THE SELF-FINANCING EQUATION IN HIGH FREQUENCY MARKETS

RENÉ CARMONA AND KEVIN WEBSTER

ABSTRACT. High Frequency Trading (HFT) represents an ever growing proportion of all financial transactions as most markets have now switched to electronic order book systems. The main goal of the paper is to propose continuous time equations which generalize the self-financing relationships of frictionless markets to electronic markets with limit order books. We use NASDAQ ITCH data to identify significant empirical features such as price impact and recovery, rough paths of inventories and vanishing bid-ask spreads. Starting from these features, we identify microscopic identities holding on the trade clock, and through a diffusion limit argument, derive continuous time equations which provide a macroscopic description of properties of the order book. These equations naturally differentiate between trading via limit and market orders. We give several applications (including hedging European options with limit orders, market maker optimal spread choice, and toxicity indexes) to illustrate their impact and how they can be used to the benefit of Low Frequency Traders (LFTs).

1. Introduction

In a series of papers ([19, 20, 21]) on the divide between high and low frequency traders, M. O'Hara and co-authors identified a number of market features that both Low Frequency Traders (LFTs for short) and most academic researchers have largely ignored, but that High Frequency Traders (HFTs from now on) exploit with great success.

> "There is no question that the goal of many HFT strategies is to profit from LFTs mistakes. [...] Part of HFTs success is due to the reluctance of LFT to adopt (or even to recognize) their paradigm."([21])

These papers also outline a program to better understand and possibly remedy these issues: in a nutshell, these authors recommend that LFTs update the strategies and models they use in order to incorporate more of the features of the high frequency markets. While the goal should not be to try to *beat* the HFTs at their own game by modeling the high frequency market microstructure in painstaking detail, it should be to *capture*, at least sparsely, the macroscopic effects of those phenomena that actually affect LFT.

This paper is in line with this program. Case in point, its main thrust is to provide forms of the *self-financing portfolio equation*, both in discrete and continuous time, consistent with the high frequency paradigm. The equations we propose are motivated by and fitted to high frequency data. They are derived theoretically from accounting rules at the high frequency level. Their continuous time limits capture

Date: December 8, 2013.

the relevant effects at the macroscopic level. From these fundamental relationships, we use the powerful tools of stochastic calculus to revisit the solutions of a certain number of standard continuous time financial problems in light of the new high frequency paradigm. We show how the latter affects for example option hedging and we highlight the different solution depending upon trading being through limit orders versus market orders. A model for market making in the spirit of [7] is solved. We also introduce, still in the same framework, an instantaneous and a cumulative toxicity indexes in the spirit of [21].

The crucial insight of [21], named 'the new paradigm', is the fact that high frequency traders do not operate on the 'calendar' clock, but instead use some form of 'event-based time', such as the trade clock, or the volume clock. This is partly due to the algorithmic nature of their strategies and the lack of direct calendar clock dependent constraints such as maturities and the likes. A fringe benefit for quantitative analysis is the well documented fact that prices *behave better* under an event-based clock than the calendar clock. A number of papers [6, 14, 15, 29, 30, 21] argue that, in addition to removing seasonal effects and resolving asynchronicity issues, this time-change makes the price returns more Gaussian-like. Even though this property is mostly irrelevant in our analysis, we choose to work in the *trade clock* in which each discrete time step corresponds to one trade. Indeed, even though our conclusions are independent of the clock used, we find the trade clock especially convenient to formulate and test the significance of our findings. With these proviso out of the way, we can outline our research agenda:

(1) Understand, at the microscopic level, structural relationships and strategies that HFTs exploit;
(2) Identify which features persist at the macroscopic level, in which form, and provide continuous time models on that scale;
(3) Use these models to update LFT strategies and provide monitoring tools: transaction cost analysis, measure of toxicity of order flow, ...

For the sake of definiteness, we focus on the self-financing portfolio equation of continuous time finance. To this effect, we review in Section 2 the role of this condition in quantitative finance, and in so doing, introduce the continuous time analysis notation used in the paper, as well as the exact form of our generalization.

The main originality of the form of the self financing condition which we propose to use, is the fact that it accounts for both price impact and price recovery, two important empirical microstructure features that are usually ignored or modeled in separate ad-hoc fashions. It also differentiates between the impacts of limit and market orders. This is important because nowadays, a large number of agents trade with both types of orders, rather than simply relying on market makers to find trades. Furthermore, our generalization of the self-financing portfolio equation can be used with a larger class of inventories models, e.g. *with infinite variation*. This allows the use of the powerful tools of stochastic calculus to retain tractability in a number of models.

The classical self-financing portfolio equation was generalized in two separate directions in the financial engineering literature. On one hand, Almgren and Chriss proposed in [4] a way to incorporate price impact and temporary transaction costs in a phenomenological model for optimal execution with market orders and finite speed of trading. On the other hand, and with a completely different point of view,

extensions of the classical self-financing equation of the Black-Scholes theory were touted by researchers attempting to include transaction costs in Merton's optimal portfolio's theory. See for example [13, 28, 36] or the recent review [27].

Two books, *Empirical market microstructure* by J. Hasbrouck ([23]) and *Market microstructure theory* by M. O'Hara ([34]) cover the state of the field prior to the advent of HFT. They contain informed trader models ([26]) and inventory-based market making ([5, 22, 24, 35]). *Three* main themes united different market structures at that time: the limit order book, adverse selection (the underlying cause of price impact) and statistical predictions. These themes are just as relevant, if not more so in the new age of high frequency trading.

Our investigations were inspired by a large number of empirical studies of high frequency data (see for example [8, 9, 10, 12, 32, 38, 39, 40]), and recent publications of theoretical models of the limit order book ([16, 17, 18, 31, 39]). However, our emphasis is different as we use limit orders as a *starting point*. Our goal is not to *explain* the evolution of the order book, but merely to analyze the *consequences* of the choices made by the liquidity providers and takers on price changes, their inventories and their wealth.

We close this introduction with a short overview of the paper. Since so much of our motivation and results depend upon the self financing condition, we devote next section to a review of the role of this condition in continuous time quantitative finance, with the goal of introducing the notation used in the paper, as well as announcing the exact form of our generalization. The remainder of the paper is structured into two parts. In the first part, we consider limit order books on which the trades take place at the best bid and best ask only. While seemingly restrictive, this assumption can be justified by looking closely at the data. Indeed, once two specific classes of executions are removed from the data [1], this assumption holds true in all the experiments reported in this paper. In the second part of the paper, we refrain from pre-processing the data in this way and we consider the case of a general order book. For the sake of completeness we derive the self-financing equations for a general order book shape. This generalization is needed for markets where a significant amount of trades happen outside the bid-ask spread. As expected, this part of the paper is more involved mathematically.

We first derive discrete versions of our self financing equation and of the price impact constraint from NASDAQ limit order book data. Our empirical studies are done in the trade clock, and we demonstrate the significance of our microscopic analysis by rigorous statistical tests. Next we take the limit as the tick size goes to zero, and obtain diffusion limits for both price and trade volumes. This leads to our proposed macroscopic continuous-time self-financing condition.

[1] We removed two specific classes of trades: 1) executions classified by NASDAQ as *type 'C'*. While we were not able to figure out what these special deals are, their numbers are very small, and on any given day, for any given stock, these executions represent less than 1% of the trades; 2) executions of hidden orders. While in very small numbers, if at all present, for small cap stocks, these trades are frequently very significant for large cap stocks. For example, on many days, the proportion of executions of hidden orders can be as large as 35 to 40% of the trades for stocks like Apple or Google. Moreover, no information is provided as to whether the execution is for a fully hidden order, or a the tip of an iceberg order. So, we decided to remove these executions for the purpose of this first empirical study of the self-financing condition from the order book.

We propose several applications of these macroscopic equations. We first revisit local volatility models for European options in our framework and obtain hedging strategies via limit or market orders. As a highlight, we show that limit orders can only hedge negative convexity options while market orders can hedge positive convexity options. This is a rare example where the theory naturally distinguishes between the roles of liquidity providers and liquidity takers. Then a model for high frequency market making is presented to uncover the relationship between optimal spread setting and price volatility. Finally, we propose two forms of toxicity of market order flow in our continuous time setting, and for the sake of illustration, we compute their empirical analogues on the pool of 120 stocks used in a recent ECB study of HFT. Following our theoretical analysis of general order book shapes, we propose for illustrative purposes, a supply and demand model based on perfect fill rates and deterministic price recovery.

2. **The self-financing equation**

In quantitative finance, the standard self financing portfolio equation is a cornerstone of the theory of frictionless markets. It plays a crucial role in many fundamental results, e.g. Merton's portfolio theory. Mathematically, speaking it is a simple equation which *constrains* the wealth process of an investor to live in a certain sub-space. This sub-space is therefore often called the space of *admissible* portfolios. New-comers to the mathematical theories of financial market often gripe with the self-financing condition and how it relates to the real world. While it can be postulated as a mathematical definition, it can also be *derived* from a limiting procedure starting from accurate descriptions of the microstructure of trades in the trade clock. This approach is at the core of our strategy.

> "The sad fact is that the self-financing condition is considerably more subtle in continuous time than it is in discrete time."[2]

When discussing market models at the macroscopic level, we assume that the mid-price p and the inventory L are given by Itô processes:

$$\begin{cases} dp_t &= \mu_t dt + \sigma_t dW_t \\ dL_t &= b_t dt + l_t dW'_t \end{cases} \tag{2.1}$$

for two Wiener processes W and W' with unspecified correlation structure. We shall also consider an adapted process s_t representing (in the continuous time limit) the bid-ask spread measured *in tick size*. The standard self-financing condition of continuous time finance can be stated as a constraint:

$$dX_t = L_t dp_t \tag{2.2}$$

between the price p of the underlying interest, the inventory L, and the wealth X of the agent. In most classical financial applications, case in point Merton's portfolio theory, the price p is exogenously given, the inventory L is the agent's input, and his wealth X appears as the output of equation (2.2).

The objective of this paper is to generalize the self-financing portfolio condition (2.2) to incorporate known idiosyncrasies of the high frequency markets including transaction costs, price impact and price recovery. Also, we want this generalization to be able to quantify the differences between trading via limit orders and market

[2]J. Michael Steele, *Stochastic Calculus and Financial Applications*, section 14.5 'Self-financing and self-doubt'.

orders. We warn the reader that the equations proposed in this paper are only *necessary* conditions and that quantifying limit order fill rates, priorities and price recovery are beyond the immediate scope of the present paper.

2.1. **Our basic formula.** The empirical analysis of NASDAQ order book data given in Section 3 and in the Appendix, together with the diffusion limit arguments of Section 4, prompt us to formulate the self-financing condition in the following form:

$$dX_t = L_t dp_t \pm \frac{s_t l_t}{\sqrt{2\pi}} dt + d[L,p]_t \tag{2.3}$$

where $\pm$ is $+$ when trading with limit orders, and $-$ when trading with market orders. Indeed, we show in Section 3 below that, when time is measured in the trade clock, the discrete time analog of formula (2.3) can be derived rigorously from a specific limit order book feature, and matches real wealth data extremely accurately. We shall also impose the constraint

$$d[L,p] < 0 \tag{2.4}$$

whenever trading with limit orders. Again, this *adverse selection* constraint is also dictated by the empirical analysis of the NASDAQ data.

We now explain how our condition (2.3) and the adverse selection constraint (2.4) relate to the conditions used in the separate sets of works reviewed in the introduction.

2.2. **The Almgren-Chriss model.** The seminal work by Almgren and Chriss [4] addresses a question closely related to ours. These authors propose a *macroscopic model* for the price impact and the change of wealth after a liquidity taker's decision. The model leads to a very tractable framework which was used by many optimal execution studies (see [2, 33] for example). This framework can be summarized by the system:

$$\begin{cases} dp_t &= f_t(l_t)dt + \sigma_t dW_t \\ dL_t &= l_t dt \\ dX_t &= L_t dp_t - c_t(l_t)dt \end{cases} \tag{2.5}$$

where f and c are two function-valued adapted processes which are positive, and in the case of c, convex.

The main advantage of this model is that price impact appears in a tractable fashion. Indeed, it comes through the function f_t, which creates a positive 'correlation' between traded volumes and the price process. However, it constrains L to be *differentiable* and for this reason, the model parameters cannot be calibrated to market data directly, making the model difficult to test empirically. As the empirical analysis of NASDAQ data reported in Section 3 and the appendix shows, there is ample evidence supporting nondifferentiable inventories. Moreover, limit orders are not part of the discussion in the Almgren-Chriss framework.

2.3. **Transaction cost literature.** The branch of classical mathematical finance most related to our paper is portfolio selection under transaction costs ([13, 28, 36] or the recent review [27]). Most of these works start from *a model for the wealth of a liquidity taker* which generalizes the self-financing equation to a setting with transaction costs. In general however, these papers do not emphasize the derivation of the model, but instead, the study of its consequences. We hope to

appeal to this side of the community by providing more accurate equations for self-financing portfolios while keeping similar tractability, leading the way to problems related to *liquidity provision*, such as market making. An interesting feature of such problems is that the agent does *not* directly control his portfolio, adding an additional modeling challenge. For the record we note that the standard equation used in this branch of the literature is

$$dX_t = L_t dp_t - \frac{s_t}{2}|dL|_t \tag{2.6}$$

where again, the inventory process L is assumed to have finite variation $\int_0^t |dL|_s < \infty$ for all finite t and s_t is the bid-ask spread.

Strengths of this model are its simplicity, relative tractability, and straightforward calibration to the market. Its weaknesses include the fact that the process L can only have finite variation. Moreover, price impact, limit orders and other microstructure considerations are absent in the model.

Formula (2.6) is much closer to our proposed equation (2.3) than it may seem at first. It merely corresponds to a different diffusion limit. It can be recovered in our framework by considering *non-vanishing* bid-ask spread, *zero* price impact and looking at market orders only. Notice that these assumptions may be more natural than ours for low frequency markets. This is presumably the reason for their introduction.

3. Empirical study and discrete equations

We first recall standard terminology from the high frequency markets.

3.1. High frequency terminology.

Trading on high frequency markets takes place on an object called *the limit order book*. An agent can interact with others via two possible trading mechanisms: limit orders and market orders. Limit orders correspond to the act of *providing liquidity* to the market, while market orders *take liquidity* from it. We will refer to agents who engage in the first type of trade as *liquidity providers*[3] while traders who trade with market orders will be referred to as *liquidity takers*. In real markets, traders often switch between liquidity providing and taking strategies, blurring this definition somewhat. The following comments can help highlight the differences.

- A liquidity taker pays a fee for his aggressiveness. This fee typically takes the form of the bid-spread, which is where most trades happen. The corresponding provider captures this bid-ask spread.
- Right after the trade happens, the price may move. If it does, it almost always moves *in favor* of the market order, compensating to some degree the transaction costs. This phenomena is called *price impact*. It is a consequence of the *adverse selection* of limit orders by takers.
- Between two successive trades, the price reverts to some value in between the impacted price and the original one. *Price recovery* is an intuitive name often used to describe this high frequency feature.
- Takers control their inventory directly. Attaining *correlation* with the market requires high frequency *predictions* of the next price move.

[3]Of which *market makers* are a special class.

- Providers do not directly control their inventory, but only their exposure to the flow of market orders. How much of the flow they are able to capture depends on their limit order *fill rate*. Flow is considered *toxic* if it leads to adverse selection. The profitability of a provider's strategy depends on the spread he captures and the toxicity of his flow.

3.2. **Data used in the Study.** The statistical tests reported in this paper were produced by the analysis of the NASDAQ ITCH data of, amongst other stocks, the pool of 120 stocks used in the recent ECB study [11] of high frequency trading. The figures included in this paper were produced using the data for Coca Cola (KO) on 18/04/13 . As explained in an earlier footnote, the only cleaning pre-processing of the raw data was to remove the special deals and the executions of hidden orders.

The data do not contain the identity of the agents involved in the transactions. For that reason, all quantities relating to the inventory L, cash K or wealth X are aggregate quantities which could be thought of as relating to a *representative aggregate liquidity provider*. The mid-price will be denoted by p and the bid-ask spread by s. The time stamps of the transactions are measured in fractions of microseconds and given in the *calendar* clock. However, the data analysis is performed in the *trade clock* $n = 1, ...N$ where each time step corresponds to one trade time. For example, $p_n = p_{t_n} = p_{t_n-}$ where t_n is the n-th trading time in the calendar clock gives the mid-price just before the n-th transaction. Limit order data happening between two trade times is the source of the changes in the best bid and best ask, (and consequently of the mid-price) and is discarded for the purpose of our analysis. More generally, if Y is a discrete process, we denote by $\Delta_n Y$ the forward-looking increment $\Delta_n Y = Y_{n+1} - Y_n$.

3.2.1. *More Notation.* We denote by s_n the bid-ask **spread** just before the n-th trade. In other words, s_n is the difference between the best ask and the best bid, just before the n-th trade. We shall argue later on that the spread is of the same order of magnitude as the change in price, namely that $s_n \approx |\Delta_n p|$. We also

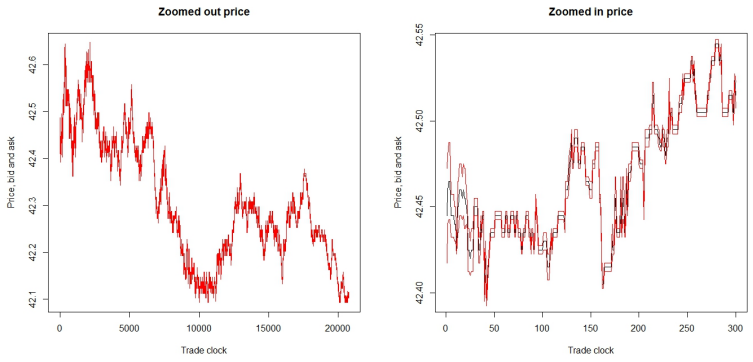


FIGURE 1. Plots of the best bid, best ask and mid-price as functions of trade time (left). Zoom into a part of the graph to see the differences between the three plots (right).

denote by L_n and K_n the inventory and the cash held by the aggregate liquidity provider just before the n-th trade. These quantities are not given explicitly with

the data provided by NASDAQ, but starting from $L_0 = K_0 = 0$, they can easily be computed after each trade. Indeed, L_n is the cumulative sum up to time n of the algebraic volumes of the trades (positive volume for a limit order executed against a sell market order, and negative volume for an execution against a buy market order). Similarly, K_n is the cumulative sum up to time n of the cash exchanged during the trades. The inventory and the cash L_n and K_n held by the aggregate liquidity provider are plotted in Figure 2 against the trade time n.

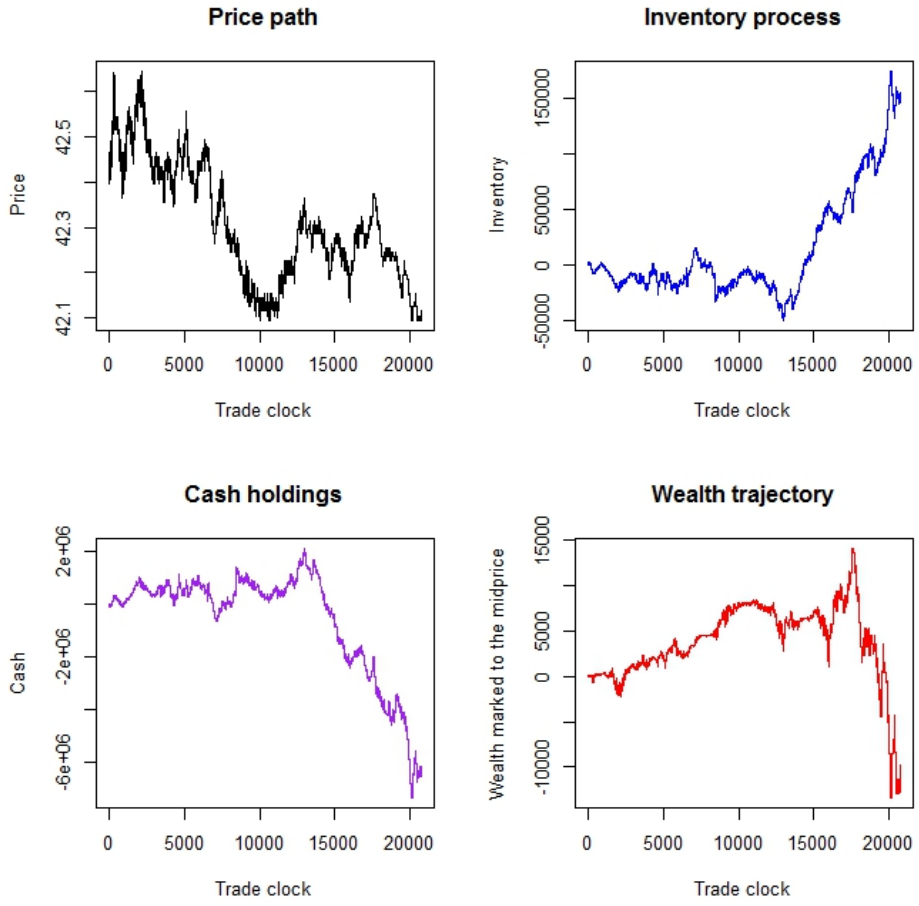


FIGURE 2. Coca Cola (KO) stock on 18/04/13. Inventory, cash and wealth are those of the aggregate liquidity provider.

3.3. **Price impact.** Empirically, price impact is the simple fact that the price moves after each trade, and tends to move in favor of the market order. There have been several empirical studies and multiple proposed measures and models for it ([2, 4, 9, 10, 12, 32, 33, 38, 39]). The main economic interpretation for price impact is adverse selection. In this study, we isolate, measure and model price impact by

a straightforward relationship.

$$\Delta_n L \Delta_n p \leq 0. \tag{3.1}$$

For all $n = 1, ...(N-1)$. This relationship states that the price cannot move up when the liquidity provider has bought and cannot move down when the provider has sold. From the taker's perspective, this means that the price always moves in his favor right after a trade.

We provide rigorous statistical tests of (3.1) in Appendix A. For the sake of illustration, we note that for Coca Cola on April 18, 2013, (3.1) holds for all but 166 of the 20742 trades of our streamlined data set, which represents 0.9% of the trades. This trade impact relationship has clear consequences for the continuous time analogs of the discrete model considered here: *the quadratic covariation* between the provider's inventory and the price process is *negative* and *decreasing*. Conversely, the quadratic covariation between the inventory of a liquidity taker and the price process is positive and increasing.

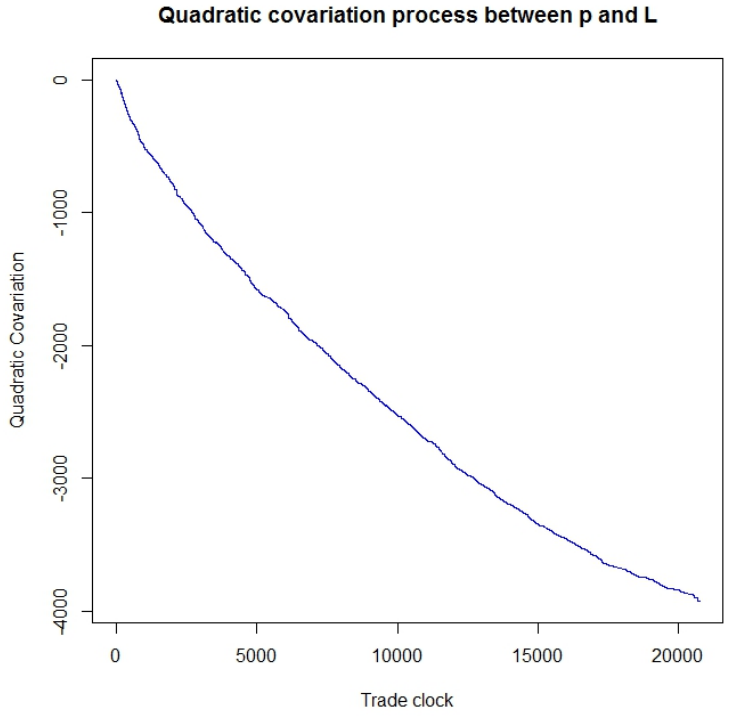


FIGURE 3. Quadratic covariation between inventory and price path.

Price impact will give us an extra compelling reason to accept trade volumes with infinite variation. Indeed, when using continuous time models, if the price path and the inventory have a non-negligible quadratic covariation, then we cannot model one as a diffusion process and the other as a finite variation process.

Remark 3.1. *The causality of price impact is unclear: do trades cause price movements, or simply predict them? While not crucial for the mathematical theory, it*

is important for interpretation purposes, and we choose to use the second option. In particular, we shall say that a liquidity taker whose changes in inventory are strongly correlated *with the price movements has a very good* short term prediction of the price. *This typically is the case of sophisticated high-frequency traders. Low-frequency traders, on the other hand, trade more slowly and acquire inventories which aren't directly correlated with the smaller price movements.*

3.4. **Price recovery.** This is another simple observation. Trades move prices, but typically move them *at most* by one bid-ask spread. If they systematically moved the price by one bid-ask spread, then the correlation between the price path and the taker inventory should be one. Otherwise, it is smaller than one and we say that the *price has recovered* from the price impact. Note that, of all our relationships, this is statistically the weakest one: it is not verified for 5% of the Coca Cola data.

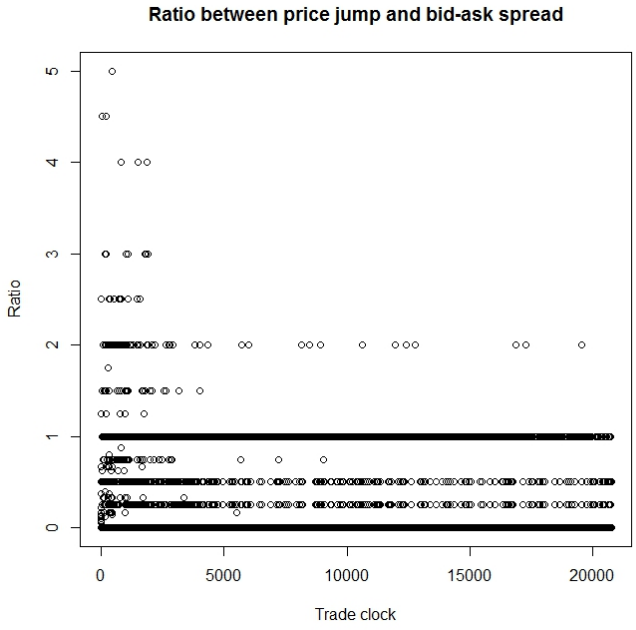


FIGURE 4. Relationship between price increments and spread.

Mathematically, this implies that $|\Delta_n p| \leq s_n$ for $n = 1, ...(N-1)$. In the continuous time version considered later, it will provide in the diffusion limit an upper bound on instantaneous price volatility based on the current spread.

3.5. **A bit of accounting.** Finally, we derive the self-financing portfolio equation from first principles in such a high frequency market.

Because after removing the special deals and the executions against hidden orders, all the trades do happen at the best bid or ask, the amount of cash exchanged is equal to

$$\Delta_n K = \begin{cases} -(p - \frac{s_n}{2})\Delta_n L & \text{if } \Delta_n L \geq 0 \\ -(p + \frac{s_n}{2})\Delta_n L & \text{else} \end{cases} \tag{3.2}$$

That is, the provider pays the bid (resp. receives the ask) when he buys (resp. sells). This can be summarized by the equation:

$$\Delta_n K = -p_n \Delta_n L + \frac{s_n}{2}|\Delta_n L| \tag{3.3}$$

3.5.1. *The aggregate liquidity provider's wealth.* We define wealth as

$$X_n = p_n L_n + K_n \tag{3.4}$$

that is, the cash held by the liquidity provider plus the value of her inventory marked to the mid-price. The wealth X_n of the aggregate liquidity provider is plotted in Figure 2 against the trade time n.

3.5.2. *The discrete self-financing equation.* We derive the dynamics of the wealth process X from equations (3.3) and (3.4):

$$\begin{aligned}\Delta_n X &= L_n \Delta_n p + p_n \Delta_n L + \Delta_n p \Delta_n L + \Delta_n K \\ &= L_n \Delta_n p + \frac{s_n}{2}|\Delta_n L| + \Delta_n p \Delta_n L \end{aligned} \tag{3.5}$$

3.5.3. *Empirical validation.* We compare four quantities: 1) the actual wealth, 2) the wealth computed from the standard self-financing equation:

$$\Delta_n X = L_n \Delta_n p \tag{3.6}$$

used in the classical Black-Scholes option pricing and Merton portfolio theories, 3) the wealth computed from the standard self-financing condition:

$$\Delta_n X = L_n \Delta_n p + \frac{s_n}{2}|\Delta_n L| \tag{3.7}$$

advocated to include transaction costs in Merton's theory of optimal portfolio choice, and finally 4) the wealth computed from our self-financing condition (3.5). The plots of these four wealth processes are given in Figure 5 for Coca Cola stock on April 18, 2013. Changing stock or changing day does not seem to affect the following facts which are easily illustrated in this figure. The wealth computed from the standard self-financing equation of the Black-Scholes theory clearly underestimates the actual wealth of the aggregate liquidity provider The wealth computed from the classic equation (3.7) tries to correct for the lack of transaction cost, but it over-shoots and over-estimates the wealth of the aggregate liquidity provider. The error is reduced and practically canceled by including the adverse selection term given by the quadratic covariation, and using our proposed formula (3.5). The quadratic covariation between inventory and price matters!

3.5.4. *Recovering the frictionless case.* A surprising property worth mentioning concerns the case $s_n = 0$. Indeed, the latter does *not* correspond to the frictionless case. Rather, choosing price jumps $|\Delta_n p| = s_n/2$ and using the fact that the price impact is negative, i.e. $\Delta_n L \Delta_n = -|\Delta_n L \Delta_n|$, yields the identity

$$\Delta_n X = L_n \Delta_n p + \frac{s_n}{2}|\Delta_n L| - |\Delta_n p||\Delta_n L| = L_n \Delta_n p$$

which is the standard self-financing portfolio equation. In our high frequency framework, it is not the absence of transaction costs that corresponds to the frictionless case, but rather the absence of price-recovery, for in that case, the price impact exactly compensates the transaction costs.

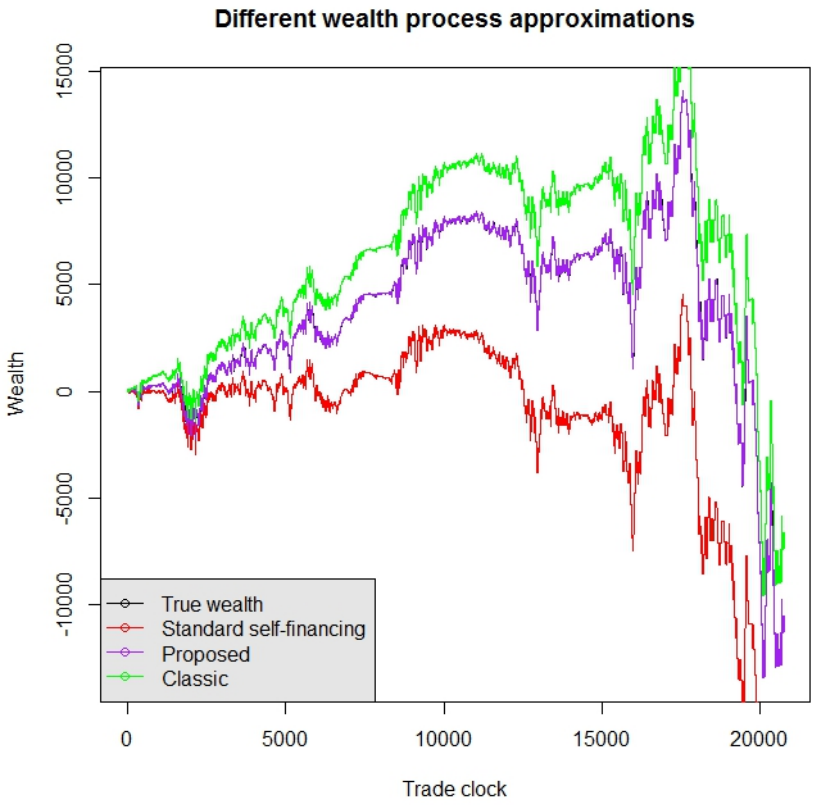


FIGURE 5. Plots of the actual wealth of the aggregate liquidity provider (as in Figure 2) together with the wealth computed from the three self-financing conditions. Red is the frictionless case. Green corresponds to (3.7). The actual wealth and the wealth computed from our self-financing condition (3.5) are indistinguishable on the graph.

3.6. **Summary.** Our empirical evidence suggests the following equations and features for the inventory L and wealth X of a liquidity provider, the bid-ask spread s and the price p:

3.6.1. *Self-financing equation.*

$$\Delta_n X = L_n \Delta_n p + \frac{s_n}{2}|\Delta_n L| + \Delta_n p \Delta_n L \tag{3.8}$$

3.6.2. *Price impact (adverse selection).*

$$\Delta_n L \Delta_n p \leq 0 \tag{3.9}$$

3.6.3. *Price recovery.*

$$|\Delta_n p| \leq s_n \tag{3.10}$$

3.6.4. *Vanishing bid-ask spread.* s_n and $\Delta_n p$ are of the same order of magnitude, namely $s_n \approx |\Delta_n p|$.

4. **Continuous equation: Bid-Ask case**

The aim of this section is to derive formula (2.3) from its discrete version (3.8) established in the previous section. In the process, we shall also derive continuous-time analogs of the price impact / adverse selection constraint, the price recovery and vanishing bid-ask spread condition equivalents of the relationships of subsection 3.6. The key is to let the tick size vanish, assume that the bid-ask spread vanish with the tick size, and assume that the price and inventory converge to diffusion limits.

Remark 4.1. *The bid-ask spread is of the same order of magnitude as the price jumps: the tick size. This implies in particular that the bid-ask spread vanishes in absolute terms in the diffusion limit and should therefore be measured in tick-size. The mathematical consequence of this simple comment is that transaction costs do not diverge as the tick size goes to zero, allowing inventories that have infinite variations in the continuous limit.*

The main technical tool we use in this section and section 6 is the functional law of large number for discretized process by Jacod and Protter [25]. Let ϕ_{σ^2} denote the density function of the Gaussian distribution with mean 0 and variance σ^2.

Theorem 4.2 ((7.2.2) from [25])**.** *Let $(t, y) \to F_t(y)$ be an $\mathcal{F}_t$-adapted random function that is a.s. continuous in (t, y) and verifies the growth condition $F_t(y) \leq Cy^2$ for some constant C. Then we have the following convergence u.c.p. as $N \to \infty$ for any continuous Itô process Y:*

$$\frac{1}{N}\sum_{n=1}^{\lfloor Nt \rfloor} F_{n/N}\left(\sqrt{N}(Y_{(n+1)/N} - Y_{n/N})\right) \to \int_0^t \int F_s(y)\phi_{\sigma_s^2}(y)dy\, ds$$

where $\sigma_t^2 = \frac{d[Y,Y]_t}{dt}$.

We proceed as follows:

(1) We begin with the continuous processes for the inventory L, price p and bid-ask spread s as our *data*.
(2) By discretizing them, we obtain the *data* to plug into the discrete relationships listed in subsection 3.6, yielding our *discrete time* output relationships.
(3) Finally, we take the limit again to obtain the diffusion limits of our discrete output to obtain our continuous-time relationships.

In discrete time, prices are a pure-jump process, and therefore have finite variations. It is common on larger time scales to consider the price as 'zoomed out' enough to be approximated by a diffusion process. Mathematically, this corresponds to a vanishing tick size. Recall that tick size is typically of the order of magnitude of the cent[4], that is 10^{-4} relative to the typical stock price. Given the relative roughness of the path of inventories when compared to prices, see for example Figure 2, it seems reasonable to also expect high-frequency inventories to be modeled by processes with infinite variation.

[4]Decibasis point for some exchanges in the foreign exchange market.

4.1. **Mathematical Setup.** Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a filtered probability space supporting two Wiener processes W and W' with unspecified correlation structure. We consider two Itô processes for the price p and provider inventory L:

$$\begin{cases} p_t &= p_0 + \int_0^t \mu_u du + \int_0^t \sigma_u dW_u \\ L_t &= L_0 + \int_0^t b_u du + \int_0^t l_u dW'_u \end{cases} \tag{4.1}$$

where p_0 and L_0 are $\mathcal{F}_0$-measurable square integrable random variables, and μ, σ, b and l are $\mathbb{F}$-adapted continuous processes. Finally, we also assume the existence of a $\mathbb{F}$-adapted continuous process s.

Now consider the discrete approximation $p_n^N = p_{n/N}$ and likewise for L, μ, σ, b and l. The interpretation is that $\frac{1}{\sqrt{N}}$ is the tick size, which we formally make vanish. For the bid-ask spread s, we define $s_n^N = \frac{1}{\sqrt{N}} s_{n/N}$ in line with our previous comments. Plugging these definitions into the equations from subsection 3.6, we obtain the discrete relationships:

$$\begin{cases} \Delta_n X^N = L_n^N \Delta_n p^N + \frac{s_{n/N}}{2} \frac{1}{\sqrt{N}} |\Delta_n L^N| + \Delta_n p^N \Delta_n L^N \\ \Delta_n L^N \Delta_n p^N \leq 0 \\ |\Delta_n p^N| \leq s_n^N \end{cases} \tag{4.2}$$

where the first equation is understood as the definition of the wealth X^N.

4.2. **Main result.**

Theorem 4.3. *Assuming that relations* (4.2) *hold for every $N \geq 1$, then the limit $\lim_{N\to\infty} X^N_{\lfloor Nt \rfloor}$ exists for the uniform convergence in probability and defines a process X_t which together with the Itô processes p_t and L_t satisfy the relationships:*

$$\begin{cases} dX_t = L_t dp_t + \frac{s_t l_t}{\sqrt{2\pi}} dt + d[L,p]_t \\ d[L,p]_t \leq 0 \\ \sigma_t \leq \sqrt{\frac{2}{\pi}} s_t \end{cases} \tag{4.3}$$

Proof. Using a localizing sequence of stopping times if needed, we can assume without any loss of generality that the process s_t is bounded by a constant. The convergence of the discrete approximations of $\int_0^t L_u dp_u$ and $[L,p]_t$ is plain, proving the second relationship.

For the last term of the self-financing equation, we have that

$$\frac{s_n^N}{2} \frac{1}{\sqrt{N}} |\Delta_n L^N| = \frac{1}{2N} s_{n/N} |\sqrt{N} \Delta_n L^N| \tag{4.4}$$

which allows us to apply Theorem 4.2 with $F_t(y) = \frac{s_t}{2}|y|$ and $Y_t = L_t$. This proves the self-financing equation.

The last relationship of (4.3) follows from applying the same theorem to the process $Y_t = p_t$ and the random function $F_t(y) = y^2 - s_t|y|$. We obtain that, for each $t_1 < t_2$, the quantity

$$\frac{1}{N} \sum_{n=\lfloor t_1 N \rfloor}^{\lfloor t_2 N \rfloor} \left((\sqrt{N} \Delta_n p^N)^2 - s_{n/N} |\sqrt{N} \Delta_n p^N| \right) \tag{4.5}$$

converges toward

$$\int_{t_1}^{t_2} (\sigma_t - \sqrt{\frac{2}{\pi}} s_t)\sigma_t dt, \tag{4.6}$$

and the fact that this process is negative for all $t_1 < t_2$ concludes the proof. □

Remark 4.4. *Technically speaking, nothing prevents us from going through with the same limiting argument for the hidden part of the order book, simply replacing p_t and s_t by their 'hidden' counterparts. Two practical problems appear however. First, measuring the hidden price and spread is difficult. Second, and more importantly, it is unclear by what to replace the price impact inequality, as adverse selection of hidden orders is not well studied or understood.*

4.3. **Time change.** Note that equation (2.3) was proved in a *trade clock*, which means that all the time-related quantities, such as volatility, must be measured per trade time. While this is a positive feature for high frequency models under this clock (e.g. [6, 9]), it is less advantageous for financial problems working under a different clock. For example, pricing an option with a fixed maturity in the calendar clock may be difficult to do directly from equation (2.3). We therefore discuss how our proposed formula behaves under time-changes, with the canonical time-change being the switch to a calendar clock. Another possible time-change is the switch from a trade clock to a volume clock.

Definition 4.5. *We define a good time change to be an $\mathcal{F}_t$-adapted stochastic process τ_t such that $\tau_0 = 0$ and*

$$d\tau_t = n_t^2 dt \tag{4.7}$$

with n_t uniformly bounded away from zero.

We start from:

$$\begin{cases} dp_t &= \mu_t dt + \sigma_t dW_t \\ dL_t &= b_t dt + l_t dW'_t \\ dX_t &= L_t dp_t + \frac{s_t}{\sqrt{2\pi}} l_t dt + d[L,p]_t \end{cases} \tag{4.8}$$

with $d[L,p]_t \leq 0$, and we study the processes $\tilde{p}_t = p_{\tau_t}$, $\tilde{L}_t = L_{\tau_t}$ and $\tilde{X}_t = X_{\tau_t}$. Note that all the time-changed processes are now adapted with respect to the time-changed filtration $\tilde{\mathcal{F}}_t = \mathcal{F}_{\tau_t}$. Note also that the processes $\tilde{W}_t = \int_0^{\tau_t} 1/n_{\tau_u^{-1}} dW_u$ and $\tilde{W}'_t = \int_0^{\tau_t} 1/n_{\tau_u^{-1}} dW'_u$ are $\tilde{\mathcal{F}}_t$ Wiener processes.

A simple chain-rule leads to the time-changed dynamics:

$$\begin{cases} d\tilde{p}_t &= \tilde{\mu}_t dt + \tilde{\sigma}_t d\tilde{W}_t \\ d\tilde{L}_t &= \tilde{b}_t dt + \tilde{l}_t d\tilde{W}'_t \\ d\tilde{X}_t &= \tilde{L}_t d\tilde{p}_t + \frac{\tilde{s}_t}{\sqrt{2\pi}} \tilde{l}_t dt + d[\tilde{L},\tilde{p}]_t \end{cases} \tag{4.9}$$

where

$$\tilde{\mu}_t = n_t^2 \mu_{\tau_t}; \quad \tilde{b}_t = n_t^2 b_{\tau_t}$$
$$\tilde{\sigma}_t = n_t \sigma_{\tau_t}; \quad \tilde{l}_t = n_t l_{\tau_t}$$

which are standard, as well as the more surprising:

$$\tilde{s}_t = n_t s_{\tau_t} \tag{4.10}$$

Remark 4.6. *Part of this result is expected: under the modified time clock, drifts and volatility must be measured by the new unit of time instead of by unit of trade, which corresponds to the factors n_t^2 and n_t. However, the unfortunate result is that the bid-ask spread must also be multiplied by n_t, which means that one needs to keep track of the process $\tilde{s}_t$ rather than the more natural process s_{τ_t}.*

Remark 4.7. *This issue is resolved when $s_t = \lambda\sigma_t$. Such a assumption would follow the conclusion of the empirical paper* [39] *which suggests a linear relationship between daily average bid-ask spread and daily average volatility* per trade. *From a theoretical perspective, this model is stable under time change, in the sense that $\tilde{s}_t = \lambda\tilde{\sigma}_t$, a desirable property.*

Remark 4.8. *One could have also from the beginning worked under the changed clock and used the law of large numbers with irregular discretization schemes found in* [25] *to recover the same result.*

4.4. **The case of a liquidity taker.** By symmetry, the corresponding equations for the inventory and wealth of a liquidity taker are

$$\begin{cases} dX_t = L_t dp_t - \frac{s_t l_t}{\sqrt{2\pi}} dt + d[L,p]_t \\ d[L,p]_t \geq 0 \end{cases} \tag{4.11}$$

Unfortunately, as we already pointed out, these equations are only *necessary* conditions. Indeed, unlike with the standard self-financing equation, it is difficult to tell which processes L and p are admissible: we can only derive X once L and p are given.

To give an example of why not all L can be attained, assume the volume on the order book is finite. Then the volatility of L must be bounded by the amount of volume available. Other factors that can come into play to determine which processes L are actually attainable by market participants are: limit order fill rate, instantaneous price recovery and for market orders the ability to predict the next price jump. These factors will directly impact the volatility of L and the possible correlation and quadratic covariation between L and p.

Ultimately, supply and demand rule the price p and volume L. X, however, stems from accounting rules.

5. **Applications**

Applications of the proposed relationships depend on models of the inventory L and the price p. Notice that, when we formulate an optimization problem, we often assume that the inventory can be any Itô process. This is an act of faith as making it happen typically requires good execution algorithms and limit order fill rates.

Reasonable models for the spread s are easier to come by. We shall typically scale the spread with the price volatility: $s_t = \sqrt{2\pi}\lambda\sigma_t$ (for some constant $\lambda > 1/2$). This is consistent with the empirical literature on the matter, e.g. [39].

5.1. **Hedging.** In this subsection we explore perfect replication of European options, assuming that the corresponding inventory can be attained via high-frequency trades. Let f be the payoff function of our option and let

$$dp_t = \mu(t,p_t)dt + \sigma(t,p_t)dW_t \tag{5.1}$$

be a Markovian stochastic differential equation for the price. Denote by L the inventory of the hedger and let us assume that it is of the form:

$$dL_t = b_t dt + l_t dW_t \tag{5.2}$$

with b_t and l_t continuous, bounded and adapted processes. Note that the dynamics of L_t are driven by the same Wiener process as the price, so the model is complete and perfect replication is possible. Note also that $l_t < 0$ corresponds to trading via limit orders and $l_t > 0$ to trading via market orders. Furthermore, when working with this signed l_t, the self-financing equation writes the same for limit and market orders:

$$dX_t = L_t dp_t - \frac{s_t l_t}{\sqrt{2\pi}} dt + d[L,p]_t \tag{5.3}$$

as when $l_t < 0$, we want to capture the transaction costs and when $l_t > 0$, we need to pay them.

Assume that interest rates are zero and define by $v(t,p)$ the price of the option knowing that $p_t = p$. Then we have the replication equation

$$d(X_t - v(t,p_t)) = (L_t - \Delta_t)\, dp_t + d[p,L]_t - \frac{s_t}{\sqrt{2\pi}} l_t dt - (\Theta_t + \frac{1}{2}\Gamma_t \sigma^2(t,p_t))dt$$

where Δ_t, Θ_t and Γ_t denote the usual Greeks evaluated at t and p_t. Delta hedging the option removes the price risk and leads to the equation

$$\begin{aligned} dL_t &= d\Delta_t \\ &= \left(\partial_t^2 v(t,p_t) + \frac{1}{2}\sigma^2(t,p_t)\partial^3_{tp^2} v(t,p_t) \right) dt + \Gamma_t dp_t \end{aligned}$$

and in particular the identity

$$l_t = \Gamma_t \sigma(t,p_t) \tag{5.4}$$

Note that therefore l_t and Γ_t must be of the same sign!

Finally, the pricing partial differential equation becomes

$$\partial_t v(t,p) + \left(\lambda - \frac{1}{2}\right) \sigma^2(t,p) \partial_p^2 v(t,p) = 0 \tag{5.5}$$

with terminal condition $v(T,p) = f(p)$. As $\lambda > 1/2$, this leads to a multiplicative factor of $\sqrt{2\lambda - 1}$ on the implied local volatility when compared to the frictionless case.

Remark 5.1. *An important point is that negative Gamma options can be replicated via limit orders, while positive Gamma options can be replicated via market orders. This is assuming that one can guarantee perfect correlation (respectively anti-correlation) with the price process for the inventory L_t to be driven by the same Wiener process as the price.*

This is consistent with the fact that one would not expect to use limit orders to delta-hedge a call option, as hedging a call option requires you to buy when the price goes up and sell when the price goes down: exactly the opposite of a limit order.

5.2. **Market making.** In this section, we adapt to our framework the key insight of the model proposed in [7]. The ultimate aim is to solve the optimization problem of a representative market maker choosing the spread and maximizing his profits. The trade-off he faces, and which is the key ingredient of the model, is the following: the smaller the spread, the likelier trades are, but the less profit he makes on each of them.

In a way similar to [7, 37], we model the probability of execution of a limit order by a decreasing function of the quoted spread. This will first be done at the microscopic level, to obtain a reasonable model for our inventory process L at the macroscopic level. A key difference with [7] is that we still impose the price impact constraint, which will further depress the market maker's profits because of adverse selection.

To guarantee the price impact constraint is satisfied, we use, at the microscopic level, a modified version of the Almgren and Chriss model [4] to relate the price to the aggregate inventory of the liquidity providers. We assume that

$$\Delta_n L = -\lambda_{n+1}\Delta_n p \tag{5.6}$$

for a $\mathcal{F}_{n+1}$-measurable, positive random variable λ_{n+1}. This is an unpredictable form of linear price impact, in the sense that, ex-post, the price increment is a linear function of the traded volume.

To capture the insight of [7], we model λ_{n+1} in such a way that

$$\mathbb{E}[\,\lambda_{n+1}|\,\mathcal{F}_n] = \rho_n(s_n)f_n(s_n); \quad \mathbb{E}[\,\lambda_{n+1}^2\big|\,\mathcal{F}_n] = (f_n(s_n))^2 \tag{5.7}$$

where s_n is the market maker's chosen spread, and ρ_n and f_n are continuous, positive function with f_n decreasing and $\rho_n \in [0,1]$. The assumption that f_n is decreasing in the spread is inherited from [7], and the fact that ρ must be smaller than 1 is due to Jensen's convexity inequality. We assume λ_{n+1} to be independent of $\Delta_n p$ conditional on $\mathcal{F}_n$. Computing the predictable quadratic variation of L_n yields:

$$\sum_{k=1}^{n-1} f_k^2(s_k)\mathbb{E}\left[\,\Delta_k p^2\big|\,\mathcal{F}_k\right], \tag{5.8}$$

while the predictable quadratic covariation of L_n and p_n is given by:

$$-\sum_{k=1}^{n-1} \rho_k(s_k)f_k(s_k)\mathbb{E}\left[\,\Delta_k p^2\big|\,\mathcal{F}_k\right]. \tag{5.9}$$

This suggests the use of the following model in the continuum limit:

$$\begin{cases} dp_t &= \mu_t dt + \sigma_t dW_t \\ dL_t &= -\rho_t(s_t)f_t(s_t)\mu_t dt + f_t(s_t)\sigma_t dW_t' \end{cases} \tag{5.10}$$

with $d[W,W']_t = -\int_0^t \rho_u(s_u)du$ for some adapted, continuous and positive functions $\rho_t(\cdot)$ and $f_t(\cdot)$ with $\rho_t \leq 1$ and f_t decreasing. Note that the equation for L_t can also be rewritten as:

$$dL_t = -\rho_t(s_t)f_t(s_t)dp_t + f_t(s_t)\sqrt{1-\rho_t^2(s_t)}\sigma_t dW_t^{\perp} \tag{5.11}$$

with a Wiener process $W_t^{\perp}$ independent from W_t. We will from now on assume that p_t is adapted to the filtration generated by W_t.

Applying our wealth equation, we obtain:

$$X_T = L_T p_T - \int_0^T p_t dL_t + \frac{1}{\sqrt{2\pi}} \int_0^T \sigma_t s_t f_t(s_t) dt. \tag{5.12}$$

For both f_t and ρ_t, a natural assumption is that they are functions of the spread rescaled by the volatility:

$$f_t(s) = f(s/\sigma_t); \quad \rho_t(s_t) = \rho(s_t/\sigma_t) \tag{5.13}$$

for some C^0 decreasing function f and C^0 function ρ. We will furthermore assume that $g(x) = xf(x)$ is a decreasing function for x large enough, that $g(x) \to 0$ as $x \to \infty$, and that $f(x) > 0$ for all $x \geq 0$.

The problem of a risk-neutral market maker attempting to set the spread optimally is to maximize:

$$\sup_s \mathbb{E} X_T. \tag{5.14}$$

We solve this control problem using the Pontryagin maximum principle. Let us define a few functions first.

Lemma 5.2. *For all $a > 0$, define the function F_a by*

$$F_a : x \mapsto \frac{x}{\sqrt{2\pi}} f(x) - a\rho(x) f(x) \tag{5.15}$$

Then the function

$$M(a) = \max_{x \in [0,\infty)} F_a(x) \tag{5.16}$$

is well defined, continuous, and decreasing in a. Furthermore, there exist a measurable selection

$$m(a) \in argmax_{x \in [0,\infty)} F_a(x) \tag{5.17}$$

and we have that $m(a) > 0$.

Proof. First, note that for all $a > 0$,

$$F_a(0) = -a\rho(0) f(0) \leq 0, \quad F_a(a+1) \geq f(a+1) > 0$$

Next, if g is decreasing on the interval $[x_0, \infty)$, then we can define the function $\beta(a)$ as $g^{-1} \circ f(a+1)$ if $f(a+1)$ is in $g[x_0, \infty)$, and x_0 otherwise. $\beta(a)$ is continuous and verifies $f(a+1) \geq g(x)$ for all $x \in (\beta(a), \infty)$.

This proves that the maximum of F_a is attained on the compact $[a+1, \beta(a)]$. The continuity of M holds by Berge's maximum theorem. It is decreasing by definition of F_a. The measurable selection result follows by Thm 18.19 of [3]. □

Proposition 5.3. *Any solution of the control problem is of the form*

$$\frac{s_t}{\sigma_t} = m(\alpha_t) \tag{5.18}$$

where

$$\alpha_t = \mathbb{E}\left[p_T - p_t \middle| \mathcal{F}_t \right] \frac{\mu_t}{\sigma_t^2} + \frac{Z_t}{\sigma_t}, \tag{5.19}$$

Z_t being the volatility of the martingale representation of p_T

Proof. We apply the necessary part of the stochastic Pontryagin maximum principle. The generalized Hamiltonian is equal to:

$$\mathcal{H}_t(s, L, Y, Z, Z^\perp) = -\rho(s/\sigma_t)f(s/\sigma_t)\left[(Y_t - p_t)\,\mu_t + \sigma_t Z\right] + \frac{\sigma_t}{\sqrt{2\pi}} s f(s/\sigma_t) + \sigma_t f(s/\sigma_t)\sqrt{1-\rho^2(s/\sigma_t)}Z^\perp$$

and the adjoint equation is solved by

$$Y_t = \mathbb{E}\left[p_T \middle| \mathcal{F}_t\right] \tag{5.20}$$

which, in particular, implies $Z_t^\perp = 0$. Z_t can be computed via the martingale representation theorem on the Brownian filtration generated by W_t.

The Hamiltonian to maximize therefore becomes

$$\sigma_t^2 F_{\alpha_t}\left(\frac{s}{\sigma_t}\right) \tag{5.21}$$

and the previous lemma concludes. □

Beyond the optimal control, one might be interested in the dependence in σ_t and α_t of the market maker's expected profits as well as the volatility of his inventory. Note that a low volatility of the inventory means that the market maker has essentially pulled out of the market.

Corollary 5.4. *The market maker's expected profits and losses are*

$$\mathbb{E}\left[\int_0^T M\left(\alpha_t\right)\sigma_t^2 dt\right] \tag{5.22}$$

while the volatility of his inventory is

$$\sigma_t f(m\left(\alpha_t\right)). \tag{5.23}$$

Proof. The expected profits can be computed by integrating the Hamiltonian along the optimal path. The rest follows from the previous proposition. □

A consequence of the corollary is that the market maker is on average short α_t and, for α_t being fixed, long volatility.

There are now two distinct problems if one looks for tractable formulas. First, an explicit model for p_T must be given for which the martingale representation term Z_t can be computed. Second, one has to propose a function g for which the maximal argument m of F can easily be characterized as a function of α_t.

5.2.1. *The martingale case.* Note that the latter problem is solved when p_t is assumed to be a martingale. Indeed, if we have

$$dp_t = \sigma_t dW_t \tag{5.24}$$

for some adapted, continuous and positive process σ_t. Then $\alpha_t = 1$ and we simply have

$$s_t = m(1)\sigma_t \tag{5.25}$$

circumventing the need for explicit functions ρ and f. This result provides a theoretical argument for the empirical claim made in [39] that the spread is a linear function of volatility.

Plugging this optimal spread back into the objective function, the market maker's expected profits and losses (P&L) are

$$M(1)\mathbb{E}\left[\int_0^T \sigma_t^2 dt\right] \tag{5.26}$$

In the martingale case, the market maker is therefore *on average*, Delta neutral, has negative Gamma but positive Vega.

5.2.2. *Explicit cases.* Other cases where α_t can be computed explicitly are:

- the Black-Scholes model

$$dp_t = \mu p_t dt + \sigma p_t dW_t \tag{5.27}$$

  in which case we obtain:

$$\mathbb{E}\left[p_T \middle| \mathcal{F}_t\right] = p_t e^{\mu(T-t)}; \quad Z_t = \sigma p_t e^{\mu(T-t)}, \tag{5.28}$$

  and hence

$$\alpha_t = \frac{\mu}{\sigma^2}\left(e^{\mu(T-t)} - 1\right) + e^{\mu(T-t)}. \tag{5.29}$$

- the case of a mean reverting (Ornstein-Uhlenbock) price process

$$dp_t = \rho(p_0 - p_t)dt + \sigma dW_t \tag{5.30}$$

  in which case:

$$\mathbb{E}\left[p_T \middle| \mathcal{F}_t\right] = p_0 + e^{-\rho(T-t)}(p_t - p_0); \quad Z_t = \sigma e^{-\rho(T-t)}, \tag{5.31}$$

  and hence

$$\alpha_t = -\frac{\rho}{\sigma^2}\left(p_t - p_0\right)^2\left(e^{-\rho(T-t)} - 1\right) + e^{-\rho(T-t)}. \tag{5.32}$$

Unlike in the martingale case, it is hard to obtain any tractable formulas without specifying a functional form for ρ and f. In the case where $\rho(x) = 1/(1+x)$ and $f(x) = 1/(1+x)^2$, the optimal spread becomes

$$s_t = \sigma_t\sqrt{1 + 3\alpha_t} \tag{5.33}$$

Note that m is an increasing function of α_t. To compare with the martingale case, where $\alpha_t = 1$, we therefore want to compare the ratio α_t to 1 to study the impact of the model assumptions on the market maker's profits and inventory volatility.

- For the Black-Scholes model, α_t is larger than 1 for $\mu > 0$. For $\mu < 0$, there exists a critical value depending on T and σ for which this ratio flips sign.
- In the case of an Ornstein-Uhlenbock process, α_t is smaller than 1 iff

$$\left(p_t - p_0\right)^2 < \frac{\sigma^2}{\rho} \tag{5.34}$$

  that is, if the current price p_t isn't too far from the long-term average p_0.

In line with intuition, the market maker quotes larger spreads, expects less profit, and captures less volume in the 'momentum' Black-Scholes model, as compared to the martingale case. In a mean-reverting market, unless the price is significantly away from its long-term trend, the market maker quotes smaller spreads, expects more profit and captures more volume than in the two other market models.

5.3. **Transaction cost analysis and measure of toxicity.** Following the suggestion of [21], one aim of the analysis is to provide macroscopic *analysis tools* of microstructure for LFTs and academics. Not everyone wants to delve into the details of high frequency rules. In this respect, this paper only scratches the surface of the microstructure relationships HFTs can uncover, but it conveniently summarizes them and compares them to the standard 'frictionless' case.

[21] identifies two particular tools that could be of use. One is what the paper calls 'transaction cost analysis', which we interpret to be the analysis of the difference between the effective wealth, and the one that would have been obtained in a frictionless market. Therefore, 'transaction costs' contain two terms:

- the spread component:
$$\pm \int_0^T \frac{s_t l_t}{\sqrt{2\pi}} dt \tag{5.35}$$
This component is positive if using limit orders, and negative if using market orders. Using one or the other affects the Gamma exposure of the trading strategy.
- and the price impact component:
$$[L, p]_T \tag{5.36}$$
which is always of the opposite sign to the spread component.

Depending on the Gamma of the LFT strategy, one or the other term will be the potential source of losses of the trader.

The second tool sought for is a measure of toxicity of the flow of market orders, preferably expressed as an index. Such an index could be used both by market makers to decide on whether it is profitable to provide liquidity and by LFTs to decide whether to execute their trades now or wait for better market conditions. The toxicity of market orders is entirely captured in our framework by the price impact term $[L, p]_T$. Two natural measures of the strength of this price impact term, and hence toxicity of market order flows, are as follows:

- The instantaneous negative correlation
$$\rho_t = -\frac{1}{\sigma_t \ell_t} \frac{d[L, p]_t}{dt} \tag{5.37}$$
between the aggregate provider's inventory and the price. In particular, this could serve as a benchmark for a particular market maker to measure if the flow of market orders he captures is more or less toxic than that of the market as a whole. For the purpose of empirical studies, when working in the discrete trade clock, we compute the discrete time toxicity index as the negative of the empirical correlation of the inventory and the mid-price over the time interval $[0, t]$, namely:
$$\rho_t^{(d)} = -\text{corr}(\Delta L_{[0,t]}, \Delta p_{[0,t]}) \tag{5.38}$$
which is nothing but a plain discretization of formula (5.37).
- The ratio between the integrated price impact and spread components of the aggregate provider's wealth.
$$r = -\sqrt{2\pi} \frac{[L, p]_T}{\int_0^T s_t l_t dt} \tag{5.39}$$

which can be discretized as

$$r^{(d)} = -2\frac{\sum \Delta_n p \Delta_n L}{\sum s_n |\Delta_n L|} \tag{5.40}$$

The market maker in particular holds an implicit option on this quantity: he can pull out of the market if the ratio is larger than 1, as in that case he loses money even in the absence of long term alpha trading by his LFT clients.

The advantage of the first measure of toxicity is that it measures the immediate proportion of toxic versus non-toxic market orders. The disadvantage is that it must be estimated via statistical procedures. The second measure, on the other hand, is more closely related to the actual P&L of a market maker but must be computed over a longer time horizon, making it an ex-post analysis tool.

We give a table illustrating these two measures across several stocks on a same given trading day.

Stock	$\rho^{(d)}$	$r^{(d)}$
AAPL	0.17270704	0.19904208
GOOG	0.23689058	0.32856196
BRCM	0.19237560	0.29776003
CELG	0.26835355	0.48287317
CTSH	0.33887494	0.51758560
CSCO	0.08393210	0.09300757
BIIB	0.27832205	0.40193651
AMZN	0.23614694	0.30494250
GPS	0.20956508	0.48908889
SFG	0.24173454	0.57253111
INTC	0.05301259	0.05574866
GE	0.10889870	0.11888714
JKHY	0.33407745	0.56987813
PFE	0.15849674	0.15958849
CBT	0.34887086	0.74490980
AGN	0.35890531	0.78020785
CB	0.38667565	0.58090719
AA	0.08046277	0.08406282
FPO	0.49598056	1.14964119

TABLE 1. Values of the toxicity indexes on sample stocks.

6. **Continuous equation: general order book shape**

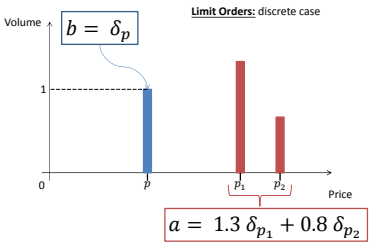
While on our particular choice of stock, most of the trades happened at the best bid or ask price, we wish to generalize our results to a general limit order book. This section starts by formally introducing the notion of limit order book and deriving some basic machinery before going through with the same diffusion limit strategy as section 4.

6.1. **Microscopic description of the order book.** Borrowing from a time-honored method in statistical physics, we first describe in depth the interactions between agents at a microscopic level before deriving effective equations holding at the macroscopic level. We consider a *single* liquidity taker and a *single* liquidity provider. They trade an asset whose possible price range is $(0, \infty)$ via a limit order system. The liquidity provider always moves first by choosing the *limit orders* she places on the limit order book. These limit orders are represented by a control variable (b, a) consisting of a pair of strictly positive measures on $(0, \infty)$. The liquidity taker then chooses the control variable $(\beta, \alpha) \in (0, \infty) \times (0, \infty)$ representing market orders that he wants to execute on that limit order book.

Throughout this section we use the liquidity provider's point of view to track changes in portfolio positions and ignore the following high frequency phenomena:

(1) *Slippage.* Market orders execute immediately at their intended price.
(2) *Partial fills.* Market orders consume all the volume present at a given price[5].
(3) *Hidden orders.* All limit orders are public.

6.1.1. *Basic relationships.* We first focus on *basic* relationships between the two agents, their orders and inventories.



The control (b, a) of the liquidity provider represents her *limit orders*. A bid for one unit of the asset placed at a price p is represented by the probability measure $b = \delta_p$, while an offer (or ask) for one unit at price p' by $a = \delta_{p'}$. If the provider places multiple limit orders, we sum these unit masses and obtain two non-negative measures b and a representing the liquidity provider's aggregate orders.

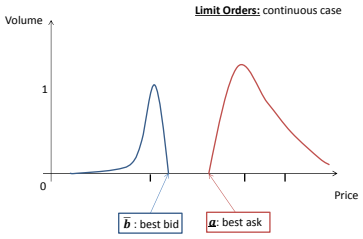
We will call (b, a) a *limit order book*, or order book. We define the best bid and ask of an order book in the following way:

Definition 6.1 (Best bid and ask)**.** *Let (b, a) be an order book. Then we define the best bid and best ask prices to be*

$$\bar{b} = \sup\{p \in supp(b)\}\,, \quad \underline{a} = \inf\{p \in supp(a)\} \tag{6.1}$$

Here we use the notation $supp(\mu)$ for the topological support of the measure μ.

Remark 6.2. *In real markets, such limit orders can only be placed on a discrete grid, and the resulting a and b are always discrete measures. The recent push of high frequency markets to refine their grid may justify considering measures a and b that are absolutely continuous with respect to the Lebesgue measure.*



The control (β, α) of the liquidity taker represents his *market orders*. A market order placed against the bids will cause all the bid orders above and including the price β to be executed. For market orders against the ask, all the limit orders *below*

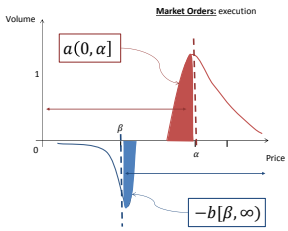
[5]This property automatically holds when you formally consider a continuous order book distribution.

the level α will be executed. The limiting cases $\alpha = 0$ and $\beta = \infty$ correspond to 'empty' market orders that do not execute any limit orders. The execution of a market order leads to the following changes in cash and inventory:

Definition 6.3 (Execution of a market order)**.** *Assume the order book is* (b, a) *and that the liquidity taker chooses the pair* (β, α)*. Then the change* ΔL *of* inventory *triggered by the trade and the change* ΔK *in* cash *that the* liquidity provider *is subject to are defined by:*

$$\Delta L = b[\beta, \infty) - a(0, \alpha] \tag{6.2}$$

$$\Delta K = \int_{(0,\alpha]} xa(dx) - \int_{[\beta,\infty)} xb(dx) \tag{6.3}$$



For the justification of this formula let us first consider a single bid $b = \delta_p$. That is, the provider expresses interest in *buying* one unit of the asset at the price p or lower. A liquidity taker's market order to *sell* will therefore execute the order if and only if its price level β is *smaller*. Should such an execution take place, the liquidity provider gains one unit of volume and loses p units of cash. The above formula is then obtained by aggregating linearly the individual limit orders.

The following assumptions will be used throughout the section.

Assumption 6.4. *The order books* (b, a) *are such that* $\bar{b} < \underline{a}$*, that is, the bid-ask spread is always positive. We will say in this case that the order book* exhibits no arbitrage.

Assumption 6.5. *It is never optimal for the liquidity taker to buy and sell simultaneously.*

In particular, we can recode the liquidity taker's control by a single real number α by making him formally send the market orders (α, α). Indeed, if $\alpha \in (\bar{b}, \underline{a})$ there is no trade, if $\alpha \geq \underline{a}$ a buy happens but no sell, and similarly for $\alpha \leq \bar{b}$.

6.1.2. ***A probabilistic model for liquidity taker behavior.*** We now provide a simple model for which Assumption 6.5 follows automatically from Assumption 6.4. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space modeling the beliefs of the liquidity taker. Let p be a random variable representing the price at which the liquidity taker values the asset at a future time. We assume that the liquidity taker is risk-neutral under $\mathbb{P}$ in the sense that he maximizes his expected wealth after the trade, in other words he solves the optimization problem:

$$\max_{\beta,\alpha} \mathbb{E}\left[-p\Delta L - \Delta K\right] \tag{6.4}$$

Proposition 6.6. *Let the order book* (a, b) *be given. Then an optimal trade for the liquidity taker is given by*

$$\beta = \alpha = \mathbb{E}[p]. \tag{6.5}$$

Proof. The liquidity taker looks for the supremum over $(0,\infty)\times(0,\infty)$ of the function

$$(\beta,\alpha)\longmapsto\int_{[\beta,\infty)}(x-\mathbb{E}[p])b(dx)-\int_{(0,\alpha]}(x-\mathbb{E}[p])a(dx) \tag{6.6}$$

This function decouples and we are left maximizing

$$\beta\longmapsto\int_{[\beta,\infty)}(x-\mathbb{E}[p])b(dx) \tag{6.7}$$

which is non-decreasing on $(0,\mathbb{E}[p]]$ and non-increasing on $[\mathbb{E}[p],\infty)$. The same result holds for

$$\alpha\longmapsto-\int_{(0,\alpha]}(x-\mathbb{E}[p])a(dx) \tag{6.8}$$

The supremum is attained for $\beta=\alpha=\mathbb{E}[p]$. □

Remark 6.7. *While we do not have uniqueness of this maximum, all the other choices of optimum market orders will lead to exactly the same executions. Indeed, the function* $\beta\longmapsto\int_{[\beta,\infty)}(x-\mathbb{E}[p])b(dx)$ *and* $\alpha\longmapsto-\int_{(0,\alpha]}(x-\mathbb{E}[p])a(dx)$ *respectively do not have a strict maximum in* $\mathbb{E}[p]$ *iff* b *and* a *respectively put zero mass on some interval including* $\mathbb{E}[p]$*. Any market orders on this interval will lead to exactly the same cash and asset transfers and we can without loss of generality replace them by market orders at* $\mathbb{E}[p]$*. A similar argument can be made to rule out partial orders. In particular, we can summarize the taker's market orders by a single number* α*.*

Corollary 6.8. *Assume the order book* (b,a) *exhibits* no arbitrage*. Then it is never optimal for the taker to buy and sell simultaneously.*

Proof. By the previous comment, we can summarize the market orders of a taker behaving optimally by a single real α. The taker's buy and sell volumes are

$$a[\underline{a},\alpha]\ \textit{and}\quad b[\alpha,\bar{b}] \tag{6.9}$$

The no arbitrage property implies that these two terms cannot both be positive. □

6.1.3. *Alternative representation of the order book.* Even though the above representation of limit and market orders is clear, we still present an alternative description which only makes sense if no arbitrage is present on the market and Assumption 6.5 is verified.

The below definitions correspond to a very intuitive 'graphic' approach. In the previous section, we have defined the order book as a pair of positive measures (b,a). The no-arbitrage condition guarantees that these two measures have disjoint supports. One is therefore tempted to 'glue' the two measures together into one. But in order to do that, we also need to keep track of where the offers starts and the bids stop. This is done in the following way.

Definition 6.9 (Quoted price)**.** *Let* (b,a) *be an order book that does not exhibit arbitrage. Then we say that* p *is a* quoted price *of the order book if* $p\in(\bar{b},\underline{a})$*.*

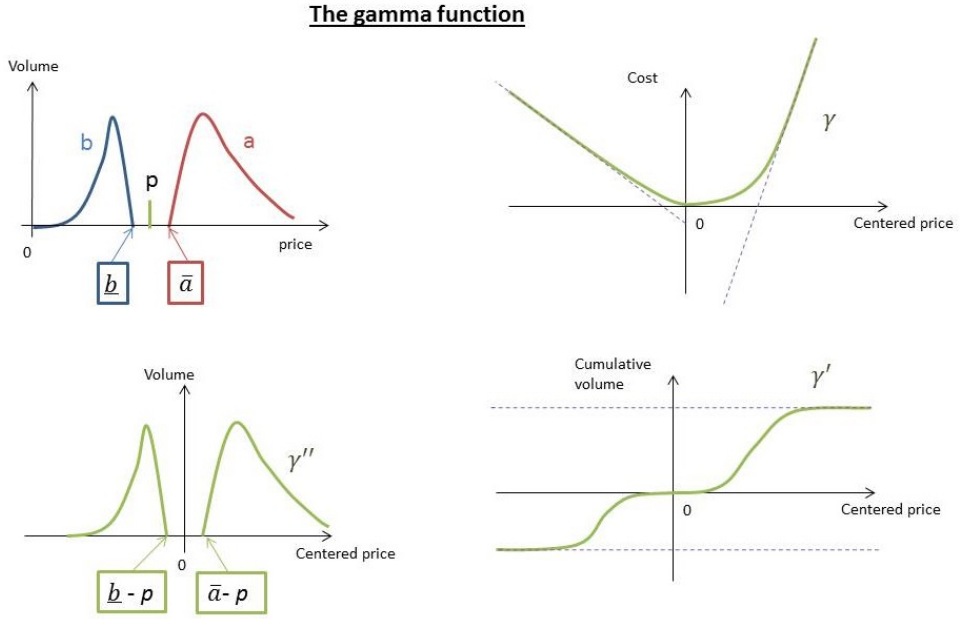
Because the bid-ask spread is positive, there is not a unique quoted price. This is an unfortunate reality of high frequency markets, and we will only be able to mathematically resolve this difficulty in the limit where the bid-ask spread vanishes. Using a quoted price as a separation point between bid and ask limit orders, we can define:

Definition 6.10 (Shape function)**.** *Let (b, a) be an order book that exhibits no arbitrage and p be one of its quoted prices. Then define the order book's* shape *function $\gamma : \mathbb{R} \longmapsto [0, \infty)$ to be*

$$\gamma(u) = \int_0^u (a(0, p + x] - b[p + x, \infty)) \, dx. \tag{6.10}$$

In particular, γ is convex, $\gamma(0) = 0$ and $\gamma'(0) = 0$. Moreover, γ' is bounded and as a result, γ has at most linear growth.

Remark 6.11. *Notice that $\gamma''(\cdot + p) = b + a$ if both measures b and a have densities, or more generally, if we understand this equality in the sense of distributions.*



The following result recasts the trade equations in terms of the function γ.

Proposition 6.12. *Let (b, a) be an order book which exhibits no arbitrage, p be one of its quoted prices and γ the associated shape function. If $\alpha = u + p$ is the liquidity taker's market order, then we have*

$$\Delta L = -\gamma'(u) \tag{6.11}$$

$$\Delta K = (u + p)\gamma'(u) - \gamma(u). \tag{6.12}$$

Proof. The first identity is immediate from the definition of γ and ΔL:

$$\begin{aligned} \Delta L &= b[\alpha, \infty) - a(0, \alpha] \\ &= -\gamma'(u) \end{aligned}$$

The second identity follows using integration by parts:

$$\begin{aligned}\Delta K &= \int_{(0,\alpha]} x a(dx) - \int_{[\alpha,\infty)} x b(dx) \\ &= \alpha\, a(0,\alpha] - \int_{(0,\alpha]} a(0,x]dx - \alpha\, b[\alpha,\infty) + \int_{[\alpha,\infty)} b[x,\infty)dx \\ &= \alpha\gamma'(u) - \gamma(u)\end{aligned}$$

□

Remark 6.13. *The liquidity provider's change in portfolio is captured by the pair* $(\Delta L, \Delta K)$ *comprising her inventory and cash positions. There are multiple ways to denote her change in* wealth*. But if there is no price recovery, then the change in price of the asset after the transaction would be* $\alpha - p$*, and we have:*

$$\begin{aligned}\Delta X &= (\alpha - p)\Delta L + \Delta K \\ &= -\gamma(u)\end{aligned}$$

for the transfer of wealth from the liquidity taker to the liquidity provider. Notice that we used (6.11) *and* (6.12) *to deduce the second equality.* ΔX *is always non-positive by construction of* γ*, and minimal at the quoted price used to define* γ*. As a result, the shape function can be seen as a measure of* adverse selection *the liquidity provider is willing to incur at a given price level if price recovery were non-existent.*

To relate the transaction costs back to the traded volume without going through the transaction price α, we use the following result:

Proposition 6.14. *(Transaction costs) Define the* transaction cost function c *as the Legendre transform of* γ*:*

$$c(l) = \sup_u \left(ul - \gamma(u)\right). \tag{6.13}$$

Then we have:

$$\Delta K = -p\Delta L + c\left(\Delta L\right), \tag{6.14}$$

and in particular, c *is convex and satisfies* $c(0) = 0$*.*

Proof. By the Fenchel identity, we have that

$$u\gamma'(u) = \gamma(u) + c(\gamma'(u))$$

and that c' is the generalized inverse of γ'. Hence, as $\Delta L = \gamma'(u)$ we have that $u = c'(\Delta L)$ and

$$\begin{aligned}\Delta K &= -p\Delta L + u\gamma'(u) - \gamma(u) \\ &= -p\Delta L + c(\Delta L)\end{aligned}$$

□

An order book (b, a) which *does not exhibit arbitrage* can therefore be represented by a pair (p, γ) with p a real and γ a differentiable, convex function with linear growth satisfying $\gamma(0) = \gamma'(0) = 0$. Note that this representation in terms of *quoted price* and order book *shape* is *not unique*, but leads to a completely equivalent description of trades and hence the same market model.

Both representations have pros and cons and unfortunately, both will need to be juggled at different times of our analysis. The advantages of the original (b, a) representation are: *uniqueness* of the decomposition, ease to derive *no-arbitrage*

relationships and *natural* interpretation of formulas. The alternative representation in terms of (p, γ) is *more tractable* and *concise* as it involves a real number and a function rather than a pair of measures.

6.1.4. *Summary.* For future reference, we summarize the different trade equations and market representations defined and derived in this section.

The liquidity provider places limit orders. If the *limit order book* formed that way presents no arbitrage, it will be represented either by a pair of measures (b, a) or a couple (p, γ) with p a real number and γ a differentiable, convex function with linear growth and $\gamma(0) = \gamma'(0) = 0$. Consistency equations between the two representations can be found above.

We call (b, a) the *order book*, p a *quoted price* and γ the *shape* of the order book. The liquidity taker's *market order* will be represented either by a real α representing a *price*, or a real u denoting a *centered* price (shifted by the quoted price p). Both representations lead to the same trades.

$$\begin{aligned} \Delta L &= b[\alpha, \infty) - a(0, \alpha] \\ &= -\gamma'(u) \end{aligned}$$

is the change in inventory of the liquidity provider, while

$$\begin{aligned} \Delta K &= \int_{(0,\alpha]} x a(dx) - \int_{[\alpha,\infty)} x b(dx) \\ &= (u + p)\gamma'(u) - \gamma(u) \\ &= p\Delta L + c\left(\Delta L\right) \end{aligned}$$

is her change in cash position.

The market order corresponding to this trade can be recovered from the limit orders and the trade volume by the relationship

$$\alpha - p = c'(-\Delta L) \tag{6.15}$$

and this is the price impact in the absence of price recovery.

6.2. **Discrete self-financing equation and other relationships.** We now give ourselves a discrete price process p and provider inventory process L. Just as in the bid-ask spread case, three necessary conditions can be derived.

6.2.1. *Self-financing equation.*

$$\Delta X = L\Delta p + c\left(\Delta L\right) + \Delta p \Delta L \tag{6.16}$$

6.2.2. *Price impact.*

$$\Delta p \Delta L \leq 0 \tag{6.17}$$

6.2.3. *Price recovery.*

$$|\Delta p| \leq |c'(-\Delta L)| \tag{6.18}$$

6.3. **Macroscopic limit.** The strategy in this section is identical to that of section 4. We start off with the data of our problem in continuous time, discretize it to apply the discrete relationships derived earlier and finally take the diffusion limit to obtain our continuous time relationships.

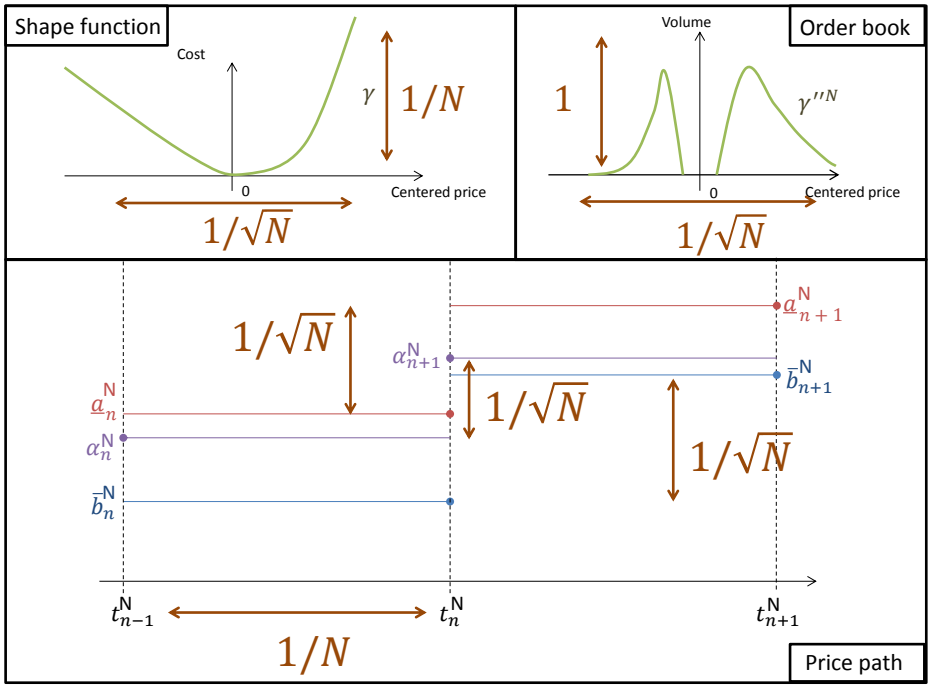


FIGURE 6. Renormalization of the model for the diffusion limit. Time is scaled by $1/N$, prices by $1/\sqrt{N}$ and volume by 1 (unchanged). For example, the y-axis of γ represents cost, that is $[volume] \cdot [price]^2$ which scales in $1/N$. The x-axis is expressed in prices and is scaled in $1/\sqrt{N}$, leading to the formula $\gamma^N(\cdot) = \gamma(\sqrt{N}\cdot)/N$.

6.3.1. *Approximation procedure.* Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a filtered probability space supporting a Wiener process (W, W') with unspecified correlation structure. We consider a fixed time interval $[0, 1]$ and give ourselves the following $\mathcal{F}$-adapted processes for the price and inventory of a provider:

$$\begin{cases} p_t &= p_0 + \int_0^t \mu_u du + \int_0^t \sigma_u dW_u \\ L_t &= L_0 + \int_0^t b_u du + \int_0^t l_u dW'_u \end{cases} \tag{6.19}$$

where p_0 and L_0 are $\mathcal{F}_0$-measurable elements of L^2 and μ, σ, b and l are $\mathcal{F}$-adapted and càdlàg processes. Finally, let $c : \Omega \times [0, 1] \times \mathbb{R} \to \mathbb{R}^d$ be a random, $\mathcal{F}_t$-adapted function that is C^0 in (t, l). Assume c to be a.s. convex for all t, with a minimum at $c_t(0) = 0$ and such that $c_t(l) < Cl^2$ for some constant C. We denote by γ_t its Legendre transform, which will represent the shape function of the order book as measured in tick size.

Let $\frac{1}{\sqrt{N}}$ be a vanishing tick size. Define the discretized price process as $p^N_n = p_{n/N}$ and likewise L^N.

We propose the following choice of renormalization for the order book.

$$\gamma_n^N(x) = \frac{1}{N}\gamma_{n/N}\left(\sqrt{N}x\right) \tag{6.20}$$

This in particular implies

$$c_n^N(l) = \frac{1}{N}c_{n/N}\left(\sqrt{N}l\right) \tag{6.21}$$

This follows from the fact that γ is defined in tick size and needs to be renormalized appropriately in the discrete approximation, where we want γ^N to be expressed in absolute terms.

6.3.2. *Main result.*

Theorem 6.15. *The continuous time relationships between provider wealth X, inventory L, price p and transaction costs c are:*

$$\begin{cases} dX_t = L_t dp_t + \Phi_{l_t}(c_t)dt + d[L,p]_t \\ d[L,p]_t \leq 0 \\ \sigma_t^2 \leq \Phi_{l_t}((c_t')^2) \end{cases} \tag{6.22}$$

where $X_t = \lim_{N\to\infty} X^N_{\lfloor Nt \rfloor}$ u.c.p.

Proof. Just as in the bid-ask spread case, the result to prove is the u.c.p. convergence of

$$\frac{1}{N}\sum_{n=1}^{\lfloor tN \rfloor} c_{n/N}\left(\sqrt{N}\Delta_n L^N\right) \tag{6.23}$$

and

$$\frac{1}{N}\sum_{n=\lfloor t_1 N \rfloor}^{\lfloor t_2 N \rfloor} \left(\Delta_n p^N\right)^2 - \left(c'_{n/N}\left(\sqrt{N}\Delta_n L^N\right)\right)^2 \tag{6.24}$$

to the integrals

$$\int_0^t \Phi_{l_u}(c_u)du \tag{6.25}$$

and

$$\int_{t_1}^{t_2} \left(\sigma_u^2 - \Phi_{l_u}((c'_u)^2)\right) du \tag{6.26}$$

This is a direct application of theorem 4.2. □

7. Naive supply and demand model

The aim of this section is to illustrate how a model for limit order fill rates and exact price recovery leads to models of the price as a function of trade volumes, or vice versa. This therefore models supply and demand in high frequency markets and closes the loop of our endeavor. However, we do not believe these models to be as accurate as the previously derived relationships and only use this section for illustrative purposes.

7.1. **Microscopic assumptions.** The proposed model is: *perfect* fill rate and *deterministic* price recovery.

7.1.1. *Disclaimer.* Unlike for the other microscopic relationships checked empirically in Section 3, the model considered now is *not always consistent with empirical data*. Fill rates are definitely *not* one, and price recovery is *not* deterministic.

7.1.2. *Setup.* Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and p and L be two discrete time processes representing the price of the market and the inventory of a liquidity provider respectively. Let γ be a C^3-function valued discrete time process representing our provider's shape function and c its associated transaction costs.

7.1.3. *Additional relationships.* We translate 'perfect fill rate' and 'deterministic price recovery' by the following equation:

$$\Delta p = \lambda c'(-\Delta L) \tag{7.1}$$

or, equivalently

$$\Delta L = -\gamma'(\lambda^{-1}\Delta p) \tag{7.2}$$

where $\lambda \in (0, 1]$ is a real that encapsulates price recovery. The bigger λ, the smaller the price recovery.

7.2. **Macroscopic limit.** Equation (7.1) allows a liquidity provider to derive the price from trade volumes and the order book, while equation (7.2) derives the trade volumes from the prices and the order book. Both lead to the same consistency relationships between p, L and γ in the continuous limit.

7.2.1. *Main tool.* The proof method is based on another result from [25]. We first summarize the hypothesis and result before imposing them on the data of our problem.

Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a filtered probability space supporting an 1-dimensional Wiener process W and Y be a 1-dimensional Itô process of the form

$$Y_t = Y_0 + \int_0^t b_t dt + \int_0^t \sigma_t dW_t \tag{7.3}$$

where we consider $t \in [0, 1]$.

Assumption 7.1. *(H)+ (K) from* [25]

Assume that b_t and σ_t are progressively measurable, b_t is locally bounded and σ_t is càdlàg.

Let now $F : \Omega \times [0, 1] \times \mathbb{R} \to \mathbb{R}$ be a random, $\mathcal{F}_t$-adapted function that is C^1 in y and C^0 in (t,y). We will shorten the notation to $y \mapsto F_t(y)$. Define the following assumption.

Assumption 7.2. *(7.2.1), (10.3.2), (10.3.3), (10.3.4) and (10.3.7) from* [25]

Assume that a.s. for all t, F_t is an odd *function in y.*

Furthermore, assume there exists a function $g : \mathbb{R} \to \mathbb{R}$ with polynomial growth and a real $\beta > 1/2$ such that, for all $\omega \in \Omega$, $(t, s) \in [0, 1]^2$ and $y \in \mathbb{R}$:

$$|F_t(y)| \leq g(y)$$
$$|F'_t(y)| \leq g(y)$$
$$|F_t(y) - F_s(y)| \leq g(y)|t - s|^\beta$$

Let us now state the new result from [25] we will use.

Theorem 7.3. *(10.3.2) from* [25] *Assume 7.1 and 7.2. Then there exists a very good filtered extension of the original space such that we have the following stable convergence in law as $N \to \infty$:*

$$\frac{1}{\sqrt{N}} \sum_{n=1}^{\lfloor Nt \rfloor} F_{n/N} \left(\sqrt{N}(X_{(n+1)/N} - X_{n/N}) \right) \to U_t$$

where

$$U_t = \int_0^t b_s \Phi_{\sigma_s} (F'_s)\, ds + \int_0^t \sqrt{\Phi_{\sigma_s} ((F_s)^2)} dW'_s \tag{7.4}$$

with W'_t a d-dimensional Wiener process such that

$$[W', W]_t = \int_0^t \frac{\Phi_{\sigma_s} \left(id\, F_s^k \right)}{\sigma_s \sqrt{\Phi_{\sigma_s} (F_s^k)^2}} ds$$

where id is the identity function.

7.2.2. *Continuous time setup.* Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a filtered probability space supporting a Wiener process W_t. We will fix either an Itô process

$$p_t = p_0 + \int_0^t \mu_s ds + \int_0^t \sigma_s dW_s \tag{7.5}$$

for the price or

$$L_t = L_0 + \int_0^t b_s ds + \int_0^t l_s dW_s \tag{7.6}$$

for the inventory.

In addition to one of these processes, we also fix an order book shape process γ_t and denote by c_t the associated transaction cost process.

Assume L (respectively p) verifies Assumption 7.1 and c (respectively γ) satisfies Assumption 7.2.

Just as previously, we define the discretized processes $L_n^N = L_{n/N}$ (respectively $p_n^N = p_{n/N}$) and $c_n^N(\cdot) = \frac{1}{N} c_{n/N} \left(\sqrt{N} \cdot \right)$ (respectively $\gamma_n^N(\cdot) = \frac{1}{N} \gamma_{n/N} \left(\sqrt{N} \cdot \right)$).

7.2.3. *Main result.* The main result is a straightforward application of Theorem 7.3. If we are given the inventory L and transaction costs c then we have:

Theorem 7.4. *There exists a very good filtered extension of the original space such that we have the stable convergence in law $p^N_{\lfloor Nt \rfloor} \to p_t$ with*

$$dp_t = -\lambda b_t \Phi_{l_t} (c''_t)\, dt + \lambda \sqrt{\Phi_{l_t} ((c'_t)^2)} dW'_t \tag{7.7}$$

where

$$[W', W]_t = -\int_0^t \frac{\Phi_{l_s} (id\, c'_s)}{l_s \sqrt{\Phi_{l_s} ((c'_s)^2)}} ds. \tag{7.8}$$

In particular,

$$d[p, L]_t = -\Phi_{l_t} (id\, c'_t)\, dt \tag{7.9}$$

A completely equivalent result is obtained if the price p and order book shape function γ are given:

Theorem 7.5. *There exists a very good filtered extension of the original space such that we have the stable convergence in law* $L^N_{\lfloor Nt \rfloor} \to L_t$ *with*

$$dL_t = -\mu_t \Phi_{\sigma_t}\left(\gamma_t''(\lambda^{-1}\cdot)\right) dt + \sqrt{\Phi_{\sigma_t}((\gamma_t')^2(\lambda^{-1}\cdot))} dW_t' \tag{7.10}$$

where

$$d[p, L]_t = -\Phi_{\sigma_s}\left(id\, \gamma_t'(\lambda^{-1}\cdot)\right) dt. \tag{7.11}$$

7.3. **A special case.** A flat order book corresponds to $\gamma_t'' = m_t$ for some adapted process m. While quite unrealistic, it is *extremely* tractable and has been proposed and used in other models ([2, 33]).

This corresponds to quadratic transaction costs and *linear* price impact:

$$\begin{cases} dp_t &= -\frac{\lambda}{m_t} dL_t \\ dX_t &= L_t dp_t + \left(\frac{1}{2} - \lambda\right) \frac{l_t^2}{m_t} dt \end{cases} \tag{7.12}$$

Note that the the sign of the effective transaction costs is that of $\frac{1}{2} - \lambda$. Indeed, in the self-financing case $\lambda = \frac{1}{2}$, price recovery and price impact perfectly cancel each other out. If $\lambda > \frac{1}{2}$, then the price impact of trades is stronger than the collected spread because of insufficient price recovery. Also, because of the uniform structure of the order book and perfect fill rate, the inventory of the provider is perfectly anti-correlated to the price.

7.3.1. *The worst case for providers.* As we have seen before, perfect anti-correlation is the worst case for the liquidity provider, making the uniform order book 'the worse' shape from the liquidity provider's perspective. Amongst uniform order books, absence of price recovery, $\lambda = 1$ is the wost case scenario.

A cute result is that if the liquidity provider provides constant liquidity ($m_t = 1$) then we have the following identity between wealth and inventory:

$$X_t = X_0 - L_t^2 + L_0^2 \tag{7.13}$$

that is, even with the most naive strategy in the worst case scenario, the liquidity provider does not lose money if she manages her inventory. Symmetrically, one can show that, even in this best case scenario for liquidity takers, there are no round-trip statistical arbitrage opportunities due to price impact only.

8. **Conclusions**

In conclusion, the present paper identifies key features of *high frequency* limit order book markets and derives corresponding *necessary* conditions on self-financing portfolios for continuous-time models of such markets. These features are:

(1) *Non-smoothness* of inventories of high frequency traders and *vanishing* bid-ask spread in high frequency markets.
(2) *Adverse selection* as given by a negative quadratic covariation between price increments and change in provider inventory, which is a consequence of the *price impact* of trades on such time-scales.
(3) *Price recovery* and the way it links the bid-ask spread and price volatility processes.
(4) Generalized formula for the wealth process of *a self-financing portfolio* when including price impact.

(5) Applications to option hedging and portfolio optimization highlighting the differences between trades via market orders and limit orders, and the differences between liquidity providers and liquidity takers.

These features were obtained by studying, both theoretically and empirically, high frequency market microstructure before summarizing it on a macroscopic level. As pointed out by [21], the crucial technical tool was the use of an *event-based* clock. We hope further research will follow this method to uncover more effects of HFT on the broader financial system.

Appendix A. **Cross-sectional analysis**

The main empirical claim of the paper is the *negative covariation* between liquidity provider inventory and the price process. This is one of many ways of identifying price impact, and is due to adverse selection of limit orders by liquidity takers. We wish to test this on a sample of stocks to identify when this relationship is verified, and when not. The data used in this appendix are 29 large cap stocks using Nasdaq ITCH data on 18/04/13. Other days and stocks have been tested with similar results.

This test will come in three forms, from the most intuitive to the most sophisticated.

We first begin by listing for each of our 29 stocks the proportion of trades not satisfying the property $\Delta L \Delta p \leq 0$.

Then we plot the empirical quadratic covariations with confidence intervals constructed using the functional central limit theorem [1] for continuous Itô processes.

Finally, we set up a rigorous statistical test based on the same functional central limit theorem. In the last case, we assume that we are given two continuous Itô processes L and p such that:

$$\begin{cases} dp_t &= \mu_t dt + \sigma_t dW_t \\ dL_t &= b_t dt + l_t dW'_t \end{cases} \tag{A.1}$$

with the quadratic covariation between W_t and W'_t being ρ_t. Assume furthermore that μ_t and b_t to be locally bounded and that σ_t, l_t and ρ_t are càdlàg.

If we then denote by p^N and L^N the discrete measurements of these processes on the uniform grid $\{1/N, 2/N, ..., 1\}$ then [1] tells us to consider the discrete processes:

$$\begin{cases} C_t^N &= \sum_{n=1}^{\lfloor Nt \rfloor - 1} \Delta_n p^N \Delta_n L^N \\ V_t^N &= N \sum_{n=1}^{\lfloor Nt \rfloor - 2} \left(\left(\Delta_n p^N \Delta_{n+1} L^N \right)^2 + \Delta_n p^N \Delta_n L^N \Delta_{n+1} p^N \Delta_{n+1} L^N \right) \end{cases} \tag{A.2}$$

and we have the functional central limit theorem

$$\mathcal{L} \left(\frac{C_t^N - [p, L]_t}{\sqrt{N^{-1} |V_t^N|}} \right) \to N(0, 1) \tag{A.3}$$

This allows the construction of confidence intervals for the quadratic covariation process. We also use this result to reject the following *null hypothesis*:

Assumption A.1. *There exists $t \in [0, 1]$ such that $\rho_t > 0$.*

by constructing confidence intervals for the quadratic covariation on small time intervals $[t_k, t_{k+1})$, we can compute rejection probabilities for the events $\rho_{t_k} > 0$ for each t_k. By multiplying these rejection probabilities, we obtain the rejection

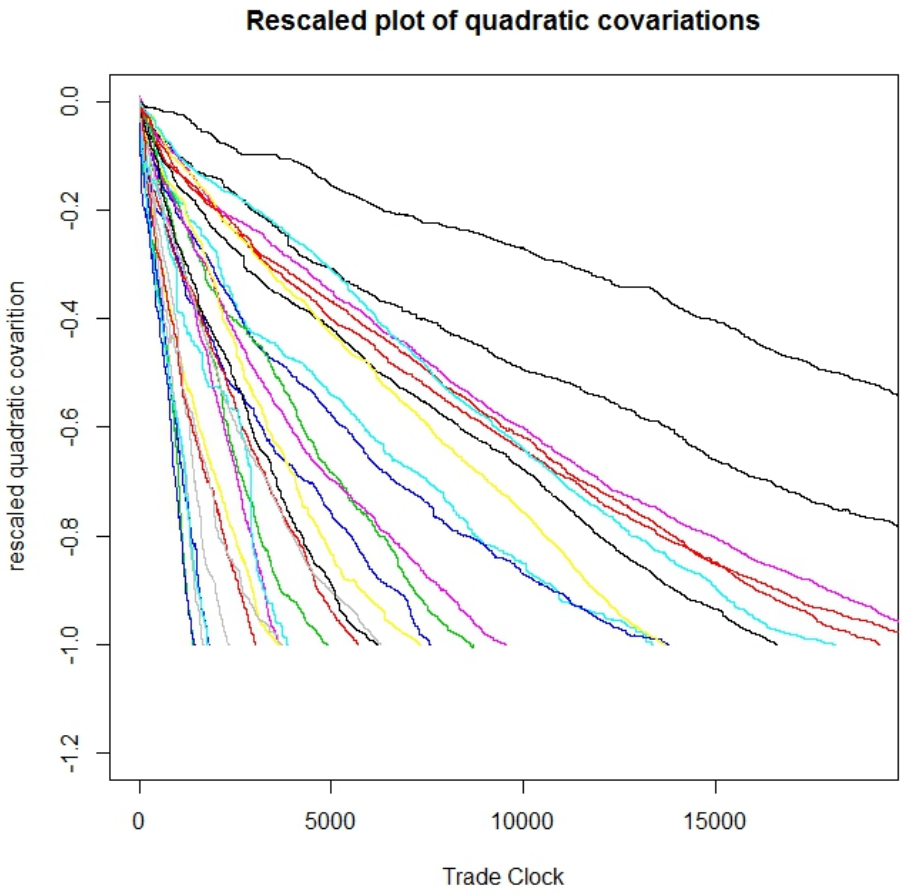


FIGURE 7. Empirical quadratic covariations (rescaled).

probability for our overall null hypothesis. Our choice of time intervals $[t_k, t_{k+1}]$ is such that we have 100 data points in each of these intervals.

Finally, we obtain the tables:

Stock	proba reject	nb false	nb trades	percent false	recovery rejection
MSFT	0.7868301	19	27540	0.06899056	6.147422
KO	0.9876695	72	20362	0.3535998	13.932816
BA	0.9999383	222	4824	4.60199	24.212272
GPS	0.9999044	97	7378	1.314719	22.445107
GE	0.9991448	4	12969	0.03084278	6.847097
CS	0.8971721	132	3621	3.645402	37.448219
CPB	0.9421457	129	3578	3.605366	26.914477
BCS	0.9625842	43	1613	2.66584	27.774334
JNJ	0.9550316	152	16114	0.9432791	19.777833
UPS	0.9983282	237	5608	4.226106	30.117689
CLX	0.9563385	118	1381	8.544533	31.643736
T	0.9996831	27	13287	0.2032061	12.139685
DELL	0.9893074	1	3742	0.02672368	5.130946
XOM	0.9998707	340	20714	1.641402	19.276818
CAT	0.9814122	397	13456	2.950357	26.575505
COF	0.8973841	131	6103	2.146485	27.117811
AAPL	0.9999987	2347	46710	5.02462	9.648897
PG	0.9998587	189	18616	1.015256	18.038247
GOOG	0.9929220	609	8595	7.085515	15.602094
HSY	0.9615380	177	1807	9.795241	35.030437
WFC	0.9129410	13	17672	0.0735627	11.854912
DTV	0.6174753	117	9334	1.253482	21.952003
BBY	0.9999374	85	7181	1.183679	22.113912
MT	0.8870935	18	2273	0.791905	21.293445
GM	0.9774693	19	5963	0.3186316	18.279390
CL	0.9833529	187	3006	6.220892	24.550898
MA	0.9996761	113	1435	7.874564	18.048780
KSU	0.9945635	118	1756	6.719818	26.765376
GIS	0.9735843	68	3624	1.87638	22.323400

TABLE 2. Rejection probability of the null hypothesis, number of trades not satisfying our main inequality, total number of trades, percentage of trades not verifying our main inequality and percentage of trades not verifying our price recovery inequality. We also noted that all the lit trades across all the stocks happened at the best bid and best ask. Note that, of all our proposed relationships, the only weak one is price recovery, which is routinely violated.

REFERENCES

[1] Y. Ait-Sahalia and J. Jacod. *High-Frequency Financial Econometrics*. Princeton University Press, 2014.

[2] A. Alfonsi, A. Fruth, and A. Schied. Optimal execution strategies in limit order books with general shape functions. *Quantitative Finance*, 10(2):143–157, 2010.

[3] C. Aliprantis and K. Border. *Infinite Dimensional Analysis*. Springer, 2006.

[4] R. Almgren and N. Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3(2):5–39, 2000.

[5] Y. Amihud and H. Mendelson. Asset pricing and the bid-ask spread. *Journal of Financial Economics*, 17(2):223–249, 1986.

[6] T. Ane and H. Geman. Order flow, transaction clock, and normality of asset returns. *The Journal of Finance*, 55(5):2259–2284, 2000.
[7] M. Avellaneda and S. Stoikov. High-frequency trading in a limit order book. *Quantitative Finance*, 8(3):217–224, 2007.
[8] B. Biais, P. Hillion, and C. Spatt. An empirical analysis of the limit order book and the order flow in the paris bourse. *Journal of Finance*, 50(5):1655–89, 1995.
[9] J. . Bouchaud, Y. Gefen, M. Potters, and M. Wyart. Fluctuations and response in financial markets: The subtle nature of 'random' price changes. *Quantitative Finance*, 4(2):176–190, 2004.
[10] J. . Bouchaud, M. Mzard, and M. Potters. Statistical properties of stock order books: Empirical results and models. *Quantitative Finance*, 2(4):251–256, 2002.
[11] J. Brogaard, T. Hendershott, and R. Riordan. High frequency trading and price discovery. Technical report, ECB, Working Papers Series, 2013.
[12] A. Chakraborti, I. Muni Toke, M. Patriarca, and F. Abergel. Econophysics: Empirical facts and agent-based models. *Quantitative Finance*, 2009.
[13] T. Chellathurai and T. Draviam. Dynamic portfolio selection with nonlinear transaction costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 461(2062):3183–3212, 2005.
[14] P.K. Clark. A subordinated stochastic process model of cotton futures prices. *Harvard University*, unpublished Ph.D. dissertation, 1970.
[15] P.K. Clark. A subordinated stochastic process model with finite variance for speculative prices. *Econometrica*, 41(1):135–155, 1973.
[16] R. Cont and a. de Larrard. Order book dynamics in liquid markets: limit theorems and diffusion approximations. *Working paper*, 2011.
[17] R. Cont and a. de Larrard. Price dynamics in a markovian limit order book market. *SIAM Journal for Financial Mathematics*, 4(1):1–25, 2013.
[18] R. Cont, S. Stoikov, and R. Talreja. A stochastic model for order book dynamics. *Operations Research*, 58(3):549–563, 2010.
[19] D. Easly, M. Lopez de Prado, and M. O'Hara. The microstructure of the flash crash: flow toxicity, liquidity crashes, and the probability of informed trading. *Journal of Portfolio Management*, 2011.
[20] D. Easly, M. Lopez de Prado, and M. O'Hara. Flow toxicity and liquidity in a high-frequency world. *Review of Financial Studies*, 2012.
[21] D. Easly, M. Lopez de Prado, and M. O'Hara. The volume clock: insights into the high frequency paradigm. *Journal of Portfolio Management*, 2012.
[22] M. B. Garman. Market microstructure. *Journal of Financial Economics*, 3(3):257–275, 1976.
[23] J. Hasbrouck. *Empirical market microstructure*. Oxford University Press, 2007.
[24] T. Ho and H.R. Stoll. Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial Economics*, 9(1):47–73, 1981.
[25] J. Jacod and P. Protter. *Discretization of Processes*. Springer, 2011.
[26] A. S. Kyle. Continuous auctions and insider trading. *Econometrica*, 53(6):1315–1335, 1985.
[27] R. Liu and J. Muhle-Karbe. Portfolio choice with stochastic investment opportunities: a user's guide. Technical report, Proc. 1st Princeton Summer School in Mathematical Finance, 2013.
[28] M. Magill and G. Constantinides. Portfolio selection with transactions costs. *Journal of Economic Theory*, 13(2):245 – 263, 1976.
[29] B. Mandelbrot. Comments on 'a subordinated stochastic process model with finite variance for speculative prices by peter k. clark'. *Econometrica*, 41(1):157–159, 1967.
[30] B. Mandelbrot and M. Taylor. On the distribution of stock price differences. *Operations Research*, 15(6):1057–1062, 1967.
[31] S. Maslov. Simple model of a limit-order driven market. *Physica A*, 278:571–578, 2000.
[32] S. Maslov and M. Mills. Price fluctuations from the order book perspective empirical facts and a simple model. *Physica A*, 299:234–246, 2001.
[33] A. Obizhaeva and J. Wang. Optimal trading strategy and supply/demand dynamics. *Preprint*, 2005.
[34] M. O'Hara. *Market microstructure theory*. Basil Blackwell, 1995.
[35] M. O'Hara and G. Oldfield. The microeconomics of market making. *Journal of Financial and Quantitative Analysis*, 21:361–376, 1986.

[36] S.E. Shreve and H.M. Soner. Optimal investment and consumption with transaction costs. *Annals of Applied Probability*, 4(3):609–692, 1994.
[37] S. Stoikov and M. Saglam. Option market making under inventory risk. *Review of Derivatives Research*, 12(1):55–79, 2009.
[38] P. Weber and B. Rosenow. Order book approach to price impact. *Quantitative Finance*, 5(4):357–364, 2005.
[39] M. Wyart, J. . Bouchaud, J. Kockelkoren, M. Potters, and M. Vettorazzo. Relation between bid-ask spread, impact and volatility in order-driven markets. *Quantitative Finance*, 8(1):41–57, 2008.
[40] I. Zovko and J. D. Farmer. The power of patience: A behavioral regularity in limit order placement. *Quantitative Finance*, 2(5):387–392, 2002.