Nonlinear price impact and portfolio choice

Published Version:
Guasoni P, Weber MH (2020). Nonlinear price impact and portfolio choice. MATHEMATICAL FINANCE, 30(2 (April)), 341-376 [10.1111/mafi.12234].

Availability:
This version is available at: https://hdl.handle.net/11585/855689 since: 2022-02-10

Published:
DOI: http://doi.org/10.1111/mafi.12234

NONLINEAR PRICE IMPACT AND PORTFOLIO CHOICE

PAOLO GUASONI†

Boston University and Dublin City University

MARKO HANS WEBER‡

National University of Singapore

In a market with price impact proportional to a power of the order flow, we find optimal trading policies and their implied performance for long-term investors who have constant relative risk aversion and trade a safe asset and a risky asset following geometric Brownian motion. These quantities admit asymptotic explicit formulas up to a structural constant that depends only on the curvature of the price impact function. Trading rates are finite as with linear impact, but are lower near the target portfolio, and higher away from the target. The model nests the square-root impact law and, as extreme cases, linear impact and proportional transaction costs.

KEY WORDS: Price Impact, Square-Root Law, Trading Volume.

MATHEMATICS SUBJECT CLASSIFICATION (2010): 91G10, 91G80

JEL CLASSIFICATION: G11, G12

1. INTRODUCTION

The impact of trades on execution prices is a critical determinant of portfolio rebalancing policies: frictionless models assume a single price, insensitive to sales and purchases of any size, resulting in policies with infinite trading volume (Merton, 1969). Models that acknowledge bid-ask spreads preclude trading when a portfolio is near its target (Constantinides, 1986; Davis and Norman, 1990), whereas linear price impact models recommend a trading rate proportional to its distance from the target (Guasoni and Weber, 2017; Moreau et al., 2017). Yet, empirical evidence suggests that price impact is nonlinear.

This paper examines the implications of nonlinear price impact for optimal portfolio rebalancing. Our model posits that a one-percent increase in trading speed increases price impact by α percent. That is, α is the elasticity of price impact to the order flow – the focus of this paper. At one extreme, $\alpha = 1$ recovers linear price impact; at the other extreme, $\alpha = 0$ reproduces constant proportional bid-ask spreads. All intermediate values generate nonlinear impact functions, with $\alpha = 1/2$ corresponding to the square-root law proposed by several authors (cf. Loeb (1983); Torre and Ferrari (1998); Grinold and Kahn (2000)).

We thank for helpful comments Thomas Cayé, Eberhard Mayerhofer, Johannes Muhle-Karbe, Mykhaylo Shkolnikov, Ronnie Sircar, and seminar participants at Carnegie Mellon University, EPFL, Fields Institute, Princeton University, University of Paris VI, Rutgers University, Bar-Ilan University, and Quant Europe. Special thanks go to the two anonymous referees, who helped correct some issues in earlier versions and improve the presentation.

†Boston University, Department of Mathematics and Statistics, 111 Cummington Mall, Boston, MA 02215, USA, and Dublin City University, School of Mathematical Sciences, Glasnevin, Dublin 9, Ireland, email `guasoni@bu.edu`. Partially supported by the ERC (279582), NSF (DMS-1412529), and SFI (16/IA/4443 and 16/SPP/3347).

‡National University of Singapore, Department of Mathematics, Singapore 119076, Singapore, email `matmhw@nus.edu.sg`

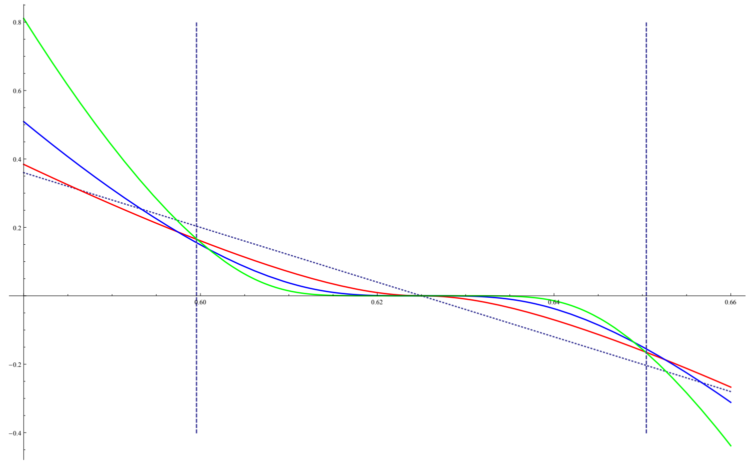

FIGURE 1. Optimal trading rates (vertical axis) against current risky weight (horizontal axis) for $\alpha = 1/8$ (green), $\alpha = 1/4$ (blue), and $\alpha = 1/2$ (red). Linear impact ($\alpha = 1$) leads to the linear rate (dotted), whereas vertical lines (dashed) identify the no-trade region arising with transaction costs ($\alpha \downarrow 0$). $\mu = 8\%$, $\sigma = 16\%$, $\gamma = 5$, $\lambda = 0.1\%$.

The paper finds the optimal rebalancing strategy, its performance, and the resulting trading volume for long-term investors with constant relative risk aversion, who trade in a market with one safe and one risky asset, and constant investment opportunities summarized by a geometric Brownian motion. The main finding is that, in the small price impact limit, the optimal policy, its performance, and trading volume are all identified, up to a change in scale and location, by the *shape equation*

$$s'_\alpha(z) = z^2 - c_\alpha - \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}} |s_\alpha(z)|^{\frac{\alpha+1}{\alpha}}, \tag{1}$$

$$\lim_{z\to-\infty} \frac{s_\alpha(z)}{|z|^{\frac{2\alpha}{\alpha+1}}} = \lim_{z\to+\infty} -\frac{s_\alpha(z)}{|z|^{\frac{2\alpha}{\alpha+1}}} = (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}}, \tag{2}$$

an ordinary differential equation, the solution of which consists of the scalar c_α and the *shape function* $s_\alpha(z)$. For $\alpha = 1$, the shape function is linear, and the model recovers the optimal policy for linear impact in Guasoni and Weber (2017). In the limit $\alpha \downarrow 0$, the shape function gives rise to the classical trading policy of Dumas and Luciano (1991) for bid-ask spreads.

The implications of nonlinear price impact are qualitatively similar to those of linear impact, but quantitatively skewed towards those of proportional transaction costs. As in models with linear impact, optimal trading policies yield finite trading rates, but these rates are lower near the target portfolio, and higher away from the target (Figure 1). These differences bring the optimal policy closer to its counterpart with proportional bid-ask spreads, in which the trading rate is zero inside the buy and sell boundaries, while it is “infinite” (in that portfolio adjustment is instantaneous) outside the boundaries. Likewise, the asymptotic long-term distribution of the portfolio weight around its target is bell-shaped, similar to the normal distribution arising with linear impact, but less peaked, and closer to the uniform distribution arising with proportional bid-ask spreads (Janeček and Shreve, 2004).

The significance of these results is fourfold. First, despite the technical challenges of nonlinear price impact, the resulting portfolio choice problem is nearly as tractable as with linear impact, and yields a solution that is less sensitive to the presence of bid-ask spreads, because nonlinear impact depresses trading when the portfolio lies near its target, similar to the no-trade regions that are implied by bid-ask spreads. Thus, in addition to being an empirically documented feature, nonlinear impact is also a flexible device to mimic the effect of bid-ask spreads. This analogy may be used as a means to attack problems in which the free-boundary problems arising from bid-ask spreads are intractable.

Second, our asymptotics show that nonlinear price impact reduces the frictionless equivalent safe rate by

$$c_\alpha \lambda^{\frac{2}{\alpha+3}} \left(\gamma \left(\frac{\sigma^2}{2}\right)^3 \bar{Y}^4 (1-\bar{Y})^4\right)^{\frac{\alpha+1}{\alpha+3}}, \tag{3}$$

where λ is the price impact for unit trading speed, $\bar{Y} = \frac{\mu}{\gamma\sigma^2}$ the frictionless target portfolio, μ and σ the risky asset's return and volatility, and γ the investor's risk aversion. This formula implies, in particular, that the square-root law ($\alpha = 1/2$) implies a cost proportional to the $4/7$-th power of the impact's magnitude λ, while for linear impact and proportional bid-ask spreads it reduces to the performance formulas in Guasoni and Weber (2017) and Gerhold et al. (2014).

Third, the shape function regulates the drift of trading volume near its frictionless level and links the parameter α to the shape of the stationary distribution of the portfolio weight, which interpolates the uniform ($\alpha \downarrow 0$) and Gaussian ($\alpha = 1$) distributions arising with bid-ask spreads and linear impact respectively. Finally, the very recent results of Cayé et al. (2018) on constant risk aversion and asset prices following general diffusions, show that the shape function identified in this paper drives the optimal policies for nonlinear price impact even in such models, thereby attesting to its central role for nonlinear impact.

This is the first paper to solve a portfolio choice model with nonlinear price impact, and we focus on a partial-equilibrium model of a large investor with constant relative risk aversion. We study power-type impact for any $\alpha \in (0, 1]$, as equilibrium models support different exponents (Kyle, 1985; Garleanu et al., 2009; Gabaix et al., 2006). In the literature, nonlinear price impact appears mainly in problems of optimal liquidation (Almgren, 2003; Vath et al., 2007; Schied et al., 2010), which are in fact a special case of portfolio choice – when the risky assets have no risk premium.

From a mathematical viewpoint, models of nonlinear impact pose challenges similar to the ones of linear impact – with a few surprises. We focus on investors with a long horizon (Dumas and Luciano, 1991; Grossman and Zhou, 1993; Guasoni and Robertson, 2012; Gerhold et al., 2014; Guasoni and Weber, 2017), thereby obtaining optimal policies that are insensitive to the investment horizon. The search for explicit formulas in the small-impact limit departs significantly from the linear case, as nonlinear impact leads to a singularity in the value function near the target weight, hence to separate scaling behaviors near to and away from the target. A suitable re-normalization of the value function and its argument removes the dependence on all model parameters except α, leading to the shape equation (15), which is akin to the corrector equation in Soner and Touzi (2013), as it describes the scale-free dynamics of the problem's state variable.

The rest of the paper is organized as follows: the next section contains a brief review of the literature that puts the paper in context. Section 3 describes the model and discusses existence and optimality results that are necessary for the main asymptotic results in section 4. Section 5 discusses the model's implications, comparing them with those of proportional bid-ask spreads and linear impact. In section 6 we provide heuristic arguments for the main result and in section 7 we discuss an extension of the model. All the proofs are deferred to the appendix.

2. Literature Review

Price impact is a broad designation that includes diverse mechanisms in which traders' actions affect current or future prices, therefore it is important to clarify the scope of this paper's results in this vast literature, which encompasses *permanent*, *transient*, and *immediate* impacts. This paper focuses on *immediate* price impact.

The seminal equilibrium model of Kyle (1985), and its extensions to continuous time (Back, 1992) and risk aversion (Baruch, 2002) among many others, have spearheaded the equilibrium literature on *permanent impact*, in which asset prices arise endogenously from the interaction of informed and uninformed traders with market makers. In this literature, price impact results from the gradual assimilation in prices of insiders' information, which accumulates over time. The models in Cuoco and Cvitanić (1998), Bank and Baum (2004), and Bank and Kramkov (2015*b*,*a*) also explore different mechanisms in which a large investor's aggregate position affects asset prices, and therefore price impact is permanent. In the same fashion, Huberman and Stanzl (2004) study a partial-equilibrium model of permanent impact and conclude that only linear price impact functions are consistent with the absence of quasi arbitrage and price manipulation.

The literature on *transient impact* specifies partial equilibrium models that capture the reversion of order books to their steady state in response to imbalances, as in Obizhaeva and Wang (2013) and Predoiu et al. (2011). Because absence of dynamic arbitrage rules out exponential decay in combination with nonlinear impact (Gatheral, 2010), the results in Gatheral et al. (2012) focus on linear impact with arbitrary decay.

The literature on *immediate impact* (sometimes also referred to as temporary or instantaneous) is generally immune to price manipulation and arbitrage issues, as it specifies partial equilibrium models in which the execution price is less favorable than some exogenous baseline price: if the latter is arbitrage-free, so is the former. Different price impact specifications are interpreted as different shapes of infinitely resilient order books, so that the execution price at each time depends only on current trading speed. (For example, linear impact corresponds to a flat order book, while square-root impact to a V-shaped book.[1])

Much of this literature follows the optimal execution model in Almgren and Chriss (2001), in which the mean-variance tradeoff is between a quicker execution at a worse execution price and a slower execution at a more uncertain price. Schied and Schöneborn (2009) develop a continuous-time version of the liquidation problem with exponential utility and immediate linear impact in the number of shares traded per unit of time, a measure of order flow. A similar specification appears in mean-variance multivariate models of dynamic trading with frictions (Gârleanu and Pedersen, 2013, 2016), where immediate price impact is linear in the number of shares traded per unit of time in all securities, in part for tractability. (Gârleanu and Pedersen (2013, 2016) refer to quadratic transaction costs, yet another name for immediate linear impact.) Collin-Dufresne et al. (2012) specify linear impact in dollar amounts traded per unit of time, which leads to a more challenging model.

[1]The observed order book reflects mostly the activity of high frequency traders. The relevant object for a price impact model is, instead, the *latent* order book introduced in Toth et al. (2011), which aggregates the total *intended* volume to buy/sell at each time.

The interest in nonlinear price impact starts from practitioners (Loeb, 1983; Torre and Ferrari, 1998; Grinold and Kahn, 2000), who posit a square-root law of market impact as a realistic rule of thumb. Subsequent empirical work by practitioners and academics confirms the concavity of price impact (Hasbrouck and Seppi, 2001; Plerou et al., 2002), with some authors reporting a dependence approximately of power type and with exponents close to $\alpha = 1/2$ (Almgren et al., 2005; Lillo et al., 2003). Nonlinear impact first appears explicitly in an optimization problem in the optimal execution model of Almgren (2003) .

3. Model and Preliminaries

3.1. **Market and Preferences.** The financial market includes a safe asset earning zero interest rate[2], and a risky asset with best quoted price S_t following the usual Samuelson model

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW_t,$$

where W_t is a standard Brownian motion, $\mu > 0$ is the expected excess return, and $\sigma > 0$ is the volatility.[3] Unlike frictionless markets, trades in the risky asset are not realized at the best quote S_t, but at a less favorable price that reflects adverse movements during the execution of a large trade. The average price for trading $\Delta\theta$ shares in the interval $[t, t+\Delta]$ is specified as

$$\tilde{S}_t := S_t \left(1 + \lambda \left|\frac{S_t \Delta\theta}{X_t \Delta t}\right|^{\alpha} \operatorname{sgn}(\Delta\theta)\right), \tag{4}$$

where X_t denotes the large investor's wealth at time t (valuing the risky position at the best quote S_t), $\lambda > 0$ is a measure of the asset's illiquidity. The shape parameter α is the elasticity of price impact with respect to trading volume – the main focus of this paper.

By the specification in (4), a larger trade $\Delta\theta$ or a smaller execution time Δt result in a higher price impact. The proportional price impact is assumed to be a function of the wealth turnover $S_t\Delta\theta/X_t$, that is the amount traded as a fraction of the investor's total wealth. The rationale for this scaling is as follows: empirical work (Engle et al., 2012; Frazzini et al., 2012) documents that price impact increases with the trade size as a fraction of daily trading volume V_t, i.e., it is some nonlinear increasing function of $S_t\Delta\theta/(V_t\Delta t)$. Daily trading volume is cointegrated with the market capitalization M_t, in that the ratio V_t/M_t is long-term stationary (typical long-term averages are of the order of 1% of market capitalization turned over daily). Furthermore, a large investor's wealth X_t is cointegrated with the market capitalization (the ratio X_t/M_t is positive and bounded from above), which in turn implies that the ratio V_t/X_t is stationary. For simplicity, (4) further approximates such a ratio with a constant, and section 7 below discusses an extension of the model to the case of a stochastic ratio, at the expense of including wealth as an additional state variable.

Passing to continuous time to make the model tractable, we assume that θ_t, the number of shares of the risky asset held at time t, is absolutely continuous with respect to time, and denote by $\dot{\theta}_t$ its derivative. The resulting self-financing condition requires that the investor's cash position $C_t = X_t - \theta_t S_t$ evolves as

$$dC_t = -\tilde{S}_t d\theta_t = -S_t \left(1 + \lambda \left|\frac{\dot{\theta}_t S_t}{X_t}\right|^{\alpha} \operatorname{sgn}(\dot{\theta}_t)\right) d\theta_t.$$

---

[2]A zero rate helps ease notation, but the results remain valid with any constant rate.

[3]For the case $\mu < 0$, see Remark 2 below.

Denoting wealth turnover as $u_t := \frac{\dot{\theta}_t S_t}{X_t}$, routine stochastic calculus derivations lead to the following joint dynamics of the investor's wealth $X_t := \theta_t S_t + C_t$ and portfolio weight $Y_t := \frac{\theta_t S_t}{X_t}$,

$$\frac{dX_t}{X_t} = Y_t(\mu dt + \sigma dW_t) - \lambda |u_t|^{\alpha+1} dt, \tag{5}$$

$$dY_t = (Y_t(1-Y_t)(\mu - Y_t\sigma^2) + u_t + \lambda Y_t |u_t|^{\alpha+1})dt + Y_t(1-Y_t)\sigma dW_t. \tag{6}$$

These joint dynamics, obtained through heuristic derivations from the self-financing conditions, are the starting point of our rigorous treatment. Consider a probability space $(\Omega, \mathcal{F}, P)$, equipped with a Brownian motion W_t and its augmented natural filtration $\mathcal{F}_t$.

An admissible strategy is defined as a pair (Y_0, u) of an initial portfolio Y_0 and a turnover process u_t such that the resulting wealth process remains solvent at all times, in analogy with the frictionless theory.

Definition 1. *For $X_0 > 0$ and $Y_0 \in \mathbb{R}$, an* admissible strategy *is a process $(u_t)_{t\geq 0}$, adapted to $\mathcal{F}_t$, such that $\int_0^T |u_t|^{1+\alpha} dt < \infty$ a.s. for all T, and for which the stochastic differential equation* (6) *has a unique strong solution on $(0, \infty)$. For any such admissible strategy, the corresponding wealth process is*

$$X_T^u = X_0 \exp\left(\int_0^T \left(\mu Y_t - \frac{\sigma^2}{2} Y_t^2 - \lambda |u_t|^{1+\alpha}\right) dt + \int_0^T \sigma Y_t dW_t\right). \tag{7}$$

In this market, the investor's objective is to maximize the equivalent safe rate, defined as

$$\mathrm{ESR}_\gamma(u) = \limsup_{T\to\infty} \frac{1}{T} \log E\left[(X_T^u)^{1-\gamma}\right]^{\frac{1}{1-\gamma}}, \tag{8}$$

where $0 < \gamma \neq 1$ denotes the investor's relative risk aversion.[4] This asymptotic objective focuses on the growth rate of the certainty equivalent of power utility at a long horizon, while neglecting the dependence on the length of the horizon and preserving stationarity.

Remark 2. *Leverage and short-selling are not admissible because short and levered positions are insolvent with positive probability.*[5] *In particular, any strategy such that $Y_t \notin [0,1]$ for some t is not admissible. If the frictionless target is short ($\mu < 0$), then full investment in the safe asset – and no trading – is optimal. If the frictionless target is levered ($\mu/(\gamma\sigma^2) > 1$), then full investment in the risky asset is optimal, again with no trading.*[6]

3.2. **Verification Theorem.** The verification theorem 3 identifies the optimal trading policy and its equivalent safe rate from the solution of the associated ergodic HJB equation.

[4] The case with logarithmic utility is analogous. In fact, all results extend to the logarithmic case by passing to the limit $\gamma \to 1$ in the statements.

[5] This fact follows from (Guasoni and Weber, 2017, Lemma A.2), which carries over to the present setting. The proof is omitted, as it is similar, replacing $-\frac{1}{4\lambda}$ with $-\frac{\alpha}{\alpha+1}\frac{1}{\lambda^{1/\alpha}(\alpha+1)^{1/\alpha}}$, and modifying the equation in the statement of Lemma A.3 to $\lim_{x\to\bar{x}} \frac{\int_\xi^x f(z)dz}{\int_\xi^x |f'(z)|^{\alpha+1}dz} = 0$, the proof of which follows from Hölder's inequality with exponent $\alpha + 1$.

[6] The statement of Theorem 2.3 in Guasoni and Weber (2017) carries over to the present setting, and follows from the validity of Lemma A.2 in Guasoni and Weber (2017).

Theorem 3. *Assume* $\bar{Y} := \frac{\mu}{\gamma\sigma^2} \in (0,1)$. *There exists* $\lambda^* > 0$ *such that for* $\lambda \in (0, \lambda^*)$ *the wealth turnover that maximizes* (8) *is*

$$\hat{u}(y) = \left| \frac{q(y)}{(\alpha+1)\lambda(1-yq(y))} \right|^{1/\alpha} \operatorname{sgn}(q(y)) \tag{9}$$

and its equivalent safe rate is

$$\mathrm{ESR}_\gamma(\hat{u}) = \hat{\beta}, \tag{10}$$

where $\hat{\beta} \in (0, \frac{\mu^2}{2\gamma\sigma^2})$ *and* $q : [0,1] \mapsto \mathbb{R}$ *are the unique pair that solves the ODE*

(11)

$$-\hat{\beta} + \mu y - \gamma\frac{\sigma^2}{2}y^2 + y(1-y)(\mu - \gamma\sigma^2 y)q + \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}} \frac{|q|^{\frac{\alpha+1}{\alpha}}}{(1-yq)^{1/\alpha}} \lambda^{-1/\alpha} + \frac{\sigma^2}{2}y^2(1-y)^2(q' + (1-\gamma)q^2) = 0$$

with the boundary conditions

$$q(0^+) = \lambda^{\frac{1}{\alpha+1}} (\alpha+1)^{\frac{1}{\alpha+1}} \left(\frac{\alpha+1}{\alpha} \hat{\beta} \right)^{\frac{\alpha}{\alpha+1}}, \tag{12}$$

(13)

$$\frac{\alpha}{\alpha+1}(\alpha+1)^{-1/\alpha} \frac{|q(1^-)|^{\frac{\alpha+1}{\alpha}}}{(1-q(1^-))^{1/\alpha}} \lambda^{-1/\alpha} = \hat{\beta} - \mu + \gamma\frac{\sigma^2}{2}.$$

Note that the boundary conditions (12)-(13) are natural in that the corresponding values of $q(0)$ and $q(1)$ are the only possible finite limits consistent with (11), as the coefficient of q' vanishes for $y = 0, 1$. Thus, a numerical procedure to solve this system is to solve an initial value problem for q starting at some arbitrary $\bar{y} \in (0,1)$ (for example, $\bar{y} = \mu/(\gamma\sigma^2)$), and then calibrate the unknown parameters $q(\bar{y})$ and $\hat{\beta}$ so that the solution has finite limits in 0 and 1. Such $\hat{\beta}$ is the equivalent safe rate of the problem, while the optimal turnover rate $\hat{u}(y)$ follows from the function $q(y)$ through the expression in (9).

The following proposition shows that liquidation can be realized in a short period of time and at a low cost relative to the final portfolio value. Accordingly, the best quote of the asset S_t is assumed constant during the liquidation.

Proposition 4 (Liquidation)**.** *Assume* $S_t \equiv S$ *constant for* $t \geq T$. *The liquidation time* $L(u) = \inf\{t \geq 0 : \theta_{T+t} = 0\}$ *of the constant selling policy* $u_t \equiv u < 0$ *equals* $L(u) = -\frac{\log(1-\lambda Y_T|u|^\alpha)}{\lambda|u|^{\alpha+1}} \sim -\frac{Y_T}{u}$ *and the proportional liquidation cost is* $\frac{X_T - X_{T+L(u)}}{X_T} = \lambda Y_T |u|^\alpha$.

For example, for a selling rate of order $O(\lambda^{-\frac{1}{\alpha+3}})$, liquidation time would be of order $O(\lambda^{\frac{1}{\alpha+3}})$ with a liquidation cost of order $O(\lambda^{\frac{3}{\alpha+3}})$.

The main results in the next section provide asymptotic formulas for the optimal trading strategies in the limit of small price impact $\lambda \downarrow 0$. This analysis departs significantly from the corresponding results for linear impact, as nonlinear impact of power type leads to qualitatively different impacts for small ($u \sim 0$) and large ($u \sim \infty$) turnover.

4. MAIN RESULT

This section contains the main result of the paper. The next Theorem calculates the asymptotic optimal trading policy, its performance, and the dynamics of the portfolio weight, which are all governed by the scale-free ordinary differential equation (15), independent of preference and market parameters other than the elasticity of impact α. The asymptotic policy depends on all other parameters only through the constants

$$l_\alpha := \left[\left(\frac{\sigma^2}{2}\right)^3 \gamma \bar{Y}^4(1-\bar{Y})^4\right]^{\frac{\alpha+1}{\alpha+3}}, \quad A_\alpha = \left(\frac{2l_\alpha}{\gamma\sigma^2}\right)^{1/2}, \quad B_\alpha = l_\alpha^{\frac{\alpha}{\alpha+1}}. \tag{14}$$

Theorem 5. *For $\alpha \in (0,1]$ there exist a unique constant $c_\alpha > 0$ and a unique function $s_\alpha \in C^1(\mathbb{R})$ that solve*

$$s'_\alpha(z) = z^2 - c_\alpha - \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}}|s_\alpha(z)|^{\frac{\alpha+1}{\alpha}} \tag{15}$$

with the growth conditions[7]

$$\lim_{z\to-\infty} \frac{s_\alpha(z)}{|z|^{\frac{2\alpha}{\alpha+1}}} = (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} \quad \textit{and} \quad \lim_{z\to+\infty} \frac{s_\alpha(z)}{|z|^{\frac{2\alpha}{\alpha+1}}} = -(\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}}. \tag{16}$$

Recalling that $q_\lambda(y)$ and $\hat{u}_\lambda(y)$ are, respectively, the reduced value function and optimal policy in Theorem 3:

(i) The rescaled value function satisfies

$$\lim_{\lambda\downarrow 0} q_\lambda(\bar{Y} + \lambda^{\frac{1}{\alpha+3}} z)\lambda^{-\frac{3}{\alpha+3}} = B_\alpha s_\alpha(z/A_\alpha). \tag{17}$$

(ii) The rescaled trading policy satisfies

$$\lim_{\lambda\downarrow 0} \lambda^{\frac{1}{\alpha+3}} \hat{u}_\lambda(\bar{Y} + \lambda^{\frac{1}{\alpha+3}} z) = v_\alpha(z) := -\left|\frac{B_\alpha s_\alpha(z/A_\alpha)}{(\alpha+1)}\right|^{1/\alpha} \operatorname{sgn}(z). \tag{18}$$

(iii) The equivalent safe rate has the asymptotic expansion

$$\mathrm{ESR}_{\gamma,\alpha}(\hat{u}_\lambda) = \frac{\mu^2}{2\gamma\sigma^2} - c_\alpha l_\alpha \lambda^{\frac{2}{\alpha+3}} + o(\lambda^{\frac{2}{\alpha+3}}). \tag{19}$$

(iv) Assume $Y_0 = \bar{Y}$. The rescaled weight $Z_s^\lambda := \lambda^{-\frac{1}{\alpha+3}}(Y_{\lambda^{2/(\alpha+3)}s} - \bar{Y})$ converges weakly to the process Z_s^0 defined by[8]

$$dZ_s^0 = v_\alpha(Z_s^0)ds + \bar{Y}(1-\bar{Y})\sigma dW_s, \qquad Z_0^0 = 0. \tag{20}$$

The above theorem summarizes the main attributes of the optimal policy: with nonlinear price impact, the optimal policy is not fully explicit even in the asymptotic regime, as it depends on the solution s_α to the asymptotic equation (15), along with the long-run dynamics of the portfolio weight in (20). Similarly, the equivalent safe rate in (19) depends on the constant c_α, which does not have a closed-form expression in terms of α.

Yet, the complexity of the asymptotic policy is limited to the one-time calculation of c_α and s_α, while the dependence on all other model parameters is explicit, and in fact

[7] It is seen immediately that if $s_\alpha(z)$ solves (15)-(16), then also $-s_\alpha(-z)$ solves (15)-(16). Hence, the uniqueness of the solution implies that $s_\alpha(z) = -s_\alpha(-z)$. In particular, one of the growth conditions can be replaced by $s_\alpha(0) = 0$.

[8] Note that the weight process Y also depends on λ, though we omit the superscript Y^λ for consistency with the previous sections.

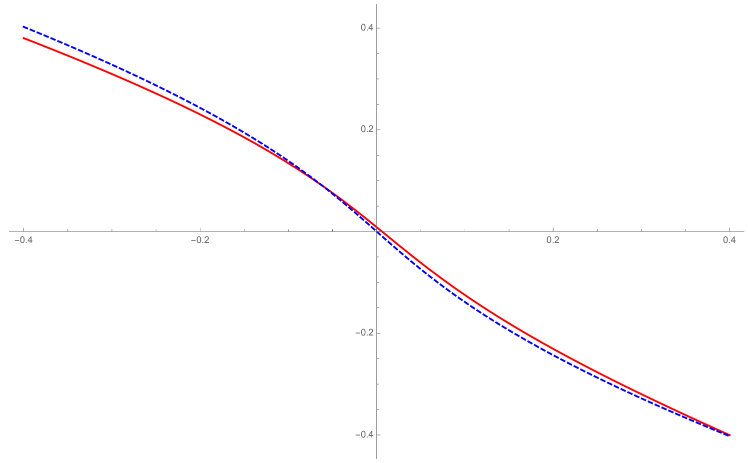

FIGURE 2. Exact rescaled value function $q_\lambda(\bar{Y} + \lambda^{\frac{1}{\alpha+3}} A_\alpha z)\lambda^{-\frac{3}{\alpha+3}}/B_\alpha$ (solid red) and its asymptotic limit $s_\alpha(z)$ (dashed blue), as functions of the rescaled portfolio z, in which 0 represents the frictionless target. The asymptotic function depends only on α, the exact curve is obtained with $\mu = 8\%$, $\sigma = 16\%$, $\alpha = 1/2$, $\gamma = 5$, and with $\lambda = 10\%$ (the curves are visually indistinguishable with $\lambda = 1\%$ or lower).

rather simple, through the constants in (14). In addition, Figure 2 shows that the asymptotic approximation is rather accurate for typical parameter values, even for relatively high λ.[9]

As nonlinear impact leads to the nonlinear trading rates in Figure 1, it also tilts the stationary density of the portfolio weight away from the normal distribution implied by linear price impact. As α declines and nonlinearity increases (Figure 3) the stationary distribution becomes flatter at the center and thinner-tailed, approaching the uniform distribution in the limit of transaction costs ($\alpha = 0$).

The confidence intervals of the portfolio weight Y_t are neighborhoods of $\bar{Y}$ of order $\lambda^{\frac{1}{\alpha+3}}$, i.e.,

$$\lim_{\lambda\to 0}\lim_{T\to\infty} \tfrac{1}{T}\int_0^T 1_{\{Y_t\in[\bar{Y}-\lambda^{\frac{1}{\alpha+3}}z,\bar{Y}+\lambda^{\frac{1}{\alpha+3}}z]\}}dt =$$

$$\lim_{\lambda\to 0}\lim_{T\to\infty} \frac{\lambda^{\frac{2}{\alpha+3}}}{T}\int_0^{T/\lambda^{\frac{2}{\alpha+3}}} 1_{\{Z_s^\lambda\in[-z,z]\}}ds = \lim_{\lambda\to 0}\mu^\lambda([-z,z]) = \mu^0([-z,z]),$$

where μ^λ is the invariant measure associated to the process Z^λ.

[9]Note that the small positive value at $z = 0$ of the exact curve is consistent with the second-order term in the linear-impact case (Guasoni and Weber, 2017, Remark A.13), and reflects the motive to contrast the current drift of the portfolio Y_t rather than its position relative to the target.

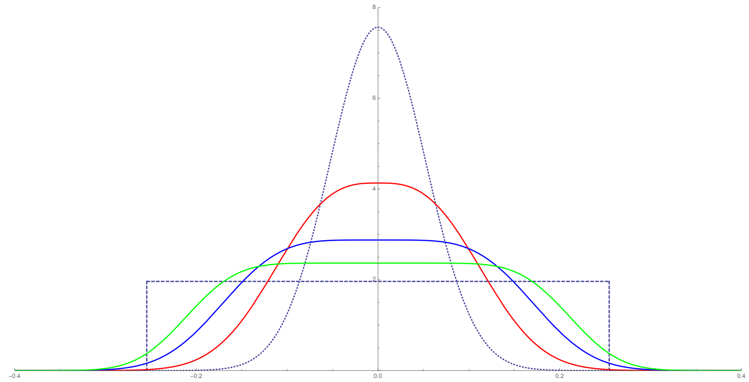

FIGURE 3. Density (vertical axis) of the invariant measure of the rescaled risky weight Z^0 in Theorem 5(iv) (horizontal axis) for $\alpha = 1/8$ (green), $\alpha = 1/4$ (blue), and $\alpha = 1/2$ (red). Linear impact ($\alpha = 1$) leads to the normal distribution (dotted) whereas transaction costs ($\alpha \downarrow 0$) to the uniform distribution (dashed). $\mu = 8\%$, $\sigma = 16\%$, $\gamma = 5$, $\bar{Y} = 62.5\%$.

5. IMPLICATIONS

5.1. **Trading Strategies.** Theorem 5 (ii) identifies the optimal wealth turnover (or trading rate) as

$$\hat{u}_\lambda(y) \sim -\lambda^{-\frac{1}{\alpha+3}} \left| \frac{B_\alpha s_\alpha(\lambda^{-\frac{1}{\alpha+3}}(y-\bar{Y})/A_\alpha)}{(\alpha+1)} \right|^{1/\alpha} \operatorname{sgn}(y-\bar{Y}). \tag{21}$$

This formula means that, at the first order, the optimal policy is to buy when the weight is below the frictionless target $\bar{Y}$ and to sell when it is above, as with linear impact. However, turnover is not proportional to the distance of the current weight y from the target $\bar{Y}$, as the function $|s_\alpha|^{1/\alpha}$ is generally nonlinear.

Figure 1 displays the typical shape of the optimal trading policy for different elasticity α: as the elasticity α declines from the linear value of one to the null level that approximates a bid-ask spread, the trading rate declines to zero near the frictionless target, while rapidly increasing away from the target. Indeed, as $\alpha \downarrow 0$, the trading rate converges to zero for the values of y that lie inside the no-trade region arising with bid-ask spreads (Gerhold et al., 2014), while diverging to plus infinity in the buy-region and to minus infinity in the sell-region.

The shape of the trading rate is a natural consequence of the concavity of the friction: As low trading rates have a proportionally large impact on the execution price, while high rates lead to higher – but proportionally lower – impact, it is rational for an investor to tolerate small deviations from the target by trading very slowly ($v_\alpha(z) \propto z^{\frac{1}{\alpha}}$ for z near 0), while reacting aggressively as the portfolio weight strays significantly from the target ($v_\alpha(z) \propto z^{\frac{2}{1+\alpha}}$ for z near $\pm\infty$). Equivalently, the portfolio weight has more frequent small deviations, but much less frequent large deviations from the target.

5.2. **Equivalent Safe Rate.** Theorem 5 (iii) implies that nonlinear impact reduces the equivalent safe rate by

$$c_\alpha \left[\left(\frac{\sigma^2}{2} \right)^3 \gamma \bar{Y}^4 (1-\bar{Y})^4 \right]^{\frac{\alpha+1}{\alpha+3}} \lambda^{\frac{2}{\alpha+3}}. \tag{22}$$

The factor $\lambda^{\frac{2}{\alpha+3}}$ reflects the cost of higher illiquidity. Such an order of magnitude admits a simple heuristic explanation, in analogy to the one offered by Rogers (2004) for proportional transaction costs, noting that illiquidity costs stem partly from the displacement of the portfolio weight from its target and partly from trading costs themselves. The displacement cost from keeping a portfolio weight at a distance ε from the frictionless proportion $\bar{Y}$ is of order ε^2: indeed, if $Y_t \equiv \bar{Y} + \varepsilon$, then $\frac{1}{T} \log E[X_T^{1-\gamma}]^{\frac{1}{1-\gamma}} = \frac{\mu^2}{2\gamma\sigma^2} - \frac{\gamma}{2}\sigma^2\varepsilon^2$.

To keep the portfolio weight at a distance of order ε from $\bar{Y}$, an investor should trade at a speed proportional to ε^{-1}, as it is gleaned from equation (6): with $u_t = k \cdot \varepsilon^{-1}$ the process $Z^\varepsilon = \varepsilon^{-1}(Y_{\varepsilon^2 s} - \bar{Y})$ converges in law as $\varepsilon \downarrow 0$ to the invariant distribution of Z_t^0 in (20). The trading cost of such strategy $u_t \equiv k \cdot \varepsilon^{-1}$ is thus $\lambda |u_t|^{1+\alpha} = \lambda |k\varepsilon^{-1}|^{1+\alpha}$. Thus, the sum of displacement and trading costs has the form

$$C\varepsilon^2 + \lambda\varepsilon^{-1-\alpha}$$

for some constant C. Minimizing such an objective with respect to ε yields $\varepsilon \propto \lambda^{\frac{1}{\alpha+3}}$, whence the overall illiquidity cost is of the order of $\lambda^{\frac{2}{\alpha+3}}$, in accordance to (22) (which, in addition, yields the correct constant). Put differently, displacement and trading costs are of the same order under the optimal trading policy, which otherwise may be improved by increasing the negligible cost to reduce the leading-order cost.

The constant c_α (see Figure 4) reflects the shape of the asymptotic trading policy, which depends on α alone through the function s_α.[10] The factor $[(\sigma^2/2)^3 \gamma \bar{Y}^4 (1 - \bar{Y})^4]^{\frac{\alpha+1}{\alpha+3}}$ accounts for the effects on performance that the asset's volatility has on an investor who keeps a portfolio close to the target weight $\bar{Y}$. These effects are minimal when the target is close to full investment in either the safe or the risky asset ($\bar{Y} = 0, 1$), hence the resulting portfolio has low volatility and requires little rebalancing.

Interestingly, the elasticities $2/(\alpha+3)$ and $(\alpha+1)/(\alpha+3)$ sum to one, which means that simultaneously multiplying the risk aversion, expected return, and illiquidity λ by the same factor, results in a cost also multiplied by the same factor, regardless of the nonlinearity α. This parameter, however, regulates the impact of the illiquidity λ relative to the combined effect of expected return and risk aversion. Whereas with linear impact the two elasticities are equal, which means that doubling illiquidity is equivalent to doubling both risk aversion and the expected return, the impact of doubling illiquidity is larger with nonlinear impact, reaching an elasticity of $2/3$ in the limit case of transaction costs.

5.3. **Trading Volume.** Nonlinearity of price impact also controls the dependence of trading volume on the illiquidity parameter λ. The starting point for the analysis of trading volume are equations (18) and (20), which describe the asymptotic turnover v_α and the dynamics of the resulting portfolio.

---

[10] To compute numerically c_α, calculate the solution to equation (15) with initial condition $s_\alpha(0) = 0$ and a sufficiently small guess for c_α. If the guess for c_α is too small, then there exists $\bar{z}$ such that the solution $s_\alpha(z)$ is increasing on $(\bar{z}, +\infty)$. By increasing c_α, the point $\bar{z}$ converges to $+\infty$. The correct guess c_α is the smallest coefficient such that $s_\alpha(z)$ is decreasing on $(0, +\infty)$.

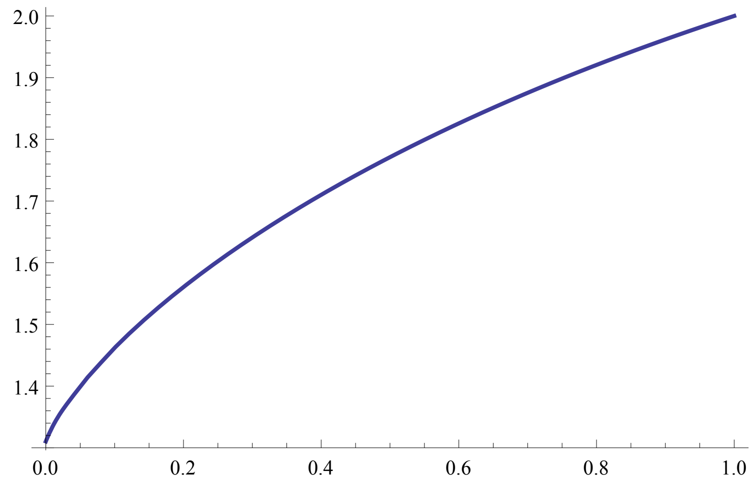

FIGURE 4. The universal constant $(0,1] \ni \alpha \mapsto c_\alpha$.

Regardless of the nonlinearity, turnover v_α increases with both risk aversion γ and volatility σ, confirming the intuition that risk aversion induces more trading through more active rebalancing.[11] Trading volume, measured by unsigned turnover (cf. Guasoni and Weber (2017)) is asymptotically inversely proportional to the width of the confidence intervals of the portfolio weight Y_t, as shown by the next proposition.

Proposition 6. *Assume $\bar{Y} \in (0,1)$. Then*

$$|ET| := \lim_{T\to\infty} \frac{1}{T}\int_0^T |\hat{u}_\lambda(Y_t)|dt = G_\alpha \left[\left(\frac{\sigma^2}{2}\right)^3 \gamma \bar{Y}^4(1-\bar{Y})^4\right]^{\frac{1}{\alpha+3}} \lambda^{-\frac{1}{\alpha+3}} + o(\lambda^{-\frac{1}{\alpha+3}}),$$

where the constant G_α depends only on α

$$G_\alpha = \frac{1}{2}\frac{1-\exp\left(-\frac{2}{(\alpha+1)^{1/\alpha}}\int_0^\infty |s_\alpha(s)|^{1/\alpha}ds\right)}{\int_0^\infty \exp\left(-\frac{2}{(\alpha+1)^{1/\alpha}}\int_0^x |s_\alpha(s)|^{1/\alpha}ds\right)dx}. \tag{23}$$

Thus the equivalent safe rate loss is

$$LOS(\lambda) \sim N_\alpha \lambda |ET|^{\alpha+1}, \tag{24}$$

where also $N_\alpha := c_\alpha G_\alpha^{-(\alpha+1)}$ depends only on α.

Note that (24) extends to nonlinear impact the performance relations in Gerhold et al. (2014) and Guasoni and Weber (2017) for proportional transaction costs and linear impact respectively. The exponent $\alpha + 1$ reflects the fact that higher volume leads to higher costs proportionally to quantities (whence 1), and less than proportional to impact (whence α).

The following proposition shows that the appropriately scaled trading volume process has a weak limit.

---

[11] $v_\alpha(z)$ increases as γ increases because $|s_\alpha(z)|$ is increasing in $|z|$, B_α increasing in γ and A_α decreasing in γ. A_α is increasing in σ, but because $|s_\alpha(z)| \sim K_1|z|$ for small $|z|$ and $|s_\alpha(z)| \sim K_2|z|^{\frac{2\alpha}{\alpha+1}}$ for large $|z|$, the effect of σ on $B_\alpha s_\alpha(z/A_\alpha)$ is always positive. In particular, $v_\alpha(z)$ is increasing in σ.

Proposition 7. *Assume $\bar{Y} \in (0,1)$. As $\lambda \downarrow 0$, the processes $(\lambda^{\frac{1}{\alpha+3}} \hat{u}_\lambda(Z^\lambda))_{\lambda \geq 0}$ converge weakly to the process $V_s := v_\alpha(Z_s^0)$, which has the dynamics*

$$dV_s = b(V_s)ds + a^{1/2}(V_s)dW_s,$$

where $b(v) = v'_\alpha(v_\alpha^{-1}(v))v + \frac{1}{2}v''_\alpha(v_\alpha^{-1}(v))\bar{Y}^2(1-\bar{Y})^2\sigma^2$ and $a(v) = (v'_\alpha(v_\alpha^{-1}(v)))^2\bar{Y}^2(1-\bar{Y})^2\sigma^2$. Thus, $b(v) \sim -2(\alpha+1)^{-1}\alpha^{-1/2}A_\alpha^{-1}B_\alpha^{\frac{\alpha+1}{2\alpha}}v^{\frac{3-\alpha}{2}}$ and $a(v) \sim 4(\alpha+1)^{-2}\alpha^{-1}A_\alpha^{-2}B_\alpha^{\frac{\alpha+1}{\alpha}}\bar{Y}^2(1-\bar{Y})^2\sigma^2 v^{1-\alpha}$ for large v.

The invariant density of the trading volume process V_s is proportional to $\frac{\exp(\int_0^x \frac{2b(s)}{a(s)}ds)}{a(x)}$. It follows from Proposition 7 that for large v it is asymptotically equivalent to

$$f \times \frac{\exp(-g|v|^{\frac{\alpha+3}{2}})}{|v|^{1-\alpha}}, \qquad \text{for some } f, g > 0.$$

Thus, the trading volume generated by the rebalancing activity of *one* large investor has fatter tails than the normal distribution, but still exponentially decaying in (a power of) the trade size. By contrast, Gabaix et al. (2003, 2006) investigate the distribution of trade sizes in aggregate microstructure data, and find tails with power decay, which they explain as arising from the actions of *a population of* funds with assets distributed according to an assumed power law.

This observation suggests that aggregate volume patterns largely depend on the cross-sectional variation among market participants rather than on the time-series variation for each participant, which has rather different statistical properties.

5.4. **Linear Impact and Bid-Ask Spreads.** The formulas in Theorem 5 recover the familiar results on linear impact for $\alpha = 1$ and on proportional bid-ask spreads in the limit $\alpha \downarrow 0$. Linear price impact follows as a straightforward special case, for which both the shape function $s_\alpha(z)$ and the constant c_α have simple explicit expressions. Substituting such expressions in Theorem 5 in turn leads to the asymptotic formulas for optimal policies, their performance, and trading volume derived by Guasoni and Weber (2017).

Lemma 8. *In Theorem 5, $c_1 = 2$ and $s_1(z) = -2z$.*

In the limit $\alpha \downarrow 0$, price impact degenerates to a market with proportional bid-ask spreads, for which the asymptotics are discussed in Gerhold et al. (2014). In the notation of this paper, the asymptotic analysis in the transaction cost literature solves models for $\alpha = 0$ and then studies the small-cost regime $\lambda \downarrow 0$. By contrast, this paper studies the regime $\lambda \downarrow 0$ while holding $\alpha > 0$ constant, and a natural question is whether the limit $\alpha \downarrow 0$ of these asymptotics coincides with the asymptotics with $\alpha = 0$ in the transaction cost literature. As we show, the answer is affirmative.

The next lemma identifies the limits of s_α, c_α in Theorem 5.

Lemma 9. *In Theorem 5, $c_0 := \lim_{\alpha \to 0} c_\alpha = \left(\frac{3}{2}\right)^{2/3}$ and*

$$s_0(z) := \lim_{\alpha \to 0} s_\alpha(z) = \begin{cases} 1, & z \in (-\infty, -\sqrt{c_0}], \\ z^3/3 - c_0 z, & z \in (-\sqrt{c_0}, \sqrt{c_0}), \\ -1, & z \in [\sqrt{c_0}, +\infty). \end{cases}$$

The next lemma shows that, as $\alpha \downarrow 0$, the optimal trading rate converges to 0 inside the no-trade region $[z_-, z_+]$, to $-\infty$ in the sell region $[z_+, +\infty]$, and to $+\infty$ in the buy-region $[-\infty, z_-]$.

Lemma 10. *Define* $z_+ := \left(\frac{3}{2\gamma}\bar{Y}^2(1-\bar{Y})^2\right)^{1/3}$ *and* $z_- := -z_+$. *For* $\alpha \downarrow 0$, *the optimal strategy* $v_\alpha(z)$ *converges to* 0 *on* (z_-, z_+), *to* $+\infty$ *on* $(-\infty, z_-)$, *and to* $-\infty$ *on* $(z_+, +\infty)$.

6. Heuristics

The proof of Theorem 5 is technical and deferred to the appendix, as it requires a careful analysis of the convergence of the rescaled value functions to their universal limit. Yet, an informal derivation of the main statement illustrates the main scaling arguments at work.

Denoting by $v(t,x,y) = E\left[\frac{X_T^{1-\gamma}}{1-\gamma}|X_t = x, Y_t = y\right]$ the value function of the finite-horizon utility maximization problem, similar arguments as those for linear impact (Guasoni and Weber, 2017) lead to the Hamilton-Jacobi-Bellman equation

$$v_t + y(1-y)(\mu - \sigma^2 y)v_y + \mu x y v_x + \frac{\sigma^2 y^2}{2}\left(x^2 v_{xx} + (1-y)^2 v_{yy} + 2x(1-y)v_{xy}\right) + \max_u \left(-\lambda x|u|^{\alpha+1}v_x + v_y\left(u + \lambda y|u|^{\alpha+1}\right)\right) = 0.$$

Next, using the ansatz $v(t,x,y) = \frac{x^{1-\gamma}}{1-\gamma}e^{(1-\gamma)(\beta(T-t)+\int_{y_0}^y q(z)dz)}$, maximizing over u, and setting $c(\lambda) := \frac{\mu^2}{2\gamma\sigma^2} - \beta$ yields

(25)
$$-\frac{\gamma\sigma^2}{2}(\bar{Y}-y)^2 + c(\lambda) + y(1-y)(\mu - \gamma\sigma^2 y)q + \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}}\frac{|q|^{\frac{\alpha+1}{\alpha}}}{(1-yq)^{1/\alpha}}\lambda^{-1/\alpha} + \frac{\sigma^2}{2}y^2(1-y)^2(q' + (1-\gamma)q^2) = 0.$$

Passing to the limit $\lambda \to 0$ and guessing that $\lim_{\lambda\to 0} q_\lambda(y) = 0$, leads to

$$\frac{\gamma\sigma^2}{2}(\bar{Y}-y)^2 = \lim_{\lambda\to 0}\alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}}|q_\lambda|^{\frac{\alpha+1}{\alpha}}\lambda^{-1/\alpha},$$

and hence to the first order approximation

$$q^{(1)}(y) = \lambda^{\frac{1}{\alpha+1}}(\alpha+1)^{\frac{1}{\alpha+1}}\left(\frac{\alpha+1}{\alpha}\frac{\gamma\sigma^2}{2}\right)^{\frac{\alpha}{\alpha+1}}|\bar{Y}-y|^{\frac{2\alpha}{\alpha+1}}\operatorname{sgn}(\bar{Y}-y). \tag{26}$$

For $\alpha \in (0,1]$, the function $q^{(1)}(y)$ is sublinear and the derivative explodes in $\bar{Y}$, even as the solution of the HJB equation is bounded and has bounded derivative. This observation suggests that to achieve nontrivial asymptotics it is necessary to "zoom in" close to the Merton proportion $\bar{Y}$, i.e. to rescale $q_\lambda(y)$ both on the horizontal and on the vertical axis.

For y far from $\bar{Y}$, the dominant term in (25) is $-\frac{\gamma\sigma^2}{2}(\bar{Y}-y)^2$, which yields the approximation (26). Instead, for $y = \bar{Y}$ the dominant term is $c(\lambda)$ in view of the guess $q_\lambda(\bar{Y}) = 0$ (a zero trading rate at the Merton proportion). In particular, if $c(\lambda) = O(\lambda^d)$, setting in (25) $y = \bar{Y}$, it follows that $q'_\lambda(\bar{Y}) = O(\lambda^d)$. The two dominant terms (for y far from $\bar{Y}$ and for $y = \bar{Y}$) match when $y \sim \bar{Y} - c(\lambda)^{1/2}$.

Setting $y = \bar{Y} - c(\lambda)^{1/2}$ in (26) yields $q^{(1)}(\bar{Y} - c(\lambda)^{1/2}) = O(\lambda^{\frac{1}{\alpha+1}+\frac{\alpha d}{\alpha+1}})$. On the other hand, as $q'_\lambda(\bar{Y}) = O(\lambda^d)$, $q_\lambda(\bar{Y} - c(\lambda)^{1/2}) \sim -q'_\lambda(\bar{Y}) \times c(\lambda)^{1/2} = O(\lambda^{\frac{3d}{2}})$. From the relation $\frac{3d}{2} = \frac{1}{\alpha+1} + \frac{\alpha d}{\alpha+1}$, the correct guess $d = \frac{2}{\alpha+3}$ follows. Therefore, assuming $c(\lambda) = \bar{c}\lambda^{\frac{2}{\alpha+3}}$ and substituting $y = \bar{Y} + \lambda^{\frac{1}{\alpha+3}} z$ and $r_\lambda(z) = q_\lambda(y)\lambda^{-\frac{3}{\alpha+3}}$ in equation (25), yields

(27)
$$-\frac{\gamma\sigma^2}{2}z^2\lambda^{\frac{2}{\alpha+3}} + \bar{c}\lambda^{\frac{2}{\alpha+3}} - \gamma\sigma^2 y(1-y)z\lambda^{\frac{4}{\alpha+3}}r_\lambda + \frac{\sigma^2}{2}y^2(1-y)^2(r'_\lambda\lambda^{\frac{2}{\alpha+3}} + (1-\gamma)r_\lambda^2\lambda^{\frac{6}{\alpha+3}}) + \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}}\frac{|r_\lambda|^{\frac{\alpha+1}{\alpha}}}{(1-yr_\lambda\lambda^{\frac{3}{\alpha+3}})^{1/\alpha}}\lambda^{\frac{2}{\alpha+3}} = 0.$$

Dividing the equation by $\lambda^{\frac{2}{\alpha+3}}$ and passing to the limit as $\lambda \downarrow 0$, it follows that $r_0(z) := \lim_{\lambda\to 0} r_\lambda(z)$ satisfies the equation

$$-\frac{\gamma\sigma^2}{2}z^2 + \bar{c} + \frac{\sigma^2}{2}\bar{Y}^2(1-\bar{Y})^2 r'_0 + \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}}|r_0|^{\frac{\alpha+1}{\alpha}} = 0.$$

Absorbing the parameters μ, σ and γ in the coefficients through the substitution $s_\alpha(w) = \frac{r_0(A_\alpha w)}{B_\alpha}$, the asymptotic HJB equation (15) follows. Likewise, reformulating the boundary conditions (12)-(13) for q_λ in terms of s_α yields the asymptotic growth conditions (16).

7. Model Extension

The arguments in the paper can be extended to more general specifications of the price impact function, as to account for dependence on a stationary process. Assume that the average execution price for the large investor is

$$S_t\left(1 + \lambda K(\xi_t)\left|\frac{S_t\Delta\theta}{X_t\Delta t}\right|^\alpha\right),$$

where ξ is an additional exogenous state variable and $0 < m_K \le K(\xi) \le M_K$, and ξ follows the SDE

$$d\xi_t = \mu_\xi(\xi_t)dt + \sigma_\xi(\xi_t)dB_t, \tag{28}$$

where μ_ξ, σ_ξ are smooth functions and B_t is another Brownian motion that has correlation ρ with W_t.

The coefficient $K(\xi_t)$ captures both the investor's size and the time-variation in illiquidity. Under this interpretation, the model introduced in Section 3 assumes a constant proportion between the daily trading volume and investor's wealth, where the constant of proportionality is incorporated into the parameter λ.

Passing to continuous time and applying Itô's Lemma yields the dynamics of the wealth process X and of the risky weight Y,

$$\frac{dX_t}{X_t} = Y_t(\mu dt + \sigma dW_t) - \lambda K(\xi_t)|u_t|^{\alpha+1}dt,$$
$$dY_t = (Y_t(1-Y_t)(\mu - Y_t\sigma^2) + u_t + \lambda K(\xi_t)Y_t|u_t|^{\alpha+1})dt + Y_t(1-Y_t)\sigma dW_t.$$

As in Definition 1, a strategy (Y_0, u) is admissible if the resulting wealth process, i.e.,

$$X_T^u = X_0\exp\left(\int_0^T\left(\mu Y_t - \frac{\sigma^2}{2}Y_t^2 - \lambda K(\xi_t)|u_t|^{1+\alpha}\right)dt + \int_0^T \sigma Y_t dW_t\right),$$

remains solvent at all times.[12]

[12] If $0 < m_K \le K(\xi) \le M_K$, then the same arguments in Remark 2 apply to the extended model. In particular, any strategy such that $Y_t \notin [0,1]$ for some t is not admissible.

The following verification theorem identifies the optimal trading policy and its equivalent safe rate in the extended model from the solution of the associated ergodic HJB equation.

Theorem 11. *Assume $\bar{Y} := \frac{\mu}{\gamma\sigma^2} \in (0,1)$. Let the process $(\xi_t)_{t\geq 0}$ satisfy* (28) *for some continuous functions $\mu_\xi, \sigma_\xi : E \mapsto \mathbb{R}$, where $E \subset \mathbb{R}$ is a bounded open set and $P(\xi_t \in E$ a.s. for all $t) = 1$. Let $K : E \mapsto \mathbb{R}$ be bounded from above and away from* 0 *(i.e., there are $m_K, M_K \in \mathbb{R}$ such that $0 < m_K \leq K(\xi) \leq M_K$ for all $\xi \in E$).*

Assume that $Q(y,\xi) \in C^2([0,1] \times E)$ (i) is bounded, (ii) satisfies the conditions $Q_y(1,\xi) < 0 < Q_y(0,\xi)$ for $\xi \in E$ and $yQ_y(y,\xi) < 1$ for $(y,\xi) \in [0,1] \times E$, (iii) is such that $\sigma_\xi(\xi)Q_\xi(y,\xi)$ is bounded, and (iv) solves the partial differential equation

$$\begin{aligned}&-\hat{\beta} + \mu y - \frac{\gamma\sigma^2}{2}y^2 + y(1-y)(\mu - \gamma\sigma^2 y)Q_y + \frac{\sigma^2}{2}y^2(1-y)^2(Q_{yy} + (1-\gamma)Q_y^2)\\ &+ \mu_\xi(\xi)Q_\xi + \frac{\sigma_\xi^2(\xi)}{2}\left((1-\gamma)Q_\xi^2 + Q_{\xi\xi}\right) + \sigma\rho\sigma_\xi(\xi)y(1-y)\left((1-\gamma)Q_yQ_\xi + Q_{y\xi}\right)\\ &+ \sigma\rho\sigma_\xi(\xi)y(1-\gamma)Q_\xi + \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}}\frac{|Q_y|^{\frac{\alpha+1}{\alpha}}}{(1-yQ_y)^{1/\alpha}(K(\xi)\lambda)^{1/\alpha}} = 0\end{aligned} \tag{29}$$

for some $\hat{\beta} \in \mathbb{R}$. Then for sufficiently small λ the optimal trading policy and its equivalent safe rate for the problem (8) *are*

$$\hat{u}(y,\xi) = \left|\tfrac{Q_y(y,\xi)}{(\alpha+1)\lambda K(\xi)(1-yQ_y(y,\xi))}\right|^{1/\alpha}\operatorname{sgn}(Q_y(y,\xi)), \qquad \mathrm{ESR}_\gamma(\hat{u}) = \hat{\beta}. \tag{30}$$

7.1. **State-dependent Asymptotics.** The asymptotic expansions extend to the state-dependent illiquidity $K(\xi)$ as follows. Denote by $\hat{\beta} = \frac{\mu^2}{2\gamma\sigma^2} - c(\lambda)$. As the frictionless limit corresponds to $Q(y,\xi) = c(\lambda) = 0$, it is reasonable to conjecture that both $Q(y,\xi)$ and $c(\lambda)$ vanish with λ, and hence that the Hamilton-Jacobi-Bellman equation (29) converges to

$$\frac{\gamma\sigma^2}{2}(y-\bar{Y})^2 = \lim_{\lambda\to 0}\alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}}\frac{|Q_y|^{\frac{\alpha+1}{\alpha}}}{(1-yQ_y)^{1/\alpha}(K(\xi))^{1/\alpha}}\lambda^{-1/\alpha}. \tag{31}$$

This expression in turn suggests for y far from $\bar{Y}$ an expansion of the form

$$Q(y,\xi) = \lambda^{1/(\alpha+1)}Q^{(1)}(y,\xi) + H^\lambda(\xi),$$

with which, for small λ, equation (29) reduces to

$$|Q_y^{(1)}(y,\xi)| = K(\xi)^{\frac{1}{\alpha+1}}(\alpha+1)^{\frac{1}{\alpha+1}}\left(\frac{\alpha+1}{\alpha}\frac{\gamma\sigma^2}{2}\right)^{\frac{\alpha}{\alpha+1}}((\bar{Y}-y)^2 + \bar{A}(\xi) + y\bar{B}(\xi))^{\frac{\alpha}{\alpha+1}},$$

where

$$\begin{aligned}\bar{A}(\xi) &= -\mu(\xi)H_\xi^\lambda(\xi) - \frac{\sigma^2(\xi)}{2}((1-\gamma)H_\xi^\lambda(\xi)^2 + H_{\xi\xi}^\lambda(\xi)),\\ \bar{B}(\xi) &= -\sigma\rho\sigma(\xi)(1-\gamma)H_\xi^\lambda(\xi).\end{aligned}$$

In view of the boundary conditions $Q_y(1,\xi) \leq 0 \leq Q_y(0,\xi)$, it follows that $\min_y((\bar{Y} - y)^2 + \bar{A}(\xi) + y\bar{B}(\xi)) = 0$. As the minimum is attained at $y^* = \bar{Y} - \frac{\bar{B}(\xi)}{2}$, H_ξ^λ satisfies the ordinary differential equation

$$0 = \bar{A}(\xi) + \bar{Y}\bar{B}(\xi) - \frac{\bar{B}(\xi)^2}{4} \tag{32}$$

with the solution $H^\lambda_\xi \equiv 0$ corresponding to a frictionless market, which suggests that H^λ_ξ is null at the first order, and hence that

$$|Q_y^{(1)}(y,\xi)| = K(\xi)^{\frac{1}{\alpha+1}}(\alpha+1)^{\frac{1}{\alpha+1}}\left(\frac{\alpha+1}{\alpha}\frac{\gamma\sigma^2}{2}\right)^{\frac{\alpha}{\alpha+1}}|\bar{Y}-y|^{\frac{2\alpha}{\alpha+1}}. \tag{33}$$

Consider now the asymptotic expansion close to $\bar{Y}$, applying the substitutions

$$z := \frac{y-\bar{Y}}{\lambda^{1/(\alpha+3)}}, \quad Q(y,\xi) := \lambda^{4/(\alpha+3)} \times R^\lambda(z,\xi) + H^\lambda(\xi). \tag{34}$$

For small λ, writing $\hat{\beta} = \frac{\mu^2}{2\gamma\sigma^2} - l\lambda^{2/(\alpha+3)}$, equation (29) reduces to

$$-\frac{\gamma\sigma^2}{2}z^2 + l + C^\lambda(\xi) + \frac{\sigma^2}{2}\bar{Y}^2(1-\bar{Y})^2 R^\lambda_{zz} + \frac{\alpha}{\alpha+1}(\alpha+1)^{-1/\alpha}|R^\lambda_z|^{\frac{\alpha+1}{\alpha}}K(\xi)^{-1/\alpha} = 0, \tag{35}$$

where $C^\lambda(\xi) := -(\bar{A}(\xi)+\bar{B}(\xi)\bar{Y})\lambda^{-2/(\alpha+3)}$. Through the state-dependent change of variable obtained by

$$A_\alpha = \left(\frac{2l_\alpha}{\gamma\sigma^2}\right)^{1/2} K(\xi)^{\frac{1}{\alpha+3}}, \qquad B_\alpha = l_\alpha^{\frac{\alpha}{\alpha+1}} K(\xi)^{\frac{3}{\alpha+3}}, \tag{36}$$

the ordinary differential equation reduces again to the same shape equation (15) (with the same constant l_α). In summary, the shape equation continues to describe the optimal trading strategy even when liquidity is stochastic, as liquidity level affects the scaling, but not the shape of the trading response.

Appendix A

A.1. Proof of Theorem 3. Recall the HJB equation

$$\begin{aligned} q' = \frac{2}{\sigma^2 y^2(1-y)^2}\Big(&\frac{\gamma\sigma^2}{2}(\bar{Y}-y)^2 - c + \gamma\sigma^2 y(1-y)(y-\bar{Y})q - \frac{\sigma^2}{2}y^2(1-y)^2(1-\gamma)q^2 \\ &- \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}}\frac{|q|^{\frac{\alpha+1}{\alpha}}}{(1-yq)^{1/\alpha}}\lambda^{-1/\alpha}\Big) =: f(y,q), \end{aligned} \tag{37}$$

where $c := \frac{\mu^2}{2\gamma\sigma^2} - \beta$.

The proof of Theorem 3 begins with some useful properties of equation (37). Lemma 13 shows the existence of a bounded solution to (37) with positive limit in 0 and negative limit in 1. The boundary conditions in 0 and 1 are required to prove in Lemma 16 that the process Y_t lives in the interval $[0,1]$. Lemma 17 provides a bound for the investor's utility. Finally, Theorem 3 follows from these results.

Lemma 12. *Assume that $c \in (0, \frac{\mu^2}{2\gamma\sigma^2})$, where c may depend on λ. Then:*

(i) There exists $K > 0$ such that $f(y, \pm K\lambda^{\frac{1}{\alpha+1}}) < 0$ on $(0,1)$ for λ small enough.

(ii) For λ small enough, the set of points $(y,h) \in \left(0, \bar{Y} - \sqrt{\frac{2}{\gamma\sigma^2}c}\right) \times (0, K\lambda^{\frac{1}{\alpha+1}})$ such that $f(y,h) = 0$ is the graph of a differentiable strictly decreasing function $y \to h(y)$. The same holds true for the set of points $(y,h) \in \left(\bar{Y} + \sqrt{\frac{2}{\gamma\sigma^2}c}, 1\right) \times (-K\lambda^{\frac{1}{\alpha+1}}, 0)$ such that $f(y,h) = 0$.

(iii) The limits of function $h(y)$ in 0 and in 1 satisfy

$$h(0^+) = \lambda^{\frac{1}{\alpha+1}}(\alpha+1)^{\frac{1}{\alpha+1}}\left(\frac{\alpha+1}{\alpha}\beta\right)^{\frac{\alpha}{\alpha+1}},$$

$$\alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}}\frac{|h(1^-)|^{\frac{\alpha+1}{\alpha}}}{(1-h(1^-))^{1/\alpha}}\lambda^{-1/\alpha} = \beta-\mu+\gamma\frac{\sigma^2}{2}.$$

Proof.

(i) Substitute $q = \pm K\lambda^{\frac{1}{\alpha+1}}$ in equation (37) and expand around $\lambda = 0$. It follows that $f(y, \pm K\lambda^{\frac{1}{\alpha+1}}) = \frac{2}{\sigma^2 y^2(1-y)^2}\left(\frac{\gamma\sigma^2}{2}(\bar{Y}-y)^2 - c - \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}}K^{\frac{\alpha+1}{\alpha}}\right) + o(1)$. The claim follows immediately if K is chosen large enough.

(ii) Because for λ sufficiently small $\frac{\partial}{\partial q}[\sigma^2 y^2(1-y)^2 f(y,q)] < 0$ on $\{\bar{Y}\} \times [0, K\lambda^{\frac{1}{\alpha+1}}]$ and $f(\bar{Y}, 0) < 0$, also $f(\bar{Y}, q) < 0$ for $q \in [0, K\lambda^{\frac{1}{\alpha+1}}]$. An explicit calculation shows that $\frac{\partial^2}{\partial y^2}[\sigma^2 y^2(1-y)^2 f(y,q)] > 0$ on $[0,1] \times [-K\lambda^{\frac{1}{\alpha+1}}, K\lambda^{\frac{1}{\alpha+1}}]$ for λ sufficiently small. In particular, for any $q \in [0, K\lambda^{\frac{1}{\alpha+1}}]$, the function $\sigma^2 y^2(1-y)^2 f(y,q)$ is strictly convex on $[0,1]$. Therefore, if $\lim_{y\to 0^+}\sigma^2 y^2(1-y)^2 f(y,\tilde{q}) < 0$, where $\tilde{q} \in [0, K\lambda^{\frac{1}{\alpha+1}}]$, then $\sigma^2 y^2(1-y)^2 f(y,\tilde{q}) < 0$ for $y \in [0, \bar{Y}]$. Instead, if $\lim_{y\to 0^+}\sigma^2 y^2(1-y)^2 f(y,\tilde{q}) > 0$, then there exists a unique $\tilde{h}(\tilde{q}) := \tilde{y} \in [0, \bar{Y}]$ such that $f(\tilde{y}, \tilde{q}) = 0$. It is verified immediately that $\tilde{h}(0) = \bar{Y} - \sqrt{\frac{2}{\gamma\sigma^2}c}$. Because $f(y, 0) > 0$ and $f(y, K\lambda^{\frac{1}{\alpha+1}}) < 0$ on $(0, \bar{Y} - \sqrt{\frac{2}{\gamma\sigma^2}c})$, it follows that $h(y) := \tilde{h}^{-1}(y)$ defines a strictly decreasing positive function on $(0, \bar{Y} - \sqrt{\frac{2}{\gamma\sigma^2}c})$. The same arguments applied on $(\bar{Y} + \sqrt{\frac{2}{\gamma\sigma^2}c}, 1)$ prove the second part of the claim.

(iii) It is enough to solve the equation $\lim_{y\to 0^+}\sigma^2 y^2(1-y)^2 f(y, h(y)) = 0$, resp. $\lim_{y\to 1^-}\sigma^2 y^2(1-y)^2 f(y, h(y)) = 0$, for $h(0^+)$, resp. $h(1^-)$, where $f(y,q)$ is defined in (37).

□

Lemma 13. *Assume $\bar{Y} = \frac{\mu}{\gamma\sigma^2} \in (0,1)$. There exists $\lambda^* > 0$ such that for $\lambda \in (0, \lambda^*)$ and for a certain $c(\lambda) \in (0, \frac{\mu^2}{2\gamma\sigma^2})$ the equation* (37) *has a unique solution $q : [0,1] \mapsto \mathbb{R}$ with boundary conditions $q(0^+) = h(0^+) > 0$ and $q(1^-) = h(1^-) < 0$, where $h(\cdot)$ is defined in Lemma 12. Furthermore, $c(\lambda)$ is a continuous increasing function on $(0, \lambda^*)$ and $\lim_{\lambda\to 0} c(\lambda) = 0$.*

Proof. The proof follows the same lines as the proof of Lemma A.8 in Guasoni and Weber (2017) and is broken into the following steps:

(i) for a fixed $c < \frac{\mu^2}{2\gamma\sigma^2}$, there exists a unique solution $q_0(\cdot)$ to (37) with limit in 0 equal to $h(0^+)$;

(ii) for a fixed $c < \frac{\mu^2}{2\gamma\sigma^2} - \mu + \frac{\gamma}{2}\sigma^2$, there exists a unique solution $q_1(\cdot)$ to (37) with limit in 1 equal to $h(1^-)$;

(iii) if $c \le 0$, then $q_0(\cdot) > 0 > q_1(\cdot)$;

(iv) for a fixed $c > 0$, there exists $\tilde{\lambda} > 0$ such that $q_0(y) < q_1(y)$ on their common domain for $\lambda \in (0, \tilde{\lambda})$;

(v) by continuity, there exists $c(\lambda) \in (0, \frac{\mu^2}{2\gamma\sigma^2})$ such that $q_0(\cdot) = q_1(\cdot)$ on $(0,1)$. The function $c(\lambda)$ is continuous, increasing and $\lim_{\lambda\to 0} c(\lambda) = 0$.

Let $h(y)$ be as in Lemma 12, i.e., $f(y, h(y)) = 0$, and let $\mathcal{D}_h := \left(0, \bar{Y} - \sqrt{\frac{2}{\gamma\sigma^2}c}\right) \cup \left(\bar{Y} + \sqrt{\frac{2}{\gamma\sigma^2}c}, 1\right)$ be its domain. Denote by $q(y; \bar{y}, \bar{w})$ the solution to (37) with initial condition $q(\bar{y}) = \bar{w}$.

Remark 14. *For every $y \in (0,1)$, $\lim_{q\to(\frac{1}{y})^-} f(y,q) = -\infty$. Thus, every solution to* (37) *starting below the curve $\frac{1}{y}$ must remain below this curve.*

Step (i).

The slope field in equation (37) satisfies $\lim_{y\to 0^+} \frac{\sigma^2 y^2}{2} f(y,q) = \frac{\mu^2}{2\gamma\sigma^2} - c - \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}} |q|^{\frac{\alpha+1}{\alpha}} \lambda^{-1/\alpha}$. In particular, if $c < \frac{\mu^2}{2\gamma\sigma^2}$, i.e., if $\beta > 0$, and y is sufficiently close to 0, then $f(y,q) > 0$ for $q > \lambda^{\frac{1}{\alpha+1}} (\alpha+1)^{\frac{1}{\alpha+1}} \left(\frac{\alpha+1}{\alpha}\beta\right)^{\frac{\alpha}{\alpha+1}} = h(0^+)$. Consider a solution $\tilde{q}(y)$ to (37) such that $\tilde{q}(\cdot) \geq h(\cdot)$, it follows that $\tilde{q}$ has a limit in 0. Assume the limit $k := \tilde{q}(0^+)$ is finite. Then $\lim_{y\to 0} y^2 \tilde{q}'(y) = \frac{2}{\sigma^2}\left(\frac{\mu^2}{2\gamma\sigma^2} - c - \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}} |k|^{\frac{\alpha+1}{\alpha}} \lambda^{-1/\alpha}\right)$. The mean-value theorem implies that $\tilde{q}(y) - \tilde{q}(0^+) = y\tilde{q}'(\eta)$ for some $\eta \in [0, y]$, multiplying both sides by y and passing to the limit $y \downarrow 0$, it follows that $y^2\tilde{q}'(y)$ converges to zero. Hence, k satisfies $\frac{\mu^2}{2\gamma\sigma^2} - c - \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}} |k|^{\frac{\alpha+1}{\alpha}} \lambda^{-1/\alpha} = 0$ and $k \geq h(0^+)$, and therefore $k = h(0^+)$.

Define $q_0(y) := \inf\{\tilde{q}(y) : \tilde{q}(\cdot) \text{ solves (37)}, \tilde{q}(\cdot) \geq h(\cdot) \text{ on } \mathcal{D}_h \cap (0, \bar{Y})\}$ for $y \in \mathcal{D}_h \cap (0, \bar{Y})$. Note that for any $\bar{y} \in \mathcal{D}_h \cap (0, \bar{Y})$ the function $q(y; \bar{y}, q_0(\bar{y}))$ coincides with $q_0(y)$ on $\mathcal{D}_h \cap (0, \bar{Y})$. In other terms, $q_0(\cdot)$ is a solution to (37). Next, assume by contradiction that $\lim_{y\to 0} q_0(y) = +\infty$. Then there exist y_1 sufficiently close to 0 and w_1 such that $q_0(y_1) > w_1 > h(0^+) > h(y_1)$. Because h is a subsolution to (37), it follows that $q_0(\cdot) > q(\cdot; y_1, w_1) > h(\cdot)$ on $\mathcal{D}_h \cap (0, \bar{Y})$, contradicting the minimality of q_0. Therefore, $\lim_{y\to 0} q_0(y) = h(0^+) = \lambda^{\frac{1}{\alpha+1}} (\alpha+1)^{\frac{1}{\alpha+1}} \left(\frac{\alpha+1}{\alpha}\beta\right)^{\frac{\alpha}{\alpha+1}}$.

To prove uniqueness, note that $\frac{\partial}{\partial q}[\sigma^2 y^2 (1-y)^2 f(y,q)] < 0$ on $[0,\varepsilon] \times [0, K\lambda^{\frac{1}{\alpha+1}}]$ for ε and λ sufficiently small. Hence, for any solution $\tilde{q}(\cdot)$ to (37) such that $0 < \tilde{q}(\cdot) < q_0(\cdot)$, it holds that $\tilde{q}'(y) > q_0'(y)$ on $(0,\varepsilon)$. It follows that $\tilde{q}(\cdot)$ and $q_0(\cdot)$ cannot have the same limit in 0. Analogously, any solution $\tilde{q}(\cdot) > q_0(\cdot)$ to (37) cannot have the same limit in 0 as $q_0(\cdot)$.

Step (ii).

This part of the proof follows the same lines as in the proof of Step (i). Analogous arguments as in Step (i) prove that for any solution $\tilde{q}(\cdot)$ to (37) such that $\tilde{q}(\cdot) \leq h(\cdot)$ the limit $\lim_{y\to 1} \tilde{q}(y)$ is either $h(1^-)$ or $-\infty$. Define $q_1(y) := \sup\{\tilde{q}(y) : \tilde{q}(\cdot) \text{ solves (37)}, \tilde{q}(\cdot) \leq h(\cdot) \text{ on } \mathcal{D}_h \cap (\bar{Y}, 1)\}$ for $y \in \mathcal{D}_h \cap (\bar{Y}, 1)$. The function $q_1(\cdot)$ is also a solution to (37). Assume by contradiction that $\lim_{y\to 1} q_1(y) = -\infty$. Then there exist y_2 sufficiently close to 1 and w_2 such that $q_1(y_2) < w_2 < h(1^-) < h(y_2)$. Because h is a subsolution to (37), it follows that $q_1(\cdot) < q(\cdot; y_2, w_2) < h(\cdot)$ on $\mathcal{D}_h \cap (\bar{Y}, 1)$, contradicting the maximality of q_1. Hence, $\lim_{y\to 1} q_1(y) = h(1^-)$. Uniqueness follows as in Step (i), after noting that $\frac{\partial}{\partial q}[\sigma^2 y^2 (1-y)^2 f(y,q)] > 0$ on $[1-\varepsilon, 1] \times [-K\lambda^{\frac{1}{\alpha+1}}, 0]$ for ε and λ sufficiently small.

Step (iii).

If $c \leq 0$, then $f(y, 0) \geq 0$ for $y \in (0,1)$. This implies that $q_0(\cdot) > 0$ and $q_1(\cdot) < 0$.

Step (iv).

The functions $q_{0,\lambda}(y) := q_0(y)\lambda^{-\frac{1}{\alpha+1}}$ and $q_{1,\lambda}(y) := q_1(y)\lambda^{-\frac{1}{\alpha+1}}$ solve the equation

$$q'_\lambda = \frac{2\lambda^{-\frac{1}{\alpha+1}}}{\sigma^2 y^2(1-y)^2}\Big(\frac{\gamma\sigma^2}{2}(\bar{Y}-y)^2 - c - \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}}\frac{|q_\lambda|^{\frac{\alpha+1}{\alpha}}}{(1-yq)^{1/\alpha}}$$
$$+\lambda^{\frac{1}{\alpha+1}}[\gamma\sigma^2 y(1-y)(y-\bar{Y})q_\lambda - \lambda^{\frac{1}{\alpha+1}}\frac{\sigma^2}{2}y^2(1-y)^2(1-\gamma)q_\lambda^2]\Big) =: f_\lambda(y,q_\lambda).$$

First, observe that from Lemma 12(i) it follows that $q_{1,\lambda}(\cdot) > -K$. Because $\frac{\gamma\sigma^2}{2}(\bar{Y}-y)^2 - c - \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}}\frac{|q_\lambda|^{\frac{\alpha+1}{\alpha}}}{(1-yq)^{1/\alpha}} < -\varepsilon_1$ for $(y,q_\lambda) \in [\bar{Y}-\varepsilon_2, \bar{Y}+\varepsilon_2]\times[-K,K]$ for some $\varepsilon_1 > 0$, $\varepsilon_2 > 0$ and with K as in Lemma 12(i), $\lim_{\lambda\to 0} f_\lambda(y,q_\lambda) = -\infty$ uniformly on $[\bar{Y}-\varepsilon_2, \bar{Y}+\varepsilon_2]\times[-K,K]$. In particular, for λ sufficiently small, either there exists $y^* < \bar{Y}-\varepsilon_2$ such that $\lim_{y\to y^*} q_{0,\lambda}(y) = -\infty$ or there exists $y_* \in [\bar{Y}-\varepsilon_2, \bar{Y}+\varepsilon_2]$ such that $q_{0,\lambda}(y_*) = -K$. Recalling that $q_{1,\lambda}(\cdot) > -K$, this proves the claim.

Step (v).

Because $q_0(\cdot)$ and $q_1(\cdot)$ depend continuously on c, there exists $c(\lambda) \in (0, \frac{\mu^2}{2\gamma\sigma^2})$ such that $q_0(\cdot) = q_1(\cdot)$ on $(0,1)$. This solution satisfies the properties of the statement.

To prove that $c(\lambda)$ is increasing, consider $\lambda_1 < \lambda_2$ and define $c_1 := c(\lambda_1)$ and $c_2 := c(\lambda_2)$. We make the dependence of the slope field function $f(y,q)$, resp. a solution q, on λ and c explicit by writing $f^{\lambda,c}(y,q)$, resp. $q^{\lambda,c}$. Because $f^{\lambda_1,c_2}(y,q) < f^{\lambda_2,c_2}(y,q)$, $q_0^{\lambda_1,c_2}(0) < q_0^{\lambda_2,c_2}(0)$ and $q_1^{\lambda_1,c_2}(1) > q_1^{\lambda_2,c_2}(1)$, it follows that $q_0^{\lambda_1,c_2} < q_0^{\lambda_2,c_2} = q_1^{\lambda_2,c_2} < q_1^{\lambda_1,c_2}$. Thus, $c(\lambda_1) = c_1 < c_2 = c(\lambda_2)$.

Assume $c(\lambda)$ is not continuous in $\bar{\lambda}$. Then there exists an $\varepsilon > 0$ and a sequence $\lambda_n \downarrow \bar{\lambda}$ such that $|c(\bar{\lambda}) - c(\lambda_n)| > \varepsilon$. For a sufficiently large n, $q^{\lambda_n}(0) < q^{\bar{\lambda}}(0)$, $f^{\lambda_n} < f^{\bar{\lambda}}$ and $q^{\lambda_n}(1) > q^{\bar{\lambda}}(1)$, which leads to a contradiction. Thus, $c(\lambda)$ must be continuous.

Assume now by contradiction that $\limsup_{\lambda\to 0} c(\lambda) = c^* > 0$. It follows that there exists a sequence $\lambda_n \downarrow 0$ such that $c(\lambda_n) > \frac{c^*}{2}$ for all n and $q_0(\cdot) = q_1(\cdot)$ for $\lambda = \lambda_n$. This contradicts the claim in Step (iv) of the proof. Hence, $\lim_{\lambda\to 0} c(\lambda) = 0$. □

Remark 15. *Let $h(y)$ be as in Lemma 12 and let $\mathcal{D}_h$ be its domain. From the definition of $q_0(y)$ and $q_1(y)$ in the proof of Lemma 13, it follows that the function $q(y)$ in Lemma 13 satisfies*

(38)

$$q(y) = \inf\{\tilde{q}(y) : \tilde{q}(\cdot) \text{ solves } (37),\ \tilde{q}(\cdot) \geq h(\cdot) \text{ on } \mathcal{D}_h \cap (0,\bar{Y})\} \quad \text{for } \bar{Y} > y \in \mathcal{D}_h$$

(39)

$$q(y) = \sup\{\tilde{q}(y) : \tilde{q}(\cdot) \text{ solves } (37),\ \tilde{q}(\cdot) \leq h(\cdot) \text{ on } \mathcal{D}_h \cap (\bar{Y},1)\} \quad \text{for } \bar{Y} < y \in \mathcal{D}_h.$$

Lemma 16. *Let $u(y)$ be a bounded continuous function on $[0,1]$ such that $u(0) > 0$ and $u(1) < 0$. For λ small enough the process with dynamics*

$$dY_t = Y_t(1-Y_t)(\mu - Y_t\sigma^2)dt + (u(Y_t) + \lambda Y_t|u(Y_t)|^{\alpha+1})dt + Y_t(1-Y_t)\sigma dW_t,$$
$$Y_0 = y \in (0,1)$$

takes values in $[0,1]$ almost surely for all t.

Proof. The claim follows from Lemma 24 below for $K(\xi) = 1$. □

Lemma 17. *Let $\hat\beta$ and $q(y)$ be as in Theorem 3, define $Q(y) := \int^y q(z)dz$. For any admissible strategy there exists a probability $\hat P$, equivalent to P, such that the terminal wealth X_T satisfies*

$$E[X_T^{1-\gamma}]^{\frac{1}{1-\gamma}} \le X_0 e^{\hat\beta T + Q(y)} E_{\hat P}[e^{-(1-\gamma)Q(Y_T)}]^{\frac{1}{1-\gamma}} \tag{40}$$

and equality holds for the optimal strategy $\hat u$ in (9)*.*

Proof. The claim follows from Lemma 25 below for $K(\xi) = 1$. □

Proof of Theorem 3. Let $c(\lambda)$ and q be as in Lemma 13 and define $\hat\beta = \frac{\mu^2}{2\gamma\sigma^2} - c(\lambda)$. The function $\hat u(y) = \left|\frac{q(y)}{(\alpha+1)\lambda(1-yq(y))}\right|^{1/\alpha} \operatorname{sgn}(q(y))$ has the same sign as $q(y)$, which is positive in 0 and negative in 1. Hence, Lemma 16 implies that $Y_t \in [0,1]$ for all $t > 0$. Thus, as Q is bounded in $[0,1]$, it follows that $\lim_{T\to\infty} \frac{1}{T} \log E_{\hat P}[e^{-(1-\gamma)Q(Y_T)}]^{\frac{1}{1-\gamma}} = 0$. This relation, combined with Lemma 17, yields

$$\lim_{T\to\infty} \frac{1}{T} \log E[X_T^{1-\gamma}]^{\frac{1}{1-\gamma}} \le \hat\beta,$$

where equality holds for the strategy $\hat u$. Uniqueness follows from the optimality of $\hat\beta$. □

A.2. **Proof of Theorem 5.** Lemma 18 shows the monotonicity of the function $q(y)$ defined in Lemma 13. This property is used in the proof of Proposition 19, which is the crucial ingredient to make the heuristic scaling argument in Section 6 rigorous. Lemmas 20 and 22 establish the existence of a solution $s_\alpha(z)$ to the reduced HJB equation (15). Finally, in the proof of Theorem 5, we show that the appropriately rescaled solution $q(y)$ defined in Lemma 13 converges to $s_\alpha(z)$.

Lemma 18. *For λ sufficiently small, the solution $q(y)$ to* (37) *defined in Lemma 13 is strictly decreasing.*

Proof. Recall that the function $h(y)$ defined in Lemma 12(ii) is decreasing and such that $f(y, h(y)) = 0$. From Remark 15 it follows that $q(y) > h(y)$ on $(0, \bar Y) \cap \mathcal{D}_h$ and $q(y) < h(y)$ on $\mathcal{D}_h \cap (\bar Y, 1)$, whence $q(y)$ decreases on $\mathcal{D}_h$.

An explicit calculation shows that $\frac{\partial^2}{\partial y^2}[\sigma^2 y^2 (1-y)^2 f(y,q)] > 0$ on $[0,1] \times [-K\lambda^{\frac{1}{\alpha+1}}, K\lambda^{\frac{1}{\alpha+1}}]$ for λ sufficiently small. Assume by contradiction that $q(y)$ is not monotonic on $(0,1)$. Then there exist $y_1 < y_2 < y_3$ such that $w_1 = q(y_1) = q(y_2) = q(y_3) \in (h(1^-), h(0^+))$ and $f(y_1, w_1) = q'(y_1) \le 0$, $f(y_2, w_1) = q'(y_2) \ge 0$, and $f(y_3, w_1) = q'(y_3) \le 0$. Because either $f(0^+, w_1) > 0$ or $f(1^-, w_1) > 0$, this contradicts the convexity of $f(y, w_1)$ as a function of y. Hence, $q(y)$ is strictly decreasing on $(0,1)$. □

Proposition 19. $0 < \liminf_{\lambda\to0} \frac{c(\lambda)}{\lambda^{\frac{2}{\alpha+3}}} \le \limsup_{\lambda\to0} \frac{c(\lambda)}{\lambda^{\frac{2}{\alpha+3}}} < +\infty.$

Proof. Let λ_n be a positive null sequence such that $\frac{c(\lambda_n)}{\lambda_n^{\frac{2}{\alpha+3}}}$ has a limit. Let $y_{*,n} := \bar Y - \sqrt{\frac{2}{\gamma\sigma^2}(1+\epsilon)c(\lambda_n)}$ and $\tilde y_n \in \left(\bar Y - \sqrt{\frac{2}{\gamma\sigma^2}c(\lambda)}, \bar Y + \sqrt{\frac{2}{\gamma\sigma^2}c(\lambda)}\right)$ be the only point such that $q(\tilde y_n) = 0$.

By definition, the solution $q(\cdot; \tilde y_n, 0)$ to (37) has to coincide with the solution q with limit $h(0^+)$ in 0. Near $y = \bar Y$, the derivative of $q(\cdot)$ is determined by $c(\lambda)$. We show that if $c(\lambda)$ were to converge to 0 too fast, then the value of $q(\cdot; \tilde y_n, 0)$ in $y = y_{*,n}$ would be too low to coincide with the solution q with limit $h(0^+)$ in 0, leading to a contradiction. Conversely, if the convergence of $c(\lambda)$ to 0 were too slow, then the value of $q(\cdot;\tilde{y}_n,0)$ in $y=y_{*,n}$ would be too large. Formally, the proof is broken into two steps.

(i) Suppose that $\lim_{n\to\infty}\frac{c(\lambda_n)}{\lambda_n^{\frac{2}{\alpha+3}}}=0$ and find a contradiction.

Up to a subsequence, assume that $\tilde{y}_n\le\bar{Y}$ (if such a subsequence does not exist, consider the same argument at $y_n^*=\bar{Y}+\sqrt{\frac{2}{\gamma\sigma^2}(1+\epsilon)c(\lambda_n)}$).

In Step (i) we prove that if $\lim_{n\to\infty}c(\lambda_n)\lambda_n^{-\frac{2}{\alpha+3}}=0$, then $q(\cdot)$ is dominated near $y=\bar{Y}$ by a linear function with slope of order $-c(\lambda)$. A solution to (37) with initial condition $q(\tilde{y}_n)=0$ and such a growth rate near $y=\bar{Y}$ is increasing at $y=y_{*,n}<\tilde{y}_n$, contradicting the monotonicity of $q(\cdot)$ proven in Lemma 18.

First, we prove that for sufficiently small λ_n and a constant $K>0$ to be chosen, $k(y):=-Kc(\lambda_n)(y-\bar{Y})>q(y)$ on $(y_{*,n},\tilde{y}_n)$. It is clear that $k(\tilde{y}_n)>q(\tilde{y}_n)=0$, so it is enough to show that $f(y,k(y))>k'(y)$ on $(y_{*,n},\tilde{y}_n)$. Substituting $q=k(y)$ in $f(y,q)$ from (37), and because $|y-\bar{Y}|\le|y_{*,n}-\bar{Y}|=\sqrt{\frac{2}{\gamma\sigma^2}(1+\epsilon)c(\lambda_n)}$ on $(y_{*,n},\tilde{y}_n)$, it follows that for some constants $A_0>0,A_1,A_2,A_3$ independent of λ

$$
\begin{aligned}
f(y,k(y))&>A_0(0-c(\lambda)-A_1Kc(\lambda)^2-A_2K^2c(\lambda)^3-A_3K^{\frac{\alpha+1}{\alpha}}c(\lambda)^{\frac{3}{2}\frac{\alpha+1}{\alpha}}\lambda^{-\frac{1}{\alpha}})\\
&=-A_0c(\lambda)\left(1+A_1Kc(\lambda)+A_2K^2c(\lambda)^2+A_3K^{\frac{\alpha+1}{\alpha}}\left(\frac{c(\lambda)}{\lambda^{\frac{2}{\alpha+3}}}\right)^{\frac{\alpha+3}{2\alpha}}\right).
\end{aligned}
$$

Thus, $q(y)<k(y)$ for $K=A_0(1+\delta)$ and sufficiently small λ. In particular, $q(y_{*,n})<k(y_{*,n})=A_4c(\lambda)^{3/2}$, for some $A_4>0$ independent of λ. To conclude the first part of the proof it suffices to show that $\frac{\sigma^2}{2}(y_{*,n})^2(1-y_{*,n})^2f(y_{*,n},q(y_{*,n}))>0$, which would lead to a contradiction, as $q(y)$ is a decreasing function by Lemma 18.

Setting $y=y_{*,n}$ in (37) and recalling that $q(y_{*,n})<A_4c(\lambda)^{3/2}$, we find constants B_1,B_2,B_3 independent of λ such that

(41)
$$
\begin{aligned}
\frac{\sigma^2}{2}(y_{*,n})^2(1-y_{*,n})^2f(y_{*,n},q(y_{*,n}))&>\epsilon c(\lambda)+B_1c(\lambda)^2+B_2c(\lambda)^3+B_3c(\lambda)^{\frac{3}{2}\frac{\alpha+1}{\alpha}}\lambda^{-1/\alpha}\\
&=c(\lambda)\left(\epsilon+B_1c(\lambda)+B_2c(\lambda)^2+B_3\left(\frac{c(\lambda)}{\lambda^{\frac{2}{\alpha+3}}}\right)^{\frac{\alpha+3}{2\alpha}}\right),
\end{aligned}
$$

and for sufficiently small λ the last part is larger than $c(\lambda)\frac{\epsilon}{2}>0$.

(ii) Assume that $\lim_{n\to\infty}\frac{c(\lambda_n)}{\lambda_n^{\frac{2}{\alpha+3}}}=+\infty$ and find a contradiction.

Up to a subsequence, assume that $\tilde{y}_n\ge\bar{Y}$ (if such a subsequence does not exist, it is sufficient to consider the same argument at $y_n^*=\bar{Y}+\sqrt{\frac{2}{\gamma\sigma^2}(1+\epsilon)c(\lambda)}$). This part of the proof is divided in two.

(ii.a) There exists a subsequence n_k such that

$$
\lim_{k\to\infty}\frac{|q(y_{*,n_k})|^{\frac{\alpha+1}{\alpha}}\lambda_{n_k}^{-1/\alpha}}{c(\lambda_{n_k})}=+\infty. \tag{42}
$$

In Step $(ii.a)$ we prove that if $\lim_{n\to\infty}c(\lambda_n)\lambda_n^{-\frac{2}{\alpha+3}}=+\infty$, then the solution $q(\cdot)$ to (37) with initial condition $q(\tilde{y}_n)=0$ has a growth rate near $y=\bar{Y}$ such that $q(y_{*,n})$ is so large that the term $|q(y_{*,n})|^{\frac{\alpha+1}{\alpha}}\lambda^{-1/\alpha}$ in the HJB equation (37) dominates the term $c(\lambda)$.

If there is $C_1 > 0$ independent of λ such that eventually $q(y_{*,n})^2 \geq C_1 c(\lambda_n)^3$, then to conclude it is enough to note that

$$\frac{|q(y_{*,n})|^{\frac{\alpha+1}{\alpha}}\lambda_n^{-1/\alpha}}{c(\lambda_n)} \geq C_1^{\frac{\alpha+1}{2\alpha}}\left(\frac{c(\lambda_n)}{\lambda_n^{\frac{2}{\alpha+3}}}\right)^{\frac{\alpha+3}{2\alpha}}.$$

Thus, up to a subsequence, assume that $q(y_{*,n})^2 < C_1 c(\lambda_n)^3$. With this assumption, and also using the monotonicity of $q(y)$, on the interval $(\bar{Y} - \sqrt{\frac{2}{\gamma\sigma^2}(1-\epsilon)c(\lambda)}, \bar{Y})$ it follows that for some constant C_2

$$\frac{\sigma^2}{2}y^2(1-y)^2 q'(y) < (1-\epsilon)c(\lambda) - c(\lambda) + 0 + C_2 c(\lambda)^3 + 0.$$

Thus, there is $C_3 > 0$ such that $q'(y) < -C_3 c(\lambda)$ on $(\bar{Y} - \sqrt{\frac{2}{\gamma\sigma^2}(1-\epsilon)c(\lambda)}, \bar{Y})$ for sufficiently small λ. Thus,

$$\begin{aligned} q(y_{*,n}) &= -\int_{y_{*,n}}^{\tilde{y}_n} q'(y)dy \geq -\int_{\bar{Y}-\sqrt{\frac{2}{\gamma\sigma^2}(1-\epsilon)c(\lambda)}}^{\bar{Y}} q'(y)dy \\ &> \int_{\bar{Y}-\sqrt{\frac{2}{\gamma\sigma^2}(1-\epsilon)c(\lambda)}}^{\bar{Y}} C_3 c(\lambda)dy = C_3\sqrt{\frac{2}{\gamma\sigma^2}(1-\epsilon)}c(\lambda)^{3/2}. \end{aligned}$$

This leads to the inequality

$$\frac{|q(y^*)|^{\frac{\alpha+1}{\alpha}}\lambda^{-1/\alpha}}{c(\lambda)} \geq C_3^{\frac{\alpha+1}{\alpha}}\left(\frac{2}{\gamma\sigma^2}(1-\epsilon)\right)^{\frac{\alpha+1}{2\alpha}}\left(\frac{c(\lambda_n)}{\lambda_n^{\frac{2}{\alpha+3}}}\right)^{\frac{\alpha+3}{2\alpha}},$$

which proves the claim.

(ii.b) For some $\delta \in (0, \bar{Y})$ and sufficiently small λ

$$q(\delta) > \lambda^{\frac{1}{\alpha+1}}(\alpha+1)^{\frac{1}{\alpha+1}}\left(\frac{\alpha+1}{\alpha}\left(\frac{\mu^2}{2\gamma\sigma^2} - c(\lambda)\right)\right)^{\frac{\alpha}{\alpha+1}} = q(0).$$

This inequality contradicts Theorem 3, thereby concluding the proof.

In Step $(ii.b)$ we prove that if in the HJB equation (37) the term $|q(y)|^{\frac{\alpha+1}{\alpha}}\lambda^{-1/\alpha}$ dominates the term $c(\lambda)$ at $y = y_{*,n}$, then $q(\cdot)$ is larger than $h(0^+)$ near 0.

Fix a large constant $K > 0$. As $\lim_{n\to\infty}\frac{|q(y_{*,n})|^{\frac{\alpha+1}{\alpha}}\lambda_n^{-1/\alpha}}{c(\lambda_n)} = +\infty$, for sufficiently small λ,

$$|q(y_{*,n})|^{\frac{\alpha+1}{\alpha}}\lambda^{-1/\alpha} > K\epsilon c(\lambda_n) = K\left(\frac{\gamma\sigma^2}{2}(y_{*,n} - \bar{Y})^2 - c(\lambda_n)\right).$$

Define $y_{1,n} := \inf\{y \in (0, \bar{Y}) : -\frac{\sigma^2}{2}y^2(1-y)^2(1-\gamma)q^2(y) > \frac{\gamma\sigma^2}{2}(y-\bar{Y})^2 - c(\lambda_n)\} > 0$, with the usual convention $\inf\emptyset = +\infty$. Assume that eventually $y_{1,n} < y_{*,n}$. Then there exist $D_1, D_2, D_3 > 0$ such that for sufficiently small λ_n

$$q(y_{1,n}) > q(y_{*,n}) > D_1(c(\lambda_n)\lambda_n^{1/\alpha})^{\frac{\alpha}{\alpha+1}} > D_2\lambda_n^{\frac{3}{\alpha+3}},$$

where the third inequality follows from the assumption $\lim_{n\to\infty} \frac{c(\lambda_n)}{\lambda_n^{\frac{2}{\alpha+3}}} = +\infty$ and the second inequality from (42). In particular,

$$|q(y_{1,n})|^{\frac{\alpha+1}{\alpha}}\lambda_n^{-1/\alpha} = q(y_{1,n})^2|q(y_{1,n})|^{\frac{1-\alpha}{\alpha}}\lambda_n^{-1/\alpha} > D_3 q(y_{1,n})^2\lambda_n^{-\frac{4}{\alpha+3}}$$
$$> -K\frac{\sigma^2}{2}y_{1,n}^2(1-y_{1,n})^2(1-\gamma)q^2(y_{1,n}) = K\left(\frac{\gamma\sigma^2}{2}(y_{1,n}-\bar{Y})^2 - c(\lambda_n)\right).$$

If eventually $y_{1,n} < y_{*,n}$, let $y_{2,n}$ be equal to $y_{1,n}$. Otherwise, up to a subsequence, $y_{*,n} < y_{1,n}$ and then define $y_{2,n} := y_{*,n}$. In both cases,

$$|q(y_{2,n})|^{\frac{\alpha+1}{\alpha}}\lambda^{-1/\alpha} > K\left(\frac{\gamma\sigma^2}{2}(y_{2,n}-\bar{Y})^2 - c(\lambda_n)\right)$$

and $y_{2,n} \le y_{1,n}$. Choose δ such that $\delta < y_{2,n}$ for every n. Thus, on the interval $(\delta, y_{2,n})$,

$$\begin{aligned}(43)\quad & \frac{\sigma^2}{2}\frac{1}{16}q'(y) \le \frac{\sigma^2}{2}y^2(1-y)^2q'(y) < \\ & \frac{\gamma\sigma^2}{2}(y-\bar{Y})^2 - c(\lambda_n) + 0 - \frac{\sigma^2}{2}y^2(1-y)^2(1-\gamma)q^2 - \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}}|q|^{\frac{\alpha+1}{\alpha}}\lambda_n^{-1/\alpha} \\ & \le 2[\frac{\gamma\sigma^2}{2}(y-\bar{Y})^2 - c(\lambda_n)] - \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}}|q|^{\frac{\alpha+1}{\alpha}}\lambda_n^{-1/\alpha},\end{aligned}$$

where the first inequality follows from $y^2(1-y)^2 \le \frac{1}{16}$, the second from $y < \bar{Y}$ and $q > 0$, the third from $y_{2,n} \le y_{1,n}$ and the definition of $y_{1,n}$. Define $\tilde{q}(y)$ on $(\delta, y_{2,n})$ as the solution of the Cauchy problem

$$\begin{aligned}\tilde{q}'(y) &= \tilde{f}(y, \tilde{q}(y)) \\ &:= \frac{32}{\sigma^2}\left(2\left[\frac{\gamma\sigma^2}{2}(y-\bar{Y})^2 - c(\lambda)\right] - \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}}|\tilde{q}|^{\frac{\alpha+1}{\alpha}}\lambda^{-1/\alpha}\right), \\ \tilde{q}(y_{2,n}) &= q(y_{2,n}).\end{aligned}$$

The inequality (43) implies that $q'(y) < \tilde{f}(y, q(y))$. Thus, $\tilde{q}(y) < q(y)$ on $(\delta, y_{2,n})$. Define the function $k(y)$ by

$$|k(y)|^{\frac{\alpha+1}{\alpha}}\lambda^{-1/\alpha} = K\left[\frac{\gamma\sigma^2}{2}(y-\bar{Y})^2 - c(\lambda)\right].$$

To conclude the proof it remains to show that $k(y)$ is a subsolution for $\tilde{q}(y)$ on $(\delta, y_{2,n})$, which implies that $k(\delta) < q(\delta)$. In particular, if K is chosen sufficiently large, the contradiction follows.

Thus, it remains to show that $\tilde{f}(y, k(y)) < k'(y)$ on $(\delta, y_{2,n})$, i.e.

$$\frac{2}{\sigma^2\delta^2(1-\delta)^2}\left(2 - \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}}K\right)\left[\frac{\gamma\sigma^2}{2}(y-\bar{Y})^2 - c(\lambda)\right] <$$
$$\lambda^{\frac{1}{\alpha+1}}K^{\frac{\alpha}{\alpha+1}}\frac{\alpha}{\alpha+1}\left[\frac{\gamma\sigma^2}{2}(y-\bar{Y})^2 - c(\lambda)\right]^{-\frac{1}{\alpha+1}}\gamma\sigma^2(y-\bar{Y}).$$

This is equivalent to

$$\frac{2(\alpha+1)}{\alpha\gamma\sigma^4\delta^2(1-\delta)^2}\frac{1}{K^{\frac{\alpha}{\alpha+1}}}\left(\alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}}K - 2\right) > \lambda^{\frac{1}{\alpha+1}}(\bar{Y}-y)\left[\frac{\gamma\sigma^2}{2}(y-\bar{Y})^2 - c(\lambda)\right]^{-\frac{\alpha+2}{\alpha+1}}.$$

In other words, there are $E_1, E_2 > 0$ independent of λ and K such that

$$E_1 \frac{(E_2 K - 2)^{\frac{\alpha+1}{\alpha+2}}}{K^{\frac{\alpha}{\alpha+2}}} > \lambda^{\frac{1}{\alpha+2}} (\bar{Y} - y)^{\frac{\alpha+1}{\alpha+2}} \left[\frac{\gamma\sigma^2}{2}(y - \bar{Y})^2 - c(\lambda) \right]^{-1} =: r(y). \tag{44}$$

A simple calculation shows that $r'(y) > 0$ on $(0, y_{*,n})$. Thus, $r(y) \leq r(y_{*,n})$ on $(\delta, y_{2,n})$, where

$$r(y_{*,n}) = \lambda^{\frac{1}{\alpha+2}} \left(\frac{2}{\gamma\sigma^2}(1+\epsilon)c(\lambda) \right)^{\frac{\alpha+1}{2(\alpha+2)}} (\epsilon c(\lambda))^{-1} = E_3 \left(\frac{\lambda^{\frac{2}{\alpha+3}}}{c(\lambda)} \right)^{\frac{\alpha+3}{2(\alpha+2)}} \downarrow 0,$$

with $E_3 > 0$. If K is large enough, the left-hand side in (44) is well defined and strictly positive. As $r(y)$ converges to 0 on $(\delta, y_{2,n})$ uniformly in λ, for sufficiently small λ inequality (44) holds true. This concludes the proof. □

Lemma 20. *There exists a unique solution to*

$$y'(x) = f(x, y(x)) := -ax^2 + b + c|y(x)|^p, \tag{45}$$

with $a, b, c > 0$ and $p \geq 2$, such that

$$\lim_{x \to \infty} \frac{y(x)}{(\frac{a}{c}x^2)^{1/p}} = 1. \tag{46}$$

Proof. The function $g(x) := (\frac{a}{c}x^2 - \frac{b}{c})^{1/p}$, defined on $(\sqrt{b/a}, +\infty)$, is such that $f(x, g(x)) = 0$. For $\bar{x} \in (\sqrt{b/a}, +\infty)$ consider the solution $y(x; \bar{x}, g(\bar{x}))$ with initial condition $(\bar{x}, g(\bar{x}))$ and define $y_*(x) := \sup\{y(x; \bar{x}, g(\bar{x})) : \bar{x} \in (\sqrt{b/a}, +\infty)\}$. The solution y_* satisfies (46) by analogous arguments as in Lemma 7.6 in Liu et al. (2017).

To show uniqueness, define for any $d > 0$ the function $w(x; d) = y_*(x) + dx^{2/p}$, and observe that it is a subsolution to (45) for $x \geq \bar{x}$ and $\bar{x}$ large enough, i.e., $w'(x; d) \leq f(x, w(x; d))$. As $y_*(x)$ solves (45), this inequality is equivalent to

$$\frac{2d}{p} \frac{x^{\frac{2-p}{p}}}{y_*(x)} \leq c \left| 1 + \frac{dx^{2/p}}{y_*(x)} \right|^p - c.$$

As $x \to +\infty$, the left-hand side converges to 0 and the right-hand to $c\left(\left|1 + \frac{d}{a^{1/p}}\right|^p - 1\right) > 0$. Thus, if $\bar{x}$ is chosen sufficiently large, the inequality holds true and $w(x; d)$ is a subsolution for any $d > 0$. In particular, let $y_2(x) > y_*(x)$ be a solution to (45) and choose d_* such that $y_2(\bar{x}) = y_*(\bar{x}) + d_*\bar{x}^{2/p}$. Then $y_2(x) \geq w(x; d_*) = y_*(x) + d_* x^{2/p}$ for $x \geq \bar{x}$ and $y_2(x)$ cannot satisfy (46). As any solution smaller than $y_*(x)$ is also – for large x – smaller than $g(x)$ and thus eventually decreasing, thereby proving uniqueness. □

Remark 21. *The function $y(x)$ in Lemma 20 is such that*

$$y(x) = \inf\{y^*(x) : y^*(\cdot) \text{ solves (45)}, y^*(\cdot) \geq g(\cdot) \text{ on } ((b/a)^{1/2}, \infty)\}$$

for $x > (b/a)^{1/2}$, where $g(x) := (\frac{a}{c}x^2 - \frac{b}{c})^{1/p}$.

Proof. Define the two functions y_1, y_2 as $y_1(x) := \sup\{y(x; \bar{x}, g(\bar{x})) : \bar{x} \in (\sqrt{b/a}, +\infty)\}$ and $y_2(x) = \inf\{\tilde{y}(x) : \tilde{y}(\cdot) \text{ solves (45)}, \tilde{y}(\cdot) \geq g(\cdot) \text{ on } ((b/a)^{1/2}, \infty)\}$. Because $g(x)$ is increasing, $y(x; \bar{x}, g(\bar{x})) < g(x)$ on $(\bar{x}, \bar{x} + \varepsilon)$ for some ε. Assume that $y_1(x) > y_2(x)$. Then there exist $\bar{x}$ and $\tilde{y}(x)$ such that $y_1(x) > y(x; \bar{x}, g(\bar{x})) > \tilde{y}(x) > y_2(x)$, leading to a contradiction. Now, assume that $y_2(x) > y_1(x)$. It suffices to show that $y_1(\cdot) > g(\cdot)$ to obtain a contradiction from the minimality of $y_2(\cdot)$. Hence, assume by contradiction that there exists $\tilde{x}$ such that

$y_1(\tilde{x}) = g(\tilde{x})$. Then $y(\tilde{x}+\varepsilon; \tilde{x}+\varepsilon, g(\tilde{x}+\varepsilon)) = g(\tilde{x}+\varepsilon) > y_1(\tilde{x}+\varepsilon)$ for some ε, contradicting the maximality of $y_1(\cdot)$. $\square$

The heuristic argument that leads to the approximation (26) suggests that the normalized value function $s_\alpha(z)$ should grow as $(\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}}|z|^{\frac{2\alpha}{\alpha+1}}\operatorname{sgn}(-z)$ for $z \to \pm\infty$. Indeed

Lemma 22. *There exist a unique constant* $c_\alpha > 0$ *and a unique solution* $s_\alpha(z)$ *to the corresponding equation* (15) *such that*

$$\lim_{z\to-\infty} \frac{s_\alpha(z)}{|z|^{\frac{2\alpha}{\alpha+1}}} = (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} \qquad \textit{and} \qquad \lim_{z\to+\infty} \frac{s_\alpha(z)}{|z|^{\frac{2\alpha}{\alpha+1}}} = -(\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}}.$$

Proof. From Lemma 20, it follows that for any parameter c in equation (15) there are unique solutions $s^l(z,c)$ and $s^r(z,c)$ such that

$$\lim_{z\to-\infty} \frac{s^l(z,c)}{|z|^{\frac{2\alpha}{\alpha+1}}} = (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} \quad \text{and} \quad \lim_{z\to+\infty} \frac{s^r(z,c)}{|z|^{\frac{2\alpha}{\alpha+1}}} = -(\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}}.$$

Note that if $s^*(z)$ solves (15), also the function $-s^*(-z)$ is a solution. In particular, from the uniqueness result in Lemma 20, $s^r(z,c) = -s^l(-z,c)$. For $c \downarrow 0$, eventually $s^l(0,c) > s^r(0,c)$, and thus $s^l(z,c) > s^r(z,c)$. In addition, for any z, $\lim_{c\to+\infty} s^l(z,c) = -\infty$ and $\lim_{c\to+\infty} s^r(z,c) = +\infty$. Thus, for $c \uparrow +\infty$ eventually $s^l_\alpha(z,c) < s^r_\alpha(z,c)$. It follows that $s^l_\alpha(z,c_\alpha) = s^r_\alpha(z,c_\alpha)$ for some constant c_α . Because $\frac{\partial s^l}{\partial c}(z,c) < 0$, such a constant is unique. From $s^l_\alpha(z,c_\alpha) = s^r_\alpha(z,c_\alpha) = -s^l_\alpha(-z,c_\alpha)$, it follows that $s^l_\alpha(0,c_\alpha) = 0$. $\square$

Proof of Theorem 5. The substitutions $y = \bar{Y} + \lambda^{\frac{1}{\alpha+3}} z$ and $r_\lambda(z) = q_\lambda(y)\lambda^{-\frac{3}{\alpha+3}}$ in equation (11) yield

$$\begin{aligned}(47)\qquad & -\frac{\gamma\sigma^2}{2} z^2 \lambda^{\frac{2}{\alpha+3}} + c(\lambda) + \frac{\sigma^2}{2} y^2(1-y)^2 (r'_\lambda \lambda^{\frac{2}{\alpha+3}} + (1-\gamma) r_\lambda^2 \lambda^{\frac{6}{\alpha+3}}) + \\ & - \gamma\sigma^2 y(1-y) z \lambda^{\frac{4}{\alpha+3}} r_\lambda + \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}} \frac{|r_\lambda|^{\frac{\alpha+1}{\alpha}}}{(1 - y r_\lambda \lambda^{\frac{3}{\alpha+3}})^{1/\alpha}} \lambda^{\frac{2}{\alpha+3}} = 0.\end{aligned}$$

Divide the equation by $\lambda^{\frac{2}{\alpha+3}}$ and consider a sequence λ_n such that $c(\lambda_n)\lambda_n^{-\frac{2}{\alpha+3}}$ has limit $\bar{c}$. By taking the limit in equation (47), it follows that $r_0(z) := \lim_{\lambda_n\to 0} r_{\lambda_n}(z)$ satisfies

$$-\frac{\gamma\sigma^2}{2} z^2 + \bar{c} + \frac{\sigma^2}{2} \bar{Y}^2 (1-\bar{Y})^2 r'_0 + \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}} |r_0|^{\frac{\alpha+1}{\alpha}} = 0.$$

The additional substitution $s_\alpha(w) = \frac{r_0(A_\alpha w)}{B_\alpha}$ yields

$$(48)\qquad -w^2 + c_* + s' + \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}} |s|^{\frac{\alpha+1}{\alpha}} = 0,$$

where $c_* = \bar{c}\left[\left(\frac{\sigma^2}{2}\right)^3 \gamma \bar{Y}^4 (1-\bar{Y})^4\right]^{-\frac{\alpha+1}{\alpha+3}}$.

Let $h(y)$ be defined as in Lemma 12 and let $\mathcal{D}_h$ be its domain. On $(-\infty, -c_*^{1/2}) \cup (c_*^{1/2}, +\infty)$ the function $\lambda_n^{-\frac{3}{\alpha+3}} B_\alpha^{-1} h(\bar{Y} + \lambda_n^{\frac{1}{\alpha+3}} A_\alpha w)$ converges to $g(w) := (w^2 - c_*)^{\frac{\alpha}{\alpha+1}} (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} \operatorname{sgn}(-w)$ for $\lambda_n \downarrow 0$. In view of Remark 15, the solution

$q(y)$ to (11) satisfies (38)-(39), and thus $s_\alpha(w)$ satisfies

$$s_\alpha(w) = \inf\{\tilde{s}(w) : \tilde{s}(\cdot) \text{ solves (48)}, \tilde{s}(\cdot) \geq g(\cdot) \text{ on } (-\infty, -c_*^{1/2})\} \quad w < -c_*^{1/2}$$

$$s_\alpha(w) = \sup\{\tilde{s}(w) : \tilde{s}(\cdot) \text{ solves (48)}, \tilde{s}(\cdot) \leq g(\cdot) \text{ on } (c_*^{1/2}, \infty)\} \qquad w > c_*^{1/2}.$$

By Remark 21, these conditions are equivalent to

$$\lim_{w\to-\infty} \frac{s_\alpha(w)}{|w|^{\frac{2\alpha}{\alpha+1}}} = (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} \qquad \text{and} \qquad \lim_{w\to+\infty} \frac{s_\alpha(w)}{|w|^{\frac{2\alpha}{\alpha+1}}} = -(\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}}.$$

Finally, Lemma 22 proves that there exists a unique constant $c_* = c_\alpha$ and a unique solution $s_\alpha(w)$ such that both these conditions are satisfied. In particular, this means that the constant c_* does not depend on the sequence λ_n chosen initially, that $r_\lambda(z)$ converges to $B_\alpha s_\alpha(z/A_\alpha)$ and that $\lim_{\lambda\to 0} c(\lambda)\lambda^{-\frac{2}{\alpha+3}} = c_\alpha \left[\left(\frac{\sigma^2}{2}\right)^3 \gamma \bar{Y}^4(1-\bar{Y})^4\right]^{\frac{\alpha+1}{\alpha+3}}$, proving part (i) and (iii) of the Theorem.

The asymptotics for $\hat{u}$ in part (ii) of the Theorem follow directly from the relation between $\hat{u}(y)$ and $q(y)$ found in Theorem 3. The drift and diffusion coefficients of Z_s^λ converge pointwise to $v_\alpha(z)$ and $\bar{Y}(1-\bar{Y})\sigma$, hence part (iv) follows from Theorem 11.1.4 in Stroock and Varadhan (1979). □

A.3. **Calculations for $\alpha = 1$ and $\alpha \downarrow 0$.**

Proof of Lemma 8. It suffices to check that $s_1(z) = -2z$ and $c_1 = 2$ solve equation (15) with the growth conditions in Lemma 22. □

Lemma 23. *Let $g_\alpha(z,s) = z^2 - c_\alpha - \alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}}|s|^{\frac{\alpha+1}{\alpha}}$ and define $h_\alpha(z) := \left(z^2 - c_\alpha\right)^{\frac{\alpha}{\alpha+1}}(\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}}\operatorname{sgn}(-z)$ (i.e. $g_\alpha(z, h_\alpha(z)) = 0$). Then*

$$\lim_{\alpha\to 0} g_\alpha(z,s) = \begin{cases} z^2 - c_0, & |s| \leq 1, \\ -\infty, & |s| > 1, \end{cases} \quad \text{and} \quad \lim_{\alpha\to 0} h_\alpha(z) = \begin{cases} 1, & |z| \leq -c_0^{1/2}, \\ -1, & |z| \geq c_0^{1/2}. \end{cases}$$

Proof. The limits follow directly from the expressions of g_α and h_α. □

Proof of Lemma 9. For all α, $s_\alpha(0) = 0$. Thus, for $\alpha \downarrow 0$, $s_\alpha(z)$ converges to $z^3/3 - cz$, for some $c > 0$ and for z such that $z^3/3 - cz \leq 1$. As $\lim_{\alpha\to 0} h_\alpha(z) = 1$ for $z \leq -c^{1/2}$, and because $s_\alpha(z) = \inf\{s^*(z) : s^*(\cdot) \text{ solves (15)}, s^*(\cdot) \geq h_\alpha(\cdot) \text{ on } (-\infty, -c^{1/2})\}$, the limit has to satisfy $\lim_\alpha s_\alpha(-c_\alpha^{1/2}) = 1$. From the relation $-c_0^{3/2}/3 - c_0(-c_0^{1/2}) = 1$ it follows that $c_0 = \left(\frac{3}{2}\right)^{2/3}$. □

Proof of Lemma 10. Recall that $v_\alpha(z) = -\left|\frac{B_\alpha s_\alpha(z/A_\alpha)}{\alpha+1}\right|^{1/\alpha}\operatorname{sgn}(z)$. As $\lim_{\alpha\to 0}(\alpha+1)/B_\alpha = 1$, $\lim_{\alpha\to 0} v_\alpha(z) = 0$ if $|s_0(z/A_0)| < 1$. For each α and $z \leq -A_\alpha\sqrt{c_\alpha}$, note that

$$v_\alpha(z) = \left|\frac{B_\alpha s_\alpha(z/A_\alpha)}{\alpha+1}\right|^{1/\alpha} \geq \left|\frac{B_\alpha h_\alpha(z/A_\alpha)}{\alpha+1}\right|^{1/\alpha} = \frac{B_\alpha^{1/\alpha}}{\alpha^{\frac{1}{\alpha+1}}}|z^2/A_\alpha^2 - c_\alpha|^{\frac{1}{\alpha+1}} \uparrow +\infty.$$

Similarly, if $z \geq A_\alpha\sqrt{c_\alpha}$, then $\lim_{\alpha\to 0} v_\alpha(z) = -\infty$. On the other hand $\lim_{\alpha\to 0} v_\alpha(z) = +\infty$ if $|s_0(z/A_0)| > 1$. A short calculation shows that $\{|s_0(z/A_0)| < 1\} = \{|z| < A_\alpha\sqrt{c_\alpha}\}$ is equivalent to

$$z \in \left(-\left(\frac{3}{2\gamma}\bar{Y}^2(1-\bar{Y})^2\right)^{1/3}, \left(\frac{3}{2\gamma}\bar{Y}^2(1-\bar{Y})^2\right)^{1/3}\right).$$

□

A.4. **Proof of Theorem 11.** The proofs proceed through several lemmas. Lemma 24 shows that under an appropriate trading strategy the process Y_t lives in the interval $[0,1]$. Lemma 25 provides the crucial tight bound for the investor's utility. Theorem 11 uses this bound, together with the boundedness of Y_t, to prove long-horizon optimality.

Lemma 24. *Let $u(y,\xi)$ be a bounded continuous function on $[0,1]$ such that $u(0,\xi) > 0$ and $u(1,\xi) < 0$ for all $\xi \in E$. Under the assumptions of Theorem 11, for λ small enough the process with dynamics*

$$\begin{aligned} dY_t &= Y_t(1-Y_t)(\mu - Y_t\sigma^2)dt + (u(Y_t,\xi_t) + \lambda K(\xi_t)Y_t|u(Y_t,\xi_t)|^{\alpha+1})dt \\ &\quad + Y_t(1-Y_t)\sigma dW_t \\ Y_0 &= y \in (0,1) \end{aligned}$$

takes values in $[0,1]$ almost surely for all t.

Proof. The claim follows from results on the stochastic invariance of diffusions (Filipovic and Mayerhofer, 2009, Lemma B.1), which require in the present model that the drift of Y_t is positive at $y=0$ and negative at $y=1$. In both cases, the drift reduces to $u(Y_t,\xi_t) + \lambda K(\xi_t)Y_t|u(Y_t,\xi_t)|^{\alpha+1}$, which has the same sign of $u(Y_t,\xi_t)$ for λ small enough, as $0 < m_K < K(\xi_t) < M_K$. □

Lemma 25. *Let $\hat\beta$ and $Q(y,\xi)$ be as in Theorem 11. For any admissible strategy there exists a probability $\hat P$, equivalent to P, such that its terminal wealth X_T satisfies*

$$E[X_T^{1-\gamma}]^{\frac{1}{1-\gamma}} \le X_0 e^{\hat\beta T + Q(y,\xi)} E_{\hat P}[e^{-(1-\gamma)Q(Y_T,\xi_T)}]^{\frac{1}{1-\gamma}} \tag{49}$$

and equality holds for the optimal strategy $\hat u$ in (30)*.*

Proof. Define

$$\begin{aligned} \log\frac{d\hat P}{dP} &:= (1-\gamma)\int_0^T Y_s\sigma(1+Q_y(1-Y_s))dW_s + (1-\gamma)\int_0^T Q_\xi\sigma_\xi dB_s \\ -\frac{(1-\gamma)^2}{2}&\int_0^T (\sigma^2Y_s^2(1+Q_y(1-Y_s))^2 + Q_\xi^2\sigma_\xi^2 + 2\rho\sigma\sigma_\xi Y_s(1+Q_y(1-Y_s))Q_\xi)ds. \end{aligned}$$

Because u_t is admissible, $Y_t \in [0,1]$ for all t (see Footnote 12). The process $\frac{d\hat P}{dP}$ satisfies Novikov's condition because $Y_t \in [0,1]$ and in view of the boundedness assumptions on $Q(y,\xi)$ and its derivatives. Hence, $\hat P$ is an equivalent probability measure. It suffices to check that

$$X_T^{1-\gamma} \le X_0^{1-\gamma} e^{(1-\gamma)(\hat\beta T + Q(y,\xi))} e^{-(1-\gamma)Q(Y_T,\xi_T)} \frac{d\hat P}{dP}$$

for $\gamma < 1$, and the reverse inequality for $\gamma > 1$. Thus, both cases follow from the inequality

$$\log X_T - \log X_0 - \frac{1}{1-\gamma}\log\frac{d\hat P}{dP} \le \hat\beta T - Q(Y_T,\xi_T) + Q(y,\xi)\ . \tag{50}$$

Recall now the self-financing condition

$$\begin{aligned} \frac{dX_t}{X_t} &= Y_t(\mu dt + \sigma dW_t) - \lambda K(\xi_t)|u_t|^{\alpha+1}dt, \\ dY_t &= (Y_t(1-Y_t)(\mu - Y_t\sigma^2) + u_t + \lambda K(\xi_t)Y_t|u_t|^{\alpha+1})dt + Y_t(1-Y_t)\sigma dW_t. \end{aligned}$$

Itô's formula implies that

$$(51)\quad \int_0^T (Q_y(Y_t,\xi_t)(Y_t(1-Y_t)(\mu - Y_t\sigma^2) + u_t + \lambda K(\xi_t)Y_t|u_t|^{\alpha+1}) + \frac{1}{2}Q_{yy}(Y_t,\xi_t)Y_t^2(1-Y_t)^2\sigma^2)dt + \int_0^T Q_y(Y_t,\xi_t)Y_t(1-Y_t)\sigma dW_t + \int_0^T (Q_\xi(Y_t,\xi_t)\mu_\xi(\xi_t) + \frac{1}{2}Q_{\xi\xi}(Y_t,\xi_t)\sigma_\xi^2(\xi_t))dt + \int_0^T Q_\xi(Y_t,\xi_t)\sigma_\xi(\xi_t)dB_t + \int_0^T Q_{y\xi}(Y_t,\xi_t)Y_t(1-Y_t)\sigma\rho\sigma_\xi(\xi_t)dt = Q(Y_T,\xi_T) - Q(y,\xi)$$

and

$$(52)\quad \log X_T - \log X_0 = \int_0^T \left(Y_t\mu - \lambda K(\xi_t)|u_t|^{\alpha+1} - \frac{Y_t^2\sigma^2}{2}\right)dt + \int_0^T Y_t\sigma dW_t.$$

Replacing (51) and (52) in (50) yields

$$\int_0^T \left(Y_t\mu - \lambda K(\xi_t)|u_t|^{\alpha+1} - \frac{Y_t^2\sigma^2}{2}\right)dt - \int_0^T Y_t\sigma Q_y(1-Y_t)dW_t - \int_0^T Q_\xi\sigma_\xi dB_t + \frac{1-\gamma}{2}\int_0^T (\sigma^2Y_t^2(1+Q_y(1-Y_t))^2 + Q_\xi^2\sigma_\xi^2 + 2\rho\sigma\sigma_\xi Y_t(1+Q_y(1-Y_t))Q_\xi)dt$$
$$\leq \int_0^T (\hat\beta - Q_y(Y_t(1-Y_t)(\mu - Y_t\sigma^2) + u_t + \lambda K(\xi_t)Y_t|u_t|^{\alpha+1}) - \frac{1}{2}Q_{yy}Y_t^2(1-Y_t)^2\sigma^2 - Q_\xi\mu_\xi(\xi_t) - \frac{1}{2}Q_{\xi\xi}\sigma_\xi^2(\xi_t) - Q_{y\xi}Y_t(1-Y_t)\sigma\rho\sigma_\xi(\xi_t))dt - \int_0^T Q_yY_t(1-Y_t)\sigma dW_t - \int_0^T Q_\xi\sigma_\xi dB_t.$$

As the stochastic integrals on both sides are equal, it remains to prove that, for all $u \in \mathbb{R}$,

$$y\mu - \lambda K(\xi)|u|^{\alpha+1} - \frac{y^2\sigma^2}{2} + \frac{1-\gamma}{2}(\sigma^2y^2(1+Q_y(1-y))^2 + Q_\xi^2\sigma_\xi^2 + 2\rho\sigma\sigma_\xi y(1+Q_y(1-y))Q_\xi) \leq \hat\beta - Q_y(y(1-y)(\mu - y\sigma^2) + u + \lambda K(\xi)y|u|^{\alpha+1}) - \frac{\sigma^2}{2}Q_{yy}y^2(1-y)^2 - Q_\xi\mu_\xi - \frac{\sigma_\xi^2}{2}Q_{\xi\xi} - Q_{y\xi}y(1-y)\sigma\rho\sigma_\xi.$$

Rearranging the terms, this inequality is equivalent to

$$\frac{\sigma_\xi^2}{2}((1-\gamma)Q_\xi^2 + Q_{\xi\xi}) + \sigma\rho\sigma_\xi y(1-y)(Q_{y\xi} + (1-\gamma)Q_yQ_\xi) + \sigma\rho\sigma_\xi(1-\gamma)yQ_\xi + \frac{\sigma^2}{2}y^2(1-y)^2((1-\gamma)Q_y^2 + Q_{yy}) - \hat\beta + \mu y - \frac{\gamma}{2}y^2\sigma^2 + \mu_\xi Q_\xi + y(1-y)(\mu - \gamma\sigma^2 y)Q_y + \lambda K(\xi)|u|^{\alpha+1}(yQ_y - 1) + Q_yu \leq 0.$$

Under the condition $yQ_y < 1$ assumed in the Theorem, the maximum of the terms in the last line is $\alpha(\alpha+1)^{-\frac{\alpha+1}{\alpha}}\frac{|Q_y|^{\frac{\alpha+1}{\alpha}}}{(1-yQ_y)^{1/\alpha}(K(\xi)\lambda)^{1/\alpha}}$, and therefore the inequality follows from the HJB equation (29). Furthermore, the inequality becomes an equality for the optimal control $\hat u$, obtained by maximizing the above terms. □

Proof of Theorem 11. Let $\hat\beta$ and Q be the solution to (29). Because the function $\hat u(y,\xi) = \left|\frac{Q_y(y,\xi)}{(\alpha+1)\lambda K(\xi)(1-yQ_y(y,\xi))}\right|^{1/\alpha}\mathrm{sgn}(Q_y(y,\xi))$ has, by definition, the same sign as $Q_y(y,\xi)$, which is positive in 0 and negative in 1 by assumption, Lemma 24 implies that $Y_t \in [0,1]$ for all $t > 0$. Thus, as Q is bounded in $[0,1] \times E$, it follows that $\lim_{T\to\infty} \frac{1}{T} \log E_{\hat{P}}[e^{-(1-\gamma)Q(Y_T,\xi_T)}]^{\frac{1}{1-\gamma}} = 0$. This relation, combined with Lemma 25, yields

$$\lim_{T\to\infty} \frac{1}{T} \log E[X_T^{1-\gamma}]^{\frac{1}{1-\gamma}} \leq \hat{\beta},$$

where equality holds for the strategy $\hat{u}$. □

A.5. **Proofs of Remaining Results.**

Proof of Proposition 4. As S_t is constant for $t \geq T$, wealth and risky weight follow the dynamics

$$\frac{dX_t}{X_t} = -\lambda |u|^{\alpha+1} dt, \qquad dY_t = (u + \lambda Y_t |u|^{\alpha+1}) dt$$

for $t \geq T$. In other words,

$$X_{T+t} = X_T e^{-\lambda|u|^{\alpha+1}t}, \qquad Y_{T+t} = \frac{1 - e^{\lambda|u|^{\alpha+1}t}}{\lambda|u|^{\alpha}} + Y_T e^{\lambda|u|^{\alpha+1}t}.$$

It follows that $Y_{T+L(u)} = 0$ and $\frac{X_T - X_{T+L(u)}}{X_T} = \lambda Y_T |u|^{\alpha}$. □

Lemma 26. *For λ^* sufficiently small and any $p \geq 1$, the absolute moment $E|Z_s^\lambda|^p$ is uniformly bounded in s and $\lambda \leq \lambda^*$.*

Proof. The processes Z^λ have drift of the form $\lambda^{\frac{1}{\alpha+3}} u_\lambda(\bar{Y} + \lambda^{\frac{1}{\alpha+3}} z) + g_1(z)$, with $g_1(z)$ bounded, and $r_\lambda(z) := \lambda^{-\frac{3}{\alpha+3}} q_\lambda(\bar{Y} + \lambda^{\frac{1}{\alpha+3}} z) = C|z|^{\frac{2\alpha}{\alpha+1}} \operatorname{sgn}(-z) + g_2(z)$, with $C > 0$ and $g_2(z)$ bounded. Thus, from the definition of $u_\lambda(\cdot)$, the dynamics of the process Z_s^λ is rewritten as

$$dZ_s^\lambda = (a_\lambda(Z_s^\lambda) + b_\lambda |Z_s^\lambda|^{\frac{2}{\alpha+1}} \operatorname{sgn}(-Z_s^\lambda)) ds + \sigma_\lambda(Z_s^\lambda) dW_s,$$

where $a_\lambda(\cdot)$, $b_\lambda > 0$ and $\sigma_\lambda(\cdot)$ are uniformly bounded in λ. Consider now the family of processes U^λ with dynamics $dU_s^\lambda = (a_\lambda^*(U_s^\lambda) - b_\lambda U_s) ds + \sigma_\lambda(U_s^\lambda) dW_s$, where $a_\lambda^*(u) := a_\lambda(u) + b_\lambda(|u|^{\frac{2}{\alpha+1}} \operatorname{sgn}(-u)) + u) 1_{|u|\leq 1}$. Applying the comparison principle to the squares of Z^λ and U^λ, it follows that $E|Z_s^\lambda|^p \leq E|U_s^\lambda|^p$. To prove that $E|U_s^\lambda|^p$ is uniformly bounded in s and λ for any $p \geq 1$, note that the process U_t^λ solves $U_t^\lambda e^{b_\lambda t} = U_0 + \int_0^t e^{b_\lambda s} a_\lambda^*(U_s^\lambda) ds + \int_0^t e^{b_\lambda t} \sigma_\lambda(U_s^\lambda) dW_s$. Hence,

$$\begin{aligned}
E|U_t^\lambda|^p e^{pb_\lambda t} &\leq 3^p |U_0|^p + 3^p E\left|\int_0^t e^{b_\lambda s} a_\lambda^*(U_s^\lambda) ds\right|^p + 3^p E\left|\int_0^t e^{b_\lambda s} \sigma_\lambda(U_s^\lambda) dW_s\right|^p \\
&\leq C_1 + 3^p E \int_0^t e^{pb_\lambda s} \left|a_\lambda^*(U_s^\lambda)\right|^p ds + 3^p E \left|\int_0^t e^{2b_\lambda s} \left|\sigma_\lambda(U_s^\lambda)\right|^2 ds\right|^{p/2} \\
&\leq C_1 + C_2(e^{pb_\lambda t} - 1) + C_3(e^{2b_\lambda t} - 1)^{p/2},
\end{aligned}$$

where the second inequality follows from Jensen's inequality and the Burkholder-Davis-Gundy inequality. Thus, $E|U_t^\lambda|^p \leq K$ uniformly in s and λ. □

Lemma 27. *Let μ^λ be the invariant measure for the process Z_s^λ and μ_α the invariant measure for the process Z_s^0. Then the family $(\mu_\lambda)_{\lambda>0}$ converges weakly to μ_α as $\lambda \downarrow 0$.*

Proof. Because Lemma 26 yields that $E|Z_s^\lambda|^2 \leq K$ uniformly in s and λ, the family of transition probabilities $(\pi_s^\lambda)_{s\geq 0,\lambda>0}$ associated to the processes Z^λ is tight. In particular, the family of invariant measures $(\mu^\lambda)_{\lambda>0}$ is tight. Tightness implies that there is a sequence of measures μ^{λ_n} (henceforth μ^n) that converges to a measure μ_α. It remains to show that μ_α is the invariant measure for Z^0, i.e. that for any bounded $\phi(z)$, $\int_{\mathbb{R}} P_t\phi(x)\mu_\alpha(dx) = \int_{\mathbb{R}} \phi(x)\mu_\alpha(dx)$, where P_t is the transition operator associated to Z^0. Note that

$$\begin{aligned}\left|\int_{\mathbb{R}} (\phi - P_t\phi)\, \mu_\alpha(dx)\right| \leq \left|\int_{\mathbb{R}} \phi d\mu_\alpha - \int_{A^n} \phi d\mu^n\right| + \left|\int_{A^n} (\phi - P_t^n\phi)\, d\mu^n\right| \\ + \left|\int_{A^n} (P_t^n\phi - P_t\phi)\, d\mu^n\right| + \left|\int_{A^n} P_t\phi d\mu^n - \int_{\mathbb{R}} P_t\phi d\mu_\alpha\right|,\end{aligned}$$

where $A^n := [-\bar{Y}\lambda_n^{-\frac{1}{\alpha+3}}, (1-\bar{Y})\lambda_n^{-\frac{1}{\alpha+3}}]$. The first and the last terms converge to 0 because μ^n converges weakly to μ_α. The second term is 0, because μ^n is invariant for P_t^n. It remains to show that the third term converges to 0. The family μ^n is tight, so for each ε there is a set A_ε such that $\mu^n(A^n \setminus A_\varepsilon) < \varepsilon$ for all n.

$$\begin{aligned}\left|\int_{A^n} (P_t^n\phi - P_t\phi)\, d\mu^n\right| &\leq \int_{A_\varepsilon} |P_t^n\phi - P_t\phi|\, d\mu^n + \int_{A^n\setminus A_\varepsilon} |P_t^n\phi - P_t\phi|\, d\mu^n \\ &\leq \sup_{x\in A_\varepsilon} |P_t^n\phi - P_t\phi| + 2||\phi||_\infty \varepsilon.\end{aligned}$$

As $\sup_{x\in A_\varepsilon} |P_t^n\phi - P_t\phi|$ converges to 0, the thesis follows. □

Proof of Proposition 6. Recall that the stationary density of Z_s^0 equals $G_\alpha \exp\{\int_0^x \frac{2b(s)}{s^2(s)} ds\}/s^2(x)$, where $b(z) = v_\alpha(z)$, $s(z) = \bar{Y}(1-\bar{Y})\sigma$ and G_α is a normalizing constant to be found. For the following calculations, observe that $\frac{A_\alpha B_\alpha^{1/\alpha}}{\bar{Y}^2(1-\bar{Y})^2\sigma^2} = 1$. That $G_\alpha = M_\alpha \frac{\bar{Y}^2(1-\bar{Y})^2\sigma^2}{A_\alpha}$, where M_α only depends on α, follows from

$$\begin{aligned}1 = \int_{\mathbb{R}} \mu_\alpha(dx) &= \frac{K}{\bar{Y}^2(1-\bar{Y})^2\sigma^2} \int_{\mathbb{R}} \exp\left\{\int_0^x \frac{2v_\alpha(s)}{\bar{Y}^2(1-\bar{Y})^2\sigma^2} ds\right\} dx \\ &= \frac{2K}{\bar{Y}^2(1-\bar{Y})^2\sigma^2} \int_0^\infty \exp\left\{-\frac{2A_\alpha B_\alpha^{1/\alpha}}{\bar{Y}^2(1-\bar{Y})^2\sigma^2(\alpha+1)^{1/\alpha}} \int_0^{x/A_\alpha} |s_\alpha(s)|^{1/\alpha} ds\right\} dx \\ &= \frac{2A_\alpha K}{\bar{Y}^2(1-\bar{Y})^2\sigma^2} \int_0^\infty \exp\left\{-\frac{2}{(\alpha+1)^{1/\alpha}} \int_0^x |s_\alpha(s)|^{1/\alpha} ds\right\} dx.\end{aligned}$$

In particular, $M_\alpha = \frac{1}{2}\left[\int_0^\infty \exp\left\{-\frac{2}{(\alpha+1)^{1/\alpha}}\int_0^x |s_\alpha(s)|^{1/\alpha}ds\right\}dx\right]^{-1}$. From Lemma 27, μ^λ converges weakly to μ_α, therefore

$$
\begin{aligned}
&\lim_{\lambda\to 0}\lim_{T\to\infty}\frac{\lambda^{\frac{1}{\alpha+3}}}{T}\int_0^T |\hat{u}(Y_t^\lambda)|dt\\
&= \lim_{\lambda\to 0}\lim_{T\to\infty}\frac{\lambda^{\frac{2}{\alpha+3}}}{T}\int_0^{T/\lambda^{\frac{2}{\alpha+3}}} \lambda^{\frac{1}{\alpha+3}}\left|\hat{u}\left(\bar{Y}+\lambda^{\frac{1}{\alpha+3}}Z^\lambda_{\lambda^{\frac{2}{\alpha+3}}s}\right)\right|ds\\
&= \lim_{\lambda\to 0}\int_{-\bar{Y}\lambda^{-1/(\alpha+3)}}^{(1-\bar{Y})\lambda^{-1/(\alpha+3)}} \lambda^{\frac{1}{\alpha+3}}|\hat{u}(\bar{Y}+\lambda^{\frac{1}{\alpha+3}}z)|\mu^\lambda(dz) = \int_{\mathbb{R}}|v_\alpha(z)|\mu_\alpha(dz)\\
&= \int_0^\infty -2\frac{M_\alpha}{A_\alpha}v_\alpha(z)\exp\left\{-\frac{2}{(\alpha+1)^{1/\alpha}}\int_0^{z/A_\alpha}|s_\alpha(s)|^{1/\alpha}ds\right\}dz\\
&= \int_0^\infty 2\frac{M_\alpha}{A_\alpha}\left|\frac{B_\alpha s_\alpha(z/A_\alpha)}{(\alpha+1)}\right|^{1/\alpha}\exp\left\{-\frac{2}{(\alpha+1)^{1/\alpha}}\int_0^{z/A_\alpha}|s_\alpha(s)|^{1/\alpha}ds\right\}dz\\
&= B_\alpha^{1/\alpha}\int_0^\infty 2M_\alpha\left|\frac{s_\alpha(x)}{\alpha+1}\right|^{1/\alpha}\exp\left\{-\frac{2}{(\alpha+1)^{1/\alpha}}\int_0^x|s_\alpha(s)|^{1/\alpha}ds\right\}dx\\
&= B_\alpha^{1/\alpha}\left[M_\alpha\left(1-\exp\left\{-\frac{2}{(\alpha+1)^{1/\alpha}}\int_0^\infty|s_\alpha(s)|^{1/\alpha}ds\right\}\right)\right].
\end{aligned}
$$

Finally, $B_\alpha^{1/\alpha} = l_\alpha^{\frac{1}{\alpha+1}} = \left(\left(\frac{\sigma^2}{2}\right)^3\gamma\bar{Y}^4(1-\bar{Y})^4\right)^{\frac{1}{\alpha+3}}$ and the term in square brackets only depends on α . □

Lemma 28. *The finite-dimensional distributions of the family of processes $(\lambda^{\frac{1}{\alpha+3}}u_\lambda(Z^\lambda))_{\lambda\le\lambda^*}$ converge weakly to the finite-dimensional distributions of $v_\alpha(Z^0)$, where $v_\alpha(\cdot)$ is defined in Theorem 5(ii) .*

Proof. The proof is identical to the proof of Lemma B.6 in Guasoni and Weber (2017). □

Proof of Proposition 7. Because $\lambda^{\frac{1}{\alpha+3}}u_\lambda(\bar{Y}+\lambda^{\frac{1}{\alpha+3}}z)\sim C|z|^{\frac{2}{\alpha+1}}\operatorname{sgn}(-z)$ for large z, Lemma 26 implies that $\lambda^{\frac{2}{\alpha+3}}E|u_\lambda(Z_s^\lambda)|^2$ is bounded uniformly in λ and s. Hence, the family $(\lambda^{\frac{1}{\alpha+3}}u_\lambda(Z^\lambda))_{\lambda>0}$ is tight. It follows now from Lemma 28 that for $\lambda\downarrow 0$ the processes $\lambda^{\frac{1}{\alpha+3}}u_\lambda(Z^\lambda)$ converge to V. The dynamics of V_t in terms of Z_t^0 follows from Itô's formula. Because $v_\alpha(\cdot)$ is invertible, the substitution $Z_t^0 = v_\alpha^{-1}(V_t)$ yields autonomous dynamics of V_t, with its drift equal to

$$
b(v) = v_\alpha'(v_\alpha^{-1}(v))v + \frac{1}{2}v_\alpha''(v_\alpha^{-1}(v))\bar{Y}^2(1-\bar{Y})^2\sigma^2.
$$

From the definition of $v_\alpha(z)$ and the limit (16), it follows that $v_\alpha(z)\sim B_\alpha^{\frac{1}{\alpha}}A_\alpha^{-\frac{2}{\alpha+1}}\alpha^{-\frac{1}{\alpha+1}}|z|^{\frac{2}{\alpha+1}}$, $v_\alpha'(z)\sim -2B_\alpha^{\frac{1}{\alpha}}A_\alpha^{-\frac{2}{\alpha+1}}\alpha^{-\frac{1}{\alpha+1}}(\alpha+1)^{-1}|z|^{\frac{1-\alpha}{\alpha+1}}$ and $v_\alpha''(z)\sim 2B_\alpha^{\frac{1}{\alpha}}A_\alpha^{-\frac{2}{\alpha+1}}\alpha^{-\frac{1}{\alpha+1}}(\alpha+1)^{-2}(1-\alpha)|z|^{-\frac{2\alpha}{\alpha+1}}$ for $z\to-\infty$. Also, $z\sim -B_\alpha^{-\frac{\alpha+1}{2\alpha}}A_\alpha\alpha^{\frac{1}{2}}v^{\frac{\alpha+1}{2}}$ for $v\to\infty$. Hence, for large v, the first term of the drift $b(v)$ is of order $v^{\frac{3-\alpha}{2}}$ and the second term is of order $v^{-\alpha}$. Therefore, for large v, $b(v)\sim v_\alpha'(v_\alpha^{-1}(v))v\sim -2B_\alpha^{\frac{\alpha+1}{2\alpha}}A_\alpha^{-1}\alpha^{-\frac{1}{2}}(\alpha+1)^{-1}v^{\frac{3-\alpha}{2}}$. The term $a(v)$ in the dynamics for V is equal to $(v_\alpha'(v_\alpha^{-1}(v)))^2\bar{Y}^2(1-\bar{Y})^2\sigma^2$. The same computations yield $a(v)\sim 4B_\alpha^{\frac{\alpha+1}{\alpha}}A_\alpha^{-2}\alpha^{-1}(\alpha+1)^{-2}\bar{Y}^2(1-\bar{Y})^2\sigma^2v^{1-\alpha}$ for large v. □

REFERENCES

Almgren, R. and Chriss, N. (2001), 'Optimal execution of portfolio transactions', *Journal of Risk* **3**, 5–40.

Almgren, R. F. (2003), 'Optimal execution with nonlinear impact functions and trading-enhanced risk', *Applied Mathematical Finance* **10**(1), 1–18.

Almgren, R., Thum, C., Hauptmann, E. and Li, H. (2005), 'Direct estimation of equity market impact', *Risk* **18**, 5752.

Back, K. (1992), 'Insider trading in continuous time', *Review of Financial Studies* **5**(3), 387–409.

Bank, P. and Baum, D. (2004), 'Hedging and portfolio optimization in financial markets with a large trader', *Mathematical Finance* **14**(1), 1–18.

Bank, P. and Kramkov, D. (2015*a*), 'A model for a large investor trading at market indifference prices. I: Single-period case', *Finance and Stochastics* **19**(2), 449–472.

Bank, P. and Kramkov, D. (2015*b*), 'A model for a large investor trading at market indifference prices. II: Continuous-time case', *The Annals of Applied Probability* **25**(5), 2708–2742.

Baruch, S. (2002), 'Insider trading and risk aversion', *Journal of Financial Markets* **5**(4), 451–464.

Cayé, T., Herdegen, M. and Muhle-Karbe, J. (2018), Trading with small nonlinear price impact.

Collin-Dufresne, P., Daniel, K., Moallemi, C. and Saglam, M. (2012), Strategic asset allocation with predictable returns and transaction costs.

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

Cuoco, D. and Cvitanić, J. (1998), 'Optimal consumption choices for a large investor', *Journal of Economic Dynamics and Control* **22**(3), 401–436.

Davis, M. H. and Norman, A. R. (1990), 'Portfolio selection with transaction costs', *Mathematics of Operations Research* **15**(4), 676–713.

Dumas, B. and Luciano, E. (1991), 'An exact solution to a dynamic portfolio choice problem under transactions costs', *The Journal of Finance* **46**(2), 577–595.

Engle, R., Ferstenberg, R. and Russell, J. (2012), 'Measuring and modeling execution cost and risk', *Journal of Portfolio Management* **38**(2), 14.

Filipovic, D. and Mayerhofer, E. (2009), 'Affine diffusion processes: Theory and applications', *Advanced Financial Modelling* **8**, 1–40.

Frazzini, A., Israel, R. and Moskowitz, T. (2012), Trading costs of asset pricing anomalies.

Gabaix, X., Gopikrishnan, P., Plerou, V. and Stanley, H. E. (2003), 'A theory of power-law distributions in financial market fluctuations', *Nature* **423**(6937), 267–270.

Gabaix, X., Gopikrishnan, P., Plerou, V. and Stanley, H. E. (2006), 'Institutional investors and stock market volatility', *The Quarterly Journal of Economics* **121**(2), 461–504.

Gârleanu, N. and Pedersen, L. H. (2013), 'Dynamic trading with predictable returns and transaction costs', *The Journal of Finance* **68**(6), 2309–2340.

Gârleanu, N. and Pedersen, L. H. (2016), 'Dynamic portfolio choice with frictions', *Journal of Economic Theory* **165**, 487–516.

Garleanu, N., Pedersen, L. H. and Poteshman, A. M. (2009), 'Demand-based option pricing', *Review of Financial Studies* **22**(10), 4259–4299.

Gatheral, J. (2010), 'No-dynamic-arbitrage and market impact', *Quantitative Finance* **10**(7), 749–759.

Gatheral, J., Schied, A. and Slynko, A. (2012), 'Transient linear price impact and Fredholm integral equations', *Mathematical Finance* **22**(3), 445–474.

Gerhold, S., Guasoni, P., Muhle-Karbe, J. and Schachermayer, W. (2014), 'Transaction costs, trading volume, and the liquidity premium', *Finance and Stochastics* **18**(1), 1–37.

Grinold, R. C. and Kahn, R. N. (2000), *Active portfolio management*, McGraw-Hill.

Grossman, S. J. and Zhou, Z. (1993), 'Optimal investment strategies for controlling drawdowns', *Mathematical Finance* **3**(3), 241–276.

Guasoni, P. and Robertson, S. (2012), 'Portfolios and risk premia for the long run', *The Annals of Applied Probability* **22**(1), 239–284.

Guasoni, P. and Weber, M. (2017), 'Dynamic trading volume', *Mathematical Finance* **27**(2), 313–349.

Hasbrouck, J. and Seppi, D. J. (2001), 'Common factors in prices, order flows, and liquidity', *Journal of Financial Economics* **59**(3), 383–411.

Huberman, G. and Stanzl, W. (2004), 'Price manipulation and quasi-arbitrage', *Econometrica* **72**(4), 1247–1275.

Janeček, K. and Shreve, S. E. (2004), 'Asymptotic analysis for optimal investment and consumption with transaction costs', *Finance and Stochastics* **8**(2), 181–206.

Kyle, A. S. (1985), 'Continuous auctions and insider trading', *Econometrica* **53**(6), 1315–1335.

Lillo, F., Farmer, J. D. and Mantegna, R. N. (2003), 'Econophysics: Master curve for price-impact function', *Nature* **421**(6919), 129–130.

Liu, R., Muhle-Karbe, J. and Weber, M. H. (2017), 'Rebalancing with linear and quadratic costs', *SIAM Journal on Control and Optimization* **55**(6), 3533–3563.

Loeb, T. F. (1983), 'Trading cost: the critical link between investment information and results', *Financial Analysts Journal* **39**(3), 39–44.

Merton, R. (1969), 'Lifetime portfolio selection under uncertainty: The continuous-time case', *The Review of Economics and Statistics* **51**(3), 247–257.

Moreau, L., Muhle-Karbe, J. and Soner, H. M. (2017), 'Trading with small price impact', *Mathematical Finance* **27**(2), 350–400.

Obizhaeva, A. A. and Wang, J. (2013), 'Optimal trading strategy and supply/demand dynamics', *Journal of Financial Markets* **16**(1), 1–32.

Plerou, V., Gopikrishnan, P., Gabaix, X. and Stanley, H. E. (2002), 'Quantifying stock-price response to demand fluctuations', *Physical Review E* **66**(2), 027104.

Predoiu, S., Shaikhet, G. and Shreve, S. (2011), 'Optimal execution in a general one-sided limit-order book', *SIAM Journal on Financial Mathematics* **2**(1), 183–212.

Rogers, L. C. G. (2004), Why is the effect of proportional transaction costs $O(\delta^{2/3})$?, *in* 'Mathematics of finance', Vol. 351 of *Contemp. Math.*, Amer. Math. Soc., Providence, RI, pp. 303–308.
URL: *https://doi.org/10.1090/conm/351/06411*

Schied, A. and Schöneborn, T. (2009), 'Risk aversion and the dynamics of optimal liquidation strategies in illiquid markets', *Finance and Stochastics* **13**(2), 181–204.

Schied, A., Schöneborn, T. and Tehranchi, M. (2010), 'Optimal basket liquidation for cara investors is deterministic', *Applied Mathematical Finance* **17**(6), 471–489.

Soner, H. M. and Touzi, N. (2013), 'Homogenization and asymptotics for small transaction costs', *SIAM Journal on Control and Optimization* **51**(4), 2893–2921.

Stroock, D. W. and Varadhan, S. S. (1979), *Multidimensional diffusion processes*, Vol. 233, Springer Verlag.

Torre, N. and Ferrari, M. J. (1998), 'The market impact model', *Horizons, The Barra Newsletter* (165).

Toth, B., Lemperiere, Y., Deremble, C., De Lataillade, J., Kockelkoren, J. and Bouchaud, J.-P. (2011), 'Anomalous price impact and the critical nature of liquidity in financial markets', *Physical Review X* **1**(2), 021006.

Vath, V. L., Mnif, M. and Pham, H. (2007), 'A model of optimal portfolio selection under liquidity risk and price impact', *Finance and Stochastics* **11**(1), 51–90.