j = x_j \cup \nu^M = e^M \right] - \psi x_j R^f. \tag{82}
   \]

Assuming \( x_j \neq 0 \),

\[
0 = 1 - E \left[ \left. \frac{\nu^M (NAV_1^M - P_{1+}^j)}{P_{1+}^j} \right| x_j = x_j \cup \nu^M = e^M \right] - \psi R^f. \tag{83}
\]

Using \( P_{1+}^j \) from index market clearing at \( t = 1^+ \) from proposition 2,
\[
0 = 1 - E \left[ \frac{\nu^M (\psi x_j - \frac{1}{2}(x_j + \sqrt{x_j^2 - 4c_j \psi x_j \nu^M}))}{\frac{1}{2}(x_j + \sqrt{x_j^2 - 4c_j \psi x_j \nu^M})} \bigg| x_j = x_j \cup \nu^M = e^M \right] - \psi R^f. \quad (84)
\]

Applying the conditional expectation operator,

\[
0 = 1 - E \left[ \frac{e^M (\psi x_j - \frac{1}{2}(x_j + \sqrt{x_j^2 - 4c_j \psi x_j \nu^M}))}{\frac{1}{2}(x_j + \sqrt{x_j^2 - 4c_j \psi x_j \nu^M})} \bigg| x_j = x_j \cup \nu^M = e^M \right] - \psi R^f. \quad (85)
\]

Simplifying

\[
0 = (1 - \psi R^f)(\kappa^M + \eta - e^M) - \left( \frac{e^M (\psi x_j - \frac{1}{2}(x_j + \sqrt{x_j^2 - 4c_j \psi x_j \nu^M}))}{\frac{1}{2}(x_j + \sqrt{x_j^2 - 4c_j \psi x_j \nu^M})} \right)
\]

\[
= (1 - \psi R^f)(\kappa^M + \eta - e^M) - \left( \frac{e^M (2\psi - (1 + \sqrt{1 - \frac{4c_j \psi \nu^M}{x_j}}))}{(1 + \sqrt{1 - \frac{4c_j \psi \nu^M}{x_j}})} \right)
\]

\[
= (1 - \psi R^f)(\kappa^M + \eta - e^M) \left( 1 + \sqrt{1 - \frac{4c_j \psi \nu^M}{x_j}} \right) - 2c_j \psi + e^M \left( 1 + \sqrt{1 - \frac{4c_j \psi \nu^M}{x_j}} \right)
\]

\[
= \left( (1 - \psi R^f)(\kappa^M + \eta - e^M) + e^M \right) \left( 1 + \sqrt{1 - \frac{4c_j \psi \nu^M}{x_j}} \right) - 2c_j \psi. \quad (86)
\]

Solving for \( x_j \),

\[
\frac{2c_j \psi}{(1 - \psi R^f)(\kappa^M + \eta - e^M) + e^M} = 1 + \sqrt{1 - \frac{4c_j \psi \nu^M}{x_j}}
\]

\[
\left( \frac{2c_j \psi}{(1 - \psi R^f)(\kappa^M + \eta - e^M) + e^M} - 1 \right)^2 = 1 - \frac{4c_j \psi \nu^M}{x_j}. \quad (87)
\]

This gives

\[
x_j = \frac{4c_j \psi e^M}{1 - \left( \frac{2c_j \psi}{(1 - \psi R^f)(\kappa^M + \eta - e^M) + e^M} - 1 \right)^2}, \quad (88)
\]

which can be simplified into
\[ x_j = \frac{4c_j \psi e^M}{1 - \left( \frac{e^M(2\psi+1) - (1-\psi \mu) \left( \kappa M + \eta - e^M \right)}{(1-\psi \mu)(\kappa M + \eta - e^M) + e^M} \right)^2} \]

Finally, this gives

\[ x_j = \frac{c_j \psi (\kappa M + \eta + e^M \psi \mu - \kappa M \psi \mu - \eta \psi \mu)^2}{(1 + \psi)(e^M(-1 + \psi(\mu - 1))) - (\kappa + \eta)(\psi(\mu - 1))} \].

With the simplifying assumption that \( \mu = 1 \),

\[ x_j = \frac{c_j \psi (\kappa M + \eta + e^M \psi - \kappa M \psi - \eta \psi)^2}{(1 + \psi)(-e^M - (\kappa + \eta)(\psi - 1))}. \]
2. Case with mean-variance preferences:

\[
0 = \bar{x}_j - \bar{x}_j E \left[ \frac{\nu^M(NAV^M_{i+1} - \rho^j_{i+1})}{\kappa^M + \eta - \nu^M} \right] \mid x_j = \bar{x}_j \cup \nu^M = e^M - \bar{x}_j \psi R^f
\]

(92)

Assuming \( Var[\Delta^M_M | x = \bar{x} \cup \nu^M = e^M] \neq 0 \) and using the fact that \( \gamma > 0 \), the above equation reduces to

\[
0 = \bar{x}_j - \bar{x}_j E \left[ \frac{\nu^M(NAV^M_{i+1} - \rho^j_{i+1})}{\kappa^M + \eta - \nu^M} \right] \mid x_j = \bar{x}_j \cup \nu^M = e^M - \bar{x}_j \psi R^f
\]

(93)

This condition is identical to equation 82 for the risk-neutral case. Hence, its solution, that is the boundary value of the lower dominance region, \( \bar{x}_j \), is independent of investors’ risk aversion. The cut-off value \( \bar{x}_j \) at which patient MF investors always run is the same for risk-neutral and risk-averse MF investors.

**Existence of the lower dominance region.** For the lower dominance region to be non-empty it must hold that \( \bar{x}_j > 0 \). That is

\[
0 < c_j \psi (\kappa^M + \eta + e^M \psi R^f - \kappa^M \psi R^f - \eta \psi R^f)^2 \frac{(1 + \psi)(e^M(-1 + \psi(R^f - 1)) - (\kappa^M + \eta)(\psi R^f - 1))}{(1 + \psi)(e^M(-1 + \psi(R^f - 1)) - (\kappa^M + \eta)(\psi R^f - 1))}
\]

(94)

From 90, it is clear that this condition is satisfied iff

\[
0 < c_j \psi \\
\text{and } (R^f - 1)\psi < 2
\]

(95)  (96)

\( 0 < c_j \psi \) ensures that the numerator of 90 is strictly positive. \( c_j \psi > 0 \) holds by definition of the model parameters. \( c_j > 0 \) implies that funds have non-zero price impact in index markets. This condition is generally satisfied except for potentially the most liquid large cap equity indices. \( 0 < \psi \) holds by assumption. If \( \psi \leq 0 \) MF investors would receive zero or negative payouts if they were to sell their MF shares at \( t = 1 \), a case for which the model does not allow.

The denominator of 90 is also strictly positive if the second condition, \( (R^f - 1)\psi < 2 \), is satisfied. The latter requirement follows from the condition

\[
(1 + \psi)(e^M(-1 + \psi(R^f - 1))) < -(\kappa^M + \eta)(\psi R^f - 1) - e^M + \psi e^M R^f - 2\psi e^M + \psi^2 e^M R^f - \psi^2 e^M < \kappa^M + \eta - \psi R^f \kappa^M - \psi R^f \eta.
\]

(97)
Note that $\kappa^M = e^M + l^M$. This condition simplifies into

$$(\psi + \psi^2)e^M R^f < \underbrace{(1 - \psi R^f)(\kappa^M + \eta)}_{> 0} + \underbrace{(1 + \psi)^2 e^M}_{> 0}$$

$$(1 - \psi R^f)(\kappa^M + \eta) + (1 - \psi R^f)e^M + ((1 - R^f)\psi^2 + 2\psi)e^M \quad (98)$$
as long as $(R^f - 1)\psi < 2$ which is true for a reasonable range for the risk-free rate and especially in the baseline case with $R^f = 1$.

Hence, given the model’s parameter restrictions, condition 94 is satisfied. The lower dominance region exists and is non-empty, $x > 0$.

**Effect of change in model parameters on size of lower dominance (run) region.**

The partial derivatives of $x_j$ with respect to the model parameters define the dependence of the size of the run region on index market segment specific liquidity and the MF’s investor base composition.

**i. Index market illiquidity, $c_j$.** First, the lower dominance region naturally increases in the illiquidity of the index market segment as reflected by $c_j$,

$$\frac{\partial x_j}{\partial c_j} = \frac{\psi(\kappa^M + \eta + e^M \psi R^f - \kappa^M \psi R^f - \eta \psi R^f)^2}{(1 + \psi)(e^M (-1 + \psi(R^f - 1)) - (\kappa^M + \eta)(\psi R^f - 1))} > 0. \quad (99)$$

The result that $\frac{\partial x_j}{\partial c_j} > 0$ directly follows from the proof of the existence of the lower dominance region under the conditions 95 and 96.

**ii. Mass of sleepy MF investors, $\eta$.** Second, the lower dominance region is [xxx] in the mass of sleepy MF investors,

$$\frac{\partial x_j}{\partial \eta} = -\frac{c_j \psi (1 - \psi R^f)(\kappa^M + \eta + e^M \psi R^f - \kappa^M \psi R^f - \eta \psi R^f)^2}{(1 + \psi)(e^M (-1 + \psi(-1 + R^f)) - (\kappa^M + \eta)(-1 + \psi R^f))^2} + \frac{2c_j \psi (1 - \psi R^f)(\kappa^M + \eta + e^M \psi R^f - \kappa^M \psi R^f - \eta \psi R^f)}{(1 + \psi)(e^M (-1 + \psi(-1 + R^f)) - (\kappa^M + \eta)(-1 + \psi R^f))} \quad (100)$$
Upper-dominance region \([2, \infty)\). Using 31 and proposition 2, the upper dominance region with respect to the fundamental \(x\) is defined by the following condition:

1. Case with risk-neutral investors:

\[
E[P_2^j - \Delta_2^{M,j} - P_1^{M,j} R^f | x_j = \pi_j \cup \nu^M = \kappa^M] = 0
\]  

Equation 101 is similar to the condition for the lower dominance region, \(\underline{z}_j\), in equation 80. The distinguishing feature between both regions is the conditioning information in the expectations operator.

From 101 follows

\[
0 = 1 - E \left[ \frac{\nu^M(\psi x_j - \frac{1}{2}(x_j + \sqrt{x_j^2 - 4c_j \psi x_j \nu^M}))}{\kappa^M + \eta - \nu^M} \right]_{x_j = \pi_j \cup \nu^M = \kappa^M} - \psi R^f. 
\]  

(102)

Applying the expectations operator and solving for \(\pi_j\)

\[
\pi_j = \frac{4c_j \psi \kappa^M}{1 - \left( \frac{\kappa^M(2\psi+1)-(1-\psi R^f)\eta}{(1-\psi R^f)\eta+\kappa^M} \right)^2}
\]  

(103)

Finally, this gives

\[
\pi_j = \frac{c_j \psi(\kappa^M + (1 - \psi R^f)\eta)^2}{(1+\psi)(\eta(1 - \psi R^f) - \kappa^M \psi)}.
\]  

(104)

**Existence of upper dominance region.** For \(\pi_j > 0\), the numerator and denominator of 104 must both be strictly positive. Beyond this necessary condition, the existence of the upper dominance region further requires \(\underline{z}_j < \pi_j\).

By definition, \(c_j \psi > 0\). Hence, the numerator of 104 is always positive. For the denominator of 104 to be strictly positive, it must hold that

\[
\eta(1 - \psi R^f) - \kappa^M \psi > 0.
\]  

(105)

Or \(\eta - \psi(R^f + \kappa^M) > 0\). By definition \(\kappa^M\) is bounded above by 1, the total mass of investors in the economy, \(\kappa^M \leq 1\). \(F^f \geq 1\). Therefore, \(\underline{z}_j > 0\) iff the mass of sleepy MF investors is large relative to the mass of MF investors that may potentially redeem their shares early.

Assuming 105 is satisfied, \(\pi_j > \underline{z}_j\) requires

52
\[ c_j \psi(\kappa^M + \eta + \varepsilon^M \psi R^f - \kappa^M \psi R^f - \eta \psi R^f)^2 \left( \frac{1}{(1 + \psi)(\varepsilon^M(1 + \psi(R^f - 1)) - (\kappa + \eta)(\psi R^f - 1))} \right) < c_j \psi(\kappa^M + (1 - \psi R^f)\eta)^2 \left( \frac{1}{(1 + \psi)(\eta(1 - \psi R^f) - \kappa^M \psi))} \right) \]  

(106)

This simplifies into

\[ \frac{(\kappa^M + (1 - \psi R^f)\eta + \psi R^f(\varepsilon^M - \kappa^M))^2}{(\varepsilon^M(1 + \psi(R^f - 1)) - (\kappa + \eta)(\psi R^f - 1))} < \frac{(\kappa^M + (1 - \psi R^f)\eta)^2}{(\eta(1 - \psi R^f) - \kappa^M \psi))} \]  

(107)

2. Case with mean-variance preferences:

\[
\frac{E[P_j^2 - \Delta_2^M - NAV_1 M R^f | x_j = \bar{z}_j \cup \nu^M = \kappa^M]}{\gamma Var[\Delta_2^M | x_j = \bar{z}_j \cup \nu^M = \kappa^M]} = 1
\]

(108)

To solve for the value of \( x_j \) that marks the upper dominance region, it is necessary to first solve for \( Var[\Delta_2^M | x = \bar{x}] \), the variance of the terminal MF tracking difference conditional on the realization of \( x = \bar{x} \). After observing the fundamental, from an individual investors’ perspective, the only residual uncertainty arises from the uncertainty regarding the overall mass of MF investors who choose to redeem their shares early. In equilibrium, patient investors’ allocation decision at \( t = 1 \) will be deterministic conditional on \( x \). All patient MF investors are ex-post identical at \( t = 1 \) and will thus follow the same equilibrium investment policy.\(^{27}\)

However, while the total mass of MF investors at \( t = 1, \kappa^M \), is known, the precise mass of impatient investors who have to liquidate their shares no matter what the fundamentals is unknown and uncorrelated with \( x \). Hence, even with perfect foresight with respect to patient investors allocations, the total mass of MF redemptions at \( t = 1 \), that is \( \nu^M \), remains uncertain from an individual investor \( i \)’s perspective. \( \nu^M \) is the only remaining uncertain variable at \( t = 1 \). It drives the residual variation in \( \Delta_2^M \). In the following, I this variance by \( \sigma^2_\Delta | \bar{x} = Var[\Delta_2^M | x = \bar{x}] \).

Formally,

\[
\sigma^2_\Delta | \bar{x} \equiv Var[\Delta_2^M | x = \bar{x}]
\]

(109)

\[
= Var\left[ -\frac{\nu^M (NAV_1^M - P_j^1)}{P_j^1} | x = \bar{x} \right]
\]

(110)

\[
= Var\left[ -\frac{\nu^M \left( \psi \left( x + \sqrt{x^2 - 4c_j \psi x} \nu^M \right) \right)}{\kappa^M + \eta - \nu^M | x = \bar{x} \right]
\]

(111)

Using the definition of the MF tracking difference \(^3\) and the terminal value function of the index, equation 108 can be rewritten as

---

\(^{27}\) In solving the model, I do not allow for mixed strategy equilibria.

53
\[ E \left[ P_j^2 - \frac{P_j^2 \left( \frac{\nu M (NAV^M - P_{j+1}^M)}{\kappa^M + \eta - \nu^M} \right)}{\kappa^M + \eta - \nu^M} - NAV^1 M R^f \mid x = \bar{x} \right] = 1 \]  

(112)

Assuming \( \text{Var} \Delta^M_{\bar{x}} \neq 0 \) and using \( \gamma > 0 \),

\[ E \left[ P_j^2 - \frac{P_j^2 \left( \frac{\nu M (NAV^M - P_{j+1}^M)}{\kappa^M + \eta - \nu^M} \right)}{\kappa^M + \eta - \nu^M} - NAV^1 M R^f \mid x = \bar{x} \right] = \gamma \sigma^2 \Delta_{\bar{x}}. \]  

(113)

Using proposition 2 and the definition of the MF NAV, \( NAV^1 M = \psi x \),

\[ E \left[ P_j^2 - \frac{P_j^2 \left( \frac{\nu M (\psi x - \frac{1}{2}(x + \sqrt{x^2 - 4c_j \psi \bar{x}^M}))}{\kappa^M + \eta - \nu^M} \right)}{\kappa^M + \eta - \nu^M} - \psi x R^f \mid x = \bar{x} \right] = \gamma \sigma^2 \Delta_{\bar{x}}. \]  

(114)

Applying the expectations operator and using \( \nu^M = \kappa^M \),

\[ \bar{x} - \bar{x} \frac{\kappa^M \left( \psi \bar{x} - \frac{1}{2}(x + \sqrt{x^2 - 4c_j \psi \bar{x} \kappa^M}) \right)}{\frac{1}{2}(x + \sqrt{x^2 - 4c_j \psi \bar{x} \kappa^M}) - \psi \bar{x} R^f} \frac{\kappa^M (2 \psi - 1) - \sqrt{1 - \frac{4c_j \psi \kappa^M}{\bar{x}}} \eta}{\eta (1 + \sqrt{1 - \frac{4c_j \psi \kappa^M}{\bar{x}}})} = 0. \]  

(115)

(116)

Simplifying and using \( \bar{x} \neq 0 \),

\[ 1 - \psi R^f - \frac{\gamma \sigma^2 \Delta_{\bar{x}}}{\bar{x}} \frac{\kappa^M (2 \psi - 1) - \sqrt{1 - \frac{4c_j \psi \kappa^M}{\bar{x}}} \eta}{\eta (1 + \sqrt{1 - \frac{4c_j \psi \kappa^M}{\bar{x}}})} = 0. \]  

(117)

There does not exist a closed form solution for \( \bar{x} \). However, it can be shown that the solution to equation 118 is unique.

**Uniqueness of \( \bar{x} \).** To proof the uniqueness of the solution for \( \bar{x} \), first note that we can rewrite 118 as follows:

\[ \frac{(\eta (1 - \psi R^f) + \kappa^M)(1 + \sqrt{1 - \frac{4c_j \psi \kappa^M}{\bar{x}}})}{\eta (1 + \sqrt{1 - \frac{4c_j \psi \kappa^M}{\bar{x}}})} - \frac{\gamma \sigma^2 \Delta_{\bar{x}}}{\bar{x}} (1 + \sqrt{1 - \frac{4c_j \psi \kappa^M}{\bar{x}}}) - 2 \kappa^M \psi = 0. \]  

(118)

(119)
Proof of Corollary 7. \( \bar{s} \) and \( \bar{s} \) are given by

\[
\begin{align*}
\bar{s} &= 0 \\
\bar{s} &= \frac{\gamma \sigma^2 + \eta - E[\nu^M|\bar{s}]}{(1 - \psi)(\kappa^M + \eta)}
\end{align*}
\]

with \( \bar{s} > 0 \). \( \bar{s} \) uses the assumption that \( R^f = 1 \) and \( \kappa^M = 1 - \kappa \). \( \bar{s} > 0 \) follows from the fact that the number of MF shareholders redeeming fund shares early at \( t = 1 \) cannot exceed the total number of MF shareholders, \( \nu^M \leq 1 - \kappa \) and \( \eta > 0 \).

Proof of Corollary 5. The impact of a change in the fraction of investors who initially invested in MFs as compared to ETFs, \( \kappa^M = 1 - \kappa \), on the size of the run region in MFs, \( \bar{s} \), follows from the partial derivative \( \frac{\partial \bar{s}}{\partial \kappa^M} \).

\[
\begin{align*}
\frac{\partial \bar{s}}{\partial \kappa^M} &= \left(2c_j(\kappa^M + \eta(1 - \psi R^f)) \right) \cdot \left( (1 - \psi)\kappa^M + \eta(1 - \psi R^f) \right)^{-2} \\
&= \frac{c_j(\kappa^M + \eta(1 - \psi R^f))}{((1 - \psi)\kappa^M + \eta(1 - \psi R^f))^2} \\
&= \left(2((1 - \psi)\kappa^M + \eta(1 - \psi R^f)) - (1 - \psi)(\kappa^M + \eta(1 - \psi R^f)) \right)
\end{align*}
\]

Since \( \frac{c_j(\kappa^M + \eta(1 - \psi R^f))}{((1 - \psi)\kappa^M + \eta(1 - \psi R^f))^2} > 0 \), the effect of a change in \( \kappa^M \) on \( \bar{s} \) depends solely on the second part of equation \( \bar{s} \).

\[
\begin{align*}
2((1 - \psi)\kappa^M + \eta(1 - \psi R^f)) - (1 - \psi)(\kappa^M + \eta(1 - \psi R^f)) \\
= 2(1 - \psi)\kappa^M + 2\eta(1 - \psi R^f) - (1 - \psi)\kappa^M - \eta(1 - \psi)(1 - \psi R^f) \\
= (1 - \psi)\kappa^M + \eta(1 + \psi)(1 - \psi R^f) > 0
\end{align*}
\]

, where the last results follows from the assumption that \( \psi < 1 \) and \( R^f = 1 \). Hence,

\[
\frac{\partial \bar{s}}{\partial \kappa^M} > 0
\]

The impact of a change in the fraction of sleepy MF investors, \( \eta \), is given by
\[
\frac{ds}{d\eta} = \left(2c_j(1 - \psi Rf)(\kappa^M + \eta(1 - \psi Rf)) \ast ((1 - \psi)\kappa^M + \eta(1 - \psi Rf)) - (1 - \psi Rf) \ast c_j(\kappa^M + \eta(1 - \psi Rf))^2 \right) \\
\ast \left((1 - \psi)\kappa^M + \eta(1 - \psi Rf)\right)^{-2}
\]
\[
= \frac{c_j(\kappa^M + \eta(1 - \psi Rf))}{((1 - \psi)\kappa^M + \eta(1 - \psi Rf))^2} \\
\ast \left(2(1 - \psi Rf)((1 - \psi)\kappa^M + \eta(1 - \psi Rf)) - (1 - \psi Rf) \ast (\kappa^M + \eta(1 - \psi Rf))\right)
\]
\[
= \frac{c_j(\kappa^M + \eta(1 - \psi Rf))}{((1 - \psi)\kappa^M + \eta(1 - \psi Rf))^2} \ast \left(\eta(1 - \psi Rf)^2 > 0 \ast \frac{\kappa^M(1 - 2\psi)(1 - \psi Rf)}{< 0 \text{ if } \psi > \frac{1}{2}}\right)
\]

(125)

Hence the effect of a change in \( \eta \) on MF investors’ run incentives depends on the model parameters. There are three different cases.

First, when \( \psi \leq \frac{1}{2} \), that is when the penalty for early liquidation is large, the run region \((0, \bar{s}]\) is strictly increasing in the number of sleepy investors

\[
\frac{ds}{d\eta} > 0 \forall \eta > 0, \kappa^M > 0
\]

(126)

Instead, in the more realistic case in which \( \frac{1}{2} < \psi < 1 \), the effect of an increase in \( \eta \) on \( \bar{s} \) depends on the relative size of \( \eta \) and \( \kappa \).

Second, when the number of MF investors who have a choice to withdraw early, \( \kappa^M \), is large relative to the number of sleepy MF investors who never withdraw early, \( \frac{ds}{d\eta} < 0 \), and \( \frac{1}{2} < \psi < 1 \), an increase in \( \eta \) decreases the run region, \( \bar{s} \). Intuitively, when there are few sleepy investors, an increase in the number of sleepy investors decreases the remaining MF investors incentive for early redemption as it increases the pool of index shares that are shares across all remaining investors at \( t = 2 \).

However, at a certain level of \( \eta \), the marginal benefit of sharing the cost of early redemptions with a greater investor based is exactly offset by the marginal benefit of sharing the remaining fund assets with a greater investor base. This occurs when \( \eta \) is given by

\[
0 = \eta(1 - \psi Rf)^2 + \kappa^M(1 - 2\psi)(1 - \psi Rf)
\]
\[
\eta(1 - \psi Rf)^2 = -\kappa^M(1 - 2\psi)(1 - \psi Rf)
\]
\[
\eta = \frac{(2\psi - 1)}{(1 - \psi Rf)}\kappa^M
\]

(127)
, where \( \frac{(2\psi-1)}{(1-\psi R_f)} > 0 \).

Third, when the number of MF investors who have a choice to withdraw early, \( \kappa^M \), is small relative to the number of sleepy MF investors who never withdraw early, \( \frac{ds}{d\eta} > 0 \), and \( \frac{1}{2} < \psi < 1 \), an increase in \( \eta \) increases the run region, \( s \). Intuitively, when \( \eta \) is already large, a further increase implies that the fund assets are shares among a greater investor base at \( t = 2 \), thereby increasing patient MF investors’ incentive to run and redeem shares early.
## C List of model variables

### Table 1: Model notation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>State variable / terminal index payoff</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Risk aversion coefficient</td>
</tr>
<tr>
<td>$\lambda_i$</td>
<td>$i$'s probability of a liquidity shock at $t = 1$</td>
</tr>
<tr>
<td>$\phi_j$</td>
<td>AP balance sheet capacity constraint parameter</td>
</tr>
<tr>
<td>$c_j$</td>
<td>Index market maker inventory cost</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Mass of sleepy investors (e.g., retirement funds)</td>
</tr>
<tr>
<td>$R_f$</td>
<td>Risk-free rate</td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>Initial endowment in terms of index shares</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\kappa_i^E$</td>
<td>Share (mass) of ETF investors at $t = 0$</td>
</tr>
<tr>
<td>$\kappa_i^M$</td>
<td>Share (mass) of MF investors at $t = 0$</td>
</tr>
<tr>
<td>$e^E_i$</td>
<td>Mass of impatient ETF investors at $t = 1$</td>
</tr>
<tr>
<td>$e^M_i$</td>
<td>Mass of impatient MF investors at $t = 1$</td>
</tr>
<tr>
<td>$P^i_t$</td>
<td>Index market price at $t$</td>
</tr>
<tr>
<td>$NAV_t^M$</td>
<td>Mutual fund net asset value, $NAV_t^M = P^M_t$</td>
</tr>
<tr>
<td>$NAV_t^E$</td>
<td>ETF net asset value</td>
</tr>
<tr>
<td>$P^E_t$</td>
<td>ETF market price</td>
</tr>
<tr>
<td>$\zeta^E_t$</td>
<td>Relative ETF mispricing (discount), $\zeta^E_t = P^i_t - P^E_t$</td>
</tr>
<tr>
<td>$\theta_{i,M}^t$</td>
<td>Units of MF shares held by agent $i$ at $t$</td>
</tr>
<tr>
<td>$\theta_{i,E}^t$</td>
<td>Units of ETF shares held by agent $i$ at $t$</td>
</tr>
<tr>
<td>$\nu^M_t$</td>
<td>Total number of MF shares redeemed at $t = 1$</td>
</tr>
<tr>
<td>$\nu^E_t$</td>
<td>Total number of ETF shares redeemed at $t = 1$</td>
</tr>
<tr>
<td>$X_t^M$</td>
<td>Total number of MF shares outstanding at $t$</td>
</tr>
<tr>
<td>$X_t^E$</td>
<td>Total number of ETF shares outstanding at $t$</td>
</tr>
</tbody>
</table>