Resilient price impact of trading and the cost of illiquidity[*]

Alexandre Roch[†] H. Mete Soner[‡]

March 29, 2011

Abstract

We construct a model for liquidity risk and price impacts in a limit order book setting, and derive a wealth equation and a characterization of illiquidity costs for liquidity providers. The model has desirable stylized facts justified by empirical studies and contains all three components identified by Kyle (1985). We give conditions under which the model is arbitrage free. By considering the standard utility maximization problem, we obtain a stochastic discount factor and an asset pricing formula which is consistent with the empirical findings of Brennan & Subrahmanyam (1996) and Amihud & Mendelson (1986). Furthermore, we show that in limiting cases for some parameters of the model, we derive many existing liquidity models present in the arbitrage pricing literature, including Çetin et al. (2004) and Rogers and Singh (2010). This offers a classification of different types of liquidity costs in terms of the depth and resilience of prices.

Key words: Liquidity risk, limit order books, asset pricing, utility maximization, resilience, price impacts.
JEL classification: D40 G11 G12

1 Introduction

The notion of liquidity in financial markets can either refer to the added costs associated to trading large quantities of a given financial security, or can address the ability to trade this asset without triggering important changes in its prices. On one hand, we may think of the level of liquidity as a measure for the added costs per transaction and model it through

[*] Research partly supported by the European Research Council under the grant 228053-FiRM. Financial support through the ETH Foundation and the Swiss Finance Institute is gratefully acknowledged.

[†] ESG UQAM, Département de finance, Case postale 8888, succursale Centre-ville, Montréal H3C 3P8, CANADA. alexandre.f.roch@gmail.com.

[‡] Corresponding author. ETH Zürich, Department of Mathematics, HG G 54.3, Rämistrasse 101, 8092 Zürich, Switzerland, and the Swiss Finance Institute, hmsoner@ethz.ch.

an exogenously defined supply curve which gives the price per share as a function of the transaction size and time. In particular, if a trade only affects the current price paid, we are effectively in this setting and the transaction costs depend mainly on the size of the trade. On the other hand, these added liquidity costs are the result of changes in the supply and demand of the asset due to volume imbalances. It is well recorded that large trades can have a lasting impact on the supply and demand of the asset in such a way that future prices will be affected. A part of this impact can be permanent and another part may decay with time. Thus, we see that these two notions are closely related. In this paper, we present a general framework for liquidity risk models for which many existing large trader and price impacts models can be seen as a particular or limiting case. We show it is arbitrage free, we obtain a stochastic discount factor and derive an asset pricing formula which relates the current price to its future payoff and the expected future levels of liquidity.

Our starting point is the seminal paper of Kyle [26] which identifies three equally-important dimensions of liquidity: depth, resilience and tightness. Depth is defined as the size of the order flow required to change prices by one monetary unit. Resilience is the degree to which prices recover from small trades. Tightness is defined as the cost per share of turning around a position. Clearly these three concepts are closely related and pertain to the second approach to liquidity described above, namely the ability to trade without triggering important changes in prices. Tightness is usually directly modeled by considering a bid and an ask price, or equivalently, a transaction cost which is proportional to the trading volume as done in the classical model of Magill & Constantinides [27] and Constantinides [14]. However, an analytical study of liquidity and liquidity risk requires proper modeling of all components of liquidity. Furthermore, liquidity becomes a risk when the future levels of depth, resilience and tightness are stochastic.

There is a large literature on equilibrium models which incorporate liquidity. In equilibrium models, liquidity is often introduced by considering different classes of investors with different information sets, as in the seminal works of Back [7], Back and Baruch [8], Easley and O'Hara [19], Glosten and Milgrom [22], Kyle [26] and Subrahmanyam [34]. Trades then occur between uninformed investors, who can be thought as information gatherers and liquidity providers, and informed investors, who profit from the illiquidity of the market. On the other hand, there is a less extensive literature on liquidity risk from an arbitrage pricing theory perspective. In this setting, liquidity costs and price processes are hypothesized rather than endogenously derived from fundamentals. See for instance Çetin et al. [12] and Roch [30].

Our contribution is to propose an arbitrage-free model, outlined in Subsection 4.3, for liquidity risk from the perspective of liquidity providers with desirable stylized facts justified by empirical studies and theoretical works on equilibrium models. The model is complex enough to model the three components of liquidity but sufficiently mathematically tractable to develop an arbitrage pricing theory and derive an asset pricing formula. To do so, we take the point of view of an investor who observes a limit order book and provides liquidity to better informed traders by making market orders. The investor hypothesizes a probability distribution for a number of dynamic processes that represent features of the limit order book that relate to the three dimensions of liquidity. Moreover, he assumes a certain relation between his trades and the future dynamic of these processes. As a result, this passive investor knows the total cost of trading due to illiquidity and can solve for his optimal holdings in this setting. From first-order conditions, we obtain a stochastic discount factor and the asset pricing relation.

Our model separates liquidity costs due to depth from those related to tightness and allows us to develop a reduced model, given in Subsection 4.4, in which the bid-ask spread is removed in order to focus entirely on the first two dimensions of liquidity which we believe are not as well understood from an arbitrage pricing perspective. We will see that this has the main benefit of obtaining the asset pricing formula. In particular, we make more apparent the interaction between depth and resilience and distinguish two different kinds of liquidity costs that can arise in this framework. We will see that, unlike proportional transaction costs, liquidity costs related to depth and resilience can be defined for strategies that do not have finite variation. This result is particularly important for options hedging and pricing problems, much like the general theory of arbitrage was essential for hedging and pricing in the frictionless case.

The paper is organized as follows. In Section 2, we discuss existing liquidity models that have a relation to our proposed model. In Section 3, we present the model and in Section 4 we derive the dynamics of the wealth process and analyze the liquidity costs in terms of the three dimensions of liquidity described above. A careful choice of the state dynamics and the wealth equation, given by Equation (4.8), is the essential step in obtaining the arbitrage theorem 4.2 and the asset pricing formula (5.4). In Section 5, we derive from first-order condition of a utility maximization problem the asset pricing formula which states that the post-liquidation asset price is the risk neutral expectation of the future post-liquidation asset price minus a weighted integral of the future optimal portfolio holdings. It is interesting to note that this result is in line with many empirical findings, namely the work of Brennan and Subrahmanyam [11] and Amihud and Mendelson [6] who find that measures of liquidity are positively correlated to equity returns. From this, we also obtain the associated stochastic discount factor from the investor's first-order conditions. In the last section, we present the connection between our model and other existing liquidity models in the arbitrage pricing literature by proving convergence results in relation to depth and resilience.

2 Existing models of liquidity

Nowadays, asset prices are frequently obtained through a limit order book (LOB), in which limit orders, i.e. orders to buy or sell a given amount of shares at a specified price, are entered by market participants and kept until a market order comes in to match one of the existing limit orders. As mentioned in the introduction, many liquidity risk models (c.f. [1], [15], [20] and [30]) directly aim to model limit order books in an arbitrage-free setting. The advantage of this approach is that the availability of historical LOB data now allows investors to estimate concerned quantities without the need to develop an equilibrium theory which takes into account many more fundamental, yet less tangible economic variables. The important features include the level of supply and demand, the amount of resilience and the bid-ask spread. There are two different kinds of liquidity costs that have been modeled in this part of the literature. Both approaches have their positive and negative aspects and the goal of this section is to present these models in order to develop a unified framework for arbitrage pricing in the presence of liquidity costs.

Limit order books are sometimes modeled implicitly through a supply curve. In supply curve models, asset prices are hypothesized to be given by stochastic processes $\{S_t(x)\}_{t\geq 0, x\in\mathbb{R}}$ giving at every point in time t the price (per share) for a transaction of size x. This idea was developed in a general semimartingale setup in the paper of Çetin, Jarrow and Protter [12]. In this setup, the liquidity cost as defined by the supply curve model is given by an integral of the square of the gamma of the portfolio with respect to time. We refer to this as the *liquidity costs of the first kind*. Due to its exogenous nature, the supply curve does not however take into account the impact of trading (however small) on its future evolution. An immediate consequence of this fact is that the supply curve effectively has infinite resilience. Indeed, if a trade is made at time t, the supply curve takes the same form at time $t+$ regardless of the size of past trades, including the one at time t. As a result, it can be shown that the optimal strategy is to divide all block trades into infinitesimally smaller ones and execute them at infinitesimally small time intervals. More specifically, Çetin et al. [12] show that only continuous strategies that have finite variation should be used. The trades are then effectively made at the best bid or best ask prices and all liquidity costs associated to large trades are avoided.

To remedy this shortcoming, two approaches were considered in the literature. The first approach is to limit the speed of trading by putting a bound on the gamma of the strategy, i.e. a bound on the speed of change of the position with respect to the asset price. This approach can be found in the paper of Çetin, Soner & Touzi [13]. The economic idea behind this approach is that the speed of resilience of the supply of an asset, namely the time it takes for the limit order book to fill up to its previous level after a market order, is limited to the speed of change of the asset price. By restricting the speed at which changes to portfolio holdings can be made, we effectively restrict the investor to wait for prices to return to their equilibrium values before making the next trade. In discrete time, this idea is more apparent. Indeed, the speed of trading is naturally restricted by the time between time steps. Since trading is not allowed between two time steps in the discrete setting, it is impossible for the investor to break down her trades into smaller ones before the next change in prices. The results of Gökay & Soner [23] show that hedging prices obtained in a discrete model version of Çetin et al. [13] converges to prices obtained in the continuous model, supporting the economic relevance of the assumptions on gamma restrictions in [13]. From our modeling point of view, this approach is, however, not fully adequate as it implicitly assumes full resilience of prices between two time steps.

The second approach is to consider the impact of trading on the supply curve, as advocated in [30]. See also [1], [2], [28] and references therein. By assuming that all trades have an impact on the supply of the asset proportionally to the size of the transaction, a change in the position in the asset will incur essentially the same positive liquidity cost in the short term whether it is passed as one block trade or divided into smaller ones and executed rapidly. The first approach indirectly models the finite resilience of the supply curve by placing restrictions on trading, whereas the second models the lasting impact of trades on the limit order book. Note that a more natural notion of resilience is also present in the papers of Alfonsi et al. [1], [2] and Obizhaeva and Wang [28], in which a portion of the permanent impact is assumed to decay exponentially with time. Most of these papers are concerned with an important but specific problem of the optimal execution of a large order as formulated by Bertsimas and Lo [9]. The typical time horizon of this problem is a few hours, and this is reflected in these models. Our goal, however, is to understand phenomena that go well beyond this time scale. Thus the similarities between these models and ours is only in the short term effects.

Another popular approach in the price impact literature is to model exogenously the liquidity premium in terms of instantaneous price impacts. In contrast to the gamma restriction of Çetin et al. [13], liquidity costs due to temporary imbalances in the supply and demand of the asset are modeled in terms of the speed of trading with respect to time in the papers of Almgren and Chriss [4], Almgren [3] and Schied and Schöneborn [32]. More specifically, the restriction imposed on trading strategies z is that they are assumed to be absolutely continuous with respect to time. Liquidity costs over the time interval $[0, T]$ are then assumed to be given by an integral of the form $\int_0^T h(\dot{z}_s)ds$, in which $\dot{z}$ is the Radon-Nikodym derivative with respect to the time variable. We refer to it as *the liquidity costs of the second kind.* This idea was taken a step further by Gatheral [20] by including a decay factor inside the integral. See also Gatheral et al. [21] in this regard.

To justify this model, Almgren and Chriss [4] first make a setup in discrete time in which the liquidity cost of transactions depends on the inverse of the time between them. In the limit as the time step goes to zero, discrete trades are not allowed so that the limiting model should be thought as an approximation of the discrete setup. Rogers and Singh [31] derive this form of liquidity costs from a construction involving a limit order book, when trades are assumed to have no impact on prices. Starting from a discrete time description of the dynamics of the LOB, it is shown that one captures a liquidity cost by letting the level of the supply vary proportionally to the size of the time step when the time step goes to zero. Clearly, strategies with jumps are very expensive in this setting. Once again, we are forced to only allow absolutely continuous strategies in this setting. In particular, it suggests that this kind of illiquidity is appropriate when depth is low and that resilience is high. We will come back to this idea later in the text.

Note that absolutely continuous strategies incur no liquidity costs of the first kind. However, we will see that when price impacts decay with time, both kinds of liquidity costs are present regardless of the type of strategies used and many models discussed so far can be obtained as limiting cases when the speed of exponential decay and/or the depth converges to infinity. Our results thus offer a straightforward economic justification for known liquidity costs models and provides a general framework that incorporates existing paradigms. In particular, a positive aspect of our setup is the lack of restrictions on the type of strategies allowed.

3 A general model

We consider a financial asset, called the stock, which is actively traded through a limit order book. In this setting, there are two types of trades possible, namely limit orders and market orders. A limit order is an order to buy or to sell the stock at a specific price which is not immediately executed. Limit orders provide liquidity by filling the limit order book. On the other hand, impatient traders can submit market orders (also known as marketable limit orders) which are executed against the existing limit orders and thereby deplete the order book. In this section, we give a static description of the order book and the impact of a trade on its composition.

We are given a trading horizon T with $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{0\leq t\leq T}, \mathbb{P})$ a filtered probability space satisfying the usual conditions. We assume that the spot rate of interest is a constant, and for simplicity we always use discounted price processes throughout.

We take the point of view of a trader (not necessarily large) who only makes market orders. The information that the investor has is represented by the filtration $(\mathcal{F}_t)_{0\leq t\leq T}$. The investor is said to be a liquidity provider in the sense he does not necessarily try to profit from less informed investors by slowly revealing his information through his trades. His trades however has an impact on prices, but rather than making assumptions about preferences and behaviors of other investors, we hypothesize a direct relation between the investor's trades and prices, which we now describe.

We assume that at any time $t \leq T$ traders observe the limit order book and know the average price to pay per share for a transaction of size z via a market order. Following the literature (for instance [1]), we represent the limit order book by two functions ρ_t^+ and ρ_t^-. The quantity $\rho_t^+(s)$ (resp. $\rho_t^-(s)$) denotes the density of the number of shares offered at the ask price (resp. bid price) s at time t. For instance,

$$\int_{s_1}^{s_2} \rho_t^+(s)ds$$

is the total number of shares offered between prices s_1 and s_2. If a trader wants to buy (resp. sell) z shares at time t through a market order then the total dollars paid (resp. obtained) for this order is

$$\int_{a_t}^{s_z} s\rho_t^+(s)ds \quad \text{(resp. } \int_{s_z}^{b_t} s\rho_t^-(s)ds)$$

where s_z solves the equation

$$\int_{a_t}^{s_z} \rho_t^+(s)ds = z \quad \text{(resp. } \int_{s_z}^{b_t} \rho_t^-(s)ds = z), \tag{3.1}$$

and a_t is the smallest price at which $\rho_t^+(s) > 0$ whereas b_t is the highest price at which $\rho_t^-(s) > 0$. The prices b_t and a_t are called best bid and best ask prices.

To understand the above expressions, note that a market order to buy will start at the quoted price a_t obtaining $\rho_t(a_t)ds$ shares at that price, then moving up the limit order book until the price of s_z is paid for the last $\rho_t(s_z)ds$ shares purchased. The total shares purchased is then z.

It is well known and very intuitive that the most optimal strategies in this setting consist in trading small quantities so that the price paid is never too far from the best bid and the best ask prices. In other words, one rarely execute orders deep in the limit order book. As a result, we make the simplifying assumption that the limit order book has a constant density $\frac{1}{2m_t}$ outside the bid ask spread at time t, in which $m = (m_t)_{t\geq 0}$ is a given adapted stochastic process. This simplifying assumption is supported by the empirical evidence in Blais and Protter [10], especially for frequently traded and large volume stocks. Furthermore, Huberman & Stanzl [24] provide theoretical justification based on the notion of quasi-arbitrage. In the constant density case, $s_z = a_t 1_{\{z>0\}} + b_t 1_{\{z<0\}} + 2m_t z$ and the total dollars paid (resp. obtained) for z shares is

$$\frac{1}{2m_t}\int_{a_t}^{a_t+2m_t z} zdz = a_t z + m_t z^2$$

$$\text{(resp. } \frac{1}{2m_t}\int_{b_t+2m_t z}^{b_t} zdz = b_t z + m_t z^2).$$

The depth at time t, i.e. the size of the order flow required to shift prices by one dollar, is $\frac{1}{2m_t}$. Consequently, m_t in our supply curve is a measure of *illiquidity*. Indeed, the larger is m_t, the larger is the price impact realized to trade z shares.

Recall that in our market, a trade occurs when a market order is placed. As a result of this market order, limit orders in the limit order book are executed starting with the cheapest to the most expensive (in the case of a buying order) until the total number of shares ordered is reached as in expression (3.1). Once the order is executed, the respective limit orders disappear and a gap is created in the limit order book. For example, after a purchase of z shares at time t, the best bid price should stay at b_t whereas the best ask price should move to $a_t + 2m_t z$. This means that immediately after this trade, the limit order book density function would be 0 for prices between a_t and $a_t + 2m_t z$ and remains unchanged elsewhere.

So far, this is a static view of the limit order book in the sense that the processes a, b and m have not yet been specified in dynamical terms. We give a precise definition of the bid and ask processes b and a in terms of z and other hypothesized processes in the next section.

3.1 Bid and ask prices

The starting point of our dynamic model, as in many other financial models, is an *equilibrium* stock price process $\{S_t^e\}_{t\geq 0}$. This process may not be observable and it is the theoretical value of the stock which will be only observed in the long run when the trader stops trading. In the context of Kyle [26] (also in Back [7]) this might be seen as the price that is known by the informed trader. See Remark 4.1 below for a more detailed discussion. In general, it can be constructed from other primitives such as preferences, endowments and specific microstructure features of the market. Since this side of the theory is very well established, we just simply assume that it is a continuous semimartingale.

The other important processes are the bid and the ask processes $\{a_t\}_{t\geq 0}$ and $\{b_t\}_{t\geq 0}$. By definition, we take $b_t \leq a_t$. It is also assumed that the bid and the ask processes converge to S_t^e *in the absence* of any further portfolio activity when t goes to infinity. This will be made more precise in the following section. The speed of this time decay is clearly a model parameter which relates to the resilience of the supply and demand of the asset.

Following the literature, we assume that there is a *"mid-quote"* stock price process $\{S_t\}_{t\geq 0}$ as well. This process again may not be observable and, as for the bid and the ask processes, it depends on the portfolio activity of the trader. It is the value to which the bid and ask processes converges when the impact of trades on the bid-ask spread has vanished. Since in general one of these two processes may converge faster than the other, this is typically not the exact arithmetic average of the bid and ask prices. However it always falls in between. The *definition of the mid-quote price is essential* in our model to separate the effects of depth from the effects of tightness or equivalently, the bid-ask spread. In particular, the wealth-like process Y in absence of bid-ask spread (defined in (4.3), below) is obtained from S.

Note that we have two types of resilience in the model, modeled as time decays, namely a time decay of the bid-ask spread and a time decay of the trader's impact of the observed prices.

4 Trading strategies and price dynamics

In this subsection, we postulate the dynamics of the price processes described in the previous subsection for simple trading strategies and extend this definition to more general trading strategies. Our goal is to obtain a mathematically tractable dynamic equation for the value of a portfolio, or wealth process, as given by Equation (4.8). The derivation in this section is thus crucial given that Equation (4.8) is the key result needed to derive the no-arbitrage theorem and the existence of a stochastic discount factor.

We assume that *all processes are càdlàg* (i.e., they are right continuous and have left limits) and if needed the initial value of any process is given as its left limit. We summarize the trading strategy of the investor by $\{z_t\}_{t\geq 0}$. It is simply the number of shares of the risky asset held at time t. It is well known that in a model with a bid-ask spread, z has to be of bounded variation. We then represent it as $z = z^+ - z^-$ with $z^\pm$ non-decreasing processes. In that case, z_t^+ represents the cumulative number of shares bought up to time t, whereas z_t^- represents the cumulative number of shares sold up to time t.

We start by describing the dynamics of the price processes and the liquidation value of the portfolio for a simple strategy of the form

$$z_t = z_{0-} + \sum_{0\leq k\leq n} \xi_k \chi_{\{\tau_k \leq t\}}, \tag{4.1}$$

where $0 = \tau_0 \leq \tau_1 \leq \cdots \leq \tau_n \leq T$ is a sequence of stopping times and ξ_k is a $\mathcal{F}_{\tau_k}$-measurable random variables for $k \geq 0$. We will then postulate a dynamic for the liquidation value for general trading strategies which will be consistent with this special case, in a sense which will be explained in Section 6. Suppose that at time τ_i a positive buy order of size $\xi_i > 0$ is executed. In view of the discussion on the limit order book, the ask price increases by an amount of $\Delta a_{\tau_i} = 2m_{\tau_i}\xi_i$. Here, we think of a_t as the observed price after the trade at time t so that the ask price in effect when the trade at time t is made is a_{t-}. Similarly, we define the increment of S at time τ_i by $2m_{\tau_i}\xi_i$. The upper bid-ask spread is the difference $\alpha_t := a_t - S_t$, whereas the lower part of the bid-ask spread is defined by $\beta_t := S_t - b_t$. Since we assume that the bid and ask prices converge to the mid-quote price exponentially fast, we have the following equations for α and β:

$$\begin{aligned}
d\alpha_t := d(a_t - S_t) &= -\hat{\kappa}_t \alpha_t dt + \sum_{i=1}^{n} 2m_{\tau_i} \xi_i \chi_{\{\tau_i = t, \xi_i < 0\}} = -\hat{\kappa}_t \alpha_t dt + 2m_t dz_t^- \\
&= -\hat{\kappa}_t \alpha_t dt + 2m_{t-} dz_t^- + 2d[m, z^-]_t, \\
d\beta_t := d(S_t - b_t) &= -\hat{\kappa}_t \beta_t dt + \sum_{i=1}^{n} 2m_{\tau_i} \xi_i \chi_{\{\tau_i = t, \xi_i > 0\}} = -\hat{\kappa}_t \beta_t dt + 2m_t dz_t^+ \\
&= -\hat{\kappa}_t \beta_t dt + 2m_{t-} dz_t^+ + 2d[m, z^+]_t,
\end{aligned}$$

where $\hat{\kappa}_t > 0$ is a relaxation parameter and $d[m, z^{\pm}]$ is the covariation between the processes m and $z^{\pm}$, see [29]. In the above discrete setting, it is simply given by $d[m, z^{\pm}] = \Delta m_t \Delta z_t^{\pm}$ for $t = \tau_i$ and is zero at other times.

The price S_t is always between the best bid and best ask prices and is interpreted as the observed mid-quote price when the bid-ask spread vanishes. Moreover, as stated before, it is assumed that this mid-quote price converges back to the equilibrium price in the long run when the trader stops trading.

Let $\ell_t = S_t - S_t^e$ represent the difference between the mid-quote price and the equilibrium price. We postulate the following connection between the mid-quote and the equilibrium price,

$$d\ell_t = -\kappa_t \ell_t dt + 2m_{t-} dz_t + 2d[m, z]_t$$

in which κ_t is the given parameter of *resilience* of the asset's mid-quote price. We allow the processes $\kappa_t, \hat{\kappa}_t$ to depend on state variables such as S_t, S_t^e and z_t. When trading stops, i.e. when z remains constant after some time t_0, then

$$S_t - S_t^e = (S_{t_0} - S_{t_0}^e) \frac{\Lambda_t}{\Lambda_{t_0}}$$

in which $\Lambda_t := \exp\left(-\int_0^t \kappa_u du\right)$. Although this is mainly a phenomenological assumption, it is in fact a very general model for decay of price impacts since we allow the process κ to depend on other state variables. Indeed, any positive decreasing process Λ_t which is absolutely continuous can be written in the form $\Lambda_0 \exp\left(-\int_0^t \kappa_u du\right)$ for some integrable process κ. Without loss of generality we can also take $\Lambda_0 = 1$. Furthermore, we will argue in Remark 4.1 below, that this choice for the mid-quote dynamics is consistent with the continuous auction equilibrium described by Kyle [26] and in Back [7].

Clearly, ℓ, α, β can be solved in terms of z. For future reference, we provide these formulae

$$\begin{aligned}
\ell_t &= \Lambda_t \left(\ell_{0-} + 2\int_0^t (\Lambda_u)^{-1}\ m_{u-} dz_u + 2\int_0^t (\Lambda_u)^{-1}\ d[m,z]_u \right) \qquad (4.2)\\
\alpha_t &= \hat{\Lambda}_t \left(\alpha_{0-} + 2\int_0^t (\hat{\Lambda}_u)^{-1}\ m_{u-} dz_u^- + 2\int_0^t (\hat{\Lambda}_u)^{-1}\ d[m,z^-]_u \right),\\
\beta_t &= \hat{\Lambda}_t \left(\beta_{0-} + 2\int_0^t (\hat{\Lambda}_u)^{-1}\ m_{u-} dz_u^+ + 2\int_0^t (\hat{\Lambda}_u)^{-1}\ d[m,z^+]_u \right)
\end{aligned}$$

with the decay factors defined as

$$\Lambda_t := \exp\,(-\int_0^t \kappa_u du), \qquad \hat{\Lambda}_t := \exp\,(-\int_0^t \hat{\kappa}_u du).$$

Note that the best bid and ask processes a and b also converge to S^e in the long run since $a = S + \alpha$ (resp. $b = S + \beta$) and α_t (resp. β_t) converges to zero when trading stops and t goes to infinity.

More extensions to this model are possible. For instance, one can assume that some part of the cumulative trade impacts ℓ_t do not decay with time so that the process S converges to a secondary process S^0 which includes permanent impacts proportional to ℓ_t. On the other hand, one can assume that the lower and upper part of the bid-ask spread converges to some given constants when t goes to infinity and the trader stops trading. This can be obtained by replacing the drift by $-\hat{\kappa}(\alpha - \alpha_0)$ for instance. Furthermore, there is no need to assume that the speed of decay of the lower part of the bid-ask spread is the same of the upper part. However, none of these generalizations would offer any new insights for our purposes, so to keep the notation as simple as possible we will not further consider these extensions.

Note that tightness in this model is time varying since the best bid and best ask prices follow different dynamics. The cost of turning around a position is directly proportional to the bid-ask spread.

Remark 4.1 We argue that the above dynamics (4.2) for the mid-quote price is consistent with the dynamic equations derived in [26]. A similar comparison to the model of Back [7] can also be made. Indeed, in Kyle's model the informed traders know the future value of the stock (which is denoted by v in that paper) and trade accordingly. In our setting the equilibrium price S_t^e is the analog of v and the trader we consider is a group of noise traders. In [26], the illiquidity parameter $\lambda(t)$ is derived as a function of the volatilities of several quantities. Then, the continuous auction equilibrium of [26] (c.f., equations (4.2) and (4.3) on page 1326 of [26]) is described by,

$$\begin{aligned}
dx(t) &= \beta(t)[v - p(t)]dt,\\
dp(t) &= \lambda(t)[dx(t) + du(t)].
\end{aligned}$$

where $p(t)$ is the price, $u(t)$ is the trading strategy of the noise traders and $x(t)$ is the strategy of the insider. So we can make the following correspondences: $\lambda(t) = 2m_t$, $u(t) = z_t$, $p(t) = S_t$, $v = S^e_t$. Using these and the above equations, we obtain the following equation for $\ell(t) = S(t) - S^e(t) = p(t) - v$,

$$\begin{aligned} d\ell(t) &= d[p(t) - v] = \lambda(t)[dx(t) + du(t)] = -\lambda(t)\beta(t)[p(t) - v]dt + \lambda(t)du(t) \\ &= -\kappa(t)\ell(t)dt + 2m_t dz(t), \end{aligned}$$

with $\kappa(t) = \beta(t)\lambda(t)$. This is precisely Equation (4.2).

4.1 Portfolio Value

Since the changes in the portfolio composition result in a liquidity cost, the value of the portfolio needs to be carefully defined. Indeed, as in the models for proportional transaction costs (c.f., Magill & Constandinides [27], Constandinides [14], Davis & Norman [16], Dumas & Luciano [18] and Soner & Shreve [33]) or in the discrete model for liquidity of Gokay & Soner [23], we need to keep track of z as well as the position in the money market account for the valuation of the wealth process. Let x_t be the position in this account. Then, in our model we either need the pair (x, z) or any other invertible transform of this pair. One choice of such a transform is

$$Y_t := x_t + z_t(S_t - m_t z_t). \tag{4.3}$$

This quantity is interpreted as the liquidation value of the portfolio after ignoring the bid-ask spread. Clearly, the right choice of the mid-quote price S is crucial in the definition of Y. Indeed, the structure of the dynamic equation (4.2) is essential in the derivation of several desired properties of Y which are outlined in the next subsection.

4.2 Dynamics

In order to obtain the dynamics for the liquidation value process Y, we first assume that z is given by (4.1). To ease the notation, for any càdlàg process v, we write

$$\Delta^k v := v_{\tau_k} - v_{\tau_k^-}.$$

In view of the discussion on the limit order book, the change in the money market account at time τ_k is given by the total cost of the trade at time τ_k, for a self-financing strategy. Namely,

$$\begin{aligned} \Delta^k x &= -\xi_k \left(a_{\tau_k^-} + m_{\tau_k}\xi_k\right) \chi_{\{\xi_k>0\}} - \xi_k \left(b_{\tau_k^-} + m_{\tau_k}\xi_k\right) \chi_{\{\xi_k<0\}} \\ &= -\xi_k S_{\tau_k^-} - m_{\tau_k} |\xi_k|^2 - |\xi_k| \left(\alpha_{\tau_k^-}\chi_{\{\xi_k>0\}} + \beta_{\tau_k^-}\chi_{\{\xi_k<0\}}\right). \end{aligned}$$

We now directly calculate that

$$\begin{aligned}
\Delta^k Y &:= \Delta^k x + (z_{\tau_k^-} + \xi_k)\left(S_{\tau_k^-} + 2m_{\tau_k}\xi_k - m_{\tau_k}(z_{\tau_k^-} + \xi_k)\right) \\
&\quad - z_{\tau_k^-}\left(S_{\tau_k^-} - m_{\tau_k^-} z_{\tau_k^-}\right) \\
&= -\left|\xi_k\right|\left(\alpha_{\tau_k^-}\chi_{\{\xi_k>0\}} + \beta_{\tau_k^-}\chi_{\{\xi_k<0\}}\right) - (z_{\tau_k^-})^2\Delta^k m,
\end{aligned}$$

where we uses the fact that $S_{\tau_k} = S_{\tau_k^-} + 2m_{\tau_k}\xi_k$. The quantity

$$|\xi_k|\left(\alpha_{\tau_k-}\chi_{\{\xi_k>0\}} + \beta_{\tau_k-}\chi_{\{\xi_k<0\}}\right)$$

is the proportional transaction cost associated to the trade at time τ_k. As a result, tightness, defined as the cost of turning around a position per share, is given by $\alpha_{\tau_k-} + \beta_{\tau_k-}$ (the bid-ask spread) at time t.

Since the interest rate is taken to be zero and since $z_t = z_{\tau_{k-1}}$ on the interval $[\tau_{k-1}, \tau_k)$,

$$\begin{aligned}
Y_{\tau_k^-} - Y_{\tau_{k-1}} &= z_{\tau_{k-1}}\left(\left(S^e_{\tau_k^-} - S^e_{\tau_{k-1}}\right) + \left(\ell_{\tau_k^-} - \ell_{\tau_{k-1}}\right)\right) - (z_{\tau_{k-1}})^2(m_{\tau_k^-} - m_{\tau_{k-1}}) \\
&= \int_{\tau_{k-1}}^{\tau_k^-} z_{u^-}\left(dS^e_u - \kappa_u\ell_u du\right) - \int_{\tau_{k-1}}^{\tau_k^-} z^2_{u^-} dm_u.
\end{aligned}$$

In the above, we also used Equation (4.2). The above calculation shows that for $k \geq 1$ and $t \in [\tau_k, \tau_{k+1})$

$$\begin{aligned}
Y_t &= Y_{0-} + \Delta^0 Y + \sum_{i=1}^{k}\left(Y_{\tau_i^-} - Y_{\tau_{i-1}} + \Delta^i Y\right) + Y_t - Y_{\tau_k} \\
&= Y_{0-} + \int_0^t z_{u^-}\left(dS^e_u - \kappa_u\ell_u du\right) - \int_0^t z^2_{u^-} dm_u - T_t \qquad (4.4)
\end{aligned}$$

where the cumulative cost of proportional transaction cost T_t is given by

$$T_t := \int_0^t \left(\alpha_{u^-} dz^+_u + \beta_{u^-} dz^-_u\right).$$

This cost is directly related to the notion of tightness of the limit order book.

For a general portfolio process z, which is càdlàg, we postulate the dynamics of Y_t to be given by (4.4). We will see in Proposition 6.3 that this definition is stable in the sense that any sequence of of liquidation values corresponding to a sequence of discrete approximations of z converge to the liquidation value with this dynamics.

4.3 Model Overview

In this short subsection, we summarize the above model. The only control process is

$$z_t := z^+_t - z^-_t = \text{the number of risky assets held at time } t,$$

where both $z_t^{\pm}$ are non-decreasing with z_t^+ is the total stock purchases up to time t and the total sales is given by z_t^-. Five state variables are

$$
\begin{aligned}
S_t^e &:= \text{equilibrium value of the stock,}\\
\ell_t &:= S_t - S_t^e, \quad S_t := \text{mid-quote price,}\\
\alpha_t &:= a_t - S_t, \quad a_t := \text{best ask price,}\\
\beta_t &:= S_t - b_t, \quad b_t := \text{best bid price,}\\
Y_t &:= x_t + z_t(S_t - m_t z_t), \quad x_t := \text{position in the money market.}
\end{aligned}
$$

Notice that one may use S instead of ℓ and also use x instead of Y. However, in what follows this choice of $(S^e, \ell, \alpha, \beta, Y)$ is the most convenient one. (In the theory of optimal control, formally, z must be considered as a state variable while the control is dz. However, we do not need this subtle difference in what follows).

The model description will be complete when the dynamics of the state variables are given. So we recall the dynamics (4.2), (4.4) for the reader's convenience.

$$
\begin{aligned}
d\alpha_t &= -\hat{\kappa}_t \alpha_t dt + 2m_{t^-} dz_t^- + 2d[m, z^-]_t, \qquad (4.5)\\
d\beta_t &= -\hat{\kappa}_t \beta_t dt + 2m_{t^-} dz_t^+ + 2d[m, z^+]_t,\\
d\ell_t &= -\kappa_t \ell_t dt + 2m_{t^-} dz_t + 2d[m, z]_t,\\
dY_t &= z_{t^-}\,(dS_t^e - \kappa_t \ell_t dt) - (z_{t^-})^2 dm_t - dT_t,
\end{aligned}
$$

where the coefficients $\hat{\kappa}$, κ and m are given adapted processes and

$$
T_t := \int_0^t [\alpha_{u^-} dz_u^+ + \beta_{u^-} dz_u^-]. \qquad (4.6)
$$

Finally, the equilibrium stock price process S^e can be taken as a general continuous semimartingale.

The above description assumes that the portfolio process is a semimartingale as dz terms appear in the dynamics. However, by an appropriate rewriting of the equations would allow us to define ℓ even for càdlàg z processes. Indeed, from (4.2), we see that the process ℓ can be written in the form

$$
\begin{aligned}
\ell_t &= \Lambda_t \left(2\int_0^t \frac{1}{\Lambda_s} m_{s-} dz_s + 2\int_0^t \frac{1}{\Lambda_s} d[z, m]_s \right)\\
&= \Lambda_t \left(-2\int_0^t \frac{1}{\Lambda_s} z_{s-} dm_s + 2\int_0^t \frac{1}{\Lambda_s} d(mz)_s \right)\\
&= 2(m_t z_t - \int_0^t z_{s-} dm_s) - 2\int_0^t \kappa_s (m_{s-} z_{s-} - \int_0^{s-} z_u dm_u) \frac{\Lambda_t}{\Lambda_s} ds\\
&= \ell_t^0 - \int_0^t \kappa_s \ell_{s-}^0 \frac{\Lambda_t}{\Lambda_s} ds, \qquad (4.7)
\end{aligned}
$$

where $\ell_t^0 = 2m_t z_t - 2\int_0^t z_{s-} dm_s$. We thus define ℓ for a general process z from (4.7) (in fact, for technical reasons from stochastic integration theory, by a general process we mean a càdlàg process, i.e, a process which left continuous with right limits). This definition of ℓ offers a lot of stability in terms of the convergence of liquidation values of approximating sequences to the liquidation value of the limit. See the results of Section 6 below in this regard.

However, for Y we still need z to be of bounded variation due to dT term. In order to define a liquidation value for general càdlàg strategies, we need to omit this term in (4.4). This is the content of the next subsection.

4.4 Reduced Model

As discussed in the introduction, we would like to separate the costs that are proportional to transactions and the continuous trading. This is achieved simply by either ignoring the T_t term in the above equations, or equivalently, by formally sending the parameter $\hat{\kappa}$ to infinity. Then, the bid and ask spreads are not needed in the description of the model. Hence, this simplified market contains three state variables S^e, ℓ, Y. The dynamics of the first two processes are as before, whereas

$$dY_t \quad = \quad z_{t^-}\left(dS_t^e - \kappa_t \ell_t dt\right) - (z_{t^-})^2 dm_t. \tag{4.8}$$

We may achieve further simplicity by assuming that κ_t is a constant and that m_t is either a constant or a constant multiple of S_t^e as done in Remark 4.3, below.

It is intuitively clear that in any model with illiquidity the portfolio value should be less than the value in an infinitely liquid model. However, in models with market impact this is not immediately obvious. Hence, we continue by showing that the liquidity cost in the reduced model is non-negative in the mean. This in turn would imply that this model is free of arbitrage and it properly isolates the effects of depth and resilience on liquidity costs.

We also would like to emphasize that the reduction of the model is not a routine process. Although it is clear that such a reduction is done by simply setting the bid and the ask spreads zero, these spreads are defined with respect to a mid-quote price S. Hence, *the definition of S is the important point in this reduction.* Indeed, our choice of (4.2) is not only consistent with [26] (as argued in Remark 4.1) but it allows us to prove the following no-arbitrage result, Theorem 4.2, and the convergence results of Section 6.

As we can see in (4.8), the difference in the liquidation value obtained in a frictionless market $Y_{0-} + \int_0^t z_u dS_u^i$ and our current liquidity risk setup is given by

$$L_t = \int_0^t z_u \kappa_u \ell_u du + \int_0^t z_{u^-}^2 dm_u. \tag{4.9}$$

Indeed, L_t is the *liquidity costs associated to the depth and the resilience of the limit order book.*

In the case $\kappa \equiv 0$, we are effectively in the liquidity risk model of Roch [30], in which it is known that the stochastic nature of the level of the order book $1/2m_t$ can lead to higher gains than in the frictionless case. In other words, the integral $\int_0^t z_{u-}^2 dm_u$ may take negative values. This is due to the fact that the impact of a trade, which is proportional to m_t, can be bigger than when this position is liquidated due to a decrease in m. As a result, Roch [30] showed that there is no arbitrage in the case $\kappa \equiv 0$ if there exists an equivalent measure that makes S^e a local martingale and m a local submartingale. On the other hand, when m is constant, only the integral $\int_0^t z_u \kappa_u \ell_u du$ remains and intuitively this term should be related to the part of the value of the portfolio lost due to time decay of temporary price impacts.

We have the following result:

Theorem 4.2 *Let L_t be as in* (4.9)*. Then $L_t \geq 0$, if m_t is constant. In general, if we assume that $\phi_t := \Lambda_t^2/m_t$ is a local super-martingale, then*

$$\mathbb{E}[L_t] \geq 0, \qquad \forall\, t \geq 0.$$

Proof. Set

$$f_t := \ell_t - 2m_t z_t,$$

so that

$$df_t = -\kappa_t \ell_t dt - 2z_{t-} dm_t.$$

We directly calculate that

$$\begin{aligned}\int_0^t z_u \kappa_u \ell_u du &= -\int_0^t z_{u-} df_u - 2\int_0^t z_{u-}^2 dm_u \\ &= \int_0^t \frac{1}{2m_{u-}}(f_{u-} - \ell_{u-}) df_u - 2\int_0^t z_{u-}^2 dm_u \\ &= \int_0^t \frac{1}{4m_{u-}} d(f_u^2) - \int_0^t \frac{1}{m_{u-}} z_{u-}^2 d[m]_u + \int_0^t \frac{1}{2m_{u-}} \kappa_u (\ell_{u-})^2 du \\ &\quad + \int_0^t \frac{1}{m_{u-}} z_{u-} f_{u-} dm_u.\end{aligned}$$

In particular, if m is constant,

$$L_t = \int_0^t z_u \kappa_u \ell_u du = \frac{(f_t)^2}{4m} + \int_0^t \frac{1}{2m} \kappa_u (\ell_u)^2 du \geq 0.$$

If m is not constant, then

$$\begin{aligned}\int_0^t z_u\kappa_u\ell_u du &= \frac{1}{4m_t}f_t^2 - \frac{1}{4}\int_0^t f_{u^-}^2 d\left(\frac{1}{m_u}\right) - \frac{1}{2}\int_0^t f_{u^-} d\left(f, \frac{1}{m_u}\right) \\ &+ \int_0^t \frac{1}{2m_u}\kappa_u(\ell_u)^2 du + \int_0^t \frac{1}{m_{u^-}} z_{u^-} f_{u^-} dm_u - \int_0^t \frac{1}{m_{u^-}} z_{u^-}^2 d[m]_u \\ &= \frac{1}{4m_t}f_t^2 - \frac{1}{4}\int_0^t \ell_{u^-}^2 d\left(\frac{1}{m_u}\right) + \int_0^t \frac{1}{2m_u}\kappa_u(\ell_u)^2 du - \int_0^t z_{u^-}^2 dm_u,\end{aligned}$$

after many simplifications. Hence,

$$L_t = -\frac{1}{4m_t}f_t^2 + \frac{1}{4}\int_0^t \ell_{u^-}^2 \Lambda_u^{-2} d\phi_u$$

in which $\phi_t := \Lambda_t^2/m_t$. From this last equation, we see that if the process ϕ is a super-martingale then L_t is non-negative in expectation. ■

In general, the proportional transaction costs given by T in Equation (4.6) are always non-negative, so the previous result also applies to the non-reduced model of Equation (4.4).

For future reference, we record the above calculations for the Y dynamics as

$$Y_t = Y_{0-} + \int_0^t z_{u^-} dS_u^e - \frac{1}{4m_t}f_t^2 + \frac{1}{4}\int_0^t \ell_{u^-}^2 \Lambda_u^{-2} d\phi_u. \tag{4.10}$$

Remark 4.3 Suppose that

$$\kappa_t \equiv \kappa, \quad \frac{m_t}{S_t^e} \equiv m, \quad \frac{dS_t^e}{S_t^e} = \mu dt + \sigma dB_t, \tag{4.11}$$

where κ, m, μ, σ are non-negative constants and B is a standard Brownian motion. Then, by a direct calculation we conclude that ϕ is a super-martingale only if

$$\kappa \geq \frac{\sigma^2}{2} - \mu.$$

■

The standard definition of an arbitrage opportunity, c.f. [25], is a wealth process Y, as defined in (4.10), which is bounded from below by a constant and which satisfies

$$\mathbb{P}\{Y_T \geq 0\} = 1 \qquad \text{and} \qquad \mathbb{P}\{Y_T > 0\} > 0. \tag{4.12}$$

We conclude this section with the following theorem which characterizes the absence of arbitrage opportunities.

Theorem 1 *Suppose there exists a measure $\mathbb{Q}$ equivalent to $\mathbb{P}$ under which S^e is a local martingale and $\phi_t := \Lambda_t^2/m_t$ is a supermartingale, then there are no arbitrage opportunities.*

Proof. By the Doob-Meyer decomposition theorem there exists a $\mathbb{Q}$-local martingale $\widetilde{M}$ and a decreasing predictable process A such that $\phi = \widetilde{M} + A$. Since,

$$Y_t = Y_{0-} + \int_0^t z_{u-} dS_u^e - \frac{1}{4m_t} f_t^2 + \frac{1}{4} \int_0^t \ell_{u-}^2 \Lambda_u^{-2} d\phi_u,$$

then

$$Y_t + \frac{1}{4m_t} f_t^2 - \frac{1}{4} \int_0^t \ell_{u-}^2 \Lambda_u^{-2} dA_u = \int_0^t z_{u-} dS_u^e - \frac{1}{4} \int_0^t \ell_{u-}^2 \Lambda_u^{-2} d\tilde{M}_u \geq -\alpha,$$

for some constant α, where we used that A is decreasing. Now, S^e and $\widetilde{M}$ are $\mathbb{Q}$-local martingales hence $Y_t + \frac{1}{4m_t} f_t^2 - \frac{1}{4} \int_0^t \ell_{u-}^2 \Lambda_u^{-2} dA_u$ is also a local martingale and because it is bounded from below it is a supermartingale. Therefore, Y is also a $\mathbb{Q}$-supermartingale and $\mathbb{E}^{\mathbb{Q}}[Y_T] \leq 0$. But, because $\mathbb{Q}$ is equivalent to $\mathbb{P}$, if Y_T were an arbitrage opportunity, Equation (4.12) would also be satisfied by $\mathbb{Q}$ as well. This would imply that $\mathbb{E}^{\mathbb{Q}}[Y_T] > 0$ which is in clear contradiction with the supermartingale property of Y under $\mathbb{Q}$. ■

One should compare the previous no arbitrage result to the classical no arbitrage result in a frictionless market for which the martingale condition only applies to the process S^e. In a liquidity risk setting, the additional supermartingale condition on the process ϕ must also be satisfied under the risk neutral measure of S^e in order to rule out arbitrage opportunities. The absence of arbitrage opportunities now allows us to solve the utility maximization problem of the following section.

5 The Asset Pricing Formula

In this section, we study the equilibrium consequences of the reduced model without the proportional transaction costs as outlined in Subsection 4.4. In particular, we consider an investor who maximizes the expected utility from final wealth, and we assume the no arbitrage conditions of Theorem 1 are satisfied. From first-order conditions, we obtain a stochastic discount factor and an asset pricing formula. Note that in most proportional transaction costs models investors may not trade or consume at the marginal rate of substitution since frictions may make it suboptimal to trade at this point. For this reason, we consider the reduced model in this section.

Let $\kappa > 0$ and $m \geq 0$ be given adapted processes and m is a semimartingale. Let S^e be a non-negative semimartingale and $\mathcal{Z}$ be the set of all càdlàg processes z such that z_- is integrable with respect to S^e and the Lebesgue measure, and z_-^2 is integrable with respect to m. Then, for $z \in \mathcal{Z}$ consider the controlled processes $Y^z, f^z, \ell^{0,z}$ given by

$$\begin{aligned}
\ell_t^{0,z} &= 2m_t z_t - 2\int_0^t z_{u-} dm_u \\
\ell_t^z &= \ell_t^{0,z} - \int_0^t \kappa_s \ell_s^{0,z} \frac{\Lambda_t}{\Lambda_s} ds, \\
df_t^z &= -\kappa_t \ell_t^z dt - 2z_{t-} dm_t \\
dY_t^z &= z_{t-}\ dS_t^e - \kappa_t \ell_t^z z_t dt - z_{t-}^2 dm_t.
\end{aligned}$$

We also have the following equivalent representation for the process f^z:

$$f_t^z = \ell_t^z - 2m_t z_t.$$

Given a concave, non-decreasing utility function U satisfying the Ineda conditions (see Karatzas & Shreve [25]), we consider the optimization problem

$$\sup_{z \in \mathcal{Z}_{adm}} \ \mathcal{U}(z), \qquad \mathcal{U}(z) := \mathbb{E}^{\mathbb{P}}[U(Y_T^z)],$$

where $\mathcal{Z}_{adm}$ is the set of all admissible portfolios so that Y^z is uniformly bounded from below by a constant. We relegate the important question of existence of a maximizer to future work and assume that there exists $z^* \in \mathcal{Z}_{adm}$ which maximizes the above problem.

Let

$$S_t^* := S_t^e + \ell_t^* - 2m_t z_t^* = S_t^e + f_t^*, \tag{5.1}$$

be the *post-liquidation stock price* (observed for strategy z^*). It is the price that we observe at time t, after the portfolio z^* is liquidated at time t. This is in fact a more accurate expression for the value of the stock for an outside investor since the observed stock price S_t is temporarily inflated due to the current position of the trader.

Now, introduce a probability measure $\mathbb{Q}$, which is the analogue of the risk neutral measure, by the following stochastic discount factor:

$$\frac{d\mathbb{Q}}{d\mathbb{P}} := \frac{U'(Y_T^*)}{\mathbb{E}^{\mathbb{P}}[U'(Y_T^*)]}, \tag{5.2}$$

in which Y_T^* is the liquidation value of the portfolio z^*. In the above, we assume that $\mathbb{E}^{\mathbb{P}}[U'(Y_T^*)]$ is finite.

Under $\mathbb{Q}$, S^* is *not* a martingale. However Theorem 5.1 below states that

$$\widetilde{S}_t^* \ = \ S_t^* + \mathbb{E}^{\mathbb{Q}}\left(L^*(t,T)|\mathcal{F}_t\right), \tag{5.3}$$

is a $\mathbb{Q}$-martingale, in which

$$L^*(t,T) = 2m_t \int_t^T \kappa_u \frac{\Lambda(u)}{\Lambda(t)} z_u^* du.$$

Equivalently,

$$S_t^* = \mathbb{E}^{\mathbb{Q}}\left(S_T^* - L^*(t,T)|\mathcal{F}_t\right) = \mathbb{E}^{\mathbb{Q}}\left(S_T^*|\mathcal{F}_t\right) - 2m_t \int_t^T \mathbb{E}^{\mathbb{Q}}\left(\kappa_u \frac{\Lambda(u)}{\Lambda(t)} z_u^* |\mathcal{F}_t\right) du. \tag{5.4}$$

Even though our asset pricing formula gives the price in terms of an endogenous variable, the optimal holdings, it still gives valuable predictions about prices without making any more specific assumptions like completeness of the market, specific distributional assumptions on specified processes or market clearing conditions. In particular, the previous equation states that the current post-liquidation stock price is the risk neutral expected value of the final post-liquidation stock price minus a weighted average of the future optimal portfolio positions. In other words, the price today exhibits a liquidity adjustment to take into account the expected effects of the trader's future holdings on future prices Moreover, the second part of the equation shows that our model predicts that the liquidity adjustment is proportional to the current value of the liquidity parameter m_t, i.e. inversely proportional to the depth of the limit order book. These two results are in line with the empirical findings of Amihud and Mendelson [6], Amihud [5], and Brennan and Subrahmanyam [11].

We have the following result.

Theorem 5.1 *Suppose that the utility maximization admits a maximizer $z^* \in \mathcal{Z}_{adm}$ such that*

$$\mathbb{E}^{\mathbb{P}}\left(U'(Y_T^*)\right) < \infty.$$

Let $\mathbb{Q}$ be given in (5.2). *Then, the liquidity-adjusted price process $\widetilde{S}^*$, defined in Equation* (5.3)*, is a $\mathbb{Q}$-martingale.*

Proof.

Let θ be a stopping time on $[0,T]$. For all $\epsilon > 0$, define $z^\epsilon = z^* + \epsilon\chi_{[\theta,T]}$. By the linearity of the equation satisfied by the ℓ process, we conclude that $\ell^{z^\epsilon} = \ell^* + \epsilon\ell^\theta$, where ℓ^* is the process ℓ^z for $z = z^*$, and ℓ^θ is the process ℓ^z for $z = \chi_{[\theta,T]}$. More specifically, we find that $l_t^\theta = 2m_\theta e^{-\int_\theta^t \kappa_s ds}\chi_{[\theta,T]}(t)$. Then, we directly calculate that

$$\begin{aligned} Y_t^{z^\epsilon} &= Y_t^* - \epsilon^2\left(2m_\theta \int_0^t \kappa_u e^{-\int_\theta^u \kappa_s ds}\chi_{[\theta,T]}(u)du + \int_0^t \chi_{(\theta,T]}(u)dm_u\right) \\ &\quad +\epsilon\left(\int_0^t \chi_{(\theta,T]}(u)dS_u^e - \int_0^t \kappa_u \ell_u^* \chi_{[\theta,T]}(u)du\right) \\ &\quad -\epsilon\left(2m_\theta \int_0^t \kappa_u e^{-\int_\theta^u \kappa_s ds}\chi_{[\theta,T]}(u)z_u^* du + 2\int_0^t \chi_{(\theta,T]}(u)z_{u-}^* dm_u\right). \end{aligned}$$

Clearly, all the integrals in the above expression are well defined, hence $z^\epsilon \in \mathcal{Z}_{adm}$. Since z^* is a maximizer,

$$\mathcal{U}(z^*) \geq \mathcal{U}(z^\epsilon), \qquad \forall\ \epsilon > 0.$$

By differentiating the above inequality, we arrive at

$$\mathbb{E}^{\mathbb{P}}\left(U'(Y^*_T)\ \mathcal{Y}_T\right) = 0, \tag{5.5}$$

where $Y^* := Y^{z^*}$ and

$$\mathcal{Y}_t := \frac{d}{d\epsilon}\bigg|_{\epsilon=0} Y^{z^\epsilon}_t .$$

Hence,

$$\begin{aligned}
\mathcal{Y}_T &= \int_0^T \chi_{(\theta,T]}(u) dS^e_u - \int_0^T \kappa_u \ell^*_u \chi_{[\theta,T]}(u) du \\
&\quad -2m_\theta \int_0^T \kappa_u e^{-\int_\theta^u \kappa_s ds} \chi_{[\theta,T]}(u) z^*_u du - 2\int_0^T \chi_{(\theta,T]}(u) z^*_{u-} dm_u \\
&= S^e_T + f^*_T - (S^e_\theta + f^*_\theta) - \frac{2m_\theta}{\Lambda(\theta)} \int_\theta^T \Lambda(u) \kappa_u z^*_u du
\end{aligned}$$

since

$$-\int_0^T \kappa_u \ell^*_u \chi_{[\theta,T]}(u) du = f^*_T - f^*_\theta + 2\int_0^T z^*_{u-} \chi_{[\theta,T]}(u) dm_u .$$

Then, (5.5) implies that

$$\mathbb{E}^{\mathbb{Q}}\left(\mathcal{Y}_T\right) = 0.$$

Recall that

$$S^*_t := S^e_t + \ell^*_t - 2m_t z^*_t = S^e_t + f^*_t .$$

Then, the above can be rewritten as

$$\mathbb{E}^{\mathbb{Q}}\left(S^*_\theta + L^*(\theta,T)\right) = \mathbb{E}^{\mathbb{Q}}\left(S^*_T\right), \quad \forall\ \theta. \tag{5.6}$$

where

$$L^*(\theta,T) = 2m_\theta \int_\theta^T \kappa_u \frac{\Lambda(u)}{\Lambda(\theta)} z^*_u du.$$

Equation (5.6) can be re-written as

$$\mathbb{E}^{\mathbb{Q}}\left(\widetilde{S}^*_\theta\right) = \mathbb{E}^{\mathbb{Q}}\left(\widetilde{S}^*_T\right), \quad \forall\ \theta \tag{5.7}$$

in which $\widetilde{S}^*_t = S^*_t + \mathbb{E}^{\mathbb{Q}}\left(L^*(t,T)|\mathcal{F}_t\right)$. Since this equality holds for all stopping times θ, we deduce that $\widetilde{S}^*$ is a martingale. Hence, the current post-liquidation stock price is the "risk neutral" expected value of the final post-liquidation stock price minus a weighted average of the future optimal portfolio positions. ■

6 Convergence of the model

Although there are many liquidity models in the literature, there are essentially two approaches that have been proposed to model liquidity in arbitrage-free settings. As discussed in the introduction, we refer to the approaches as the liquidity costs of the first kind, given by $\int_0^T m_t d[z]_t$ as in the paper of Çetin et al. [12] (in the case of a linear supply curve), and the second kind, given by liquidity costs of the form $\int_0^T h(\dot{z}_t)dt$.

In this section, we demonstrate the connection between these two approaches by showing that both models can be obtained as a limit of our current setup. More precisely, we are interested in the behavior of the model as the speed of decay goes to infinity. In order to do this, we define Λ^K for $K > 0$ as the process Λ defined in (4) with κ replaced by $K\kappa$. Furthermore, we denote the associated processes by x^K, ℓ^K, Y^K, etc.

6.1 Liquidity costs of the first kind

In Çetin et al. [12], it is postulated that the value of the money market account x_t^{CJP} at time t is given (in our notation) by

$$x_t^{CJP} = Y_{0-} - z_t S_t^e + \int_0^t z_{u-} dS_u^e - \int_0^t m_u d[z]_u$$

in the case of a linear supply curve, when z is a semimartingale.

When the speed of exponential decay is large (i.e., when K is large), the temporary impact decays quickly so that we expect that the liquidation value is close to that of the Çetin-Jarrow-Protter liquidity risk model. Indeed, we have the following result:

Proposition 6.1 *Let z be a semimartingale. Suppose ℓ^0 is uniformly bounded and κ strictly positive. Then, $x^K \to x^{CJP}$ uniformly on compact time intervals in probability (ucp, for short) as $K \to \infty$, i.e. for all $\epsilon > 0$ and all $t \geq 0$*

$$\lim_{K\to\infty} \mathcal{P}\left(\sup_{0\leq s\leq t} |x_s^K - x_s^{CJP}| > \epsilon\right) = 0.$$

Proof. The expression for Y^K in (4.8) can be re-written, using (4.2), as follows:

$$\begin{aligned} Y_t^K &= Y_{0-} + \int_0^t z_{u-} dS_u^i + \int_0^t z_{u-} d\ell_u^K - \int_0^t \left(2m_{u-}z_{u-}dz_u + 2z_{u-}d[m,z]_u + z_{u-}^2 dm_u\right) \\ &= Y_{0-} + \int_0^t z_{u-} dS_u^i + \int_0^t z_{u-} d\ell_u^K - \int_0^t \left(2m_{u-}z_{u-}dz_u + d[m,z^2]_u + z_{u-}^2 dm_u\right) \\ &= Y_{0-} + \int_0^t z_{u-} dS_u^i + \int_0^t z_{u-} d\ell_u^K - \int_0^t \left(2m_{u-}z_{u-}dz_u - m_{u-}dz_u^2\right) - z_t^2 m_t \\ &= Y_{0-} + \int_0^t z_{u-} dS_u^i - \int_0^t \ell_{u-}^K dz_u + \ell_t^K z_t - [\ell^K, z]_t + \int_0^t m_{u-} d[z]_u - m_t z_t^2 \end{aligned}$$

by integration by parts. As a result, we find that

$$Y_t^K \quad = \quad Y_{0-} + \int_0^t z_{u-} dS_u^i - \int_0^t \ell_{u-}^K dz_u - \int_0^t m_{u-} d[z]_u + \ell_t^K z_t - m_t z_t^2$$

since $[\ell^K, z]_t = \int_0^t 2m_{u-} d[z]_u$, using (4.2).

From (4.7), recall that the process ℓ^K can be written in the form

$$\ell_t^K \quad = \quad \ell_t^0 - \int_0^t K\kappa_s \ell_{s-}^0 \frac{\Lambda_t^K}{\Lambda_s^K} ds.$$

This integral is defined a.s. Furthermore, since ℓ^0 is bounded, so is the sequence $(\ell^K)_{K \geq 0}$. Note that $\Lambda_t^K = \left(\Lambda_t^1\right)^K$. Since ℓ_-^0 is bounded and left-continuous, we find that

$$\int_0^t K\kappa_s \ell_{s-}^0 \frac{\Lambda_t^K}{\Lambda_s^K} ds \to \ell_{t-}^0$$

for all t as $K \to \infty$ by applying Lemma 6.2 below to $g(s) = 1/\Lambda_s$. As a result, we find that $\ell_-^K \to 0$ a.s.

By the Dominated Convergence Theorem for stochastic integrals (see Theorem 32 of Chapter IV in [29]), we find that $\int \ell_-^K dz$ converges to zero in ucp.

Now, $x_t^K = Y_t^K - z_t(S_t - m_t z_t)$ by definition (see (4.3)). Hence,

$$x_t^K = x_t^{CJP} - \int_0^t \ell_{u-}^K dz_u + \ell_t^K z_t - m_t z_t^2 + z_t S_t^e - z_t S_t + m_t z_t^2 = x_t^{CJP} - \int_0^t \ell_{u-}^K dz_u,$$

since $S^e = S - \ell^K$. ■

In the preceding proof, we used the following lemma.

Lemma 6.2 *Let $g : [0, \infty] \to \mathbb{R}$ be a positive, strictly increasing continuously differentiable function. Let $t > 0$. If $f : [0, t) \to \mathbb{R}$ is bounded and has a left limit at t, then*

$$\lim_{K \to \infty} \int_0^t \tilde{g}_K(s) f(s) ds = \lim_{s \uparrow t} f(s),$$

in which $\tilde{g}_K(s) = K \left(\frac{g(s)}{g(t)}\right)^K \frac{g'(s)}{g(s)}$.

Proof. Let $\epsilon > 0$. Take $\delta > 0$ such that $|f(s) - f(t-)| < \epsilon$ for $t - \delta < s < t$. Since g is strictly increasing, we can take K large enough so that $\tilde{g}_K(s) < \frac{\epsilon}{t\|f\|_\infty}$ for all $s < t - \delta$. Then, $|\int_0^t \tilde{g}_K(s) f(s) ds - \int_{t-\delta}^t \tilde{g}_K(s) f(s) ds| < \epsilon$. Furthermore,

$$\begin{aligned} |\textstyle\int_{t-\delta}^t \quad & \tilde{g}_K(s) f(s) ds - f(t-) \int_{t-\delta}^t \tilde{g}_K(s) ds| \\ & \leq \epsilon \int_{t-\delta}^t \tilde{g}_K(s) ds. \end{aligned}$$

To finish the proof, it suffices to notice that

$$\int_{t-\delta}^{t} \tilde{g}_K(s)ds = 1 - \left(\frac{g\,(t-\delta)}{g(t)}\right)^K$$

converges to 1 when $K \to \infty$. ■

One of the properties of the liquidation value in the CJP model, is that, under mild conditions on the semimartingale z, one can always find a sequence of continuous processes z^n with finite variation such that $z^n_T = 0$ and

$$Y_{0-} + \int_0^T z^n_{u-} dS_u - \int_0^T m_{u-} d[z^n]_u - m_T (z^n_T)^2 \to Y_{0-} + \int_0^T z_{u-} dS_u$$

in L^2. In other words, the liquidation value obtained using the strategy z^n in the liquidity risk model of Çetin et al. [12] converges to the liquidation value obtained using z in an infinite liquidity model. However, since z^n is continuous and has finite variation the liquidity costs associated to these strategies are zero. Although, this is in line with the common practice of dividing big trades into smaller ones in practice, the model has the undesirable property that liquidity costs can be completely avoided in the limit by doing this. In our model, liquidity costs of an approximating trading strategy is approximately the same as the liquidity cost of the approximated strategy as stated in the following proposition, thus showing that the total liquidity cost of a strategy is a continuous function on the space of strategies as defined below.

Proposition 6.3 *Let $(z^n)_{n\geq 1}$ be a sequence of continuous processes converging in probability to z. Let $m = \tilde{m} + a$, with $\tilde{m}$ a local martingale and a an adapted process of finite variation. Suppose m is bounded from above and below, a is of integrable bounded variation and $E[\tilde{m}] < \infty$. Let ℓ^n and f^n be the processes associated to z^n. Then, the liquidity costs associated to the processes z^n,*

$$L^n = \frac{1}{4m}(f^n)^2 + \int \frac{1}{2m_u} \kappa_u (\ell^n_u)^2 du,$$

converges in probability to the liquidity costs associated to z, i.e. $L = \frac{1}{4m} f^2 + \int \frac{1}{2m_u} \kappa_u (\ell_u)^2 du$, as $n \to \infty$.

Proof. Let $\ell^{0,n}_t = 2m_t z^n_t - 2\int_0^t z^n_{s-} dm_s$. Then $\ell^n_t = \ell^{0,n}_t - \int_0^t \kappa_s \ell^{0,n}_s \frac{\Lambda_t}{\Lambda_s} ds$. By Theorem 4.1 of [17], we know that $\ell^{0,n}$ converges in probability to ℓ^0. Since $E\int_0^T \left|\kappa_s \frac{\Lambda_T}{\Lambda_s}\right| ds = E(1-\Lambda_T) < \infty$, we also find that ℓ^n converges in probability to ℓ. In particular, f^n converges to f. Also, $(\ell^n)^2$ converges to ℓ^2 by the continuous mapping theorem. Another application of Theorem 4.1 of [17] yields the result. ■

6.2 Liquidity costs of the second kind

In models of the second kind, liquidity costs are of the form $\int_0^T h(\dot{z}_s)ds$. Almgren [3] justifies this model with a discrete time construction. He assumes there is a temporary part to price impacts for the average price paid for a transaction which is proportional to the size of the transaction and inversely proportional to the length of time between trades, i.e. it is proportional to $\frac{z_{t+\delta}-z_t}{\delta}$ in which δ is the time between two trades. In the formal limit $\delta \to 0$, this quantity is interpreted as the speed of trading $\dot{z}_t$. In his model, the temporary impact does not affect the price paid at the next transaction. In this sense prices are highly resilient with respect to the temporary price impact. However, due to the assumption that the temporary impact is inversely proportional to the time between trades, thus modeling the fact that when one trades faster he pays more liquidity costs, the depth of the order book is also implicitly assumed to be going to zero as the time between trades goes to zero. The analogy with our current model is obtained be taking the parameters m and κ arbitrarily large. Note that in this case the liquidity cost of a block trade, or more generally the liquidity cost of trading with a non-zero quadratic variation, is arbitrarily large. We are thus forced to use absolutely continuous strategies. We have the following result.

Proposition 6.4 *For $K \geq 1$, let $m_t^K = Km_t$ and $\kappa_t^K = K\kappa_t$. Suppose κ is càdlàg and is bounded away from zero. Suppose z is absolutely continuous and denote by $\dot{z}$ its Radon-Nykodym derivative. Assume further that $\dot{z}$ is càdlàg, and $z_T = 0$. For $K \geq 1$, denote by L^K the associated liquidity cost process. Then,*

$$\lim_{K\to\infty} L_T^K = \int_0^T 2\frac{m_s}{\kappa_s}\dot{z}_s^2 ds$$

Proof. Since $z_T = 0$,

$$L_T^K = \frac{1}{4Km_T}\left(\ell_T^K\right)^2 + \int_0^T \frac{1}{2Km_s}K\kappa_s(\ell_s^K)^2 ds$$

in which $\ell_t^K = 2\int_0^t K\kappa_s\left(\frac{\Lambda_t}{\Lambda_s}\right)^K \frac{m_s}{\kappa_s}\dot{z}_s ds$. This implies that $\lim_{K\to\infty}\ell_t^K = 2\frac{m_{t-}}{\kappa_{t-}}\dot{z}_{t-}$ a.s. by Lemma 6.2.

Since κ is bounded away from zero, and m, κ and $\dot{z}$ are càdlàg, $m_t\dot{z}_t/\kappa_t$ is bounded on $[0,T]$ a.s. Also, since $0 \leq \int_0^t K\kappa_s\left(\frac{\Lambda_t}{\Lambda_s}\right)^K ds \leq 1$, we have that ℓ_t^K is also bounded a.s., uniformly in K. We obtain the result with an application of the bounded convergence theorem. ■

Remark 6.5 In the preceding proposition, it is necessary to assume that $z_T = 0$, because otherwise there is a final block trade to liquidate the position at time T, which would be infinitely costly. Mathematically, this appears from the term $m_T^K z_T^2$ in the liquidation value Y_T. In particular, since K is arbitrarily large, we find that this last liquidity cost is arbitrarily large. By assuming $z_T = 0$, we are implicitly assuming that the position is liquidated before time T in an absolutely continuous fashion. Note that in all models of the second kind we have mentioned in the introduction, it is either assumed that $z_T = 0$ or that $m_T = 0$ to exclude the liquidity cost of the liquidation of the portfolio.

7 Concluding Remarks

In this paper, we have developed a general model for illiquidity, outlined in Subsection 4.3, for an agent who sends market orders to a limit order. This model has the three important quantities identified earlier in the literature. We also argued in Remark 4.1 that the proposed dynamics are consistent with the seminal model of Kyle [26]. Moreover, the model allows to separate the liquidity costs associated to the bid-ask spread and to continuous trading. This separation allows us to derive a reduced model which ignores the transactions costs by properly identifying a mid-quote price process S and a wealth process Y. In Theorem 4.2 we also showed that the simplified model is arbitrage free.

We then used the reduced model in a classical utility maximization setting. This consideration shows that the post-liquidation asset price is the risk neutral expectation of the future post-liquidation asset price minus a weighted integral of the future optimal portfolio holdings. It is consistent with empirical findings of Brennan and Subrahmanyam [11] and Amihud and Mendelson [6]. In the final section, we obtained several other models as limiting cases of this reduced model.

References

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

[2] A. Alfonsi, A. Fruth, and A. Schied. Constrained portfolio liquidation in a limit order book model. *Banach Center Publ.*, 83:9–25, 2008.

[3] R. Almgren. Optimal execution with nonlinear impact functions and trading-enhanced risk. *Applied Mathematical Finance*, 10:1–18, 2003.

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

[5] Y. Amihud. Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets*, 5(1):31–56, 2002.

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

[7] K. Back. Insider trading in continuous time. *Review of Financial Studies*, 5/3:387–409, 1992.

[8] K. Back and S. Baruch. Information in securities markets: Kyle meets glosten and milgrom. *Econometrica*, 72/2:433–465, 2004.

[9] D. Bertsimas and A. Lo. Optimal control of execution costs. *Journal of Financial Markets*, 1:1–50, 1998.

[10] M. Blais and P. Protter. An analysis of the supply curve for liquidity risk through book data. *International Journal of Theoretical and Applied Finance*, 13:821–838, 2010.

[11] A. Brennan, M.J. and Subrahmanyam. Market microstructure and asset pricing. *J. Financial Economics*, 41:441–464, 1996.

[12] U. Çetin, R. Jarrow, and P. Protter. Liquidity risk and arbitrage pricing theory. *Finance and Stochastics*, 8:311–341, 2004.

[13] U. Çetin, H. M. Soner, and N. Touzi. Option hedging for small investors under liquidity costs. *Finance and Stochastics*, 4:317–341, 2010.

[14] G.M. Constantinides. Capital market equilibrium with transaction costs. *J. Political Economy*, 94:842–862, 1986.

[15] R. Cont, S. Stoikov, and R. Talreja. A stochastic model for order book dynamics. *Quantitative Finance*, 58:549–563, 2010.

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

[17] D. Duffie and P. Protter. From discrete- to continuous-time finance: Weak convergence of the financial gain process. *Mathematical Finance*, 2(1):1–15, 1992.

[18] B. Dumas and E. Luciano. An exact solution to a dynamic portfolio choice problem under transaction costs. *Journal of Finance*, 46:577–595, 1991.

[19] D. Easley and M. O'Hara. Price, trade size and information in securities markets. *Journal of Financial Economics*, 19:69–90, 1987.

[20] J. Gatheral. No-dynamic-arbitrage and market impact. *Quantitative Finance*, 10:749–759, 2010.

[21] J. Gatheral, A. Schied, and A. Slynko. Exponential resilience and decay of market impact. preprint, 2010.

[22] L. R. Glosten and P. R. Milgrom. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14:71–100, 1985.

[23] S. Gökay and H.M. Soner. Liquidity in a binomial market. *To appear in Mathematical Finance*, 2010.

[24] G. Huberman and W. Stanzl. Optimal liquidity trading. *Review of Finance*, 9:165–200, 2005.

[25] I Karatzas and S.E.. Shreve. *Methods of Mathematical Finance*. Springer-Verlag, second edition edition, 1998.

[26] A. Kyle. Continuous auctions and insider trading. *Econometrica*, 53:1315–1335, 1985.

[27] M.J.P. Magill and G.M. Constantinides. Portfolio selection with transaction costs. *J. Economic Theory*, 13:254–263, 1976.

[28] A. Obizhaeva and J. Wang. Optimal trading strategy and supply/demand dynamics. Preprint, 2005.

[29] P. Protter. *Stochastic Integration and Differential Equations*. Springer-Verlag, second edition edition, 2005.

[30] A. Roch. Liquidity risk, price impacts and the replication problem. to appear in Finance and Stochastics, 2011.

[31] L.C.G. Rogers and S. Singh. The cost of illiquidity and its effects on hedging. *Mathematical Finance*, 20:597–615, 2010.

[32] A. Schied and T. Schöneborn. Risk aversion and the dynamics of optimal liquidation strategies in illiquid markets. *Finance and Stochastics*, 13(2):181–204, 2009.

[33] S. E. Shreve and H. M. Soner. Optimal investment and consumption with transaction costs. *The Annals of Applied Probability*, 4(3):609–692, 1994.

[34] A. Subrahmanyam. Risk aversion, market liquidity, and price efficiency. *Review of Financial Studies*, 4:417–441, 1991.