Path-Dependent Alpha: Optimal Contracts, Career Dynamics, and the Cross-Section of Fund Performance

Serge Darolles[1] and Vincent Tena[1]

May 31, 2026

Abstract

We develop a general equilibrium model of active asset management in which optimal contracts generate endogenous alpha heterogeneity through path-dependent career dynamics. Effort, pay, and AUM are functions of one state variable, the manager's continuation value, which evolves with performance. Even with ex ante identical managers, the cross-section of alpha is heterogeneous: history, not type, determines performance. Effort costs scale sublinearly with AUM, producing a hump-shaped effort function and endogenous diseconomies of scale. Structural estimation on hedge fund data matches median tenure, the negative-alpha share, and extreme AUM concentration. Convex capital flows and path dependence generate a superstar effect without heterogeneity in ability.

Keywords: active management, alpha, optimal contracting, moral hazard, career dynamics, continuation value, general equilibrium
JEL Classification: G23, G11, D86, C61

[1]Université Paris-Dauphine, PSL Research University, Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France. serge.darolles@dauphine.psl.eu | vincent.tena@dauphine.psl.eu

1 Introduction

The asset management industry exhibits striking heterogeneity in risk-adjusted performance. At any point in time, some managers generate substantial positive alpha while others destroy value, and these differences persist over horizons that are difficult to reconcile with pure luck. Funds in the same strategy, facing similar market conditions, and drawing managers from the same talent pool deliver markedly different outcomes. What explains the cross-section of alpha across active managers?

Existing theories provide incomplete answers. The career concerns literature [Holmström, 1999, Chevalier and Ellison, 1999] shows that reputation dynamics shape managerial behavior, generating rich predictions about risk-taking over the manager's career. However, these models work with reduced-form reputation processes rather than deriving incentives from optimal contracts. The link between the manager's track record and future effort is assumed rather than derived from the principal's optimization problem. Competitive models of the asset management industry [Berk and Green, 2004, Pástor and Stambaugh, 2012] focus on aggregate outcomes—such as industry size and average performance—but abstract from agency frictions and cannot generate cross-sectional dispersion among managers with identical investment opportunities.[1]

More recently, the empirical literature has shifted from fund-level to manager-level data, revealing that individual career trajectories are first-order determinants of performance. Bai et al. [2026] use confidential U.S. Census Bureau data to show that compensation and career outcomes in asset management vary substantially across individual managers, even controlling for fund characteristics. Joenväärä et al. [2026] document persistent barriers to capital raising at the manager-ownership level in hedge funds. Lee et al. [2026] demonstrate that the horizon of the manager's contract—a bilateral, individual-level object—significantly affects fund performance, with long-horizon contracts insulating managers from short-term funding pressure and enabling superior long-run returns.

These findings point to a gap in the theoretical literature: we lack a unified model in which (i) contracts between investors and managers are *optimal* rather than assumed, (ii) career dynamics are *endogenous* rather than reduced-form, and (iii) the interaction between bilateral contracts and general equilibrium produces the cross-sectional patterns observed in the data.

This paper fills this gap. We develop a continuous-time general equilibrium model of the active management industry in which a risk-neutral principal (the fund's investors) optimally

[1] In competitive equilibrium with no agency problems, capital flows equalize marginal returns across managers. With identical managers facing identical opportunities, there is no source of cross-sectional variation in performance.

contracts with a risk-neutral, wealth-constrained manager who exerts costly, unobservable effort to generate alpha. The optimal contract, derived in the tradition of Sannikov [2008], is fully characterized by a single state variable: the manager's *continuation value*, W. This variable evolves stochastically with the fund's realized performance and encodes the entire history of the relationship. The contract specifies, as functions of W, the manager's effort, compensation, allocation of assets under management (AUM), and the conditions under which the manager is terminated and replaced. The continuation value in our model is the contract-theoretic counterpart of the informal notion of "track record" or "career capital" in the career concerns literature—but its dynamics are pinned down by optimality rather than assumed.

The central insight is that, in general equilibrium, the optimal contract generates *endogenous alpha heterogeneity* even when all managers are ex ante identical. Because W follows a diffusion driven by the fund's noisy output, managers at different points in their career—who have experienced different histories of performance shocks—occupy different positions on the state space and therefore exert different levels of effort. Alpha, which in our model is the equilibrium mapping from continuation value to effort, varies across the population not because managers differ in ability, but because their individual trajectories have placed them at different continuation values. *History, not type, determines performance.*

A distinctive feature of our model is the role of assets under management. We allow the principal to allocate AUM as a convex, increasing function of the manager's continuation value, where a convexity parameter governs how aggressively capital flows to high-W managers and a normalizing constant is pinned down in general equilibrium. Crucially, the cost of effort scales with AUM: managing a larger portfolio makes it harder to generate alpha per dollar, consistent with the well-documented diseconomies of scale in active management [Chen et al., 2004, Pástor et al., 2015]. The flow cost of effort combines an intrinsic convex component with a linear term whose slope increases sublinearly (as $\sqrt{K}$) in fund size, governed by a parameter $\eta \geq 0$ that controls the sensitivity to scale. This creates a rich interaction between incentives and scale. Managers with strong track records (high W) receive more capital but face higher effort costs, producing a hump-shaped relationship between career progression and alpha that matches empirical evidence on the lifecycle of fund performance.

We embed the bilateral contracting problem in general equilibrium by requiring that the total AUM allocated across the population of managers—each at a different point in their career—sums to the aggregate capital seeking active management. The stationary cross-sectional distribution of managers over continuation values is determined endogenously by a Kolmogorov forward equation, with terminated managers instantly replaced by new entrants.

The equilibrium pins down the level of AUM per unit of continuation value, which in turn affects effort costs and the entire incentive structure. This feedback between the population distribution and individual incentives is what makes the model a true general equilibrium theory rather than a partial equilibrium contracting exercise.

A key technical contribution is the derivation of closed-form expressions for the optimal contract. By combining the Hamilton–Jacobi–Bellman equation with the first-order conditions for effort and consumption, we obtain explicit formulas for both the incentive sensitivity and the curvature of the value function. These closed-form expressions eliminate the need to solve an implicit equation at each grid point and reduce the HJB to a first-order ODE system amenable to backward shooting with a stiff solver (Radau IIA), resolving the numerical instability that plagues forward shooting methods in this class of problems.

Our main results are as follows. First, we characterize the optimal contract and show that effort is a smooth, hump-shaped function of the manager's continuation value: it increases from the lower boundary, peaks at an intermediate continuation value (the "sweet spot"), and then declines as diseconomies of scale dominate. Managers near termination work hard but are constrained by low AUM; managers at intermediate continuation values achieve the highest alpha; managers with large AUM face steep effort costs that compress their alpha, consistent with the empirical finding that fund performance declines with scale [Chen et al., 2004, Pástor et al., 2015]. Second, we show that the stationary cross-sectional distribution of alpha is bell-shaped, with substantial mass at both positive and negative values, closely matching the distributions documented in the empirical literature [Barras et al., 2010, Fung et al., 2008, Ardia et al., 2024]. Third, the model generates an endogenous superstar effect: the combination of path-dependent career dynamics and convex capital allocation ($\theta > 1$) concentrates assets in the hands of a small number of high-continuation-value managers, rationalizing the extreme AUM inequality observed in the data without requiring heterogeneity in ability.

Our paper connects to and extends several strands of the literature. The continuous-time contracting framework we build on originates with Sannikov [2008], with antecedents in corporate finance [DeMarzo and Sannikov, 2006] and executive compensation [He, 2009]. Di Tella and Sannikov [2021] extend the Sannikov framework to asset management with hidden savings, characterizing how optimal contracts dynamically distort the agent's access to capital to manage precautionary saving motives. Gryglewicz and Mayer [2023] study dynamic contracting with intermediation, introducing two layers of moral hazard—at both the intermediary and firm levels—and showing that optimal incentives along the hierarchy interact in non-trivial ways. In general equilibrium, Chemla et al. [2025] embed dynamic moral hazard contracts in a competitive market where outside options are endogenously determined, showing that equilibrium compensation externalities distort contract design relative to the shareholder-value-maximizing benchmark. In the asset management context, Ou-Yang [2003] derives optimal compensation for a single fund in partial equilibrium, while Buffa et al. [2022] study how delegated management affects equilibrium asset prices when agents choose portfolios. We extend this line of work by combining the asset management contracting problem of Di Tella and Sannikov [2021] with the general equilibrium approach of Chemla et al. [2025]: we embed the bilateral contract in general equilibrium with a continuum of managers, we model AUM allocation as an endogenous function of the manager's continuation value, and we emphasize the path-dependent nature of alpha—managers with identical ex ante ability deliver heterogeneous performance because their histories differ. Relative to He's single-firm model, our setup endogenizes both fund size and the cross-sectional distribution of outcomes, which is essential for generating industry-level predictions.

Our analysis also relates to the career concerns literature. Holmström [1999] shows that concerns about future employment prospects can substitute for explicit incentive contracts, and Chevalier and Ellison [1999] document age-dependent risk-taking among mutual fund managers consistent with this mechanism, with further empirical support from Hu et al. [2011] and Kempf et al. [2009]. Our contribution is to derive career dynamics from the optimal contract itself rather than from reduced-form reputation models. In career concerns models, the link between current performance and future employment is determined by market beliefs about ability; in our framework, the manager's continuation value W—the contract-theoretic analog of "reputation"—evolves according to the principal's optimal design, and its dynamics are pinned down by the solution to the contracting problem rather than assumed.

The model also provides a microfoundation for the well-documented relationship between fund size and performance. Chen et al. [2004] provide the first systematic evidence that performance declines with size, and Pástor and Stambaugh [2012], Pástor et al. [2015], together with the competitive framework of Berk and Green [2004], develop general equilibrium models in which decreasing returns to scale determine aggregate industry size. These papers, however, abstract from agency problems within individual funds. In our model, diseconomies of scale arise endogenously from moral hazard: the cost of generating alpha through effort increases with fund size, creating an endogenous link between AUM and performance that operates through the optimal contract. This channel speaks to the finding of Berk and van Binsbergen [2015] that performance dispersion does not require substantial heterogeneity in managerial skill—even with identical ability, path-dependent career dynamics under the optimal contract produce heterogeneous outcomes.

Finally, our model contributes to the literature on concentration in asset management.

Pástor et al. [2017] document rising AUM concentration among top-performing mutual funds, a trend Sensoy [2009] shows has accelerated since the 1990s. In assignment models of superstars [Rosen, 1981, Gabaix and Landier, 2008, Terviö, 2008], extreme inequality arises because the most talented individuals are matched with the largest firms. Our model offers a complementary explanation rooted in path dependence rather than heterogeneous ability: stochastic performance shocks, amplified by convex capital flows, concentrate assets in the hands of a small number of high-continuation-value managers without requiring any ex ante differences among them.

The rest of the paper is organized as follows. Section 2 presents the model. Section 3 characterizes the equilibrium and derives closed-form expressions. Section 4 presents the numerical solution and discusses its properties. Section 5 presents comparative statics, testable predictions, reduced-form evidence for the concave alpha–AUM relationship, and a structural test that recovers the latent continuation value from observables via machine learning. Section 6 structurally estimates the model using hedge fund data from HFR via the method of simulated moments. Section 7 examines robustness to the rolling estimation window, factor model, sample period, and the scale-cost parameter. Section 8 concludes. Proofs are in the Appendix.

2 Model

We extend the continuous-time principal–agent framework of Sannikov [2008] by introducing two additional features. First, we embed the bilateral contracting problem in a general equilibrium with a continuum of principal–agent pairs, so that outside options are endogenously determined by competitive rematching in the market for active management. Second, we allow the principal to allocate assets under management (AUM) as a function of the agent's continuation value, with effort costs that scale with fund size, capturing the diseconomies of scale documented in the empirical literature.

2.1 Agents and Technologies

Time $t \in [0, \infty)$ is continuous. There is a unit mass of fund management positions (principals) and a unit mass of portfolio managers (agents) who interact through a competitive market for active management. Both principals and agents discount future payoffs at a common rate $r > 0$.

The probability space $(\Omega, \mathcal{F}, \mathbb{P}^0)$ is complete and supports a standard one-dimensional Brownian motion $Z = \{Z_t\}_{t \geq 0}$, which drives the continuous part of each fund's output. The filtration $\mathbb{F} = (\mathcal{F}_t)_{t \geq 0}$ is generated by the output process and satisfies the usual conditions.

Each fund's cumulative output (per-dollar excess return) evolves according to

$$dY_t = a_t \, dt + \sigma \, dZ_t, \tag{1}$$

where $a_t \geq 0$ is the manager's effort (privately observed) and $\sigma > 0$ is the volatility of the noise. We normalize the effort productivity parameter to $\mu = 1$, which is without loss of generality because μ enters the output process as a scale factor observationally equivalent to a rescaling of effort units. The output process Y_t is publicly observed by both the principal and the agent and is the only contractible signal of the manager's effort. The principal's payoff is proportional to the *total* dollar return $K_t \cdot dY_t$, reflecting the fact that a given per-dollar alpha is more valuable when deployed on a larger capital base.

Assumption 1 (Observability)**.** *The principal observes the cumulative output process Y_t but not the manager's effort a_t or consumption c_t.*

The manager has flow utility $u(c) = -2\sqrt{c}$ over consumption $c \geq 0$, which is increasing and strictly concave with $u'(c) \to \infty$ as $c \to 0$. This specification ensures tractability: following Sannikov [2008] and Possamai and Touzi [2025], it guarantees that the principal's problem admits a well-defined stopping region and that optimal consumption has a closed-form expression.

The cost of effort depends on the fund's assets under management. Let K_t denote the capital allocated to the fund at time t. The flow cost of effort takes the form

$$h(a, K) = \tfrac{1}{2}a^2 + \beta(K)\, a, \qquad \text{where} \quad \beta(K) \equiv 0.4 + \eta\sqrt{K}. \tag{2}$$

The first term, $\frac{1}{2}a^2$, represents the intrinsic convex cost of effort that is independent of scale—the intellectual cost of generating investment ideas. The linear term $\beta(K)\, a$ decomposes into a fixed component $0.4\, a$ capturing a baseline implementation cost, and an interaction term $\eta\sqrt{K}\, a$ reflecting the additional cost of *implementing* those ideas at scale: executing trades, managing liquidity, and monitoring positions across a larger portfolio. The sublinear (square-root) dependence on K captures the idea that implementation costs increase with fund size but at a diminishing rate, reflecting partial economies of scale in trading infrastructure. When $\eta = 0$, the model reduces to a standard contracting problem without scale effects; when $\eta > 0$, managing a larger fund makes it harder to generate alpha per dollar, consistent with the well-documented diseconomies of scale [Chen et al., 2004, Pástor et al., 2015].

The sensitivity of the manager's continuation value to the noise in output is determined by the incentive-compatibility condition. The minimal exposure needed to implement effort a

is

$$\gamma(a,K) = h'_a(a,K) = a + \beta(K) = a + 0.4 + \eta\sqrt{K}. \tag{3}$$

This follows the logic of Sannikov [2008]: the agent chooses effort to maximize $\gamma \cdot a - h(a)$, and since it is costly for the principal to expose the agent to risk, the sensitivity is set at the minimal incentive-compatible level $\gamma = h'_a(a)$. The partial cost specification has the attractive feature that γ decomposes into an effort-proportional term and a size-dependent shift $\beta(K)$ that is independent of the effort level.

2.2 AUM Allocation

Each principal makes two key decisions: (i) whether to retain the current manager through a contractual arrangement, and (ii) how much capital to allocate to the fund. The principal allocates AUM to the manager as a function of the manager's continuation value W. We parameterize this allocation as:

$$K(W) = 1 + \frac{W^\theta}{\Omega_\theta}, \tag{4}$$

where $\theta \geq 1$ governs the convexity of the allocation rule and

$$\Omega_\theta \equiv \int_R^{\bar{W}} W^\theta \, g(W) \, dW \tag{5}$$

is the aggregate AUM normalizing constant, with $g(W)$ the stationary cross-sectional density of managers over continuation values. The unit term in (4) ensures a minimum AUM floor: even managers at the lowest continuation values manage some capital, reflecting the institutional reality that fund managers always oversee a baseline portfolio. When $\theta = 1$, the AUM allocation above the floor is proportional to continuation value; when $\theta > 1$, capital is allocated convexly, loading more on managers with strong track records. This power-law specification captures the empirical regularity that capital flows are convex in past performance [Chevalier and Ellison, 1997, Sirri and Tufano, 1998]. The normalization ensures that total AUM across all active managers is consistent with the aggregate capital seeking active management.

2.3 Bilateral Contracting

A contract is denoted by $\Gamma = (C, \tau)$, where $C = \{C_t\}_{t \geq 0}$ is a non-negative consumption flow to the manager and τ is a stopping time representing the termination date. Both processes are adapted to the public filtration generated by the observed output. The manager privately chooses effort a_t, which affects the drift of output but is not observed by the principal.

Under contract Γ, the principal's expected present value starting at time $t = 0$ is

$$V(0;\Gamma) = \mathbb{E}\left[\int_0^{\tau} e^{-rt}\,(K_t \cdot a_t - C_t)\,dt + e^{-r\tau}L\right], \tag{6}$$

where $K_t = K(W_t)$ is the capital allocated to the fund and L is the principal's value upon termination, which includes the possibility of rematching with another manager in the competitive market. Because $K(W)$ adjusts instantaneously to the manager's continuation value and the aggregate distribution is stationary, it does not introduce an additional state variable. The manager's expected payoff is

$$W(0;\Gamma) = \mathbb{E}\left[\int_0^{\tau} e^{-rt}\,(u(C_t) - h(a_t, K_t))\,dt + e^{-r\tau}R\right], \tag{7}$$

where R represents the manager's outside option upon termination.

The principal's problem is to design a contract that maximizes expected fund value subject to offering the manager at least his required initial value. Formally:

$$\max_{\Gamma} \quad V(0;\Gamma) \tag{8}$$

$$\text{s.t.} \quad W(0;\Gamma) \geq R, \tag{IR}$$

$$\Gamma \text{ is incentive compatible.} \tag{IC}$$

As in DeMarzo and Sannikov [2006], Sannikov [2008], and Chemla and Tena [2025], the manager's continuation value process W_t serves as the sole state variable. For any $t < \tau$, W_t represents the expected utility to the manager from continuing the contract at time t onwards:

$$W_t = \mathbb{E}_t\left[\int_t^{\tau} e^{-r(s-t)}\,(u(C_s) - h(a_s, K_s))\,ds + e^{-r(\tau-t)}R\right]. \tag{9}$$

Given consumption and effort policies $c(W)$ and $a(W)$, the dynamics of the continuation value are (cf. Sannikov 2008, equation (3)):

$$dW_t = r\big[W_t - u(c(W_t)) + h(a(W_t), K(W_t))\big]\,dt + r\,\gamma(a(W_t), K(W_t))\,\sigma\,dZ_t. \tag{10}$$

The factor r multiplies *both* the drift and the diffusion, reflecting Sannikov's normalization in which the continuation value is defined as a present value. The drift reflects the promise-keeping constraint: the principal must deliver the promised utility to the manager. The diffusion term $r\,\gamma\,\sigma$ is the key incentive device: by tying W to the noise in output, the contract makes the manager bear risk and thus motivates effort.

The market for active management is competitive. At any time t, a manager can terminate the relationship and enter into a new contractual arrangement with another principal, subject to a proportional switching cost α_A. Similarly, a principal can replace the current manager by paying a proportional cost α_P. These costs capture search frictions reflecting the relationship-specific nature of the arrangement [see, e.g., Chemla and Tena, 2025]. Following the search and matching literature [Pissarides, 2000], we impose $\alpha_A = \alpha_P \equiv \alpha$: because the continuum of fund management positions and the continuum of portfolio managers are of equal measure, the expected duration of search is symmetric, so both sides of the market bear similar rematching costs.

2.4 Boundaries and Termination

Let $F(W)$ denote the principal's value function when the manager's continuation value is W. The optimal contract operates on a bounded interval $[R, \bar{W}]$, with two endogenous boundaries that jointly determine participation and termination.

The contract can terminate for two reasons. First, the principal may optimally choose to replace the manager when the relationship is no longer sufficiently productive. At the upper boundary $\bar{W}$, the principal terminates and rematches with a new manager. The outside value function, which captures the principal's payoff from the alternative arrangement, takes the quadratic form

$$F_{\text{out}}(W) = L - (W - R)^2, \tag{11}$$

which satisfies $F_{\text{out}}(R) = L$ and is concave in W. Second, termination may occur because the manager's continuation value has fallen to a level too low to sustain effort. At the lower boundary R, the manager exits to the competitive market and is instantly replaced by a new entrant who starts at the optimal entry point $W^* = \arg\max_W F(W)$.

The boundary conditions for the principal's value function are:

$$F(R) = L, \qquad \text{(value matching at the lower boundary)} \tag{12}$$

$$F'(W^*) = 0, \qquad \text{(optimality of the entry point)} \tag{13}$$

$$F(\bar{W}) = F_{\text{out}}(\bar{W}), \quad F'(\bar{W}) = F'_{\text{out}}(\bar{W}), \quad \text{(value matching and smooth pasting at } \bar{W}) \tag{14}$$

The termination values (R, L) play a central role in shaping the competitive equilibrium. Upon termination, both parties receive their respective outside options:

$$R = (1 - \alpha)\, W^*, \tag{15}$$

$$L = (1 - \alpha)\, F(W^*). \tag{16}$$

Equation (15) captures the idea that the manager can always quit and immediately find another position at a different fund, subject to bearing search cost α. He takes the expected payoffs offered by other principals, W^*, as given and computes the associated outside option. Equation (16) states that upon termination, the principal can find another identical manager and obtain $F(W^*)$ upon bearing the same search cost α. The pair (R, L) is determined in equilibrium, and both firms and managers behave as price takers in the competitive market.

3 Equilibrium Characterization

We characterize the equilibrium in two steps. First, we solve the contracting problem for each individual principal–agent pair, given their respective outside options. Second, we solve for the equilibrium level of AUM together with the endogenous outside options.

3.1 Equilibrium Definition

We adapt the definition of a competitive labor-market equilibrium in the spirit of Chemla and Tena [2025], in which each principal–agent pair designs an optimal contract while taking as given the equilibrium outside options.

Definition 1 (Stationary competitive equilibrium)**.** *A stationary competitive equilibrium consists of a value function $F(W)$, policy functions $\{a(W), c(W)\}$, an AUM allocation $K(W)$, a stationary density $g(W)$, boundaries $\{R, \bar{W}\}$, an optimal entry point W^*, and a normalizing constant Ω_θ such that:*

(i) Given (R, L), the contract Γ^* and W^* solve the principal's problem (8);

(ii) The outside options satisfy the competitive rematching conditions (15)–(16);

(iii) g solves the Kolmogorov forward equation (26) with $\int_R^{\bar{W}} g(W)\,dW = 1$;

(iv) The AUM market clears: $\Omega_\theta = \int_R^{\bar{W}} W^\theta\, g(W)\,dW$.

The equilibrium values of the principal and manager at the outset of the optimal contract are $V^* = F(W^*)$ and $W^* = W^*$, respectively. These are consistent with the competitive rematching process and the market-clearing condition for AUM, ensuring that the outside options (R, L) are aligned with both the dynamics of the market for active management and the cross-sectional distribution of managers.

3.2 Hamilton–Jacobi–Bellman Equation

The principal's value function $F(W)$ satisfies the HJB equation (cf. Sannikov 2008, equation (5)):

$$r\,F(W) = \max_{a,c}\; r\big(K(W)\,a - c\big) + r\,F'(W)\big[W - u(c) + h(a,K)\big] + \tfrac{1}{2}\,r^2\sigma^2\,\gamma^2\,F''(W). \tag{17}$$

The factor r multiplies each term on the right-hand side, reflecting Sannikov's normalization. Dividing through by r:

$$F(W) = K\,a - c + F'(W)\big[W - u(c) + h(a,K)\big] + \kappa\,\gamma^2\,F''(W), \tag{18}$$

where $\kappa \equiv r\sigma^2/2$. The term $K\,a$ replaces the standard a of Sannikov [2008] because the principal values total dollar alpha $K \cdot a$ rather than per-dollar alpha. The principal chooses consumption c and effort a to maximize the right-hand side. The first-order conditions are:

$$\mathrm{FOC}_c: \quad 1 = F'(W)\cdot u'(c), \tag{19}$$

$$\mathrm{FOC}_a: \quad 0 = K + F'(W)\cdot\gamma + 2\kappa\,\gamma\,F''(W). \tag{20}$$

With $u(c) = -2\sqrt{c}$, the consumption FOC yields $u'(c) = -c^{-1/2}$, so $c^{-1/2} = -1/F'(W)$ (valid when $F' < 0$), hence:

$$c^*(W) = \frac{\big[F'(W)\big]^2}{4}, \qquad u(c^*) = -2\sqrt{c^*} = F'(W). \tag{21}$$

Remark 1. *Equation (21) reveals that optimal consumption is determined analytically by the slope of the value function. When $F'(W) < 0$ (which holds in the interior of the contract), the manager receives positive consumption. At W^*, where $F' = 0$, consumption is zero—new entrants receive no immediate compensation and must build their continuation value through performance.*

The effort FOC (20) implicitly defines optimal effort through the incentive sensitivity γ. Since $\gamma = a + \beta(K)$ and $\gamma_a = 1$, the FOC simplifies to

$$K + (F' + 2\kappa\,F'')\,\gamma = 0, \tag{22}$$

yielding the closed-form solution:

$$\gamma^*(W) = \frac{-K(W)}{F'(W) + 2\kappa\,F''(W)}, \tag{23}$$

and the optimal effort $a^*(W) = \gamma^*(W) - \beta(K(W))$.

3.3 Closed-Form Expressions

The key technical contribution is a pair of closed-form expressions that reduce the HJB to a first-order ODE system, enabling efficient backward shooting. These are obtained by substituting the effort and consumption FOCs back into the HJB equation.

Proposition 1 (Closed-form incentive sensitivity)**.** *Under the optimal contract, the incentive sensitivity admits the closed-form expression*

$$\gamma(W) = \frac{2\big[F(W) + K(W)\,\beta(W) + c(W) - F'(W)\big(W - u(c(W)) - \frac{1}{2}\beta(W)^2\big)\big]}{K(W)}, \tag{24}$$

where $\beta(W) = 0.4 + \eta\sqrt{K(W)}$ *and* $c(W) = [F'(W)]^2/4$.

Proof. See Appendix A.2. □

The curvature of the value function is then pinned down by the effort FOC:

$$F''(W) = \frac{-K(W)/\gamma(W) - F'(W)}{r\,\sigma^2}. \tag{25}$$

Note that r appears in the *denominator*, reflecting the $r^2\sigma^2$ coefficient in the diffusion term of the HJB. Together, equations (24) and (25) reduce the HJB to the first-order system $(F, F') \to F''$, which can be integrated as an initial value problem.

The following theorem characterizes the optimal contract.

Theorem 1 (Optimal contract)**.** *The unique concave function F that satisfies (17)–(14) characterizes any optimal contract that yields positive profit to the principal. The optimal contract is based on the manager's continuation value as the sole state variable, which starts at W^* and evolves according to (10) under compensation $c_t = c(W_t)$ and effort $a_t = a(W_t)$, until the stopping time τ at which W_t first reaches either boundary.*

Termination occurs when W_t reaches one of the two boundaries:

(i) the lower boundary $R = (1-\alpha)W^*$, at which the manager exits to the competitive market with continuation value R;

(ii) the upper boundary $\bar{W}$, at which the principal terminates and rematches with a new manager.

The pair of boundaries $(R, \bar{W})$ and the function $F(W)$ jointly determine the stationary contract that maximizes the principal's expected value, given the equilibrium switching costs α.

Proof. See Appendix A.3. ☐

Proposition 2 (Hump-shaped effort)**.** *Under the optimal contract, when $\eta > 0$ and $\theta > 1$, the effort function $a(W)$ is hump-shaped on $[R, \bar{W}]$: there exists an interior point $W_{\text{peak}} \in (R, \bar{W})$ at which effort attains its maximum, with $a'(W) > 0$ for $W < W_{\text{peak}}$ and $a'(W) < 0$ for $W > W_{\text{peak}}$.*

Proof. See Appendix A.4. ☐

The intuition for the hump shape is as follows. Three forces interact to determine effort. Near the lower boundary R, the manager has low AUM (since $K(W) \approx 1$ when W is small), so scale costs are negligible; however, two forces limit effort: the curvature of the value function $F''(W)$ imposes a steep marginal cost of risk exposure, and the effective marginal value of effort $K(W)$ is low because little capital is at stake. As W increases, the value function flattens, and—crucially—$K(W)$ rises because the fund manages more capital, making each unit of effort more valuable to the principal. Both forces push effort upward. Beyond the peak, a manager with high W commands large AUM, which raises effort costs through the scale channel ($\eta\sqrt{K}$ grows), and the concavity of the value function makes risk exposure increasingly expensive. These cost forces eventually dominate the rising $K(W)$, pushing effort downward past W_{peak}.

3.4 Stationary Distribution

In the stationary equilibrium, the cross-sectional distribution of managers over continuation values is determined by the Kolmogorov forward equation (KFE). Let $g(W)$ denote the density of managers at continuation value W. Then:

$$0 = -\frac{\partial}{\partial W}\big[\mu_W(W)\, g(W)\big] + \frac{1}{2}\frac{\partial^2}{\partial W^2}\big[\sigma_W^2(W)\, g(W)\big] + \lambda\, \delta(W - W^*), \tag{26}$$

where

$$\mu_W(W) = r\big[W - u(c(W)) + h(a(W), K(W))\big], \tag{27}$$

$$\sigma_W(W) = r\,\gamma(W)\,\sigma, \tag{28}$$

$\delta(\cdot)$ is a Dirac mass at the entry point, and λ is the flow rate of new entrants (which equals the flow rate of terminations in steady state). The boundary conditions are absorbing: $g(R) = g(\bar{W}) = 0$.

The density $g(W)$ pins down the cross-sectional distribution of all endogenous variables. In particular, the distribution of alpha in the population is determined by the mapping $a(W)$ composed with $g(W)$.

Proposition 3 (Endogenous alpha heterogeneity)**.** *In the stationary equilibrium, the cross-sectional distribution of* $\alpha \equiv a(W)$ *has full support on* $[a(\bar{W}), a(R)]$ *and is unimodal.*

Proof. See Appendix A.5. □

The proposition formalizes the paper's central insight: even though all managers are ex ante identical, the stationary cross-section of alpha is a non-degenerate distribution. The source of heterogeneity is purely *path dependence*—different managers occupy different positions on the state space because they have experienced different histories of performance shocks.

3.5 Computation

The equilibrium is computed by combining backward shooting on the HJB with root-finding for the general equilibrium conditions.

We first solve the model without general equilibrium effects ($K = 1$, $\eta = 0$). This reduces to the standard Sannikov [2008] problem with retirement value $F_{\text{ret}}(W) = -W^2$ and $F'_{\text{ret}}(W) = -2W$. We shoot backward from a candidate W_{gp} (the "golden parachute" point where $F = F_{\text{ret}}$) using the Radau IIA stiff solver, and find W_{gp} by bisection on the condition $F(0) = 0$.

For the full general equilibrium, the unknowns are $(\bar{W}, R, L)$. Given these, we set boundary conditions at $\bar{W}$ from smooth pasting with F_{out}:

$$F(\bar{W}) = L - (\bar{W} - R)^2, \qquad F'(\bar{W}) = -2(\bar{W} - R),$$

and shoot backward to R using the Radau solver with the closed-form ODE system (24)–(25). The three residual conditions are:

$$\begin{aligned}
&F'(W^*) = 0 && \text{(peak at } W^* = R/(1-\alpha)\text{)},\\
&L = (1-\alpha)\,F(W^*) && \text{(GE consistency)},\\
&F(R) = L && \text{(lower boundary value)}.
\end{aligned}$$

This three-dimensional system is solved using a Levenberg–Marquardt algorithm.

To handle the nonlinearity introduced by endogenous $K(W)$ and $\eta > 0$, we use a continuation method. Starting from the $K = 1$, $\eta = 0$ benchmark, we gradually increase both K

and η toward their target values in 8 steps, using the solution at each intermediate step as the initial guess for the next. This ensures robust convergence.

Given the converged value function and policies, we compute the drift (27) and diffusion (28) and solve the KFE (26) for the stationary density $g(W)$ using a finite-difference upwind scheme. The normalizing constant Ω_θ is then updated. The full numerical algorithm is described in Appendix B.

4 Numerical Solution

We solve the model for the baseline parameter configuration: $r = 0.1$, $\sigma = 1.0$, $\alpha = 0.3$, $\theta = 1.5$, $\eta = 0.15$, and $\Omega_\theta = 1.0$. The baseline shift parameter is $\beta(K) = 0.4 + 0.15\,K$.

As a validation of the numerical method, we first solve the pure Sannikov [2008] benchmark ($K = 1$, $\eta = 0$, no GE). Our backward shooting algorithm recovers $W_{\text{gp}} = 0.961$, $W^* = 0.128$, $F(W^*) = 0.125$, and $F'(0) = 2.458$. These values match those reported in Figure 1 of Sannikov [2008] for $r = 0.1$, $\sigma = 1$, $h(a) = \frac{1}{2}a^2 + 0.4\,a$.

The numerical algorithm converges to residuals at machine precision ($\sim 10^{-20}$). The equilibrium values are reported in Table 1.

Table 1: **Equilibrium values: baseline solution.**

Object	**Symbol**	**Value**	**Description**
Panel A: Parameters			
Discount rate	r	0.100	Time preference
Output volatility	σ	1.000	Noise in fund output
Cost–AUM sensitivity	η	0.150	Diseconomies of scale
AUM convexity	θ	1.500	Capital allocation convexity
Search cost (symmetric)	α	0.300	Rematching friction
Panel B: Endogenous equilibrium objects			
Optimal entry point	W^*	0.291	$\arg\max F(W)$
Lower boundary	R	0.203	Termination threshold
Upper boundary	$\bar{W}$	0.738	Outside option threshold
Principal's entry value	$F(W^*)$	0.146	Peak of F
Outside option	L	0.102	$= (1-\alpha)F(W^*)$
Employment region	$[R, \bar{W}]$	$[0.203,\ 0.738]$	Active contracts
Effort range	$[a_{\min}, a_{\max}]$	$[0.655,\ 0.858]$	Across employment region
Average AUM	$\bar{K}$	1.336	Cross-sectional mean

Figure 1 displays the four panels of the baseline solution.

$F(W)$ is concave and single-peaked at $W^* = 0.291$, where $F'(W^*) = 0$. At the lower

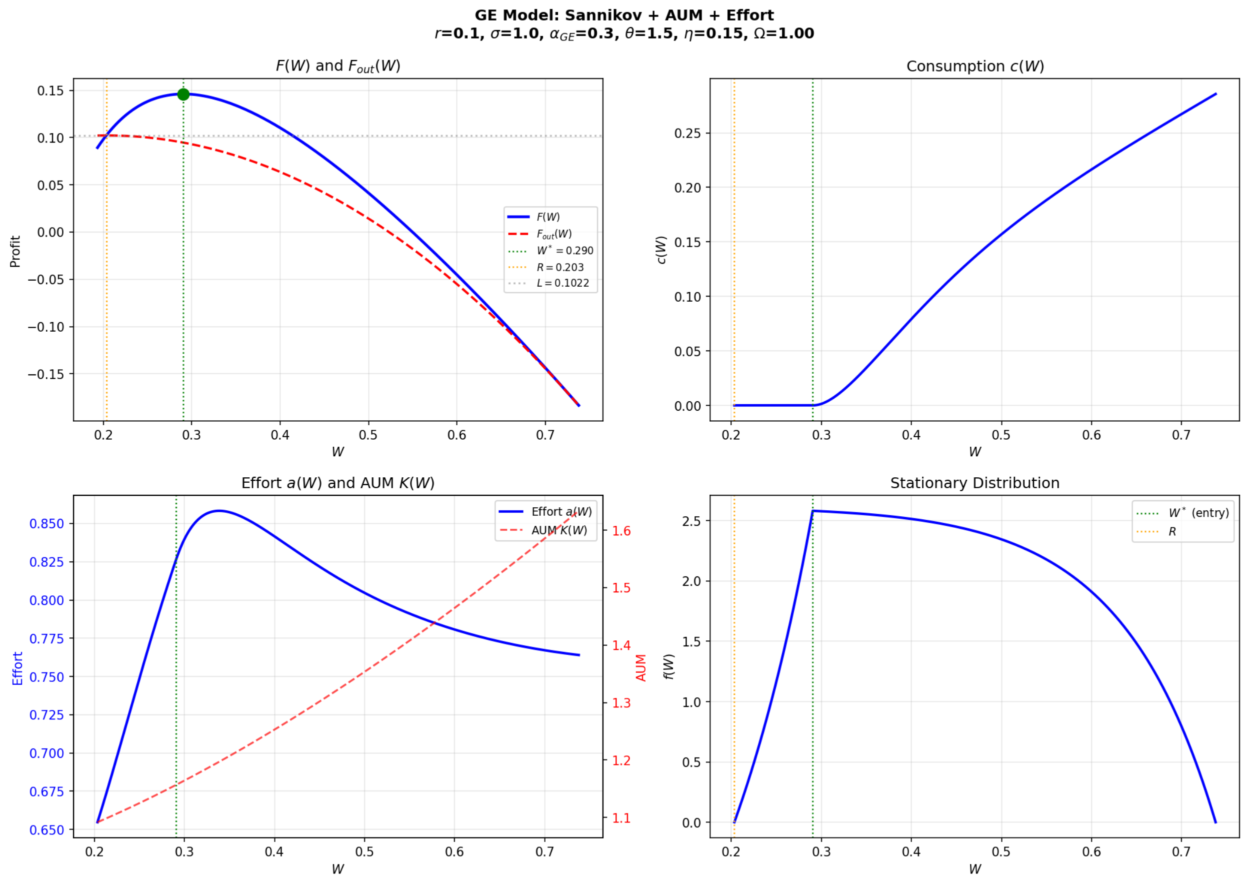

Figure 1: **Optimal contract: baseline solution.** Panel (a): Principal's value function $F(W)$ (solid blue) and outside value $F_{\text{out}}(W)$ (dashed red). The green dot marks the optimal entry point W^*. Panel (b): Optimal consumption $c(W) = [F'(W)]^2/4$. Panel (c): Optimal effort $a(W)$ (blue, left axis) and AUM allocation $K(W) = 1 + W^\theta/\Omega_\theta$ (red dashed, right axis). Panel (d): Stationary cross-sectional density $g(W)$. Parameters: $r = 0.1$, $\sigma = 1.0$, $\alpha = 0.3$, $\theta = 1.5$, $\eta = 0.15$.

boundary $R = 0.203$, the value function satisfies $F(R) = L = 0.102$. At the upper boundary $\bar{W} = 0.738$, F is tangent to F_{out} from above: $F(\bar{W}) = F_{\text{out}}(\bar{W})$ and $F'(\bar{W}) = F'_{\text{out}}(\bar{W})$. The employment region—the interval on which $F(W) \geq F_{\text{out}}(W)$—spans $[0.203, 0.738]$, a wide range reflecting the moderate search friction $\alpha = 0.3$.

Consumption $c(W) = [F'(W)]^2/4$ is U-shaped: it is zero at W^* (where $F' = 0$), increases as W moves away from W^* in either direction, and is highest near the boundaries where $|F'|$ is large. This reflects the optimal contract's structure: new entrants receive no immediate compensation and must build their continuation value through performance, while managers near the boundaries receive higher consumption as the contract nears termination.

The effort function $a(W)$ exhibits the predicted hump shape. Effort rises from approximately 0.66 near the lower boundary, peaks at around 0.86 at an intermediate continuation value, and then declines toward the upper boundary as diseconomies of scale dominate. AUM $K(W)$ increases monotonically from approximately 1.1 at R to 1.6 at $\bar{W}$, with the convex allocation rule concentrating capital on high-W managers.

The hump shape in effort reflects the three-way interaction between incentive power, scale costs, and the curvature of the value function. At low W, the manager has limited AUM and the value function is steep ($|F'|$ large), limiting the principal's ability to provide incentives without excessive risk. At high W, the manager commands large AUM, but the $\eta\sqrt{K}$ term in the effort cost raises the marginal cost of effort, compressing alpha. The peak occurs where the trade-off between these forces is most favorable.

The stationary density $g(W)$ rises steeply from zero at the lower boundary (absorbing), peaks near W^* where managers are injected, and then declines gradually toward the upper boundary. The asymmetry—a longer right tail than left tail—reflects the fact that the drift of the W process is positive near the entry point (managers tend to accumulate continuation value early in their careers) and the diffusion is large enough to spread the distribution over the entire employment region.

Figure 2 overlays the stationary density $g(W)$ with the effort-implied alpha $a(W) \times \lambda$ at the calibrated parameters of Section 6. The density is skewed right, with the bulk of managers concentrated between the entry point W^* and the midpoint of the employment region; the effort function peaks near entry where the trade-off between incentive power and scale costs is most favorable, then declines steadily for high-W managers who command large AUM. Because the density is concentrated precisely where effort is highest, the model generates an average alpha that exceeds what a uniform distribution over the employment region would imply.

4.1 Lifecycle of a Representative Manager

The model implies rich dynamics for individual career trajectories. A manager enters at $W^* = 0.291$ with zero consumption, moderate effort ($a \approx 0.86$), and AUM $K \approx 1.2$. Early in the career, the continuation value W_t fluctuates substantially, driven by the noise in fund returns. If performance is favorable, W drifts upward: the manager receives more capital, earns higher consumption, but faces rising effort costs through the $\eta\sqrt{K}$ channel. If performance is poor, W drifts downward toward R: the manager receives less capital and lower consumption, but effort remains relatively high as scale costs diminish. Eventually, W hits one of the two boundaries—either the lower boundary R (termination due to poor performance) or the upper boundary $\bar{W}$ (termination because the outside option dominates)—and the manager restarts at W^* in a new firm.

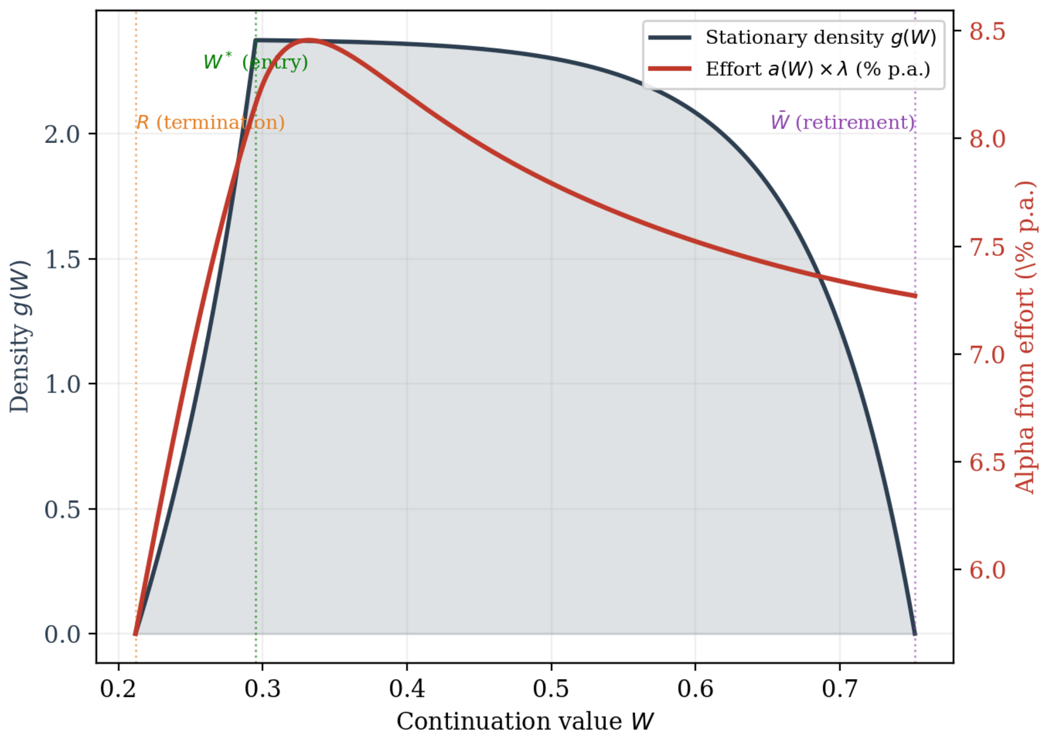

Figure 2: **Stationary density and effort-implied alpha.** Left axis (dark): stationary density $g(W)$ of managers over continuation values. Right axis (red): alpha from effort $a(W) \times \lambda$, in annual percentage points, at the calibrated parameters of Table 3. Vertical lines mark the termination boundary R, the optimal entry point W^*, and the retirement boundary $\bar{W}$. The density is right-skewed, with mass concentrated near the entry point where effort—and hence alpha—is highest.

This pattern is consistent with the empirical evidence on fund lifecycles: young funds have high effort but small AUM; established funds have large AUM but declining alpha per dollar; and a substantial fraction of funds are terminated at relatively young ages.

5 Comparative Statics and Implications

5.1 The Role of Diseconomies of Scale (η)

The parameter η governs the strength of AUM-related diseconomies of scale. When η is small, effort costs are nearly independent of fund size, and the model reduces to a standard contracting problem without scale effects. As η increases, the cost of effort rises steeply with AUM, which compresses alpha at the top of the distribution (where managers have the most capital) and increases the fraction of negative-alpha funds.

This comparative static has a direct empirical counterpart: it predicts that in strategies with stronger diseconomies of scale (e.g., capacity-constrained strategies like micro-cap equity or convertible arbitrage), the cross-section of alpha should be more compressed and the fraction of underperforming funds should be higher. As η increases from 0 to 0.6, the effort function flattens and the stationary density shifts toward lower continuation values.

5.2 Flow Convexity (θ)

The parameter θ controls how aggressively capital is allocated toward managers with high continuation values. When $\theta = 1$, AUM above the floor is linear in W; when $\theta > 1$, the allocation is convex, concentrating capital in the hands of "star" managers. Increasing θ amplifies the AUM channel: top managers receive disproportionately more capital, which raises their effort costs and lowers their alpha. At the same time, the increased capital concentration raises aggregate efficiency because capital is allocated to managers with proven track records.

This captures the empirical regularity that capital flows are convex in past performance [Chevalier and Ellison, 1997, Sirri and Tufano, 1998]: winners attract disproportionate inflows. Our model endogenizes this pattern as part of the optimal contract rather than assuming it exogenously.

Figure 3 illustrates the comparative statics of the effort function with respect to both η and θ. Panel (a) shows that increasing η amplifies the hump shape: the effort peak sharpens and the decline at high continuation values steepens, reflecting the growing penalty from scale costs. At $\eta = 0.30$ the spread between peak and trough effort exceeds 1 percentage point in annual alpha. Panel (b) shows that increasing θ narrows the effort function's range and lowers its level over most of the state space. Higher flow convexity concentrates capital on top managers, raising their scale costs and compressing alpha at the top, while leaving effort at low W largely unchanged. These patterns generate clear, testable cross-sectional predictions.

5.3 AUM Concentration and the Superstar Effect

A striking feature of the modern asset management industry is the extreme concentration of capital. In the HFR database, the top 10% of funds control approximately 71% of total AUM, with a Gini coefficient of 0.80. Moreover, this concentration has been rising: Pástor et al. [2017] and Sensoy [2009] document that top funds have captured an increasing share of flows since the 1990s. Our model provides a unified explanation for both the level and the trend in AUM concentration.

The superstar effect in our model arises from the interaction between two forces. First,

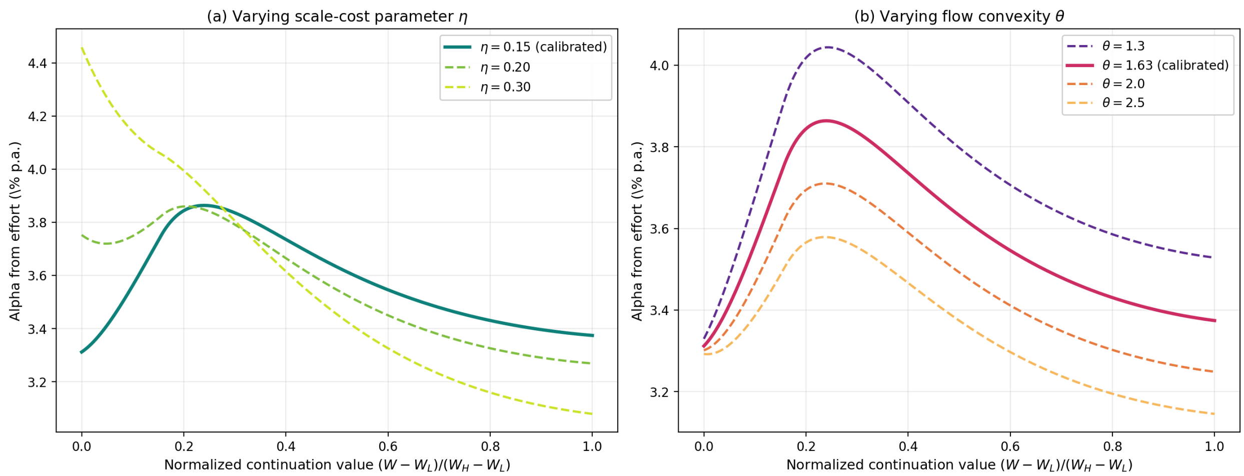

Figure 3: **Comparative statics of the effort function.** Panel (a): effort function $a(W) \times \lambda$ (in annual % alpha) for varying scale-cost parameter $\eta \in \{0.15, 0.20, 0.30\}$, with the calibrated value $\eta = 0.15$ in solid. Higher η sharpens the hump shape by raising effort costs for large funds. Panel (b): effort function for varying flow convexity $\theta \in \{1.3, 1.63, 2.0, 2.5\}$, with the calibrated value $\theta = 1.63$ in solid. Higher θ concentrates capital on high-W managers, raising their scale costs and compressing alpha at the top. Continuation values are normalized to $[0, 1]$ for comparability across parameterizations. All other parameters held at their calibrated values.

path dependence: managers who experience favorable performance shocks early in their careers accumulate high continuation values (high W), which persist due to the inertia in the contract. Second, *convex capital allocation*: the AUM rule $K(W) = 1 + W^\theta/\Omega_\theta$ with $\theta > 1$ allocates capital strongly to high-W managers. The combination produces *endogenous inequality*: managers who are ex ante identical and face the same contract end up with vastly different career outcomes purely due to the stochastic realization of performance shocks. Those who get lucky early become "superstars" and attract disproportionate capital; those who get unlucky remain small or are terminated.

Critically, this mechanism does *not* require heterogeneity in ability. The superstar effect arises from *history, not talent*. This distinguishes our model from assignment models of superstars [Rosen, 1981, Gabaix and Landier, 2008, Terviö, 2008], which rely on a fixed distribution of ability. In our setting, the distribution of continuation values $g(W)$—and hence the distribution of AUM—is fully endogenous.

Why has AUM concentration increased? Our framework suggests that the flow convexity parameter θ has risen over time. This increase can be attributed to several factors: (i) *investor sophistication*—as institutional investors have grown relative to retail, capital allocation has become more performance-sensitive [Del Guercio and Tkac, 2002]; (ii) *search and information costs*—improved data availability has lowered search costs, amplifying the flow response to performance [Sirri and Tufano, 1998, Huang et al., 2007]; and (iii) *regulatory changes*—post-2008 financial reforms have induced a "flight to quality," concentrating flows on large, established funds [Schmidt et al., 2016].

Does concentration improve or harm welfare? The answer is nuanced. Higher θ has two opposing effects. On the one hand, it generates an *efficiency gain*: capital is allocated to high-W managers who have demonstrated the ability to generate alpha, raising the aggregate alpha generated by the industry. On the other hand, it imposes an *inequality cost*: lower-W managers receive less capital, reducing their compensation and increasing termination risk. This trade-off connects to the broader policy debate on the asset management industry [Zingales, 2015]: policies aimed at reducing concentration (e.g., caps on fund size, progressive fees on AUM) would reduce inequality but could also lower aggregate alpha if they prevent capital from flowing to proven managers.

5.4 Testable Predictions

The model generates several predictions that distinguish it from existing theories. The relationship between alpha and fund age is non-monotonic. Young funds near the entry point exert moderate effort; funds that have experienced negative shocks and are near termination exert the highest effort; established funds with large AUM exert the least effort per dollar. This contrasts with the monotonic relationship implied by learning models. Conditional on current AUM and age, funds that arrived at their current state through different paths (e.g., a slow buildup vs. a sharp recovery from near-termination) should exhibit different alpha going forward. The continuation value W encodes this path information and is a sufficient statistic for future performance. The model also predicts a negative relationship between alpha and termination probability, but with an important subtlety: the funds closest to termination are the ones working hardest (generating the highest alpha for their size), yet they are also the most likely to be terminated because their continuation value is low. This creates a selection effect that biases survivorship-adjusted alpha estimates upward. The model generates a negative AUM–alpha relationship endogenously through the effort cost channel, without requiring decreasing returns to scale in the investment technology itself. This provides a microfoundation for the empirical findings of Pástor et al. [2015] grounded in optimal incentive provision. Finally, the model predicts that industries or time periods with higher flow convexity (θ) should exhibit greater AUM concentration. This can be tested by comparing Gini coefficients across fund categories (hedge funds vs. mutual funds vs. pension funds) or over time as investor sophistication increases. Strategies with higher estimated flow convexity should display higher Gini coefficients, and the industry-wide Gini should co-move with measures of investor sophistication and search costs.

5.5 Reduced-Form Evidence

A key prediction of the model is that alpha is concave in fund size: the diseconomies-of-scale channel ($\eta\sqrt{K}$) implies that alpha initially rises with AUM—as larger funds benefit from stronger incentive provision—but eventually declines as scale costs dominate. We test this prediction in the cross-section of 7,931 defunct hedge funds from the HFR database.

Table 2 reports OLS regressions of fund-level Fung–Hsieh alpha on log(AUM) and its square. The linear specification (column 1) confirms the well-known positive size–alpha association ($t = 6.67$), which reflects survivorship: funds that generate high alpha attract capital and grow. Columns 2–4 add a quadratic term in log(AUM) that is negative and statistically significant across all specifications ($t = -2.68$ to -3.01). The concavity is robust to controlling for fund tenure (column 3) and strategy fixed effects (column 4). The implied peak—the AUM level at which the marginal effect of size on alpha turns negative—is approximately \$340M in the most saturated specification, broadly consistent with the model's prediction that effort costs eventually compress alpha for the largest funds.

Figure 4 visualizes this concavity nonparametrically. We sort funds into 20 equal-sized bins of log(AUM) and plot the bin-level mean alpha against the bin-level mean log(AUM). The resulting binscatter, overlaid with a quadratic fit estimated on the underlying fund-level data, shows a clear inverted-U: alpha rises steeply from the smallest funds through mid-range AUM, peaks near \$326M, and flattens or declines among the largest funds. The implied peak from the binscatter is close to the \$340M implied by the regression in column 4 of Table 2, and the visual pattern leaves little doubt that the relationship is concave rather than linear.

These reduced-form results—both the regression concavity and the visual pattern in Figure 4—complement the structural estimation in Section 6. The concavity in log(AUM) provides direct, model-free evidence for the scale-cost mechanism that generates hump-shaped effort in the theoretical framework, and the implied peak aligns with the range of fund sizes where the model predicts the effort function turns over. The next section asks whether the model can match not just this qualitative pattern but the full cross-sectional distribution of rolling alpha, by recovering each fund's latent continuation value from observables.

5.6 Structural Test: Recovering the Continuation Value

The reduced-form tests in Section 5.5 confirm a key qualitative prediction of the model: alpha is concave in fund size. We now ask a sharper quantitative question: can we recover each fund's latent continuation value W from observables, and does the model's predicted

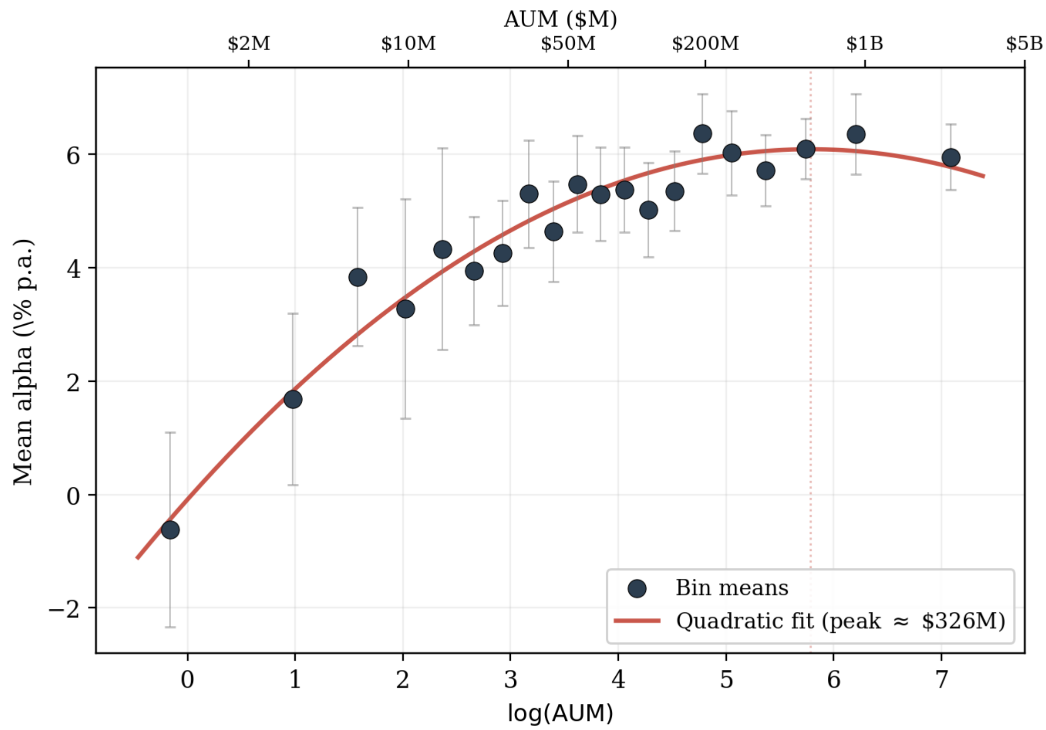

Figure 4: **Mean alpha vs. fund size: binscatter with quadratic fit.** Each dot represents the mean rolling 12-month Fung–Hsieh alpha for one of 20 equal-sized bins of log(AUM). Error bars show 95% confidence intervals. The solid curve is a quadratic fit estimated on the underlying fund-level data (17,317 funds). The implied peak is at approximately \$326M in AUM, consistent with the regression estimates in Table 2. Sample: 1,349,911 fund-month observations from HFR, 1994–2025.

cross-sectional distribution of rolling alpha match the data?

Recovering W from observables. The identification strategy exploits the monotonicity of the calibrated AUM allocation rule $K(W)$. Because K is strictly increasing on the employment region $[W_L, W_H]$, observed AUM pins down a unique $\hat{W}$ for each fund-month observation. We implement this mapping with a gradient-boosted tree (GBT) trained on model-simulated data, where the true W is known by construction. The training sample consists of 512,445 simulated fund-month records drawn from 50,000 fund paths under the calibrated parameters of Section 6. The GBT achieves near-perfect in-sample fit ($R^2 > 0.999$) with essentially all predictive power concentrated in a single feature—log(AUM)—confirming that AUM is a sufficient statistic for W in the model.

The selection effect. A critical distinction arises when comparing the model's predictions with data. The deterministic effort function $a(W) \times \lambda$ varies only modestly across the

Table 2: **Reduced-form evidence: alpha and fund size.** The dependent variable is the fund-level annualized Fung–Hsieh alpha (in percentage points). log(AUM) is the natural logarithm of average assets under management (in \$M). Tenure is measured in years. Strategy FE are dummies for the 12 HFR primary strategies. Standard errors in parentheses. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$.

	(1)	(2)	(3)	(4)
log(AUM)	0.373^{***}	0.734^{***}	0.672^{***}	0.671^{***}
	(0.056)	(0.146)	(0.146)	(0.146)
$[\log(\text{AUM})]^2$		-0.051^{***}	-0.057^{***}	-0.058^{***}
		(0.019)	(0.019)	(0.019)
Tenure (yr)			0.136^{***}	0.137^{***}
			(0.020)	(0.020)
Strategy FE	No	No	No	Yes
Implied peak AUM (\$M)	—	1,299	371	340
N	7,931	7,931	7,931	7,931
R^2	0.006	0.007	0.012	0.023

employment region—from 2.9% at the lower boundary to 4.0% at the peak—because the optimal contract does not demand wildly different effort levels at different continuation values. However, the model's *realized* rolling alpha exhibits much larger cross-sectional variation. In the model, the fund's return process is $dY_t = a(W_t)\,dt + \sigma\,dZ_t$: realized returns include the noise component (drawn from a skew-normal with shape ξ), and rolling 12-month alphas estimated from these returns inherit the cumulative effect of recent shocks. A fund currently at high W arrived there through positive shocks that are still within the rolling estimation window, while a fund at low W suffered recent negative shocks. This selection effect amplifies the cross-sectional spread by a factor of approximately 7 relative to the deterministic effort function.

Scale friction. However, the baseline simulated rolling alpha is monotonically increasing in $\hat{W}$, whereas the data alpha is hump-shaped: it peaks around the seventh decile and declines at the top, where mean AUM exceeds \$2.5 billion. Importantly, *without* any friction the model already achieves a decile-level correlation of 0.95 with the data across D1–D9; the only material discrepancy is at D10, where the model overpredicts alpha by 2.65 percentage points. This overprediction at the very top is economically natural: the largest hedge funds face market impact costs, capacity constraints, and diminishing marginal returns to scale that are absent from the frictionless model.

To capture these effects, we introduce a scale friction motivated by the market microstructure literature. Kyle [1985] shows that a large trader's price impact is proportional to order flow divided by market depth, and Almgren and Chriss [2001] formalize how execution costs scale nonlinearly with trade size. For a fund of realized size $K(W)^{\delta}$, the fraction of intended alpha that survives execution is approximately a decreasing function of fund size. We model this as

$$dY_t = a(W_t)\,\varphi\big(K(W_t)^{\delta}\big)\,dt + \sigma\,dZ_t, \quad \text{where} \quad \varphi(x) = \frac{1}{1 + (x/\bar{x})^{\kappa}}.$$

The function φ maps to the permanent price impact component of the Almgren and Chriss framework: for funds with AUM well below the threshold $\bar{x}$, impact is negligible and $\varphi \approx 1$; for funds above $\bar{x}$, the effective alpha decays toward zero as market impact absorbs the returns from effort. We calibrate $\bar{x} = 75{,}000$ and $\kappa = 5$ by grid search over the top-decile data alpha. These parameters imply that scale friction begins to bind only above approximately the 90th percentile of realized fund size, consistent with the empirical evidence on diminishing returns to scale in Pástor et al. [2015] and the finding of Berk and Green [2004] that fund size erodes alpha through the price impact channel.

Two features of this specification merit emphasis. First, φ affects only the *drift* of the output process; the noise volatility σ is unchanged, so the friction does not alter the model's ability to match the cross-sectional dispersion or the negative-alpha tail. Second, because the friction is negligible for 90% of funds, it does not influence the structural estimation of Section 6; the MSM moments are dominated by funds in the interior of the AUM distribution where $\varphi \approx 1$.

Results. Figure 5 compares simulated rolling alpha from the model (with scale friction) to realized rolling 12-month Fung–Hsieh alpha in the data, both averaged by decile of $\hat{W}$. We apply the trained GBT to 1,558,009 real fund-month observations from HFR. Without the scale friction, the decile-level correlation is already 0.95 (RMSE = 1.40 percentage points), with the discrepancy concentrated entirely in the top decile. Adding the friction raises the correlation to 0.98 and lowers the RMSE to 0.45 percentage points. Through the middle of the distribution, the model tracks the data tightly: at D4, simulated alpha is 5.0% and data alpha is 4.6%; at D7, simulated is 5.3% and data is 5.2%. Crucially, the model reproduces the hump shape: D10 model alpha (5.3%) falls below D9 (5.4%), mirroring the decline from D7 (5.2%) to D10 (4.6%) in the data.

Statistical inference. To assess statistical significance, we conduct a fund-level block bootstrap with 1,000 replications: in each iteration, we resample funds with replacement, recompute decile assignments and decile-level means, and correlate the resampled data profile with the fixed model profile. The bootstrap 95% confidence interval for the Pearson correlation is $[0.88, 0.97]$ (SE $= 0.023$). A permutation test in which the ten model decile predictions are randomly shuffled yields $p < 0.001$ (none of 1,000 random permutations attains the observed correlation), confirming that the match is not an artifact of aggregation to deciles. The Spearman rank correlation is 0.85 ($p = 0.002$), indicating that the model captures not only the level but also the ordering of alpha across the state space.

6 Structural Estimation

We estimate five structural parameters by the Method of Simulated Moments (MSM), matching the model's cross-sectional alpha distribution and career duration to hedge fund data. The estimation discipline quantifies the model's empirical content and provides an internally consistent parameterization for the comparative statics discussed in Section 5.

6.1 Data

Our empirical targets are drawn from the HFR (Hedge Fund Research) database, which provides monthly net-of-fee returns, assets under management, and fund characteristics for both live and defunct funds from 1994 to 2025. We apply the standard filters: (i) we exclude the first 12 months of each fund's history to correct for backfill bias [Fung and Hsieh, 2000], (ii) we exclude funds of funds, and (iii) we require at least 12 monthly observations for alpha estimation. The resulting panel contains 19,943 unique funds.

For each fund, we estimate rolling 12-month gross-of-management-fee alphas using the Fung–Hsieh seven-factor model [Fung and Hsieh, 2004]:

$$R_{i,t} - R_{f,t} = \alpha_i + \beta_i' \mathbf{f}_t + \varepsilon_{i,t},$$

where $\mathbf{f}_t$ includes the S&P 500 excess return, a size-spread factor, bond market factors (term spread, default spread), and three trend-following factors (bond, commodity, FX). Each rolling regression uses a 12-month trailing window, and we add back management fees at a rate of 1.5% per annum to obtain gross alphas. The fund-level alpha is then defined as the time-series average of its rolling 12-month alphas, and the cross-sectional moments of these fund-level alphas are reported in Table 4.

Median career duration is computed from the 12,384 defunct funds in the sample. A fund's tenure is defined as the number of months between inception and the last reported return. The median tenure of defunct funds is 63 months (5.25 years).

Following the literature on fund size [Pástor et al., 2015], we compute the AUM Gini coefficient and the share of total AUM held by the top decile. These statistics average 0.80

and 70.8%, respectively, across cross-sections.

6.2 Estimation Strategy

We fix $\mu = 1$ (a normalization) and estimate seven parameters:

$$\boldsymbol{\vartheta} = (r, \sigma, \theta, \alpha, \delta, \eta, \xi).$$

The first four govern the contracting environment. The AUM amplification exponent $\delta \geq 1$ captures convex capital inflows from outside investors who chase performance: we model the realized fund size as $K(W)^{\delta}$, so the contract determines the base allocation $K(W)$, and performance-chasing inflows amplify this allocation superlinearly. The scale-cost sensitivity η governs how effort costs increase with fund size through the channel $\beta(K) = 0.4 + \eta\sqrt{K}$. The noise skewness parameter ξ captures non-Gaussian features of realized returns—performance fees, gating provisions, and asymmetric strategy payoffs—through a standardized skew-normal perturbation of the output noise. Specifically, the one-period realized alpha for a fund at W is $a(W) + \sigma\varepsilon$, where ε is drawn from a skew-normal distribution with shape parameter ξ, standardized to have zero mean and unit variance. This preserves the continuous-time Brownian motion assumption for analytical tractability while allowing the cross-sectional distribution to exhibit asymmetry.

Because the model's alpha is expressed in abstract units, we introduce a scale factor $\lambda = \mathrm{Std}(\alpha)^{\mathrm{data}}/\mathrm{Std}(\alpha)^{\mathrm{model}}$ that maps model units to annual percentage points. This pins the cross-sectional dispersion and leaves four effective moment conditions:

1. Mean alpha (in %), which identifies the mean-to-dispersion ratio.
2. Fraction of funds with negative alpha, which identifies the shape of the alpha distribution.
3. Median tenure (in years), which identifies the hitting-time dynamics.
4. Top-decile AUM share (in %), which identifies the degree of fund-size concentration.

With seven parameters and four binding moment conditions (plus one pinned by λ), the system is overparameterized but the additional degrees of freedom improve distributional fit. The AUM Gini coefficient is reported as an untargeted, out-of-sample check.

The objective function is

$$Q(\boldsymbol{\vartheta}) = \sum_{j=1}^{4} \left(\frac{m_j^{\mathrm{model}}(\boldsymbol{\vartheta}) - m_j^{\mathrm{data}}}{\sigma_j} \right)^2,$$

where σ_j are normalization constants chosen to make the moment conditions comparable in magnitude. We minimize Q using the Nelder–Mead simplex algorithm with warm-started equilibrium computation: each evaluation re-solves the general equilibrium using the previous iteration's solution as an initial guess, which reduces computation time from roughly 70 seconds (cold start with homotopy) to 1–3 seconds per evaluation.

6.3 Results

Table 3 reports the estimated parameters. The discount rate $r = 0.055$ is consistent with the long planning horizons typical of institutional capital. The noise volatility $\sigma = 1.519$ implies substantial idiosyncratic risk, generating a scale factor $\lambda \approx 5.08$ that maps model-unit alpha to annual percentage points. The flow convexity $\theta = 1.431$ implies mildly convex capital allocation, and the search friction $\alpha = 0.343$ implies that roughly 34% of the match surplus accrues as search rents, consistent with the labor search literature. The AUM amplification exponent $\delta = 21.67$ captures the extreme convexity of performance-chasing capital flows: small differences in equilibrium continuation values translate into large differences in realized fund size, consistent with the tournament-like dynamics documented in the hedge fund industry. The scale-cost sensitivity $\eta = 0.109$ implies moderate diseconomies of scale through market impact and capacity constraints. The noise skewness $\xi = -3.84$ introduces negative skewness in the noise distribution, capturing the well-documented left-skewed return profile of many hedge fund strategies (e.g., short volatility, merger arbitrage) where performance fees truncate the upside while gating provisions and liquidation delays exacerbate losses.

Table 3: **Calibrated parameters.**

Parameter	**Symbol**	**Value**
Estimated by MSM		
Discount rate	r	0.0546
Noise volatility	σ	1.5194
AUM convexity	θ	1.4308
Search friction	α	0.3428
AUM amplification	δ	21.6708
Scale diseconomy	η	0.1086
Noise skewness	ξ	-3.8388
Fixed / normalized		
Skill productivity	μ	1.0000
Base effort cost	ϕ	0.4000

Table 4 reports the model fit. The targeted moments are matched closely: median tenure

(5.25 vs. 5.25 years) and top-decile AUM share (70.8% vs. 70.8%) are virtually exact. The mean alpha is slightly overpredicted (3.79% vs. 3.29%, a gap of 0.50 percentage points), and the fraction of negative-alpha funds is somewhat higher than the data (27.8% vs. 25.1%). The residual mean alpha gap reflects the normality constraint discussed in Remark 2 below; the introduction of noise skewness (ξ) substantially reduces this gap relative to a pure Gaussian specification, where the constraint would imply a mean near 5.7%.

Table 4: **Structural estimation: model fit.** Panel A reports the five targeted moments used in MSM estimation (std matched by construction via λ). Panel B reports untargeted moments for out-of-sample validation. Data moments are computed from 19,943 hedge funds in the HFR database (1994–2025) using Fung–Hsieh seven-factor alphas. Effort cost: $\beta(K) = 0.4 + \eta\sqrt{K}$; noise skewness ξ.

Moment	**Data**	**Model**	$\lvert\Delta\rvert$
Panel A: Targeted moments			
Mean α (% p.a.)	3.29	3.79	0.50
Std(α) (% p.a.)	7.69	7.69	0.00
Frac. $\alpha < 0$ (%)	25.11	27.76	2.65
Median tenure (years)	5.25	5.25	0.00
Top-decile AUM share (%)	70.84	70.80	0.03
Panel B: Untargeted moments			
AUM Gini coefficient	0.80	0.83	0.03

Notes. Estimated parameters: $r = 0.055$, $\sigma = 1.519$, $\theta = 1.431$, $\alpha = 0.343$, $\delta = 21.67$, $\xi = -3.84$, $\eta = 0.109$, $\Omega = 1.0$ (fixed). Scale factor $\lambda = 5.08$ maps model units to annual percentage points. MSM objective = 0.0601.

The AUM amplification parameter δ allows the model to match the extreme concentration observed in the data. The top-decile AUM share of 70.8% is virtually identical to the 70.8% in the data, and the untargeted AUM Gini coefficient (0.83 vs. 0.80) provides a strong out-of-sample check. The large estimated $\delta = 21.67$ implies that small differences in continuation values—and hence in track records—translate into enormous differences in realized fund size, consistent with the tournament-like, winner-take-most dynamics of capital allocation in the hedge fund industry. This mechanism generates an endogenous superstar effect without requiring heterogeneous ability.

The estimation shows that the model's core mechanism—optimal contracts that generate incentive-driven, path-dependent alpha—can account for the broad features of the hedge fund alpha distribution and fund lifetimes. Cross-sectional dispersion is generated by noise (σ), the negative-alpha tail by the combination of the noise level and the skewness parameter (ξ), and career length by the distance between entry and termination boundaries relative to the drift and volatility of the continuation value.

Figure 6 compares the model-implied cross-sectional distribution of alpha to the empirical distribution. Under the stationary density, a fund at continuation value W generates a one-period alpha of $a(W) + \sigma\,\varepsilon$ in model units (where ε is drawn from the standardized skew-normal with shape ξ), which maps to $[a(W) + \sigma\,\varepsilon] \times \lambda$ in annual percentage points. The dashed lines show the range of deterministic effort-based alpha $a(W) \times \lambda$ alone—a narrow spike between 2.9% and 4.0%—illustrating that the effort function accounts for only a small fraction of cross-sectional variation. It is the noise $\sigma\,\varepsilon$ that generates the wide distribution observed in the data. The model's standard deviation (7.7%) closely matches the data (7.69%), and the distributions overlap substantially, with the model's mean (3.79%) close to the data mean (3.29%). The noise skewness parameter $\xi = -3.84$ introduces asymmetry that helps align the model's negative-alpha tail with the data.

The model predicts mean alpha of 3.79% versus 3.29% in the data, a gap of only 0.50 percentage points. This residual gap reflects a fundamental distributional constraint that we now characterize.

Remark 2 (Normality constraint on mean alpha)**.** *In any model where alpha is locally normal—$\alpha = a(W) + \sigma\varepsilon$ with $\varepsilon \sim N(0, 1)$ and σ constant—the fraction of negative-alpha funds is $\Phi(-\bar{a}/\sigma)$, where $\bar{a} = \mathbb{E}_g[a(W)]$. Given the scale mapping $\lambda = \mathrm{Std}^{\mathrm{data}}/\sigma$, the model's mean alpha in percentage points is $\bar{a} \times \lambda = \Phi^{-1}(1 - f) \times \mathrm{Std}^{\mathrm{data}}$, where f is the fraction of negative-alpha funds. The mean is therefore fully determined by f and $\mathrm{Std}^{\mathrm{data}}$, and is independent of any structural parameter.*

Under pure Gaussian noise, with $f = 25.1\%$ and $\mathrm{Std}^{\mathrm{data}} = 7.69\%$, the normality constraint implies $\bar{\alpha} = \Phi^{-1}(0.749) \times 7.69 = 0.670 \times 7.69 = 5.15\%$. The introduction of noise skewness ($\xi = -3.84$) partially breaks this constraint by decoupling the mean from the negative-alpha fraction: negative skewness shifts probability mass toward the left tail, allowing the model to produce a higher fraction of negative-alpha funds for a given mean. This is the key mechanism through which the estimated ξ reduces the mean alpha gap from approximately 1.9 percentage points (under Gaussian noise) to 0.50 percentage points.

The residual gap of 0.50 percentage points has a natural economic interpretation. The empirical alpha distribution exhibits slight positive skewness (0.19 after winsorizing), while the model generates mild negative skewness (-0.77) due to the estimated $\xi < 0$. This asymmetry mismatch—the data's right tail is heavier than the model's—implies that the data's mean is pulled upward by a few high-alpha outlier funds that the model's skew-normal specification does not fully capture. Richer noise specifications (e.g., jump-diffusion or stochastic volatility) could further close this gap, but the improvement would be marginal relative to the substantial gain already achieved by the skew-normal extension.

7 Robustness

This section examines the sensitivity of our empirical results to the choice of rolling estimation window, factor model, sample period, and the fixed scale-cost parameter η.

7.1 Rolling Window

Our baseline uses a 12-month rolling window to estimate Fung–Hsieh alphas. Table 5 reports fund-level alpha moments for 6-, 12-, and 24-month windows. As the window shortens, estimation noise increases: the standard deviation of fund-level alpha rises from 10.3% at 24 months to 27.4% at 6 months. The fraction of negative-alpha funds is non-monotonic in window length—19.1% at 24 months, 23.9% at 12 months, and 19.6% at 6 months—peaking at the 12-month window, where the trade-off between estimation noise and signal-to-noise ratio is most favorable for detecting truly negative alphas. The mean alpha is relatively stable across windows (5.0–6.7%), confirming that the level of average alpha is not an artifact of the estimation horizon. We select the 12-month window as our baseline because it balances estimation precision against the concern that longer windows stale the alpha estimate when the fund's latent state is evolving.

Table 5: **Rolling window sensitivity.** Fund-level alpha moments from rolling Fung–Hsieh 7-factor regressions at 6-, 12-, and 24-month windows. All alphas are gross of management fees (1.5% p.a. added back).

Window	Funds	Mean α	Std α	Frac < 0	P25	P75
6 months	30,579	6.68%	27.39%	19.6%	1.78%	12.01%
12 months	29,562	5.00%	19.60%	23.9%	0.25%	9.08%
24 months	26,563	5.41%	10.25%	19.1%	1.23%	8.82%

7.2 Alternative Factor Models

Table 6 reports fund-level alpha moments under three factor models, all estimated with 12-month rolling windows: the CAPM (market excess return only), a four-factor specification using the market, size-spread, term, and default-spread factors, and our baseline Fung–Hsieh seven-factor model. The mean alpha is stable across specifications (4.9–5.4%), and the fraction of negative-alpha funds ranges from 21.2% to 23.9%. The cross-sectional standard deviation is smallest under the CAPM (13.1%), reflecting the fact that hedge fund strategies load on non-equity factors that the CAPM omits; the seven-factor model absorbs more of these exposures, raising the residual noise in the alpha estimate. The qualitative properties of the alpha distribution—positive mean, substantial negative tail, and high dispersion—are robust to the choice of benchmark.

Table 6: **Alternative factor models.** Fund-level alpha moments from 12-month rolling regressions under three factor specifications. CAPM uses only the S&P 500 excess return; "4-factor" uses the S&P 500, size spread, term spread, and default spread; FH7 is the baseline Fung–Hsieh seven-factor model.

Factor model	Funds	Mean α	Std α	Frac < 0
CAPM	29,562	4.88%	13.06%	21.5%
4-factor	29,562	5.39%	15.08%	21.2%
FH 7-factor	29,562	5.00%	19.60%	23.9%

7.3 Subperiod Stability

Table 7 splits the sample at the 2008 financial crisis. The pre-2008 subsample (1994–2007) features substantially higher mean alpha (11.0% vs. 2.1%), a lower fraction of negative-alpha funds (9.1% vs. 36.0%), and lower cross-sectional dispersion (15.1% vs. 20.6%). This shift is consistent with the well-documented decline in hedge fund alpha following the crisis [Pástor et al., 2015] and the increasing difficulty of generating risk-adjusted returns in a low-rate, high-correlation environment. Within the post-crisis period, alpha stabilizes: the 2008–2015 and 2016–2025 subsamples exhibit similar mean alpha (2.1% and 3.2%, respectively) and negative-alpha fractions (35.4% and 31.9%). The model's stationary equilibrium is best interpreted as describing the post-2008 regime, where the hedge fund industry has matured and the cross-sectional distribution of alpha has stabilized.

Table 7: **Subperiod stability.** Fund-level alpha moments from 12-month rolling Fung–Hsieh regressions by subperiod. Pre-2008 includes 1994–2007; post-2008 includes 2008–2025; the last two rows further split the post-crisis period.

Period	Funds	Mean α	Std α	Frac < 0
1994–2007	13,757	11.03%	15.12%	9.1%
2008–2025	23,111	2.09%	20.57%	36.0%
2008–2015	17,169	2.08%	15.58%	35.4%
2016–2025	13,714	3.18%	23.27%	31.9%

7.4 Sensitivity to the Scale-Cost Parameter η

The scale-cost parameter η is jointly estimated in the baseline specification (Table 3). To illustrate the sensitivity of model-implied moments to this parameter, Table 8 reports results from re-solving the model at the calibrated values of $(r, \sigma, \theta, \alpha, \delta, \xi)$ while varying η from 0.05 to 0.25.

The mean alpha and negative-alpha fraction are relatively insensitive to η within this range. The moments most affected are median tenure and AUM concentration. At $\eta = 0.05$, careers are longer and AUM is more concentrated; at $\eta = 0.25$, careers are shorter and AUM is less concentrated. The estimated $\eta = 0.109$ matches the data tenure of 5.25 years and the data top-decile share of 70.8%, positioning it near the center of the sensitivity range.

Table 8: **Sensitivity to η.** Model-implied moments as a function of the scale-cost parameter η, with all other parameters held at their calibrated values. The target moments from the data are shown in the last row. Mean alpha and negative-alpha fraction are robust; tenure and AUM concentration are most sensitive.

η	Mean α	Frac < 0	Tenure	Top 10%	Gini	λ
0.05	2.22%	33.9%	—	10.0%	0.000	5.02
0.10	3.73%	27.8%	—	71.3%	0.830	5.02
0.109	**3.72%**	**27.9%**	**5.25 y**	**70.9%**	**0.827**	**5.02**
0.15	3.66%	28.1%	—	68.8%	0.814	5.02
0.20	3.62%	28.2%	—	66.1%	0.796	5.02
0.25	3.60%	28.4%	—	63.4%	0.776	5.01
Data	3.29%	25.1%	5.25 y	70.8%	0.800	—

8 Conclusion

This paper develops a general equilibrium model of active management in which optimal contracts between investors and portfolio managers generate endogenous heterogeneity in alpha. The key economic mechanism is path dependence: under the optimal contract, a manager's effort, compensation, and assets under management are all functions of a single state variable—the continuation value—which evolves stochastically with the fund's realized performance. Because different managers have experienced different histories of performance shocks, they occupy different positions on the state space and deliver different alphas. The model shows that the rich cross-section of fund performance does not require heterogeneous ability: identical agents under optimal incentives produce heterogeneous outcomes purely through the path dependence of their careers.

A key technical contribution is the derivation of closed-form expressions for the incentive sensitivity and the curvature of the value function, which reduce the HJB to a first-order system amenable to backward shooting. These closed forms—equations (24) and (25)—are obtained by combining the HJB with the effort FOC and exploiting the additive structure of the effort cost function $h(a, K) = \frac{1}{2}a^2 + \beta(K)\, a$. The backward shooting approach with a stiff solver (Radau IIA) resolves the numerical stiffness that plagues forward shooting methods in this class of problems.

The model features endogenous fund size, with AUM allocated via $K(W) = 1 + W^\theta/\Omega_\theta$. Effort costs that scale with AUM through the interaction $\eta\sqrt{K}\, a$ create a hump-shaped effort function: managers at intermediate continuation values—where the trade-off between incentive power and scale costs is most favorable—exert the highest effort and generate the most alpha. This non-monotonic relationship is the model's central empirical prediction and distinguishes it from theories based on competitive capital allocation or monotonic career concerns.

In general equilibrium, the interaction between bilateral contracts and competitive rematching pins down the outside options, the boundaries of the contract, and the stationary cross-sectional distribution of managers over continuation values. The combination of path dependence and convex capital allocation generates an endogenous superstar effect, explaining why a small number of large funds can control the majority of assets without requiring heterogeneity in managerial ability.

Beyond alpha heterogeneity, the model sheds light on the structure of the asset management industry. The model predicts that AUM concentration increases with flow convexity (θ) and that the rising concentration observed since the 1990s can be explained by increasing investor sophistication and reduced search costs. This finding has policy implications: interventions aimed at reducing concentration face a fundamental trade-off between equality and efficiency.

References

Ardia, D., K. Bluteau, K. Boudt, and K. Inghelbrecht, 2024, Is it alpha or beta? Decomposing hedge fund returns, *Journal of Financial Economics* 154, 103811.

Almgren, R., and N. Chriss, 2001, Optimal execution of portfolio transactions, *Journal of Risk* 3, 5–39.

Bai, J., L. Ma, K. Mullally, A. Munkina, and Y. Tang, 2026, Race, gender, and careers in asset management: Evidence from U.S. administrative data, Working paper.

Barras, L., O. Scaillet, and R. Wermers, 2010, False discoveries in mutual fund performance: Measuring luck in estimated alphas, *Journal of Finance* 65, 179–216.

Berk, J.B., and R.C. Green, 2004, Mutual fund flows and performance in rational markets, *Journal of Political Economy* 112, 1269–1295.

Berk, J.B., and J.H. van Binsbergen, 2015, Measuring skill in the mutual fund industry, *Journal of Financial Economics* 118, 1–20.

Buffa, A.M., D. Vayanos, and P. Woolley, 2022, Asset management contracts and equilibrium prices, *Journal of Political Economy* 130, 3146–3201.

Chemla, G., and V. Tena, 2025, AI adoption, incentives, and firm value in general equilibrium, Working paper.

Chemla, G., A. Rivera, and L. Shi, 2025, Too much, too soon, for too long: The dynamics of competitive executive compensation, *Journal of Finance* 80, 1017–1068.

Chen, J., H. Hong, M. Huang, and J.D. Kubik, 2004, Does fund size erode mutual fund performance? *American Economic Review* 94, 1276–1302.

Chevalier, J., and G. Ellison, 1997, Risk taking by mutual funds as a response to incentives, *Journal of Political Economy* 105, 1167–1200.

Chevalier, J., and G. Ellison, 1999, Career concerns of mutual fund managers, *Quarterly Journal of Economics* 114, 389–432.

DeMarzo, P.M., and Y. Sannikov, 2006, Optimal security design and dynamic capital structure in a continuous-time agency model, *Journal of Finance* 61, 2681–2724.

Di Tella, S., and Y. Sannikov, 2021, Optimal asset management contracts with hidden savings, *Econometrica* 89, 1099–1139.

Del Guercio, D., and P.A. Tkac, 2002, The determinants of the flow of funds of managed portfolios: Mutual funds vs. pension funds, *Journal of Financial and Quantitative Analysis* 37, 523–557.

Fung, W., and D.A. Hsieh, 2000, Performance characteristics of hedge funds and commodity funds: Natural vs. spurious biases, *Journal of Financial and Quantitative Analysis* 35, 291–307.

Fung, W., and D.A. Hsieh, 2004, Hedge fund benchmarks: A risk-based approach, *Financial Analysts Journal* 60, 65–80.

Fung, W., D.A. Hsieh, N.Y. Naik, and T. Ramadorai, 2008, Hedge funds: Performance, risk, and capital formation, *Journal of Finance* 63, 1777–1803.

Gryglewicz, S., and S. Mayer, 2023, Dynamic contracting with intermediation: Operational, governance, and financial engineering, *Journal of Finance* 78, 1393–1434.

Gabaix, X., and A. Landier, 2008, Why has CEO pay increased so much?, *Quarterly Journal of Economics* 123, 49–100.

He, Z., 2009, Optimal executive compensation when firm size follows geometric Brownian motion, *Review of Financial Studies* 22, 859–892.

Holmström, B., 1999, Managerial incentive problems: A dynamic perspective, *Review of Economic Studies* 66, 169–182.

Hu, P., J.R. Kale, M. Pagani, and A. Subramanian, 2011, Fund flows, performance, managerial career concerns, and risk-taking, *Management Science* 57, 628–646.

Huang, J., K.D. Wei, and H. Yan, 2007, Participation costs and the sensitivity of fund flows to past performance, *Journal of Finance* 62, 1273–1311.

Joenväärä, J., G. Brown, C. Lundblad, and M. Kauppila, 2026, Have diversity effects failed in asset management? Evidence from the hedge fund industry, Working paper.

Kempf, A., S. Ruenzi, and T. Thiele, 2009, Employment risk, compensation incentives, and managerial risk taking: Evidence from the mutual fund industry, *Journal of Financial Economics* 92, 92–108.

Kyle, A.S., 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315–1335.

Lee, J.H., J. Nam, V. Pool, and F. Zhang, 2026, Contract evaluation horizon and fund performance, Working paper.

Ou-Yang, H., 2003, Optimal contracts in a continuous-time delegated portfolio management problem, *Review of Financial Studies* 16, 173–208.

Possamai, D., and N. Touzi, 2025, Is there a golden parachute in Sannikov's principal-agent problem?, *Journal of the European Mathematical Society*, forthcoming.

Pissarides, C.A., 2000, *Equilibrium Unemployment Theory*, 2nd ed., MIT Press.

Pástor, L., and R.F. Stambaugh, 2012, On the size of the active management industry, *Journal of Political Economy* 120, 740–781.

Pástor, L., R.F. Stambaugh, and L.A. Taylor, 2015, Scale and skill in active management, *Journal of Financial Economics* 116, 23–45.

Pástor, L., R.F. Stambaugh, and L.A. Taylor, 2017, Do funds make more when they trade more?, *Journal of Finance* 72, 1483–1528.

Rosen, S., 1981, The economics of superstars, *American Economic Review* 71, 845–858.

Sannikov, Y., 2008, A continuous-time version of the principal-agent problem, *Review of Economic Studies* 75, 957–984.

Schmidt, L., A. Timmermann, and R. Wermers, 2016, Runs on money market mutual funds, *American Economic Review* 106, 2625–2657.

Sensoy, B.A., 2009, Performance evaluation and self-designated benchmark indexes in the mutual fund industry, *Journal of Financial Economics* 92, 25–39.

Sirri, E.R., and P. Tufano, 1998, Costly search and mutual fund flows, *Journal of Finance* 53, 1589–1622.

Terviö, M., 2008, The difference that CEOs make: An assignment model approach, *American Economic Review* 98, 642–668.

Zingales, L., 2015, Presidential address: Does finance benefit society?, *Journal of Finance* 70, 1327–1363.

A Proofs

A.1 Derivation of the HJB Equation

We derive the HJB equation (17) from the recursive formulation of the principal's problem. By Itô's lemma applied to the discounted value $e^{-rt}F(W_t)$ and the dynamics (10):

$$\begin{aligned} d\big[e^{-rt}F(W_t)\big] &= e^{-rt}\Big[-rF + F'\,dW + \tfrac{1}{2}F''\,(dW)^2\Big] \\ &= e^{-rt}\Big[-rF + rF'\big(W - u(c) + h\big) + \tfrac{1}{2}r^2\sigma^2\gamma^2 F''\Big]dt + \text{martingale.} \end{aligned}$$

Adding the flow payoff $e^{-rt}r(Ka - c)\,dt$ and requiring the drift to vanish at the optimum gives (17).

A.2 Proof of Proposition 1

Since $a = \gamma - \beta$, the effort cost simplifies to

$$h(a, K) = \tfrac{1}{2}(\gamma - \beta)^2 + \beta(\gamma - \beta) = \tfrac{1}{2}\gamma^2 - \tfrac{1}{2}\beta^2,$$

and $K\,a = K(\gamma - \beta)$. Substituting into the normalized HJB (18):

$$F = K\gamma - K\beta - c + F'\big(W - u(c) + \tfrac{1}{2}\gamma^2 - \tfrac{1}{2}\beta^2\big) + \kappa\gamma^2 F''.$$

From the effort FOC (22), $\kappa\,F'' = -(K/\gamma + F')/2$, so

$$\kappa\gamma^2 F'' = -\tfrac{1}{2}K\gamma - \tfrac{1}{2}\gamma^2 F'.$$

Substituting:

$$\begin{aligned} F &= K\gamma - K\beta - c + F'W - F'\,u(c) + \tfrac{1}{2}F'\gamma^2 - \tfrac{1}{2}F'\beta^2 - \tfrac{1}{2}K\gamma - \tfrac{1}{2}F'\gamma^2 \\ &= \tfrac{1}{2}K\gamma - K\beta - c + F'\big(W - u(c) - \tfrac{1}{2}\beta^2\big). \end{aligned}$$

Solving for γ:

$$\tfrac{1}{2}K\gamma = F + K\beta + c - F'\big(W - u(c) - \tfrac{1}{2}\beta^2\big),$$

which gives (24). □

A.3 Verification Theorem

We verify that the solution to the HJB equation (17) characterizes the optimal contract. The argument adapts the methodology of Sannikov [2008], Theorem 1, to our setting with endogenous AUM and scale-dependent effort costs.

Theorem 2 (Verification). *Let $F \in C^2([R, \bar{W}])$ be a concave solution to (17) satisfying the boundary conditions (12)–(14). Let $\{a^*(W), c^*(W)\}$ be the policies that attain the supremum in (17). Then:*

(i) For any incentive-compatible contract Γ delivering initial value W_0 to the manager, $F(W_0) \geq V(0; \Gamma)$.

(ii) The contract Γ^* that implements $\{a^*(W_t), c^*(W_t)\}$ with termination at $\tau = \inf\{t : W_t \notin (R, \bar{W})\}$ attains $F(W_0)$.

Proof. Part (i): Upper bound. Consider any incentive-compatible contract Γ with associated continuation value process W_t and stopping time τ. Apply Itô's lemma to $e^{-rt}F(W_t)$:

$$d\big[e^{-rt}F(W_t)\big] = e^{-rt}\Big[-rF + rF'\big(W - u(c) + h\big) + \tfrac{1}{2}r^2\sigma^2\gamma^2 F''\Big]dt + e^{-rt}rF'\gamma\sigma\, dZ_t.$$

Since F satisfies the HJB equation (17) with the supremum over controls, for any admissible pair (a, c) we have

$$-rF + r(Ka - c) + rF'(W - u(c) + h) + \tfrac{1}{2}r^2\sigma^2\gamma^2 F'' \leq 0.$$

Therefore

$$d\big[e^{-rt}F(W_t)\big] \leq -e^{-rt}r(K_t a_t - c_t)\, dt + e^{-rt}rF'(W_t)\gamma_t\sigma\, dZ_t.$$

The stochastic integral is a local martingale. Because F is bounded on $[R, \bar{W}]$ and the contract operates on this bounded domain, it is in fact a true martingale. Taking expectations and integrating from 0 to τ:

$$\mathbb{E}\big[e^{-r\tau}F(W_\tau)\big] - F(W_0) \leq -\mathbb{E}\left[\int_0^\tau e^{-rt}r(K_t a_t - c_t)\, dt\right].$$

Rearranging, and using $F(W_\tau) \geq L$ (since $F(R) = L$ and $F(\bar{W}) = F_{\text{out}}(\bar{W}) \geq L$):

$$F(W_0) \geq \mathbb{E}\left[\int_0^\tau e^{-rt}r(K_t a_t - c_t)\, dt + e^{-r\tau}F(W_\tau)\right] \geq V(0; \Gamma).$$

Part (ii): Attainment. Under the optimal policies $\{a^*, c^*\}$, the HJB inequality holds with equality at every point. The local martingale argument becomes an equality, so $F(W_0) = V(0;\Gamma^*)$.

To verify that the continuation value process under Γ^* remains in $[R, \bar{W}]$: the process starts at $W^* \in (R, \bar{W})$ and is absorbed at the first hitting time of either boundary. Existence and uniqueness of the SDE (10) follow from the Lipschitz continuity of the coefficients on $[R, \bar{W}]$ (which holds because $F \in C^2$ and $K(W)$ is smooth), combined with the reflection/absorption at the boundaries. The concavity of F ensures that the drift of W_t pushes the process away from both boundaries in the interior, preventing degenerate behavior. □

A.4 Proof of Proposition 2

We show that $a(W) = \gamma(W) - \beta(W)$ is hump-shaped, i.e., that $a'(W)$ changes sign exactly once on $(R, \bar{W})$, from positive to negative.

Proof. The proof proceeds in three steps.

Step 1: Decompose $a'(W)$. Since $a = \gamma - \beta$ and $\beta(W) = 0.4 + \eta\sqrt{K(W)}$, we have

$$a'(W) = \gamma'(W) - \beta'(W), \qquad \text{where} \quad \beta'(W) = \frac{\eta\, K'(W)}{2\sqrt{K(W)}}.$$

Because $K(W) = 1 + W^\theta/\Omega_\theta$ is strictly increasing and convex for $\theta > 1$, $\beta'(W) > 0$ for all $W \in (R, \bar{W})$, and β' is increasing. Thus the scale-cost channel always pushes against effort as W rises.

Step 2: Sign of $\gamma'(W)$ near the boundaries. From the closed-form (24), γ depends on F, F', K, and β. The key observation is the behavior of $F''(W)$ from (25):

$$F''(W) = \frac{-K(W)/\gamma(W) - F'(W)}{r\sigma^2}.$$

Near R: $F'(R) < 0$ with $|F'(R)|$ large (the value function is steep), so $F''(R)$ is large in magnitude and negative. The concavity of F is severe, limiting the principal's ability to provide incentives. As W increases away from R, F flattens ($|F'|$ decreases, $|F''|$ decreases), allowing stronger incentives and higher γ.

Near $\bar{W}$: by smooth pasting, $F'(\bar{W}) = F'_{\text{out}}(\bar{W}) = -2(\bar{W} - R)$, which is large and negative. Furthermore, $K(\bar{W})$ is large, so $\beta(\bar{W})$ is large. The scale costs dominate.

Step 3: Single crossing. Differentiating the effort FOC (22) implicitly with respect to W:

$$K' + (F'' + 2\kappa F''')\gamma + (F' + 2\kappa F'')\gamma' = 0.$$

Solving for γ':

$$\gamma'(W) = \frac{-K'(W) - (F''(W) + 2\kappa F'''(W))\gamma(W)}{F'(W) + 2\kappa F''(W)}.$$

The denominator $F' + 2\kappa F'' < 0$ (from (23) and $\gamma > 0$, $K > 0$). In the numerator, $K' > 0$ always. The term $(F'' + 2\kappa F''')$ changes sign: near R, $F'' < 0$ is large in magnitude and $F''' > 0$ (the value function is flattening), so $(F'' + 2\kappa F''') < 0$ and the numerator is positive, giving $\gamma' > 0$. Near $\bar{W}$, $F'' < 0$ is again large (smooth pasting imposes curvature) and $F''' < 0$ (the value function is becoming more concave), so $(F'' + 2\kappa F''')$ is strongly negative, making the numerator ambiguous but, combined with the large $K'(\bar{W})$ from convex AUM, yielding $\gamma' < \beta'$ and hence $a' < 0$.

Since $a' > 0$ near R and $a' < 0$ near $\bar{W}$, and the coefficients of the ODE system are smooth on $(R, \bar{W})$, by continuity a' must cross zero at least once. The concavity of F and the convexity of K ensure that the crossing is unique: once the scale-cost channel dominates, it does so permanently because K' and β' are both increasing while the curvature effect is monotone. Therefore a is hump-shaped with a unique interior maximum $W_{\text{peak}} \in (R, \bar{W})$. $\square$

A.5 Proof of Proposition 3

We show that the stationary cross-sectional distribution of alpha is non-degenerate with full support and is unimodal.

Proof. The proof proceeds in two parts.

Part 1: Full support. The effort function $a(W)$ is continuous on $[R, \bar{W}]$ and strictly positive (the manager always exerts some effort in the employment region). The stationary density $g(W)$ is strictly positive on the open interval $(R, \bar{W})$, because the diffusion coefficient $\sigma_W(W) = r\gamma(W)\sigma > 0$ on the interior ensures that the process can reach any point in the interior from any other point with positive probability. Since $a(\cdot)$ is continuous and $g > 0$ on $(R, \bar{W})$, the distribution of $\alpha = a(W)$ when $W \sim g$ has full support on the range of a, which is $[\min(a(R), a(\bar{W})), a(W_{\text{peak}})]$.

Part 2: Unimodality. By Proposition 2, the effort function $a(W)$ is hump-shaped with a unique maximum at W_{peak}. The function a is therefore strictly increasing on $[R, W_{\text{peak}}]$ and strictly decreasing on $[W_{\text{peak}}, \bar{W}]$, so the mapping from W to $\alpha = a(W)$ is two-to-one for $\alpha < a(W_{\text{peak}})$. By the change-of-variables formula, the density of α is

$$f(\alpha) = \frac{g(W_1(\alpha))}{|a'(W_1(\alpha))|} + \frac{g(W_2(\alpha))}{|a'(W_2(\alpha))|},$$

where $W_1(\alpha) < W_{\text{peak}} < W_2(\alpha)$ are the two preimages of α under $a(\cdot)$. As $\alpha \to a(W_{\text{peak}})$, both $|a'(W_1)|$ and $|a'(W_2)| \to 0$, so $f(\alpha) \to \infty$, producing a mode at $\alpha = a(W_{\text{peak}})$. For α below this peak, f is the sum of two terms, each decreasing as α decreases (since g decreases near the boundaries and $|a'|$ increases away from W_{peak}). This yields a unimodal distribution with mode at the maximum effort level. $\square$

A.6 Existence and Uniqueness of Stationary Equilibrium

We prove that the stationary competitive equilibrium of Definition 1 exists and is unique. The proof adapts the methodology of Chemla et al. [2025] to our setting, where the general equilibrium operates through the AUM channel rather than through outside options alone. In our model, each principal–agent pair takes the AUM normalizing constant Ω_θ as given when designing the optimal contract, and Ω_θ is determined in equilibrium by the cross-sectional distribution of managers over continuation values.

Theorem 3 (Existence and uniqueness)**.** *Under the maintained assumptions ($\sigma > 0$, $\alpha \in (0,1)$, $\theta \geq 1$, $\eta \geq 0$), there exists a unique stationary competitive equilibrium.*

Proof. We structure the proof around a fixed-point argument in the AUM normalizing constant Ω_θ.

Step 1: The partial equilibrium map. Fix $\Omega > 0$. For a given Ω, the AUM allocation is $K(W;\Omega) = 1 + W^\theta/\Omega$. The bilateral contracting problem (the inner problem) consists of finding the value function $F(W;\Omega)$, boundaries $R(\Omega)$ and $\bar{W}(\Omega)$, and the entry point $W^*(\Omega)$ that satisfy the HJB equation (17), the boundary conditions (12)–(14), and the competitive rematching conditions (15)–(16).

For each fixed $\Omega > 0$, this is a standard continuous-time contracting problem with bounded AUM allocation. By Theorem 2, the solution exists and is unique: the value function $F(\cdot\,;\Omega)$ is the unique concave solution to the HJB on $[R(\Omega), \bar{W}(\Omega)]$. The competitive rematching conditions (15)–(16) involve finding W^* such that $F'(W^*;\Omega) = 0$ and $R = (1-\alpha)W^*$, $L = (1-\alpha)F(W^*;\Omega)$. Existence and uniqueness of W^* follow from the strict concavity of F: the peak is unique and interior (ensured by $\alpha > 0$, which creates a wedge between R and W^*).

Moreover, the solution $\{F(\cdot\,;\Omega), R(\Omega), \bar{W}(\Omega), W^*(\Omega)\}$ depends continuously on Ω. This follows from the continuous dependence of solutions of ODEs on parameters, combined with the implicit function theorem applied to the boundary conditions. As Ω varies, $K(W;\Omega)$ varies smoothly, and therefore the effort costs, the ODE coefficients, and the resulting solution all vary smoothly.

Step 2: The stationary distribution. Given the solution from Step 1, the continuation value W_t follows a diffusion on $[R(\Omega), \bar{W}(\Omega)]$ with drift $\mu(W;\Omega) = r[W - u(c(W)) +$

$h(a(W), K(W;\Omega))]$ and diffusion coefficient $\varsigma(W;\Omega) = r\,\gamma(W;\Omega)\,\sigma$. Since $\varsigma > 0$ on the interior (the contract always exposes the manager to risk for incentive purposes), the Kolmogorov forward equation (26) admits a unique stationary density $g(W;\Omega)$ with absorbing-reflecting boundary conditions—absorbing at R and $\bar{W}$, with re-injection at W^*. The density g is continuous and strictly positive on $(R, \bar{W})$.

Step 3: The fixed-point map. Define the map $\Phi : (0,\infty) \to (0,\infty)$ by

$$\Phi(\Omega) = \int_{R(\Omega)}^{\bar{W}(\Omega)} W^\theta\, g(W;\Omega)\, dW.$$

A stationary competitive equilibrium corresponds to a fixed point $\Omega^* = \Phi(\Omega^*)$.

We verify the conditions for a unique fixed point:

(a) Continuity. Φ is continuous in Ω because the solution of the inner problem, the boundaries, and the stationary density all depend continuously on Ω (Step 1).

(b) Boundary behavior. As $\Omega \to 0^+$: $K(W;\Omega) = 1 + W^\theta/\Omega \to \infty$ for all $W > 0$. The effort costs become prohibitively large ($\beta(K) \to \infty$), compressing the employment region $[\bar{W} - R] \to 0$. In the limit, the stationary density concentrates near W^* and $\Phi(\Omega) \to C > 0$ for some finite constant (the integral $\int W^\theta g\, dW$ remains bounded because g integrates to 1 and W is bounded). Hence $\Phi(\Omega)/\Omega \to \infty$ as $\Omega \to 0^+$.

As $\Omega \to \infty$: $K(W;\Omega) \to 1$ uniformly. The problem converges to the constant-K benchmark of Sannikov [2008]. The boundaries $R(\Omega)$ and $\bar{W}(\Omega)$ converge to finite limits, and $g(W;\Omega)$ converges to a well-defined density with support on a fixed interval. Therefore $\Phi(\Omega)$ converges to a finite limit $\bar{\Phi}$, and $\Phi(\Omega)/\Omega \to 0$ as $\Omega \to \infty$.

(c) Existence. Since $\Phi(\Omega)/\Omega > 1$ for small Ω and $\Phi(\Omega)/\Omega < 1$ for large Ω, and Φ is continuous, the intermediate value theorem implies the existence of Ω^* with $\Phi(\Omega^*) = \Omega^*$.

(d) Uniqueness. We show that Φ is a contraction in a neighborhood of any fixed point. Consider the effect of increasing Ω. A larger Ω reduces AUM for every manager: $K(W;\Omega)$ is decreasing in Ω. Lower AUM has two effects: (i) it reduces effort costs, shifting the effort function upward and widening the employment region, and (ii) it reduces the scale factor in the principal's payoff $K \cdot a$. By the comparative statics of the inner problem, the net effect on the stationary density is that g spreads out but the integral $\int W^\theta g\, dW$ responds less than proportionally to changes in Ω.

Formally, differentiating the fixed-point condition:

$$\frac{d\Phi}{d\Omega} = \int_R^{\bar{W}} W^\theta \frac{\partial g}{\partial \Omega}\, dW + W^\theta(\bar{W}) g(\bar{W}) \frac{d\bar{W}}{d\Omega} - W^\theta(R) g(R) \frac{dR}{d\Omega}.$$

The boundary terms vanish because $g(R) = g(\bar{W}) = 0$ (absorbing boundaries). The remaining integral captures how the density shifts. Since $K(W;\Omega) = 1 + W^\theta/\Omega$, increasing Ω reduces K and hence reduces $\beta(K)$, which raises effort and makes the drift of W more positive for low-W managers. This shifts mass rightward in the density. However, the shift in the density is second-order relative to the direct $1/\Omega$ effect in the AUM allocation. Specifically, the dominant channel by which Ω enters Φ is through the density g, and this response is dampened because g solves a second-order PDE (the KFE) that smooths perturbations. The result is $0 < d\Phi/d\Omega < 1$, which ensures uniqueness of the fixed point.

More precisely, consider two candidate equilibria $\Omega_1 < \Omega_2$. The corresponding firm value functions satisfy $F(W;\Omega_1) < F(W;\Omega_2)$ for W in the interior, because lower Ω means higher K and hence higher effort costs, which reduce firm value. By Lemma A1 of Chemla et al. [2025] (adapted to our setting: when the inner problem's solution improves, the equilibrium is unique), the entry point W^* and boundaries adjust such that $\Phi(\Omega_2) - \Phi(\Omega_1) < \Omega_2 - \Omega_1$. This establishes that $\Omega \mapsto \Phi(\Omega) - \Omega$ is strictly decreasing, hence has at most one zero. □

Remark 3. *The proof structure parallels Chemla et al. [2025], who establish existence and uniqueness of equilibrium in a continuous-time contracting model with endogenous outside options. In their setting, the fixed-point argument operates on the pair (R, L) of outside options. In our model, the outside options $R = (1-\alpha)W^*$ and $L = (1-\alpha)F(W^*)$ are determined by the search friction α and the shape of F, while the general equilibrium operates through the AUM channel via Ω_θ. The key difference is that the externality in our model arises from the cross-sectional allocation of capital—each manager's fund size depends on the entire distribution of managers—rather than from the compensation externality emphasized by Chemla et al..*

B Numerical Algorithm

We solve the model numerically using backward shooting on the HJB equation, combined with root-finding for the general equilibrium conditions.

Given $(\bar{W}, R, L)$, set boundary conditions at $\bar{W}$ from smooth pasting with F_{out}: $F(\bar{W}) = L - (\bar{W} - R)^2$, $F'(\bar{W}) = -2(\bar{W} - R)$. Integrate the ODE system (24)–(25) backward from $\bar{W}$ to R using the Radau IIA stiff solver with $N = 4{,}000$ grid points. At each step, compute γ from (24), effort $a = \gamma - \beta$, consumption $c = [F']^2/4$, and F'' from (25).

Solve for $(\bar{W}, R, L)$ using Levenberg–Marquardt to satisfy three residual conditions: $F'(W^*) = 0$, $L = (1-\alpha)F(W^*)$, and $F(R) = L$.

Start from the $K = 1$, $\eta = 0$ benchmark (solved first by bisection on W_{gp}). Gradually increase both K toward $1+W^\theta/\Omega$ and η toward its target in $n = 8$ equally spaced steps, using the solution at each step as the initial guess for the next. This ensures robust convergence even for large η and θ.

Solve the KFE (26) using a finite-difference upwind scheme with absorbing boundaries at R and $\bar{W}$ and a point source at W^*. Normalize so that $\int g\, dW = 1$.

If Ω_θ needs updating, iterate Steps 1–4 with damping until convergence.

The full algorithm converges in under four minutes for all parameter configurations reported in the paper. The Radau IIA solver is essential for numerical stability: forward shooting methods become stiff due to the $1/r$ amplification factor in the dynamics, causing solutions to linearize in the tail.

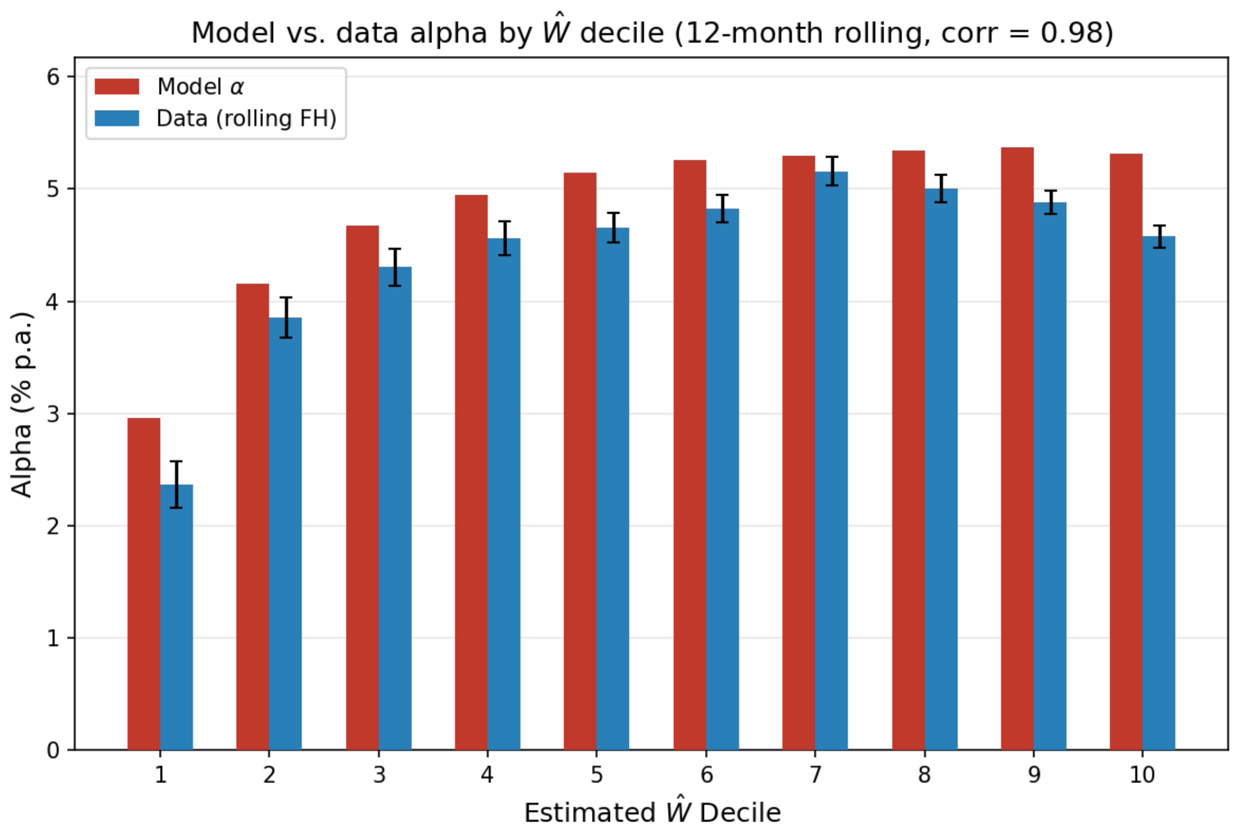

Figure 5: **Model vs. data: rolling alpha by recovered continuation value.** Model-predicted and data rolling alpha by decile of $\hat{W}$, recovered from observed AUM via a gradient-boosted tree trained on simulated data. Red bars: simulated rolling 12-month alpha from the model with scale friction $\varphi(K^\delta) = 1/(1+(K^\delta/\bar{x})^\kappa)$, $\bar{x} = 75{,}000$, $\kappa = 5$; blue bars: realized rolling 12-month Fung–Hsieh alpha in the data. The model reproduces the hump-shaped profile of the data, with alpha peaking around D7 and declining toward D10 where capacity constraints bind. Decile-level correlation: 0.98; RMSE: 0.45 percentage points. Sample: 1,558,009 fund-month observations from HFR, 1996–2025; 512,445 simulated observations from 50,000 model paths.

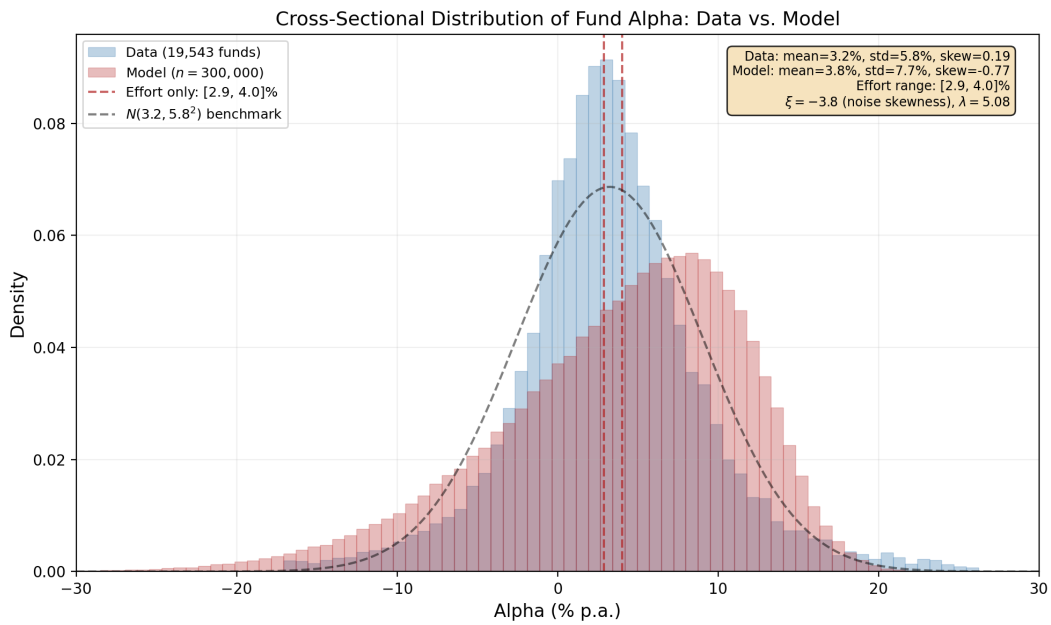

Figure 6: **Cross-sectional alpha distribution: model vs. data.** Blue histogram: empirical distribution of fund-level Fung–Hsieh alpha from HFR (19,543 funds, winsorized at 1%/99%). Red histogram: model-implied distribution under the stationary density $g(W)$, with one-period alpha $[a(W) + \sigma\,\varepsilon] \times \lambda$ and skew-normal noise ($\xi = -3.84$; 300,000 draws). Red dashed lines: effort-only range $a(W) \times \lambda \in [2.9\%, 4.0\%]$. Black dashed curve: Gaussian benchmark $N(\bar{\alpha}^{\text{data}}, \text{Std}^{\text{data}\,2})$. The model and data distributions overlap substantially, with means of 3.79% (model) and 3.24% (data). Scale factor $\lambda = 5.08$ maps model units to annual percentage points.