International Journal of Financial Engineering
Vol. 2, No. 3 (2015) 1550029 (30 pages)

DOI: 10.1142/S2424786315500292

World Scientific
www.worldscientific.com

Patience vs. impatience of traders: Formation of the value-at-price distribution through competition for liquidity

Peter Lerner

SciTech Analytical Services, LLC Woodland Drive
State College, PA 16803, United States

Received: 19 May 2015; Revised: 18 August 2015; Accepted: 18 August 2015
Published: 30 September 2015

Abstract

The ability to postpone one's execution in the market without penalty in search of a better price is an important strategic advantage in high-frequency trading. To elucidate competition between traders one has to formulate to a quantitative theory of formation of the execution price from market expectations and quotes. Equilibrium theory was provided in 2005 by Foucault, Kadan and Kandel. I derive an asymptotic distribution of the bids/offers as a function of the ratio of patient and impatient traders using the dynamic version of the Foucault, Kadan and Kandel limit order book (LOB) model. Our version of the LOB model allows stylized, but sufficiently realistic representation of the trading markets. In particular, dynamic LOB allows simulation of the distribution of execution times and spreads from high-frequency quotes. Significant analytic progress is made toward framing of short-term trading as competition for immediacy of execution between traders under imperfect information. The results are qualitatively compared with empirical volume-at-price distribution of highly liquid stocks.

Keywords: Limit order book; Foucault, Kadan and Kandel model; bid-ask spread; market micro-structure.

JEL Classifications: G14; G12; C63.

1. Introduction

The propagation of high-frequency trading (HFT), along with the algorithmic executions and extreme events usually associated with HFT such as the "flash crash" of May 6, 2010 and several interruptions of Nasdaq trading in 2013, have

Email address: pblerner@syr.edu

brought financial microstructure studies from the backburner to the forefront of economic theory.[1] However, precise theoretical reasons for abnormal reactions of electronic trading systems in quickly changing supply/demand conditions, as well as their responses to changing beliefs and preferences, remain elusive. Thus, it is important to have dynamic equations, however simplified, to be able to model transitional regimes in the formation of bid-ask prices within trading systems.

Currently, a large number of microstructure models, each with their own level of realism, empirical accuracy and analytic tractability, are in existence. In general, microstructure models are difficult subjects for empirical verification because they frequently express unobservable parameters (such as asset volatility or microstructure noise) in terms of other unobservable parameters such as the fraction of informed traders (O'Hara, 1995; Hasbrouck, 2007; Lerner, 2009). Yet, they are necessary if we are ever to move beyond random walk prices and perfectly efficient markets. The report of the Swiss Finance Institute proclaims: "One of the most pressing subjects is to come up with a realistic agent-based model, where crisis and complexity arise from simple rules and interactions in a universal way." (Sornette and von der Becke, 2011).

The simplest "behavioral" effect in market microstructure is the distinction between patient and impatient traders. It is well known that, in the absence of changing economic fundamentals, the traders who have freedom to postpone their execution receive a more favorable execution price than those who have no such opportunity (Harris, 2003). To elucidate this effect of micro-liquidity, one has to resort to a quantitative theory of formation of the terminal execution price from market expectations and quotes.

In this paper, we propose a theory based on the general assumption that trading can be represented as a random walk of successive bids (offers) over the state space of LOB. This insight allows the author to replace a very complicated "negotiation" process price with a family of 1D random walks. This approach was pioneered by Obizhaeva and Wang (2005) in the context of optimal execution of quotes. Our approach is completely independent of Wang and Obizhaeva, but instead makes a dynamic extension of the static model provided in the same year 2005 by Foucault, Kadan and Kandel (FKK).

This paper is structured as follows. Section 2 is a literature review. Sections 3 and 4 discuss empirical data on volume-at-average-price (VWAP) distributions. Section 5 is an exposition of the FKK theory. Section 6 is my own dynamic extension of the FKK model. Section 7 compares simulated distributions with

[1]For instance, the preprint of "Flash Crash: Flow Toxicity, Liquidity Crashes and the Probability of Informed Trading" was downloaded from ssrn.com more than 15,000 times by January 21, 2013 from the date of its release in 2011 (Easley *et al.*, 2011).

empirical VWAPs. Section 8 discusses biological and social analogies of the proposed theory. In Sec. 9, we comment on the numerical size of the dynamical limit order book effects. The paper ends with a conclusion and has two appendices dealing with analytical features of the model.

2. Literature Review

There are several, mutually complementary approaches to market microstructure. One of them is a dynamic modeling of the limit order book (LOB), which is described in Chapter 8 of the book by De Jong and Rindi (2010). The most analytically tractable of this class of models has been proposed by Foucault, Kadan and Kandell in 2005 (Foucault, Kadan and Kandell, further FKK 2005). FKK can be placed in a class of barrier-diffusion models first proposed by Harris (1998). In this class of model, the order submission evolves as diffusion and the limit orders execute when the price hits a barrier (Hasbrouck, 2007).

The FKK model describes LOB in terms of two state variables: execution time T_h and the proportion of so-called "patient" and "impatient" traders, θ_P and $(1 - \theta_P)$. In the FKK model, all trading is in immediacy. Namely, patient traders benefit at the expense of impatient traders who cannot wait for a more favorable price for execution of their trades.

This heuristic and intuitive picture of trading is supported by the empirical results of Menkveld (2010) on the HFT. Namely, Menkveld identified the role of HFT, i.e., supposedly well-informed and well-connected traders as a substitute, electronic version of the market maker (see also Foucault, 2012). That line of research finds its further development in Jovanovic and Menkveld's (2011) study of HFT as latter-day middlemen.

Execution time is difficult to observe and the proportion of impatient traders is completely unobservable. While, recently, exchanges and "dark pools" started to provide information on all limit orders with their time stamps on a proprietary basis, this information is unlikely to be available soon for commodities, which are mostly transacted on a party-to-party basis (see e.g., Skjeltorp *et al.*, 2012). Skjeltorp *et al.* (2012) performed the first empirical tests of FKK. But even they acknowledged that direct clarification of liquidity externalities is "extremely challenging." In real markets, liquidity is determined endogenously—while in most models it can be introduced as a functional of the state variables—and there are few good instruments with which one can identify this functional (Barclay and Henderschott, 2004; Henderschott and Jones, 2005).

Skjeltorp *et al.* (2012, further SST) built a complete LOB from Nasdaq OMX BX data with millisecond accuracy. Thus, in view of the results of SST, the

necessity for a simplified dynamic model as in this paper might seem obviated, but there are extenuating circumstances.

First, the increased proliferation of the "dark pools" as alternative venues for execution (see e.g., Angel *et al.*, 2010) makes it unlikely that LOBs of most markets will become directly accessible. Furthermore, recent studies (Switzer and Fan, 2010) demonstrate that the most frequently used measures of trading costs such as bid-ask spreads and volumes (but not the number of trades) have very low instantaneous correlation with the actual cost of trading. In the view of the Switzer and Fan study, it seems advantageous to develop an endogenous theory of formation of the bid-ask spread, which is independent of the empirical metrics involved.

Second, one cannot exclude the emergence of derivative instruments contingent on liquidity. Pricing of such securities can benefit from closed-form analytical equations. Finally, a relatively parsimonious analytic model can be important despite oversimplifications.

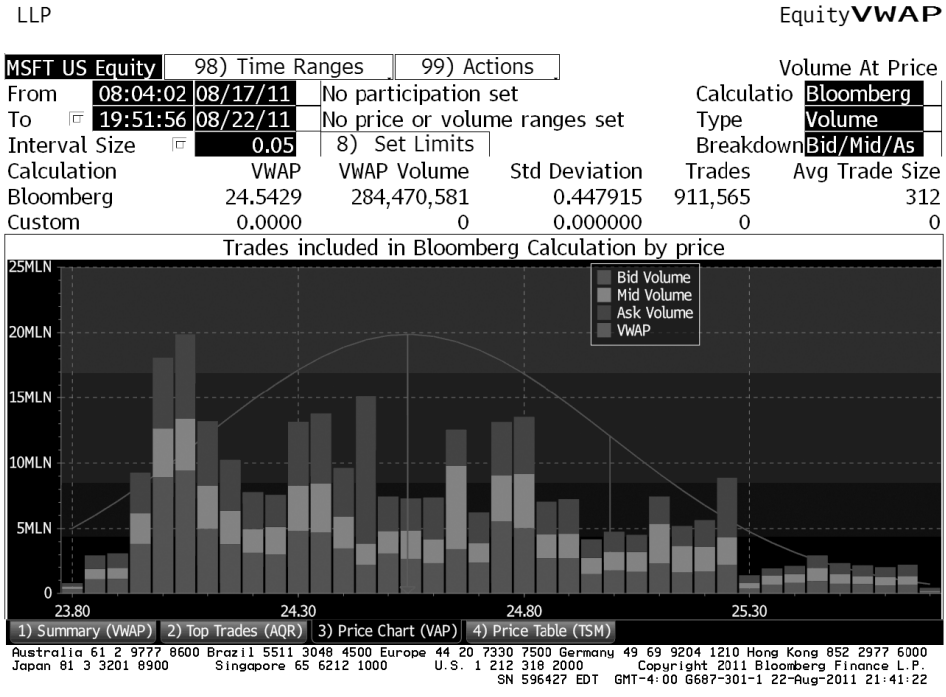
A possibility for the empirical verification of a dynamic LOB model would be to develop functionals of these state variables, which simulate the behavior of the observable quantities related to prices and volumes of traded securities. Once the probability distribution (measure) of the price-by-volume has been established, its integrals provide expectations for market-observed values such as average price and intraday volatility (see Sec. 5). My approach has an advantage over the time-series analysis (Aït-Sahalia *et al.*, 2012, 2013) in that the integrated quantities are robust with respect to the jumps of the state variables and particular statistics of the microstructure noise.

In this paper, I derive dynamic equations for the proportion of impatient traders in time. Before we can discuss the observable consequences of the FKK, we must provide a brief exposition of this theory, which we then extend to a fully dynamical version. In exposition of the FKK, we mostly follow De Jong and Rindi (2009).

3. Empirical Evidence on Integrated Volume-at-Price Distributions

My theory is proposed to describe one stylized fact, namely, that the volume-at-weighted-average-price (VWAP) distributions of liquid stocks integrated over several ordinary days have rather similar shapes. Specifically, during a single day, they typically form a unimodal structure, which crudely resembles a normal or lognormal distribution (Fig. 1).

During the next few days, a peak in the distribution widens and forms into a random structureless shape (Fig. 2(a)). Finally, after more than one week, a typical VWAP settles into a bimodal distribution (Fig. 2(b)).



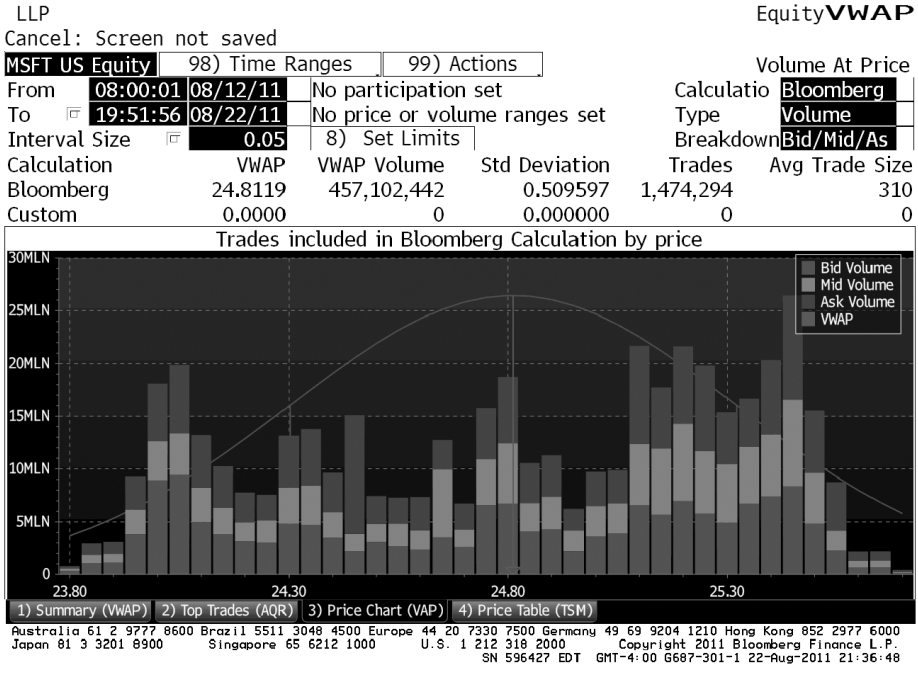
(a)

(b)

Fig. 1. Evolution of volume-at-price distributions with sampling days. (a) Volume-at-price distribution integrated for one day (August 22, 2011) for the Microsoft stock. Distribution is approximately bell-shaped and (b) VWAP distribution of AOL orders for the same day (blue) approximated by lognormal distribution (red) (color online).

In Fig. 3, I show VWAP distributions of quotes for Microsoft stock at Nasdaq during 1, 5 and 10 days of information collection in mid-August 2011, respectively. Ten days of collecting was the practical limit of the Bloomberg© VWAP function for that time and a longer observation interval was impossible. During these days in August, the trading volume was between 50 and 100 million shares. Cumulative data on trading are collected in Table 1.

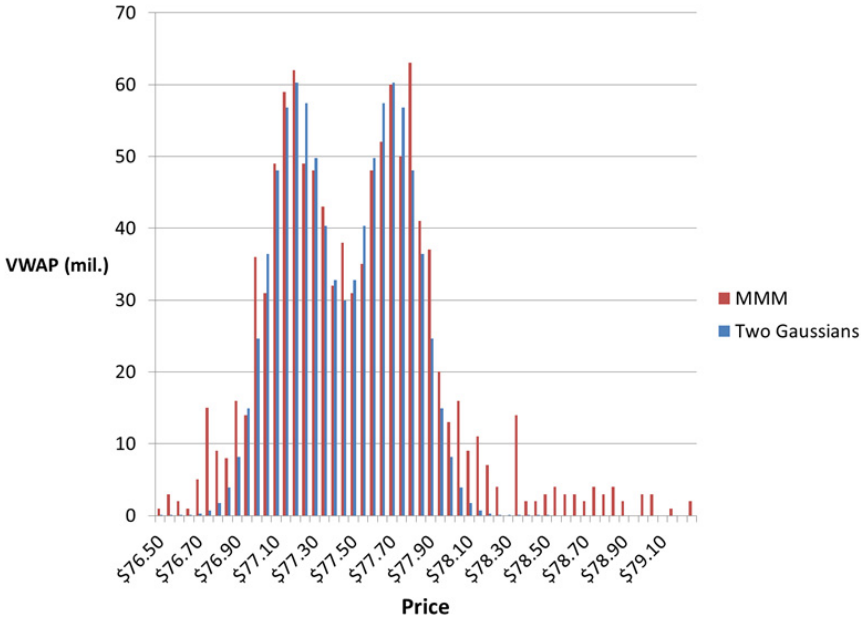
To exclude the influence of the real macroeconomic events from short-horizon (1–10 days) VWAP distributions, we use the following method inspired by the work of Barinov (2013).[2] At the first stage, we model the intraday NTSE-100 index. Simultaneously, we take the list of all substantial economic events in their intraday sequential order according to the WECO function of Bloomberg© and use Bloomberg-assigned weights and the news sign (positive-negative with respect to



(a)

Fig. 2. (a) Integrated volume-at-price distribution of the Microsoft stock for five days (08/17/11–08/22/11). The distribution is extended along the price axis. Maximum is poorly discernible and (b) Integrated VWAP for the MMM stock for the same period (red). Two overlapping peaks start to resolve (red). Approximation by two Gaussians is added as a guide (blue).

[2]Suggestion to use analysts disagreements as a measure of predictable volatility belonged to Lorne Switzer (Concordia University), see the introductory footnote.



(b)

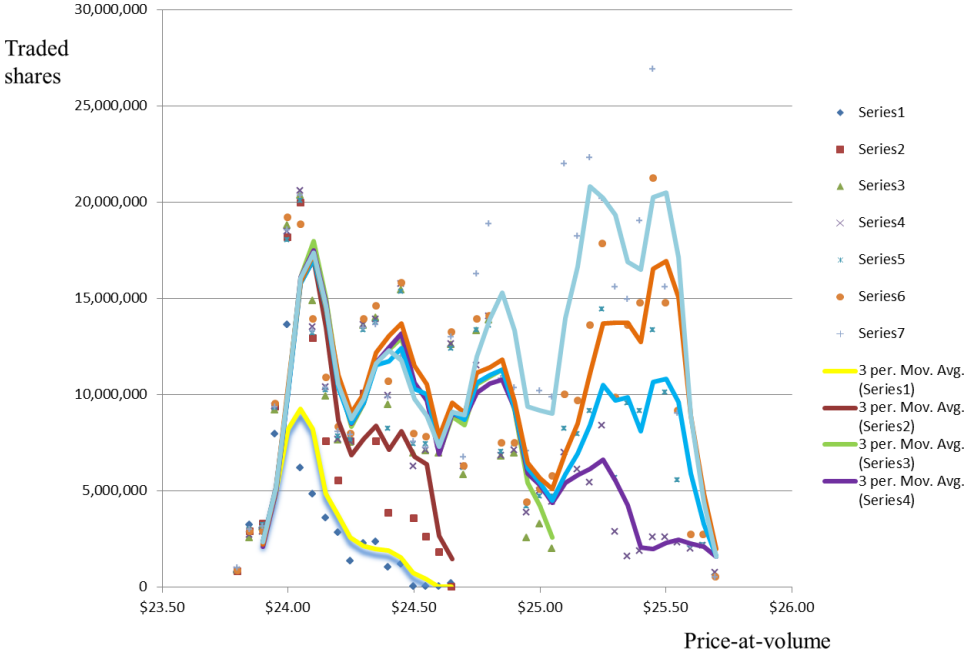
Fig. 2. (*Continued*)

prediction by the analysts' community) and

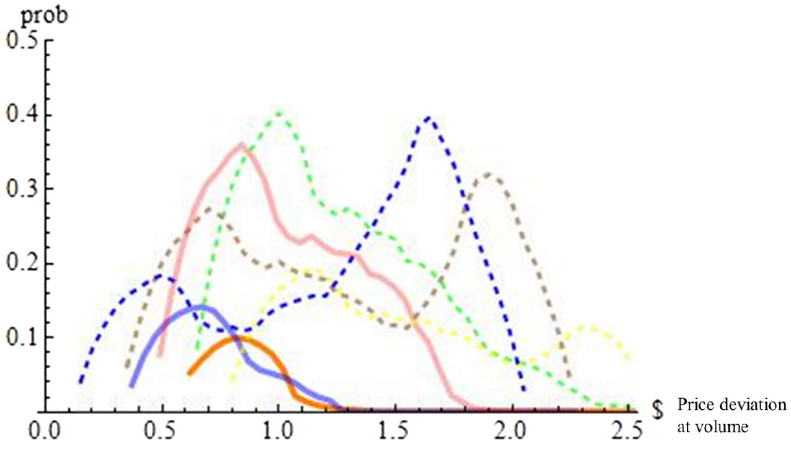
$$\mathbf{X}_t = \hat{B} r_t + \hat{C} F_t + \varepsilon_t$$
$$\mathbf{X}' = (r, F)$$
$$\hat{B}' = \left(\sum_{j=1}^{5} b_j L^j, 0 \right), \quad \hat{C}' = \left(\sum_{m=1}^{3} f_m, f_4 \rho L \right). \tag{1}$$

Here, in Eq. (1), r_t is the intraday index returns (taken with the standard 6 min interval). L is the one-period lag operator. We follow price evolution for 5 lags, i.e., for 30 min of trading. Economic factors f_m, $m = 1 \div 4$ are the coefficients measuring response of the index to the economic announcements F. The announcements are assigned signed numerical weights according to the Bloomberg© classification. Signs reflect the sign of surprise (positive or negative) with respect to the analysts' expectations. Also included as the explanatory variables their product (cross-influence) and integrated product with a weight ρ^t for the lag t put on the past events.[3] The distribution of thus constructed vector of economic news F (850 announcements during 140 trading days) is approximately normal.

[3]This is, of course, equivalent to the auxiliary AR(1) process for the factor f_4.



(a)



(b)

Fig. 3. (a) Microsoft VWAP for the 10 days (08/11/11-08/22/11). 10 days was a practical limit of quote collection allowed by the Bloomberg© system in 2011. We observe that the distribution has a tendency to concentrate in two peaks — near the minimum and maximum price for an observed period and (b) De-trended and filtered VWAP distributions in accordance with precepts of the Sec. 5.

Table 1. Descriptive statistics of trading in Microsoft stock during the period 08/11/11–08/22/11. They demonstrate a significant stability of an average price with a tendency to grow but a significant increase of standard deviation in price with subsequent flattening out.

Trading days	Average price ($)	Standard deviation ($)	Trading volume (in 1000 shares)	Cumulative trading
1	24.0617	0.145818	54,183	54,183
2	24.1733	0.190368	76,505	130,688
3	24.3891	0.316262	104,198	234,886
4	24.5429	0.447915	49,585	284,471
5	24.673	0.512251	52,847	337,318
6	24.737	0.536929	55,365	392,683
7	24.818	0.509597	64,419	457,102

The last factor reflects noninstant incorporation of economic events in the index price. Our estimates below (see Table 2) indicate that current economic news is fully incorporated into the index during 25–30 min of trading.

Explanatory power of this model is measured by OLS regression of predicted (Eq. (1) without the error term) and the actual returns. Explanatory power is low ($R^2 \approx 1\%$) and we conclude that the bulk of intraday fluctuations of the stock prices within the bid-ask window (see Cao *et al.*, 2009, Fig. 1 for the sketch) is produced by microstructure effects. The fact that the slope of the OLS regression is close to unity and the intercept is close to zero suggests that the model of Eq. (1) correctly describes essential features of the intraday price process despite its low explanatory power.

We have tried several alternatives of the regression of Eq. (1) explaining stock fluctuations by the intervening economic events. For instance, we tried to predict daily residuals of CAPM regression from the moving average of the strengths of the past announcements

$$\begin{aligned} r_t &= \alpha + \beta r_{Mt} + \delta \sum_1^5 f_m \cdot h_{t-m} + \varepsilon_t \\ r_t &= \alpha + \beta r_{Mt} + \delta \sum_1^5 \tanh(f_m \cdot h_{t-m}) + \varepsilon_t. \end{aligned} \tag{1'}$$

In Eq. (1′), r_{Mt} is a return of a stock on some broad market index, h_m are the measured strengths of the preceding announcements and δ is a scaling factor. Our two regressions: linear (first equation of (1′)) and logistic (second equation) produce roughly similar results. The regressions of Eq. (1′) for the first 140 days

Table 2. Intraday regressions of NTSE-100 for 19 trading days (07/25/11–08/18/11) with 6 min interval according to Eq. (1) and daily regressions for 140 trading days (01/04/11–08/17/11). (A) Parameter estimates for regression of Eq. (1) is denoted as (1) and the Eq. (1′)—as (2). (B) Relationship between the predictable and actual returns by OLS regression.

(A)		
Estimation parameter	Magnitude (1)	Magnitude (2)
b_1	−0.0218	—
b_2	0.0214	—
b_3	−0.0122	—
b_4	0.0166	—
b_5	0.0181	—
f_1	−9.00E-006	1.64E-003
f_2	−4.60E-003	−1.28E-003
f_3	6.91E-003	1.04E-003
f_4	−5.54E-003	−3.80E-004
f_5	—	7.70E-004
ρ/δ	0.2360	1.2033
(B)		
Parameter	Regression (1) magnitude (Std err at 5%)	Regression (2) magnitude (Std err at 5%)
Intercept	−8.11E-05	−2.1E-004
	(7.64E-05)	(1.61E-004)
Slope	0.974	0.019
	(0.222)	(0.012)
F	19.33	2.597
R^2_{adj}	1.30%	1.14%
Regression valid at 5%	Yes	No

preceding our measurement interval have similar predictive quality as the regression of Eq. (1). The coefficients describing influence of past economic announcements have the same order of magnitude. Unlike the intraday model, the coefficients of the OLS comparison of the daily model with actual returns differ insignificantly from zero (see Table 2). Hence, we conclude that (i) intraday returns significantly depends on the announcements of substantial economic news but they contribute relatively little to volatility and (ii) these announcements are completely incorporated into the returns within one day and do not have any predictive value for the daily movements of the stock. Consequently, expected

influence of the macroeconomic news on the shape of VWAP distributions from day to day is relatively small.

4. Aliasing of the VWAP Distributions

At the second stage, we use method of aliasing borrowed from image processing techniques (Mallat, 1999). Namely, we use a Gaussian smoothing filter with the width corresponding to the standard deviation of predictable (past history and economic-defined) events as a smoothing kernel (Tsay, 2002) for the empirical distributions both in state variable (price) and frequency domains. This intuitively corresponds to "averaging out" substantial economic events for the entire period.

$$F_j(p,\omega) = \sum_{p',\omega} f_j(p-p',\omega-\omega')\hat{F}_j(p',\omega'). \tag{2}$$

In Eq. (2), $\hat{F}_j(p,\omega) = P_{VWAP}(p) \cdot \varphi_{v\to\omega}[P_{VWAP}(p)]$, is an observed volume-by-price distribution for the j integration days and $\varphi(\cdot)$ is its Fourier transform. Furthermore, f_j is a Gaussian aliasing kernel with the Gaussian width σ_p in price and σ_ω frequency space (Mallat, 1999, Sec. 2.3.2) obtained from the first-stage regression. For the squares of Gaussian widths, we use sample estimators of the variances of the predictor of the (Eq. (1)) of the original signal and its Fourier transform, respectively. These sample estimators are scaled in proportion to the actual variance of the distribution of stock returns.

Intuitively, $F_j(p,\omega)$ is a distribution, for which actual macroeconomic events happening on observation days are replaced by their average intensity (volatility). The kernel f_j has conventional diffusion scaling with the number of integration days:

$$f_j(p,\omega) = f_1\left(p\sqrt{j}, \frac{\omega}{\sqrt{j}}\right) \propto \exp\left(-\frac{p^2 j}{2\sigma_p^2} - \frac{\omega^2}{2j\sigma_\omega^2}\right).$$

The above procedure is somewhat similar to the one we used in our paper (Lerner, 2013) to separate noise, microstructure effects and economic events in intraday returns on NTSE-100 index.

5. Formulation of the LOB Theory by Foucault, Kadan and Kandel

Foucault *et al.* (2005) assume that a risky security is traded in a continuous double auction. Information is symmetric and all participants are liquidity traders, i.e.,

they trade independently of the market fundamentals. The only difference between traders is their tolerance for the speed of execution of their orders, quantified by the "patience" parameter, $0 < \theta_P < 1$, or the proportion of patient traders in the crowd.

In FKK, as well as in empirical reality the number of trades matters much more than the volume (Glaser and Weber, 2007). In FKK, the buyer always follows the seller because all liquidity is supplied endogenously. Consecutive orders are numbered by an integer j, $j\{0, 1, \ldots, s-1\}$, where s is the length of a session. If the tick size is Δ, then the updated prices will be

$$\begin{aligned} P_{\text{buy}}(j) &= a - \Delta \cdot j, \\ P_{\text{sell}}(j) &= b + \Delta \cdot j, \end{aligned} \tag{3}$$

where a and b, respectively, are the ask and bid prices in the beginning of the trading session. Equation (3) assumes that the price has been updated j times to a moment at which the prices were observed.

In the FKK model there is a technical assumption that $a, b \in [\text{A}, \text{B}]$, where A and B are acceptable price limits. This assumption is not entirely unrealistic because in many markets there are circuit breakers preventing extreme price movements. For market orders, we automatically presume $j = 0$.

The time for their execution of each order (a waiting time) is a random function $T(j)$. Traders optimize the waiting losses:

$$c_j = j_i - \delta_i \cdot T(j) \geq 0, \tag{4}$$

where $i = \text{I}, \text{P}$ ("Impatient", "Patient") indicates the degree of immediacy (delay of execution) that each type of trader is willing to tolerate. The variables δ_i determine the potential loss expected by each type of trader for the unit time delay of the execution of their order. By construction $\delta_I > \delta_P > 0$. When, for a certain $j*$, the inequality in Eq. (4) reduces the equality of Eq. (4) to zero, the corresponding price according to Eq. (3) is called the reservation price and the time of execution the reservation time. The meaning of this equality is that the reservation price by definition is the price at which the trader is indifferent between the limit and the market order.

Further, FKK assume that a waiting time for the jth sell order, which follows the jth buy order, is distributed according to a Poisson distribution with the rate constant λ having the unit of inverse time.

For the execution time, FKK derive an equation:

$$T(j) = \frac{\alpha(j)}{\lambda} + \sum_{k=1}^{j-1} \alpha_k(j) \left[\frac{1}{\lambda} + T(k) + T(j) \right]. \tag{5}$$

In Eq. (5), $\alpha_k(j)$ is the probability of a limit order with the spread k on the jth step. The meaning of Eq. (5) is as follows. The first term indicates the probability that all orders up to $(j-1)$th were market orders. The expected delay for the execution of the limit order with the observed spread k is larger than the expected delay of the market order $1/\lambda$ by the sum of the expected delay of the trader of the opposite kind (buyer for seller and seller for buyer) $T(k)$ and the expected delay of the trader of the same kind $T(j)$. Furthermore, by the classic rule for probabilities:

$$\sum_{k=1}^{j-1} \alpha_k(j) = 1, \tag{6}$$

and the equation for the $T(j)$ acquires the form

$$T(j) = \frac{1}{\alpha_0(j)} \left[1/\lambda + \sum_{k=1}^{j-1} \alpha_k(j) T(k) \right]. \tag{7}$$

Then, FKK proves that the patient trader $i = P$ facing the spread within the limits $< n_h + 1, n_h >$, for $h = 1, \ldots, q-1$ where $n_1 = j'_P$ is the reservation spread for a patient trader, and the trader numbered $n_q = K \equiv a - b$ submits a limit order at her reservation price $j = n_h \cdot \Delta$. Hence, the equilibrium distribution of delay times follows from Eq. (7) as:

$$T(n_h) = \frac{1}{\lambda} \left[1 + 2 \sum_{k=1}^{h-1} \left(\frac{\theta_P}{1-\theta_P} \right)^k \right], \tag{8}$$

where $h = 2, \ldots, q-1$. This equation relies on the observation that, on their way to h's trade, the patient trader randomly met patient and impatient traders in direct proportion of their occurrence θ_P and $(1-\theta_P)$ in the sample.

Because of Eq. (4), the distribution of delay times is proportional to the cost function for the traders

$$c_j(n_h) = j_i - \delta_i \cdot T(n_h). \tag{9}$$

Expected prices in the efficient market are expressed through the proportion of patient traders θ_P as follows:

$$\left\{ \begin{array}{l} P_{\text{buy}} = a - \Delta \cdot \theta_P \cdot j_P - \Delta(1-\theta_P) j_P \\ P_{\text{sell}} = a + \Delta \cdot \theta_P \cdot j_P + \Delta(1-\theta_P) j_P \end{array} \right\}. \tag{10}$$

Equation (10) reflects the fact that, due to the presence of two groups of traders, selling pressure reduces the ask price and raises the bid price, in equal amounts.

The mid-price, in the original formulation of the FKK model, stays the same and we continue this convention for clarity, though this assumption can be modified for enhanced realism.

For a very large number of traders/adjustment steps $h \to \infty$ (very liquid stock), distribution of waiting times (Eq. (8)) tends to constant, which depends only on the fraction of impatient traders

$$T(n_h) \to T_\infty = \frac{1}{\lambda(1-2\theta_P)}, \quad T_\infty^{-1} = \lambda \cdot (1-2\theta_P). \tag{11}$$

We note that, while the process of quotes update is not a Markov process (waiting time depends on prehistory), asymptotically, for the large number of traders it is Markov. We shall exploit this property in the next section.

In the absence of intervening economic events, the expected mid-price always stays the same. Equation (8) and its derivatives such as Eq. (11) cannot easily be compared with empirical data, at least, not unless all limit orders with time stamps are known. Therefore, we propose a theoretical setting, which can potentially lead to observable quantities. Namely, suppose that each trading session has a fixed number of patient and impatient traders, but that what we observe in the market is a representative number of these trading sessions. In that case, averaging produces statistical distribution of the patient/impatient traders in time, which we identify with VWAP.

6. PDE Governing Traders' Distribution

To express the FKK theory in a potentially verifiable form, we assume that the quotes are governed by a two-state Markov stochastic process θ_t instead of a static parameter θ_P. This process indicates a kind of trader (patient or impatient) executing on the market at a given time. Two values of this process are equal to θ_P and $(1-\theta_P)$ and are independent from the market noise. The transition probability matrix describes the following situation:

Departure of patient traders from patient state	Arrival of impatient traders in place of patient ones
Arrival of patient traders in place of impatient ones	Departure of impatient traders from impatient state

Markov dynamics are fully determined by the transition matrix between states. In our case, following the usual assumptions, we choose the transition matrix consistent with Eq. (11) in its simplest form:

$$\hat{T} = \begin{pmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{pmatrix}. \tag{12}$$

Then, the probability of the patient trader continuing to wait $\vartheta = P[\theta_t \leq \vartheta]$ obeys the equation

$$\frac{d\vartheta}{dt} = \lambda(1 - 2\vartheta). \tag{13}$$

This equation indicates that with a rate $\lambda(1 - 2\vartheta)$, the patient trader is being replaced by an impatient trader and vice versa. The given form of the transition matrix (12) is not unique, on the contrary, there is infinite number of Markov transition matrices corresponding to the asymptotic law of Eq. (11), but we shall use the 2×2 transition matrix for simplicity. Equation (13) is deterministic, albeit with the stochastic initial condition — initial distribution of θ_P — and can be associated with the state equation of the filtering problem. The meaning of Eq. (13) is that asymptotically, in the limit of the large number of trades, a convergence to the true price is negative-exponential.

One cannot observe real-time proportion of patient and impatient traders. We assume that we can glean it with some inevitable stochastic error through the volume-at-price distribution (see the next section for explanation)

$$d\omega = \vartheta d\tau + \sigma dW_\tau. \tag{14}$$

In Eq. (14), as usual, W_τ is a stochastic noise process assuming a continuous Gaussian Markov process for analytic tractability,[4] the variable $\tau = T - t$. Equation (14) performs a role of an observation equation in the filtering problem.

As the state variable we chose the conditional expectation that at the inverse time $\tau = T - t$, a randomly chosen trader from the sample of traders active during the interval $[t, t + d\tau]$ belongs to either the "Impatient" or "Patient" group: $\omega(\tau) = E[\vartheta = \theta_P \,|\, F_\tau^W]$. We need only one variable because of the dichotomy of patient and impatient traders in the FKK model. The set F_τ^W refers to the microstructure noise. To derive a dynamic equation for the state variable, we replaced

[4]Aït-Sahalia and Jacod (2012) suggest that price jumps play a fundamental role in time-series statistics of asset returns. Microstructure noise in continuous time is much more likely to resemble white noise. We shall partially correct the restrictions of our assumption for the microstructure noise by assuming Poisson statistics for the short term fluctuations of arrival of orders (see Sec. 5).

discretized spreads of Eq. (3) by a random walk with an instant coordinate ϑ corresponding to $n = n_h$.

Equation (11) expresses delay of execution through a fraction of patient traders, θ_P. Inverse to the time T_∞ is the average rate at which orders get executed in the limit of a large number of trades. Impatient traders, for instance buyers, want to execute their successive trades during a characteristic time $T_1 = 1/\lambda$. However, they might not find equally impatient seller and have to wait until a patient seller arrives. Yet, patient traders have a distribution of execution times between $T_1 = 1/\lambda$ and $T_\infty(\theta_P) \to \infty$, for $\theta_P \to 1/2$.

It turns out that starting from Eq. (13) we can derive, in the continuous limit, a PDE governing the distribution function of ω (see Appendix A). Equation (10) permits us to express the price distribution from this function.

A conventional Kalman–Bucy filtering procedure (Liptser and Shiryayev, 1977, Theorem 9.1) leads to a backward Kolmogorov equation for the distribution function (see Appendix A)

$$\frac{\partial \pi(\omega,\tau)}{\partial \tau} = -\lambda\left[(1-2\omega) - \frac{\mu\omega(1-\omega)}{\lambda}\right]\frac{\partial \pi(\omega,\tau)}{\partial \omega} + \frac{\lambda^2\sigma^2}{2}[\omega^2(1-\omega^2)]\frac{\partial^2 \pi(\omega,\tau)}{\partial \omega^2}. \tag{15}$$

The analogues of the evolution Eq. (15) and the Hermitian conjugate of the evolution operator (forward Kolmogorov equation) appear in many contexts (except the already quoted Liptser and Shiryayev (1977), see for instance Bharucha-Reid (1960, Chapter 4.5) and Çetin and Verschuere (2010)). We shall discuss biological analogies and interpretations below, in Sec. 7.

The terminal condition for $t = T$ for this equation can, for instance, approximate a stationary solution of Eq. (15) (Appendix A). One of the simplest approximations of bimodal distribution is

$$P(\omega, n_h \to \infty) = A \cdot \delta(\omega - \theta_1) + B \cdot \delta(\omega - \theta_2), \tag{16}$$

where θ_1 and θ_2 are the lower and upper limits on the proportion of patient traders. If $A + B = 1$, the constants A and B can be identified as the proportion of patient buyers and patient sellers, respectively.

The terminal condition of Eq. (16) simply means that for a sufficiently long trading session all patient traders-buyers concentrate near a bid price and patient traders-sellers—near an ask price. Note that the asymptotic value of the limiting proportion of patient traders, θ_P, cannot exceed 1/2 (and not one) because the series for $T(n_h)$ diverges for a large number of executed trades when the proportion of patient traders exceeds one half.

Discontinuous terminal conditions of Eq. (16) can pose a difficulty for numerical modeling. In our simulation we replace them with a suitable continuous approximation of a δ-function, for instance, a narrow Gaussian shape. The question of the boundary conditions requires additional considerations, which are presented in Appendix B.

7. Price Distribution in the Dynamic LOB Model

Equation (15) with appropriate terminal and boundary conditions gives a probability distribution at a given time τ for a fraction of patient traders during the trading session. This equation is, in principle, not very different from the celebrated Black–Scholes equation. However, unlike the Black–Scholes where the state variable is the stock price S, here, the state variable is the probability of the next trader being patient or impatient. This is not an empirically observable quantity.

To compare the computed distribution of patient traders with any empirical price distribution, one needs to use Eq. (3) or (10) and the Bayes formula. Of course, this price distribution depends on the lower and upper limits of prices, which are the results of economic fundamentals. In our presentation, for clarity, we consider that the economic fundamentals change infrequently. For practical calibration, this assumption might be revised.

The Bayesian weighting method used in this paper is similar to the one used by Merton (Merton, 1973). The foundations of Merton's method using modern stochastic calculus techniques can be found in Jeanblanc *et al.* (2009) and the references therein.

We define a price distribution for bids in a real time t as an expectation that at a time t a trade will be executed at a reservation price $P(j')$, compared with Eqs. (3) and (4)

$$\pi_a(p,t) = E_t^X[p = a - P_{\text{buy}}(j') \,|\, \theta_P = \omega(t)]. \tag{17}$$

Similarly, the price distribution for the ask prices can be defined as

$$\pi_a(p,t) = E_t^X[p = P_{\text{sell}}(j') - b \,|\, \theta_P = \omega(t)]. \tag{18}$$

In Eqs. (17) and (18), X is a symbolic notation for the terminal and boundary conditions in Eq. (15) and j' is the reservation spread.

In the FKK framework, the bid prices are always adjusted upwards (Eq. (3)) and the ask prices are always adjusted downwards with respect to the trading limits a and b. These limits depend on economic fundamentals, which we consider as being infrequently revised, which is the practical case for almost all stocks. Even for the most liquid companies big news does not arrive every day. As one can glean from Eq. (10) — for instance from the fact that in FKK, mid-price is

unchanged during trading — these distributions are identical, so we omit the upper and lower indexes.

The function is obviously a probability distribution because (a) the expected price adjustment p is non-negative — nobody would bid at a price above an already available selling price or ask below a current buy price — and (b), it is normalized to unity

$$\int_0^\infty \pi(p,t)dp = 1. \tag{19}$$

Normalization follows from the simple fact that for any moment in real time we expect the price change $p > 0$ to be equal to *something*. Actual limits of integration consistent with the FKK in the integral of Eq. (19) lie between a and b — the starting prices at the beginning of the stylized trading session.

Further, we can use the Bayes formula and Eq. (3) to derive the observable price distribution in real time $P_{\text{obs}}(p,\tau)$

$$P_{\text{obs}}\left(p \equiv \delta \cdot \frac{\omega}{\Delta}, \tau\right) = \sum_{i=1}^{\infty} \omega\pi(\omega, \tau = \lambda i) \cdot P_{n,P}(i,t) + \sum_{i=1}^{\infty} (1-\omega)\pi(\omega, \tau = \lambda i) \cdot P_{n,I}(i,t), \tag{20}$$

where $P_{n,P}(i,t) = P_{\text{patient}}(t \,|\, n_h = i)$, $P_{n,I}(i,t) = P_{\text{impatient}}(t \,|\, n_h = i)$ are the prior price distributions conditional on the average number of price adjustments at a time $t < T$.

In agreement with the FKK, the priors have the following (Poisson) form

$$\begin{aligned} P_{n,I}(t \sqcup n_h = i) &= \frac{(\lambda t)^i}{i!} e^{-\lambda t} \\ P_{n,P}(t \sqcup n_h = i) &= \frac{(\lambda t(1-2\theta))^i}{i!} e^{-\lambda t(1-2\theta)}. \end{aligned} \tag{20'}$$

In our case, prior distributions of impatient and patient traders are assumed to be Poisson distributions with rate constants equal to $T_1^{-1} = \lambda$ and $T_\infty^{-1} = \lambda(1-2\vartheta)$, respectively.[5] This is, of course, a crude approximation, but I expect that the difference with the true distribution for a few days would be small. Equation (20′) means that the patient trader on the average waits $1/(1-2\cdot\theta)$ times longer for her execution than the impatient trader in accordance with Eq. (13). If we were to have only patient traders, this model of order volume would

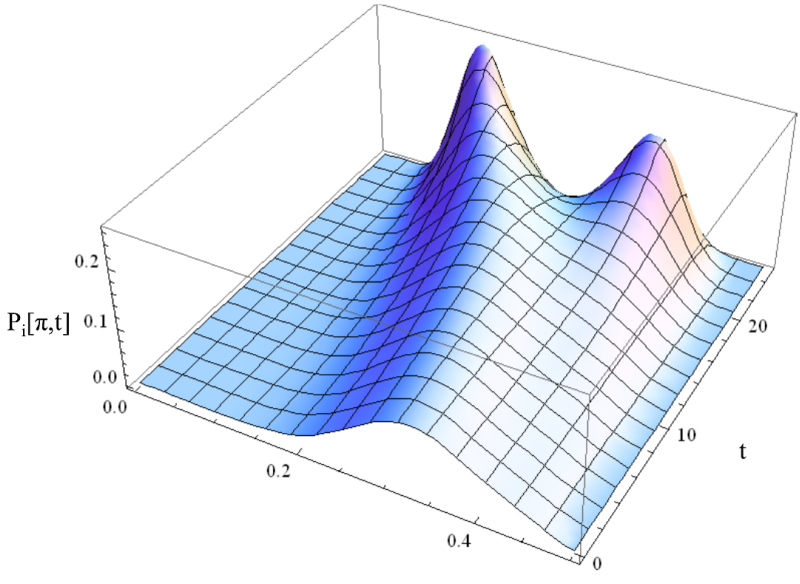
[5]For modeling using Poisson distributions, see e.g., Gourieroux and Jasiak (2001) or Chernobai *et al.* (2007).

be classified as self-exciting point process (Gourieroux and Jasiak, 2001; Chernobai, 2007).

For convenience, in Eq. (20), we rescaled price into units of Δ — the jump size—in Eq. (3). In writing down Eq. (19), we assumed that the distributions of the price adjustments are independently normalized on unity. In Eq. (20), $\pi(\omega, \tau)$ is the result of numerical integration of Eq. (15).

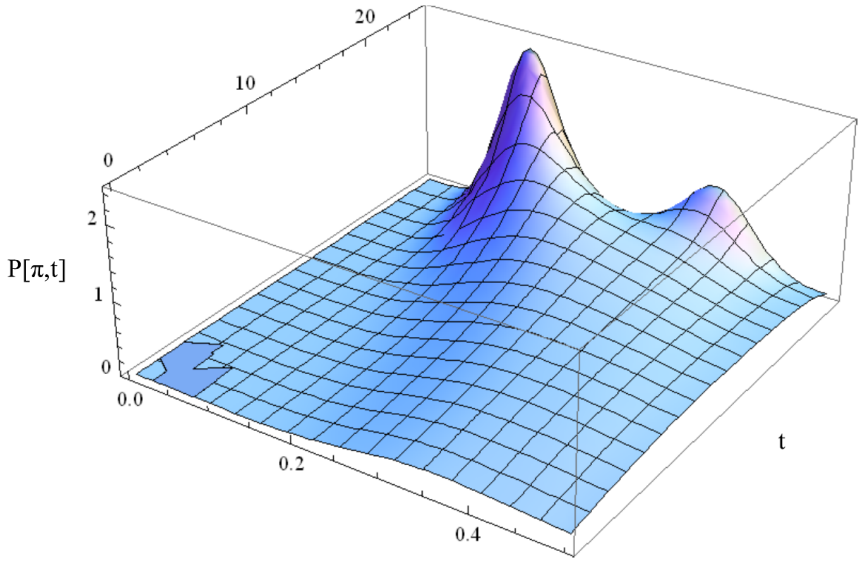
Equation (20) gives a price distribution in a parametric form dependent on a fraction of impatient traders. To compare it with observable volume-at-price distributions, one needs to complement the Bayesian formula (20) with an appropriate prior for the state variable.

In Fig. 4, I present the simulations of the trading market model of Eqs. (15) and (19). They reproduce the main qualitative feature of the empirical volume-at-price

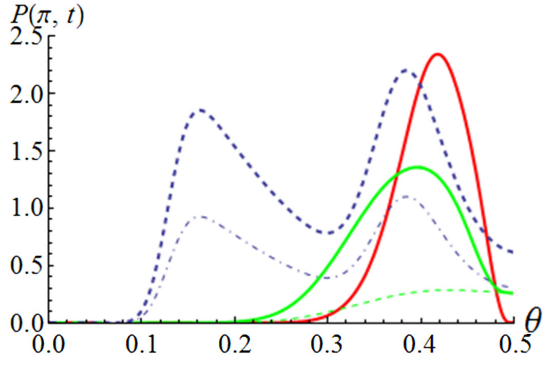


(a)

Fig. 4. (a) The solution of Eq. (15) of the main text. Simulated VWAP distribution is plotted as a function of normalized price and time. In the beginning of time, there is unimodal distribution, which converges to a terminal bimodal distribution. (b) Weighted θ-distribution (Eq. (18)), which we identify with the volume-at-price distribution. (c) Simulated price distribution in arbitrary units as a function of time (arbitrary scale). Fraction of participating patient traders is shown by the thin line of the same color. Distribution integrated for 0.5 days displays almost a bell-shaped form (bold red curve compare to Fig. 1(a)), the trading volume of patient traders will not be discernible at this scale. For the medium integration time of 5.5 days, the structure is relatively absent (bold green curve; compare to Fig. 1(b)). For 10 days the structure converges to bimodal shape similar to the real quotes in Fig. 2(c) (bold blue curve).



(b)



(c)

Fig. 4. (*Continued*)

distributions: nearly bell-shaped distribution for a small quote integration time, a shapeless form for a few days and progression to a distinct bimodal distribution with modes near the upper and lower limits of trading for the period. The existence of the two humps (otherwise known as bimodal structure of the distribution) is not surprising — it is an artifact of the terminal condition of Eq. (16) — but Eq. (15) demonstrates an evolution from this distribution into a familiar bell-shaped curve.

The simulations were performed for relatively arbitrary values of the parameters (unit of trading time, duration of the period, proportions of patient buyers and

Table 3. First two moments of the analytical distributions. Numerical analytic distributions were approximately rescaled to conform to the scale of the empirical distributions of Table 1. Numerical data, corresponding to the values of parameters (unit of trading time, duration of the period, proportions of patient buyers and sellers and maxima for terminal distribution of patient buyers and sellers, respectively) $\lambda = 0.5$, $T = 10$, $A = B$ and $\theta_1 \approx 0.13$, $\theta_2 \approx 0.38$ were not calibrated to the empirical data and are given here for illustrative purposes. Arbitrary conversion factor into dollars was the same for prices and their standard deviation.

Rounded-off time (days)	Average price ($)	Standard deviation ($)
2	24.28	0.240
3	24.35	0.282
6	24.50	0.453
7	24.91	0.646

sellers and maxima for terminal distribution of patient buyers and sellers, respectively) $\lambda = 0.25$, $T = 10$, $A = B$ and $\theta_1 \approx 0.13$, $\theta_2 \approx 0.38$. The moments of the simulated distribution behave approximately as the moments of the real distribution integrated for 10 days (Table 3).

8. Biological Analogues of the Competition Between Patient and Impatient Traders

Many biological/ecological systems express features similar to those of the dynamic LOB evolution. Their review can shed light on the intuitive nature of the results of previous sections. In the problem of competition between two alleles of the same species, if an appropriate mutation increases the longevity of the fertile period of an organism, it increases intergenerational representation of a longer-living species and thus assures the dominance of this species in a distant future. In the beginning, there will be a random distribution of alleles, but as time goes by only the allele with a useful mutation survives. In the end, the age distribution will be dominated by the longer-living, i.e., "patient" species. This view is a variation of the well-studied Kimura (1954) model in mathematical biology. (Bharucha-Reid, 1960, Chapter 4.5) Different versions of the Kimura model are possible and were developed in the context of mathematical genetics.

If, for instance, instead of increased survival, we assume that mutation increases only the longevity (but not, for instance, the average fertility rate) of the mutated species, the model will be described, instead by the Kimura equation, by an equation very close to the Eq. (15).

In application to finance, traders who have a relatively small penalty if they decide to wait for a more advantageous price benefit more at the expense of traders who value immediacy. Eventually, however, with the depletion of the pool of impatient traders, patient traders are forced to trade with one another and the VWAP pattern settles into the buyers concentrating near the selling price and sellers near the buying price with occasional clearance happening when a surviving impatient trader completes a deal (consult Fig. 4 for extinction of impatient traders during a trading process).

One of the more profound analogies of this situation has been modeled by the Nobel Prize Winner in biology Edward O. Wilson in his 1980 paper with Lumsden (Lumsden and Wilson, 1980). In that paper, the authors postulated the existence of evenly distributed "cultural" patterns between societies, which can be taught and learned. Conformity to these patterns, called "trend-watching", improves procreative chances of an individual (similar to competition for the best price of execution in the financial literature (Cont *et al.*, 2008) resulting in a broadened distribution of genetic markers. Through generations, small initial differences in learning bias can be significantly amplified. The affinity of Lumsden–Wilson simulated distributions and our simulations (compare Figs. 3 and 4 in Lumsden and Wilson (1980) and our Fig. 4) in the case they called "non-saturable trend-watching" is striking. The authors modeled the preference by an exponential "assimilation function" while its analog in our treatment is parabolic.

Lumsden and Wilson were concentrating on a stationary distribution (Lumsden and Wilson, 1980, Eq. (5)), which can be easily inferred from a forward Fokker–Plank equation

$$P(\theta) = C\left(\frac{\theta}{1-\theta}\right)^{\alpha} \exp\left[\frac{1}{\lambda\sigma^2}\left(\frac{1}{\theta}+\frac{1}{1-\theta}\right)\right], \tag{21}$$

where $\alpha = 2\mu/\lambda\sigma^2$ and C is a normalization constant. In our case, this solution does not play an important role because in real trading markets, the equilibrium is systematically violated by the arrival of substantive news.

9. Note on the Predictable Size of the Effect

First, I comment on the alternative hypothesis: namely that the one peak-two peak structure is the result of buying or selling pressure stimulated by the arrival of real economic news in a course of a few days (five or so in the case of Microsoft) and that multiple peaks will be observed for a longer period of observation impossible

with a current system.[6] The Bloomberg© system could produce separate distributions for "buy" and "sell" orders and the bimodal structure can be observed in either of them. Yet, if the bimodal structure were simply the result of the overlap of economic news from different days, one peak will be much more likely registered in "buy" and another — in "sell" distributions (or both in one and none in the other) caused by variable price pressures and order imbalance. We use cumulative statistics only for smoother VWAP distributions because separate buy and sell distributions can have "holes" in them.

We used a very liquid stock (Microsoft) for our illustration. Furthermore, only a 10-day integration horizon was available for the VWAP function in the 2011 version of Bloomberg©, which limited our lookback capacity to seven trading days.[7] Thus, a size of infrastructure effects observed in our example is likely to be underestimated with respect to longer samples and less liquid securities.

A natural estimate of the trading loss of impatient traders as a group to the patient traders would be the standard deviation of VWAP for the unimodal distribution ($\sim\$0.15$ in our example) or half of the standard deviation—an approximate width of a single peak—$\$0.25$ for the bimodal distribution. Given that the typical price of Microsoft stock in 2011 hovered around $\$25$, a trading loss because of immediacy concerns can be crudely estimated as -0.7–1% per day. Hence, the front-runners with inside information must gain as much to break even (see Da *et al.* (2011) and Switzer and Fan (2007) for empirical estimates from the actual trading).

Accrual of this loss throughout longer periods is harder to estimate because during longer periods the same agent can appear on the patient as well as the impatient side. If we identify the accumulation of transaction cost losses with the cost of maintenance of the portfolios, i.e., this cost scales approximately as a square root of the number of trading periods (Bollen *et al.*, 2004; Lerner, 2009, Chapter 2.2), then during one month assumed as 22 trading days

$$L \simeq -70 \div 100\,bp\sqrt{22} \simeq -300 \div 470\,bp.$$

This estimate demonstrates that even for a very liquid stock, trading losses due to microstructure frictions can quite substantially eat into the return if trades are scheduled unwisely. Here, I must provide a cautionary note that the above considerations do not contradict profitability of the "front-running" strategies such as detected by Da, Gao and Jagannathan (Da *et al.* (2011), see also preface to Lerner

[6]This suggestion was made by Alexander Skobelin (Goldman Sachs) at the World Finance Conference in Venice, 2014. I am thankful for his question producing the train of thought below.

[7]Intraday, full-quote VWAP was not available for all stocks in the 2011 version of Bloomberg© subscribed by Cornell University.

(2009) for the description of some of the abusive methods of front-running). Front-running relies on the speed of reaction to real, imaginary or manufactured economic events within the bid-ask window and the traders can benefit from even a very unfavorable execution price if predictable movement of the stock price in response to the information event is significant.

10. Conclusion

In this paper, I develop a plausible argument that a relatively parsimonious extension of the Foucault–Kadan–Kandel (FKK, 2005) theory can explain some qualitative features of volume-at-price distribution for stocks with a large depth. I demonstrate that asymptotically, in the limit of a large trade size, FKK leads to a relatively simple PDE, which can be solved numerically.

Simulations demonstrate the same qualitative feature as the real VWAP plots, namely, that the bid-ask quotes eventually concentrate on the "wings" of the distribution without intervening economic news, which is absent in our stylized model. The standard deviation of price distribution grows despite the fact that the "true" price becomes clearer for the participants of trading as more and more traders are willing to wait.

Because the model is symmetric with respect to buying and selling, but its buying and selling distributions can be independently calibrated with the available market data (see Fig. 1), one can hope to quantify notions of "buying" and "selling pressure" in the framework of the presented model. Quantitative comparisons are currently a work in progress. The impediment is that the calibration of parameters in a Poisson-weighted family of solutions of partial differential equations is not a straightforward econometric problem.

We estimated the size of the expected effect. If we assume that in the end of the trading period, all patient sellers concentrate near the ask price and all patient buyers concentrate near the bid price, then the size of the effect is roughly a half of the bid-ask spread for the regular days (days without much imbalance for buy and sell orders). Daily accumulation of these losses can reduce the performance of portfolios quite substantially.

Even for the most liquid stocks, such as AOL, Microsoft and 3M, impatient but poorly informed traders, lose on the order of magnitude of a dozen-or-so cents for the share price of several tens of dollars. A loss of only a few percent results in several million a day transferred from impatient to patient traders for a typical turnover of a highly liquid S&P or Nasdaq stock.

Acknowledgments

I am thankful to the members of the finance seminar at Illinois Institute of Technology for their valuable comments on the first (2012) version and to my discussants at Global Finance Conference (Monterey, May 2013), Hisham Farag (University of Birmingham), the 5th Meeting on the Behavioral Finance and Economics (Chicago, September 2013), Mikaela Pagel (Berkeley) and the World Finance Conference (Venice, 2014), David-Jan Jensen (Chair), Yushi Yoshida (Shiga University) and Al Skobelin (Goldman Sachs). I express deep gratitude to Lorne Switzer (Concordia University), Chair at the Monterey Meeting, for his advice during and after the conference, especially for using the WECO Bloomberg© function for economic events. I bear responsibility for all errors.

Appendix A. Filtering Equation for the Observation of Quote-Updating Process

A derivation of the PDE for the observation process described in Sec. 4 is performed by the standard techniques of Kalman–Bucy filtering. In the treatment below, we closely follow Liptser and Shiryaev (1977). Let (ϑ, ω) be the random process introduced in Sec. 4, where ϑ is a "true" traders' arrival process and ω is the process as it appears to the observer. Then, one can define a conditional probability

$$\omega_\beta(t) = P[\theta_t = \beta \,|\, F_t^W], \tag{A.1}$$

$\beta = i, p$ which is the best guess of the type of trader based on the previous trading information.

Then, under some generic conditions on the nature of probability space, Liptser and Shiryaev prove the following theorem. If the processes (ϑ, ω) are diffusions driven by the Brownian noise term, then the probability $\omega_\beta(t)$ obeys the equation of diffusion

$$d\omega_\beta(t) = a(\omega_\beta, t)dt + b(\omega_\beta, t)d\bar{W}_t, \tag{A.2}$$

where the coefficients a and b are given by the following equations:

$$a(\omega, t) = \sum_\beta \lambda_\beta \omega_\beta(t), \tag{A.3}$$

$$b(\omega, t) = \sigma[A_t(\omega, \xi) - \sum_\beta A_t(\xi)\omega_\beta(t)]. \tag{A.4}$$

The terms λ_β in Eq. (A.3) are the elements of the Markov transition matrix (Eq. (12) of the main text). A_t is the transition rate from Eq. (12). The Brownian motion $\bar{W}_t$ is a Wiener process for innovations, the exact nature of which is not relevant for us now (e.g., Liptser and Shiryaev, 1977, Theorem 9.1). An explicit expression for $\bar{W}_t$ is given in Theorem 8.1, *ibid*. In (A.2), we presume that the number of discrete values of our process is at least countable (it is 2—patient and impatient—in our simplified formulation). Because we have only two states, $\omega_p = 1 - \omega_i = \omega$ and we can take the probability of patient trader as the only state variable. The differentials of these variables are related as $d\omega_p = -d\omega_i$. Using (A.3), (A.4) and Eqs. (12)–(14) of the main paper we get

$$a(\omega, t) = -\lambda\omega + \lambda(1 - \omega) = \lambda(1 - 2\omega) \tag{A.5}$$

and

$$b(\omega, t) = \lambda\sigma\omega(1 - \omega) \tag{A.6}$$

The resulting diffusion takes a form

$$d\omega(t) = \lambda(1 - 2\omega)dt + \lambda\sigma\omega(1 - \omega)d\bar{W}_t. \tag{A.7}$$

However, Eq. (A.3) reflects the situation where the choice between immediate execution and waiting is risk-neutral. But there is no logic in assuming that, given that the order of execution is not tradable between parties.[8]

Because of that, $\bar{W}_t$ is no longer a Brownian motion under $P[\cdot \mid F_t^W]$. Generally, its form can be quite arbitrary. However, if we assume that the distribution of ϑ_T at the final moment of the trading session is known, in an extended algebra $F^{W \sqcup \vartheta_T}$ there exists a numerical constant μ such that a transformed process is a Brownian motion under $P[\cdot \mid F_t^{W \sqcup \vartheta_T}]$.

$$d\bar{\bar{W}}_t \to d\bar{W}_t - \mu\omega(1 - \omega)dt, \tag{A.8}$$

(see e.g., the treatment in Çetin and Verschuere, 2010). Constant μ has the intuitive meaning of a market price of delay risk. The final equation for the diffusion becomes

$$d\omega(t) = \lambda\left[(1 - 2\omega) - \frac{\mu}{\lambda}\omega(1 - \omega)\right]dt + \lambda\sigma\omega(1 - \omega)d\bar{\bar{W}}_t. \tag{A.9}$$

Equation (A.9) can be rendered in the form of the forward Kolmogorov diffusion for $\pi(\omega, t) = dP(\omega_t \leq \omega)/d\omega$ — the probability distribution for the fraction of

[8]Of course, this assumption would be violated if high-frequency traders are allowed to bid higher prices for speed of the execution. Consequences of more complete high-frequency markets are unclear at this point but are certainly worth investigating.

patient vs. impatient traders:

$$\frac{\partial \pi(\bar{\omega}, t)}{\partial \tau} = \lambda \frac{\partial}{\partial \bar{\omega}} \left[(1 - 2\bar{\omega}) - \frac{\mu \bar{\omega}(1 - \bar{\omega})}{\lambda} \right] \pi(\bar{\omega}, t) + \frac{\lambda^2 \sigma^2}{2} \frac{\partial^2}{\partial \bar{\omega}^2} [\bar{\omega}^2 (1 - \bar{\omega}^2)] \pi(\bar{\omega}, t) \tag{A.10}$$

We use the backward form of Eq. (A.10), $\tau = T - t$ (Eq. (15) of the main text) in our simulations and the forward form in Appendix B. Elsewhere, for brevity we omit the bar over the argument, because a subtle difference between $\omega_t = \omega(\tau)$ — the state variable and ω — the argument of the distribution function, cannot cause confusion in other contexts.

Appendix B. Boundary Conditions for Kolmogorov–Fokker–Planck Equation

Equation (15) has the diffusion-dissipative type like all the Black–Scholes-type (or Fokker–Kolmogorov) equations. Because of that it is irreversible in time. In particular, only the forward equation is likely to preserve the probability normalization to unity.

We *postulate* that the boundary conditions for the backward (i.e., standard for finance) PDE must preserve probability for the conjugate forward equation. This postulate is not a consequence of any mathematical theory, but it is consistent with a common sense. In particular, it is a direct analog of the smooth-pasting condition in the endogenous default models in credit risk analysis (Mella-Barral and Perraudin, 1997).

Hence, the problem of probability conservation and the boundary conditions must be investigated for the forward KFP. We assume homogeneous boundary conditions of a Dirichlet or Neumann type.

The forward version of Eq. (8) is as follows:

$$\frac{\partial P(\omega, \tau)}{\partial \tau} = -\lambda \left[(1 - 2\omega) - \frac{\mu \omega(1 - \omega)}{\lambda} \right] \frac{\partial P(\omega, \tau)}{\partial \omega} + \frac{\lambda^2 \sigma^2}{2} [\omega^2 (1 - \omega^2)] \frac{\partial^2 P(\omega, \tau)}{\partial \omega^2} \tag{B.1}$$

In the Fick's (Klages *et al.*, 2008) form

$$\frac{\partial P(\omega, \tau)}{\partial \tau} = -\frac{\partial}{\partial \omega} j(\omega, \tau). \tag{B.2}$$

The probability current is equal to

$$j(\omega,\tau) = -\lambda\left[(1-2\omega) - \frac{\mu\omega(1-\omega)}{\lambda}\right]P(\omega,t) + \frac{\lambda^2\sigma^2}{2}\frac{\partial}{\partial\omega}[\omega^2(1-\omega^2)P(\omega,\tau)] \tag{B.3}$$

The conservation of the probability condition

$$\int_0^{1/2} P(\omega,\tau)d\omega = 1 \tag{B.4}$$

can be differentiated in τ to yield

$$\int_0^{1/2}\frac{\partial P(\omega,\tau)}{\partial\tau}d\omega = -\int_0^{1/2}\frac{\partial j(\omega,\tau)}{\partial\omega}d\omega = j(0,\tau) - j(1/2,\tau) = 0. \tag{B.5}$$

Using Eq. (B.3) to express the probability current on the boundaries through the probability, we arrive at the expression:

$$P(0,\tau) + \frac{\lambda\sigma^2}{8}\frac{\partial}{\partial\omega}P(\omega = 1/2,\tau) = 0. \tag{B.6}$$

This equation connects the probability distribution on the lower boundary $\theta_P = 0$ with the first derivative of the probability on the upper boundary $\theta_P = 1/2$. Assuming that if only impatient traders are present in the market, a boundary condition reads as

$$P(0,\tau) = 0. \tag{B.7}$$

The intuitive meaning of the Eq. (B.7) is that, in the beginning of trading session, the numbers of impatient traders completely dominate the patient trading and only though continuing trading the presence of patient traders becomes visible. Then, Eq. (B.7) immediately gives

$$\frac{\partial}{\partial\omega}P(\omega = 1/2,\tau) = 0. \tag{B.8}$$

The last equation provides a condition on the spatial derivative of the probability distribution on the upper boundary. Another assumption fixing the same boundary conditions would be that Eq. (B.6) must hold for arbitrary λ.

References

Aït-Sahalia, Y, J Fan and Y Li (2013). The leverage effect puzzle: Disentangling source of bias at high frequency at high frequency. *Journal of Financial Economics*, 109: 224–249.

Aït-Sahalia, Y and J Jacod (2012). Analyzing the spectrum of asset returns and volatility components in high-frequency data. *Journal of Economic Literature*, 50(4): 1007–1050.

Angel, J, L Harris and CS Spatt (2010). Equity trading in the 21st century. Marshall School of Business Working Paper No. FBE 09-10.

Barclay, M and T Henderschott (2004). Liquidity externalities and adverse selection: Evidence from trading after hours. *Journal of Finance*, 59: 681–710.

Barinov, A (2013). Analyst disagreement and aggregate volatility risk. *Journal of Financial and Quantitative Analysis*, 48(6): 1877–1900.

Bharucha-Reid, AT (1960). *Elements of the Theory of Markov Processes and their Applications*. New York: McGraw Hill.

Bollen, NP, T Smith and RE Whaley (2004). Modeling the bid/ask spread: Measuring the inventory-holding premium. *Journal of Financial Economics*, 72: 97–141.

Çetin, U and M Verschuere (2010). Pricing and hedging in carbon emissions markets. *International Journal of Theoretical and Applied Finance*, 12(7): 949–967.

Chernobai, AS, S Rachev and F Fabozzi (2007). *Operational Risk: A Guide to Basel II Capital Requirements, Models, and Analysis*, Frank J. Fabozzi Series, Vol. 152.

Cont, R, S Stoikov and R Talreja (2008). A stochastic model of order book dynamics. Columbia University Working Paper.

Da, Z, P Gao and R Jagganathan (2011). Impatient trading, liquidity provision and stock selection in mutual funds. *Review of Financial Studies*, 24(3): 675–720.

De Jong, F and B Rindi (2009). *The Microstructure of Financial Markets*. Cambridge, UK: Cambridge University Press.

Easley, D, M Lopez de Prado and M O'Hara (2011). The microstructure of the "flash crash": Flow toxicity, liquidity crashes and the probability of informed trading. *The Journal of Portfolio Management*, 37(2), 118–128.

Foucault, T, O Kadan and E Kandel (2005). Limit order book as a market for liquidity. *Review of Financial Studies*, 4: 1171–1217.

Glaser, M and M Weber (2007). Overconfidence and trading volume. *Geneva Risk and Insurance*, 32: 1–36.

Gourieroux, C and J Jasiak (2001). *Financial Econometrics: Problems, Models, and Methods*. Princeton: Princeton University Press.

Harris, LE (1998). Optimal dynamic order submission strategies in some trading problems. *Financial Markets, Institutions and Instruments*, 7: 1–76.

Harris, L (2003). *Trading and Exchanges: Market Microstructure for Practitioners*. New York: Oxford University Press.

Hasbrouck, J (2007). *Empirical Market Microstructure*. New York: Oxford University Press.

Henderschott, T and CM Jones (2005). Island goes dark: Transparency, fragmentation and regulation. *Review of Financial Studies*, 18: 743–793.

Jeanblanc, M, M Yor and M Chesney (2009). *Mathematical Methods for Financial Markets*. Heidelberg, Germany: Springer.

Jovanovic, B and A Menkveld (2011). Middlemen in limit order markets. New York University Working Paper.

Klages, R, G Radons and IM Sokolov (2008). *Anomalous Transport*. Weinheim, Germany: Wiley.

Lerner, P (2009). *Microstructure and Noise in Financial Markets*. Saarbrucken, Germany: VDM Verlag.

Liptser, RS and AN Shiryayev (1987). *Statistics of Random Processes*, Vols. 1 and 2. New York: Springer-Verlag.

Mella-Barral, P and W Perraudin (1997). Strategic debt service. *Journal of Finance*, 50(2): 531–556.

Menkveld, AJ (2010). High frequency traders and the New-Market makers, VU University Amsterdam.

Merton, RC (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science*, 4(1): 141–183.

Obizhaeva, A and J Wang (2005). Optimal trading strategy and supply/demand dynamics, NBER Working Paper No. 11144.

O'Hara, M (1995). *Market Microstructure Theory.* Blackwell, UK: Oxford University Press.

Skjeltorp, J, E Sojli and W Wah Tam (2012). Identifying Cross-Sided Liquidity Externalities, Erasmus University Working Paper.

Sornette, D and S von der Becke (2011). Crashes and high frequency trading. Swiss Finance Institute Series No. 11–63.

Switzer, L and H Fan (2007). The transaction costs of risk management vs. speculation in an electronic trading environment: Evidence from the Montreal exchange. *Journal of Trading*, 2(4): 82–100.