Dynamic Portfolio Choice with Frictions

Garleanu, Nicolae; Heje Pedersen, Lasse

Document Version
Final published version

Published in:
Journal of Economic Theory

DOI:
10.1016/j.jet.2016.06.001

Publication date:
2016

Available online at www.sciencedirect.com

ScienceDirect

Journal of Economic Theory 165 (2016) 487–516

JOURNAL OF Economic Theory

www.elsevier.com/locate/jet

Dynamic portfolio choice with frictions ☆

Nicolae Gârleanu [a,b,c,*], Lasse Heje Pedersen [d,e,f,c]

[a] *Haas School of Business, University of California, Berkeley, United States*
[b] *NBER, United States*
[c] *CEPR, United Kingdom*
[d] *Copenhagen Business School, Denmark*
[e] *New York University, United States*
[f] *AQR Capital Management, United States*

Received 7 November 2014; final version received 1 April 2016; accepted 2 June 2016

Available online 23 June 2016

Abstract

We show how portfolio choice can be modeled in continuous time with transitory and persistent transaction costs, multiple assets, multiple signals predicting returns, and general signal dynamics. The objective function is derived from the limit of discrete-time models with endogenous transaction costs due to optimal dealer behavior. We solve the model explicitly and the intuitive solution is also the limit of the solutions of the corresponding discrete-time models. We show how the optimal high-frequency trading strategy depends on the nature of the trading costs, which in turn depend on dealers' inventory dynamics. Finally, we provide equilibrium implications and illustrate the model's broader applicability to micro- and macro-economics, monetary policy, and political economy.

☆ We are grateful for helpful comments from Kerry Back, Darrell Duffie, Pierre Collin-Dufresne, Andrea Frazzini, Esben Hedegaard, Brian Hurst, David Lando, Hong Liu (discussant), Anthony Lynch, Ananth Madhavan (discussant), Stavros Panageas, Andrei Shleifer, and Humbert Suarez, as well as from seminar participants at Stanford GSB, AQR Capital Management, UC Berkeley, Columbia University, NASDAQ OMX Economic Advisory Board Seminar, University of Tokyo, New York University, the University of Copenhagen, Rice University, University of Michigan Ross School, Yale University SOM, the Bank of Canada, the Journal of Investment Management Conference, London School of Economics, and UCLA. Pedersen gratefully acknowledges support from the European Research Council (ERC grant No. 312417) and the FRIC Center for Financial Frictions (grant No. DNRF102).
* Corresponding author.
E-mail address: garleanu@haas.berkeley.edu (N. Gârleanu).
URL: http://www.lhpedersen.com/ (L.H. Pedersen).

http://dx.doi.org/10.1016/j.jet.2016.06.001

JEL classification: G11; G12; G23; C6; D53; E44

Keywords: Dynamic trading; Frictions; Transaction costs; Continuous time; Predictability; Equilibrium

A fundamental question in financial economics is how to choose an optimal portfolio. Investors must choose their portfolio in light of the current risks, expected returns, and transaction costs of all available assets, as well as how often they can trade in the future and the future evolution of the risks and returns. This portfolio choice depends crucially on the future trading opportunities for several reasons: First, expected returns are driven by multiple economic factors that vary over time, leading to variation in the optimal portfolio.[1] Second, transaction costs imply that an investor must consider the portfolio's optimality both currently and in the future. Third, investors must decide how often to trade and how much to trade. A number of questions arise from these dynamic considerations: What is the difference between trading in markets that are open continuously versus discrete markets? How do transaction costs change when markets are open continuously rather than at discrete times? Why do high-frequency trading (HFT) firms and other investors trade continuously throughout the day when most existing models with transaction costs imply infrequent, lumpy trading? What are the implications for asset-price dynamics?

We provide a general and tractable framework to address these issues. First, to study portfolio choice for high- and low-frequency trading, we show how to formulate the problem in continuous time such that the objective function is a limit of discrete-time models in which transaction costs arise endogenously from dealer behavior. As a result, we clarify how transaction costs can be captured consistently for high- and low-frequency trades and how model parameters scale with time. Second, we solve the continuous-time model and derive a simple expression for the optimal high-frequency portfolio choice. The tractability of our framework contrasts with that of standard models in the literature based on proportional transaction costs.[2] These standard models are complex and rely on numerical solutions even in the case of a *single* asset with *i.i.d.* returns (i.e., no return predicting factors).[3] In contrast, our framework based on quadratic costs allows a closed-form optimal portfolio choice with multiple assets and multiple return-predicting factors. The assumption that transaction costs are quadratic in the number of securities traded is natural since it is equivalent to a linear price impact. Third, we show how the continuous-time solution obtains as the limit of optimal discrete-time portfolios. Fourth, we derive implications for equi-

[1] See, e.g., Campbell and Viceira (2002) and Cochrane (2011) and references therein.

[2] In discrete time, quadratic costs have been shown to provide tractability, and we rely in particular on Gârleanu and Pedersen (2013). In addition to introducing a continuous-time model, our contributions are to generalize the framework, consider a micro foundation for trading costs, derive the connection between discrete and continuous time, and provide equilibrium implications. See also Heaton and Lucas (1996) and Grinold (2006) who also assume quadratic costs, Glasserman and Xu (2013) who extend the model of Gârleanu and Pedersen (2013) to account for robust optimization, and Collin-Dufresne et al. (2014) who show how to linearize — and thus solve approximately — a more general and useful class of portfolio-choice models.

[3] There is an extensive literature on proportional transaction costs following Constantinides (1986). Davis and Norman (1990) provide a more formal analysis and Liu (2004) determines the optimal trading strategy for an investor with constant absolute risk aversion (CARA) and many independent securities with both fixed and proportional costs (without predictability). The assumptions of CARA and independence across securities imply that the optimal position for each security is independent of the positions in the other securities. Also, our paper is related to the literature on optimal trade execution (e.g., Perold, 1988; Bertsimas and Lo, 1998; Almgren and Chriss, 2000; Obizhaeva and Wang, 2006; Engle and Ferstenberg, 2007, and Gatheral and Schied, 2011), although this literature treats the total traded quantity as given exogenously while it is part of our solution.

librium expected returns. Finally, we provide several additional applications of our framework to other issues in social science.

To understand the intricacies in studying high-frequency trading, consider first the following apparent puzzle, namely whether market impact costs matter at all in continuous time. For instance, in the model of Cetin et al. (2006), transaction costs are irrelevant in continuous time. To see why quadratic costs might be irrelevant in continuous time, consider splitting a trade into two equal parts. The quadratic transaction cost of each part of the trade is $(\frac{1}{2})^2 = \frac{1}{4}$ of the cost of the original trade, leading to a total cost that is half (two times $\frac{1}{4}$) what it was before. This insight leads to two apparent conclusions, the latter of which we wish to dispel: (i) Splitting orders up and trading gradually over time is optimal, as is the case in our optimal strategy and in real-world electronic markets. (ii) One can continue to halve one's cost by splitting the trade up further, so the cost goes to zero as trading approaches continuous time. We refute point (ii) under certain conditions, as it relies on an implicit assumption that, when the trading frequency increases, the parameter of the quadratic cost function remains unchanged. This assumption does not hold in general when trading costs are micro-founded; in this case, instead, the unchanging quantity is the transaction cost incurred per time unit if trading a given number of shares per time unit.

To provide an economic foundation for a continuous-time model with transaction costs, we discretize the model and let transaction costs arise endogenously due to dealers' inventory considerations à la Grossman and Miller (1988).[4] We consider both persistent and transitory costs, corresponding to dealers who can lay off their inventory gradually or in one shot. We show that the discrete-time persistent market-impact costs converge to a continuous-time model with the same persistent market-impact parameter and a resiliency parameter that depends on the length of the time periods to the first order.

There are two ways to model the dependence of the transitory costs on the trading frequency: (a) If dealers can always lay off their inventory in one time period, then shorter time periods imply that dealers need only hold inventories for a shorter time and, in this case, transitory costs vanish in the limit. (b) If, instead, the time it takes dealers to unload inventories does not go to zero even as the trading frequency increases, then transitory costs survive in the limit. In this case, the limit transaction costs are quadratic in the *trading intensity*, i.e., the number of shares traded per time unit.

In either case, we show that trading costs and, more broadly, the objective function, converge to their continuous-time counterparts as trading frequencies increase. We derive the optimal portfolio in discrete and continuous time and naturally the optimal discrete-time portfolio also converges to the continuous-time solution. In the case with vanishing transitory costs, the optimal continuous-time portfolio has positive quadratic variation. With transitory costs, however, our optimal continuous-time strategy is smooth and has a finite turnover. Hence, while the trading strategies appear to have the same structure in discrete time (in fact, given by the same equation), their continuous-time limits are qualitatively different in the two cases.

The more realistic case is arguably the one with a smooth trading strategy with finite turnover, corresponding to finite turnover speed of dealer inventories. This case provides an economic foundation for quadratic transaction costs that matter in continuous time.[5] At the same time, the model shows how continuous trading (e.g., high frequency trading) can be optimal, yet accomplished with a limited turnover. Further, it shows how to scale transaction costs parameters with

[4] Inventory models with multiple correlated assets include Greenwood (2005) and Gârleanu et al. (2009).

[5] We thus offer a justification for the specification employed in such studies as Carlin et al. (2008) and Oehmke (2009).

time frequencies such that the model solution is (almost) the same independent of whether we study trading at the second or millisecond frequencies (in contrast, if transaction costs did not matter in continuous time, then it would follow either that the discrete-time models rely heavily on the sufficient length of the time period or that transaction costs also have a small effect in these models).

Our optimal strategy is qualitatively different from the strategy with proportional or fixed transaction costs, which exhibits long periods of no trading. Our strategy resembles the method used by many real-world traders in electronic markets, namely to continuously post limit orders close to the best bid or ask. The trading speed is the limit orders' "fill rate," which naturally depends on the price-aggressiveness of the limit orders, i.e., on the cost that the trader is willing to accept — just as in our model. Our strategy has several advantages in the real world, according to discussions with people who design trading systems: Trading continuously minimizes the order sizes at each point in time and exploits the liquidity that is available throughout the day (or week, month, etc.), rather than submitting large infrequent orders when limited liquidity may be available. Consistently, the empirical literature generally finds transaction costs to be convex (e.g., Engle et al., 2008; Lillo et al., 2003), with some researchers estimating quadratic trading costs (e.g., Breen et al., 2002 and Kyle and Obizhaeva, 2011), including for large orders (Kyle and Obizhaeva, 2012). Huberman and Stanzl (2004) show that the persistent price impact must be linear to rule out arbitrage opportunities. Chacko et al. (2008) model transaction costs as a monopolistic market-maker's price of immediacy and find evidence of a market impact that increases with the square root of the order size, corresponding to a total cost that increases with order size raised to power 3/2 (rather than quadratic). Nevertheless, the main features of their model, namely transaction costs that are convex and do not vanish in continuous time, are the very ingredients our theory and micro foundation rely on.

We also show how our continuous-time model can help analyze equilibrium asset price dynamics.[6] We study an equilibrium model in which rational investors facing transaction costs trade with several groups of noise traders who provide a time-varying excess supply or demand of assets. We show that, in order for the market to clear, the investors must be offered return premiums depending on the properties of the noise-traders' positions. In particular, the noise-trader positions that mean revert more quickly generate larger alphas in equilibrium, as the rational investors must be compensated for incurring higher transaction costs per time unit. Long-lived supply fluctuations, on the other hand, give rise to smaller and more persistent alphas. This can help explain the short-term return reversals documented by Lehman (1990) and Lo and MacKinlay (1990), and their relation to transaction costs documented by Nagel (2012).

Finally, linear–quadratic models are widely used in social science (see Ljungqvist and Sargent, 2004; Hansen and Sargent, 2014, and references therein). We contribute to this broader literature in two ways. First, the general solution comes down to algebraic matrix Riccati equations requiring numerical solutions, while we solve our model explicitly, including the Riccati equations. Second, we consider how to act optimally in light of frictions and several signals with varying mean-reversion rates in the linear–quadratic framework. This leads to insights with broad implications across social science as we discuss in the concluding section of the paper. For example, a central bank may receive several signals about inflation (e.g., core inflation versus headline inflation, or in several regions, or across several product markets) and face costs of changing

[6] The literature on equilibrium asset pricing with trading costs includes Amihud and Mendelson (1986), Vayanos (1998), Vayanos and Vila (1999), Lo et al. (2004), Jang et al. (2007), Gârleanu (2009), and Lagos (2010), and the literature on asset pricing with time-varying trading costs includes Acharya and Pedersen (2005), Lynch and Tan (2011).

monetary policy. A politician may face varying signals from several constituents and incur costs from political changes. A firm may receive several signals about consumer preferences and face costs to adjusting its products. The macro economy may face different signals about total factor productivity (TFP) and capital adjustment costs. In each of these examples, our framework can be used to see how to optimally weight the signals in light of their dynamics and costs. Our model shows that the optimal policy moves gradually in the direction of an aim, which incorporates an average of current and future expected signals. This means that the decision maker should give more weight to persistent signals. Specifically, the model shows explicitly how a firm should weight persistent consumer trends, a central bank should weight core inflation over transitory inflation, and the macro capital adjustment should be based on persistent TFP shocks.

1. Optimal trading in continuous markets

We start by introducing our tractable continuous-time framework and illustrating its solution, the optimal "high-frequency" trading strategy. We first consider the case of purely transitory transaction costs, then introduce persistent transaction costs, and finally consider the case of purely persistent costs.

We show that the nature of the optimal trading strategy is fundamentally different in the cases with and without transitory transaction costs, and we discuss at a deeper level in Sections 2–3 when each case is most likely to apply. Further, Sections 2–3 also lay the foundation for a number of modeling choices, including the investor's continuous-time objective function, as these are far from obvious when starting directly from continuous time, as we shall see. All proofs are in the (on-line) appendix.

1.1. Purely temporary transaction costs

An investor must choose an optimal portfolio among S risky securities and a risk-free asset. The risky securities have prices p with dynamics

$$dp_t = \left(r^f p_t + Bf_t\right) dt + du_t. \tag{1}$$

Here, f_t is a $K \times 1$ vector which contains the factors that predict excess returns, B is an $S \times K$ matrix of factor loadings, and u is an unpredictable "noise term," i.e., a martingale (e.g., a Brownian motion or a jump diffusion) with instantaneous variance–covariance matrix $\text{var}_t(du_t) = \Sigma dt$.[7]

The return-predicting factors follow a general Markovian jump diffusion:

$$df_t = \mu_f(f_t)dt + d\varepsilon_t, \tag{2}$$

where ε is a martingale. Like the innovation u, ε can contain both Brownian and jump components. We impose on the dynamics of f conditions sufficient to ensure that it is stationary, with finite first two moments. Occasionally, we also make use of the following assumption that specifies the matrix Φ of mean-reversion rates for the factors.

[7] We note that, in the interest of tractability, we model returns per security, i.e., absolute changes in price levels, rather than proportional returns. This choice is conducive to a linear–quadratic solution. Collin-Dufresne et al. (2014) compute approximate solutions in a proportional-return discrete-time model, and find little quantitative difference from our (discrete-time) solution when returns are heteroscedastic.

Assumption A1. The drift of f_t is given by $\mu_f(f) = -\Phi f$.

The agent chooses his trading intensity $\tau_t \in \mathbb{R}^S$, which determines the rate of change of his position x_t:

$$dx_t = \tau_t dt. \tag{3}$$

We only consider smooth portfolio policies here because discrete jumps in positions or quadratic variation would be associated with infinite trading costs in this setting. This idea is based on the discrete-time foundation for temporary transaction costs in Proposition 4 below, which shows that such non-smooth strategies would have infinite transaction costs when the length of the trading periods approaches zero.[8]

The transaction cost TC per time unit of trading with intensity τ_t is

$$TC(\tau_t) = \frac{1}{2}\tau_t^\top \Lambda \tau_t. \tag{4}$$

Here, Λ is a symmetric positive-definite matrix measuring the level of trading costs.[9] As seen in the micro foundation in Section 3, this quadratic transaction cost arises as a trade Δx_t shares moves the price by $\frac{1}{2}\Lambda\Delta x_t$, and this results in a total trading cost of Δx_t times the price move. This is a multi-dimensional version of Kyle's lambda. Most of our results hold with this general transaction cost function, but some of the resulting expressions are simpler in the following special case.

Assumption A2. Transaction costs are proportional to the amount of risk: $\Lambda = \lambda\Sigma$ for a scalar $\lambda > 0$.

This assumption is natural and, in fact, implied by the micro-foundation that we provide in Section 3.2. To understand this, suppose that a dealer takes the other side of the trade Δx_t, holds this position for a period of time dt, and "lays it off" at the end of the period. Then the dealer's risk is $\Delta x_t^\top \Sigma \Delta x_t\, dt$ and the trading cost is the dealer's compensation for risk, depending on the dealer's risk aversion reflected by λ. Section 3.2 further analyzes the conditions under which the compensation for risk is strictly positive.

The investor chooses his optimal trading strategy to maximize the present value of the future stream of expected excess returns, penalized for risk and trading costs:

$$\max_{(\tau_s)_{s\geq t}} E_t \int_t^\infty e^{-\rho(s-t)} \left(x_s^\top B f_s - \frac{\gamma}{2} x_s^\top \Sigma x_s - \frac{1}{2}\tau_s^\top \Lambda \tau_s \right) ds. \tag{5}$$

This objective function means that the investor has mean-variance preferences over the change in his wealth W_t each time period. The objective can be shown to approximate a standard utility function or it can be viewed as that of an asset manager who seeks a high Sharpe ratio. Also, this type of objective function is widely used in macro-economics (see Hansen and Sargent, 2014 and references therein).

[8] E.g., if the agent trades n shares over a time period of Δt, then the cost according to (4) is $\int_0^{\Delta t} TC(\frac{n}{\Delta t})dt = \frac{1}{2}\Lambda\frac{n^2}{\Delta t}$, which approaches infinity as Δt approaches 0.

[9] The assumption that Λ is symmetric is without loss of generality. To see this, suppose that $TC(\Delta x_t) = \frac{1}{2}\Delta x_t^\top \hat{\Lambda} \Delta x_t$, where $\hat{\Lambda}$ is not symmetric. Then, letting Λ be the symmetric part of $\hat{\Lambda}$, i.e., $\Lambda = (\hat{\Lambda} + \hat{\Lambda}^\top)/2$, generates the same trading costs as $\hat{\Lambda}$.

We conjecture and verify that the value function is quadratic in x:

$$V(x, f) = -\frac{1}{2}x^\top A_{xx}x + x^\top A_x(f) + A(f). \tag{6}$$

We solve the model explicitly, as the following proposition states. It is helpful to compare our result with the optimal portfolio under the classical no-friction assumption, for which we use the notation *Markowitz* as a reference to the classical findings of Markowitz (1952):

$$Markowitz_t = (\gamma\Sigma)^{-1} Bf_t. \tag{7}$$

The Markowitz portfolio has an optimal trade-off between risk (Σ) and expected excess return (Bf_t), leveraged to suit the agent's risk aversion (γ). We show that the optimal portfolio in light of transaction costs moves gradually towards an "aim portfolio," which incorporates current and future expected Markowitz portfolios.

Proposition 1.

(i) *There exists a unique optimal portfolio strategy.*

(ii) *The optimal portfolio x_t tracks a moving "aim portfolio" $\bar{M}^{aim}(f_t)$ with a tracking speed of $\bar{M}^{rate}$. That is, the optimal trading intensity $\tau_t = \frac{dx_t}{dt}$ is*

$$\tau_t = \bar{M}^{rate}\left(\bar{M}^{aim}(f_t) - x_t\right), \tag{8}$$

where the coefficient matrix $\bar{M}^{rate}$ is given by

$$\bar{M}^{rate} = \Lambda^{-1}A_{xx} \tag{9}$$

$$A_{xx} = -\frac{\rho}{2}\Lambda + \Lambda^{\frac{1}{2}}\left(\gamma\Lambda^{-\frac{1}{2}}\Sigma\Lambda^{-\frac{1}{2}} + \frac{\rho^2}{4}I\right)^{\frac{1}{2}}\Lambda^{\frac{1}{2}} \tag{10}$$

and the aim portfolio by

$$\bar{M}^{aim}(f_t) = A_{xx}^{-1}A_x(f_t), \tag{11}$$

and $A_x(f)$ satisfies a second-order ODE given in the Appendix.

(iii) *The aim portfolio $M^{aim}(f)$ has the intuitive representation*

$$M^{aim}(f) = b\int_0^\infty e^{-bt}\,\mathrm{E}\left[Markowitz_t \mid f_0 = f\right]dt \tag{12}$$

with $b = \gamma A_{xx}^{-1}\Sigma$.

(iv) *Under Assumption A2, the solution simplifies: $A_{xx} = a\Sigma$, $b > 0$ is a scalar, and*

$$\bar{M}^{rate} = a/\lambda = \frac{1}{2}(\sqrt{\rho^2 + 4\gamma/\lambda} - \rho) \tag{13}$$

$$\bar{M}^{aim} = \gamma^{-1}\Sigma^{-1}B\,(I + a/\gamma\Phi)^{-1}, \tag{14}$$

where the last equation also requires Assumption A1, $\mu(f) = -\Phi f$.

This proposition provides an intuitive method of portfolio choice. The optimal portfolio can be written in a simple closed-form expression. In light of the literature on portfolio choice with

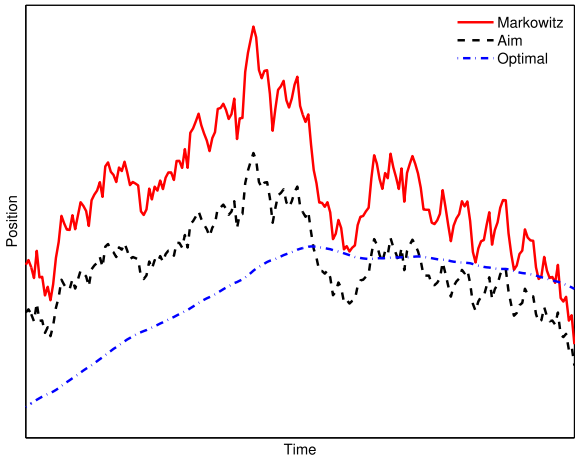

Fig. 1. **Optimal portfolio with one asset and temporary impact costs.** The solid line shows the evolution over time of the Markowitz portfolio for a simulated path of the model, that is, the optimal portfolio in the absence of transaction costs. The dashed line shows the corresponding aim portfolio. Lastly, the dash-dotted line shows the optimal portfolio. The optimal portfolio is smooth, to save on transaction costs, and moves gradually in the direction of the aim portfolio.

proportional transaction costs (Constantinides, 1986), which relies on numerical results even for a single asset with i.i.d. returns, our framework offers remarkable tractability even with correlated multiple assets and multiple signals.

It is intuitive that the optimal portfolio trades toward an aim portfolio, which is a weighted average of future expected Markowitz portfolios. This means that persistent signals are given more weight as they affect the Markowitz portfolio for a longer time period. This result is seen most clearly in Equation (14). The aim portfolio is seen to be almost of the same form as the Markowitz portfolio, except that the signals are scaled down by $(I + a/\gamma\Phi)^{-1}$. Given that Φ contains the signals' mean-reversion rates, this means that quickly mean-reverting signals are scaled down more while more persistent ones receive more weight. This dependence on the signals' mean-reversion rate is greater with larger transaction costs, that is, a/γ (which multiplies the mean-reversion Φ) increases with λ.

This intuitive portfolio strategy is the continuous-time counterpart of the discrete-time solution of Gârleanu and Pedersen (2013), but here we solve it for general factor dynamics and, under Assumption A1, the solution for trading rate (13) is even simpler in continuous time. Indeed, for a patient agent with $\rho \cong 0$, we see that the trading rate is approximately $\sqrt{\gamma/\lambda}$. More generally, we see directly that the trading rate is decreasing in the transaction cost λ and increasing in risk aversion γ.

Example. To illustrate the optimal portfolio choice with frictions, we consider a specific example as seen in Fig. 1. We solve the model and simulate of path over time based on the following parameters: $\Phi = 1$, $\Sigma = 1$, $f_0 = 1$, $\mathrm{Var}(df_t)/dt = 1$, $\gamma = 0.4$, $\rho = 0.05$, $B = 1$, $x_0 = 0$, and $\lambda = 0.1$.

Fig. 1 plots the evolution of the Markowitz portfolio in an economy with a single asset, say an equity-market index. The agent must decide on his equity allocation in light of his time-varying estimate of the equity premium while being mindful of transaction costs. To do this, he constructs an aim portfolio as seen in the figure. The aim portfolio is a more conservative version of the Markowitz portfolio due to transaction costs and because the agent anticipates that the Markowitz

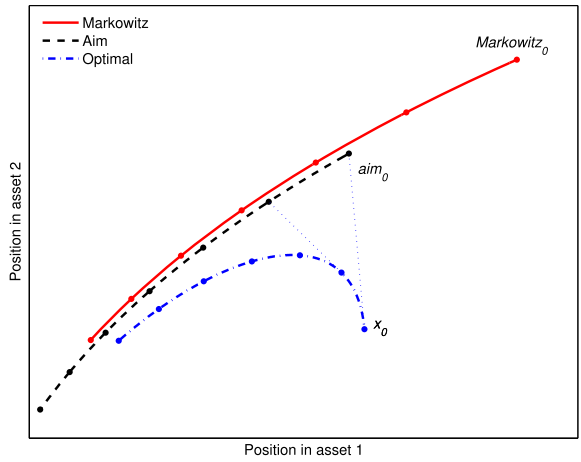

Fig. 2. **Expected optimal portfolio with two assets.** The point $Markowitz_0$ would be the optimal current portfolio in the absence of transaction costs. The solid line shows the expected future evolution of the Markowitz portfolio as the return-predicting factors mean-revert. The current aim portfolio is labeled aim_0, and it is derived as a weighted average of the current and future expected Markowitz portfolios. The dashed line shows the expected evolution of the aim portfolio, which seeks to lead the Markowitz portfolio. The current portfolio is labeled x_0. As the trader optimally rebalances toward the aim portfolio, the portfolio is expected to evolve along the dash-dotted line.

portfolio will mean-revert. Finally, the agent's optimal portfolio, also plotted in Fig. 1, smoothly moves toward the aim portfolio, thus saving on transaction costs while capturing most of the benefits on the Markowitz portfolio.

Interesting, we see that there are times when the optimal portfolio is below the Markowitz portfolio and above the aim, such that the optimal strategy is selling (i.e., negative $dx/dt = \tau$) even though the best risk-return trade-off is above. This selling is motivated by the agent's anticipation that the Markowitz portfolio will go down in the future, and, to save on transaction costs, it is cheaper to start selling gradually already now.

While it may not be easy to see in the figure, the distance between the aim portfolio and the Markowitz portfolio varies over time. This distance depends on which signal is driving the current high Markowitz portfolio — a persistent signal increases the aim portfolio more than a signal that will quickly mean-revert, while the signal's mean-reversion rates are irrelevant for Markowitz portfolio (and have not been studied by the portfolio choice literature more broadly, with the exception of Gârleanu and Pedersen, 2013).

Fig. 2 illustrates the optimal portfolio choice with two assets. There are several differences in this illustration. First, the horizontal axis is now the allocation to asset one and the vertical axis the allocation to asset two. Second, rather than considering how the optimal portfolio evolves over time as shocks hit the economy, we consider its *expected* path. Third, the parameters for this two-asset case are $\gamma = 0.4$, $\rho = 0.05$, $\lambda = 2$, $\Sigma = \mathrm{Var}(df_t)/dt = B = \mathrm{diag}(1, 1)$, $\Phi = \mathrm{diag}(0.2, 0.05)$, $f_0 = (1, 1)$, and $x_0 = (1.5, 1.5)$.

We see that the Markowitz portfolio is expected to mean-revert along a concave curve. The concavity reflects that the signal that currently predicts a high return of asset 2 is more persistent. The current aim portfolio (aim_0 in the figure) is an average of the current and future expected Markowitz portfolios so it lies in between the points on the solid curve. The optimal portfolio trades in the direction of the aim and it is expected to eventually approach the future Markowitz

portfolio. Intuitively, the initial trading process focuses on buying shares of asset two, which is expected to have a high return over an extended time period.

1.2. Temporary and persistent transaction costs

We modify the set-up above by adding persistent transaction costs. Specifically, the agent transacts at price $\bar{p}_t = p_t + D_t$, where the distortion D_t evolves according to

$$dD_t = -RD_t\,dt + C dx_t = -RD_t\,dt + C\tau_t\,dt, \tag{15}$$

where the scalar R is the price resiliency. The agent's objective now becomes

$$\max_{(\tau_s)_{s\geq t}} E_t \int_t^\infty e^{-\rho(s-t)} \left(x_s^\top \left(Bf_s - (r^f + R)D_s + C\tau_s \right) - \frac{\gamma}{2} x_s^\top \Sigma x_s - \frac{1}{2}\tau_s^\top \Lambda \tau_s \right) ds. \tag{16}$$

This objective function is similar to the one from above, but it has some new terms that multiply the position x_s. Indeed, since the price including the persistent distortion is $\bar{p}_s = p_s + D_s$, the expected excess return is now given by the expected excess return of p_s, which is Bf_s as before, plus the expected excess return of D_s, which is given by (15) in excess of the risk-free rate r^f. It might appear odd that the agent seems to benefit from buying and pushing the price higher, but this benefit leads to a loss as the distortion D decays.

It is no longer true in general that the objective (16) is concave in $\{\tau_t\}_t$, since the gain from the immediate increase in the mark-to-market value of the portfolio may exceed the loss from the (discounted) round-trip transaction costs. We therefore have to restrict attention to parameter configurations for which the objective is, indeed, concave. The fact that such configurations — with $C \neq 0$ — exist is ensured by Lemma 1, which provides sufficient conditions for concavity.

Lemma 1. *The objective function (16) is concave in $\{\tau_t\}_t$ if the persistent-impact matrix C is symmetric positive definite and $\gamma \geq (\rho - r^f)\|\Sigma^{-\frac{1}{2}} C \Sigma^{-\frac{1}{2}}\|$.*

The second condition roughly states that the price impact C is not too large relative to the trader's perceived risk cost $\gamma\Sigma$; the condition is automatically satisfied if $\rho \leq r^f$.

We conjecture, as before, a value function that is quadratic in the endogenous state variable (x_t, D_t) and the factor f_t. Specifically, we write

$$\begin{aligned} V(x, D, f) = &-\frac{1}{2} x^\top A_{xx} x + x^\top A_{xy} D + \frac{1}{2} D^\top A_{DD} D + x^\top A_x(f) \\ &+ D^\top A_D(f) + A_{ff}(f). \end{aligned} \tag{17}$$

Under an appropriate transversality condition, the value function exists and must have this form. We now focus on the optimal trading strategy.

Proposition 2.

(i) *The optimal trading intensity has the form*

$$\tau_t = \bar{M}^{rate} \left(\bar{M}_f^{aim}(f_t) + \bar{M}_D^{aim} D_t - x_t \right) \tag{18}$$

for appropriate matrices $\bar{M}^{rate}$ and $\bar{M}_D^{aim}$ and function $\bar{M}_f^{aim}$ solving an ODE given in the proof.

(ii) *An equivalent representation of the portion of the aim due to f is*

$$\bar{M}_f^{aim}(f) = N_2 \int_0^\infty e^{-N_1 t} N_3 \, \mathrm{E}\left[Markowitz_t | f_0 = f\right] dt \tag{19}$$

for matrices N_i given explicitly in the appendix.

We see that the optimal portfolio choice continues to have the same intuitive characteristics as in the model with only temporary impact costs. The optimal portfolio trades toward an aim, which now depends both on the current signals and the current persistent price distortions. The current signals affect the aim through a combination of their implied current and future Markowitz portfolios.

1.3. Purely persistent costs

The set-up is as above, but now we take $\Lambda = 0$. Under this assumption, it no longer follows from the micro foundation in Section 2 that x_t has to be of the form $dx_t = \tau_t \, dt$ for some τ. Indeed, with purely persistent price-impact costs, the optimal portfolio policy can have jumps and infinite quadratic variation (i.e., "wiggle" like a Brownian motion).

The price distortion D evolves as before, except that $\tau_t \, dt$ is replaced by dx_t:

$$dD_t = -RD_t \, dt + C dx_t. \tag{20}$$

We define the objective of the trader to be

$$E_t \int_t^\infty e^{-\rho(s-t)} \left(x_s^\top \left(Bf_s - (r^f + R) D_s \right) - \frac{\gamma}{2} x_s^\top \Sigma x_s \right) ds \tag{21}$$

$$+ \, E_t \int_t^\infty e^{-\rho(s-t)} x_{s-}^\top C dx_s + \frac{1}{2} E_t \int_t^\infty e^{-\rho(s-t)} d\left[x, Cx\right]_s .$$

The notation $[X, Y]_t$ stands for the quadratic variation of processes X and Y and $d[X, Y]_t$ is the corresponding innovation, which can be interpreted as the instantaneous covariance between the two processes. This objective function is formally justified by Proposition 4 below, but here we provide some intuition. While the overall objective function appears complex, when broken down in its components, we see that each term arises naturally from the micro foundation.

The terms in the first row of (21) are as before. Also, the first term in the second row is as before, although here it is written more generally. Indeed, $x_{s-}^\top C dx_s = x_s^\top C \tau_s ds$ when the portfolio is continuous and of bounded variation as it was above. This term captures the mark-to-market profit on the old position due to the market impact of the new trade, as before.

The last term is new. It records the instantaneous mark-to-market gain on the just-purchased units dx_s. Specifically, the new trade moves the price distortion by $C dx_s$ and we assume that the trade is executed at an average of the pre- and post-trade prices, which leads to a mark-to-market

profit of $\frac{1}{2}$ times the price move. As the price distortion eventually disappears, this short-term gain is more than reversed later.[10]

A helpful observation in this case is that making a large trade Δx over an infinitesimal time interval has an easily described impact on the value function. Suppose that the investor decides to trade from x to 0 instantaneously — thus, with $x_{t-} = x$, $dx_t = -x$ (a jump). As specified by (20), the distortion D_t decreases by Cx. The trade also has a direct impact on the P&L flow at time t, via the last two terms of (21), which capture jumps. Specifically, the first of the two is $x^\top C(-x)$, while the second $\frac{1}{2}(-x)^\top C(-x)$, so that they combine for a net mark-to-market financial loss of $\frac{1}{2}x^\top Cx$ (plus a change in the value function due to the new situation). Putting all the elements together, we arrive at the conjecture

$$V(x, D, f) = V(0, D - Cx, f) - \frac{1}{2}x^\top Cx. \tag{22}$$

We prove this intuitive conjecture[11] by providing a verification argument for the optimal control and value function we propose, as part of the proof of the following result.

Proposition 3.

(i) *A quadratic value function exists of the form* (A.42) *in the appendix. The optimal portfolio is given by*

$$x_t = \bar{M}_f(f_t) - \bar{M}_D(D_{t-} - Cx_{t-}), \tag{23}$$

where the matrices $\bar{M}_f$ and $\bar{M}_D$ are given in the appendix in terms of solutions to algebraic Riccati equations or an appropriate ODE.

(ii) *It holds that*

$$\bar{M}_f(f) = \hat{N}_0 Markowitz_0 + \hat{N}_2 \int_0^\infty e^{-\hat{N}_1 t}\hat{N}_3 \,\mathrm{E}\left[Markowitz_t \,|\, f_0 = f\right] dt \tag{24}$$

for appropriate matrices $\hat{N}_i$ given in the appendix.

We see that the optimal strategy is qualitatively different from the strategies that we derived above. Indeed, with purely persistent costs, the optimal strategy is no longer to trade toward an aim, but, rather, to choose a portfolio directly based on the current signals. Further, while the optimal portfolio continues to depend on the current and future expected Markowitz portfolios,

[10] The assumption that the trade is made at the average of the pre- and post-trade prices is seen in the last term of (33) below. One could alternatively assume that the entire new trade is executed at the post-trade price, thus eliminating this term. However, such an assumption would imply that the objective function has no solution since a trader would prefer trades that are arbitrarily fast, but continuous and of bounded variation, rather than the solution we derive. These strategies would be arbitrarily close to the optimal strategy that we derive. Other alternative assumptions suffer from similar issues. Under our assumption, a concavity result similar to Lemma 1 holds.

[11] More generally, for an arbitrary trade Δx we have

$$V(x, D, f) = V(x + \Delta x, D + C\Delta x, f) + x^\top C\Delta x + \frac{1}{2}\Delta x^\top C\Delta x.$$

Direct computation readily confirms that, according to this conjecture, the effect of trading Δx and then $\Delta x'$ at the same time t is the same as that of trading $\Delta x + \Delta x'$.

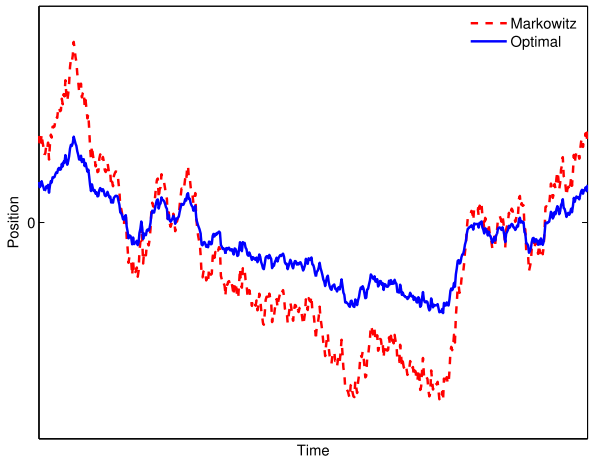

Fig. 3. **Optimal portfolio with one asset and purely persistent.** The dashed line shows the evolution over time of the Markowitz portfolio for a simulated path of the model, that is, the optimal portfolio in the absence of transaction costs. The solid line shows the optimal portfolio with purely persistent transaction costs. The optimal position is smaller in magnitude to reduce transaction costs, but it is not smooth as in Fig. 1 due to the zero transitory transaction costs.

the current one now has a distinct impact as seen in Equation (24). These qualitative differences between the solutions to Proposition 3, respectively Propositions 1 and 2, are immediately apparent in continuous time, but harder to detect in discrete time, where they are given by the same functional form, as seen in Proposition 5 (below).

The reason for the qualitative difference is the cost of buying and immediately selling. With transitory costs, such immediate round-trip trades are costly, but with purely persistent costs, they are not — transaction costs arise from buying now and selling only later when the price pressure has diminished. As a result, with transitory costs, the trader optimally chooses a portfolio strategy that is smooth over time to limit turnover. With purely persistent costs, the trader can afford quick moves in and out of the market, but limits his typical portfolio size to limit persistent impact costs.

Example. Fig. 3 illustrates this result graphically. The parameters used to make the figure are $\Phi = 0.5$, $\Sigma = 1$, $\mathrm{Var}(df_t)/dt = 1$, $\gamma = 0.5$, $\rho = 0.05$, $\lambda = 0.05$, $B = 1$, $x_0 = 0$, $f_0 = 1$, $D_0 = 0$, $R = 0.2$, $C = 2$, $r^f = 0$.

We see that the optimal portfolio varies significantly even over short time periods, that is, it has positive quadratic variation. The optimal portfolio follows the Markowitz portfolio, but moderates the position to economize on persistent transaction costs.

2. Foundation for continuous trading

We next turn to the discrete-time foundation for our continuous-time model. Considering the discrete-time foundation is important for two reasons. First, in reality, most traders update their orders at discrete times, but the trading frequency has increased over time. A decade or two ago, most traders updated their positions only monthly or annually, but gradually institutional investors started updating their positions daily, then several times intraday, and by now many investors trade throughout the day. We show how increasingly frequent trading converges to continuous trading.

Second, we need a mathematical foundation for the continuous-time model of Section 1. As noted in that section, we need to resolve a number of issues such as: What is the trading cost of trading smoothly vs. with quadratic variation? What is cost of changing the position via a "jump"? How should the mark-to-market profit/loss be handled when trading has persistent market impact? Said differently, what is the correct specification of the objective function in continuous time? These issues cannot be resolved be starting directly in continuous time, but must be understood by starting in discrete time and then taking the limit.

To accomplish these objectives, Section 2.1 first lays out the discrete-time model for any horizon Δt. Then we show in Section 2.2 how the objective function converges as $\Delta t \to 0$, thus providing a foundation for the continuous-time model studied above. Lastly, Section 2.3 provides the optimal discrete-time trading strategy and shows how it evolves as the trading frequency increases.

2.1. Model of trading in discrete time

We start by presenting a model of discrete-time trading. Securities are now traded at dates indexed by $n \in \{0, 1, 2, ...\}$, corresponding to calendar times $0, \Delta t, 2\Delta t, \ldots$, where Δt is the length of the time periods. We will actually abuse notation somewhat by indexing the same variable in two different ways: when we use the letter n in the index, then we are referring to the counting index giving the number of periods of length Δt — thus calendar time $n\Delta t$ — while when we use the letters t or s we are referring to calendar time.

The S securities' price changes between times n and $n+1$ in excess of the risk-free return, $p_{n+1} - e^{r^f \Delta t} p_n$, are collected in an $S \times 1$ vector r_{n+1}. We could let p_n be given by the continuous-time model, sampled on the time grid $0, \Delta t$, etc. For simplicity of exposition, though, we drop terms of higher order in Δt from the dynamics of the random variables. Thus, excess returns, which are predicted by the factors f_n, evolve according to

$$r_{n+1} = B f_n \Delta t + u_{n+1}, \tag{25}$$

where u_{t+1} is the unpredictable zero-mean noise term with variance $\mathrm{var}_n(u_{n+1}) = \Sigma \Delta t$. The return-predicting factor f_n is known to the investor at date n and it evolves according to

$$\Delta f_{n+1} = -\Phi f_n \Delta t + \varepsilon_{n+1}, \tag{26}$$

where $\Delta f_{n+1} = f_{n+1} - f_n$ is the change in the factors, Φ is the matrix of mean-reversion coefficients, and ε_{n+1} is the factor shock with variance $\mathrm{var}_t(\varepsilon_{n+1}) = \Omega \Delta t$. (We note that we have imposed Assumption A1 to simplify the dynamics of f, but this is just for ease of exposition as our results extend more generally.)

An investor in the economy faces transaction costs. The transitory transaction cost (TC) associated with trading $\Delta x_n = x_n - x_{n-1}$ shares is given by

$$TC(\Delta x_n) = \frac{1}{2} \Delta x_n^\top \Lambda(\Delta t) \Delta x_n, \tag{27}$$

where $\Lambda(\Delta t)$ is the matrix of transitory market impact costs. The literature does not offer guidance for how $\Lambda(\Delta t)$ depends on Δt. To address this issue, Section 3.2 provides this dependence of transaction costs on Δt in a model of endogenous dealer behavior, but for now we consider a general $\Lambda(\Delta t)$ function.

To handle persistent transaction costs, we proceed as follows. The "reference price" p_n is distorted by a persistent market impact, giving rise to an observed price

$$\bar{p}_n = p_n + D_n. \tag{28}$$

Hence, the price $\bar{p}_t$ is the sum of the price p_t without the persistent effect of the investor's own trading (as before) and the new "distortion" term D_t, which captures the accumulated price distortion due to the investor's (previous) trades.

As a consequence, the investor incurs the cost associated with the persistent price distortion D_t in addition to the temporary trading cost TC discussed above. Trading an amount Δx_t pushes prices by $C\Delta x_t$ such that the price distortion becomes $D_t + C\Delta x_t$, where $C(\Delta t)$ is Kyle's lambda for persistent price moves. Further, the price distortion mean reverts at a speed (or "resiliency") $R(\Delta t)$. Section 3.2 studies how persistent price impact arises in an economic model of inventory risk, showing that, to the leading term, C does not depend on Δt and $R(\Delta t) = R\Delta t$, for constants C and R (with the usual abuse of notation). Given the persistent price impact and resilience, the price distortion at the following date $(n+1)$ is

$$D_{n+1} = (I - R\Delta t)\,(D_n + C\,\Delta x_n)\,. \tag{29}$$

The investor's objective is derived as follows. We start by noting that, within time t, the investor realizes a mark-to-market profit on the beginning-of-period position x_{n-1} of

$$\Delta x_{n-1}^\top C\Delta x_n. \tag{30}$$

The newly-purchased shares trade at the average price $D_n + \frac{1}{2}C\Delta x_n$, so they also experience a mark-to-market gain at the end of the period, namely

$$\frac{1}{2}\Delta x_n^\top C\Delta x_n. \tag{31}$$

Even though the new shares are purchased at the average price and are marked at the post-trade price, the persistent impact ends up being a cost because the price is expected to mean-revert toward the pre-trade price. However, if shares are bought and immediately sold (before this mean reversion happens), then the persistent impact cost is zero (because both the buys and sells happen at the average price) — the cost would be only the transitory impact cost captured by Λ. Hence, the assumption of execution at the average price provides a nice way to separate transitory and persistent costs.

Finally, between n and $n+1$, the entire position x_n experiences an expected gain per share

$$\begin{aligned} E_n[\bar{p}_{n+1}] - (1 + r^f\Delta t)(\bar{p}_n + C\Delta x_n) &= Bf_n\Delta t + E_n[D_{n+1}] - (1 + r^f\Delta t)\,(D_n + C\Delta x_n) \\ &= Bf_n\Delta t - \left(R + r^f\right)(D_n + C\Delta x_n)\,\Delta t. \end{aligned} \tag{32}$$

Putting all these pieces together, we obtain the investor's objective function. Specifically, the investor seeks to choose the dynamic trading strategy $(x_0, x_1, ...)$ to maximize the present value of all future expected excess returns, penalized for risks and trading costs:

$$\begin{aligned} E_0\Bigg[\sum_n (1-\rho\Delta t)^{t+1} \left(x_n^\top \left[Bf_n - \left(R + r^f\right)(D_n + C\Delta x_n) \right] \Delta t - \frac{\gamma}{2} x_n^\top \Sigma x_n \Delta t \right) \\ + (1-\rho\Delta t)^t \left(-\frac{1}{2}\Delta x_n^\top \Lambda \Delta x_n + x_{n-1}^\top C\Delta x_n + \frac{1}{2}\Delta x_n^\top C\Delta x_n \right) \Bigg], \end{aligned} \tag{33}$$

where the discount rate is $\rho\Delta t$ with $\rho \in (0,1)$, and γ is the risk-aversion coefficient (which naturally does not depend on Δt).

We note that, in discrete time, there is no need to distinguish the case of pure persistent costs — it provides a qualitatively different solution only in continuous time.

2.2. Convergence of objective: understanding continuous markets

We now consider what happens to the objective function as the time horizon Δt approaches zero or, equivalently, as the trading frequency rises. The answer depends on the nature of the transitory transaction costs, $\Lambda(\Delta t)$:

Proposition 4.

(i) *If $\Lambda(\Delta t) \to \infty$ as $\Delta t \to 0$ and $\Lambda(\Delta t)\Delta t$ has a finite limit, which we also denote by Λ, then the objective function (33) converges to the continuous-time objective (16) with transitory-cost parameter Λ and persistent transaction costs C. Specifically, for any continuous-time strategy x_t and the discretely sampled counterparts $x_t^{(\Delta t)}$, the objective (33) tends to (16) for any strategy x_t satisfying $dx_t = \tau_t dt$, and for all other strategies the limit objective equals negative infinity.*

(ii) *If $\Lambda(\Delta t) \to 0$ as $\Delta t \to 0$ then, for any continuous-time strategy x_t, the objective (33) evaluated at the discretely-sampled $x_t^{(\Delta t)}$ tends to the continuous-time objective (21) with purely persistent costs.*

(iii) *If $\Lambda(\Delta t) \to \Lambda$ for a constant $\Lambda \neq 0$, then the conclusion of part (ii) holds, except that the objective is augmented with the term $-\frac{1}{2} E_t \int_t^\infty e^{-\rho(s-t)} d\,[x, \Lambda x]_s$.*

Parts (i) and (ii) of the proposition establish that, for small Δt, the discrete-time model is fundamentally the same as one of the two continuous-time models introduced in Section 1.2, respectively Section 1.3. This result provides a foundation for the continuous-time model, which is not self-evident given the intricacies of handling transaction costs in continuous time.

The micro-founding model of Section 3.2 yields as natural outcomes $\Lambda(\Delta t) \cong \frac{\Lambda}{\Delta t}$, covered by part (i) of the proposition, and $\Lambda(\Delta t) \cong \Lambda \Delta t$, covered by part (ii). We concentrate on these cases for the rest of the analysis, in particular the convergence of the optimal trading strategies given in Proposition 6.

For the sake of completeness, though, let us make a brief remark about case (iii) in the proposition. Under the conditions of this case, the trader faces no transitory costs in the continuous-time limit, but quadratic-variation trades are costly. As a consequence, the trader prefers to trade with arbitrarily high intensity, but zero quadratic variation; in doing so, her utility approaches the one achieved in case (ii) — pure persistent costs — but can only come arbitrarily close, rather than equal it. Strictly speaking, therefore, the trader's problem does not have a solution in this case. That said, a "smooth version" of the solution in case (ii) also provides an approximate solution in case (iii). Hence, our continuous-time solutions in Sections 1.2–1.3 provide the solutions for the two cases where solutions exist and an approximate solution when none exists.[12]

2.3. Optimal trading at higher and higher frequency

To solve the optimal trading strategy in discrete time, we consider the value function V, which is quadratic in the state variable $(x_{t-1}, y_t) \equiv (x_{t-1}, f_t, D_t)$:

[12] Similarly, if we are in case (i) and $\Lambda(\Delta t)\Delta t$ tends to zero (rather than having a limit greater than zero), then there is also no solution since quadratic variation trades are infinitely costly, but can be approached by bounded variation strategies. (As mentioned before, however, this case is not what our microfoundation generates.)

$$V(x,y) = -\frac{1}{2}x^\top A_{xx}x + x^\top A_{xy}y + \frac{1}{2}y^\top A_{yy}y + A_0.$$

Based on quadratic programming methods, we see that there exists a unique solution to the Bellman equation and the following proposition characterizes the optimal portfolio strategy.

Proposition 5. *The optimal portfolio x_t is*

$$\Delta x_t = M^{rate}(\Delta t)\left(M^{aim}(\Delta t)y_t - x_{t-1}\right), \tag{34}$$

which tracks an aim portfolio, $M^{aim}(\Delta t)y_t$, that depends on the return-predicting factors and the price distortion, $y_t = (f_t, D_t)$. The coefficient matrices $M^{rate}(\Delta t)$ and $M^{aim}(\Delta t)$, which depend on the length Δt of the time periods, are stated in the appendix.

We see that the optimal trading strategy bears close resemblance to the optimal continuous-time trading strategy with transitory costs in Proposition 2, but no obvious connection to the optimal strategy with purely persistent costs in Proposition 3. Nevertheless, the optimal strategy converges to either one, depending on the limiting properties of transitory trading costs:

Proposition 6.

(i) *If $\Lambda(\Delta t)\Delta t \to \Lambda > 0$ as $\Delta t \to 0$, then the optimal discrete-time trading strategy from Proposition 5 tends to the continuous-time solution from Proposition 2. In particular, the continuous-time matrix coefficients $\bar{M}^{rate}$ and $\bar{M}^{aim} \equiv \left(\bar{M}_f^{aim}, \bar{M}_D^{aim}\right)$ are the limits of the discrete-time coefficients M^{rate} and M^{aim} as follows:*

$$\lim_{\Delta t\to 0} \frac{M^{rate}(\Delta t)}{\Delta t} = \bar{M}^{rate} \tag{35}$$

$$\lim_{\Delta t\to 0} M^{aim}(\Delta t) = \bar{M}^{aim}. \tag{36}$$

(ii) *If $\Lambda(\Delta t)$ goes to zero with Δt, then, for the optimal discrete-time trading strategy converges to the continuous-time strategy described in Proposition 3.*

A key observation is that the limit model, and consequently the qualitative properties of the optimal strategy, are different depending on the behavior of the function $\Lambda(\Delta t)$ as Δt vanishes, and thus on the nature of the liquidity provision by the intermediaries in discussed in Section 3. Specifically, in case (i), transitory transaction costs persist at high frequencies, leading to smooth optimal trading obtains in the limit. In case (ii) on the other hand, there are no transitory costs in the limit and the optimal trading eventually becomes very "jagged" with non-zero quadratic variation.

Example. This convergence result is illustrated in Fig. 4. The figure plots the optimal position in discrete time when $\Delta t = 1$ and $\Delta t = 0.25$ and in continuous time.

The continuous-time parameters are $\Phi = 0.5$, $\Sigma = 1$, $f_0 = 1$, $\mathrm{Var}(df_t)/dt = 1$, $\gamma = 0.5$, $\rho = 0.05$, $x_0 = 0$, $\lambda = 0.05$, $B = 1$, and the discrete-time parameters are chosen consistently under case (i) in Proposition 6. Also, the outcome of the random shocks are chosen consistently in the sense that the discrete-time models use the discretized versions of the shocks to the return-predicting signals f_t.

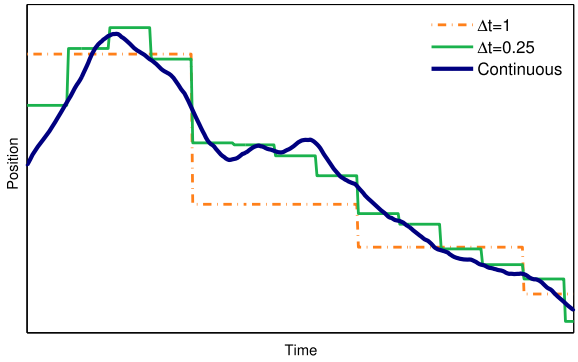

Fig. 4. **Convergence of discrete time to continuous time.** The dash-dotted step-function shows the optimal portfolio over time when the trader can trade once per time unit, say once per month. The solid step-function shows the optimal portfolio when the trader can trade 4 times per month. Naturally, the portfolio has a similar shape, but now the portfolio has shorter periods with a constant portfolio, that is, the flat line segments are 4 times shorter. Lastly, the smooth curve shows the optimal portfolio with continuous trading. We see the portfolios with discrete trading get closer and closer to the one with continuous trading as the trading frequency rises.

Fig. 4 illustrates how discrete-time trading corresponds to a step-function for the portfolio. As the trading frequency increases, the step function becomes smoother and, in the limit, converges to the continuous-time solution as shown.

3. High-frequency transaction costs

We have seen that, to understand the nature of continuous trading, we need to be able to evaluate transaction costs of high-frequency trading strategies. Said differently, we need to know how each model parameter depends on Δt. For the statistical-distribution parameters, the dependence on Δt is standard from the literature on discretizing continuous-time models, of course. For instance, the variance of a shock is linear in Δt, and so on. The one set of parameters where the literature does *not* offer guidance as to their dependence on Δt concerns the transaction cost.

We address these issues by considering the economics of transaction costs in two ways. First, Section 3.1 describes how real-world markets work via some actual empirical examples, and discusses how our model seeks to describe the real world in a stylized way.

Second, Section 3.2 lays out a stylized inventory-based model to understand the economics of how transaction costs depend on time frequencies. This micro foundation is based on a standard model of market makers in the spirit of the large literature that follows Grossman and Miller (1988). The central ingredients in this framework are as follows. First, market makers are competitive. In the real world, market making is indeed a competitive business. Numerous firms compete to profit from providing liquidity, including a number of investment banks, market making firms like Getco LLC and Citadel, high-frequency traders, hedge funds, and many others. A second assumption is that market makers intermediate between end users who are not always available in the market. Again, it is a standard assumption that markets are "thin" in the sense that not all buyers and sellers arrive to the market at the same time, which provides a role for intermediation. A third important assumption is that market makers are risk averse and therefore focus on their inventory. This behavior is well documented in the literature, and we follow the inventory-based models by abstracting from adverse-selection issues. Regarding the specific form of risk aversion, we assume that market makers have mean-variance preferences. This assumption simplifies

the analysis, but is not essential, as we are concerned with the dependence of transaction costs on Δt and our results continue to hold, for instance, under standard constant-absolute-risk-aversion preferences paired with normal shocks. In summary, our micro foundation of transaction costs is based on realistic assumptions that are standard in the literature. What we add is to derive its implications for the dependence of transaction costs on the length of the trading intervals.

3.1. Real-world transaction costs and trading

To relate our model more directly to real-world trading, we consider here the example of trading in an electronic limit order book. This form of trading dominates most modern markets such as those for individual equities, equity index futures, currencies, commodity futures, bond futures, and increasingly other markets as well. In such markets, market participants can post buy and sell orders that are listed in the limit order book, which is simply the list of orders that are yet to be executed. Orders to buy at a given price are called bids and orders to sell at a given price are called asks (or offers).

The top left panel of Fig. 5 shows an example of a real-world limit order book. In particular, the figure shows the 20 best "bids" (the bars pointing left) and the 20 best "asks" (the bars pointing right) for an arbitrary liquid stock at an arbitrary time, namely IBM on 1/28/2015 at 09:55 AM (a large stock listed on New York Stock Exchange). A market participant arriving at the market at this time has several choices: He can "hit" some of the displayed orders or post his own limit order. Starting with the first option, suppose that a trader wants to buy immediately. Then he can hit the ask with the lowest price, trading up to 200 shares (the width of the lowest bar going to the right) at the price of 154.09. If he wants to buy more shares than that, he can buy another 400 shares at 154.10 (the second-best ask), and so on.

The top right panel of Fig. 5 shows the marginal prices that the trader must pay as a function of the order size. The step function to the right of zero shows the increasing prices for larger and larger buy orders, starting at 154.09, then 154.10 as described above. If the trader wants to sell, then we label his orders as having a negative order size (following standard conventions), giving rise to the step function to the left of zero. In this case, he must hit the bids at lower and lower prices as his order gets larger in magnitude (i.e., more and more negative).

The average price that the trader must pay is naturally the weighted average of the marginal prices, displayed in the bottom left panel of Fig. 5. The average price is relatively smooth, in contrast to the step function seen in the marginal price. Also, the average price is close to linear, except that there is a discontinuity around zero (no trade) due to the so-called "bid–ask spread." The bid–ask spread is the distance between the best bid at 154.02 and the best ask at 154.09, in this case of 7 cents. The bid–ask spread measures the cost of buying and immediately selling a small number of shares, say 1 share. The "half-spread" (i.e., the bid–ask spread divided by two) can be considered the cost of simply buying 1 share as it is the difference between the ask price and the "mid quote," which is the average of the best bid and the best ask, here at 154.055.

Lastly, the bottom right panel of Fig. 5 shows the total transaction cost as a function of the order size. The total transaction cost is the order size multiplied by the difference between the average execution price and the mid quote. Given an order size of 1, the total transaction cost is the half-spread, but the transaction cost increases more than proportionally with the order size since larger orders move the price, giving rise to the approximately quadratic curve segments.

Recall that so far we have only discussed the trader's ability to hit existing orders (using so-called "market orders" or "marketable limit orders"). Hence, the depicted measure of total

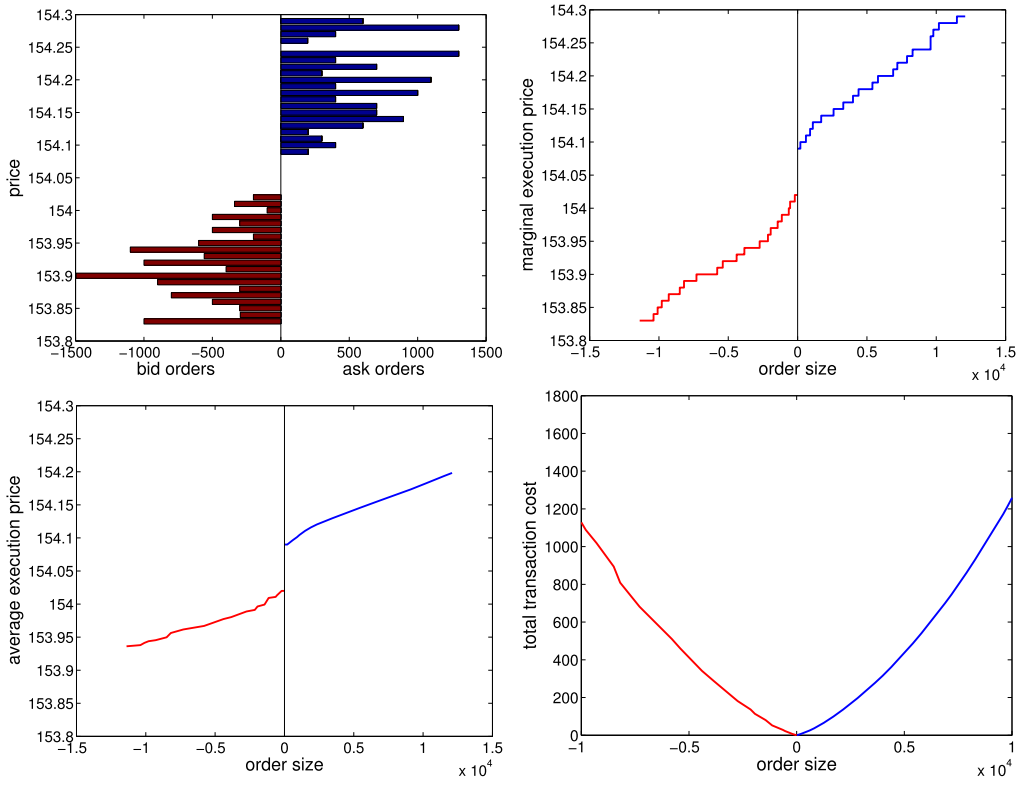

Fig. 5. **Transaction costs in a real-world limit-order book: IBM.** The top left panel shows the top of the limit order book for IBM stock on 1/28/2015 at 09:55 AM. The top right plot shows the corresponding marginal execution price for a trader who is "walking the book" as a function of the market order size. The bottom left panel shows the corresponding average execution price as a function of order size. Lastly, the bottom right panel shows the total transaction cost as a function of order size, calculated as the order size multiplied by the difference between the average execution price and the mid quote.

transaction costs refer to immediate execution. However, another possibility is that the trader posts a non-marketable limit order, that is, a limit order that will not be immediately executed because it does not hit an existing order. For instance, the trader could post a limit order to buy at 154.06, which would become the new best ask. In this case, the trader hopes that another investor will hit the order, allowing him to buy at his posted price.

Next, let's see how this real-world example corresponds to our model, which is illustrated in Fig. 6. The key element of the model is the total transaction cost, which is illustrated in the bottom right panel. Clearly, the total transaction cost is quadratic in order size Δx, that is, $TC(\Delta x) = \frac{1}{2}\Lambda(\Delta x)^2$.

To achieve this quadratic total cost, the average execution price must be linear in the order size as seen in the bottom left panel. More specifically, average price(Δx) = mid quote $+ \frac{1}{2}\Lambda\Delta x$ and this linear price impact recovers the quadratic transaction cost,

$$TC(\Delta x) = \Delta x\,(\text{average price}(\Delta x) - \text{mid quote}) = \frac{1}{2}\Lambda(\Delta x)^2.$$

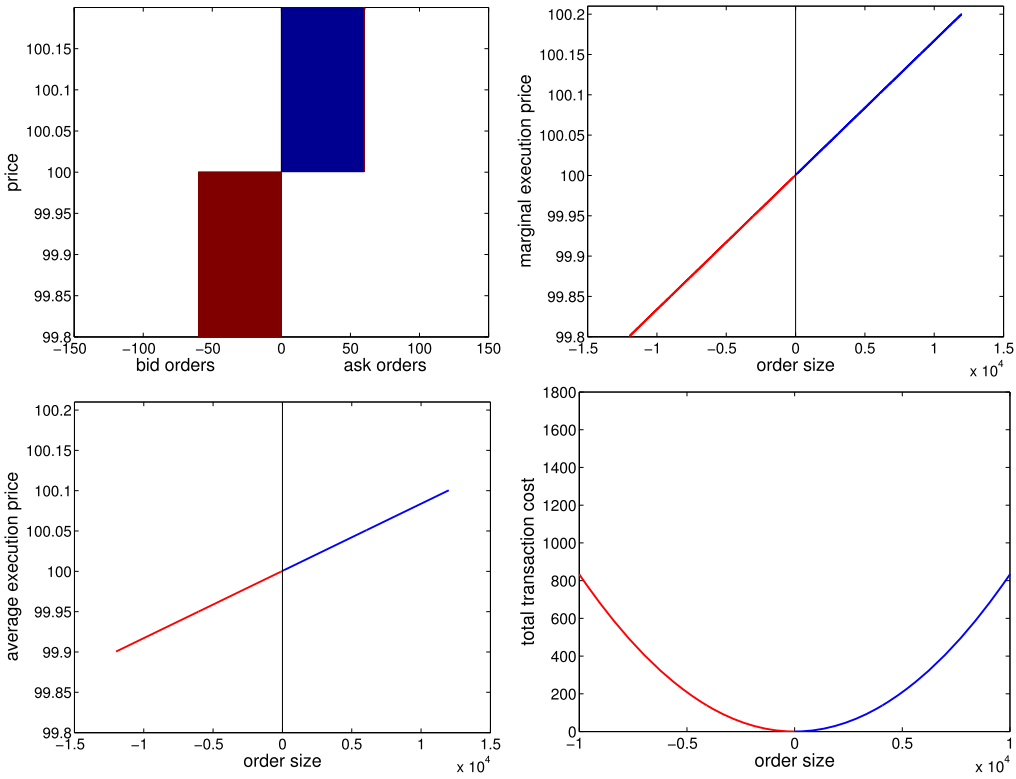

Fig. 6. **Transaction costs in our model.** The top left panel illustrates that, if our model is interpreted as a limit order book, then the limit order book has constant depth and an arbitrarily small bid–ask spread. As a result, the marginal execution price is linear in the order size (top right plot). Also, the average execution price is linear, though with half the slope (bottom left plot). The total transaction cost (bottom right panel) is quadratic in the order size as it is the order size multiplied by the difference between the average execution price and the mid quote.

This average cost structure means that the marginal execution prices must also be linear in the order size as seen in the top right panel of Fig. 6. The slope of the marginal price function must be double that of the average price, namely Λ.

Lastly, to achieve this marginal price schedule in a limit order book, the top left panel shows what the limit order book could look like in our model. We see that the envisioned limit order book has constant depth and consists of many little limit orders at each price point (more precisely, we consider the whole continuum of prices).

The real-world depicted in Fig. 5 and the model depicted in Fig. 6 share many similarities, but there are also differences. First, the real-world limit order book is more "noisy" and the marginal price is a step function. In contrast, the model has a limit order book of constant depth with finely sliced limit orders that lead to marginal prices that form a straight line. These differences are not important, however, as ultimately the average prices are close to straight lines in either case.

A potentially more important difference is that the real-world average price is discontinuous at zero due to the bid–ask spread, while the model has a zero bid–ask spread and an average price which is continuous at zero. As a result of this difference, the model-implied total transaction cost is quadratic with a slope of zero at zero, whereas the real-world total transaction cost appears well

approximated by two quadratic curves for, respectively, buy and sell orders with a kink where they meet at zero (because neither has a zero absolute slope at zero).

However, this difference between the real world and the model may not be as significant as it appears. Remember that the trader has the option to post his own limit orders. In the way that we envisioned the model-implied limit order book, the trader has no need to post non-marketable limit orders since the bid–ask spread is zero (there is no room for non-marketable orders), but real-world traders who know that they want to trade in a certain direction often find it worthwhile to post limit orders and wait to have them executed. By posting a limit order to buy below the mid quote, the trader can hope to trade at a negative cost, but execution may be delayed, never happen, or be subject to adverse selection. By posting a small limit order at the midquote, the trader is likely to have the order executed at essentially zero cost in a very liquid market. Larger orders may need to be posted at worse prices above the midquote and be broken up over time. Further, in the real world there may be hidden orders that are hit when the trader submits his order and there could be latent orders in the sense that other market participants could be waiting in the wings, ready to hit a new limit order between the best bid and best ask.

Also, the trader may be able to execute at or near the mid quote by using so-called "dark pools," crossing networks, or by accessing other sources of liquidity. Going into the details of these real-world market structures is beyond the scope of the paper, but it suffices to say that the real world may be closer to the model than what is seen from assuming that the trader must be "walking" the visible part of the limit order book, that is, mechanically hitting the visible orders, one by one.

As one way to make this point explicitly, we could envision the model-implied limit order book in a different way that would correspond more closely to the data: Suppose that the limit order book looks as in the top left panel of Fig. 6, except that there is a strictly positive bid–ask spread in that there is some space between the bid and ask orders. Suppose further that small limit orders posted inside the bid–ask spread are immediately hit by latent orders, such that the resulting marginal pricing scheme remains as in top right panel of Fig. 6. In this case, the economic situation remains exactly as in the model, but an empiricist who only observes the orders posted in the limit order book (not the latent orders), would conclude that there is a difference.

In summary, the model matches a number of features of the real world, but it is clearly not perfect. The quadratic total transaction cost makes the model highly tractable and while the notion of linear market impact may not be unrealistic, the implied vanishing cost of a very small order may not be completely realistic. Nevertheless, the model-implied optimal trading pattern captures how some institutional investors are now often trading. As the model implies, many real-world traders are trading continuously throughout the day.

Fig. 7 shows the position of a real-world asset manager over three arbitrary days in 2015, trading an arbitrary liquid futures contract in an electronic limit order book. The figure plots the minute-by-minute positions, which continuously evolve throughout each day, reflecting continuous trading. The smoothness of the curve is not due to interpolation as the data are unsmoothed and the minute-by-minute data frequency is so high that the graph looks the same whether we plot each data point as a dot or connect the dots with a line. Further, the position is measured in contracts (not dollars) so it evolves only due to trading (not due to price changes). While our model of quadratic transaction costs implies such continuous trading, models of constant or linear transaction costs imply no-trade regions and only intermittent trading. Hence, the real-world continuous trading is consistent with the qualitative insights of our model (but it is of course not a proof of the realism of the details of the model).

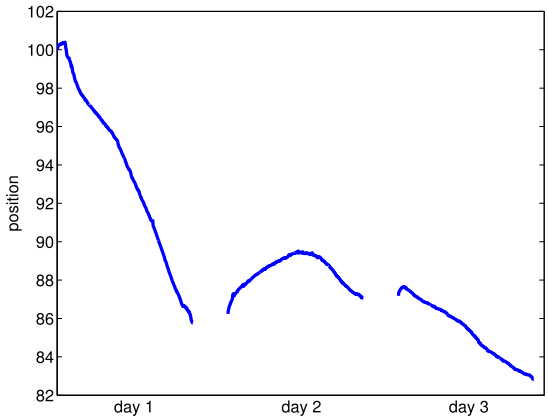

Fig. 7. **Real-world trading example.** This figure shows the position (measured in contracts, as a percentage on the initial number of contracts) of a given futures contract of an actual asset manager over the course of three days in 2015. The graph is based on minute-by-minute data so the smoothness is not due to interpolation, but, rather, an expression of continuous trading throughout the day.

3.2. Micro model of transaction costs: modeling the cost of high-frequency trading

We next consider the economic foundation for quadratic transaction costs and, importantly, how transaction costs depend the trading frequency and evolve in the limit as the trading frequency increases. The model is stylized, but captures the notion that transaction costs arise due to the risk faced by intermediaries who hold inventories as counterparties access markets at different times. Inventory risk is one of the important risks faced by intermediaries who provide liquidity for instance in a limit-order book as discussed above. (Another risk is adverse selection, from which we abstract, following Grossman and Miller, 1988 and the rest of the inventory-based intermediation literature.) The micro model is general and may also capture the economics of an over the counter market or other market structures — our focus is on how trading costs depend on trading frequencies.

Transitory transaction costs. To obtain a temporary price impact of trades endogenously, we consider an economy populated by three types of investors: (i) the trader whose optimization problem we study in the paper, referred throughout this section as "the trader," (ii) "market makers," who act as intermediaries, and (iii) "end users," on whom market makers eventually unload their positions as described below.

The temporary price impact arises as compensation for market makers' inventory risk as follows. There are a continuum of competitive market makers of mass one indexed by the set $[0, h]$, and they arrive for the first time at the market at a time equal to their index. The market operates only at discrete times Δt apart,[13] and the market makers trade at the first trading opportunity. Hence, there are $h/\Delta t$ types of market makers that trade a successive time periods.

Once they trade — say, at time t — market makers must spend h units of time gaining access to end users. At time $t + h$, therefore, they can trade with end users at an exogenous price p_{t+h}, which corresponds to the fundamental price in Section 2.1 with associated return given

[13] We make the simplifying assumption that $\frac{h}{\Delta t}$ is an integer.

by equation (25). After trading with end users, the market makers rejoin the market. It follows that at each market trading date there is always a mass $\frac{\Delta t}{h}$ of competing market makers that clear the market.

The price $\hat{p}_t$ at which the trader buys from the market makers is determined as follows. For any price $\hat{p}_t$, the market maker determines their optimal position q, and then the market-clearing $\hat{p}_t$ is the price that ensures that the total position of the $\frac{\Delta t}{h}$ market makers balances the trader's trade, $q(\hat{p}_t)\frac{\Delta t}{h} + \Delta x_t = 0$. Market makers' supply/demand schedule $q(\hat{p}_t)$ arises to maximize their quadratic utility:

$$q(\hat{p}_t) = \arg\max_q \left\{ \hat{E}_t \left[q^\top (p_{t+h} - e^{rfh}\hat{p}_t) \right] - \frac{\gamma^M}{2} Var_t \left[q^\top (p_{t+h} - e^{rfh}\hat{p}_t) \right] \right\}, \tag{37}$$

where γ^M is each market maker's risk aversion and the symbol $\hat{E}$ denotes expectations under the market makers' beliefs. For simplicity we specify the market makers' beliefs $\hat{E}$ about future trade prices p by assuming that market makers do not think that returns are predictable (that is, they may observe f, but they believe that $B = 0$). This assumption is not central; the main point is that the price impact $\hat{p}_t - p_t$ gives market makers an incentive to change their position from whatever position they would otherwise have taken — for simplicity, we essentially assume that the alternative position is zero (i.e., market makers only intermediate).

Hence, maximizing the objective (37) for a given $\hat{p}$ gives a market-maker demand of $q \cong (\gamma^M h\Sigma)^{-1}(p_t - \hat{p}_t)$, where $h\Sigma$ is the price variance over the time period (we perform the formal calculations in the appendix, but here we can provide an intuitive reasoning). Imposing the market-clearing condition $q(\hat{p}_t)\frac{\Delta t}{h} + \Delta x_t = 0$ yields the trader's execution price $\hat{p}$ and the market impact

$$\hat{p}_t - p_t \cong \frac{\gamma^M h^2 \Sigma}{\Delta t} \Delta x_t. \tag{38}$$

Hence, the total market-impact cost of the Δx_t shares is approximately $\Delta x_t^\top \frac{\gamma^M h^2 \Sigma}{\Delta t} \Delta x_t$, corresponding to a quadratic transaction cost function with parameter $\Lambda(\Delta t) \cong \frac{2\gamma^M h^2 \Sigma}{\Delta t}$.

This intuitive calculation shows how transaction costs depend on the trading intervals Δt, but, before we can state the result formally, we note that the time h market makers must hold their position could also depend on the trading frequency Δt. Two cases suggest themselves naturally: (i) a constant h, meaning that market makers need a fixed amount of time to lay off a position regardless of the trading frequency, and (ii) h decreasing in Δt because a lower Δt is thought of as an improvement in the trading technology (or in the attention of market participants) — in its simplest form, $h = \Delta t$.

Proposition 7. *Under the model described above, the trader faces quadratic transitory transaction costs* $\frac{1}{2}\Delta x_t^\top \Lambda(\Delta t)\Delta x_t$.

(i) *If dealers need a fixed amount of calendar time to lay off their inventory, i.e., h does not depend on Δt, then the transitory market-impact satisfies $\Lambda(\Delta t)\Delta t \to \Lambda$, where the limit parameter is strictly positive: $\Lambda \neq 0$.*

(ii) *If dealers can lay off their inventory during each time period, i.e., $h = \Delta t$, then the transitory market-impact parameter converges to zero: $\Lambda(\Delta t) \to 0$.*

We see that the endogenous trading costs give rise to two cases that correspond exactly to the two cases considered in Propositions 4 and 6 and we have come full circle. To summarize,

in each case (i) or (ii), transaction costs have a micro foundation, and the trader faces a linear average price $\hat{p}$ as a function of the traded position Δx_t and a quadratic transaction cost (just as in the bottom panels of Fig. 6). Further, in both cases the objective function has a continuous-time limit, the model can be solved in discrete or continuous time, and the discrete-time solutions converge to their continuous-time counterparts. Case (i) corresponds to a constant holding period for market makers, whereas case (ii) corresponds to market makers having vanishing holding periods as trading frequencies rise.

The proposition also shows how transaction costs depend on the time scale. In case (i), the transitory market-impact cost scales inversely with the trading intervals, $\Lambda(\Delta t) \cong \Lambda/\Delta t$ while, in case (ii), the impact cost is proportional to the trading interval, $\Lambda(\Delta t) \cong \Lambda \Delta t$.

While $\Lambda(\Delta t)$ varies with Δt in both cases, we note that transaction costs are "stable" in case (i) in the following sense: Say that the trader buys 10,000 shares over 5 hours at a constant number of shares per time interval. Then the transaction cost per calendar time unit — and the total transaction cost over the entire 5-hour period — is the same for all Δt (to the first order). This is a natural conclusion since the economics of the situation are the same regardless of the specific modeling choice of Δt. Indeed, the trader behaves essentially the same regardless of whether Δt is a minute or a second or a millisecond.

Persistent transaction costs. We model persistent price impact costs similarly, but with a different specification of the market makers. Consider therefore the same model as in the previous section, but suppose now that market makers do not hold their inventories for a deterministic number h of time units, but rather manage to deplete them at a constant rate ψ, through trade with end users at price p.

Since market makers are now long lived, they must have expectations over future trades $\Delta x_{t+k\Delta t}$. For simplicity, we assume that market makers cannot predict $\Delta x_{t+k\Delta t}$ (e.g., because these trades are hidden among many other trades) so that, at time t, they face the following conditional moments for $k > 0$:

$$\hat{E}_t\left[\Delta x_{t+k\Delta t}\right] = 0 \tag{39}$$

$$\hat{E}_t\left[\Delta x_{t+k\Delta t}^{\top} \Delta x_{t+k\Delta t}\right] = v. \tag{40}$$

Furthermore, for technical reasons that we discuss in detail in Remark 1 of the appendix, we assume that each unit of the asset is traded at its marginal price so that the market makers do not make any utility gains. Remark 1 makes it clear that dispensing with the assumption has a very limited impact on the results.

We use the symbol I_t to denote the market maker's inventory at any time t, after all trades at t. Between two trading dates with the trader, this inventory evolves according to

$$\Delta I_t = -\psi I_{t-\Delta t} \Delta t + q_t. \tag{41}$$

Here, the first part of the inventory change reflects that the market makers reduce their inventory through trades with end users. The second part reflects new liquidity-providing trades with the trader. In equilibrium, market makers buy what the trader sells and vice versa, $q_t = -\Delta x_t$.

The market makers continue to maximize a quadratic objective:

$$\max_{\{q_s\}_{s=t+\mathbb{N}\Delta t}} \left\{ \hat{E}_t \sum_{s=t+\mathbb{N}\Delta t} e^{-r^f(s-t)} \left(\psi I_{s-\Delta t}^{\top} \Delta t \; p_s - q_s^{\top} \bar{p}_s - \frac{\gamma^M}{2} I_{s-\Delta t}^{\top} \mathrm{Var}_{s-\Delta t}(r_s) \, I_{s-\Delta t} \right) \right\}, \tag{42}$$

which depends positively on the expected cash flows due to future trades with the end users ($\psi I_{s-\Delta t}^{\top} \Delta t\, p_s$) at the exogenous price p, similarly for trades with the trader ($-q_s^{\top} \bar{p}_s$) at the endogenous price $\bar{p}$, and negatively on the inventory risk.

We seek to determine the endogenous transaction price between the trader and market makers, $\bar{p}_t = p_t + D_t$, that clears the market. The key to the results we are pursuing is that the market makers charge the trader for the contribution of his trade to an inventory that is expected to diminish only gradually (exponentially). The gradual depletion of the inventory translates into a price distortion that also decays, at the same rate. The refreshment of the inventory due to the trader's trade Δx_t naturally gives rise to an innovation to the distortion.

We summarize the implications for the specification of the price impact faced by the trader in the following proposition, which also shows how transaction cost parameters depend on the trading interval.

Proposition 8. *Under the gradual-decay-inventory model, the trader transacts at prices given by (28) and (29), where the resiliency parameter R is of the form $R(\Delta t) = R\Delta t$, while the persistent market impact C does not depend on Δt to the leading order as $\Delta t \to 0$.*[14]

Transitory and persistent transaction costs. The two types of price impact can obtain simultaneously in this model, so that we can have both kinds of transaction costs and consider their separate or simultaneous convergence to continuous time using Propositions 7–8. As one way to see this, we can consider an economy populated by the trader and *two* kinds of intermediaries, say high-frequency traders (HFTs) and hedge funds. The trader transacts with the HFTs and, after a period of length h, the HFTs clear their inventories with the hedge funds, who specialize in holding the inventory over a longer time period and gradually trading out of it as end users arrive over time. The hedge funds deplete their inventories only gradually (at a constant rate, as above), giving rise to a persistent impact. The trader must compensate both groups of market makers for the risk taken, resulting in the two price-impact components. Indeed, this hot-potato framework for market making, as it is sometimes called, under our assumptions gives rise to exactly the model with transitory and persistent price impact that we use above. In the interest of space, we describe the environment precisely and prove the formal result in the appendix.

4. Equilibrium implications

In this section, we study how transaction costs and supply/demand pressure can affect security prices and returns in equilibrium. More specifically, we consider a situation in which an investor facing transaction costs absorbs the residual supply from a group of "noise traders" and analyze the relationship implied between the characteristics of the supply dynamics and the excess return.

We consider a model set in continuous time, as detailed in Section 1.1, featuring one security, but now we wish to determine the price p endogenously. There exist $L \geq 1$ groups of noise traders, where group l holds the (exogenously given) portfolio z_t^l. As in the literature following Kyle (1985), these noise traders are interpreted as investors with various hedging demands or other liquidity reasons to trade. The noise traders positions are mean reverting around a stochastic mean df_t^l:

[14] To the leading order means that $R(\Delta t)/\Delta t \to R$ and $C(\Delta t) \to C$ as $\Delta t \to 0$.

$$dz_t^l = \kappa\left(f_t^l - z_t^l\right)dt \tag{43}$$

$$df_t^l = -\psi_l f_t^l dt + d\varepsilon_t^l. \tag{44}$$

The only other investors in the economy are the investors considered in Section 1.1, facing transaction costs given by $\Lambda = \lambda\Sigma =: \lambda\sigma^2$.

In this simple context, an equilibrium is defined as a price process and market-clearing asset holdings that are optimal for all agents given the price process. Since the noise traders' positions are optimal by assumption as specified by (43)–(44), the restriction imposed by equilibrium is that the dynamics of the price are such that, for all t, the trader wants to take a position

$$x_t = -z_t \tag{45}$$

which is the opposite of the aggregate noise-trader holding, $z_t \equiv \sum_l z_t^l$.

To achieve this equilibrium, the trader must have an incentive to buy what the noise traders are selling and, therefore, we conjecture that returns are driven by the vector of factors given by $f \equiv (f^1, ..., f^L, z)$. This factor f has the right structure since it contains all information about the noise traders' aggregate position and future dynamics, and it satisfies the Markovian structure (2) as in our baseline model. To solve the equilibrium, we also need to determine how this factor affects expected returns, more specifically the factor loadings B from equation (1). We find B as part of Proposition 9 below, which formally states the equilibrium.

Proposition 9. *The market is in equilibrium if $x_0 = -z_0$ and the security's expected excess return is given by*

$$\frac{1}{dt}E_t[dp_t - r^f p_t\, dt] = \sum_{l=1}^{L} \lambda\sigma^2\kappa(\psi_l + \rho + \kappa)(-f_t^l) + \sigma^2(\rho\lambda\kappa + \lambda\kappa^2 - \gamma)z_t. \tag{46}$$

The coefficients $\lambda\sigma^2\kappa(\psi_k + \rho + \kappa)$ are positive and increase in the mean-reversion parameters ψ_k and κ and in the trading costs $\lambda\sigma^2$. In other words, noise trader selling ($f_t^k < 0$) increases expected excess returns, and especially so if its mean reversion is faster and if the trading cost is larger.

Naturally, noise-trader selling increases the expected excess return, while noise-trader buying lowers it, since the arbitrageurs need to be compensated to take the other side of the trade. Interestingly, the effect is larger when trading costs are larger and for noise-trader shocks with faster mean reversion because such shocks are associated with larger trading costs for the arbitrageurs.

5. Conclusion and broader implications

This paper provides a general framework for optimal portfolio choice with frictions and multiple time-varying signals about expected returns. While the framework is very general, allowing rich dynamics for returns and signals, it is nevertheless highly tractable. Indeed, the optimal portfolio is derived as an intuitive closed-form expression.

The optimal portfolio strategy trades gradually toward an aim portfolio that depends on the current and expected future optimal portfolios in the absence of transaction costs. Hence, financial frictions imply that signals' dynamics, in particular their persistence over time, are important. Intuitively, a signal is given more weight if it is more persistent, since a longer-lasting effect should be incorporated more in light of frictions.

We show how our continuous-time model is approached by discrete-time models of vanishing time-period length if the model parameters are scaled appropriately. The key innovation

in this respect is to determine the correct time-scaling of the transaction-cost parameter. We provide an economic foundation for this time-scaling of transaction costs, and show that the convergence happens naturally in this economic setting. Further, we derive implications for equilibrium expected returns, showing why high-frequency movements in expected returns are larger than low-frequency movements, as documented empirically. Finally, as we elaborate below, the model's tractability makes it a powerful tool with many potential applications in other areas of economics and, indeed, even more broadly.

General dynamic models. Before outlining a few specific applications, we note that many dynamic models in the social sciences are special cases of the linear–quadratic framework or can be approximated well by this framework as discussed by Hansen and Sargent (2014) and references therein. By incorporating frictions and multiple signals with varying mean-reversion rates into the linear–quadratic framework, our model shows, at a high level, how the optimal policy moves gradually toward an aim that overweights persistent information. Our solution is explicit and makes it particularly easy to link the crucial input parameters, such as persistences and adjustment costs, to the output. Further, our model shows how the answer is robust to the frequency of policy changes when the policy parameters are scaled appropriately depending on the time horizon. Indeed, in each of the settings described below, we think that the case where transitory transaction costs survive in the limit is most natural (unless technology improves such that adjustment costs vanish).

Macroeconomics. Many macroeconomic models rely on the linear–quadratic framework (see, e.g., Ljungqvist and Sargent, 2004). As an illustration, consider an economy with different signals about total factor productivity (TFP) and capital adjustment costs. In this case, our model can be applied to show how to gradually adjust the capital stock towards an aim that overweights persistent signals about TFP shocks.

Monetary policy. The linear–quadratic framework has also been employed extensively in models of monetary economics (Benigno and Woodford, 2003). Our model can be recast as describing a central bank receiving multiple signals about inflation pressures, e.g., across regions, and facing adjustment costs (capturing what is often termed "policy inertia"). In this case, our results mean that monetary policy should move gradually towards an aim that optimally weights the different signals. This can help explain why central banks focus on core inflation rather than headline inflation, which includes such transitory shock as oil price changes. As another example, a highly persistent signal of deflationary pressures in southern Europe should be weighted more heavily than a transitory signal of inflation in (an equal-sized region of) Germany.

Political economy. As another potential application, the model could describe a political party receiving different signals from various constituents. In this case, our model's insight shows how the party should move its politics gradually toward an aim that optimally incorporates all signals, giving more weight to persistent political trends and less to shorter-lived fads. Let us sketch how to capture this in our model, as this framework may be less standard in political economics. A political party must choose its views x_t on each of several issues, e.g., x_t^1 is the view on economic policy, x_t^2 is the view on social issues, and so on. The party receives signals f_t about the views of different constituents, which can be aggregated to a vector of average views about all the issues, Gf_t. The policy maker faces quadratic costs of deviating from the current average view:

$$(x_t - Gf_t)'\Sigma(x_t - Gf_t) = -x_t'Bf_t + x_t'\Sigma x_t + f_t'G'\Sigma Gf_t,$$

where $B = -2\Sigma G$. The first two terms correspond to our objective function and the last can be ignored as it is independent of the choice x_t. Further, the party faces quadratic costs of changing its views (the cost of "flip-flopping"), making the model a special case of our framework.

Microeconomic model of product design. Consider a monopolistic firm, which must choose the design x_t of its product, where x_t^1 could be the color, x_t^2 the marketing expense, etc. Customers' preferences for different products change over time such that the firm faces the following demand curve:

$$Demand(price; x_t, f_t) = x_t'Hf_t \cdot price^{-s}.$$

Here, $s > 1$ is the price elasticity, H is a matrix with positive elements, and f_t is a positive process capturing how consumers value each product attribute. Renting a machine that can produce the good with design x_t costs $1/2\, x_t'\Sigma x_t$ and the marginal production cost is c. Given a product design, the profit is derived from the optimal price setting:

$$x_t'Bf_t := x_t'Hf_t \cdot \max_{price} price^{-s} \cdot (price - c).$$

With a quadratic cost of changing the product design, we see that this model is a special case of our general framework. Hence, our results show that the product design should be adjusted towards a combination of the signals of consumer tastes that gives higher weight to the more persistent trends.

In summary, the model presents a highly tractable framework that gives rise to several insights concerning optimal trading in financial markets, and it can be applied to other dynamic problems featuring frictions and signals of varying persistence.

Appendix A. Proofs

The proofs of the results stated in this article can be found online at http://dx.doi.org/10.1016/j.jet.2016.06.001.

References

Acharya, V., Pedersen, L.H., 2005. Asset pricing with liquidity risk. J. Financ. Econ. 77, 375–410.
Almgren, R., Chriss, N., 2000. Optimal execution of portfolio transactions. J. Risk 3, 5–39.
Amihud, Y., Mendelson, H., 1986. Asset pricing and the bid–ask spread. J. Financ. Econ. 17, 223–249.
Benigno, P., Woodford, M., 2003. Optimal monetary and fiscal policy: a linear–quadratic approach. NBER Macroecon. Annu. 18, 271–364.
Bertsimas, D., Lo, A.W., 1998. Optimal control of execution costs. J. Financ. Mark. 1, 1–50.
Breen, W.J., Hodrick, L.S., Korajczyk, R.A., 2002. Predicting equity liquidity. Manag. Sci. 48, 470–483.
Campbell, J.Y., Viceira, L.M., 2002. Strategic Asset Allocation Portfolio Choice for Long-Term Investors. Oxford University Press, Oxford, UK.
Carlin, B.I., Lobo, M., Viswanathan, S., 2008. Episodic liquidity crises: cooperative and predatory trading. J. Finance 62, 2235–2274.
Cetin, U., Jarrow, R., Protter, P., Warachka, M., 2006. Pricing options in an extended Black Scholes economy with illiquidity: theory and empirical evidence. Rev. Financ. Stud. 19, 493–529.
Chacko, G.C., Jurek, J.W., Stafford, E., 2008. The price of immediacy. J. Finance 63, 1253–1290.
Cochrane, J.H., 2011. Presidential address: discount rates. J. Finance 66, 1047–1108.
Collin-Dufresne, P., Daniel, K., Moallemi, C., Saglam, M., 2014. Strategic asset allocation in the presence of transaction costs. Working paper. Columbia University.

Constantinides, G.M., 1986. Capital market equilibrium with transaction costs. J. Polit. Econ. 94, 842–862.
Davis, M., Norman, A., 1990. Portfolio selection with transaction costs. Math. Oper. Res. 15, 676–713.
Engle, R., Ferstenberg, R., 2007. Execution risk. J. Portf. Manag. 33, 34–45.
Engle, R., Ferstenberg, R., Russell, J., 2008. Measuring and modeling execution cost and risk. Working paper. University of Chicago.
Gârleanu, N., 2009. Portfolio choice and pricing in imperfect markets. J. Econ. Theory 144, 532–564.
Gârleanu, N., Pedersen, L.H., 2013. Dynamic trading with predictable returns and transaction costs. J. Finance 68, 2309–2340.
Gârleanu, N., Pedersen, L.H., Poteshman, A., 2009. Demand-based option pricing. Rev. Financ. Stud. 22, 4259–4299.
Gatheral, J., Schied, A., 2011. Optimal trade execution under geometric Brownian motion in the Almgren and Chriss framework. Int. J. Theor. Appl. Finance 14, 353–368.
Glasserman, P., Xu, X., 2013. Robust portfolio control with stochastic factor dynamics. Oper. Res., 1–20.
Greenwood, R., 2005. Short and long term demand curves for stocks: theory and evidence. J. Financ. Econ. 75, 607–650.
Grinold, R., 2006. A dynamic model of portfolio management. J. Invest. Manag. 4, 5–22.
Grossman, S., Miller, M., 1988. Liquidity and market structure. J. Finance 43, 617–633.
Hansen, L.P., Sargent, T., 2014. Recursive Models of Dynamic Linear Economies. Princeton University Press, Princeton, NJ.
Heaton, J., Lucas, D., 1996. Evaluating the effects of incomplete markets on risk sharing and asset pricing. J. Polit. Econ. 104, 443–487.
Huberman, G., Stanzl, W., 2004. Price manipulation and quasi-arbitrage. Econometrica 74, 1247–1276.
Jang, B.-G., Koo, H.K., Liu, H., Loewenstein, M., 2007. Liquidity premia and transaction costs. J. Finance 62, 2329–2366.
Kyle, A.S., 1985. Continuous auctions and insider trading. Econometrica 6, 1315–1335.
Kyle, A.S., Obizhaeva, A., 2011. Market microstructure invariants: theory and implications of calibration. Working paper. University of Maryland.
Kyle, A.S., Obizhaeva, A.A., 2012. Large bets and stock market crashes. Working paper. University of Maryland.
Lagos, R., 2010. Asset prices and liquidity in an exchange economy. J. Monet. Econ. 57, 913–930.
Lehman, B.N., 1990. Fads, martingales, and market efficiency. Q. J. Econ. 105, 1–28.
Lillo, F., Farmer, J.D., Mantegna, R.N., 2003. Master curve for price-impact function. Nature 421, 129–130.
Liu, H., 2004. Optimal consumption and investment with transaction costs and multiple assets. J. Finance 59, 289–338.
Ljungqvist, L., Sargent, T., 2004. Recursive Macroeconomic Theory, 2nd edition. MIT Press, Cambridge, MA.
Lo, A., MacKinlay, A., 1990. When are contrarian profits due to stock market overreaction? Rev. Financ. Stud. 3, 175–205.
Lo, A., Mamaysky, H., Wang, J., 2004. Asset prices and trading volume under fixed transaction costs. J. Polit. Econ. 112, 1054–1090.
Lynch, A., Tan, S., 2011. Explaining the magnitude of liquidity premia: the roles of return predictability, wealth shocks, and state-dependent transaction costs. J. Finance 66, 1329–1368.
Markowitz, H.M., 1952. Portfolio selection. J. Finance 7, 77–91.
Nagel, S., 2012. Evaporating liquidity. Rev. Financ. Stud. 25, 2005–2039.
Obizhaeva, A., Wang, J., 2006. Optimal trading strategy and supply/demand dynamics. Working paper. MIT.
Oehmke, M., 2009. Gradual arbitrage. Working paper. Columbia.
Perold, A., 1988. The implementation shortfall: paper versus reality. J. Portf. Manag. 14, 4–9.
Vayanos, D., 1998. Transaction costs and asset prices: a dynamic equilibrium model. Rev. Financ. Stud. 11, 1–58.
Vayanos, D., Vila, J.-L., 1999. Equilibrium interest rate and liquidity premium with transaction costs. Econ. Theory 13, 509–539.