Asset Management Contracts and Equilibrium Prices

Andrea M. Buffa

University of Colorado Boulder

Dimitri Vayanos

London School of Economics, Center for Economic and Policy Research, and National Bureau of Economic Research

Paul Woolley

London School of Economics

We model asset management as a continuum between active and passive: managers can deviate from benchmark indices to exploit noise trader–induced distortions, but agency frictions constrain these deviations. Because constraints force managers to buy assets that they underweight when these assets appreciate, overvalued assets have high volatility, and the risk-return relationship becomes inverted. Distortions are more severe for overvalued assets than for undervalued ones because trading against the former entails more risk and tighter constraints. We provide empirical evidence supporting our model's main mechanisms. Using the data, we infer the constraints' tightness and compute a measure of effective arbitrage capital.

We thank Ricardo Alonso, Daniel Andrei, Oliver Bought, Adrian Buss, Jennifer Carpenter, Sergey Chernenko, Chris Darnell, Peter DeMarzo, Philip Edwards, Ken French, Willie Fuchs, Diego Garcia, Jeremy Grantham, Zhiguo He, Apoorva Javadekar, Ron Kaniel, Ralph Koijen, Dong Lou, John Moore, Dmitry Orlov, Marco Pagano, Anna Pavlova, Gabor Pinter, Christopher Polk, Andy Skrzypacz, Harald Uhlig, and four anonymous referees; seminar

Electronically published September 28, 2022

Journal of Political Economy, volume 130, number 12, December 2022.
© 2022 The University of Chicago. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0), which permits non-commercial reuse of the work with attribution. For commercial use, contact journalpermissions@press.uchicago.edu. Published by The University of Chicago Press.
https://doi.org/10.1086/720515

I. Introduction

Financial markets have become highly institutionalized. For example, individual investors were holding directly 47.9% of US stocks in 1980 but only 21.5% in 2007, with most of the remainder held by financial institutions, such as mutual funds and pension funds (French 2008). The portfolios of these institutions are chosen by professional asset managers.

The institutionalization of financial markets has stimulated research on the performance of professional managers and their effects on equilibrium asset prices and market efficiency. A vast literature examines whether actively managed funds outperform passively managed ones. A related literature investigates whether the growth of passive funds has made markets less efficient and whether efficiency increases in the ratio of active to passive.[1]

Drawing a sharp distinction between passive funds constrained to hold specific portfolios and active funds investing without constraints can be misleading. This is because much of active management is done around benchmark indices, with managers being constrained in how much they can deviate from them. A common constraint is a bound on tracking error (TE), defined as the standard deviation of the difference between a fund's return and the return of its benchmark index. Bounds can also be imposed on the deviation between a fund's portfolio weight in each asset class, geographical area, or industry sector and the corresponding index weight.[2] Viewing asset management as a *continuum* between active and

participants at Bank of England, Banque de France, Bocconi University, Boston University, Central European University, Cheung Kong Graduate School of Business, Collegio Carlo Alberto, Dartmouth College, Duke University, Einaudi Institute for Economics and Finance, Federal Reserve Bank of New York, Indiana University, Institut Européen d'Administration des Affaires, London School of Economics, Massachusetts Institute of Technology, Purdue University, Rice University, Southern Methodist University, Stanford University, University of Arizona, University of British Columbia, Université Catholique de Louvain, University College London, University of Glasgow, University of Maryland, University of Michigan, University of Minnesota, University of Naples, University of Oregon, University of Texas at Austin, University of Toulouse, Yale University, University of Vienna, and University of Virginia; and conference participants at Adam Smith Asset Pricing, American Economic Association, Bank of England Macro-Finance, Bank for International Settlements, Center for Financial Frictions, Center for Monetary and Financial Studies, Conference on Research on Economic Theory and Econometrics, European Summer Symposium in Financial Markets Gerzenzee, Financial Conduct Authority, Financial Intermediation Research Society, Finance Theory Group, Institute for Quantitative Investment Research, Jackson Hole, National Bureau of Economic Research Asset Pricing, Society for Financial Studies Cavalcade, Utah, and Western Finance Association for helpful comments. Taisiya Sikorskaya provided outstanding research assistance. Replication files are available in a zip file. This paper was edited by Harald Uhlig.

[1] See, e.g., Elton and Gruber (2013) for a survey of the literature on mutual fund performance and Franzoni, Ben-David, and Moussawi (2017) for a survey of exchange-traded funds and their effects on market performance.

[2] For example, the Norwegian Sovereign Wealth Fund, one of the largest institutional investors globally, reports the following regarding its TE constraint: "The Ministry of Finance has set limits for how far the fund may deviate from the benchmark index. The most

passive, depending on the tightness of managers' constraints, seems a better description of reality. In this paper, we flesh out that alternative view, provide empirical evidence for it, and explore theoretically its implications for equilibrium asset prices and market efficiency. We show that these implications differ significantly from the conventional view.

A simple example helps motivate our analysis. Suppose that some asset managers must keep their portfolio weight in each industry sector within 5% of the sector's weight in a benchmark index. Suppose also that a sector that the managers view as overvalued has 10% weight in the index, while the managers give it 5% weight. If the sector appreciates and reaches 20% weight in the index, then its weight in the managers' portfolio reaches (approximately) 10% but must rise further to 15% so that the constraint is met. Buying pressure by the managers amplifies the sector's appreciation, raising its volatility. Overvalued sectors thus have high volatility in addition to their low expected return, causing the risk-return relationship to become weak or inverted, consistent with empirical evidence.[3] Amplification does not arise when managers are constrained to hold the index or when they are unconstrained. The example implies additionally that overvaluation is harder to correct than undervaluation. Indeed, managers must stick closer to the index in overvalued sectors: a 5% difference in weight allows less leeway in relative terms when the sector's benchmark weight is large.

Section II presents evidence on the portfolio constraints of asset managers and the behavior they induce. Active funds investing in US equities exhibit large differences in their TE: the average TE for funds in quintile 5 of TE is about four times as large as for funds in quintile 1. Moreover, these differences persist over time and can thus be viewed as a fund characteristic: a fund in quintile 1 of TE lies on average in quintile 1.63 after 3 years. Persistence is even stronger for active share (AS), computed by summing across assets the deviations between an asset's portfolio weight in a fund and in the fund's benchmark index. These findings extend Cremers and Petajisto (2009).

While the persistence of TE and AS could, in principle, be due to inertia, we present new evidence that it reflects constraints and that the constraints' effects are consistent with our model's main mechanisms. Funds buy stocks that they underweight relative to their benchmark indices and do so *procyclically*: they buy to a larger extent after the stocks perform well.

---

important limit is expressed using the statistical concept of expected relative volatility, or TE. The limit for expected relative volatility has been set at 125 basis points. This means that the difference between the return on the fund and the return on the benchmark portfolio is expected to be more than 1.25 percentage points in only one out of every 3 years" (https://www.nbim.no/en/the-fund/how-we-invest/risk-management/).

[3] References to the empirical literature on risk-return inversion are in sec. IV.D.

Conversely, funds sell stocks that they overweight and do so slightly *countercyclically*: they sell to a larger extent after the stocks perform well. The procyclical (momentum) buying of underweighted stocks is stronger for funds with low TE or AS. Funds in quintile 1 of TE or AS eliminate 40% of their underweight in overperforming stocks in the two semesters during and following the overperformance. By comparison, they eliminate 20% of their underweight in underperforming stocks.

Section III presents the model. Investors can invest in a riskless asset and in multiple risky assets over an infinite horizon. The riskless rate is constant, and each risky asset's dividend flow per share follows a square root process. Investors maximize a mean-variance objective over instantaneous changes in wealth. Some investors are unconstrained, while others face a constraint limiting the deviation between the portfolio weight they give to each asset and the asset's weight in a benchmark index. Investors deviate from the index to exploit price distortions created by noise traders.

Section IV derives the equilibrium prices of the risky assets, taking the constraint as exogenous and not distinguishing between investors and the asset managers they employ. We analyze two polar cases first: no constraint, in which case all investors are fully active, and an infinitely tight constraint, in which case constrained investors hold the index and are fully passive. In both cases, we derive a closed-form solution for each asset's price and show that it is affine in the asset's dividend flow. An increase in noise trader demand raises price and lowers expected return. It does not affect, however, return volatility: the price becomes more sensitive to the dividend flow, but the effect is proportional to the increase in the price level. Moving from no constraint (all investors fully active) to an infinitely tight constraint (constrained investors fully passive) exacerbates the price distortions created by noise traders. This is because the constraint prevents constrained investors from absorbing noise trader demand. The constraint does not affect return volatility, however, because volatility is independent of demand.

We next analyze the general case. For each asset, the equilibrium involves a region where the constraint on that asset's portfolio weight does not bind and a region where it binds. The constraint binds for high values of the asset's dividend flow because the asset's portfolio weight is high.

The equilibrium price of each risky asset is convex in the asset's dividend flow if the asset is in high noise trader demand and concave if it is in low demand. The convexity reflects the amplification effect. The concavity reflects the opposite dampening effect: since constrained investors give higher weight to an undervalued asset relative to the asset's index weight, they must sell the asset when it appreciates, dampening the appreciation. These effects map to the evidence in section II on the procyclical buying of underweighted stocks and countercyclical selling of overweighted stocks. They generate a negative cross-sectional relationship between

volatility and expected return. That relationship is most pronounced for intermediate levels of the constraint (and is absent in the polar cases of no or infinitely tight constraint, where volatility is independent of noise trader demand).

Consistent with empirical evidence, the inverted risk-return relationship in our model is driven primarily by the overvalued assets, and distortions for these assets are larger. Our model is also consistent with evidence that return momentum is more pronounced within overvalued assets. Overvaluation is often attributed to a combination of heterogeneous beliefs and short selling costs (e.g., Harrison and Kreps 1978; Scheinkman and Xiong 2003; Hong and Stein 2007). Our results suggest that short selling costs are not necessary for overvaluation distortions to be more severe than undervaluation ones. Indeed, short selling costs are not present in our model, and all investors hold long positions in our calibrated example. Overvaluation is harder to correct than undervaluation because overvalued assets make up a larger fraction of the market, so trading against them entails more risk and tighter constraints.

Because overvaluation is more severe than undervaluation, market segments with more heterogeneous noise trader demand across their component assets earn lower expected returns than segments with less heterogeneity and same average demand. An analogous result is shown in the literature on heterogeneous beliefs and short selling costs, but our result assumes no short selling costs. Our model implies additionally that the relationship between heterogeneity and overvaluation is stronger when managers are more constrained to remain close to their benchmark indices.

Section V endogenizes the parameters of the constraint within a simple contracting model in which investors employ asset managers. Managers can be skilled and observe noise trader demand and the dividend flow or unskilled and trade on uninformative signals. Investors optimize over the wealth they allocate to their managers, a performance-based fee they pay the managers, and an investment restriction that limits how much the managers' portfolio weight in each asset can deviate from the index weight. The optimal fee aligns managers' risk preferences with those of their investors. Investors must guard, however, against the possibility that their managers are unskilled and do so through the investment restriction. The restriction that investors impose becomes tighter when the fraction of unskilled managers increases.

In a calibrated example, we infer the constraints' tightness from the data. Interpreting assets as industry sector portfolios, we find that observed differences in AS across funds are consistent with a bound on deviations from sector index weights by managers of constrained funds of around 5%. Investors find it optimal to impose such a bound when they believe that the fraction of unskilled managers ranges between 20% and 40%.

One interpretation of the high inferred fraction of unskilled managers is that 5% reflects not only an explicit bound that investors impose on managers but also an implicit bound that managers impose on themselves to limit their reputational risk from underperforming the index. Our calibrated example also allows us to compute *effective capital*, defined as the capital that—if managed without constraints—would reduce price distortions to the same extent as a given capital managed with empirically plausible constraints.

Our paper relates to several strands of work on asset management and asset pricing. One literature concerns the performance of active versus passive funds, and their impact on market efficiency. That literature builds on the seminal paper by Grossman and Stiglitz (1980), in which informed and uninformed investors trade with noise traders, there is a cost to becoming informed, and price informativeness increases in the fraction of the informed. In Subrahmanyam (1991), the introduction of index futures induces noise traders to trade the index rather than the component assets. This lowers liquidity for the component assets and has ambiguous effects on market efficiency. Related mechanisms are at play in Cong and Xu (2016) and Bhattacharya and O'Hara (2018), who study how exchange-traded funds affect market efficiency and liquidity, and Bond and Garcia (2021), who study the effects of lowering the costs of passive investing. Pastor and Stambaugh (2012) and Stambaugh (2014) explain the decline in active funds' expected returns based on the increase in the assets they manage and the decline in noise trading, respectively.[4] In Garleanu and Pedersen (2018), active funds' expected returns decline when investors are better able to locate skilled managers. In these papers, active funds invest without investor-imposed constraints, while constraints are central to our analysis.

In emphasizing constraints, our paper is related to the literature on leverage constraints and fire sales (for surveys, see Gromb and Vayanos 2010; Shleifer and Vishny 2011). In that literature, constraints tighten when asset prices fall, generating procyclicality. Moreover, distortions are largest in down markets. In our model, by contrast, constraints tighten when asset prices rise, and this generates counterclicality for undervalued assets—which managers overweight—and procyclicality for overvalued assets. Moreover, distortions are largest for overvalued assets and in up markets.

Another related literature studies asset management contracts. Within its strand that takes asset prices as given, our paper relates most closely to He and Xiong (2013) and Parlour and Rajan (2019), in which investors constrain managers' choice of assets to better incentivize them to acquire

[4] Berk and Green (2004) show that decreasing returns to scale at the level of individual funds help explain why investors flow into funds with good past performance even though performance does not persist.

information or to guard against other forms of moral hazard. Investors in our model constrain managers to guard against the possibility that they are unskilled.

Within the strand of the asset management contracts literature that endogenizes prices, our paper relates most closely to papers that examine the effects of compensating managers on the basis of their performance relative to a benchmark index. A common theme in several papers is that such compensation raises the price of the benchmark index and of assets covarying highly with it. Brennan (1993), Basak and Pavlova (2013), and Buffa and Hodor (2018) show this result in settings where managers derive direct utility from relative performance. Kapur and Timmermann (2005) and Cuoco and Kaniel (2011) show a similar result in settings where managers receive a linear fee. The latter paper also finds that the result can reverse when the fee has option-like components. Kashyap et al. (2021a) explore the result's implications for real investment.[5] Tighter constraints in our model can instead lower the price of the benchmark index because investors respond to overvaluation by cutting down their investment with asset managers.

An alternative explanation for risk-return inversion is based on leverage constraints (Black 1972; Frazzini and Pedersen 2014): investors prefer assets with high capital asset pricing model (CAPM) beta because they provide leverage, which investors cannot replicate by investing in low-beta assets and borrowing. Leverage constraints generate a negative relationship between CAPM beta and alpha but a positive one between beta and expected return. In our model, both relationships can be negative.

II. Evidence

A basic premise of our theory is that investment funds must maintain their deviations from benchmark indices within bounds. Our theory treats the bound for each fund as a characteristic of the fund, similar to other basic characteristics, such as the fees the fund charges and the asset classes the fund invests in. Our theory implies additionally that some of the trades that fund managers make are triggered by the requirement to maintain deviations from indices within bounds. These trades are procyclical for assets that funds underweight relative to the indices and countercyclical for assets that funds overweight. In this section, we provide supportive evidence for the basic premise and mechanisms of our theory.

[5] Other papers on the equilibrium effects of benchmarking include Qiu (2017) and Cvitanic and Xing (2018). See Garcia and Vanden (2009), Gorton, He, and Huang (2010), Kyle, Ou-Yang, and Wei (2011), Malamud and Petrov (2014), Sato (2016), Huang (2018), and Sockin and Xiaolan (2018) for other models that determine jointly asset management contracts and equilibrium prices.

A. *Deviation from Benchmark as a Fund Characteristic*

A fund's deviation from its benchmark index can be measured by comparing the return of the fund to that of the index or by comparing the portfolio weights. A measure that reflects the first comparison is TE. TE is commonly defined (e.g., Roll 1992; Grinold and Kahn 2000; Jorion 2003) as the standard deviation of the difference between the return of a fund and the return of its benchmark index:

$$\mathrm{TE}_{\mathrm{fund},t} \equiv \sqrt{\mathbb{V}\mathrm{ar}(R_{\mathrm{fund},t} - R_{\mathrm{index},t})}. \tag{1}$$

A measure that reflects the second comparison is AS. AS is computed (Cremers and Petajisto 2009) by taking the absolute value of the difference between an asset's portfolio weight in the fund and in the fund's benchmark index, summing across assets, and dividing by 2:

$$\mathrm{AS}_{\mathrm{fund},t} = \frac{1}{2}\sum_{n=1}^{N} |w_{\mathrm{fund},n,t} - w_{\mathrm{index},n,t}|. \tag{2}$$

When portfolio weighs are nonnegative, AS lies between 0 and 1.

Table 1 presents evidence on the heterogeneity of TE across funds and on the persistence of TE over time for a given fund. Table 2 does the same for AS. These tables extend findings of Cremers and Petajisto (2009) to our sample period.

Tables 1 and 2 as well as the rest of the empirical exercise in section II are calculated for actively managed mutual funds that invest in US stocks. Our sample period is January 1999 to December 2018. We express TE, AS, returns, and portfolio weights as percentages: for example, an AS of 0.9 as 90 and a portfolio weight of 1% as 1. We compute TE using daily fund returns over a 1-year lookback window.

We source asset and fund returns from the Center for Research in Security Prices (CRSP), fund portfolio weights from Thomson Reuters, fund benchmark indices from Morningstar, index returns from Morningstar, and index weights from CRSP, Morningstar, and Russell. Because our data set includes only Standard and Poor's (S&P) 500 and Russell index weights, we exclude funds benchmarked on other indices. We also exclude funds

TABLE 1
Cross-Sectional Heterogeneity and Persistence of TE

		Quintile After		
	TE	1 Year	2 Years	3 Years
TE quintile 1	2.11	1.37	1.53	1.63
TE quintile 5	8.49	4.65	4.51	4.43

Note.—The table shows the average TE and average TE quintile 1, 2, and 3 years later for active funds in quintiles 1 and 5 of TE.

TABLE 2
Cross-Sectional Heterogeneity and Persistence of AS

		Quintile After		
	AS	1 Year	2 Years	3 Years
AS quintile 1	52.24	1.27	1.40	1.50
AS quintile 5	91.11	4.81	4.72	4.66

Note.—The table shows the average AS and average AS quintile 1, 2, and 3 years later for active funds in quintiles 1 and 5 of AS.

whose combined portfolio weight in stocks does not always lie between 80 and 120, funds younger than 1 year, and funds that CRSP flags as index funds or exchange-traded funds or that have index-related words in their name.

Our main sample consists of funds whose benchmark indices include large stocks. These indices in our data are the S&P 500 and the Russell 200, 1000, and 3000 and their value and growth versions. There are 1,118 funds with these indices as their benchmarks (as reported by Morningstar). We also report findings in appendix C for an additional sample of 677 funds whose benchmark indices include only small or midcap stocks. These indices are the Russell 2000 and 2500, Midcap and Small Cap Completeness, and their value and growth versions.

We compute the statistics in table 1 by sorting funds into quintiles at the end of each year on the basis of their TE. The average TE for funds in quintile 1 is 2.11, while the average TE for funds in quintile 5 is 8.49, about four times as large. The large differences in TE across funds seem to reflect an underlying characteristic that persists over time in relative terms. The average quintile where quintile 1 funds in a given year lie 1, 2, and 3 years later is 1.37, 1.53, and 1.63, respectively.

The columns in table 2 are constructed in the same manner as those in table 1, except that funds are sorted into quintiles at the end of each year on the basis of their AS. AS is somewhat more persistent than TE in relative terms. The average quintile where quintile 1 funds in a given year lie 1, 2, and 3 years later is 1.27, 1.40, and 1.50, respectively.[6]

B. Funds Trade to Maintain Deviations within Bounds

One might argue that the persistence of TE and AS shown in tables 1 and 2 is due to inertia. If a fund holds a portfolio close to the benchmark at

[6] The high persistence of TE and AS for funds in quintile 1 is not driven by closet indexers, defined as active funds that invest in a near-passive manner. Cremers and Petajisto (2009) take closet indexers to be the funds with AS below 60. When excluding these funds from our sample and recomputing the quintiles, we find that the average quintile where quintile 1 funds in a given year lie 3 years later rises from 1.63 to 1.76 in the case of TE and from 1.50 to 1.68 in the case of AS. Hence, TE and AS remain quite persistent.

the end of a given year and turns over a small fraction of its portfolio during the following year, then the portfolio at the end of that year will be close to the benchmark as well. The significant persistence of TE and AS over 2- and 3-year horizons is evidence against the inertia explanation. We next present more direct evidence by showing that funds with low TE or AS trade actively to maintain their deviations from benchmark indices within bounds.

We measure the extent to which a fund underweights or overweights an asset by active weight, defined as the asset's portfolio weight in the fund minus the weight in the fund's benchmark index:

$$Aw_{\text{fund},n,t} = w_{\text{fund},n,t} - w_{\text{index},n,t}.$$

We compute funds' trading activity at a semiannual frequency. If a fund does not trade during semester t, then asset n's portfolio weight $\hat{w}_{\text{fund},n,t}$ in the fund at the end of semester t is

$$\hat{w}_{\text{fund},n,t} = \frac{w_{\text{fund},n,t-1}(1 + R_{n,t})}{\sum_{n'=1}^{N} w_{\text{fund},n',t-1}(1 + R_{n',t})}, \tag{3}$$

where $w_{\text{fund},n,t-1}$ is asset n's weight at the end of semester $t - 1$ and $R_{n,t}$ is asset n's return during semester t.[7] The change in asset n's weight due to trading during semester t is asset n's weight at the end of semester t minus asset n's no-trade weight:

$$\Delta w_{\text{fund},n,t} = w_{\text{fund},n,t} - \hat{w}_{\text{fund},n,t}.$$

The assets held by the funds in our sample are mainly stocks but also include cash and bonds. We observe the returns on stocks. We take the return on cash to be the 1-month US Treasury bill rate (rolled over a semester) and the return on bonds to be that of the Bloomberg (Lehman) Aggregate Bond Index.

Table 3 presents evidence on how funds trade stocks that they underweight or overweight relative to their benchmark index, as a function of the stocks' return. At the end of each semester $t - 1$, we compute active weight for each fund/stock pair such that the stock belongs to the 100 largest stocks in the fund's benchmark index. We sort these fund/stock pairs into deciles based on active weight and then sort within each decile into quintiles based on the stock's return during the next semester t. We compute the change in the stock's portfolio weight in the fund due to trading during semester t (col. 3) and during the year formed by semesters t and

[7] Equation (3) remains valid when the fund trades because of inflows or outflows or because of assets paying dividends. In the case of flows, we must assume that when the fund experiences inflows (outflows), it buys (sells) assets according to its current portfolio weights. In the case of dividends, we must assume that they are reinvested in the assets paying them and define the return $R_{n,t}$ to include the dividends.

TABLE 3
FUNDS' TRADING OF STOCKS AS FUNCTION OF STOCKS' ACTIVE WEIGHT AND RETURN

			Change in Weight	
	Active Weight (1)	Return (2)	6 Months (3)	1 Year (4)
Active weight decile 1 and return quintile 1	−1.66	−15.53	.08	.16
Active weight decile 1 and return quintile 5	−1.66	21.03	.20	.33
Active weight decile 10 and return quintile 1	2.09	−15.76	−.51	−.87
Active weight decile 10 and return quintile 5	2.09	28.25	−.56	−1.03

NOTE.—The table shows the change in weight due to trading for fund/stock pairs in deciles 1 and 10 of active weight and in quintiles 1 and 5 of stock return. The change in weight is computed over the same semester or over the same and subsequent semester as the return.

$t + 1$ (col. 4). We consider only the 100 largest stocks in each fund's benchmark index so that stock returns have a significant effect on index weights. As a robustness check, we extend the set of stocks to include the 250 largest. The effects are in the same direction but slightly weaker (table C1 in app. C). For the indices in our main sample, the 100 largest stocks account for 54%–72% of index value and the 250 largest stocks for 71%–89%.

A first observation from table 3 is that funds tend to buy stocks that they underweight and sell stocks that they overweight—an effect also documented in DeVault, Sias, and Starks (2019). The change in portfolio weight due to trading is positive for fund-stock pairs in decile 1 of active weight and negative for pairs in decile 10, regardless of whether the stock earns a low return (quintile 1 of return) or high return (quintile 5).

A second observation from table 3, which is particularly relevant for our theory, is that funds' buying of underweighted stocks is procyclical and selling of overweighted stocks is countercyclical. Funds increase the portfolio weight of an underweighted stock that underperforms (decile 1 and quintile 1) by 0.16 in the semesters during and following the underperformance. They increase the weight of an equally underweighted but overperforming stock (decile 1 and quintile 5) by 0.33, so twice as much. Conversely, funds decrease the weight of an overweighted stock that underperforms (decile 10 and quintile 1) by 0.87. They decrease the weight of an equally overweighted but overperforming stock (decile 10 and quintile 5) more, by 1.03.

Funds' buying of underweighted stocks is procyclical for all such stocks and not only for the stocks with the most negative active weight in decile 1. Active weight is negative for fund/stock pairs in deciles 1–7, is approximately zero in decile 8, and is positive in deciles 9 and 10. Active weight is positive in only two deciles because funds hold a relatively small fraction of the stocks in their benchmark index—the average fund in our sample

gives nonzero weight to only 23 of the largest 100 stocks in its benchmark index. In all deciles 1–7, funds buy stocks on average, and buying is procyclical. In deciles 9 and 10, funds sell stocks on average. Selling is procyclical in decile 9 but becomes countercyclical in decile 10. Aggregating across deciles 9 and 10, we find that selling of overweighted stocks is slightly countercyclical. (Table C2 extends table 3 to all deciles.[8]) Thus, procyclical buying of underweighted stocks is more pervasive than countercyclical selling of overweighted stocks.

Averaging across all deciles of active weight, our findings imply that mutual funds engage in procyclical trading. Momentum (procyclical) trading by mutual funds is documented in Nofsinger and Sias (1999) and Wermers (1999). We complement these papers by showing that momentum trading is driven by the stocks that funds underweight.

We next examine how funds' procyclical trading of underweighted stocks depends on their TE and AS. We proceed as in table 3, sorting fund/stock pairs into deciles based on active weight at the end of each semester $t - 1$. We then sort within each decile into two sets of quintiles based on the stock's return during the next semester t and based on the TE or AS of the fund at the end of semester $t - 1$. We compute the change in the stock's portfolio weight in the fund due to trading during semester t and during the year formed by semesters t and $t + 1$. Figure 1 plots the change in weight as a function of the quintile of TE (fig. 1*A*) and of AS (fig. 1*B*) for the stocks in decile 1 of active weight. The dashed lines represent the stocks with the lowest return (quintile 1 of return), and the solid lines represent the stocks with the highest return (quintile 5). The gray lines represent the change in weight over semester t (during which return is calculated), and the black lines represent the change in weight over semesters t and $t + 1$.

Figure 1 shows that procyclical trading of underweighted stocks is driven by the funds with low TE or AS. Indeed, for funds in the highest TE or AS quintile, purchases of underweighted stocks are almost independent of performance (the dashed and solid lines are close to each other). When moving to the lower TE or AS quintiles, purchases of underweighted underperforming stocks rise slightly (the dashed line has a small negative slope), while purchases of underweighted overperforming stocks rise sharply (the solid line has a large negative slope). The negative relationship between the procyclical buying of underweighted stocks and TE or AS (larger negative slope of the solid line than of the dashed line) is statistically significant at the 1% level (table C4).

[8] We construct a counterpart of table C2 that excludes stocks with zero weights (table C3). Table C3 resembles a truncated version of table C2 to its larger deciles, with procyclical buying of underweighted stocks and countercyclical selling of overweighted stocks.

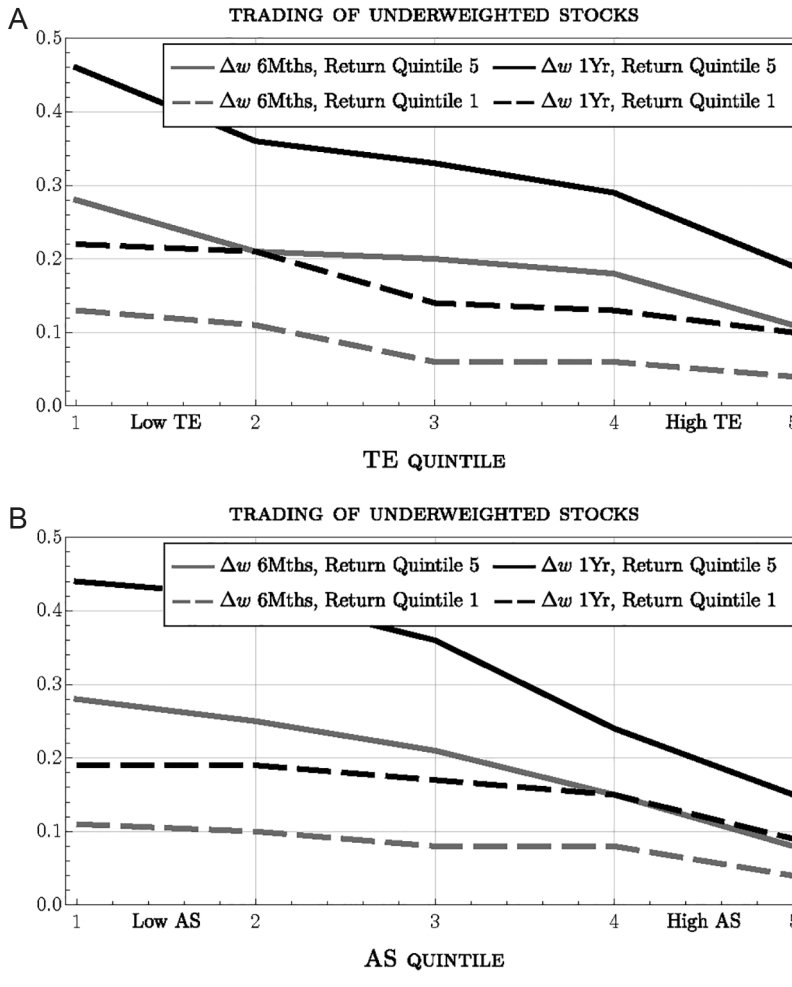

FIG. 1.—Funds' trading of stocks in decile 1 of active weight as a function of stocks' return and funds' TE and AS.

The negative relationship between procyclical buying and TE or AS extends beyond the stocks with the most negative active weight in decile 1. It remains negative and statistically significant at the 1% level when deciles 1–7 are pooled together, except for TE and the 6-month horizon where significance drops to 5% (table C5). When each decile is considered in isolation, significance is at the 1% level for AS and deciles 1–7 and for TE and deciles 1 and 2. In terms of economic significance, funds in the lowest TE or AS quintile eliminate 40% of their underweight in overperforming stocks in deciles 1–7 in the semesters during and following the overperformance. By

comparison, they eliminate 20% of their underweight in underperforming stocks in the same deciles.

When restricting our analysis to funds whose benchmark indices include only small or midcap stocks, we still find procyclical buying of underweighted stocks, countercyclical selling of overweighted stocks, and a negative relationship between procyclical buying and TE or AS (table C6; fig. C1). The procyclicality is weaker, which is consistent with our theory because the weights of the largest stocks in small or midcap indices are smaller than for large-cap indices.

We finally perform the same analysis at the level of industry sectors rather than stocks. We classify the stocks held by each active fund into 11 industry sectors, using the Global Industry Classification Standard (GICS), and compute the fund's industry portfolio weights. We compute the same weights for the fund's benchmark index. Active weight for sectors is distributed more symmetrically around zero than for stocks: it is negative for fund/sector pairs in deciles 1–5, is approximately zero in decile 6, and is positive in deciles 7–10. Thus, funds include stocks from most sectors in their portfolio, even though each sector may be represented by a few stocks. Funds buy underweighted sectors procyclically. They sell overweighted sectors procyclically in deciles 7–9 and countercyclically in decile 10 (table C7). Funds' procyclical buying of underweighted sectors is more pronounced for funds with low TE or AS (fig. C2).

C. Asset-Level versus Portfolio-Level Constraints

TE and AS are portfolio-level measures of a fund's deviation from its benchmark index, and bounds on these measures are portfolio-level constraints. In our model, we assume instead constraints at the level of individual assets, which we interpret as stocks or industry sectors. The evidence on procyclical trading of underweighted stocks or sectors—and on a negative relationship between that trading and TE or AS—is consistent with portfolio- or asset-level constraints. We next provide evidence that asset-level constraints matter.

One test is whether trading of underweighted stocks remains procyclical when TE or AS does not change. As in table 3, we sort fund/stock pairs into deciles based on active weight at the end of each semester $t - 1$. We focus on decile 1 and sort within it into two sets of quintiles based on the stock's return during the next semester t and based on the no-trade change in AS during semester t. The no-trade change in AS is the change computed under the assumption that the fund does not trade during semester t, in which case portfolio weights are given by (3). If the no-trade change in AS is zero, then constraints on AS should not induce trade by the fund. We do not compute a no-trade change for TE because of measurement

issues explained in appendix C, but we supplement our analysis with actual change in TE and AS.

We find that changes in weights are similar to those in table 3 across the first four quintiles of no-trade change in AS—including in the third quintile, in which the no-trade change in AS is almost zero (table C8). This suggests that asset-level constraints matter. Changes in weights are larger in the fifth quintile, and procyclicality is about 50% larger than in the other quintiles. Since the no-trade change in AS is largest in the fifth quintile, portfolio-level constraints seem to matter as well. We corroborate these findings with a regression analysis, with no-trade change in AS as a continuous variable (table C9). The findings are similar when using actual change in TE or AS.

Another test is whether portfolio spillover effects are small. Under portfolio-level constraints, a high return by a heavily underweighted stock n should induce a fund to not only buy stock n but also execute similar-size purchases of other heavily underweighted stocks and sales of heavily overweighted stocks. We characterize spillovers associated with stock n by adding (negative) weight changes due to trading for stocks other than n that the fund overweights and subtracting (positive) weight changes for stocks other than n that the fund underweights. We compute the procyclicality of the resulting quantity by comparing it for high and low values of stock n's return. We find that the high minus low difference aggregated across all industry sectors other than stock n's is less than 40% of the same difference in stock n's sector (table C10). Thus, spillover effects across sectors appear to be small, providing further evidence for asset-level constraints.

III. Model

Time t is continuous and goes from zero to infinity. The riskless rate is exogenous and equal to $r > 0$. There are N risky assets. Asset $n = 1, \ldots, N$ pays a dividend flow D_{nt} per share and is in supply of $\theta_n > 0$ shares. The price S_{nt} per share of the risky asset is determined endogenously in equilibrium.

The return per share of risky asset n in excess of the riskless rate is

$$dR^{sh}_{nt} \equiv D_{nt}\,dt + dS_{nt} - rS_{nt}\,dt, \tag{4}$$

and its return per dollar in excess of the riskless rate is

$$dR_{nt} \equiv \frac{dR^{sh}_{nt}}{S_{nt}} = \frac{D_{nt}\,dt + dS_{nt}}{S_{nt}} - r\,dt. \tag{5}$$

We refer to dR^{sh}_{nt} as share return, omitting that it is in excess of the riskless rate. We refer to dR_{nt} as return, omitting that it is per dollar and in excess of the riskless rate.

The dividend flow D_{nt} of risky asset n follows the square root process

$$dD_{nt} = \kappa_n(\bar{D}_n - D_{nt})dt + \sigma_n\sqrt{D_{nt}}dB_{nt}, \tag{6}$$

where $\{\kappa_n, \overline{D}_n, \sigma_n\}_{n=1,\ldots,N}$ are positive constants and B_{nt} is a Brownian motion. For simplicity, we take the Brownian motions $\{B_{nt}\}_{n=1,\ldots,N}$ to be independent, thus assuming that assets have independent dividends.

The square root specification (6) allows for closed-form solutions while also ensuring that dividends remain positive. A property of the square root specification that is important for our results is that the volatility (standard deviation) of dividends per share D_{nt} increases with the level of dividends. This property is realistic: if a firm becomes larger and keeps the number of its shares constant, then its dividends per share become more uncertain.[9]

Investors form a continuum with measure 1. They are of two types: *unconstrained investors*, who can invest in all assets without any limitations, and *constrained investors*, who are limited in the risk they can take. Unconstrained investors are in measure $1 - x \in (0, 1)$, and constrained investors are in the complementary measure x. We denote by W_{1t} and W_{2t} the wealth of an unconstrained and a constrained investor, respectively, and by z_{1nt} and z_{2nt} the number of shares of risky asset n that they hold.

At time t, investors choose their asset positions to maximize the mean-variance objective

$$\mathbb{E}_t(dW_{it}) - \frac{\rho}{2}\mathbb{V}\mathrm{ar}_t(dW_{it}), \tag{7}$$

subject to the budget constraint

$$\begin{aligned} dW_{it} &= \left(W_{it} - \sum_{n=1}^{N} z_{int}S_{nt}\right)rdt + \sum_{n=1}^{N} z_{int}(D_{nt}dt + dS_{nt}) \\ &= W_{it}rdt + \sum_{n=1}^{N} z_{int}dR^{sh}_{nt}, \end{aligned} \tag{8}$$

where ρ is a risk aversion coefficient common to all investors, $i = 1$ for unconstrained investors, and $i = 2$ for constrained investors. The mean

[9] Two alternative common specifications of dividends are geometric Brownian motion (GBM) and arithmetic Brownian motion (ABM). The volatility of dividends per share is proportional to the dividend level under GBM and is independent of the dividend level under ABM. Our main results would hold under GBM, but we do not adopt that specification because it does not yield closed-form solutions. ABM yields closed-form solutions, but our main results would hold only under the AS-based constraint (10) and not under the TE-based constraint (12). Indeed, under ABM and the TE-based constraint, equilibrium prices would be linear functions of dividends, volatility per share would be constant, and the constraint would not tighten when dividends increase. An additional drawback of ABM is that dividends and prices can become negative, complicating calculations of returns. Dividends remain positive under GBM (as they do under the square root specification).

and variance in the objective (7) are computed over the infinitesimal change in investor wealth. That change is equal to the riskless rate paid on wealth between t and $t + dt$ plus the sum over risky assets of the capital gains from each risky asset in excess of the riskless rate. The capital gains for risky asset n are equal to the number of shares z_{int} times the share return dR^{sh}_{nt}.

The objective (7) renders our equilibrium analysis tractable because of two key properties: (i) the coefficient of absolute risk aversion is independent of wealth and (ii) there is no intertemporal hedging demand. Investors with the objective (7) can be interpreted as infinitely lived, but in that case (7) does not follow from a Von Neumann-Morgenstern (VNM) utility over intertemporal consumption.[10] Alternatively, investors can be interpreted as overlapping generations living over infinitesimal periods. In that case, (7) follows from all VNM utilities, with ρ equal to $-U''(W)/U'(W)$ and W equal to the investors' initial wealth.

The constraint limits the deviation of each constrained investor's risky asset portfolio from a benchmark index. We denote by $\hat{\eta}_n$ the number of shares of risky asset n in the index and interpret it as the number of shares sold by the issuing firm. The supply θ_n of asset n can differ from $\hat{\eta}_n$ because it includes demand by other (unmodeled) traders. We refer to these traders as noise traders. Their demand can differ across assets in a way not proportional to $\hat{\eta}_n$. We take their demand to be constant over time (capturing slow mean reversion) and treat it as a model parameter in section IV. Because demand differs across assets, however, it is effectively random at the stage when the parameters of the constraint are determined in section V.

We consider two specifications of the constraint. The first specification is in the spirit of AS, defined in (2). Suppose that a constrained investor allocates wealth W_{2zt} in a fund investing in the risky assets and possibly in the riskless asset and allocates the remaining wealth $W_{2t} - W_{2zt}$ in the riskless asset. The weight of asset n in the fund is $z_{2nt}S_{nt}/W_{2zt}$. The weight of asset n in the benchmark index is $\hat{\eta}_n S_{nt}/\Sigma^N_{n'=1}\hat{\eta}_{n'}S_{n't}$. Requiring the two weights to differ by no more than $\hat{L} \geq 0$ yields the constraint

$$\left| \frac{z_{2nt}S_{nt}}{W_{2zt}} - \frac{\hat{\eta}_n S_{nt}}{\sum^N_{n'=1}\hat{\eta}_{n'}S_{n't}} \right| \leq \hat{L}, \tag{9}$$

a simplified version of which we adopt below. The constraint (9) is in the spirit of AS because AS is based on the difference in portfolio weights. We impose a bound not on AS but on the difference in portfolio weights asset

[10] In the polar cases where the constraint is absent or is infinitely tight, our equilibrium analysis remains tractable and the results are similar when investors have negative exponential utility over intertemporal consumption. This is shown in an early version of this paper (Buffa, Vayanos, and Woolley 2014).

by asset. This helps eliminate spillover effects, whereby price movements in one asset impact the constraint's tightness for other assets. The evidence in section II motivates an asset-level constraint because spillover effects are small. To fully eliminate spillover effects, rendering our analysis more tractable, we approximate $\Sigma_{n'=1}^{N}\hat{\eta}_{n'}S_{n't}$ in (9) by its unconditional expectation—an approximation that becomes more accurate as the number N of independent risky assets increases. We also take W_{2zt} to be a constant W_{2z}—an assumption that is in the spirit of the previous approximation and consistent with the overlapping generations interpretation. Under these assumptions, (9) simplifies to

$$|z_{2nt} - \eta_n|S_{nt} \leq L \tag{10}$$

for all n, where $\eta_n \equiv [W_{2z}/\mathbb{E}(\Sigma_{n'=1}^{N}\hat{\eta}_{n't}S_{n't})]\hat{\eta}_n$ and $L \equiv \hat{L}W_{2z}$. Equation (10) is our first specification of the constraint. We refer to η_n as the benchmark position.

The second specification is in the spirit of TE, defined in (1). As with the AS-based constraint, we impose a bound not on TE but on the standard deviation of the difference between fund and index return that is generated by each specific asset. Since assets are independent, the standard deviation that is generated by asset n is the absolute value of the difference in portfolio weights times the standard deviation of the asset return. The constraint for asset n thus is

$$\left|\frac{z_{2nt}S_{nt}}{W_{2zt}} - \frac{\hat{\eta}_n S_{nt}}{\sum_{n'=1}^{N}\hat{\eta}_{n'}S_{n't}}\right|\sqrt{\frac{\mathbb{V}\mathrm{ar}_t(dR_{nt})}{dt}} \leq \hat{L}. \tag{11}$$

When $\Sigma_{n'=1}^{N}\hat{\eta}_{n'}S_{n't}$ is approximated by its unconditional expectation and W_{2zt} is a constant W_{2z}, (11) simplifies to

$$|z_{2nt} - \eta_n|\sqrt{\frac{\mathbb{V}\mathrm{ar}_t(dR_{nt}^{sh})}{dt}} \leq L \tag{12}$$

for all n. Equation (12) is our second specification of the constraint.

The constraints (10) and (12) depend on the parameters $(W_{2z}, \hat{L})$. These parameters determine L and η_n for all n. When $\hat{L}$ is infinite, there is no constraint and constrained investors are fully active. When instead $\hat{L} = 0$, constrained investors hold the benchmark index and are fully passive. For intermediate values $\hat{L} \in (0, \infty)$, constrained investors combine elements of active and passive: they are active in the sense that they have some leeway when choosing their position in each risky asset but passive in the sense that they cannot deviate much from the index. We take the parameters $(W_{2z}, \hat{L})$ as exogenous in section IV and endogenize them in section V.

IV. Equilibrium with Exogenous Constraint

A. *No Constraint*

We first derive the equilibrium when the bound $\hat{L}$ in the constraint is infinite and constrained investors are identical to unconstrained investors. We look for an equilibrium in which the price S_{nt} of each risky asset n is a function of that asset's dividend flow D_{nt} only. Denoting that function by $S_n(D_{nt})$ and assuming that it is twice continuously differentiable, we can write the share return dR^{sh}_{nt} as

$$\begin{aligned} dR^{sh}_{nt} &= D_{nt}dt + dS_n(D_{nt}) - rS_n(D_{nt})dt \\ &= \left[D_{nt} + \kappa_n(\bar{D}_n - D_{nt})S_n'(D_{nt}) + \frac{1}{2}\sigma_n^2 D_{nt}S''(D_{nt}) - rS_n(D_{nt})\right]dt \\ &\quad + \sigma_n\sqrt{D_{nt}}S'(D_{nt})dB_{nt}, \end{aligned} \tag{13}$$

where the second step follows from (6) and Ito's lemma.

Using the budget constraint (8) and the mutual independence of the Brownian motions $\{B_{nt}\}_{n=1,\ldots,N}$, we can write the objective (7) as

$$\sum_{n=1}^{N} z_{int}\mathbb{E}_t(dR^{sh}_{nt}) - \frac{\rho}{2}z_{int}^2\mathbb{V}\mathrm{ar}_t(dR^{sh}_{nt}).$$

The first-order condition with respect to z_{int} is

$$\mathbb{E}_t(dR^{sh}_{nt}) = \rho z_{int}\mathbb{V}\mathrm{ar}_t(dR^{sh}_{nt}). \tag{14}$$

The expected share return $\mathbb{E}_t(dR^{sh}_{nt})$ is the drift term in (13), and the share return variance $\mathbb{V}\mathrm{ar}_t(dR^{sh}_{nt})$ is the square of the diffusion term.

Since unconstrained and constrained investors are identical, the market-clearing condition

$$(1 - x)z_{1nt} + xz_{2nt} = \theta_n \tag{15}$$

implies $z_{1nt} = z_{2nt} = \theta_n$. Each investor's position in asset n is thus equal to the asset's supply θ_n, which coincides with the supply per investor since investors form a continuum with mass 1. Setting $z_{int} = \theta_n$ in (14), we find the following ordinary differential equation (ODE) for the function $S_n(D_{nt})$:

$$D_{nt} + \kappa_n(\bar{D}_n - D_{nt})S_n'(D_{nt}) + \tfrac{1}{2}\sigma_n^2 D_{nt}S_n''(D_{nt}) - rS(D_{nt}) = \rho\theta_n\sigma_n^2 D_{nt}S_n'(D_{nt})^2. \tag{16}$$

The ODE (16) is second order and nonlinear and must be solved over $(0, \infty)$. We require that its solution $S_n(D_{nt})$ has a derivative that converges to finite limits at zero and infinity. This yields one boundary condition at zero and one at infinity.

We look for an affine solution to the ODE (16):

$$S_n(D_{nt}) = a_{n0} + a_{n1}D_{nt}, \tag{17}$$

where (a_{n0}, a_{n1}) are constant coefficients. This function satisfies the boundary conditions since its derivative is constant. Substituting this function into (16) and identifying terms, we can compute (a_{n0}, a_{n1}).

PROPOSITION 1. Suppose $\hat{L} = \infty$ and $\theta_n > -[(r + \kappa_n)^2/4\rho\sigma_n^2]$. An affine solution $S_n(D_{nt}) = a_{n0} + a_{n1}D_{nt}$ to (16) exists, with

$$a_{n0} = \frac{\kappa_n}{r} a_{n1}\bar{D}_n, \tag{18}$$

$$a_{n1} = \frac{2}{r + \kappa_n + \sqrt{(r + \kappa_n)^2 + 4\rho\theta_n\sigma_n^2}}. \tag{19}$$

The price $S_n(D_{nt})$ of asset n and the sensitivity $S_n'(D_{nt})$ of the price to changes in the dividend flow D_{nt} are decreasing and convex functions of the asset's supply θ_n.

The intuition for (18) and (19) is as follows. The coefficient a_{n1} is the sensitivity $S_n'(D_{nt})$ of the price of risky asset n to changes in the asset's dividend flow D_{nt}. Consider a unit increase in D_{nt}. When the asset's supply θ_n is equal to zero, (19) implies that the price S_{nt} increases by $a_{n1} = 1/(r + \kappa_n)$. This is the present value of the increase in future expected dividends discounted at the riskless rate r. Indeed, a unit increase in D_{nt} raises the expected dividend flow $E_t(D_{nt'})$ at time $t' > t$ by $e^{-\kappa_n(t'-t)}$. Hence, the present value of future expected dividends increases by

$$\int_t^\infty e^{-\kappa_n(t'-t)} e^{-r(t'-t)}\, dt' = \frac{1}{r + \kappa_n}.$$

When the supply θ_n is positive, the price $S_n(D_{nt})$ increases by $a_{n1} < 1/(r + \kappa_n)$ in response to a unit increase in D_{nt}. This is because the increase in D_{nt} not only raises expected dividends but also makes them riskier because of the square root specification of D_{nt}. Moreover, since investors hold a long position, the increase in risk makes them more willing to unwind their position and sell the asset. This results in a smaller price increase than when $\theta_n = 0$. When instead $\theta_n < 0$, investors hold a short position, and the increase in risk makes them more willing to buy the asset. This results in a larger price increase than when $\theta_n = 0$, that is, $a_{n1} > 1/(r + \kappa_n)$. Equation (19) confirms that a_{n1} decreases in θ_n.

The effect of θ_n on a_{n1} is stronger when θ_n is small, implying that the price sensitivity $S_n'(D_{nt})$ is convex in θ_n. Convexity is related to a_{n1} being decreasing in θ_n and bounded below by zero. Indeed, these properties imply that the derivative of a_{n1} with respect to θ_n converges to zero when θ_n becomes large (while it is negative for smaller values of θ_n).

The coefficient a_{n0} is equal to the price level when the dividend flow D_{nt} is zero. If the mean-reversion parameter κ_n were equal to zero—and thus the dividend flow were to stay at zero forever—then a_{n0} would be equal to zero. Because, however, κ_n is positive—and thus the dividend flow returns with certainty to positive values—a_{n0} is positive. Moreover, a_{n0} inherits properties of a_{n1} since the larger a_{n1} is, the more the price increases when the dividend flow becomes positive. In particular, a_{n0} is decreasing and convex in the supply θ_n of the risky asset, and so is the price $S_{nt} = a_{n0} + a_{n1}D_{nt}$. Corollary 1 examines how θ_n affects the asset's expected return and the return volatility.

COROLLARY 1. Suppose $\hat{L} = \infty$ and $\theta_n > -[(r + \kappa_n)^2/4\rho\sigma_n^2]$. An increase in the supply θ_n of risky asset n raises the asset's conditional expected return $\mathbb{E}_t(dR_{nt})$ and leaves the return's conditional volatility $\sqrt{\mathbb{V}\text{ar}_t(dR_{nt})}$ unaffected. The effects on the unconditional values $\mathbb{E}(dR_{nt})$ of expected return and $\sqrt{\mathbb{V}\text{ar}(dR_{nt})}$ of volatility are the same as on the conditional values.

Recall from (5) that the return of the risky asset is

$$dR_{nt} = \frac{D_{nt}}{S_{nt}}\,dt + \frac{dS_{nt}}{S_{nt}} - r dt.$$

Return volatility is caused by the term dS_{nt}/S_{nt}, that is, the capital gains per dollar invested. Since an increase in θ_n lowers the sensitivity a_{n1} of the price S_{nt} to changes in the dividend flow D_{nt}, it makes the capital gains $dS_{nt} = a_{n1}\,dD_{nt}$ per share less volatile. At the same time, the share price $S_{nt} = a_{n0} + a_{n1}D_{nt}$ decreases. Because θ_n has the same percentage effect on a_{n0} and a_{n1}, the capital gains dS_{nt}/S_{nt} per dollar invested do not change, and neither does return volatility $\sqrt{\mathbb{V}\text{ar}_t(dR_{nt})}$. On the other hand, expected return $\mathbb{E}(dR_{nt})$ increases because of the term $(D_{nt}/S_{nt})dt$, that is, the dividends per dollar invested. An increase in θ_n does not affect the dividend flow D_{nt} per share but lowers the share price S_{nt}.

B. *Infinitely Tight Constraint*

We next derive the equilibrium when the bound $\hat{L}$ in the constraint is zero and constrained investors hold the benchmark position of η_n shares in each risky asset n. Since the constrained investors' position z_{2nt} is equal to η_n, the market-clearing condition (15) implies $z_{1nt} = (\theta_n - x\eta_n)/(1 - x)$. Substituting z_{1nt} into (14) for $i = 1$, we find the ODE

$$\begin{aligned} D_{nt} + \kappa_n(\bar{D}_n - D_{nt})S_n'(D_{nt}) + \frac{1}{2}\sigma_n^2 D_{nt} S_n''(D_{nt}) - rS_n(D_{nt}) \\ = \frac{\rho(\theta_n - x\eta_n)}{1 - x}\sigma_n^2 D_{nt} S_n'(D_{nt})^2. \end{aligned} \tag{20}$$

The ODE (20) is identical to (16) except that supply θ_n is replaced by $(\theta_n - x\eta_n)/(1 - x)$. The solution $S_n(D_{nt})$ of the ODE can be derived from proposition 1 with the same substitution.

PROPOSITION 2. Suppose $\hat{L} = 0$ and $\theta_n > x\eta_n - \{[(1 - x)(r + \kappa_n)^2]/4\rho\sigma_n^2\}$. An affine solution $S_n(D_{nt}) = a_{n0} + a_{n1}D_{nt}$ to (20) exists, with a_{n0} given by (18) and

$$a_{n1} = \frac{2}{r + \kappa_n + \sqrt{(r + \kappa_n)^2 + \{[4\rho(\theta_n - x\eta_n)]/(1 - x)\}\sigma_n^2}}. \quad (21)$$

Relative to the case $\hat{L} = \infty$:

- $S_n(D_{nt})$ is lower when $\theta_n > \eta_n$ and higher when $\theta_n < \eta_n$.
- $S_n'(D_{nt})$ is lower when $\theta_n > \eta_n$ and higher when $\theta_n < \eta_n$.

Under an infinitely tight constraint ($\hat{L} = 0$), noise trader demand has a larger effect on the price than under no constraint ($\hat{L} = \infty$). Recall from proposition 1 that when $\hat{L} = \infty$, the price decreases in the supply θ_n of the risky asset. In particular, the price is higher when $\theta_n < \eta_n$, corresponding to high noise trader demand, than when $\theta_n > \eta_n$, corresponding to low noise trader demand. When $\hat{L} = 0$, the difference is exacerbated: the price is even higher when $\theta_n < \eta_n$ and is even lower when $\theta_n > \eta_n$. Intuitively, the constraint exacerbates the effect that noise trader demand has on the price because it prevents constrained investors from absorbing that demand. Indeed, if the constraint is imposed, constrained investors must change their position from θ_n to η_n. Therefore, they must buy the asset when $\theta_n < \eta_n$, which is when noise trader demand is high, and must sell the asset when $\theta_n > \eta_n$, which is when noise trader demand is low.

The constraint exacerbates the effects of noise trader demand not only on the price level but also on the price sensitivity to changes in the dividend flow D_{nt}. Recall from proposition 1 that when $\hat{L} = \infty$, the price is more sensitive to D_{nt} (i.e., $S_n'(D_{nt})$ is larger) when $\theta_n < \eta_n$ than when $\theta_n > \eta_n$. When $\hat{L} = 0$, the difference in sensitivities is exacerbated because θ_n is replaced by $(\theta_n - x\eta_n)/(1 - x)$: the price becomes more sensitive to D_{nt} when $\theta_n < \eta_n$ because $(\theta_n - x\eta_n)/(1 - x) < \theta_n$, and it becomes less sensitive to D_{nt} when $\theta_n > \eta_n$ because $(\theta_n - x\eta_n)/(1 - x) > \theta_n$.

While an infinitely tight constraint exacerbates the mispricing, it does not affect return volatility. Indeed, since volatility is independent of θ_n, it does not change when θ_n is replaced by $(\theta_n - x\eta_n)/(1 - x)$.

COROLLARY 2. Suppose $\hat{L} = 0$ and $\theta_n > x\eta_n - \{[(1 - x)(r + \kappa_n)^2]/4\rho\sigma_n^2\}$. The conditional volatility $\sqrt{\mathbb{V}\text{ar}_{nt}(dR_{nt})}$ and the unconditional volatility $\sqrt{\mathbb{V}\text{ar}(dR_{nt})}$ of risky asset n's return are independent of the asset's supply θ_n and are the same as when $\hat{L} = \infty$.

C. General Case

We next derive the equilibrium for $\hat{L} \in (0, \infty)$. The equilibrium is described by an unconstrained region, where the constraint does not bind, and a constrained region, where it binds. We nest the constraints (10) and (12) into

$$|z_{2nt} - \eta_n| G_n(D_{nt}) \leq L, \tag{22}$$

where $G_n(D_{nt}) \equiv S_n(D_{nt})$ in the case of the AS-based constraint (10) and $G_n(D_{nt}) \equiv \sigma_n \sqrt{D_{nt}} S_n'(D_{nt})$ in the case of the TE-based constraint (12). The value of $G_n(D_{nt})$ for the TE-based constraint follows from (13) and assuming that $S_n(D_{nt})$ increases in D_{nt} (which we confirm is the case in equilibrium).

In the unconstrained region, all investors are identical. Therefore, their positions z_{1nt} and z_{2nt} are equal to the supply θ_n, and the function $S_n(D_{nt})$ solves the same ODE (16) as when the constraint never binds. Substituting $z_{2nt} = \theta_n$ into (22), we find that the unconstrained region is defined by

$$|\theta_n - \eta_n| G_n(D_{nt}) \leq L. \tag{23}$$

In the constrained region, (22) holds as an equality. Using the market-clearing condition to write z_{1nt} as a function of z_{2nt} and substituting into (14) for $i = 1$, we find

$$\begin{aligned} D_{nt} + \kappa_n(\bar{D}_n - D_{nt}) S_n'(D_{nt}) + \frac{1}{2}\sigma_n^2 D_{nt} S_n''(D_{nt}) - r S_n(D_{nt}) \\ = \rho \frac{\theta_n - x z_{2nt}}{1 - x} \sigma_n^2 D_{nt} S_n'(D_{nt})^2. \end{aligned} \tag{24}$$

A binding constraint forces the position of constrained investors closer to η_n while keeping it on the same side of η_n as for unconstrained investors. When, for example, $\theta_n < \eta_n$, unconstrained investors hold a position $z_{1nt} < \eta_n$, and constrained investors hold a position $z_{2nt} \in (z_{1nt}, \eta_n)$. Substituting z_{2nt} from (22), which holds as an equality in the constrained region, into (24) and noting that $z_{2nt} - \eta_n$ has the same sign as $\theta_n - \eta_n$, we find the ODE

$$\begin{aligned} D_{nt} + \kappa_n(\bar{D}_n - D_{nt}) S_n'(D_{nt}) + \frac{1}{2}\sigma_n^2 D_{nt} S_n''(D_{nt}) - r S_n(D_{nt}) \\ = \frac{\rho(\theta_n - x\eta_n)}{1 - x} \sigma_n^2 D_{nt} S_n'(D_{nt})^2 - \frac{\rho \operatorname{sgn}(\theta_n - \eta_n) x L}{1 - x} \frac{\sigma_n^2 D_{nt} S_n'(D_{nt})^2}{G_n(D_{nt})}, \end{aligned} \tag{25}$$

where $\operatorname{sgn}(\theta_n - \eta_n)$ is the sign function equal to 1 if $\theta_n > \eta_n$ and to -1 if $\theta_n < \eta_n$. The constrained region is defined by the opposite inequality to (23), that is,

$$|\theta_n - \eta_n| G_n(D_{nt}) > L. \tag{26}$$

The price function $S_n(D_{nt})$ solves the ODE (16) in the unconstrained region (23) and (25) in the constrained region (26). The two ODEs are second order and nonlinear and must be solved as a system over $(0, \infty)$. As in sections IV.A and IV.B, we require that $S_n'(D_{nt})$ converges to finite limits at zero and infinity.

At a boundary point D_n^* between the constrained and the unconstrained region, the values of $S_n(D_n^*)$ implied by the two ODEs must be equal, and the same is true for the values of $S_n'(D_n^*)$. These are the smooth-pasting conditions, and they follow from $S_n(D_{nt})$ being twice continuously differentiable. The boundary point(s) between the constrained and the unconstrained region must be solved together with the ODEs. This makes the problem a free-boundary one.

The system of ODEs (16) and (25) does not have a closed-form solution. We can prove, however, that in the case of the TE-based constraint (12), a solution exists and has a number of key properties. In the case of the AS-based constraint (10), we do not have a general proof but compute the solution numerically and show that it has the same properties.

THEOREM 1. Suppose $G_n(D_{nt}) = \sigma_n \sqrt{D_{nt}} S_n'(D_{nt})$, $\hat{L} \in (0, \infty)$, $\theta_n > x\eta_n - \{[(1 - x)(r + \kappa_n)^2]/4\rho\sigma_n^2\}$, and $\kappa_n \bar{D}_n > \sigma_n^2/4$. A solution $S_n(D_{nt})$ to the system of ODEs (16) in the unconstrained region (23) and of (25) in the constrained region (26), with a derivative that converges to finite limits at zero and infinity, exists and has the following properties:

- It is positive and increasing in D_{nt}.
- It lies between the affine solution derived for $\hat{L} = \infty$ and that derived for $\hat{L} = 0$.
- Its derivative $S'(D_{nt})$ lies between the derivative of the affine solution derived for $\hat{L} = \infty$ and that derived for $\hat{L} = 0$.
- It is concave when $\theta_n > \eta_n$ and convex when $\theta_n < \eta_n$.
- The unconstrained and constrained regions are separated by only one boundary point D_n^*.

Theorem 1 confirms that an increase in the dividend flow D_{nt} raises the price S_{nt}. It also shows that S_{nt} lies between the values that it takes in the polar cases $\hat{L} = \infty$ and $\hat{L} = 0$. For given D_{nt}, the difference in price between $\theta_n < \eta_n$ and $\theta_n > \eta_n$ is positive when there is no constraint ($\hat{L} = \infty$), higher when there is a constraint ($\hat{L} \in (0, \infty)$), and even higher when the constraint is infinitely tight ($\hat{L} = 0$). The same comparisons hold for the difference in price sensitivity $S_n'(D_{nt})$ between $\theta_n < \eta_n$ and $\theta_n > \eta_n$.

A key difference with the polar cases $\hat{L} = \infty$ and $\hat{L} = 0$ is that the price is nonlinear in D_{nt}: it is concave for $\theta_n > \eta_n$ and convex for $\theta_n < \eta_n$, while it is affine in the polar cases. The nonlinearities are driven by the trading

that the constraint induces and in turn drive the risk-return inversion. In the polar cases, there is no constraint-induced trading either because the constraint never binds ($\hat{L} = \infty$) or because constrained investors hold the index ($\hat{L} = 0$).

The intuition for the nonlinearities is as follows. Suppose that $\theta_n > \eta_n$ and D_{nt} is in the constrained region. Following an increase in D_{nt}, investors' positions go up in value and their volatility rises. To continue meeting the constraint, constrained investors must bring their positions closer to η_n. Since $\theta_n > \eta_n$, they must sell some shares of asset n to unconstrained investors. This *dampens* the price rise. The dampening effect is weaker when D_{nt} is smaller and in the unconstrained region because it concerns not actual sales but an expectation that sales might occur in the future. The price increase is thus larger for smaller D_{nt}, resulting in concavity. Conversely, suppose that $\theta_n < \eta_n$ and D_{nt} is in the constrained region. Following an increase in D_{nt}, constrained investors must bring their positions closer to η_n. Since $\theta_n < \eta_n$, they must buy some shares of asset n from unconstrained investors. This *amplifies* the price rise. The amplification effect is weaker when D_{nt} is smaller and in the unconstrained region, resulting in convexity.

To illustrate our results in this and subsequent sections, we use a calibrated example. We set the risk aversion coefficient ρ and the number of shares $\hat{\eta}_n$ of each risky asset $n = 1, \ldots, N$ in the benchmark index to 1. These are normalizations. In the case of ρ, we redefine the numeraire in the units of which wealth is expressed. In the case of $\hat{\eta}_n$, we redefine one share of each asset by rescaling the dividend flow. We set the wealth W_{2z} to $\mathbb{E}(\Sigma_{n'=1}^{N} \hat{\eta}_{n't} S_{n't})$. This implies that the benchmark position η_n in the constraints (10) and (12) is $\eta = \hat{\eta}_n = 1$. We assume that assets differ in their supply θ_n and that supply is distributed symmetrically around $\eta = 1$.

We set the number N of risky assets to 10 and interpret them as industry sector portfolios. An advantage of calibrating our model on industry sectors rather than on individual stocks is that the former are of comparable size (measured in our model by the average dividend per share $\bar{D}_n$), while size varies sharply across the latter. The evidence in section II is consistent with constraints at the level of industry sectors.

We assume that the values of $(\kappa_n, \bar{D}_n, \sigma_n)$ are identical across assets and set them to $(\kappa, \bar{D}, \sigma) = (0.05, 0.15, 0.4)$. We choose these values on the basis of the asset-level volatility and Sharpe ratio. Within our model, we compute unconditional versions of these quantities using the stationary distribution of D_{nt}, which is gamma, with mean $\bar{D} = 0.15$ and 95th percentile $D_{95} = 0.873$. The unconditional volatility averaged across assets is 20.07, and the unconditional Sharpe ratio is 0.27. (As in sec. II, we express TE, AS, returns, and portfolio weights as percentages, and we do the same for return moments.) For comparison, the average volatility of the 11 value-weighted GICS industry sector portfolios within the S&P 500 index is

18.72 during our sample period. Moreover, the average Sharpe ratio of these portfolios is 0.34. While we use three parameters $(\kappa, \bar{D}, \sigma)$ to target two moments (volatility and Sharpe ratio), the third degree of freedom has a small effect on our numerical results. We set the interest rate r to 3%. This parameter has a small effect on our main results.

We set the fraction x of constrained investors to $x = 0.6$; that is, 60% of investors are constrained and 40% are unconstrained. Identifying the set of unconstrained and constrained investors with that of active funds, we can interpret unconstrained investors as the funds in the top two AS quintiles and constrained investors as the funds in the bottom three quintiles.

We assume that the supply θ_n of the 10 assets takes the 10 values (0.6, 0.7, 0.8, 0.9, 1, 1, 1.1, 1.2, 1.3, 1.4). We choose the spread of the distribution of θ_n around $\eta = 1$ on the basis of the AS of the aggregate portfolio of unconstrained and constrained investors. (The supply θ_n corresponds to the aggregate holdings of asset n by unconstrained and constrained investors.) Under our chosen values, this AS is 10.17. The empirical counterpart of this quantity is the AS of the aggregate portfolio of all active funds, constructed at the industry sector level. We construct that portfolio in two steps. First, we compute industry portfolio weights for each active fund in our sample by classifying the stocks held by the fund into the 11 GICS industry sectors. Second, we aggregate these portfolio weights across all active funds by weighting the portfolio of each fund by that fund's assets under management. We compute the AS of the resulting portfolio weights relative to the same weights for the S&P 500 index. AS is 10.81 during our sample period. Repeating this exercise only for funds with S&P 500 as their benchmark yields an AS of 9.78.

We assume the AS-based constraint (10). The results for the TE-based constraint (12) are similar (app. D). We set the upper bound $\hat{L}$ on the deviation between the weight of an asset in the portfolio of constrained investors and in the index to 5 (same as in the simple example in the introduction). We choose $\hat{L}$ on the basis of the difference between the AS of the portfolio of unconstrained investors and that of constrained investors: a smaller $\hat{L}$ implies a tighter constraint and a larger AS difference. The AS, constructed at the industry sector level, of the aggregate portfolio of all S&P 500–benchmarked active funds in the top two (stock-level) AS quintiles is 13.57 during our sample period. Its counterpart for the funds in the bottom three quintiles is 9.93. The difference thus is 3.64, and it rises to 9.00 when using only the top quintile instead of the top two quintiles. For $\hat{L} = 5$, the difference is 4.97 (AS of unconstrained investors is 13.74 and of constrained investors is 8.78). We also consider the value $\hat{L} = 4$, under which the difference is 6.23 (AS of unconstrained investors is 14.77 and of constrained investors is 8.55).

Figure 2*A* plots the price of an asset n as a function of the asset's dividend flow D_{nt}. The thick lines represent the price when there is a constraint

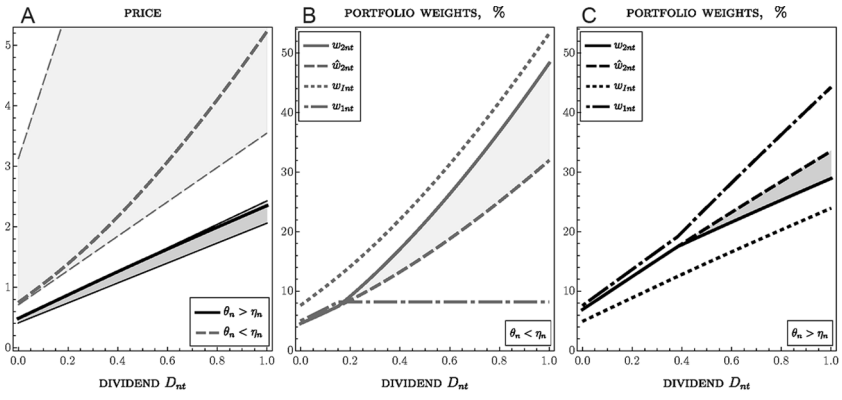

FIG. 2.—Effect of AS-based constraint on prices and portfolio weights.

($\hat{L} \in (0, \infty)$). The thin lines represent the price in the two polar cases where there is no constraint ($\hat{L} = \infty$) and where the constraint is infinitely tight ($\hat{L} = 0$), with the price in the latter case corresponding to the more extreme values. In all three cases, the dashed gray line is drawn for the asset n with $\theta_n = 0.6$ (highest noise trader demand), and the solid black line is drawn for the asset n with $\theta_n = 1.4$ (lowest noise trader demand). The area between the price in the two polar cases is shaded. Consistent with theorem 1, the gray lines lie above the black lines, and the thick lines lie inside the shaded area.

Figure 2*A* shows additionally that noise trader demand has larger effects on prices when it is high ($\theta_n = 0.6$) than when it is low ($\theta_n = 1.4$). Moreover, the asymmetry is more pronounced when the constraint is tighter. We return to the asymmetry in subsequent sections.

Figure 2*B* plots the portfolio weight of the asset n with $\theta_n = 0.6$ as a function of the asset's dividend flow D_{nt}. The dotted line represents the asset's benchmark weight. The solid and dash-dotted lines represent the weight in a constrained and an unconstrained investor's portfolio, respectively. The dashed line represents the weight in a constrained investor's portfolio when he does not trade away from his $D_{nt} = 0$ position when D_{nt} increases. That weight almost coincides with its counterpart no-trade weight for an unconstrained investor (not plotted). When D_{nt} increases within the constrained region, constrained investors buy asset n from unconstrained investors: the difference between the solid and the dashed line (shaded area) increases, and so does the difference between the dashed and the dash-dotted line. Figure 2*C* plots the counterpart portfolio weights of the asset n with $\theta_n = 1.4$. When D_{nt} increases within the constrained region for that asset, constrained investors sell the asset to unconstrained investors. A key implication of the portfolio weight panels is

that constrained investors hold larger positions than unconstrained investors in overvalued assets and smaller positions in undervalued assets.

D. *Risk-Return Inversion*

In the polar cases $\hat{L} = \infty$ and $\hat{L} = 0$, the volatility of an asset's return is the same and is independent of the asset's supply θ_n. For intermediate values of $\hat{L}$ in $(0, \infty)$, volatility differs from the polar cases and depends on supply. When $\theta_n < \eta_n$, volatility is higher than in the polar cases. This is because of the amplification effect, which generates the price convexity in theorem 1. When instead $\theta_n > \eta_n$, volatility is lower than in the polar cases. This is because of the dampening effect, which generates the price concavity. Hence, volatility is higher for an asset n with $\theta_n < \eta_n$ than for an asset n' with $\theta_{n'} > \eta_{n'}$ and identical other characteristics.

PROPOSITION 3. Consider assets (n, n') with $\theta_n < \eta_n$, $\theta_{n'} > \eta_{n'}$, and identical other characteristics. If their prices have the properties in theorem 1, then:

- asset n has higher conditional volatility $\sqrt{\mathbb{V}\text{ar}_t(dR_{nt})}$ and unconditional volatility $\sqrt{\mathbb{V}\text{ar}(dR_{nt})}$ than asset n'; and
- the conditional and unconditional volatilities of asset n are higher than their counterparts for $L = \infty$ and $L = 0$, while those of asset n' are lower.

Proposition 3 implies a negative cross-sectional relationship between volatility and expected return. For asset n with $\theta_n < \eta_n$, expected return is low, so that investors are induced to hold small positions, and volatility is high. For asset n' with $\theta_{n'} > \eta_{n'}$ instead, expected return is high, so that investors are induced to hold large positions, and volatility is low. High volatility goes together with overvaluation (low expected return) because they are both driven by high noise trader demand. Indeed, to accommodate the high demand, investors underweight asset n relative to the benchmark position η_n. When the market goes up, the constraint forces them to underweight less and hence to buy the asset. This yields amplification and high volatility.

A negative cross-sectional relationship between volatility and expected return has been documented empirically and is known as the *volatility anomaly* because it is at odds with standard theories. Haugen and Baker (1996) and Ang et al. (2006) document the volatility anomaly in the cross section of US stocks.

Figure 3 illustrates risk-return inversion in the calibrated example. Figure 3*A* and 3*B* plot the unconditional average of the price and of return volatility, respectively, as functions of $\hat{L}$. In both panels, the dashed gray line is drawn for the asset n with $\theta_n = 0.6$ (highest noise trader demand),

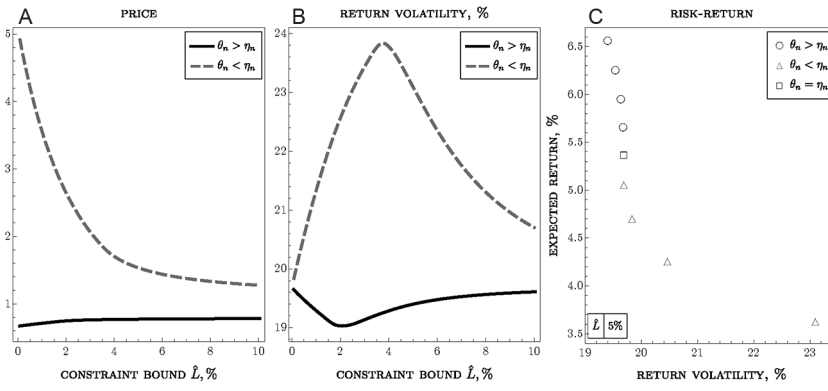

FIG. 3.—Risk-return inversion.

and the solid black line is drawn for the asset n' with $\theta_{n'} = 1.4$ (lowest noise trader demand). Consistent with propositions 1 and 2, the difference between the prices of the two assets increases when the constraint tightens ($\hat{L}$ decreases). Consistent with proposition 3, the difference between the assets' return volatilities is largest for intermediate values of $\hat{L}$. When $\hat{L}$ is below 0.5 or above 10, the volatility difference does not exceed 1, and both volatilities are close to 20. When instead $\hat{L}$ lies between 2.5 and 4.5, the volatility difference exceeds 4. The increase is driven primarily by the amplification effect for the asset in high noise trader demand ($\theta_n = 0.6$). Its volatility rises to above 23, while the volatility of the asset in low demand drops to about 19.

Figure 3*C* plots unconditional expected return as a function of return volatility for $\hat{L} = 5$. The triangles correspond to the assets with $\theta_n < \eta = 1$, the circles to the assets with $\theta_n > \eta = 1$, and the square to the assets with $\theta_n = \eta = 1$. Consistent with proposition 3, variation driven by θ_n generates a negative relationship between volatility and expected return.

An additional measure of risk that we can relate to expected return is CAPM beta. The CAPM predicts a positive relationship between beta and expected return. Empirically, however, a flat or negative relationship has been documented and is known as the *beta anomaly*. Black (1972), Black, Jensen, and Scholes (1972), and Frazzini and Pedersen (2014) document a flat relationship in the cross section of US stocks. Baker, Bradley, and Wurgler (2011) find that the relationship turns negative in recent decades. Asness, Frazzini, and Pedersen (2014) find that the beta anomaly holds across industry sectors as well as within sectors.

Our model generates a negative cross-sectional relationship between beta and expected return. This is because with independent dividend flows, an asset's beta is proportional to the asset's return variance times

the asset's weight in the market portfolio. An asset n with $\theta_n < \eta_n$ has higher beta than an asset n' with $\theta_{n'} > \eta_{n'} = \eta_n$ because it has both higher return volatility (proposition 3) and higher market portfolio weight due to its higher price (theorem 1). In the calibrated example, the beta of asset n with $\theta_n = 0.6$ (highest noise trader demand) exceeds that of asset n' with $\theta_{n'} = 1.4$ (lowest noise trader demand) by 0.65 when $\hat{L} = 4$ and by 0.56 when $\hat{L} = 5$.

PROPOSITION 4. Consider assets (n, n') with $\theta_n < \eta_n = \eta_{n'} < \theta_{n'}$ and identical other characteristics. If their prices have the properties in theorem 1, then asset n has higher conditional and unconditional CAPM beta than asset n'.

In addition to generating volatility and beta anomaly patterns, our model makes two predictions about these anomalies, both of which are borne out in the data. The first is that investors whose deviations from indices are constrained more tightly give larger weight to high-volatility and high-beta assets than less constrained investors. This prediction is consistent with the empirical finding in Christoffersen and Simutin (2017) that mutual fund managers who manage pension fund assets—and hence are evaluated more tightly relative to benchmarks—hold a larger fraction of their portfolios in high-beta stocks. Christoffersen and Simutin (2017) also find that these managers achieve lower CAPM alphas, consistent with our model.

The second prediction is that the profitability of the volatility and beta anomalies derives primarily from the overvalued assets. For example, in figure 3*C*, the expected return of the highest-volatility assets lies at a larger distance below the median than the expected return of the lowest-volatility assets lies above. This reflects the asymmetric effects of noise trader demand: larger effects when demand is high than when it is low. The asymmetry is consistent with the empirical finding in Stambaugh, Yu, and Yuan (2012) that the profitability of anomalies comes primarily from the stocks that are sold short. It is also consistent with the finding in Stambaugh, Yu, and Yuan (2015) that the cross-sectional relationship between volatility and expected return is negative for overvalued stocks and positive for undervalued stocks. Indeed, the negative cross-sectional relationship driven by θ_n dominates the standard positive relationship, driven in our model by the dividend volatility coefficient σ_n, when noise trader demand has large effects, which is when it is high.

A third prediction can be derived from an extension of our model in which constrained investors can trade only gradually to meet their constraint, consistent with the evidence in section II. Assuming that gradual future trading is not fully reflected in current prices (as, e.g., in the rational theory of momentum in Vayanos and Woolley 2013), procyclical buying for overvalued, underweighted assets would generate return momentum. This is consistent with the empirical finding in Favilukis and Zhang

(2021) that the momentum anomaly is more profitable within the set of overvalued (low alpha) stocks. It is also consistent with the finding in Lou, Polk, and Skouras (2019; table 8, panel C) that momentum in intraday returns is more profitable for stocks that mutual funds underweight. Indeed, according to their finding, intraday returns are driven by institutional trading to a larger extent than overnight returns.

E. Overvaluation Bias

Since noise trader demand has asymmetric effects on prices, it does not cancel out when aggregating assets into portfolios, but it introduces an *overvaluation bias*. To show overvaluation bias, we assume that the asset market consists of segments and each segment consists of two subsegments. We identify the assets in our model with the subsegments and assume that noise trader demand differs across them. Propositions 1 and 2 imply that in the polar cases $\hat{L} = \infty$ and $\hat{L} = 0$, the price is a convex function of θ_n. Hence, a segment in which θ_n varies across subsegments trades at a higher price than a segment with lower such variation and same average θ_n.

PROPOSITION 5. Suppose $\hat{L} = \infty$ or $\hat{L} = 0$, and $\theta_n > x\eta_n - \{[(1 - x)(r + \kappa_n)^2]/4\rho\sigma_n^2\}$ for all $n = 1, \ldots, N$. For a segment consisting of assets (n, n') and a segment consisting of assets $(\hat{n}, \hat{n}')$ with $\theta_n < \theta_{\hat{n}} \leq \theta_{\hat{n}'} < \theta_{n'}$, $(\theta_n + \theta_{n'})/2 = (\theta_{\hat{n}} + \theta_{\hat{n}'})/2 \equiv \bar{\theta}$, $\eta_n = \eta_{n'} = \eta_{\hat{n}} = \eta_{\hat{n}'} \equiv \eta$ and other characteristics being identical across assets,

$$\mathcal{O}(D_t) \equiv [S_n(D_t) + S_{n'}(D_t)] - [S_{\hat{n}}(D_t) + S_{\hat{n}'}(D_t)] > 0. \tag{27}$$

Moreover, $\mathcal{O}(D_t)$ is larger when $\hat{L} = 0$ than when $\hat{L} = \infty$ under the sufficient condition $\bar{\theta} \leq \eta$.

Proposition 5 implies a negative relationship between the variability of noise trader demand—or, equivalently, of expected returns—within a segment and the segment's own expected return. The negative relationship arises because the price sensitivity $S_n'(D_{nt})$ to the dividend flow D_{nt} decreases in θ_n (proposition 1). When θ_n is large, volatility per share is low because price sensitivity is low. Hence, an increase in the number of shares θ_n causes a small price drop. When instead θ_n is small, volatility per share is high, and hence an equal decrease in θ_n causes a large price rise. When we average across the two cases, a segment with more extreme values of θ_n trades at a higher price than a segment with less extreme values.

Proposition 5 shows additionally that the negative relationship between within-segment variability of noise trader demand and segment expected return strengthens when the constraint tightens (from $\hat{L} = \infty$ to $\hat{L} = 0$). This is because the constraint prevents constrained investors from absorbing noise trader demand, increasing the demand's effective variability.

Figure 4 illustrates overvaluation bias in the calibrated example. We group the 10 assets into the five segments (0.6, 1.4), (0.7, 1.3), (0.8, 1.2),

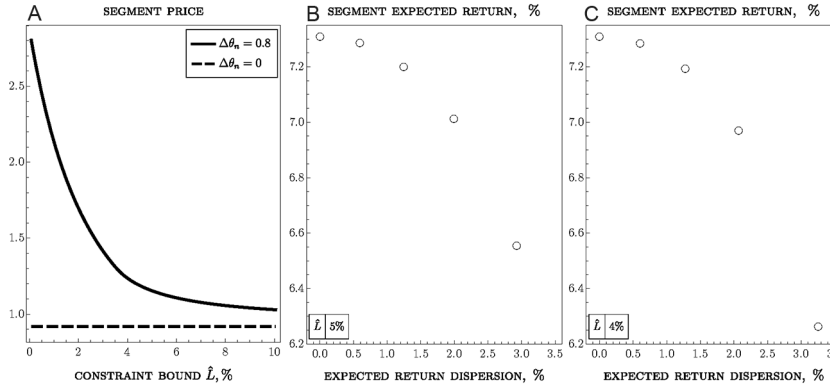

FIG. 4.—Overvaluation bias.

(0.9, 1.1), and (1, 1). Figure 4*A* plots the unconditional averages of the prices of the two extreme segments as functions of $\hat{L}$: the segment with $(\theta_n, \theta_{n'}) = (0.6, 1.4)$, represented by the thick line, and the segment with $(\theta_{\hat{n}}, \theta_{\hat{n}'}) = (1, 1)$, represented by the thin line. Consistent with proposition 5, the former segment trades at a higher price, and the price difference increases when $\hat{L}$ decreases. Since the price of the latter segment does not depend on $\hat{L}$ (the constraint does not bind for assets $(\hat{n}, \hat{n}')$ because $\theta_{\hat{n}} = \theta_{\hat{n}'} = \eta = 1$), the price of the former segment increases when $\hat{L}$ decreases. This reflects the asymmetry shown in figures 2 and 3: the constraint raises the price of the asset with $\theta_n = 0.6$ more than it lowers the price of the asset with $\theta_{n'} = 1.4$.

Figure 4*B* and 4*C* plot expected return at the segment level as a function of the dispersion in expected returns within the segment. Figure 4*B* is drawn for $\hat{L} = 5$ and figure 4*C* for $\hat{L} = 4$. Both show a negative relationship between within-segment dispersion in expected returns and segment expected return. The slope of the relationship is steeper (more negative) when $\hat{L} = 4$, consistent with the comparison that proposition 5 derives between $\hat{L} = \infty$ and $\hat{L} = 0$.

V. Equilibrium with Endogenous Constraint

A. Information and Contracts

In this section, we endogenize the parameters of the constraint within a static contracting model, which we next embed into our dynamic equilibrium model. We sketch the contracting model briefly here and develop it more fully in appendix A. An investor can invest in multiple independent risky assets through a fund run by a manager. Contracting between the investor and the manager takes place in period 0, information is observed

and assets are traded in period 1, and assets pay off in period 2. The manager is either skilled or unskilled. A skilled manager observes an informative signal about the payoff distribution of each asset. An unskilled manager observes an uninformative signal that she wrongly treats as informative. The probability that the manager is unskilled is $\lambda \in [0, 1)$. The uninformative signal makes the unskilled manager excessively optimistic or excessively pessimistic, with equal probabilities. The investor allocates wealth W_z to the fund in period 0. The manager can invest W_z in the risky assets and possibly also in the riskless asset.

The contract between the investor and the manager consists of a fee and an investment restriction. The fee can be any nonnegative and increasing function $f(W_{z2})$ of the investor's wealth W_{z2} held by the fund in period 2. The investment restriction requires that the distance between the fund's portfolio weight in each risky asset and the asset's weight in a benchmark index lies in a closed set $\mathcal{L}$. The investor chooses his wealth W_z allocated to the fund and the contract parameters $(f(W_{z2}), \mathcal{L})$ to maximize his expected utility. He is subject to the manager's incentive compatibility constraint, whereby the manager chooses positions in the risky assets to maximize her expected utility derived from the fee.

Proposition A1 characterizes the solution to the investor's optimization problem. The optimal set $\mathcal{L}$ has the form $[0, \hat{L}]$, with $\hat{L} > 0$. Hence, the investor allows portfolio weights in the fund to differ from index weights as long as the distance does not exceed a positive bound $\hat{L}$. The position chosen by the skilled manager in each asset is the optimal position given the investor's risk preferences if the resulting distance in portfolio weights is smaller than $\hat{L}$. Otherwise, the position is the maximum or the minimum allowed by the constraint. The unskilled manager chooses the maximum or the minimum position.

Intuitively, the optimal fee $f(W_{z2})$ aligns the manager's risk preferences with the investor's. Absent the constraint, the skilled manager would choose the investor's optimal position for all realizations of her signal, but the unskilled manager would choose extreme positions. The investment restriction limits extreme positions. This is desirable when the extreme positions are chosen by the unskilled manager but undesirable when they are chosen by the skilled manager (observing an extreme informative signal). An increase in the probability λ that the manager is unskilled results in a smaller value for the optimal $\hat{L}$, that is, a tighter constraint. When λ goes to 1, the optimal $\hat{L}$ goes to 0; that is, the investor renders the constraint infinitely tight, replicating passive investing.

The contracting model endogenizes the constraint (9) only in a parametric sense; that is, it yields optimal values for the parameters $(W_{2z}, \hat{L})$. The general form of the constraint—as a function of the distance in portfolio weights asset by asset—remains exogenous. Nevertheless, the exercise has two advantages. First, by deriving the parameters $(W_{2z}, \hat{L})$ in the

constraint as a function of more primitive parameters, such as the probability λ that the manager is unskilled, we can map our asset pricing analysis to these primitives. Second, the exercise resolves the tension in section IV that a constrained investor optimizes over positions but constrains himself over that choice. Under the contracting model, the skilled manager chooses the position that is optimal for the investor because the optimal fee aligns her risk preferences with the investor's. Moreover, the constraint exists to guard against the unskilled manager.

We next embed the static contracting model into our dynamic equilibrium model. As in our calibrated example, we set the number of shares $\hat{\eta}_n$ of each risky asset n in the benchmark index to 1 (a normalization) and assume that assets differ in their supply θ_n but not in their other characteristics $(\kappa_n, \bar{D}_n, \sigma_n)$. We denote the latter characteristics by $(\kappa, \bar{D}, \sigma)$). We assume that unconstrained investors can invest in the risky assets directly and observe (θ_n, D_{nt}) for all n. By contrast, constrained investors do not observe (θ_n, D_{nt}) and can invest through a fund. As in the static contracting problem, they choose their wealth W_z allocated to the fund and the contract parameters $(f(W_{z2}), \mathcal{L})$ to maximize their expected utility. They compute expected utility using the cross-sectional distribution for θ_n and the unconditional time series distribution for D_{nt}. Since they do not observe (θ_n, D_{nt}), their optimal values for $(W_z, f(W_{z2}), \mathcal{L})$ do not depend on (θ_n, D_{nt}). Fund managers can be skilled or unskilled. Skilled managers observe (θ_n, D_{nt}) for all n. The parameters $(W_z, \hat{L})$ determine the parameters (η_n, L) in the constraint, as in section III. The benchmark position η_n is the same across assets, since $\hat{\eta}_n$ is, and we denote it by η.

Implicit in our formulation is that constrained investors do not contract dynamically and do not learn over time. These assumptions can be imposed as restrictions on infinitely lived investors or can follow more directly by interpreting investors as overlapping generations living over infinitesimal periods.

B. Equilibrium

Solving for equilibrium involves a fixed-point problem: asset prices must clear the market given the constraint, and the parameters (η, L) in the constraint must be optimal given equilibrium prices. The determination of equilibrium prices given the constraint is as in section IV. The only change is that constrained investors whose manager turns out to be unskilled do not invest optimally subject to the constraint. Half of them employ a manager who is excessively optimistic about asset n and invests the maximum value of z_{2nt} that meets the constraint. The remaining half employ a manager who is excessively pessimistic and invests the minimum value. Since the average of the maximum and the minimum value is η

and the measure of uninformed investors employing an unskilled manager is λx, the market-clearing condition (15) is replaced by

$$(1 - x)z_{1nt} + (1 - \lambda)xz_{2nt} + \lambda x\eta = \theta_n. \tag{28}$$

The definition of the unconstrained and the constrained regions is modified similarly. Since in the unconstrained region $z_{1nt} = z_{2nt}$, (28) implies $z_{2nt} = (\theta_n - \lambda x\eta)/(1 - \lambda x)$. Substituting into the constraint (22), we find that the unconstrained region is defined by

$$\frac{|\theta_n - \eta|}{1 - \lambda x} G_n(D_{nt}) \leq L, \tag{29}$$

which replaces (23). The ODE system is modified similarly, as shown in the proof of proposition 6. The investment restriction in the static contracting model yields the AS-based constraint with $G_n(D_{nt}) = S_n(D_{nt})$. Modifying that model so that the investment restriction concerns the standard deviation of the difference between the fund and the index return yields the TE-based constraint with $G_n(D_{nt}) = \sigma\sqrt{D_{nt}}S'_n(D_{nt})$. Under either constraint, the determination of the optimal values of (η, L) follows in proposition A1. Proposition 6 characterizes these values.

Proposition 6. The optimal values of (η, L) have the following properties:

- When θ_n is the same across assets, $\eta = \theta$ and $L = 0$, where θ is the common value of θ_n.
- When θ_n differs across assets, $\eta \in (\theta_{\min}, \theta_{\max})$, where $\theta_{\min}$ is the minimum and $\theta_{\max}$ is the maximum value of θ_n. Moreover, $L = \infty$ when $\lambda = 0$, $L \in (0, \infty)$ when $\lambda > 0$, and $\lim_{\lambda \to 1} L = 0$.

When asset supply θ_n takes the same value θ for all assets, holding an equal number of shares of each asset is optimal for an investor. Moreover, the optimal number of shares of each asset is θ: since investors form a mass 1 continuum, θ is the asset supply per investor. Constrained investors achieve the optimal outcome by requiring the fund to hold the index ($\hat{L} = 0$) and allocating to it wealth W_z such that the number of shares that it holds of each asset is $\eta = \theta$.

When θ_n differs across assets, holding an equal number of shares of each asset is not optimal. If all managers are skilled, then constrained investors achieve the optimal outcome by not restricting the fund ($\hat{L} = \infty$). If instead some managers are unskilled, then constrained investors impose a restriction ($\hat{L} \in (0, \infty)$). The wealth W_z they allocate to the fund is such that the benchmark position η is smaller than $\mathbb{E}(\theta_n)$. This choice of W_z reflects an optimal response of (rational) constrained investors to the asymmetric effects of noise trader demand. To explain the intuition,

suppose for simplicity that constrained investors always require the fund to hold the index ($\hat{L} = 0$), in which case η is the number of shares that the fund holds of each asset. Suppose also that θ_n takes the same value θ for all assets, in which case constrained investors set W_z such that $\eta = \theta$. Suppose next that noise traders buy some shares of one asset and sell an equal number of shares of another asset, so that $\mathbb{E}(\theta_n)$ remains equal to θ. Because of the asymmetry, the price of the latter asset rises more than the price of the former asset drops. Constrained investors respond to the aggregate overvaluation by changing W_z so that the fund holds a smaller number of shares η of each asset.

The optimal response of constrained investors tempers the asymmetry but does not eliminate it. This is shown in figure 5*A* and 5*B*, which are the counterparts of figure 3*A* and 3*B* for the endogenous constraint. The two panels plot the unconditional average of the price and of return volatility as functions of the fraction λ of unskilled managers. Parameter values are as in the calibrated example of section IV, except that (W_z, $\hat{L}$) are solutions to the contracting problem and functions of λ. Noise trader demand has larger effects on price and volatility when it is high ($\theta_n = 0.6$) than when it is low ($\theta_n = 1.4$). The asymmetry thus remains but is weaker than in figure 5. In particular, the maximum spread in return volatilities, which is mainly driven by the high-demand asset, drops to 1.82% from 4.49% in figure 3.

Figure 5*C* plots the optimal values of $\hat{L}$ (left scale) and η (right scale) as functions of λ. Consistent with proposition 6, the optimal $\hat{L}$ becomes large when λ goes to 0 and decreases to 0 when λ goes to 1. The optimal η is smaller than $\mathbb{E}(\theta) = 1$ and decreases in λ.

Figure 5*C* shows that for a fraction λ of unskilled managers up to 30%, the constraint is laxer than implied by the data. When λ = 10 %, the bound $\hat{L}$ in the constraint is 10.1, and the difference between the AS of

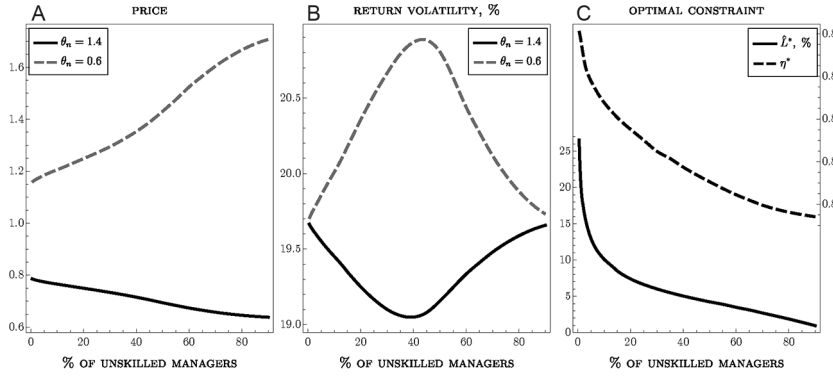

FIG. 5.—Endogenous constraint.

unconstrained and of constrained investors is 1.67—smaller than its empirical counterpart of 3.64 (or 9.00 when identifying unconstrained investors with the funds in the top AS quintile). When λ rises to 20%, $\hat{L}$ drops to 7.4, and the AS difference rises to 2.62. It takes λ to rise all the way to 40% for $\hat{L}$ to drop to 5, as in the calibrated example of section IV. For that value of λ, the AS difference becomes 4.12.

When the constraint is specified in TE rather than AS terms, its data-implied tightness is consistent with values of λ smaller than 30%. For example, when λ = 20%, the AS difference is 4.05.[11] Nevertheless, the fraction of unskilled managers that it takes to generate a constraint of the observed tightness remains significant. One interpretation of this result is that constraints incorporate not only an explicit bound that investors impose on managers but also a bound that managers impose on themselves to limit their reputational risk from underperforming the index.

C. *Effective Capital*

We finally use our model to compute effective capital. According to the market efficiency view, noise trader–induced distortions should be small because institutions such as mutual funds and pension funds can deploy large pools of capital to trade against them. According to the limits of arbitrage view, that capital can be ineffective because agency problems between the managers in these institutions and the investors who own the capital limit the managers' ability to take risk. Our model can inform the debate between the two views because it determines how agency-induced constraints on asset managers affect equilibrium asset prices. Suppose that a given amount of capital is invested with constrained asset managers. What is the equivalent amount of capital that—if managed without constraints—would result in the same price distortions?

To compute effective capital, we assume that a subset of constrained investors with measure $y < x$ can invest with skilled managers, to whom they (optimally) impose no constraints, and the remaining subset, with measure $x - y$, invests in the index. We determine the value of y such that price distortions are the same as when all constrained investors can invest with managers of unknown skill, to whom they impose constraints. We refer to y as effective capital and express it as a fraction of x, to which we refer as total capital. We measure price distortions by the average difference

[11] The AS- and TE-based specifications differ in the relationship between λ and the optimal $\hat{L}$ because of the dependence of conditional return volatility on the dividend flow. Since the volatility of dividend per share D_{nt} goes to zero when D_{nt} goes to zero, and the price S_{nt} does not go to zero, return volatility goes to zero. With low return volatility in the unconstrained region, the cost to investors of a large investment by unskilled managers is small, and investors can afford to raise $\hat{L}$. This effect is absent under the TE-based constraint, as that constraint is specified in volatility terms and thus becomes laxer when volatility drops.

TABLE 4
EFFECTIVE CAPITAL

	ΔAS = 4		ΔAS = 6	
	AS Based	TE Based	AS Based	TE Based
Fraction of unskilled managers λ (%)	38.4	19.6	55.8	37.0
Constraint bound $\hat{L}$ (%)	5.2	1.6	3.8	1.1
Effective to total capital y/x (%)	42.6	52.0	23.2	28.4

NOTE.—Effective to total capital y/x is shown as a function of the type and severity of the constraint.

between the price of assets in high noise trader demand and the price of assets in low demand.

We compute y/x within our calibrated example under both the AS-based and the TE-based constraint. Under either constraint, we consider two values of λ: one that generates a difference between the AS of unconstrained and of constrained investors of 4 and one that generates a difference of 6. The implied value of λ is sensitive to the specification of the constraint, but effective capital (shown in table 4) is less so. For an AS difference of 4, effective capital is around 50% of total capital. For an AS difference of 6, effective capital drops to around 25%.

VI. Conclusion

We argue that asset management should be viewed as a continuum between active and passive rather than as two polar extremes. Active managers are not required to hold benchmark indices. Yet they are often required to maintain their deviations from indices within bounds. These bounds differ significantly across funds and can be viewed as a characteristic of each fund.

We provide new empirical evidence supporting the continuum view and interpret findings in the literature in that light. We also explore theoretically the implications of the continuum view for equilibrium asset prices and market efficiency. We show that constrained asset managers buy underweighted assets procyclically, and this generates a positive association between overvaluation and high volatility. We also show that overvaluation is harder to correct than undervaluation, even in the absence of short sale costs. These mechanisms have attracted policy attention because of their links with asset bubbles.[12]

[12] For example, a 2003 report by the Committee on the Global Financial System (Bank for International Settlements 2003, 19) notes, "Overvalued assets/stocks tend to find their way into major indices, which are generally capitalization-weighted and therefore will more likely include overvalued securities than undervalued securities. Asset managers may therefore need to buy these assets even if they regard them as overvalued; otherwise they risk

Our research can be extended in a number of directions. One direction is to explore the optimal design of benchmark indices—which we show matter not only for passive funds but also for active funds.[13] For example, should asset managers be required to remain close to an index portfolio or to an average portfolio of other managers? A related and broader direction is to explore the optimal design of asset manager contracts and constraints. Linking the contracts to equilibrium asset prices, as we do in this paper, raises welfare questions as well. Would a social planner internalizing the links between contracts and prices employ the same contracts as private investors?[14]

An additional extension is to introduce dynamic contracts and reputational concerns. Our calibrated example suggests that constraints may partly reflect managers' concern to limit their reputational risk from underperforming their benchmark index. A number of papers show that reputational concerns of asset managers generate herding and a preference for negative skewness.[15] The links between reputational concerns and the asset pricing effects that we derive in this paper, such as risk-return inversion, could be explored.

Appendix A

Endogenous Constraint

There are three periods: 0, 1, and 2. In period 0, contracts are written. In period 1, information is observed and assets are traded. In period 2, assets pay off. There is one riskless asset and N risky assets. The riskless asset has return r. Risky asset $n = 1, \dots, N$ trades at price S_n per share and has return R_n in excess of the riskless asset.

An investor has wealth W. In period 0, he allocates $W - W_z$ to the riskless asset and $W_z \geq 0$ to a fund run by a manager. The manager can invest W_z in the risky assets and possibly also in the riskless asset. The investor has prior distribution Π_0 on $\{(S_n, R_n)\}_{n=1,\dots,N}$ in period 0. Under Π_0, the pair (S_n, R_n) is independent across assets. We denote expectations under Π_0 by $\mathbb{E}_0$.

---

violating agreed tracking errors. In a similar spirit, a 2015 International Monetary Fund working paper (Jones 2015) notes, "Another source of friction capable of amplifying bubbles stems from the 'captive buying' of securities in momentum-biased market capitalization-weighted financial benchmarks. Underlying constituents that rise most in price will see their benchmark weights increase irrespective of fundamentals, inducing additional purchases from fund managers seeking to minimize benchmark tracking error."

[13] Recent evidence that benchmarks matter for active funds is in Pavlova and Sikorkaya (2022), who examine active funds' trading around dates when benchmark composition changes.

[14] Kashyap et al. (2021b) compare privately and socially optimal contracts in a two-period model of asset management.

[15] See, e.g., Froot, Scharftstein, and Stein (1992), Dasgupta and Prat (2008), Dasgupta, Prat, and Verardo (2011), and Guerrieri and Kondor (2012).

The manager is either skilled or unskilled. A skilled manager observes informative signals $\{s_n\}_{n=1,\ldots,N}$ about asset returns $\{R_n\}_{n=1,\ldots,N}$. An unskilled manager observes uninformative signals about returns, but she wrongly treats them as informative.[16] Signals are independent across assets. Both the skilled and the unskilled manager observe prices $\{S_n\}_{n=1,\ldots,N}$. Signals and prices are observed in period 1. The posterior distribution that a skilled manager has on R_n after observing signal s_n is $\Pi_n(s_n)$. The posterior distribution Π_n that an unskilled manager has on R_n is either an optimistic one Π_n^o or a pessimistic one Π_n^p, with the two outcomes equally likely. The probability that the manager is unskilled is $\lambda \in [0, 1)$. The investor and the manager have negative exponential utility over consumption in period 2, with coefficient of absolute risk aversion equal to ρ for the investor and $\bar{\rho}$ for the manager.

The signal s_n that a skilled manager observes about R_n is continuous, and yields a posterior distribution $\Pi_n(s_n)$ that gives positive probability to positive and to negative values of R_n. As a consequence, the position $z_n^*(s_n)$ that maximizes the investor's expected utility conditional on s_n is finite. We take the range of $z_n^*(s_n)$ to be the real line. The unskilled manager gives positive probability only to positive values of R_n under the optimistic posterior distribution Π_n^o and only to negative values under the pessimistic distribution Π_n^p. Thus, the unskilled manager believes that each asset n either has no downside or has no upside.

If the investor allocates wealth $W_z > 0$ to the fund in period 0, then he offers the manager a contract. If the manager accepts the contract, then she observes prices and her private signals about returns in period 1 and chooses a portfolio for the investor. The portfolio consists of $\{z_n\}_{n=1,\ldots,N}$ shares in the risky assets and $W_z - \Sigma_{n=1}^N z_n S_n$ dollars in the riskless asset. The investor's wealth held by the fund in period 2 is $W_{z2} = W_z(1 + r) + \Sigma_{n=1}^N z_n S_n R_n$.

The contract consists of a fee, which depends on the investor's wealth W_{z2} held by the fund in period 2 and on an investment restriction. The fee can be a general function $f(W_{z2})$, subject to a nonnegativity and a monotonicity constraint. The nonnegativity constraint is $f(W_{z2}) \geq 0$ and arises because the manager has limited liability. The monotonicity constraint is that $f(W_{z2})$ is increasing and could arise from moral hazard in period 2. Indeed, a decreasing fee could incentivize the manager to engage in wasteful activities that reduce W_{z2} so to increase her fee. A nondecreasing fee could also incentivize such activities if they yield an infinitesimally small private benefit to the manager. An additional reason to assume an increasing fee is to rule out the implausible outcome that the investor can induce the manager to choose any positions $\{z_n\}_{n=1,\ldots,N}$ just by offering her a constant fee and exploiting her indifference. To ensure that an optimal fee exists, we formulate the monotonicity constraint as a weak rather than a strict inequality: $f'(W_{z2}) \geq \epsilon g'(W_{z2}) > 0$, where ϵ is a positive constant and $g(W_{z2})$ is an increasing and bounded function defined over $(-\infty, \infty)$. We derive the optimal fee for each ϵ and take the limit when ϵ goes to zero.

The investment restriction concerns the positions $\{z_n\}_{n=1,\ldots,N}$ chosen by the manager in period 1. We assume that the investor observes the distance

[16] See Vayanos (2018) for a related model in which the unskilled manager is rational, and takes extreme positions despite observing uninformative signals because she is less risk-averse than the skilled manager.

$|(z_n S_n/W_z) - (\hat{\eta}_n S_n/\Sigma_{n'=1}^N \hat{\eta}_{n'} S_{n'})|$ between the portfolio weight $z_n S_n/W_z$ of each risky asset n in the fund and the asset's weight $\hat{\eta}_n S_n/\Sigma_{n'=1}^N \hat{\eta}_{n'} S_{n'}$ in a benchmark index that includes $\hat{\eta}_n$ shares of asset n. The investor restricts that distance to lie in a closed set $\mathcal{L}$, same for all assets. To ensure that the investor allocates positive wealth in the fund, we assume that the index earns nonnegative unconditional expected return in excess of the riskless asset, that is, $\mathbb{E}_0(\Sigma_{n=1}^N \hat{\eta}_n S_n R_n/\Sigma_{n=1}^N \hat{\eta}_n S_n) \geq 0$.

The investor chooses in period 0 wealth $W_z \geq 0$ allocated to the fund as well as contract parameters $(f(W_{z2}), \mathcal{L})$ if $W_z > 0$ to maximize his expected utility. He is subject to the manager's incentive compatibility constraint, whereby the manager chooses positions in the risky assets to maximize her expected utility derived from the fee. He must also ensure that the fee satisfies nonnegativity and monotonicity. Nonnegativity ensures that the manager's individual rationality constraint is satisfied.

Since the constraint $|(z_n S_n/W_z) - (\hat{\eta}_n S_n/\Sigma_{n'=1}^N \hat{\eta}_{n'} S_{n'})| \in \mathcal{L}$ implies $z_n = 0$ when $W_z = 0$, we can nest the case $W_z = 0$ within the case $W_z > 0$ when solving the investor's maximization problem. That is, we can assume that even when the investor allocates wealth $W_z = 0$ to the fund, he offers the manager a contract. Under that contract, positions z_n are zero, and so is the manager's fee $f(W_{z2})$ (for the only possible value $W_{z2} = 0$).

Our contracting model is in the spirit of the literature on optimal delegation (e.g., Alonso and Matouschek 2008; Amador and Bagwell 2013). A key result in that literature is that instead of taking an action based on information sent by the agent, the principal can equivalently let the agent take the action within a restricted *delegation set*. The delegation literature generally precludes monetary transfers between the principal and the agent. We allow monetary transfers but—in the spirit of the delegation literature—restrict the fee function $f(W_{z2})$ to not depend on information sent by the agent. We also restrict the delegation set $\mathcal{L}$ to depend on only some statistics of the agent's action. That restriction could be arising from investors' limited ability to observe or process information.

Proposition A1 shows that in the limit when ϵ goes to zero, the investor allocates wealth $W_z > 0$ to the fund and chooses a delegation set $\mathcal{L}$ of the form $[0, \hat{L})$, with $\hat{L} > 0$. The proposition also characterizes the optimal values of $(W_z, \hat{L})$ and the optimal fee $f(W_{z2})$. We denote by $\bar{z}_n$ and $\underline{z}_n$ the maximum and minimum values, respectively, of z_n that meet the constraint $|(z_n S_n/W_z) - (\hat{\eta}_n S_n/\Sigma_{n'=1}^N \hat{\eta}_{n'} S_{n'})| \leq \hat{L}$.

Proposition A1. In the limit when ϵ goes to zero:

- The investor allocates wealth $W_z > 0$ to the fund.
- The optimal delegation set $\mathcal{L}$ has the form $[0, \hat{L}]$, with $\hat{L} > 0$.
- The position $z_{Gn}(s_n)$ in asset n chosen by the skilled manager is $\bar{z}_n$ when $z_n^*(s_n) > \bar{z}_n$, $\underline{z}_n$ when $z_n^*(s_n) < \underline{z}_n$, and $z_n^*(s_n)$ otherwise.
- The position z_{Bn} in asset n chosen by the unskilled manager is $\bar{z}_n$ when her posterior is Π_n^O and $\underline{z}_n$ when her posterior is Π_n^P.
- The optimal values of $(W_z, \hat{L})$ solve

$$\max_{(W_z, \hat{L}) \in [0,\infty)^2} \mathbb{E}_0\left[-(1-\lambda)e^{-\rho\left(W(1+r)+\sum_{n=1}^N z_{Gn}(s_n)S_n R_n\right)} - \lambda e^{-\rho\left(W(1+r)+\sum_{n=1}^N z_{Bn} S_n R_n\right)}\right]. \quad \text{(A1)}$$

- The optimal fee $f(W_{z2})$ converges to zero for all W_{z2}.

Appendix B

Proofs

B1. Proof of Proposition 1

Substituting the affine price function (17) into the ODE (16), we find

$$D_{nt} + \kappa_n(\bar{D}_n - D_{nt})a_{n1} - r(a_{n0} + a_{n1}D_{nt}) = \rho\theta_n\sigma_n^2 D_{nt}a_{n1}^2. \tag{B1}$$

Equation (B1) is affine in D_{nt}. Identifying the terms that are linear in D_{nt} yields the equation

$$\rho\theta_n\sigma_n^2 a_{n1}^2 + (r + \kappa_n)a_{n1} - 1 = 0. \tag{B2}$$

Equation (B2) is quadratic in a_{n1}. When $\theta_n > 0$, the left-hand side is increasing for positive values of a_{n1}, and (B2) has a unique positive solution, given by (19). When $\theta_n < 0$, the left-hand side is hump shaped for positive values of a_{n1}, and (B2) has two positive solutions, one positive solution, or no solution. Condition $\theta_n > -[(r + \kappa_n)^2/4\rho\sigma_n^2]$ ensures that two positive solutions exist when $\theta_n < 0$. Equation (19) gives the smaller of the two solutions, which is the continuous extension of the unique positive solution when $\theta_n > 0$. Identifying the constant terms yields the equation

$$\kappa_n\bar{D}_n a_{n1} - ra_{n0} = 0,$$

whose solution is (18).

To show that $S_n(D_{nt})$ and $S_n'(D_{nt})$ are decreasing and convex in θ_n, we note that a_{n1} takes the form

$$\Psi(\theta_n) \equiv \frac{1}{A + \sqrt{B + C\theta_n}}$$

for positive constants (A, B, C). The function $\Psi(\theta_n)$ is decreasing. It is also convex because its derivative

$$\Psi'(\theta_n) = -\frac{C}{2\sqrt{B + C\theta_n}}\frac{1}{\left(A + \sqrt{B + C\theta_n}\right)^2}$$

is increasing. Hence, a_{n1} is decreasing and convex in θ_n. These properties extend to a_{n0} from (18) and to $S'(D_{nt}) = a_{n1}$ and $S(D_{nt}) = a_{n0} + a_{n1}D_{nt}$. QED

B2. Proof of Corollary 1

Substituting the price from (17) into (13), we find that the share return of asset n is

$$\begin{aligned} dR_{nt}^{sh} &= [D_{nt} + \kappa_n(\bar{D}_n - D_{nt})a_{n1} - r(a_{n0} + a_{n1}D_{nt})]\,dt + \sigma_n\sqrt{D_{nt}}a_{n1}\,dB_{nt} \\ &= \rho\theta_n\sigma_n^2 D_{nt}a_{n1}^2\,dt + \sigma_n\sqrt{D_{nt}}a_{n1}\,dB_{nt}, \end{aligned} \tag{B3}$$

where the second step follows from (B1). Substituting the share return from (B3) and the price from (17) into (5), we find that the (dollar) return of asset

n is

$$\begin{aligned} dR_{nt} &= \frac{\rho\theta_n\sigma_n^2 D_{nt}a_{n1}^2\,dt + \sigma_n\sqrt{D_{nt}}a_{n1}\,dB_{nt}}{a_{n0} + a_{n1}D_{nt}} \\ &= \frac{\rho\theta_n\sigma_n^2 D_{nt}a_{n1}\,dt + \sigma_n\sqrt{D_{nt}}\,dB_{nt}}{(\kappa_n/r)\bar{D}_n + D_{nt}} \\ &= \frac{\left[(2\rho\theta_n\sigma_n^2 D_{nt}\,dt)/\left(r + \kappa_n + \sqrt{(r+\kappa_n)^2 + 4\rho\theta_n\sigma_n^2}\right)\right] + \sigma_n\sqrt{D_{nt}}\,dB_{nt}}{(\kappa_n/r)\bar{D}_n + D_{nt}}, \end{aligned} \tag{B4}$$

where the second step follows from (18) and the third step follows from (19).

The conditional expected return is the drift coefficient in (B14) times dt,

$$\mathbb{E}_t(dR_{nt}) = \frac{2\rho\theta_n\sigma_n^2 D_{nt}\,dt}{\left(r + \kappa_n + \sqrt{(r+\kappa_n)^2 + 4\rho\theta_n\sigma_n^2}\right)[(\kappa_n/r)\bar{D}_n + D_{nt}]}.$$

It takes the form $\Phi(\theta_n)\{(2\rho\sigma_n^2 D_{nt}\,dt)/[(\kappa_n/r)\bar{D}_n + D_{nt}]\}$, where

$$\Phi(\theta_n) \equiv \frac{\theta_n}{A + \sqrt{B + C\theta_n}}$$

for positive constants (A, B, C). The function $\Phi(\theta_n)$ is increasing, and hence the conditional expected return is increasing in θ_n. (The derivative of $\Phi(\theta_n)$ has the same sign as

$$A + \sqrt{B + C\theta_n} - \frac{C}{2\sqrt{B + C\theta_n}}\theta_n = A + \frac{1}{\sqrt{B + C\theta_n}}\left(B + \frac{C\theta_n}{2}\right).$$

This expression is positive for $B + C\theta_n > 0$, a condition that is required for the term in the square root to be positive.) The unconditional expected return is the unconditional expectation of the conditional expected return,

$$\mathbb{E}(dR_{nt}) = \mathbb{E}(\mathbb{E}_t(dR_{nt})),$$

because of the law of iterative expectations. Since $\mathbb{E}_t(dR_{nt})$ is increasing in θ_n for any given D_{nt}, $\mathbb{E}(dR_{nt})$ is increasing in θ_n.

The return's conditional volatility is the diffusion coefficient in (B4) times $\sqrt{dt}$,

$$\sqrt{\mathbb{V}\mathrm{ar}_t(dR_{nt})} = \frac{\sigma\sqrt{D_{nt}\,dt}}{(\kappa_n/r)\bar{D}_n + D_{nt}}. \tag{B5}$$

It is independent of θ_n. The return's unconditional variance is the unconditional expectation of the return's conditional variance,

$$\mathbb{V}\mathrm{ar}(dR_{nt}) = \mathbb{E}(\mathbb{V}\mathrm{ar}_t(dR_{nt})). \tag{B6}$$

Since $\mathbb{V}\mathrm{ar}_t(dR_{nt})$ is independent of θ for any given D_{nt}, $\mathbb{V}\mathrm{ar}(dR_{nt})$ is independent of θ, and so is the return's unconditional volatility $\sqrt{\mathbb{V}\mathrm{ar}(dR_{nt})}$. Equation (B6) is implied by the law of total variance

$$\mathbb{V}\text{ar}(dR_{nt}) = \mathbb{E}(\mathbb{V}\text{ar}_t(dR_{nt})) + \mathbb{V}\text{ar}(\mathbb{E}_t(dR_{nt})) \tag{B7}$$

and because in continuous time the second term in the right-hand side of (B7) is negligible relative to the first: the second term is of order dt^2 while the first is of order dt. QED

B3. Proof of Proposition 2

Since the ODE (20) is identical to (16) except that θ_n is replaced by $(\theta_n - x\eta_n)/(1 - x)$, (21) can be derived from (19) with the same substitution. The comparisons with the case $\hat{L} = \infty$ follow because the function $\Psi(\theta_n)$ defined in the proof of proposition 1 is decreasing. Since $(\theta_n - x\eta_n)/(1 - x) > \theta_n$ when $\theta_n > \eta_n$, (19) and (21) imply that a_{n1} is smaller in the case $\hat{L} = 0$ than in the case $\hat{L} = \infty$. Conversely, since $(\theta_n - x\eta_n)/(1 - x) < \theta_n$ when $\theta_n < \eta_n$, (19) and (21) imply that a_{n1} is larger in the case $\hat{L} = 0$ than in the case $\hat{L} = \infty$. These comparisons of a_{n1} extend to a_{n0}, $S'(D_{nt}) = a_{n1}$, and $S_n(D_{nt}) = a_{n0} + a_{n1}D_{nt}$. QED

B4. Proof of Corollary 2

The price in the case $\hat{L} = 0$ can be derived from the price in the case $\hat{L} = \infty$ by replacing θ_n by $(\theta_n - x\eta_n)/(1 - x)$. Since the conditional and unconditional volatility in the case $\hat{L} = \infty$ are independent of θ_n (corollary 2), they are also independent of θ_n in the case $\hat{L} = 0$, and they are equal across the two cases. QED

B5. Proof of Theorem 1

The proof of theorem 1 is in appendix E. The existence part of the proof is along similar lines as in Kondor and Vayanos (2019). We start with a compact interval $[\epsilon, M] \subset (0, \infty)$ and show that there exists a unique solution to the ODEs with one boundary condition at ϵ and one at M. The boundary conditions are derived from the limits of $S'(D_t)$ at zero and infinity. In the case of M, for example, the requirement that $S'(D_t)$ has a finite limit at infinity determines that limit uniquely, and we set $S'(M)$ equal to that value. To construct the solution over $[\epsilon, M]$, we use $S'(M)$ and an arbitrary value for $S''(M)$ as initial conditions for the ODEs at M and show that there exists a unique $S''(M)$ such that the boundary condition at ϵ is satisfied. Showing uniqueness uses continuity of solutions with respect to the initial conditions as well as a monotonicity property with respect to the initial conditions that follows from the structure of the ODEs. We next show that when ϵ converges to zero and M to infinity, the solution over $[\epsilon, M]$ converges to a solution over $(0, \infty)$. The monotonicity property of solutions with respect to the initial conditions is key to the convergence proof because it yields monotonicity of the solution with respect to ϵ and M. QED

B6. Proof of Proposition 3

Substituting asset n's share return from (13) into (5) and setting $S_{nt} = S_n(D_{nt})$, we find that the asset's dollar return is

$$dR_{nt} = \frac{[D_{nt} + \kappa_n(\bar{D}_n - D_{nt})S_n'(D_{nt}) + (1/2)\sigma_n^2 D_{nt} S_n''(D_{nt})]dt + \sigma_n\sqrt{D_{nt}}S_n'(D_{nt})dB_{nt}}{S_n(D_{nt})} - rdt. \quad \text{(B8)}$$

The conditional volatility of asset n's return is the diffusion coefficient in (B8) times $\sqrt{dt}$:

$$\sqrt{\mathbb{V}\text{ar}_t(dR_{nt})} = \frac{\sigma_n\sqrt{D_{nt}}S_n'(D_{nt})\sqrt{dt}}{S_n(D_{nt})}. \quad \text{(B9)}$$

The conditional volatility under the affine solutions derived for $\hat{L} = 0$ and $\hat{L} = \infty$ is given by (B5). Comparing (B5) and (B9), we find that the conditional volatility of asset n's return is higher than under the affine solutions if

$$Z_n(D_{nt}) \equiv S_n'(D_{nt})(\kappa_n\bar{D}_n + rD_{nt}) - rS_n(D_{nt}) > 0.$$

Likewise, the conditional volatility of asset n''s return is lower than under the affine solutions if

$$Z_{n'}(D_{n't}) \equiv S_{n'}'(D_{n't})(\kappa_{n'}\bar{D}_{n'} + rD_{n't}) - rS_{n'}(D_{n't}) < 0.$$

Since $S_n'(D_t)$ converges to a finite limit when D_{nt} goes to zero, $D_{nt}S_n''(D_{nt})$ converges to zero. Since, in addition, $G_n(D_{nt}) > L/(|\theta_n - \eta_n|) > 0$ in the constrained region, (16) and (25) imply $Z_n(0) = 0$. Convexity of $S_n(D_{nt})$ and $Z_n(0) = 0$ imply $Z_n(D_{nt}) > 0$, and hence the conditional volatility of asset n's return is higher than under the affine solutions. Likewise, concavity of $S_{n'}(D_{n't})$ and $Z_{n'}(0) = 0$ imply that the conditional volatility of asset n''s return is lower than under the affine solutions. The comparison of conditional volatility across assets n and n' follows from the comparison of each case with the affine solutions since volatility under the affine solutions is the same for the two assets.

Since the return's unconditional variance is the unconditional expectation of the return's conditional variance, the comparisons derived for conditional volatility carry over to unconditional volatility. QED

B7. Proof of Proposition 4

The conditional beta of asset n is

$$\beta_{nt} = \frac{\mathbb{C}\text{ov}_t(dR_{nt}, dR_{Mt})}{\mathbb{V}\text{ar}_t(dR_{Mt})}, \quad \text{(B10)}$$

where dR_{nt} denotes the return of asset n and dR_{Mt} denotes the return of the market portfolio. Assuming that the market portfolio includes η_m shares of asset $m = 1, \ldots, N$, its return is

$$dR_{Mt} = \frac{dR_{Mt}^{sh}}{S_{Mt}} = \frac{\sum_{m=1}^N \eta_m dR_{mt}^{sh}}{\sum_{m=1}^N \eta_m S_{mt}} = \sum_{m=1}^N \frac{\eta_m S_{mt}}{\sum_{m=1}^N \eta_m S_{mt}} dR_{mt} = \sum_{m=1}^N \omega_{mt} dR_{mt}, \quad \text{(B11)}$$

where S_{Mt} denotes the market portfolio's price and

$$\omega_{mt} \equiv \frac{\eta_m S_{mt}}{\sum_{m'=1}^N \eta_{m'} S_{m't}}$$

denotes asset m's weight in the market portfolio. Equation (B10) implies that the conditional beta of asset n exceeds that of asset n' if

$$\begin{aligned} &\mathrm{Cov}_t(dR_{nt}, dR_{Mt}) > \mathrm{Cov}_t(dR_{n't}, dR_{Mt}) \\ &\Leftrightarrow \omega_n \mathbb{V}\mathrm{ar}_t(dR_{nt}) > \omega_{n'} \mathbb{V}\mathrm{ar}_t(dR_{n't}) \\ &\Leftrightarrow \eta_n S_{nt} \mathbb{V}\mathrm{ar}_t(dR_{nt}) > \eta_{n'} S_{n't} \mathbb{V}\mathrm{ar}_t(dR_{n't}), \end{aligned} \tag{B12}$$

where the second step follows from (B11) and the independence of returns across assets.

Suppose next that $\theta_n < \eta_n = \eta_{n'} < \theta_{n'}$ and that other characteristics of assets n and n' are identical ($\kappa_n = \kappa_{n'}, \bar{D}_n = \bar{D}_{n'}, \sigma_n = \sigma_{n'}$, and $D_{nt} = D_{n't}$). Since a_{1n} decreases in θ_n (proposition 1), the affine solution derived for $\hat{L} = \infty$ is larger for θ_n than for θ'_n. Since, in addition, S_{nt} lies above the affine solution for θ_n while $S_{n't}$ lies below the affine solution for θ'_n, $S_{nt} > S_{n't}$. Since, finally, $\mathbb{V}\mathrm{ar}_t(dR_{nt}) > \mathbb{V}\mathrm{ar}_t(dR_{n't})$ (proposition 3), (B12) implies $\mathrm{Cov}_t(dR_{nt}, dR_{Mt}) > \mathrm{Cov}_t(dR_{n't}, dR_{Mt})$ and hence $\beta_{nt} > \beta_{n't}$.

The unconditional beta of asset n is

$$\beta_{nt} = \frac{\mathrm{Cov}(dR_{nt}, dR_{Mt})}{\mathbb{V}\mathrm{ar}(dR_{Mt})} = \frac{\mathbb{E}(\mathrm{Cov}_t(dR_{nt}, dR_{Mt}))}{\mathbb{E}(\mathbb{V}\mathrm{ar}_t(dR_{Mt}))}.$$

Since the conditional covariance of $\mathrm{Cov}_t(dR_{nt}, dR_{Mt})$ is larger for asset n than for asset n', the same is true for the unconditional covariance and hence for the unconditional beta. QED

B8. Proof of Proposition 5

Since (18) implies $S(D_t) = a_1[(\kappa/r)\bar{D} + D_t]$, (27) is equivalent to

$$a_{1n} + a_{1n'} - (a_{1\hat{n}} + a_{1\hat{n}'}) > 0. \tag{B13}$$

When $\hat{L} = \infty$, proposition 1 implies that (B13) is equivalent to

$$\Psi(\theta_n) + \Psi(\theta_{n'}) - [\Psi(\theta_{\hat{n}}) + \Psi(\theta_{\hat{n}'})] > 0, \tag{B14}$$

where the function $\Psi(\theta)$ is defined in the proof of proposition 1. Setting $\ell \equiv \bar{\theta} - \theta_n = \theta_{n'} - \bar{\theta} > 0$ and $\hat{\ell} \equiv \bar{\theta} - \theta_{\hat{n}} = \theta_{\hat{n}'} - \bar{\theta} \in (0, \ell)$, we can write (B14) as

$$\begin{aligned} &\Psi(\bar{\theta} - \ell) + \Psi(\bar{\theta} + \ell) - \left[\Psi(\bar{\theta} - \hat{\ell}) + \Psi(\bar{\theta} + \hat{\ell})\right] > 0 \\ &\int_{\hat{\ell}}^{\ell} \Psi'(\bar{\theta} + x)dx - \int_{\hat{\ell}}^{\ell} \Psi'(\bar{\theta} - x)dx > 0 \\ &\Leftrightarrow \int_{\hat{\ell}}^{\ell} \left(\int_{-x}^{x} \Psi''(\bar{\theta} + y)dy \right) dx > 0. \end{aligned} \tag{B15}$$

Equation (B15) holds because $\Psi(\theta)$ is convex. When $\hat{L} = 0$, proposition 2 implies that (B13) is equivalent to (B14), with the function $\Psi((\theta - x\eta)(1 - x))$ instead of $\Psi(\theta)$. Since $\Psi((\theta - x\eta)(1 - x))$ is convex, the modified (B14) holds.

Propositions 1 and 2 imply that the comparison between $\hat{L} = \infty$ and $\hat{L} = 0$ in the corollary is equivalent to

$$\begin{aligned}
&\Psi\left(\frac{\theta_n - x\eta}{1-x}\right) + \Psi\left(\frac{\theta_{n'} - x\eta}{1-x}\right) - \left[\Psi\left(\frac{\theta_{\widehat{n}} - x\eta}{1-x}\right) + \Psi\left(\frac{\theta_{\widehat{n}} - x\eta}{1-x}\right)\right] \\
&> \Psi(\theta_n) + \Psi(\theta_{n'}) - [\Psi(\theta_{\widehat{n}}) + \Psi(\theta_{\widehat{n}})] \\
&\Leftrightarrow \Psi\left(\frac{\overline{\theta} - x\eta}{1-x} - \frac{\ell}{1-x}\right) + \Psi\left(\frac{\overline{\theta} - x\eta}{1-x} + \frac{\ell}{1-x}\right) \\
&\qquad - \left[\Psi\left(\frac{\overline{\theta} - x\eta}{1-x} - \frac{\widehat{\ell}}{1-x}\right) + \Psi\left(\frac{\overline{\theta} - x\eta}{1-x} + \frac{\widehat{\ell}}{1-x}\right)\right] \\
&> \Psi(\overline{\theta} - \ell) + \Psi(\overline{\theta} + \ell) - \left[\Psi(\overline{\theta} - \widehat{\ell}) + \Psi(\overline{\theta} + \widehat{\ell})\right] \\
&\Leftrightarrow \int_{\widehat{\ell}/(1-x)}^{\ell/(1-x)} \left(\int_{-x}^{x} \Psi''\left(\frac{\overline{\theta} - x\eta}{1-x} + y\right) dy\right) dx > \int_{\widehat{\ell}}^{\ell} \left(\int_{-x}^{x} \Psi''(\overline{\theta} + y)\, dy\right) dx.
\end{aligned} \tag{B16}$$

Since $\Psi(\theta)$ is convex and $x \in [0, 1)$,

$$\begin{aligned}
\int_{\ell/(1-x)}^{\ell/(1-x)} \left(\int_{-x}^{x} \Psi''\left(\frac{\overline{\theta} - x\eta}{1-x} + y\right) dy\right) dx &> \int_{\ell/(1-x)}^{[\hat{\ell}/(1-x)]+\ell-\hat{\ell}} \left(\int_{-x}^{x} \Psi''\left(\frac{\overline{\theta} - x\eta}{1-x} + y\right) dy\right) dx \\
&> \int_{\hat{\ell}}^{\ell} \left(\int_{-x}^{x} \Psi''\left(\frac{\overline{\theta} - x\eta}{1-x} + y\right) dy\right) dx.
\end{aligned}$$

Since, in addition,

$$\Psi''(\theta) = \frac{C^2}{4(B + C\theta)^{3/2}} \frac{1}{\left(A + \sqrt{B + C\theta}\right)^2} + \frac{C^2}{2(B + C\theta)} \frac{1}{\left(A + \sqrt{B + C\theta}\right)^3}$$

is decreasing, (B16) holds under the sufficient condition $(\overline{\theta} - x\eta)/(1 - x) \leq \bar{\theta}$, which is equivalent to $\bar{\theta} \leq \eta$. QED

B9. Proof of Proposition A1

We proceed in three steps. In the first step, we show that for any $\epsilon > 0$, the investor's expected utility under any contract does not exceed the utility (A1). This is the utility that the investor achieves when $(W_z, \mathcal{L})$ are as in the proposition and the fee $f(W_{z2})$ is zero. In the first step, we also show that the difference in utilities is bounded away from zero when the parameters $(W_z, \mathcal{L})$ in the contract or the manager's positions are not as in the proposition. In the second step, we show that there exists a contract with parameters $(W_z, \mathcal{L})$ and manager's positions as in the proposition, under which the investor's expected utility converges to the utility (A1) when ϵ goes to zero. In the third step, we show that the maximum in (A1) is achieved for $W_z > 0$ and $\hat{L} > 0$.

The first and second steps imply that in a contract maximizing the investor's expected utility when ϵ goes to zero, the parameters $(W_z, \mathcal{L})$ and the manager's positions are as in the proposition. Indeed, if a utility-maximizing contract involved different parameters $(W_z, \mathcal{L})$ or manager's positions, then it would yield a utility bounded away from (A1), while the contract involving these parameters

and positions yields (A1) when ϵ goes to zero. Adding the third step implies that the investor employs the manager. Indeed, not employing her is equivalent to setting $W_z = 0$, but this generates a utility bounded away from (A1) because $W_z = 0$ is not as in the proposition.

B9.1. Step 1

Since the fee $f(W_{z2})$ is increasing and Π_n^O gives positive probability only to positive values of R_n, an unskilled manager with posterior distribution Π_n^O on R_n chooses the maximum value of z_n that meets the constraint $|(z_n S_n / W_z) - (\hat{\eta}_n S_n / \Sigma_{n'=1}^N \hat{\eta}_{n'} S_{n'})| \in \mathcal{L}$. Conversely, since $f(W_{z2})$ is increasing and Π_n^P gives positive probability only to negative values of R_n, an unskilled manager with posterior distribution Π_n^P on R_n chooses the minimum value of z_n such that $|(z_n S_n / W_z) - (\hat{\eta}_n S_n / \Sigma_{n'=1}^N \hat{\eta}_{n'} S_{n'})| \in \mathcal{L}$. We denote these maximum and minimum values by $\bar{z}_n$ and $\underline{z}_n$, respectively, using the same notation as in the case $\mathcal{L} = [0, \hat{L}]$. We denote the resulting position in asset n chosen by an unskilled manager by z_{Bn}. We denote the position in asset n chosen by a skilled manager by $z_n(s_n)$.

Since $f(W_{z2})$ is nonnegative, the investor's expected utility is smaller than the utility achieved when the manager's positions remain the same and the fee is zero. The latter utility is

$$\begin{aligned}
&\mathbb{E}_0\left[-(1-\lambda)e^{-\rho\left(W(1+r)+\sum_{n=1}^N z_n(s_n)S_nR_n\right)} - \lambda e^{-\rho\left(W(1+r)+\sum_{n=1}^N z_{Bn}S_nR_n\right)}\right] \\
&\quad = -(1-\lambda)e^{-\rho W(1+r)}\prod_{n=1}^N \mathbb{E}_0\left[e^{-\rho z_n(s_n)S_nR_n}\right] - \lambda e^{-\rho W(1+r)}\mathbb{E}_0\left[e^{-\rho\sum_{n=1}^N z_{Bn}S_nR_n}\right],
\end{aligned} \tag{B17}$$

where the second step follows from independence across assets. Since the conditional expected utility $-\mathbb{E}_{s_n}\left[e^{-\rho z_n S_n R_n}\right]$ is concave in z_n, it is increasing for $z_n < z_n^*(s_n)$ and decreasing for $z_n > z_n^*(s_n)$. Therefore, when $z_n^*(s_n) > \bar{z}_n$, conditional expected utility for $z_n(s_n)$ is smaller than for $\bar{z}_n > z_n(s_n)$. Conversely, when $z_n^*(s) < \underline{z}_n$, conditional expected utility for $z_n(s_n)$ is smaller than for $\underline{z}_n < z_n(s_n)$. Since, in addition, conditional expected utility is maximum for $z_n^*(s_n)$, the law of iterative expectations implies

$$-\mathbb{E}_0\left[e^{-\rho z_n(s_n)S_nR_n}\right] \le -\mathbb{E}_0\left[e^{-\rho z_{Gn}(s_n)S_nR_n}\right], \tag{B18}$$

where we denote by $z_{Gn}(s_n)$ the position that is equal to $\bar{z}_n$ when $z_n^*(s_n) > \bar{z}_n$, $\underline{z}_n$ when $z_n^*(s_n) < \underline{z}_n$, and $z_n^*(s_n)$ when $z_n^*(s_n) \in [\underline{z}_n, \bar{z}_n]$, using the same notation as in the case $\mathcal{L} = [0, \hat{L}]$. Equation (B18) implies that (B17) does not exceed

$$\begin{aligned}
&-(1-\lambda)e^{-\rho W(1+r)}\prod_{n=1}^N \mathbb{E}_0\left[e^{-\rho z_{Gn}(s_n)S_nR_n}\right] - \lambda e^{-\rho W(1+r)}\mathbb{E}_0\left[e^{-\rho\sum_{n=1}^N z_{Bn}S_nR_n}\right] \\
&\quad = \mathbb{E}_0\left[-(1-\lambda)e^{-\rho\left(W(1+r)+\sum_{n=1}^N z_{Gn}(s_n)S_nR_n\right)} - \lambda e^{-\rho\left(W(1+r)+\sum_{n=1}^N z_{Bn}S_nR_n\right)}\right].
\end{aligned} \tag{B19}$$

Equation (B19) describes also the expected utility when the set $\mathcal{L}$ is replaced by $[0, L]$ with $L \equiv \sup \mathcal{L}$, since $(\bar{z}_n, \underline{z}_n)$ are the same for both sets. Since replacing $\mathcal{L}$

by $[0, L]$ yields the term in square brackets in (A1), and since (A1) is the maximum of that term over $(W_z, \hat{L})$, it exceeds the utility under any contract.

For $(W_z, \hat{L})$ not maximizing the term in square brackets in (A1), (B19) is smaller than (A1). For $\mathcal{L}$ differing from $[0, \hat{L}]$ by a positive measure set, (B17) is smaller than (B19) and hence also than (A1). Indeed, since the range of $z_n^*(s_n)$ is the real line, $z_n(s_n)$ differs from $z_n^*(s_n)$ in a positive measure set. For $z_n(s_n)$ differing from the values in (B19) in a positive measure set (while meeting the constraint $|(z_n S_n / W_z) - (\hat{\eta}_n S_n / \Sigma_{n'=1}^N \hat{\eta}_{n'} S_{n'})| \in \mathcal{L}$), (B17) is smaller than (B19) and hence also than (A1). Since (A1), (B17), and (B19) are independent of ϵ, the difference between (A1) and the utility under a contract in which the parameters $(W_z, \mathcal{L})$ or the manager's positions are not as in the proposition is bounded away from zero.

B9.2. Step 2

Suppose that $(W_z, \mathcal{L})$ are as in the proposition and

$$f(W_{z2}) = \epsilon g(W_{z2}) + \epsilon^{1/2}\left(\epsilon^{-1/8} - e^{-\rho W_{z2}}\right) 1_{\{W_{z2} > (1/8\rho)\log(\epsilon)\}}. \tag{B20}$$

(The term $1_{\{W_{z2} > \underline{W}\}}$ is the indicator function equal to 1 if $W_{z2} > \underline{W}$ and 0 otherwise.) Since the function $\epsilon^{-1/8} - e^{-\rho W_{z2}}$ is positive and increasing for $W_{z2} > (1/8\rho)\log(\epsilon)$, the fee $f(W_{z2})$ satisfies the nonnegativity and monotonicity constraints. Since the function $g(W_{z2})$ is bounded over $(-\infty, \infty)$ and the function $1 - \epsilon^{1/8} e^{-\rho W_{z2}}$ is bounded over $W_{z2} > (1/8\rho)\log(\epsilon)$, $f(W_{z2})$ converges uniformly to zero when ϵ goes to zero.

Equation (B20) implies that the manager's utility is

$$-e^{-\bar{\rho} f(W_{z2})} = -1 + \bar{\rho}\epsilon^{1/2}\left(\epsilon^{-1/8} - e^{-\rho W_{z2}}\right) 1_{\{W_{z2} > (1/8\rho)\log(\epsilon)\}} + \epsilon^{3/4} k(W_{z2}), \tag{B21}$$

where the function $k(W_{z2})$ is uniformly bounded when ϵ goes to zero. Since the dominant term in (B21) in the interval $W_{z2} > (1/4\rho)\log(\epsilon)$ is an affine transformation of the investor's utility, the position that maximizes the skilled manager's expected utility when ϵ goes to zero converges to the position that maximizes the investor's expected utility. Hence, when ϵ goes to zero, the investor's expected utility is given by (A1).

B9.3. Step 3

Using the definitions of $(z_{Gn}(s_n), z_{Bn})$, we find that the derivative of (A1) with respect to $y \in \{W_z, \hat{L}\}$ is

$$\begin{aligned}\rho \mathbb{E}_0 \Bigg[& (1-\lambda) e^{-\rho\left(W(1+r) + \sum_{n=1}^N z_{Gn}(s_n) S_n R_n\right)} \sum_{n=1}^N \left(\frac{\partial \bar{z}_n}{\partial y} 1_{\{z_n^*(s_n) > \bar{z}_n\}} + \frac{\partial \underline{z}_n}{\partial y} 1_{\{z_n^*(s_n) < \underline{z}_n\}} \right) S_n R_n \\ & + \lambda e^{-\rho\left(W(1+r) + \sum_{n=1}^N z_{Bn} S_n R_n\right)} \sum_{n=1}^N \left(\frac{\partial \bar{z}_n}{\partial y} 1_{\{\Pi_n = \Pi_n^o\}} + \frac{\partial \underline{z}_n}{\partial y} 1_{\{\Pi_n = \Pi_n^p\}} \right) S_n R_n \Bigg].\end{aligned} \tag{B22}$$

The definitions of $(\bar{z}_n, \underline{z}_n)$ imply

$$\bar{z}_n = \frac{W_z\hat{\eta}_n}{\sum_{n'=1}^{N}\hat{\eta}_{n'}S_{n'}} + \frac{W_z\hat{L}}{S_n}, \tag{B23}$$

$$\underline{z}_n = \frac{W_z\hat{\eta}_n}{\sum_{n'=1}^{N}\hat{\eta}_{n'}S_{n'}} - \frac{W_z\hat{L}}{S_n}. \tag{B24}$$

Differentiating (B23) and (B24) with respect to $(W_z, \hat{L})$, we find

$$\frac{\partial\bar{z}_n}{\partial W_z} = \frac{\hat{\eta}_n}{\sum_{n'=1}^{N}\hat{\eta}_{n'}S_{n'}} + \frac{\hat{L}}{S_n}, \tag{B25}$$

$$\frac{\partial\underline{z}_n}{\partial W_z} = \frac{\hat{\eta}_n}{\sum_{n'=1}^{N}\hat{\eta}_{n'}S_{n'}} - \frac{\hat{L}}{S_n}, \tag{B26}$$

$$\frac{\partial\bar{z}_n}{\partial\hat{L}} = -\frac{\partial\underline{z}_n}{\partial\hat{L}} = \frac{W_z}{S_n}. \tag{B27}$$

When $W_z = 0$, (B23) and (B24) imply $(\bar{z}_n, \underline{z}_n) = 0$ for all n, and hence $(z_{Gn}, z_{Bn}) = 0$ for all n. Substituting into (B22) and using (B25) and (B26), we find that the derivative of (A1) with respect to W_z at $W_z = 0$ is

$$\rho e^{-\rho W(1+r)}\mathbb{E}_0\left[\frac{\sum_{n=1}^{N}\hat{\eta}_n S_n R_n}{\sum_{n=1}^{N}\hat{\eta}_n S_n} + (1-\lambda)\hat{L}\sum_{n=1}^{N}\left(1_{\{z_n^*(s_n)>0\}}R_n - 1_{\{z_n^*(s_n)<0\}}R_n\right)\right]. \tag{B28}$$

Since the unconditional expected return on the index is nonnegative, the first term in (B28) is nonnegative. The second term in (B28) is positive since $z_n^*(s_n)$ has the same sign as $E_{s_n}(R_n)$. Hence, (B28) is positive, which means that the maximum in (A1) is achieved for $W_z > 0$.

When $\hat{L} = 0$, (B23) and (B24) imply $\bar{z}_n = \underline{z}_n = (W_z\hat{\eta}_n/\Sigma_{n'=1}^{N}\hat{\eta}_{n'}S_{n'})$ for all n, and hence $z_{Gn}(s_n) = z_{Bn} = (W_z\hat{\eta}_n/\Sigma_{n'=1}^{N}\hat{\eta}_{n'}S_{n'})$ for all n. Substituting into (B22) and using (B27), we find that the derivative of (A1) with respect to $\hat{L}$ at $\hat{L} = 0$ is

$$\rho(1-\lambda)W_z\mathbb{E}_0\left[e^{-\rho\left[W(1+r)+\left(W_z\sum_{n=1}^{N}\hat{\eta}_n S_n R_n/\sum_{n=1}^{N}\hat{\eta}_n S_n\right)\right]}\sum_{n=1}^{N}\left(1_{\{z_n^*(s_n)>\bar{z}_n\}}R_n - 1_{\{z_n^*(s_n)<\underline{z}_n\}}R_n\right)\right]. \tag{B29}$$

Since $\mathbb{E}_{s_n}(e^{-\rho\bar{z}_n S_n R_n}R_n) > 0$ for all s_n such that $z_n^*(s_n) > \bar{z}_n$, and $\mathbb{E}_{s_n}(e^{-\rho\underline{z}_n S_n R_n}R_n) < 0$ for all s_n such that $z_n^*(s_n) < \underline{z}_n$, (B29) is positive, which means that the maximum in (A1) is achieved for $\hat{L} > 0$. QED

B10. Proof of Proposition 6

The maximum position $\bar{z}_n$ and minimum position $\underline{z}_n$ that meet the constraint $|z_{2nt} - \eta|G_n(D_{nt}) \le L$ are

$$\bar{z}_n = \eta + \frac{L}{G_n(D_{nt})}, \tag{B30}$$

$$\underline{z}_n = \eta - \frac{L}{G_n(D_{nt})}, \tag{B31}$$

respectively. The position $z_n^*(s_n)$ that maximizes an investor's expected utility conditional on s_n is the position that unconstrained investors hold in equilibrium. In the unconstrained region, defined by (29), $z_n^*(s_n)$ can be derived by setting $z_{1nt} = z_{2nt}$ in the market-clearing condition (28) and is

$$z_{1nt} = z_n^*(s_n) = \frac{\theta_n - \lambda x\eta}{1 - \lambda x}.$$

To derive $z_n^*(s_n)$ in the constrained region, defined by

$$\frac{|\theta_n - \eta|}{1 - \lambda x} G_n(D_{nt}) > L,$$

we distinguish cases. When $\theta_n > \eta$, $z_n^*(s_n)$ can be derived by setting $z_{2nt} = \bar{z}_n$ in (28) and is

$$z_{1nt} = z_n^*(s_n) = \frac{\theta_n - x\eta - [(1 - \lambda)xL/G_n(D_{nt})]}{1 - x}.$$

When instead $\theta_n < \eta$, $z_n^*(s_n)$ can be derived by setting $z_{2nt} = \underline{z}_n$ in (28) and is

$$z_{1nt} = z_n^*(s_n) = \frac{\theta_n - x\eta + [(1 - \lambda)xL/G_n(D_{nt})]}{1 - x}.$$

Hence, when $\theta_n > \eta$, $z_n^*(s_n) \in (\eta, \bar{z}_n]$ in the unconstrained region, and $z_n^*(s_n) > \bar{z}_n$ in the constrained region. When instead $\theta_n < \eta$, $z_n^*(s_n) \in [\underline{z}_n, \eta)$ in the unconstrained region, and $z_n^*(s_n) < \underline{z}_n$ in the constrained region.

The ODEs in the unconstrained and constrained region can be derived from (14) by replacing $\mathbb{E}_t(dR_t^{sh})$ by the drift term in (13), $\mathbb{V}\mathrm{ar}_t(dR_t^{sh})$ by the square of the diffusion term, and z_{1nt} by $z_n^*(s_n)$. This yields

$$\begin{aligned} &D_{nt} + \kappa(\bar{D} - D_{nt})S_n'(D_{nt}) + \frac{1}{2}\sigma^2 D_{nt} S_n''(D_{nt}) - rS(D_{nt}) \\ &\quad = \frac{\rho(\theta_n - \lambda x\eta)}{1 - \lambda x}\sigma^2 D_{nt} S_n'(D_{nt})^2 \end{aligned} \tag{B32}$$

in the unconstrained region and

$$\begin{aligned} &D_{nt} + \kappa(\bar{D} - D_{nt})S_n'(D_{nt}) + \frac{1}{2}\sigma^2 D_{nt} S_n''(D_{nt}) - rS(D_{nt}) \\ &\quad = \frac{\rho(\theta_n - x\eta)}{1 - x}\sigma^2 D_{nt} S_n'(D_{nt})^2 - \frac{\rho \mathrm{sgn}(\theta_n - \eta)(1 - \lambda)xL}{1 - x}\frac{\sigma^2 D_{nt} S_n'(D_{nt})^2}{G_n(D_{nt})} \end{aligned} \tag{B33}$$

in the constrained region.

The derivative of (A1) with respect to η and L in the continuous-time limit can be derived from (B22) by replacing $S_n R_n$ by dR_{nt}^{sh} and is

$$\begin{aligned} &\rho\mathbb{E}_0\Bigg[(1 - \lambda)e^{-\rho\left(W(1+r) + \sum_{n=1}^N z_{2n}(s_n)dR_{nt}^{sh}\right)}\sum_{n=1}^N \left(\frac{\partial \bar{z}_n}{\partial y} 1_{\left\{z_n^*(s_n) > \bar{z}_n\right\}} + \frac{\partial \underline{z}_n}{\partial y} 1_{\left\{z_n^*(s_n) < \underline{z}_n\right\}}\right) dR_{nt}^{sh} \\ &\quad + \lambda e^{-\rho\left(W(1+r) + \sum_{n=1}^N z_{2n} dR_{nt}^{sh}\right)}\sum_{n=1}^N \left(\frac{\partial \bar{z}_n}{\partial y} 1_{\{\Pi_n = \Pi_n^o\}} + \frac{\partial \underline{z}_n}{\partial y} 1_{\{\Pi_n = \Pi_n^o\}}\right) dR_{nt}^{sh}\Bigg]. \end{aligned} \tag{B34}$$

To simplify (B34), we use

$$\begin{aligned}\mathbb{E}_{s_n}\left(e^{-\rho z_n dR_{nt}^{sh}} dR_{nt}^{sh}\right) &= \mathbb{E}_{s_n}((1-\rho z_n dR_{nt}^{sh}) dR_{nt}^{sh}) \\ &= \mathbb{E}_{s_n}(dR_{nt}^{sh}) - \rho z_n \mathbb{V}\mathrm{ar}_{s_n}(dR_{nt}^{sh}) \\ &= \rho[z_n^*(s_n) - z_n]\mathbb{V}\mathrm{ar}_{s_n}(dR_t^{sh}) \\ &= \rho[z_n^*(s_n) - z_n]\sigma^2 D_{nt} S_n'(D_{nt})^2,\end{aligned} \tag{B35}$$

where the third step follows because $z_n^*(s_n)$ is optimal for unconstrained investors and hence satisfies the first-order condition (14). We also use

$$\frac{\partial \bar{z}_n}{\partial \eta} = 1, \tag{B36}$$

$$\frac{\partial \underline{z}_n}{\partial \eta} = 1, \tag{B37}$$

$$\frac{\partial \bar{z}_n}{\partial L} = \frac{1}{G_n(D_{nt})}, \tag{B38}$$

$$\frac{\partial \underline{z}_n}{\partial L} = -\frac{1}{G_n(D_{nt})}, \tag{B39}$$

which follow by differentiating (B30) and (B31).

Using (B35)–(B37), we can write (B34) for $y = \eta$ as

$$\begin{aligned}\rho\epsilon^{-\rho(1+r)W}\mathbb{E}_0\Bigg[(1-\lambda)\Bigg(&\sum_{n:\theta_n>\eta}[z_n^*(s_n) - \bar{z}_n]1_{\{z_n^*(s_n)>\bar{z}_n\}}\sigma^2 D_{nt}S_n'(D_{nt})^2 \\ &+ \sum_{n:\theta_n<\eta}[z_n^*(s_n) - \underline{z}_n]1_{\{z_n^*(s_n)<\underline{z}_n\}}\sigma^2 D_{nt}S_n'(D_{nt})^2\Bigg) + \frac{\lambda}{2}\sum_{n=1}^{N}[2z_n^*(s_n) - \bar{z}_n - \underline{z}_n]\sigma^2 D_{nt}S_n'(D_{nt})^2\Bigg].\end{aligned} \tag{B40}$$

Setting (B40) to zero and using (B30) and (B31) to simplify the third term, we find

$$\begin{aligned}\mathbb{E}_0\Bigg[&\sum_{n:\theta_n>\eta}[z_n^*(s_n) - \bar{z}_n]1_{\{z_n^*(s_n)>\bar{z}_n\}}\sigma^2 D_{nt}S_n'(D_{nt})^2 + \sum_{n:\theta_n<\eta}[z_n^*(s_n) - \underline{z}_n]1_{\{z_n^*(s_n)<\underline{z}_n\}}\sigma^2 D_{nt}S_n'(D_{nt})^2 \\ &+ \frac{\lambda}{1-\lambda}\sum_{n=1}^{N}[z_n^*(s_n) - \eta]\sigma^2 D_{nt}S_n'(D_{nt})^2\Bigg] = 0.\end{aligned} \tag{B41}$$

Using (B35), (B38), and (B39), we can write (B34) for $y = L$ as

$$\begin{aligned}\rho\epsilon^{-\rho(1+r)W}\mathbb{E}_0\Bigg[(1-\lambda)\Bigg(&\sum_{n:\theta_n>\eta}[z_n^*(s_n) - \bar{z}_n]1_{\{z_n^*(s_n)>\bar{z}_n\}}\frac{\sigma^2 D_{nt}S_n'(D_{nt})^2}{G_n(D_{nt})} \\ &- \sum_{n:\theta_n<\eta}[z_n^*(s_n) - \underline{z}_n]1_{\{z_n^*(s_n)<\underline{z}_n\}}\frac{\sigma^2 D_{nt}S_n'(D_{nt})^2}{G_n(D_{nt})}\Bigg) + \frac{\lambda}{2}\sum_{n=1}^{N}[\underline{z}_n - \bar{z}_n]\frac{\sigma^2 D_{nt}S_n'(D_{nt})^2}{G_n(D_{nt})}\Bigg].\end{aligned} \tag{B42}$$

Setting (B42) to zero and using (B30) and (B31) to simplify the third term, we find

$$\mathbb{E}_0\left[\sum_{n:\theta_n>\eta}[z_n^*(s_n)-\bar{z}_n]1_{\{z_n^*(s_n)>\bar{z}_n\}}\frac{\sigma^2 D_{nt}S_n'(D_{nt})^2}{G_n(D_{nt})}-\sum_{n:\theta_n>\eta}[z_n^*(s_n)-\underline{z}_n]1_{\{z_n^*(s_n)<\underline{z}_n\}}\frac{\sigma^2 D_{nt}S_n'(D_{nt})^2}{G_n(D_{nt})}\right. \tag{B43}$$

$$\left.-\frac{\lambda L}{1-\lambda}\sum_{n=1}^{N}\frac{\sigma^2 D_{nt}S_n'(D_{nt})^2}{G_n(D_{nt})^2}\right]=0.$$

When $\eta > \theta_{\max}$, the first term on the left-hand side of (B41) is zero because the summation is over an empty set of n, the second term is negative because the summation is over a nonempty set of n and the set of values of D_{nt} such that $z_n^*(s_n) < \underline{z}_n$ has positive measure, and the third term is negative because $z_n^*(s_n) < \eta$ when $\theta_n < \eta$. Hence, the left-hand side of (B41) is negative, which means that the investor can raise his utility by lowering η. When instead $\eta < \theta_{\min}$, the first term is positive because the summation is over an nonempty set of n and the set of values of D_{nt} such that $z_n^*(s_n) > \bar{z}$ has positive measure, the second term is zero because the summation is over an empty set of n, and the third term is positive because $z_n^*(s_n) > \eta$ when $\theta_n > \eta$. Hence, the left-hand side of (B41) is positive, which means that the investor can raise his utility by raising η. Therefore, $\eta \in [\theta_{\min}, \theta_{\max}]$.

When θ_n can take only one value, $\theta_{\min}$ and $\theta_{\max}$ coincide with that value, and so does $\eta \in [\theta_{\min}, \theta_{\max}]$. Moreover, the first and second terms on the left-hand side of (B43) are zero because the summations are over empty sets of n. Hence, $L = 0$.

When θ_n can take multiple values, the argument showing that the left-hand side of (B41) is negative when $\eta > \theta_{\max}$ can be extended to $\eta \geq \theta_{\max}$ because the set of n such that $\theta_n < \eta$ is nonempty. Likewise, the argument showing that the left-hand side of (B41) is positive when $\eta < \theta_{\min}$ can be extended to $\eta \leq \theta_{\min}$ because the set of n such that $\theta_n > \eta$ is nonempty. Therefore, $\eta \in (\theta_{\min}, \theta_{\max})$. Fixing $\eta \in (\theta_{\min}, \theta_{\max})$, the first and second terms on the left-hand side of (B43) are positive and bounded for $L \geq 0$ and converge to zero when L goes to infinity. When $\lambda = 0$, the third term is zero. Hence, the left-hand side of (B43) is positive, which means that the investor can raise his utility by raising L to infinity. When $\lambda \in [0, 1)$, the third term is a linear and decreasing function of L. Hence, the solution L to (B43) is finite. When λ goes to 1, the third term converges to infinity for any finite L. Hence, the solution L to (B43) converges to zero. QED

References

Alonso, Ricardo, and Niko Matouschek. 2008. "Optimal Delegation." *Rev. Econ. Studies* 75:259–93.

Amador, Manuel, and Kyle Bagwell. 2013. "The Theory of Optimal Delegation with an Application to Tariff Caps." *Econometrica* 81:1541–99.

Ang, Andrew, Robert Hodrick, Yuhang Xing, and Xiaoyan Zhang. 2006. "The Cross-Section of Volatility and Expected Returns." *J. Finance* 61:259–99.

Asness, Clifford, Andrea Frazzini, and Lasse Pedersen. 2014. "Low-Risk Investing without Industry Bets." *Financial Analysts J.* 70:24–41.

Baker, Malcolm, Brendan Bradley, and Jeffrey Wurgler. 2011. "Benchmarks as Limits to Arbitrage: Understanding the Low-Volatility Anomaly." *Financial Analysts J.* 67:40–54.

Basak, Suleyman, and Anna Pavlova. 2013. "Asset Prices and Institutional Investors." *A.E.R.* 103:1728–58.

Berk, Jonathan, and Richard Green. 2004. "Mutual Fund Flows and Performance in Rational Markets." *J.P.E.* 112:1269–95.
Bhattacharya, Ayan, and Maureen O'Hara. 2018. "Can ETFs Increase Market Fragility? Effects of Information Linkages in ETF Markets." Working paper, Cornell University.
Bank for International Settlements. 2003. "Incentive Structures in Institutional Asset Management and Their Implications for Financial Markets." Basel: Bank Internat. Settlements.
Black, Fischer. 1972. "Capital Market Equilibrium with Restricted Borrowing." *J. Business* 45:444–55.
Black, Fischer, Michael Jensen, and Myron Scholes. 1972. "The Capital Asset Pricing Model: Some Empirical Tests." In *Studies in the Theory of Capital Markets*, edited by Michael Jensen, 79–121. New York: Praeger.
Bond, Philip, and Diego Garcia. 2021. "The Equilibrium Consequences of Indexing." *Rev. Financial Studies* 35:3175–230.
Brennan, Michael. 1993. "Agency and Asset Pricing." Working Paper no. 1147, Anderson Graduate School Management, Univ. California Los Angeles.
Buffa, Andrea, and Idan Hodor. 2018. "Institutional Investors, Heterogeneous Benchmarks and the Comovement of Asset Prices." Working paper, Boston Univ.
Buffa, Andrea, Dimitri Vayanos, and Paul Woolley. 2014. "Asset Management Contracts and Equilibrium Prices." Working paper, London School Econ.
Christoffersen, Susan, and Mikhail Simutin. 2017. "On the Demand for High-Beta Stocks: Evidence from Mutual Funds." *Rev. Financial Studies* 30:2596–620.
Cong, William, and Douglas Xu. 2016. "Rise of Factor Investing: Asset Prices, Informational Efficiency and Security Design." Working paper, Cornell Univ.
Cremers, Martijn, and Antti Petajisto. 2009. "How Active Is Your Fund Manager? A New Measure That Predicts Performance." *Rev. Financial Studies* 22:3329–65.
Cuoco, Domenico, and Ron Kaniel. 2011. "Equilibrium Prices in the Presence of Delegated Portfolio Management." *J. Financial Econ.* 101:264–96.
Cvitanic, Jaksa, and Hao Xing. 2018. "Asset Pricing under Optimal Contracts." *J. Econ. Theory* 173:142–80.
Dasgupta, Amil, and Andrea Prat. 2008. "Information Aggregation in Financial Markets with Career Concerns." *J. Econ. Theory* 143:83–113.
Dasgupta, Amil, Andrea Prat, and Michela Verardo. 2011. "The Price Impact of Institutional Herding." *Rev. Financial Studies* 24:892–925.
DeVault, Luke, Richard Sias, and Laura Starks. 2019. "Sentiment Metrics and Investor Demand." *J. Finance* 74:985–1024.
Elton, Edwin, and Martin Gruber. 2013. "Mutual Funds." In *Handbook of the Economics of Finance*, edited by George M. Constantinides, Milton Harris, and Rene M. Stulz, 1011–61. London: Elsevier.
Favilukis, Jack, and Terry Zhang. 2021. "One Anomaly to Explain Them All." Working paper, Univ. British Columbia.
Franzoni, Francesco, Itzhak Ben-David, and Rabih Moussawi. 2017. "Exchange-Traded Funds." *Ann. Rev. Financial Econ.* 9:169–89.
Frazzini, Andrea, and Lasse Heje Pedersen. 2014. "Betting against Beta." *J. Financial Econ.* 111:1–25.
French, Kenneth. 2008. "Presidential Address: The Cost of Active Investing." *J. Finance* 63:1537–73.
Froot, Kenneth, David Scharftstein, and Jeremy Stein. 1992. "Herd on the Street: Informational Inefficiencies in a Market with Short-Term Speculation." *J. Finance* 47:1461–84.

Garcia, Diego, and Joel Vanden. 2009. "Information Acquisition and Mutual Funds." *J. Econ. Theory* 144:1965–95.
Garleanu, Nicolae, and Lasse Pedersen. 2018. "Efficiently Inefficient Markets for Assets and Asset Management." *J. Finance* 73:1663–712.
Gorton, Gary, Ping He, and Lixin Huang. 2010. "Security Price Informativeness with Delegated Traders." *American Econ. J. Microeconomics* 2:137–70.
Grinold, Richard, and Ronald Kahn. 2000. *Active Portfolio Management: A Quantitative Approach for Providing Superior Returns and Controlling Risk*. New York: McGraw Hill.
Gromb, Denis, and Dimitri Vayanos. 2010. "Limits of Arbitrage." *Ann. Rev. Financial Econ.* 2:251–75.
Grossman, Sanford, and Joseph Stiglitz. 1980. "On the Impossibility of Informationally Efficient Markets." *A.E.R.* 70:393–408.
Guerrieri, Veronica, and Peter Kondor. 2012. "Fund Managers, Career Concerns, and Asset Price Volatility." *A.E.R.* 102:1986–2017.
Harrison, Michael, and David Kreps. 1978. "Speculative Investor Behavior in a Stock Market with Heterogeneous Expectations." *Q.J.E.* 92:323–36.
Haugen, Robert, and Nardin Baker. 1996. "Commonality in the Determinant of Expected Stock Returns." *J. Financial Econ.* 41:401–39.
He, Zhiguo, and Wei Xiong. 2013. "Delegated Asset Management, Investment Mandates, and Capital Immobility." *J. Financial Econ.* 107:239–58.
Hong, Harrison, and Jeremy Stein. 2007. "Disagreement and the Stock Market." *J. Econ. Perspectives* 21:109–28.
Huang, Shiyang. 2018. "Delegated Information Acquisition and Asset Pricing." Working paper, Hong Kong Univ.
Jones, Bradley. 2015. "Asset Bubbles: Re-Thinking Policy for the Age of Asset Management." Working paper, Internat. Monetary Fund.
Jorion, Philippe. 2003. "Portfolio Optimization with Constraints on Tracking Error." *Financial Analysts J.* (September): 70–82.
Kapur, Sandeep, and Allan Timmermann. 2005. "Relative Performance Evaluation Contracts and Asset Market Equilibrium." *Econ. J.* 115:1077–102.
Kashyap, Anil, Natalia Kovrijnykh, Jian Li, and Anna Pavlova. 2021a. "The Benchmark Inclusion Subsidy." *J. Financial Econ.* 142:756–74.
———. 2021b. "Is There Too Much Benchmarking in Asset Management?" Working paper, Univ. Chicago.
Kondor, Peter, and Dimitri Vayanos. 2019. "Liquidity Risk and the Dynamics of Arbitrage Capital." *J. Finance* 74:1139–73.
Kyle, Albert, Hui Ou-Yang, and Bin Wei. 2011. "A Model of Portfolio Delegation and Strategic Trading." *Rev. Financial Studies* 24:3778–812.
Lou, Dong, Christopher Polk, and Spyros Skouras. 2019. "A Tug of War: Overnight versus Intraday Expected Returns." *J. Financial Econ.* 134:192–213.
Malamud, Semyon, and Evgeny Petrov. 2014. "Portfolio Delegation and Market Efficiency." Working paper, Swiss Finance Inst.
Nofsinger, John, and Richard Sias. 1999. "Herding and Feedback Trading by Institutional and Individual Investors." *J. Finance* 54:2263–95.
Parlour, Christine, and Uday Rajan. 2019. "Contracting on Credit Ratings: Adding Value to Public Information." *Rev. Financial Studies* 33:1412–44.
Pastor, Lubos, and Robert Stambaugh. 2012. "On the Size of the Active Management Industry." *J.P.E.* 120:740–81.
Pavlova, Anna, and Taisiya Sikorkaya. 2022. "Benchmarking Intensity." *Rev. Financial Studies*, hhac055.
Qiu, Zhigang. 2017. "Equilibrium-Informed Trading with Relative Performance Measurement." *J. Financial and Quantitative Analysis* 52:2083–118.

Roll, Richard. 1992. "A Mean-Variance Analysis of Tracking Error." *J. Portfolio Management* 18:13–22.
Sato, Yuki. 2016. "Delegated Portfolio Management, Optimal Fee Contracts, and Asset Prices." *J. Econ. Theory* 165:360–89.
Scheinkman, Jose, and Wei Xiong. 2003. "Overconfidence and Speculative Bubbles." *J.P.E.* 111:1183–219.
Shleifer, Andrei, and Robert Vishny. 2011. "Fire Sales in Finance and Macroeconomics." *J. Econ. Perspectives* 25:29–48.
Sockin, Michael, and Mindy Zhang Xiaolan. 2018. "Delegated Learning in Asset Management." Working paper, Univ. Texas Austin.
Stambaugh, Robert. 2014. "Presidential Address: Investment Noise and Trends." *J. Finance* 69:1415–53.
Stambaugh, Robert, Jianfeng Yu, and Yu Yuan. 2012. "The Short of It: Investor Sentiment and Anomalies." *J. Financial Econ.* 104:288–302.
———. 2015. "Arbitrage Asymmetry and the Idiosyncratic Volatility Puzzle." *J. Finance* 70:1903–48.
Subrahmanyam, Avanidhar. 1991. "A Theory of Trading in Stock Index Futures." *Rev. Financial Studies* 4:17–51.
Vayanos, Dimitri. 2018. "Risk Limits as Optimal Contracts." Working paper, London School Econ.
Vayanos, Dimitri, and Paul Woolley. 2013. "An Institutional Theory of Momentum and Reversal." *Rev. Financial Studies* 26:1087–145.
Wermers, Russ. 1999. "Mutual Fund Herding and the Impact on Stock Prices." *J. Finance* 54:581–622.