1

Liquidity Models in Continuous and Discrete Time*

Selim Gökay, Alexandre F. Roch, and H. Mete Soner

[1] ETH Zürich selim.gokay@math.ethz.ch
[2] ETH Zürich alexandre.f.roch@gmail.com
[3] ETH Zürich and Swiss Finance Institute hmsoner@ethz.ch

Abstract. We survey several models of liquidity and liquidity related problems such as optimal execution of a large order, hedging and super-hedging options for a large trader, utility maximization in illiquid markets and price impact models with price manipulation strategies.

1.1 What Is Illiquidity?

The study of liquidity in financial markets either invokes the ease with which financial securities can be bought and sold, or addresses the ability to trade without triggering important changes in asset prices. More specifically, one can think of liquidity as an exogenous measure of the added costs per transaction associated to trading large quantities of the asset. This is the approach advocated by Çetin et al. [9], in which an exogenously defined supply curve gives the price per share as a function of transaction size. On the other hand, one can take this idea a step further and recognize that these added costs are the product of imbalances in the supply and demand of the asset due to the trading of large quantities. If the imbalance is temporary and only affects the current price paid, we are effectively in the previous setting and the transaction costs depend mainly on the size of the trade. On the other hand, these imbalances can have a lasting effect in such a way that future prices will be affected by previous trades. For instance, Jarrow [21, 22] considers the price per share as a function of the holdings of the large trader. As we can see, these two notions are closely related and one approach can be more convenient or realistic than the other depending on the setting.

* Research partly supported by the European Research Council under the grant 228053-FiRM. Financial support from Credit Suisse through the ETH Foundation and by the National Centre of Competence in Research "Financial Valuation and Risk Management" (NCCR FINRISK) are also gratefully acknowledged.

There are four main themes present in the current mathematical literature on liquidity. The first one pertains to the problem of optimal execution of large orders. Consider the situation in which a trader plans to sell a large number of units of a risky asset before a predetermined time horizon. Since the size of the order is large, this trader may find it more optimal to work the order in several smaller slices to minimize her impact on prices by trading during times of higher liquidity and taking advantage of the resilience of the supply and demand of the asset. On the other hand, delaying the orders for too long increases the exposure to other risks. The goal is to find the right balance between liquidity risk and other market risks. Many papers have been written on this question and we survey some of the main results in Section 1.2.

The second theme we discuss in this survey relates to the familiar problem of option pricing. On one hand, the existence of a supply curve that governs the liquidity cost of a transaction clearly suggests that the hedging of derivatives will be more costly that in the classical frictionless setting. On the other hand, the hedger's capacity to have an impact on prices may influence her into manipulating prices in her favor. The classical hedging problem gains a new level of complexity as the hedger's strategy, which is chosen in terms of the option payoff, has a repercussion on the future evolution of prices on which the option payoff is calculated. The different approaches commonly used in this setting are reviewed in Section 1.3 and 1.4. In Section 1.3 we review the results on hedging for a large trader, including the papers of Cvitanic and Ma [14], Platen, Schweizer [27], Bank and Baum [7] and Roch [28]. In Section 1.4 we introduce the supply curve model introduced in [9], discuss the super-replication problem in this context and focus on the works of Çetin, Soner and Touzi [11] and Gökay and Soner [18].

The third theme is related to the expected utility maximization problem with permanent or temporary price impacts. We briefly summarize some of the main results in this line of research in Section 1.5.

The introduction of price impacts on the evolution of the price processes evokes the possibility of *price manipulations*, defined as trading strategies with negative expected execution costs. For instance, by making the price go up after a purchase, a large trader has the possibility of making higher profits than average by re-selling the shares purchased if the average impact on prices is smaller for sell orders than buy orders. This is only one example of a price manipulation and it has lead some authors to investigate these types of irregularities in terms of the price impact functions. It is the focus of Section 1.6.

1.2 Optimal Execution Problem

The optimal execution problem consists in allocating a large buy or sell order of a risky asset over a fixed time horizon with the aim of minimizing the expected cost of the order due to the relative illiquidity of the asset. The main challenge in this kind of allocation is to choose a trading program which is executed on a period of time short enough to reduce the risk of the uncertainty of future prices while dividing the large order in smaller ones distributed over time to reduce the liquidity costs associated to this trading program.

There are mainly two approaches in the literature which we summarize in this section. The first approach, proposed in the papers of Bertsimas and Lo [8], Almgren [4], Almgren and Chriss [5,6], and Schied and Schöneborn [30], measures the associated cost of a sequence of transactions in terms of a permanent price impact and/or a temporary price impact which are exogenously determined and depend on the size of the transaction and the speed of change of the position in the asset. On the other hand, the second approach presumes the existence of a limit order book through which the orders of the large trader are executed. In this setting, the cost of an execution strategy depends on endogenous variables such as the density of the number of shares being offered at each price and the resilience of the order book. The main references that we will summarize for this approach are the papers of Obizhaeva and Wang [26], Alfonsi, Fruth and Schied [1, 2], and Alfonsi and Schied [3].

1.2.1 The First Approach

In the optimal execution problem, the investor wants to liquidate a certain number $X_0 > 0$ of units of an asset before a fixed finite time horizon T. Dividing the trading period $[0, T]$ into N equal intervals of length $\tau = T/N$, the investor chooses quantities $\xi_k \geq 0$ to sell at discrete times $t_k = k\tau$ for $k = 1, \ldots, N$ such that $\sum_{k=1}^{N} \xi_k = X_0$. The number of units still held by the investor at time t_k is given by $X_k = X_0 - \sum_{j=1}^{k} \xi_j$. Note that the case $X_0 < 0$ can be treated in a similar way.

Bertimas and Lo [8] approach this problem by minimizing expected execution costs, whereas Almgren [4], and Almgren and Chriss [5,6] extend this idea by also incorporating the risk into the execution problem using the variance of the associated costs.

Bertsimas and Lo [8] propose a general formulation for the price process of the asset, of which two special cases stand out. One special case proposed in [8] gives a stock price of the form

$$\tilde{S}_k = \tilde{S}_{k-1} - \gamma\xi_k + \epsilon_k \quad \gamma > 0 \tag{1.1}$$

in which $\{\epsilon_k\}_{k=1}^{N}$ is a sequence of independent and identically distributed random variables with mean zero and variance σ_ϵ^2, whereas ξ_k is the size of the transaction at time t_k. The profit obtained from a strategy, also commonly called the *capture*, is given by $\sum_{k=1}^{n} \xi_k \tilde{S}_k(\xi_k)$. The *total cost of trading* associated to a strategy X is defined as the difference between the book value $X_0 S_0$ and the capture, and is computed as

$$C(X) = X_0 S_0 - \sum_{k=1}^{N} \xi_k \tilde{S}_k.$$

In this setup, the goal is to minimize the expected execution cost

$$\min_{\{\xi_k\}_{k=1}^N} E\left[C(X)\right]$$

subject to the constraint

$$\sum_{k=1}^N \xi_k = X_0.$$

The price impact due to the trade ξ_k is said to be permanent in (1.1) since the price at time t_k is defined in terms of the price at time t_{k-1} which is also affected by the trade ξ_{k-1} at time t_{k-1}. For this special case there exists an explicit optimal strategy. It is called the naive strategy and is obtained by dividing the total order X_0 into N equal slices, i.e. $\xi_k = \frac{X_0}{N}$.

Bertsimas and Lo [8] also consider a linear temporary price impact model. In this setup the execution price $\tilde{S}_k$ at time k, i.e. the price paid for the transaction at time k, is decomposed into an exogenous unaffected price S_k and a price impact as a function of the trade size. The unaffected price, also called publicly-available price, can be interpreted as the price that would be obtained in absence of price impacts. The execution price at time t_k is a function of ξ_k and assumed to be given by

$$\tilde{S}_k(\xi_k) = S_k - (\eta\xi_k + \gamma Y_k)S_k, \quad \eta > 0$$

in which Y is an adapted process. In the special case that the unaffected price process $\{S_k\}_{k=1}^N$ follows

$$S_k = S_{k-1}\exp(\alpha_k),$$

and the state vector $\{Y_k\}_{k=1}^N$ satisfies

$$Y_k = \rho Y_{k-1} + \zeta_k,$$

in which $\{\zeta_k\}_{k=1}^N$ and $\{\alpha_k\}_{k=1}^N$ are i.i.d. normal random variables with mean 0, the authors show that the best execution strategy consists in trade sizes which are linear functions of the remaining number of shares X_k and the state variable Y_k.

The implicit assumption in the paper of Bertsimas and Lo [8] is that the investor is not risk averse as she only aims to minimize the expected cost of the execution. In the optimal execution model of Almgren [4], and Almgren and Chriss [5, 6], the investor's tolerance for risk influences her trading decisions. To illustrate this point, consider the two following execution strategies. On one hand, a risk averse agent may choose to trade everything now. The advantage of this strategy is that the cost is known and all risks regarding the future prices of the asset are eliminated. On the other hand, the cost is high and the investor may be willing to take some risk by dividing her orders and executing them through time in order to have a lower expected cost. Almgren and Chriss characterize this trade-off between the cost and the variance of optimal execution strategies by an efficient frontier. They show that the points on the frontier are determined by the level of risk aversion of the agent. They argue that the optimal strategies for the execution problem are static, i.e. these decisions can be fully determined at the beginning of the trading period, and give explicit solutions for some specific cases.

In addition to the above mathematical setup, we denote by $v_k = \frac{\xi_k}{\tau}$ the speed of trades on the k-th interval. In [5], the publicly-available price per share S_k is modeled as follows. Let $\{\zeta_k\}_{k=1}^N$ be i.i.d. random variables with zero mean and unit variance. We assume that

$$S_k = S_{k-1} + \sigma\sqrt{\tau}\zeta_k - \tau g(v_k), \quad k = 1, \ldots, N,$$

where $\sigma > 0$ is a volatility parameter and $g : \mathbf{R} \to \mathbf{R}$ is a permanent impact function. The price per share paid by the investor at time k is

$$\tilde{S}_k(\xi_k) = S_{k-1} - h\left(\frac{\xi_k}{\tau}\right), \quad k = 1, \ldots, N$$

in which h is a given function, called the temporary impact function. The capture is computed as

$$\begin{aligned} C(X) &= X_0 S_0 - \sum_{k=1}^{N} \xi_k \tilde{S}_k\left(\xi_k\right) \\ &= \sum_{k=1}^{N} \tau X_k g(v_k) + \sum_{k=1}^{N} \tau v_k h(v_k) - \sigma\sqrt{\tau} \sum_{k=1}^{N-1} X_k \zeta_k, \end{aligned} \tag{1.2}$$

with expected value and variance at time 0 given by

$$E(C(X)) = \sum_{k=1}^{N} \tau X_k g(v_k) + \sum_{k=1}^{N} \tau v_k h(v_k), \quad \mathrm{Var}(C(X)) = \sum_{k=1}^{N} \tau \sigma^2 X_k^2$$

when the strategy X is deterministic.

A strategy is called *efficient* if there is no strategy that has a lower expected value for a level of variance which is equal or lower. The family of efficient strategies is given by the solutions $X^*(\lambda)$ of the optimization problem

$$\min_X \{E(C(X)) + \lambda \mathrm{Var}(C(X))\}$$

for different values of $\lambda \geq 0$. The family of solutions $(X^*(\lambda))_{\lambda \geq 0}$ is called the efficient frontier. The parameter λ measures the risk aversion of the investor. Every point on the frontier corresponds to a pair

$$(\mathrm{Var}(C(X^*(\lambda))), E(C(X^*(\lambda))))$$

for some λ. The efficient frontier gives rise to a smooth and convex function, which we denote by $\mathcal{E}(V)$, assigning the optimal expected cost $\mathcal{E}(V)$ to each possible value of the variance V, i.e. there exists $\lambda \geq 0$ such that $(V, \mathcal{E}(V)) = (\mathrm{Var}(C(X^*(\lambda))), E(C(X^*(\lambda))))$.

In [6], the permanent impact function is taken to be linear, i.e. $g(v) = \gamma v$ (with $\gamma > 0$) and the temporary price impact function consists of the sum of a fixed cost function and a linear impact function so that

$$h(v) = \theta \ \mathrm{sign}(v) + \eta v \quad (v \in \mathbf{R}) \tag{1.3}$$

for some positive constants $\theta, \eta > 0$. In this case, it is easy to see that the expectation of the cost becomes

$$E(C(X)) = \frac{1}{2}\gamma X_0^2 + \epsilon \sum_{k=1}^{N} |\xi_k| + \frac{\eta - \frac{1}{2}\gamma\tau}{\tau} \sum_{k=1}^{N} \xi_k^2.$$

Almgren and Chriss [6] show that the optimal solution for the case of g linear and h given by (1.3) can be written in terms of $\lambda > 0$ as

$$X_j^* = \frac{\sinh(\kappa(T - t_j))}{\sinh(\kappa T)} X_0, \quad j = 0, \ldots, N,$$

in which

$$\kappa \sim \sqrt{\frac{\lambda \sigma^2}{\eta}} + O(\tau), \quad \tau \to 0.$$

If the agent is risk-neutral ($\lambda = 0$), she only wants to minimize the expected cost. Then her optimal strategy is the naive strategy $\xi_k = \frac{X_0}{N}$ as we have seen earlier. In this case, the expected cost and variance of this strategy are given by

$$\begin{aligned} E_0 &:= \tfrac{1}{2}\gamma X_0^2 + \epsilon X_0 + \left(\eta - \tfrac{1}{2}\gamma\tau\right) \tfrac{X_0^2}{T}, \\ V_0 &:= \tfrac{1}{3}\sigma^2 X_0^2 T \left(1 - \tfrac{1}{N}\right)\left(1 - \tfrac{1}{2N}\right). \end{aligned}$$

The naive strategy corresponds to the minimal point of the efficient frontier, in the sense that $\frac{d\mathcal{E}}{dV}$ evaluated at (V_0, E_0) is equal to zero. Thus for (V, E) in the vicinity of (V_0, E_0),

$$E - E_0 \approx \frac{1}{2}(V - V_0)^2 \left.\frac{d^2\mathcal{E}}{dV^2}\right|_{V=V_0},$$

in which $d^2\mathcal{E}/dV^2\big|_{V=V_0}$ is positive by the convexity of the efficient frontier.

1.2.2 Continuous-Time Models

Let us now consider non-linear impact functions and analyze the continuous-time limit of the previous model as $\tau \to 0$. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbf{P})$ be a given filtered probability space on which a Brownian motion W is defined. In the continuous setup, the publicly-available price will be assumed to be given by

$$S_t = \sigma W_t - \int_0^t g(\dot{X}(t))dt. \tag{1.4}$$

Here, $\dot{X}_t$ is the derivative of X_t with respect to t, it corresponds to v_k in the previous discrete setup. The proceeds associated to a trading strategy X and an initial position of X_0 in the risky asset and y in the riskless asset are given by

$$\begin{aligned}\mathcal{R}_T(X) = X_0 S_0 + y - \int_0^T X_t g(\dot{X}_t)dt \\ - \int_0^T \dot{X}_t h(\dot{X}_t)dt + \sigma \int_0^T X_t dW_t.\end{aligned} \tag{1.5}$$

The cost of the strategy X is defined as $C(X) := X_0 S_0 + y - \mathcal{R}_T(X)$. This can be formally obtained as a limit of (1.2). The expectation and the variance of the cost are given by

$$E(C(X)) = \int_0^T X(t)g(\dot{X}_t) + \dot{X}_t h(\dot{X}_t)dt, \quad \mathrm{Var}(C(X)) = \int_0^T \sigma^2 X_t^2 dt$$

when X is deterministic. The problem then consists in finding a deterministic absolutely-continuous strategy $(X_t)_{t\in[0,T]}$ that minimizes $E(C(X)) + \lambda \mathrm{Var}(C(X))$ for a given risk-aversion level λ.

To obtain explicit solutions to the above minimization problem, Almgren [4] considers a linear permanent impact $g(v) = \gamma v$ and a temporary impact in the form of a power law $h(v) = \eta v^k$ with $k > 0$. For each trading horizon T, there is an optimal strategy. Almgren finds that the optimal strategy which takes the longest to execute can be expressed as

$$\frac{X_t}{X_0} = \begin{cases} \left(1 - \frac{k-1}{k+1}\frac{t}{T_*}\right)^{\frac{k+1}{k-1}} & \text{if } k \neq 1, \\ \exp\left(-\frac{t}{T_*}\right) & \text{if } k = 1, \end{cases}$$

in which T_*, called the *characteristic time*, is given by

$$T_* = \left(\frac{k\eta X_0^{k-1}}{\lambda\sigma^2}\right)^{1/(k+1)}.$$

For the linear case, $k = 1$, the characteristic time is independent of the initial portfolio size X_0 and corresponds to the amount of time needed for the portfolio position to decrease by a factor of e^{-1}. If $k < 1$, volatility risk dominates the expected cost as the portfolio size increases and the speed of trading decreases with time. When $k > 1$, the trading cost dominates volatility risk.

When $k \leq 1$, the execution time is infinite, i.e. $X_t > 0$ for all $t < \infty$. On the other hand, when $k > 1$, the trading stops after a finite time given by

$$T = \frac{k+1}{k-1} T_*.$$

Next consider the same wealth equation as (1.5) with $T = \infty$, $h(x) = \lambda x$ and $g(x) = \gamma x$. This is the setup considered by Schied and Schöneborn [30]. The admissible portfolios $(X_t)_{t \geq 0}$ considered are more general than in the previous setups as they are assumed to satisfy the following conditions:

- X_t is absolutely continuous and $\xi(t) := -\dot{X}(t)$,
- $X_T = 0$,
- ξ is progressively measurable with respect to the filtration $(\mathcal{F}_t)_{t \geq 0}$ with $\int_0^t \xi_s^2 ds < \infty$ for all $t > 0$,
- $X_t(\omega)$ is uniformly bounded in t and ω.

The class of admissible strategies starting with X_0 units of the risky asset and y shares in the riskless asset is denoted by $\mathcal{X}(X_0, r)$ in which $r = X_0 S_0 + y - \frac{\gamma}{2} X_0^2$. The goal is to maximize the expected utility of the capture $\mathcal{R}_t(X)$ over the class of admissible strategies. Assume the utility function u is smooth with risk aversion factor

$$A(r) = -\frac{u_{rr}(r)}{u_r(r)},$$

satisfying

$$0 < A_{min} := \inf_{r \in \mathbf{R}} A(r) \leq \sup_{r \in \mathbf{R}} A(r) := A_{max} < \infty.$$

We consider two different maximization problems. The first problem is given by the following value function:

$$v_1(x, r) = \sup_{X \in \mathcal{X}(x,r)} E\left[u(R_\infty(X))\right],$$

where

$$R_\infty(X) = r + \sigma \int_0^\infty X_s dB_s - \lambda \int_0^\infty \dot{X}_s^2 ds.$$

In the above equation we avoid the technical limiting argument and the associated admissibility class. The second problem involves the value function

$$v_2(x, r) = \sup_{X \in \mathcal{X}(x,r)} \lim_{t \to \infty} E\left[u(\mathcal{R}_t(X))\right],$$

where

$$R_t(X) = r + \sigma \int_0^t X_s dB_s - \lambda \int_0^t \dot{X}_s^2 ds.$$

It can be shown that v_1 and v_2 are equal and solve the Hamilton-Jacobi-Bellman equation

$$-\frac{1}{2}\sigma^2 x^2 v_{rr} + \inf_c \left\{ \lambda v_r c^2 + v_x c \right\} = 0, \quad \text{for } x > 0, r \in \mathbf{R} \tag{1.6}$$

together with the boundary condition

$$v(0,r) = u(r), \quad r \in \mathbf{R}.$$

The unique optimal control $\dot{X}_t^*$ is Markovian and is given in feedback form by

$$\dot{X}_t^* = c\left(X_t^*, \mathcal{R}_t(X^*)\right) \tag{1.7}$$

in which $c(x,r) = -\frac{v_x(x,r)}{2\lambda v_r(x,r)}$.

To prove the above statements, the authors show that there exists a sufficiently smooth solution $\tilde{c} : (y,r) \in \mathbf{R}_0^+ \times \mathbf{R} \to \tilde{c}(y,r) \in \mathbf{R}$ of the partial differential equation

$$\tilde{c}_y = -\frac{3}{2}\lambda \tilde{c}\tilde{c}_r + \frac{\sigma^2}{4\tilde{c}}\tilde{c}_{rr}$$

with initial value

$$\tilde{c}(0,r) = \sqrt{\frac{\sigma^2 A(r)}{2\lambda}}.$$

Moreover, the solution satisfies

$$\sqrt{\frac{\sigma^2 A_{min}}{2\lambda}} \leq \tilde{c}(y,r) \leq \sqrt{\frac{\sigma^2 A_{max}}{2\lambda}}. \tag{1.8}$$

Also, there exists a sufficiently smooth solution $\tilde{w} : \mathbf{R}_0^+ \times \mathbf{R} \to \mathbf{R}$ of the transport equation

$$\tilde{w}_y = -\lambda \tilde{c}\tilde{w}_r$$

with initial value

$$\tilde{w}(0,r) = u(r).$$

Then the function $w(x,r) := \tilde{w}(x^2, r)$ solves the HJB equation (1.6) and the unique minimum is attained at

$$c(x,r) := \tilde{c}(x^2, r)x.$$

A verification argument concludes that the solution of the HJB equation (1.6) must be equal to the value functions v_1 and v_2 and the unique optimal control satisfies (1.7) in which $c(x,r) = -\frac{v_x(x,r)}{2\lambda v_r(x,r)}$. Then in view of (1.7) the asset position $X_t^{\hat{\xi}}$ at time t under the optimal control $\hat{\xi}_t$ is given as

$$X_t^{\hat{\xi}} = X_0 \exp\left(-\int_0^t \tilde{c}\left((X_s^{\hat{\xi}})^2, R_s^{\hat{\xi}}\right) ds\right)$$

and because of (1.8) it is bounded as follows:

$$X_0 \exp\left(-t\sqrt{\frac{\sigma^2 A_{max}}{2\lambda}}\right) \le X_t^{\hat{\xi}} \le X_0 \exp\left(-t\sqrt{\frac{\sigma^2 A_{min}}{2\lambda}}\right).$$

In the case with constant absolute risk aversion $A = A_{min} = A_{max}$, the optimal adaptive liquidation strategy is static and is given by

$$X_t^{\tilde{\xi}} = X_0 \exp\left(-t\sqrt{\frac{\sigma^2 A}{2\lambda}}\right).$$

Since the absolute risk aversion of the utility function determines the initial condition of the partial differential equation for $\tilde{c}$, it is a key factor for the optimal trading strategy. In particular, the optimal strategy inherits monotonicity properties of the absolute risk aversion. Let u^0 and u^1 be two utility functions with corresponding absolute risk aversion $A^0(r)$ and $A^1(r)$. If $A^1(r) \ge A^0(r)$ for all r, then an investor with utility function u^1 liquidates the same portfolio X_0 faster than an investor with utility function u^0. More precisely, we get

$$c^1 \ge c^0 \quad \text{and} \quad \hat{\xi}_t^1 \ge \hat{\xi}_t^0 \quad \mathbf{P}-\text{a.s.},$$

where c^i and $\hat{\xi}_t^i$ are obtained from the utility function u^i with $i \in \{0,1\}$. As a corollary, it follows that $c(x,r)$ is increasing (decreasing) in r for all values of x if and only if the absolute risk aversion parameter $A(r)$ is increasing (decreasing) in r. Therefore, an investor with increasing absolute risk aversion $A(r)$ would sell faster when prices rise, since an increase in prices lead to an increase in r. In this case, the investor is called aggressive in-the-money. On the other hand, an investor having a decreasing absolute risk aversion $A(r)$ is passive in-the-money, i.e. she would sell slower when prices increase.

1.2.3 Models of Limit Order Books

We now analyze the limit order book (LOB) models and focus on the papers by Obizhaeva and Wang [26], and Alfonsi, Fruth and Schied [1,2]. As before, we take the point of view of a large trader who needs to liquidate a certain fixed number of units of a risky asset. In limit order books, as opposed to modeling the price process directly, one models the dynamics of supply and demand for the asset in the market and its impact on the execution cost. Then the supply and demand levels determine the magnitude of price impacts.

A limit order is an order to sell or buy a certain number of shares of an asset at a specified price. The limit order book consists of the collection of all sell and buy limit orders. A market order is an order to buy or sell a certain number of shares at the most favorable price available in the limit order book. The lowest specified price in the LOB for a sell order is called the best ask price, whereas the highest price of a buy order in the LOB is the best bid price. A market order to buy (resp. sell) is executed against the limit orders to sell (resp. buy). In LOB models, the dynamics of the LOB is assumed to only be affected by noise traders when the large trader is inactive, and their actions determine the unaffected best ask price A_t^0 and the unaffected bid price B_t^0. The processes $A^0 = (A_t^0)_{t \geq 0}$ and $B^0 = (B_t^0)_{t \geq 0}$ are adapted, exogenously defined stochastic processes on the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbf{P})$. Clearly, a natural condition to impose on these two processes is $A_t^0 \geq B_t^0$ for all $t \geq 0$. We denote the density of the LOB at the price $A_t^0 + x$ (resp. $B_t^0 + x$) by $f(x)$ for $x > 0$ (resp. $x < 0$), i.e. the number of shares offered at the price $A_t^0 + x$ (resp. $B_t^0 + x$) is given by $f(x)dx$. It is assumed that $f : \mathbf{R} \to (0, \infty)$ is a bounded continuous function, called the shape function of the LOB. The large trader makes buy and sell orders, thereby temporarily depleting parts of the LOB. We denote by F the antiderivative of f, i.e.

$$F(y) = \int_0^y f(x)dx.$$

The actual best ask price at time t, denoted by A_t, takes into account the price impacts of the previous market orders of the large trader. The positive difference between the actual and the unaffected best ask prices $D_t^A = A_t - A_t^0$ is called the extra spread. A buy order of size $\xi > 0$ at time t consumes all shares in the LOB from the actual best ask price A_t to

$$A_{t^+} = A_t + D_t^A(\xi) - D_t^A,$$

where $D_t^A(\xi)$ is determined by the relation

$$\int_{D_t^A}^{D_t^A(\xi)} f(x)dx = \xi.$$

The process D^A specifies the market impact of orders on the best ask price. For a general shape function f, the market impact $D_t^A(\xi) - D_t^A$ is non-linear. However, if we assume a block shaped LOB, i.e. an LOB in which the shape function is equal to a constant q above the actual best ask price, then the market impact $D_t^A(\xi) - D_t^A$ is linear and equal to ξ/q.

We now describe the admissible trading strategies for the large trader. Assume the trader wants to buy $x > 0$ shares in $N+1$ trades within the time interval $[0,T]$. The trading strategies considered by Alfonsi and Schied [3] are simple strategies of the form

$$X_t = \xi_0 + \sum_{n=1}^{N} \xi_n \mathbf{1}_{\{t \geq \tau_n\}} \quad (0 \leq t < T),$$

where $\tau_0, \ldots, \tau_N$ are stopping times satisfying $0 = \tau_0 \leq \tau_1 \leq \cdots \leq \tau_N$ and every ξ_n is bounded below and measurable with respect to $\mathcal{F}_{\tau_n}$. The quantity ξ_n represents the size of the market order placed at time τ_n. We denote this set of admissible strategies by $\mathcal{X}_N$. In [1, 2], the admissible strategies considered are special cases of the above setup, i.e. the trading times are not stopping times, but they are predetermined. For convenience, we denote by $X_t^+ = \xi_0 + \sum_{i=1}^{N} \xi_n \mathbf{1}_{\{t \geq \tau_n, \xi_n > 0\}}$ the cumulative buy orders, and $X_t^- = X_t - X_t^+$ the cumulative sell orders.

It is assumed that the market impact decays with time as the result of new sell orders coming in the order book. This phenomenon is known as the resilience of the LOB. In [2, 3] there are two different approaches to model resilience. Either the volume of the order book consumed by the large trader, denoted at time t by E_t^A, is assumed to recover exponentially or the extra spread D_t^A decays exponentially. The assumption regarding resilience is stated as follows: there is a deterministic rate process $(\rho_t)_{t \geq 0}$ such that either

$$dE_t^A = -\rho_t E_t^A dt + dX_t^+$$

or

$$dD_t^A = -\rho_t D_t^A dt + D_t^A(\Delta X_t^+)$$

In the specific case of a block-shaped LOB, it can be shown that

$$D_t^A = \frac{1}{q} \sum_n e^{-\int_{\tau_n}^t \rho_s ds} \xi_n \mathbf{1}_{\{\tau_n \leq t, \xi_n > 0\}}. \tag{1.9}$$

It is easy to see that these two approaches of resilience coincide for block-shaped LOBs. The dynamics of the bid side of the LOB are modeled identically. As before, the density of the number of shares offered at the price $B_t^0 + x$ for $x < 0$ are given by the shape function f. The extra spread D_t^B is the difference between the actual best bid price and the best unaffected bid price $D_t^B = B_t - B_t^0$, which is non-positive. A sell order of size $\xi < 0$ will move the actual best bid price to

$$B_{t^+} = B_t + D_t^B(\xi) - D_t^B,$$

where $D_t^B(\xi)$ is defined as before

$$\xi = \int_{D_t^B}^{D_t^B(\xi)} f(x)dx.$$

As before, the resilience is either modeled in terms of the volume consumed by the large trader or the extra spread as follows:

$$\begin{aligned} dE_t^B &= -\rho_t E_t^B dt + dX_t^-, \text{ or} \\ dD_t^B &= -\rho_t D_t^B dt + D_t^B(\Delta X_t^-) \end{aligned}$$

The difference $A_t - B_t$ between the best ask and the best bid price is called the bid-ask spread.

A buy order of size $\xi > 0$ at time t consumes the $f(x)dx$ available shares at price $A_t^0 + x$, where x ranges from D_t^A to $D_t^A(\xi)$. The cost associated to this transaction is given by

$$\pi_t(\xi) = \int_{D_t^A}^{D_t^A(\xi)} (A_t^0 + x)f(x)dx = A_t^0\xi + \int_{D_t^A}^{D_t^A(\xi)} xf(x)dx.$$

Similarly for a sell order $\xi < 0$ we have

$$\pi_t(\xi) = \int_{D_t^B}^{D_t^B(\xi)} (B_t^0 + x)f(x)dx = B_t^0\xi + \int_{D_t^B}^{D_t^B(\xi)} xf(x)dx.$$

The expected cost $\mathcal{C}(X)$ of an admissible strategy X can then be obtained by

$$\mathcal{C}(X) = E\left[\sum_{n=0}^{N} \pi_{\tau_n}(\xi_n)\right].$$

The goal is then to minimize $\mathcal{C}(X)$ among all admissible strategies X. Note that, in contrast with the works of Almgren [4] and Almgren and Chriss [5,6], intermediate sell orders (resp. buy orders) are allowed for execution orders of $x > 0$ (resp. $x < 0$) shares.

In [3], it is established that minimizing $\mathcal{C}(X)$ over the set of admissible strategies $\mathcal{X}_N$ is equivalent, under some assumptions on the density function f, to minimizing $\mathcal{C}(X)$ under the constraint that the trading times sequence $(\tau_0, \tau_1, \dots, \tau_n)$ is given by the time spacing $\mathcal{T}^* = (t_0^*, t_1^*, \dots, t_N^*)$ defined by

$$\int_{t_{i-1}^*}^{t_i^*} \rho_s ds = \frac{1}{N}\int_0^T \rho_s ds, \quad i = 1, \dots, N.$$

The unique optimal strategy for the first model of resilience is given by

$$\xi_1^* = \cdots = \xi_{N-1}^* = \xi_0^*(1 - a^*),$$

in which

$$a^* = \exp\left(-\frac{1}{N}\int_0^T \rho_s ds\right)$$

and ξ_0^* solves

$$F^{-1}(x - N\xi_0^*(1-a^*)) = \frac{F^{-1}(\xi_0^*) - a^* F^{-1}(a^*\xi_0^*)}{1-a^*}.$$

The last order ξ_N^* is determined so that

$$\xi_N^* = X_0 - \xi_0^* - (N-1)(1-a^*)\xi_0^*.$$

When f is constant,

$$\xi_0^* = \frac{x}{(N-1)(1+a^*)+2}.$$

In the asymptotic limit, i.e. as $N \to \infty$, of the block-shaped LOB, the optimal execution strategy is a combination of discrete and continuous trades when the resilience factor ρ is constant. The initial and final trades are discrete, whereas the intermediate ones are continuous. The optimal strategy is given by

$$\xi_0^* = \xi_T^* = \frac{X_0}{\rho T + 2}, \quad \frac{d}{dt}\xi_t^* = \frac{\rho X_0}{\rho T + 2}.$$

Note that in the LOB price impact model described above the impact of a trade is not permanent: the extra spread decays with time. Alfonsi et al. [1] and Obizhaeva and Wang [26] include an additional permanent impact factor in the block-shaped LOB model. More specifically, they let the density function $f = q \in \mathbf{R}$ and assume the extra spread D^A caused by a strategy X satisfies

$$D_t^A = \gamma \sum_n \xi_n \mathbf{1}_{\{\tau_n \le t, \xi_n > 0\}} + \kappa \sum_n \exp\left(-\int_{\tau_n}^t \rho_s ds\right) \xi_n \mathbf{1}_{\{\tau_n \le t, \xi_n > 0\}},$$

in which $0 \le \gamma \le 1/q$ is the permanent effect factor and $\kappa = 1/q - \gamma$ is the proportion of the market impact that decays with time. Similar dynamics holds for sell orders. Comparing this to (1.9), we see that a proportion $\frac{\gamma}{1/q}$ of the consumed volume by the large trader does not recover in the long run. It turns out that the minimization problem with permanent impact has the same optimal trading strategy as the minimization problem with $\gamma = 0$.

In [1], Alfonsi et al. consider this problem under convex constraints and obtain closed-form solutions. The set of strategies considered is smaller however than $\mathcal{X}_N$ as trading is only permitted on a pre-determined time grid $t_0, t_1, \ldots, t_n$. The aim is to reduce the constrained optimization problem to the minimization of a positive definite quadratic form on a convex subset of Euclidean space. As a special case, they obtain closed-form solutions for the unconstrained problem.

1.3 Option Hedging for Large Traders

In this section we survey the large trader models for hedging options. The trades of the large trader are assumed to have an impact on the prices so that she has to take this effect into account when considering hedging options. There are various approaches to incorporate the trading decisions of the large trader into the price process of the underlying. Jarrow [21, 22] considers the price process expressed in terms of reaction functions of the holdings of the large trader. This turns out to be a generalization of Huberman and Stanzl's model for price manipulation. In [14] and [13], the coefficients of the price process depend exogenously on the large trader's portfolio. Platen and Schweizer [27], Frey and Stremme [16], and Sircar and Papanicolau [31] use an equilibrium approach to derive the reaction function for the price process. Frey [15] assumes that this reaction function describing the price process as a function of the holdings of the large trader is exogenously given. Bank and Baum [7] model the price process of the risky asset in terms of a smooth family of semimartingales $(S^z)_{z\in\mathbf{R}}$, where S^z describes the evolution of the stock price process for constant z, which represents the size of the large trader's holdings. Roch [28] considers a setup similar to the limit order book models described above in which the parameter of the linear permanent impact function is given by a stochastic process.
Throughout the remaining sections, we work with a filtered probability space $(\Omega, \mathcal{F}, \mathbf{F}, \mathbf{P})$, which supports a standard Brownian motion $(W_t)_{0\leq t\leq T}$. We also fix a finite time horizon $T > 0$. Unless otherwise specified, there is one risky asset and one riskless asset in the market. We normally think of the risky asset as a stock and the riskless asset as a money market account. The money market account is taken to be a numéraire so that its price is normalized to unity. The discounted price of the stock process at time t is denoted by S_t. There are two types of traders in the economy, one large trader and reference traders. The large trader can be a speculator, a program trader or a portfolio insurer. The reference traders are typically noise traders or arbitrage-based speculators. Let X_t be the number of money market units, Y_t the book value of the stock position and Z_t be the number of stocks the large trader holds at time t. The processes X, Y and Z are assumed to be adapted to the filtration $\mathbf{F}$.

In classical settings based on the Black-Scholes model, the stock price process S_t is modeled as a solution of a linear stochastic differential equation (SDE). The drift and volatility coefficients of the SDE are not influenced by the agents portfolio and wealth processes. This is based on the assumption that agents are price takers in this framework. Cvitanic and Ma [14] model the price process of the underlying asset by a SDE taking into account that large trader's decisions have a price impact. In particular, they assume that the drift and volatility coefficients depend on the large trader's portfolio and wealth process. The authors consider a market with d risky assets (stocks) and one riskless asset (money market account). Let S^0_t be the price process of the money market account, S_t^i be the price process of the ith stock. Then the dynamics of these processes are given as

$$\begin{aligned} dS_t^0 &= S_t^0 r(t, Y_t, Z_t)dt,\ 0 \le t \le T, \quad S_0^0 = 1, \\ dS_t^i &= b_i(t, S_t, Y_t, Z_t)dt + \sum_{j=1}^{d} \sigma_{ij}(t, S_t, Y_t, Z_t)dW_t^j,\ 0 \le t \le T,\ S_t^i = s_i > 0, \\ dY_t &= \hat{b}(t, S_t, Y_t, Z_t)dt + \hat{\sigma}(t, S_t, Y_t, Z_t)dW_t,\ 0 \le t \le T,\ Y_0 = y > 0, \end{aligned}$$

where

$$\begin{aligned} \hat{b}(t, s, y, z) &= \left(y - \sum_{i=1}^{d} s_i z_i \right) r(t, y, z) + \sum_{i=1}^{d} z_i b_i(t, s, y, z), \\ \hat{\sigma}_j(t, s, y, z) &= \sum_{i=1}^{d} z_i \sigma_{ij}(t, s, y, z) \quad j = 1, \ldots, d. \end{aligned}$$

Under additional assumptions on the coefficients of the above SDE's, the authors show that the replication of European options with payoff in the form $g(S_T)$ has a solution. The method is based on forward-backward stochastic differential equations and the well-known 4-step scheme of Ma et al. [25].

Platen and Schweizer [27], Frey and Stremme [16], Frey [15] and Sircar and Papanicolau [31] do not model the price process explicitly as in [13] and [14]. However, they follow a microeconomic equilibrium approach to understand the feedback effects from hedging strategies. As before, there are two types of investors in the market, a large trader and reference traders. The aggregate demand of the reference trader at time t is given by $D(t, F_t, S_t)$, where $F = (F_t)_{0 \le t \le T}$ is the fundamental state process and S_t is the price for stock. The fundamental state process can represent various things, for instance noise or misspecifications in the model, demand for liquidity or aggregated income of the reference trader. Supposing that at time t the large trader possesses a fraction α_t of the total supply of the stock, then the market clearing condition states that

$$D(t, F_t, S_t) + \alpha_t = 1.$$

Under some assumptions it can be shown that there is a unique solution for S_t in terms of t, α_t and F_t, i.e. we can express $S_t = \psi(t, F_t, \alpha_t)$. The function ψ is called the *reaction* function.

Frey and Stremme [16] investigate the impact of dynamic hedging on the price process in a general discrete time economy with the equilibrium model. They pass to the diffusion limit and investigate the continuous-time equilibrium price process and its volatility. The price process is still represented by an Itô process, but the volatility increases and becomes time and price dependent.

Sircar and Papanicolau [31] analyze the increases in market volatility of asset prices. Many investors use Black-Scholes trading strategies to hedge derivatives. The use of these hedging strategies is so extensive that they have an impact on the price of the asset, which in turn influences the price of the derivative. In their framework, there is an interaction between reference traders and large traders who follow a dynamic Black-Scholes hedging strategy. Following an equilibrium analysis, they derive a stochastic process for the price of the asset that depends on the hedging strategy of the large trader. Then they derive a nonlinear partial differential equation for the derivative price and the hedging strategy. They observe that the increase in volatility can be attributed to the feedback effect of Black-Scholes hedging strategies.

Platen and Schweizer [27] aim to study the implied volatility structure in the above reaction setup. In other words, instead of taking an exogenously given price process, they develop a diffusion model for stock prices that incorporates the technical demand induced by the hedgers. The diffusion model is endogenously determined by the trading decisions in the economy. With their modeling, they can explain volatility smiles and skews as a result of feedback effects from hedging derivatives. They consider the following specification of the demand function:

$$D(t, F_t, S_t) = F_t + \gamma(\log(S_t) - \log(S_0))$$

where $F_t = vW_t + mt$ is a random error term and $\gamma > 0$ represents how reference traders react to changes in logarithmic stock prices. The last term can be interpreted as the demand created by trading decisions of hedging options. The option hedgers work under the assumption that the stock price $S^{(0)}$ is given by a geometric Brownian motion with constant drift μ_0 and volatility σ_0 to hedge a given number of call options with different maturities and strikes. This determines the term α_t in the market clearing equation. Then the market equilibrium condition determines the resulting price process $S_t^{(1)}$ by

$$dS_t^{(1)} = S_t^{(1)} \left(\sigma(S_t^{(1)}) dW_t + \mu(S_t^{(1)}) dt \right),$$

where

$$\begin{aligned} \sigma(s) &= -\frac{v}{\gamma + \xi'(\log(s))} \\ \mu(s) &= \frac{m}{v}\sigma(s) + \frac{1}{2}\sigma^2(s) + \frac{\xi''(\log(s))}{2v}\sigma^3(s) \end{aligned}$$

and the term $\xi(\log(s))$ represents the hedging demand created in the market. Observe that we started with a model $S_t^{(0)}$ for stock price process and derived another model $S_t^{(1)}$ by equilibrium approach that incorporates the hedging decisions of the large trader. However, the sophisticated large traders could also use the model $S_t^{(1)}$ to hedge derivatives so that we would obtain another model $S_t^{(2)}$ in equilibrium. In general, one can start from $S^{(k)}$ and use this to compute option values and hedging strategies. The equilibrium argument will yield a new model $S^{(k+1)}$. In the end, one wonders if there exists a fixed point $S^{(\infty)}$ of this transformation. Such a model $S^{(\infty)}$ would be used by the hedgers to compute their hedging strategy and would also be the one obtained in equilibrium.

Frey [15] takes the reaction function $S_t = \psi(t, F_t, \alpha_t)$ as given. He considers replicating the payoff of certain non-path dependent derivatives. In this continuous-time setup, there is a nonlinear partial differential equation for the hedge of the option replication problem. In particular, these hedging strategies take the feedback effect of their implementation on the price process into account. Therefore, Frey argues that the existence of these hedging strategies for certain payoffs corresponds to the fixed point of the volatility transformation introduced in [27].

Bank and Baum [7] assume that there exists a smooth family of semimartingales S^z for $z \in \mathbf{R}$ that specify the price process of the risky asset when the large trader's holdings are kept at a constant size z. For fixed z, the semimartingale S^z can be interpreted as the fluctuations of the asset prices when the large trader is not active in the market. If the large trader follows a semimartingale strategy $(Z_t)_{0\le t\le T}$, then the asset price obtained is given by

$$S_t = S_t^{Z_t} =: S(Z_t, t).$$

The self-financing portfolio strategies are characterized by

$$X_t = X_{0^-} - \int_0^t S(Z_{s^-}, s)dZ_s - [S(Z, \cdot), Z]_t.$$

Bank and Baum assume that asset prices are non-decreasing with respect to the position of the large trader, i.e. for $z \le z'$ we have $S^z \le S^{z'}$. In an illiquid market, there are many possible ways to value the large trader's portfolio. One can consider the book value Y_t of the portfolio evaluated at current prices,

$$Y_t = X_t + S(Z_t, t)Z_t,$$

or the real wealth achieved by the trading strategy Z until time t given by

$$V_t = X_t + L(Z_t, t),$$

where

$$L(z, t) = \int_0^z S(x, t)dx.$$

The term $L(z, t)$ represents the liquidation value of z shares by splitting the order into infinitesimally small packages and selling them over an infinitesimally small time period. By the Itô-Wentzell formula for smooth family of semimartingales, the real wealth process has the dynamics

$$V_t = V_{0^-} + \int_0^t L(Z_{s^-}, ds) - \frac{1}{2}\int_0^t S_z(Z_{s^-}, s)d[Z]_s^c - \sum_{0\le s\le t}\int_{Z_{s-}}^{Z_s} \{S(Z_s, s) - S(x, s)\}\, dx.$$

The term $\int_0^t L(Z_{s^-}, ds)$ represents the profit or loss coming from price fluctuations caused by exogenous random shocks. The term $\frac{1}{2}\int_0^t S'(Z_{s^-}, s)d[Z]_s^c$ gives the transaction costs due to continuous trading and

$$\sum_{0\le s\le t}\int_{Z_{s-}}^{Z_s} \{S(Z_s, s) - S(x, s)\}\, dx$$

sums up the transaction costs due to discrete block orders. These two transaction terms disappear if one follows trading strategies that are continuous and of bounded variation. As in [21,22], Bank and Baum investigate the possibility of arbitrage opportunities for the large trader. On one hand, the large trader has the power to influence the market prices, on the other hand, her trading incurs transaction costs, i.e. her orders affect the stock price before they are exercised. If there exists a measure $\mathbf{P}^* \approx \mathbf{P}$ which is a local martingale measure for all the processes P^θ simultaneously, then there are no arbitrage opportunities for the investor.

A natural problem in this setting is to describe the set of payoffs the large trader can attain with continuous strategies of bounded variation. To answer this question, Bank and Baum introduce two definitions. A contingent claim $H \in L^0(\mathcal{F}_T)$ is *attainable modulo transaction costs* for initial capital v if

$$H = v + \int_0^T L(Z_s, ds)$$

almost surely for some L-integrable predictable process Z such that $\int_0^\cdot L(Z_s, ds)$ is uniformly bounded from below. The claim H is *approximately attainable* for initial capital v if for any $\epsilon > 0$, there exists a self-financing strategy Z^ϵ such that $\int_0^\cdot L(Z_s^\epsilon, ds)$ is uniformly bounded from below, and

$$|H - V_T^\epsilon| \le \epsilon$$

holds $\mathbf{P}$ almost surely, in which V_T^ϵ is the real wealth process associated to strategy Z^ϵ. To this end, the authors establish an approximation scheme for stochastic integrals. Let $\epsilon > 0$. If Z is an L-integrable, predictable process with $Z_0 \in L^0(\mathcal{F}_0)$ and $Z_T \in L^0(\mathcal{F}_{T^-})$, then there exists a predictable process Z^ϵ with continuous paths of bounded variation such that $Z_0^\epsilon = Z_0$, $Z_T^\epsilon = Z_T$ and

$$\sup_{0\le t\le T}\left|\int_0^t L(Z_s, ds) - \int_0^t L(Z_s^\epsilon, ds)\right| \le \epsilon \quad \mathbf{P} - a.s.$$

From this, it is easy to see that any contingent claim $H \in L^0(\mathcal{F}_T)$ which is attainable modulo transaction costs is approximately attainable with the same initial capital. Furthermore, under some further assumptions the attainable claims in a suitable small investor model become approximately attainable for the large trader. Moreover, the authors show that to compute the superreplication cost of a claim $H(\omega, Z_T(\omega)) \in \mathcal{F}_{T^-} \otimes \mathcal{B}(\mathbf{R})$ one first determines the terminal position Z^*_T which minimizes the payoff, i.e. $Z^*_T(\omega) = argmin_{z\in\mathbf{R}} H(\omega, z)$, and then compute the small investor superreplication price of the claim $H(\omega, Z^*_T(\omega))$.

Roch [28] extends the linear version of the liquidity risk model of Çetin et al. [9] to allow for price impacts. The author considers the hedging problem faced by a large trader who makes market order through a limit order book with stochastic density. More specifically, it is assumed that the limit order book has a constant density at time t given by $\frac{1}{2M_t}$, in which M is an adapted stochastic process. Liquidity becomes a risk factor when the magnitude of the impact of these phenomena changes randomly over time. We denote by S the observed marginal price process, i.e. S_t is the price per share for an infinitesimal order size at time t. By the constant density property of the LOB, it is clear that a transaction of size ΔZ_t at time t has a cost of $\Delta Z_t(S_t + \lambda M_t \Delta Z_t)$. The model proposed in [28] is based on a well-documented feature of asset prices that volatility is high when liquidity is low, and low when liquidity is high. Since M is a measure of illiquidity, we can expect the instantaneous variance of the log-returns of the stock price to be in part correlated with M. Consequently, we let F denote the unaffected marginal price process. It is the equilibrium (or fundamental) price process observed in absence of large traders. It is defined by the following stochastic volatility model:

$$dF_t = \Sigma_t F_t dW_{1,t},$$

in which W_1 is the first component of the three-dimensional Brownian motion W, and Σ_t is the stochastic volatility. We are working directly under a risk neutral measure $\mathbf{Q}$ for unaffected prices, hence F has no drift term. We model M and Σ as follows. Define V and U as the solutions of

$$dU_t = \gamma(U_t + \eta)dt + \Phi(U_t)dW_{2,t},$$
$$dV_t = \alpha(V_t + a)dt + \Theta(V_t)dW_{3,t}$$

in which $W = (W_{j,t})_{j\le 3, t\le T}$ is a three dimensional Brownian motion, and $\alpha, \gamma, \eta, a \in \mathbf{R}$. We define $\Sigma^2_t = U_t + V_t$ and let $M = \varepsilon\Gamma(U)$, in which Γ is strictly increasing and twice continuously differentiable. Φ and Θ are given real-valued functions. We are using a three dimensional Brownian motion since there are three different sources of risk in this model, namely the stock price, the liquidity level and the volatility, which is, in practice, only partially dependent on the level of liquidity.

The specification of the process S is similar to the one of the LOB models described above. Indeed, it is assumed that the observed marginal price process S is obtained from the unaffected process F by directly adding the impact of the large trader as follows:

$$S_{t+} = F_t + 2\lambda \int_0^t M_{u-} dZ_u + 2\lambda \int_0^t d[M, Z]_u \quad (t \leq T)$$

for a semimartingale trading strategy Z. S_{t+} denotes the observed price after the trade at time t. λ is a resilience parameter, and should be taken between 0 and 1. It measures the proportion buy (resp. sell) limit orders versus sell (resp. buy) limit orders that come in to fill up the LOB after a market order to buy (resp. sell).

It can be shown that the money market account position X and the position Z in the stock satisfy the following identity:

$$X_t + Z_t(S^0_{t+} - \lambda M_t Z_t) = X_{t_0-} + Z_{t_0-}(S^0_{t_0} - \lambda M_{t_0} Z_{t_0-}) + \int_{t_0}^T Z_{u-} dS_u - \lambda \int_{t_0}^T Z^2_{u-} dM_u - \int_{t_0}^T (1-\lambda) M_u d[Z, Z]_u. \quad (1.10)$$

One can think of $Y_t + x(S^0_t - \lambda M_t x)$ as the liquidation value of a portfolio with x shares at time t. Similar to the infinitely-liquid case ($M = 0$), (1.10) states that the difference in the liquidation values between time t_0 and t is equal the cumulative gains in the risky asset $\int_{t_0}^t X_{u-} dS_u$, except that in this case there are added costs coming from the finite liquidity of the asset. First note that if $\lambda = 0$ we get a linear version of the CJP model. The integral with respect to M is related to the impact of trading. If $\lambda = 0$, the limit order book is automatically refilled after a market order, as in the CJP model. At the other extreme, when $\lambda = 1$ the impact of trading is at its fullest. It is interesting to notice that whatever the trading strategy used an investor always has a partial benefit from the asset becoming more liquid. Indeed, when M_t decreases, the associated integral is positive no matter what the sign of X_t is. To understand this, it is important to remember that the hedger's trades have a permanent impact on the quoted price which is proportional to the level of liquidity. If the liquidity is low when he purchases a share and high when she sells it, the price goes up higher after her purchase then it comes down after the sale. As a result, the hedger has a partial gain from this trade. This is a typical characteristic of large trader models. Note that unless the hedger uses a trading strategy with zero quadratic variation this is only a partial benefit because there is always a liquidity cost associated to her trades.

Equation (1.10) allows us to obtain a sufficient condition to rule out arbitrage opportunities in this setting. Indeed, Roch [28] shows that the existence of an equivalent measure $\mathbf{Q}$ under which the unaffected price process F is a local martingale and M is a local submartingale suffices to exclude the existence of arbitrage opportunities. For a precise statement, we refer the reader to Definition 2.5 and Theorem 2.6 of [28]. The advantage of this result is that it is stated in terms of the exogenously defined processes F and M. Note that in the terminology of Section 1.2 the impact of the hedger's trade in the above model is linear, i.e. a trade of size ΔZ_t at time t is in the form $g_t(\Delta Z_t) = 2\lambda M_t \Delta Z_t$. The case of M_t constant corresponds to the linear permanent impact models of Huberman and Stanzl [20], Almgren and Chriss [5] and others. In this case M clearly is a local submartingale under any risk-neutral measure for S. In this sense, the no arbitrage condition in [28] extends the results of Huberman and Stanzl [20] in the case of a stochastic linear permanent impact function.

We now turn to the replication problem. The relation between liquidity and volatility risk is a key observation which enables us to hedge derivatives. Indeed, we will see that one can hedge against the liquidity risk by trading variance swaps. Since volatility is one of the most correlated quantities to liquidity risk, this is a very natural choice. We thus consider contingent claims denoted by G_i $(i = 1, 2)$ for which the payoff at time $T_i > T$ $(T_1 \neq T_2)$ equals the difference between the realized variance over the time interval $[0, T_i]$ and a strike K_i, i.e.,

$$\begin{aligned} G_{i,T_i} &= \int_0^{T_i} \Sigma_s^2 ds - K_i \\ &= \int_0^{T_i} (U_s + V_s) ds - K_i. \end{aligned}$$

To rule out arbitrage opportunities, we assume the unaffected price processes G^i are $\mathbf{Q}$-martingales $(i = 1, 2)$.

Let h be the payoff function of a European option with maturity T. Suppose h is a Lipschitz function. For $x \in \mathbf{R}$, define $\widetilde{S}_T^x := F_T - 2\lambda \int_0^T x \hat{Z}_{u-} dM_u$ in which $\hat{Z}$ is the solution of the replication problem in the case $\lambda = 0$, $\varepsilon = 0$ and $x = 1$. It can be shown that $\widetilde{S}_T^x$ is an approximation of the observed price process S obtained when the large trader hedges the option with payoff h. Jarrow [22] used a similar approach and interpreted $\hat{Z}_t$ as the market's perception of the option's "delta" Z_t. The main result of the paper states that $xh(\widetilde{S}_T^x)$ can be approximately replicated in L^2 for all $x \in \mathbf{R}$ in the sense that there exists a sequence of trading strategies Z^n for which the terminal wealth X_T after liquidation converges to $xh(\widetilde{S}_T^x)$ in L^2.

Due to the non-additivity of liquidity costs, it is clear that the replicating cost of x units of the option h is not x times the replicating price of one unit. Let $H_t^n(x)$ denotes the approximate-replication cost per unit of x units of h, then $H_t^n(0) = \mathbf{E}(h(S_T)|\mathcal{F}_t)$ and $H_t^n(x)$ is a.s. differentiable at $x = 0$. Furthermore, it can be shown that

$$\begin{aligned} \lim_{n\to\infty} \frac{d}{dx} H_t^n(0) = \lambda\ &\mathbf{E}\left(\int_t^T \mu(M_s) \hat{Z}_s^2 ds \Big| \mathcal{F}_t \right) \\ &- 2\lambda \mathbf{E}\left(h'(S_T) \left(\int_t^T \hat{Z}_s dM_s \right) \Big| \mathcal{F}_t \right) \end{aligned}$$

when h is differentiable everywhere except at a finite number of points.

Jarrow et al. [23] have used ideas from the above setup to construct a liquidity based model for financial bubbles which explains both bubble formation and bubble bursting. In contrast with the classical approach to bubbles based on local martingales, the authors define the asset's fundamental price process exogenously and asset price bubbles are endogenously determined by market trading activity. More specifically, they assume that the stock price is governed by the following dynamic:

$$S_t = F_t + 2\int_0^t \Lambda_u M_{u-} dZ_u \quad (t \leq T)$$

in which F is the fundamental price process, Λ is a process version of the resilience parameter λ in [28] and Z represents the signed volume of aggregate market orders (volume of market buy orders minus volume of market sell orders). The bubble at time t is then defined by $S_t - F_t$, the difference between the market price of the stock and its fundamental value. In their model, the impact of trading activity on the fundamental price process - derived in terms of a liquidity risk process M, the resilience process Λ and the market orders - is what generates price bubbles. They study conditions under which asset price bubbles are consistent with no arbitrage opportunities.

1.4 Supply Curve Models

Çetin, Jarrow and Protter [9] model illiquidity with a supply curve model. This supply curve incorporates the temporary impact of the trade size into the price of the security. Assume that the marginal price process S is given. Then the price deviation at time t from S_t is determined by the supply curve in terms of the size of the trade. We denote the price per share for a trade of ν shares at time t by $\mathbf{S}(t, S_t, \nu)$. For instance, for a supply curve of the form

$$\mathbf{S}(t, S_t, \nu) = S_t \exp(\Lambda \nu) \tag{1.11}$$

a trade of size ν would deviate from the marginal price process by a factor of $\exp(\Lambda\nu)$. Since Λ measures the price impact, it is called the liquidity parameter of the market. $\Lambda = 0$ corresponds to a infinitely liquid market. Investors are price-takers with respect to the curve and their trading decisions affect the price only instantaneously, hence they have no lasting impact. Therefore, the Çetin-Jarrow-Protter model (henceforth called CJP model) belongs to a temporary price impact setting. An order of size $\nu > 0$ is a buy and $\nu < 0$ is a sell. $\mathbf{S}(t, S_t, 0)$ is equal to the marginal price S_t. Apart from measurability and smoothness assumptions, we assume $\mathbf{S}(t, S_t, \nu)$ is non-decreasing in ν for $\nu > 0$ and non-increasing for $\nu < 0$. It is also non-negative.

Consider a finite horizon economy with $T > 0$. Take a filtered probability space $(\Omega, \mathcal{F}, \mathbf{F}, \mathbf{P})$ satisfying the usual conditions. We let $(W_t)_{0 \leq t \leq T}$ be a standard Brownian motion with respect to the filtration $\mathbf{F} = (\mathcal{F}_t)_{0 \le t \le T}$. Assume there are two assets in the economy, one risk-free asset and one risky asset. We consider a money market account as the risk-free asset and normalize its price to unity. The risky asset is by convention a stock and the price per share of stock is $\mathbf{S}(t, S_t, \nu)$ with the marginal price process S_t. Let X_t and Z_t represent the holdings of the trader at time t in the money market account and in the stock respectively. There are various ways to value the wealth process of the investor. One way is to look at the block liquidation value

$$X_t + Z_t \mathbf{S}(t, S_t, -Z_t).$$

Another way is to consider the book or paper value of the portfolio

$$Y_t := X_t + Z_t S_t$$

evaluated at the marginal process S. It is shown in [33] that this value Y_t also corresponds to infinitesimal liquidation value. In the remainder of the section we focus on the book value Y_t and specify its dynamics. It is natural to define the self-financing condition for simple strategies of the form $Z_t = \sum_{i=1}^{N} \Delta Z_{\tau_i} \mathbf{1}_{\{t \ge \tau_i\}}$ with a sequence of stopping times $0 = \tau_0 < \tau_1 < \cdots < \tau_N = T$ by

$$X_{\tau_{k+1}} = X_{\tau_k} - \Delta Z_{\tau_{k+1}} \mathbf{S}\left(\tau_{k+1}, S_{\tau_{k+1}}, \Delta Z_{\tau_{k+1}}\right), \tag{1.12}$$

where $\Delta Z_{\tau_{k+1}} = \left(Z_{\tau_{k+1}} - Z_{\tau_k}\right)$. Then the dynamics of the book value Y for simple strategies is described as

$$\begin{aligned} Y_{\tau_{k+1}} = Y_{\tau_k} &+ Z_{\tau_k}\left(S_{\tau_{k+1}} - S_{\tau_k}\right) \\ &- \Delta Z_{\tau_{k+1}}\left[\mathbf{S}\left(\tau_{k+1}, S_{\tau_{k+1}}, \Delta Z_{\tau_{k+1}}\right) - S_{\tau_{k+1}}\right]. \end{aligned} \tag{1.13}$$

Formally, for general semimartingale strategies Z, one can pass to the limit as $N \to \infty$ to obtain the dynamics

$$Y_t = y + \int_0^t Z_{u-} dS_u - \sum_{0 \le u \le t} \Delta Z_u \left[\mathbf{S}(u, S_u, \Delta Z_u) - S_u\right] \tag{1.14}$$

$$- \int_0^t \frac{\partial \mathbf{S}}{\partial \nu}(u, S_u, 0) d[Z, Z]_u^c \tag{1.15}$$

for $0 \le t \le T$. The term $\int_0^t Z_{u-} dS_u$ represents the capital gains and losses. The other terms in the above equation appear because of liquidity effects, the first one is a result of block orders and the second one of continuous trading. These liquidity costs can be eliminated by using continuous strategies of finite variation. Furthermore, Çetin et al. [9] prove that for any appropriately integrable predictable process Z, there exists a sequence $\{Z^n\}_{n \ge 0}$ of predictable continuous strategies of finite variation such that

$$\int_0^T Z_u^n dS_u \to \int_0^T Z_u dS_u \quad \text{in} \quad L^2.$$

This approximation also follows from the Bank and Baum [7] result as well.

Çetin et al. find sufficient conditions to rule out arbitrage in the CJP model. They generalize the first fundamental theorem of asset pricing to their setting. They show that there is no free lunch with vanishing risk in their framework if and only if there exists an equivalent local martingale measure for the marginal price process S_t. They also establish that if there exists an equivalent local martingale measure $\mathbf{Q}$ for the marginal price process S, then any appropriately integrable claim C can be attained in the L^2 sense. Then the above approximation result shows that all liquidity costs can be avoided in this setting and the value of the claim is the classical one given by $E^{\mathbf{Q}}[C]$.

The previous result is sometimes seen as a shortcoming of the CJP model. As a response, Rogers and Singh consider a temporary price impact model in [29] in which the liquidity cost cannot be avoided by the use of continuous strategies of finite variation. In their setup, the admissible portfolio processes $Z = (Z_t)_{0\le t\le T}$ are taken to be absolutely continuous with density $\dot{Z} = (\dot{Z}_t)_{0\le t\le T}$. The cost of liquidity enters into their wealth dynamics $Y = (Y_t)_{0\le t\le T}$ as a penalization of the speed of trading like in the framework of Almgren and Chriss [6]:

$$dY_t = Z_t dS_t - S_t l(\dot{Z}_t)dt.$$

They take S_t as a geometric Brownian motion with zero drift and l a convex, non-negative function with $l(0) = 0$. In [7] and [9], all transaction costs due to illiquidity can be eliminated by using continuous strategies of finite variation. However, in the setup of Rogers and Singh [29], the use of these strategies induces a liquidity cost. Assume that an investor holds Z_0 number of shares, x units of money market account and she wants to replicate a European contingent claim with payoff $g(S_T)$. Since the Black-Scholes hedge $\theta(t, S_t)$ of a European contingent claim is of infinite variation, it will incur infinite liquidity costs. As a result the authors propose to minimize the mean squared hedging error and the associated liquidity costs incurred over portfolio processes $Z = (Z_t)_{0\le t\le T}$

$$\frac{1}{2}E\left[\left(x + Z_0 S_0 + \int_0^T Z_t dS_t - g(S_T)\right)^2\right] + E\left[\int_0^T S_t l\left(\dot{Z}_t\right) dt\right].$$

They solve the Hamilton-Jacobi-Bellman equation for the associated optimal control problem in almost closed form and study it numerically.

Çetin, Soner and Touzi [11] study the superreplication problem using the CJP model under the additional constraint on the boundedness of the quadratic variation and the absolute continuous parts of the portfolio processes. Their driving motivation is the lack of liquidity premium, i.e. the extra amount one has to pay due to illiquidity, in the papers by Bank and Baum [7], and Çetin et al. [9] as a result of using continuous strategies of bounded variation. They link the absence of the liquidity premium to the choice of admissible strategies and show that one can find a nonzero liquidity premium in continuous-time for a set of admissible strategies appropriately defined. Their results and the justification for the set of admissible strategies they consider are well supported by a convergence result of the discrete-time setting in [18]. In fact, there are no restrictions on the portfolio strategies in [18]. As the dynamics of the paper value of the portfolio Y in (1.14) is obtained as a limit of the natural discrete time self-financing conditions, this is a justification of the validity of the constraints placed on the portfolio processes in [11]. In particular, Gökay and Soner [18] analyze the asymptotic limit of the Binomial version of the CJP model both numerically and theoretically. Although there are no constraints placed on the portfolio processes in their model, Gökay and Soner recover the same super-replicating cost as in [11] in the limit, hence show that the liquidity premium persists in the continuous-time.

Çetin et al. [11] consider a marginal process S satisfying the stochastic differential equation

$$S_r = s + \int_t^r S_u \sigma(u, S_u) dW_u^0,$$

which has a strong solution denoted by $S^{t,s}$ with the initial condition $S_t = s$. Moreover, they take the portfolio process Z to be of the form

$$Z_r = \sum_{n=0}^{N-1} z_n \mathbf{1}_{\{r \geq \tau_{n+1}\}} + \int_t^r \alpha_u du + \int_t^r \Gamma_u dS_u^{t,s},$$

where $t = \tau_0 < \tau_1 < \cdots$ is an increasing sequence of $[t, T]$-valued $\mathbf{F}$-stopping times, the random variable

$$N := \inf\{n \in \mathbf{N} : \tau_n = T\}$$

indicates the number of jumps and z_n is $\mathcal{F}(\tau_n)$-measurable. The infinite variation part of this trading strategy consists of an integral with respect to the marginal price process S, where the integrand is the gamma $\Gamma = (\Gamma_t)_{0 \leq t \leq T}$ of the portfolio. The integrands α and Γ are $\mathbf{F}$-progressively measurable processes. Moreover, there are additional constraints imposed on the processes Z, α and Γ similar to those in [12] and [32]. Then the authors consider superreplicating a European contingent claim with payoff g. The payoff g is continuous, non-negative and satisfies $g(s) \leq C(1+s)$ for some constant C. If the supply curve is of the form (1.11), then the superreplicating cost $\phi(t,s)$ is the unique viscosity solution of the following dynamic programming equation

$$-\phi_t(t,s) + \sup_{\beta \geq 0} \left(-\frac{1}{2} s^2 \sigma^2 (\phi_{ss}(t,s) + \beta) - \Lambda s^2 \sigma^2(t,s) (\phi_{ss}(t,s) + \beta)^2 \right) = 0$$

and satisfies the terminal condition $\phi(T,\cdot) = g(\cdot)$ along with the growth condition $0 \leq \phi(t,s) \leq C(1+s)$ for some constant C. With constant volatility σ, one can rewrite it as

$$-\phi_t(t,s) - s^2\sigma^2 H(\phi_{ss}(t,s)+\beta) = 0, \tag{1.16}$$

in which

$$H(\gamma) = \begin{cases} \frac{1}{2}\gamma + \Lambda\gamma^2 & \gamma \geq -\frac{1}{4\Lambda}, \\ -\frac{1}{16\Lambda} & \gamma < -\frac{1}{4\Lambda}, \end{cases}$$

and the liquidity parameter Λ is given as $\frac{\partial \mathbf{S}}{\partial \nu}(t,0)$. For $\Lambda = 0$, one recovers the Black-Scholes setting. In fact, if ϕ_{BS} is the Black-Scholes value of the claim g, then by a maximum principle argument one concludes that $\phi(t,s) \geq \phi_{BS}(t,s)$. Moreover, ϕ and ϕ_{BS} coincide if and only if the payoff is an affine function. This implies that there exists a liquidity premium, a difference between the superreplicating cost ϕ and the Black-Scholes value ϕ_{BS}, for non-trivial claims g. This result conflicts with the statement that in an illiquid market all liquidity costs can be avoided by approximating with continuous strategies with finite variation. The intuitive reasoning is that such an approximation neutralizes the gamma of the portfolio process, however it makes α infinitely large in the limit so that it no longer satisfies the imposed constraints.

Çetin et al. [11] also study the associated super-hedging strategy under liquidity costs. They characterize a set $\mathcal{C}$ such that outside $\mathcal{C}$, the hedging strategy is given by $\phi_s(t,s)$ and in $\mathcal{C}$ the strategy is a mixture of dynamically replicating an auxiliary function ψ and applying a buy and hold strategy to $\phi - \psi$. The set $\mathcal{C}$ is determined by a level of concavity on the value function ϕ.

Gökay and Soner [18] study a discrete version of the supply curve model. For a fixed step size $h > 0$, they divide the trading period $[0,T]$ into equal intervals of length h. The evolution of the marginal price process is given by a Binomial tree, i.e. at any node (t, S_t) it either goes up by a factor of $1+\sigma\sqrt{h}$ or down by a factor of $1-\sigma\sqrt{h}$. We use the notation

$$S_{t+h} = S_t(1 \pm \sigma\sqrt{h}).$$

The filtration $\mathbf{F}$ is generated by the marginal price process S and the portfolio process Z is taken to be adapted with respect to $\mathbf{F}$. They consider a supply curve of the form

$$\mathbf{S}(t,s,\nu) = S_t + \Lambda\nu$$

with liquidity parameter Λ. Observe that this supply curve may take negative values, so one may consider $\mathbf{S}(t,S_t,\nu) = (S_t + \Lambda\nu)^+$, however the analysis in [18] shows that both supply curves yield to the same partial differential equation in the limit. The self-financing condition is given as in (1.12) and the book value Y has the dynamics of (1.13). We introduce the notation $Z_\cdot^{t,z}$

to denote the portfolio process with initial condition $Z_t = z$ and $Y_t^{t,y,Z}$ the book value that starts $Y_t = y$ and uses the control $Z.$. The authors study the super-replication problem of a European contingent claim with payoff g. As in [11], the payoff g is continuous, non-negative and satisfies the linear growth condition $g(s) \leq C(1+s)$ for some $C > 0$. The minimal super-replicating cost $\phi^h(t,s)$ at time t and $S_t = s$ is given by

$$\phi^h(t,s) = \inf \left\{ y \mid \exists \, \mathbf{F} - \text{adapted } \{Z.\} \text{ so that } Y_T^{t,y,Z.} \geq g(S_T^{t,s}) \ a.s. \right\}.$$

The main observation is that dynamic programming approach fails for the value function $\phi^h(t,s)$, therefore to restore dynamic programming one needs introduce the dependence of the value function on the portfolio position z in addition to the current stock price and time. So we define

$$\begin{aligned} v^h(t,s,z) := \inf \{ y \mid \exists \, \mathbf{F} - &\text{adapted } \{Z.\} \\ &\text{so that } Z_t = z \text{ and } Y_T^{t,y,Z.} \geq g(S_T^{t,s}) \ a.s. \}. \end{aligned}$$

Clearly,

$$\phi^h(t,s) = \inf_z v^h(t,s,z).$$

The following dynamic programming is the key element of the analysis of Gökay and Soner [18]

$$\begin{aligned} v^h(t,s,z) = \inf \{ y \mid \exists \, \mathbf{F} - &\text{adapted } \{Z.\} \\ &\text{s.t } Z_t = z \text{ and } Y_\tau^{t,y,Z.} \geq v^h(\tau, S_\tau^{t,s}, Z_\tau) \ a.s. \}, \end{aligned}$$

in which $t = nh < \tau = mh \leq T$ for some $n, m \in \mathbf{N}$. In particular for $\tau = t + h$ we have the following form

$$\begin{aligned} v^h(t,s,z) = \max \Bigg(&\min_a \left\{ v^h(t+h, su, z+a) - zs\sigma\sqrt{h} + \Lambda a^2 \right\}, \\ &\min_b \left\{ v^h(t+h, sd, z+b) + zs\sigma\sqrt{h} + \Lambda b^2 \right\} \Bigg). \end{aligned}$$

This equation is complemented by the terminal data

$$v^h(T,s,z) = g(s).$$

Using the theory of viscosity solutions, the authors pass to the limit by letting the time step $h \downarrow 0$. In particular, they show that $v^h(t,s,z)$ converges to the solution $\phi(t,s)$ of the partial differential equation (1.16) locally uniformly as $h \downarrow 0$. To this aim they consider the standard upper and lower relaxed limits in the theory of viscosity solutions

$$\phi^*(t,s,z) = \limsup_{\substack{h \to 0 \\ (t',s',z') \to (t,s,z)}} v^h(t',s',z'),$$

$$\phi_*(t,s,z) = \liminf_{\substack{h\to 0\\(t',s',z')\to(t,s,z)}} v^h(t',s',z').$$

The authors prove that $\phi^*(t,s,z)$ is independent of z and set

$$\phi^*(t,s) := \phi^*(t,s,z).$$

However, it is difficult to derive directly a similar claim for $\phi_*(t,s,z)$. In fact, the challenge in proving this convergence result is that in discrete-time the value function $v^h(t,s,z)$ depends on the initial portfolio value z, whereas this dependence becomes irrelevant in the limit $\phi(t,s)$. Therefore, the authors overcome this difficulty by defining

$$\phi_*(t,s) = \inf_z \left\{ \liminf_{\substack{h\to 0\\(t',s',z')\to(t,s,z)}} v^h(t',s',z') \right\}$$

and developing further the idea of corrector functions as in the applications of viscosity solutions to homogenization. The authors proceed by showing that the upper semi-continuous relaxed limit $\phi^*(t,s)$ is a viscosity subsolution and the lower semi-continuous relaxed limit $\phi_*(t,s)$ is a viscosity supersolution of the partial differential equation (1.16). Moreover, both ϕ_* and ϕ^* are growing almost linearly and attain $\phi_*(T,s) = \phi^*(T,s) = g(s)$. So by the comparison argument established in [11], they conclude that $\phi_* = \phi^*$ and it is equal to the unique viscosity solution of (1.16). Now the local uniform convergence of $v^h(t,s,z)$ to $\phi(t,s)$ will follow from the definitions of ϕ_* and ϕ^*.

1.5 Expected Utility Maximization in Illiquid Markets

In this section, we briefly review some results regarding the problem of expected utility maximization in illiquid markets. We first consider the permanent price impact setting of Ly Vath et al. [24], and then the setup of temporary price impacts in discrete time, as done in Çetin and Rogers [10].

Ly Vath et al. [24] solve the expected utility maximization problem with permanent price impacts in continuous time with admissible strategies of the form

$$Z_t = \xi_0 + \sum_{i=1}^{N} \xi_n \mathbf{1}_{\{t\geq\tau_n\}} \quad (0 \leq t < T), \tag{1.17}$$

in which $\{\tau_n\}_{n\geq 1}$ is a sequence of stopping times and $\xi_n \in \mathcal{F}_{\tau_n}$ for all $n \geq 1$. A trade of size ξ at time t is assumed to have a permanent impact of the exponential form. Furthermore, they assume that the stock price evolves as a geometric Brownian motion between trades, i.e.

$$dS_t = \mu S_{t-}dt + \sigma S_{t-}dW_t + \lambda S_{t-}dZ_t$$

for some positive constants $\lambda, \sigma > 0$ and $\mu \in \mathbf{R}$. Each time a transaction is made, the investor pays a fixed transaction cost k so that the money market account obeys the following equation:

$$X_t = \int_0^t rX_{s-} ds - \int_0^t S_{s-} e^{\lambda \Delta Z_s} dZ_s - \sum_{i \geq 1} k \mathbf{1}_{\{\tau_n \leq t\}}.$$

A strategy Z belongs to the set of admissible strategies $\mathcal{A}(t, x, z, s)$ started at time t with $X_t = x$, $Z_t = z$ and $S_t = s$ if it satisfies the solvency constraint

$$X_s + S_{s-} e^{-\lambda Z_s} Z_s - k \geq 0$$

for all $t \leq s \leq T$. The second term in the above inequality is the liquidation value of a position of size Z_t in the risky asset S. The solvency constraint states that the liquidation value of an admissible portfolio is always positive. Due to this fixed cost at each transaction, the authors show that the optimal trading strategy which maximizes the expected utility is indeed in the form of (1.17) and they describe the optimal trading times τ_n in terms of the value of the money market account, the position in the risky asset and the current price. Their main result is to show that the value function

$$v(t, x, z, s) = \sup_{Z \in \mathcal{A}(t,x,z,s)} \mathbf{E} U(X_T + S_{T-} e^{-\lambda Z_T} Z_T - k)$$

is a viscosity solution of the following quasi-variational Hamilton Jacobi Bellman inequality:

$$\min\{-\frac{\partial v}{\partial t} - rx \frac{\partial v}{\partial x} - \mathcal{L}v, v - \mathcal{H}v\} = 0$$

in which $\mathcal{L}$ is the infinitesimal generator of a geometric Brownian motion and $\mathcal{H}$ is an impulse generator of the form

$$\mathcal{H}v(t, x, z, s) = \sup_{\xi} v(t, x - se^{\lambda \xi}\xi, z - \xi, se^{\lambda \xi})$$

with the supremum taken over the set of transactions that satisfy the solvency condition.

Çetin and Rogers [10] study the discrete-time utility maximization problem using a supply curve of the form

$$\mathbf{S}(t, S_t, \nu) = \varphi(\nu) S_t,$$

in which φ is a strictly increasing and strictly convex function. Their objective is to maximize utility from terminal liquidation value $Y_N = X_N + Z_N S_N$, where $Z_N = 0$ and U is a strictly concave and strictly increasing utility function. They show that this problem has a solution. Moreover, the marginal utility of optimal terminal wealth $U'(Y_N)$ is an equivalent martingale measure and the process $M_n = \varphi'(\Delta Z_n) S_n$ becomes a martingale under this measure.

1.6 Price Manipulation strategies in Price Impact Models

So far, there is one fundamental notion of finance we have not addressed: arbitrage from price manipulations. The assumption that the large trader has a temporary and permanent impact on the prices clearly suggests the possibility that she can manipulate the prices in her favor. In Section 1.2, this issue has been partly avoided by either assuming a priori that the execution of the large sell (resp. buy) order is restricted to smaller sell (buy) orders or that this condition is satisfied a posteriori as a consequence of the assumptions made. Indeed, in the former case, arbitrage is not possible since a sell order makes the price lower so that the next sell order will come at a less favorable price. In more general models, however, there sometimes exists weaker version of the arbitrage condition. For instance, the widespread concept of quasi-arbitrage and price manipulations which correspond to strategies with a negative expected cost is often considered in the literature. This particular approach can be found in the papers of Huberman and Stanzl [19], Gatheral [17], and Jarrow [21, 22].

To make the notion of quasi-arbitrage more precise, Huberman and Stanzl [19] define the notion of a *round trip*, a trading strategy that starts with zero shares and terminates with zero shares of the risky asset. They consider a model in discrete time, with n time steps. There are noise traders and we denote by η_k the number of shares of the risky asset they purchase at time k. As before, ξ_k denotes the trade size of the large trader at time k. Let $\{\zeta_k\}_{k=1,\dots,N}$ be i.i.d. random variables with zero expectation. We also assume $\{\eta_k\}_{k=1,\dots,N}$ are i.i.d. random variables with zero expectation. The authors consider the following dynamic for the marginal price of the risky asset:

$$S_k = S_{k-1} + g(\xi_k + \eta_k) + \zeta_k.$$

They also hypothesize the existence of a temporary price impact function h, so that the large trader pay a total of $\xi_k(S_k + h(\xi_k + \eta_k))$ at time k. The temporary impact includes the noise traders' trading volume η_k, and the η_k's are assumed to be unknown by the large trader at the moment of the transaction at time k. The profit of a round trip is given by $\pi(\xi) = -\sum_{k=1}^{n} \xi_k(S_k + h_k(\xi_k + \eta_k))$. Huberman and Stanzl [19] define a *price manipulation* as a round trip with positive expected value. They also define a *quasi-arbitrage* as a sequence of round trips $\xi^m = \{\xi_k^m\}_{k=1,\dots,n}$ for $m = 1, 2, \dots$ such that $\lim_{m\to\infty} \mathbf{E}\pi(\xi^m) = \infty$ and

$$\lim_{m\to\infty} \frac{\mathbf{E}\pi(\xi^m)}{\sqrt{\mathrm{Var}(\pi(\xi^m))}} = \infty.$$

Their main result states that if $\mathbf{P}(\eta_k = 0) = 1$ $(k = 1, \dots, n)$ or the η_k's are normally distributed then the absence of price manipulation implies that the permanent impact function g is linear. On the other hand, no restrictions is required on the temporary impact function h.

Gatheral [17] considers models for stock prices with price impacts that decay with time. More specifically, he focuses on models on the following form:

$$S_t = S_0 + \int_0^t g(\dot{X}_s)G(t-s)ds + \sigma W_t,$$

in which g is the permanent impact function and G is the decay factor. In words, the permanent impact of a trade at time t decays with time due to the function G. The setting is the same as in (1.4) when $G = 1$. The author finds a relationship between the shape of the market impact function f and the resilience function G under the no-dynamic-arbitrage assumption. In particular, he obtains similar results to Huberman and Stanzl [19] regarding the linearity of the price impact function.

In [21], Jarrow considers a discrete-time economy. In his model, the stock price can be expressed in terms of a sequence $\{g_{t_k}\}_{0\leq k\leq N}$ with $g_{t_k} : \Omega \times \mathbf{R}^{t+1} \to \mathbf{R}$ such that

$$S_{t_k}(\omega) = g_{t_k}(\omega, Z_{t_k}(\omega), \ldots, Z_0(\omega)) \quad \forall\, \omega \in \Omega,\ 0 \leq k \leq N.$$

The functions $\{g_{t_k}\}_{0\leq k\leq N}$ are the *reaction* functions, which reflect how the participants of the market react to large trader's portfolio decisions. Particular cases of these functions are the permanent and temporary impact function described in Section 1.2. These reaction functions provide the reduced form equilibrium relationship between relative prices and the large trader's trades. In [21], Jarrow concentrates on market manipulation strategies for the large trader. In Jarrow's terminology, a *market manipulation strategy* is a strategy that can generate positive real wealth for the large trader without taking any risk. The real wealth for the large trader is characterized as the value of her portfolio after liquidation. Market manipulation strategies are shown to sometimes exist in this economy. Sufficient conditions are provided that restrict the market manipulation strategies. These conditions include the requirement that the stock price process is independent of the past holdings of the large trader and depends only on her instantaneous holdings, i.e.

$$S_{t_k}(\omega) = g_{t_k}(\omega, Z_{t_k}(\omega))$$

and that if the large trader is not active in the time interval $[t_k, t_{k+1}]$, then there are no arbitrage opportunities available for the reference traders in this time period. In [22], Jarrow extends this framework for markets that include a derivative security. He finds sufficient conditions to exclude market manipulation strategies, after showing that market manipulation strategies can exist after the introduction of the derivative security. To avoid market manipulation strategies, the market must be in synchrony. This means that the number of shares, whether bought in the stock market or acquired jointly in the stock and derivative market, should yield the same stock price. Moreover, Jarrow shows that one can hedge options using the standard method based on the binomial model with random volatilities.

References

1. A. Alfonsi, A. Fruth, and A. Schied. Constrained portfolio liquidation in a limit order book model. *Banach Center Publ.*, 83:9–25, 2008.
2. A. Alfonsi, A. Fruth, and A. Schied. Optimal execution strategies in limit order books with general shape functions. *Quantitative Finance*, Volume 10, Issue 2:143–157, 2010.
3. A. Alfonsi and A. Schied. Optimal execution and absence of price manipulations in limit order book models. Preprint, 2009.
4. R. Almgren. Optimal execution with nonlinear impact functions and trading-enhanced risk. *Applied Mathematical Finance*, 10:1–18, 2003.
5. R. Almgren and N. Chriss. Value under liquidation. *Risk*, 12(12):61–63, 1999.
6. R. Almgren and N. Chriss. Optimal execution of portfolio transactions. *J. Risk*, 3:5–39, 2000.
7. P. Bank and D. Baum. Hedging and portfolio optimization in financial markets with a large trader. *Mathematical Finance*, 14:1–18, 2004.
8. D. Bertsimas and A. Lo. Optimal control of execution costs. *Journal of Financial Markets*, 1:1–50, 1998.
9. U. Çetin, R. Jarrow, and P. Protter. Liquidity risk and arbitrage pricing theory. *Finance and Stochastics*, 8:311–341, 2004.
10. U. Çetin and L.C.G Rogers. Modelling liquidity effects in discrete time. *Mathematical Finance*, 17:15–29, 2007.
11. U. Çetin, H. M. Soner, and N. Touzi. Option hedging for small investors under liquidity costs. to appear in Finance and Stochastics, 2007.
12. P. Cheridito, H.M. Soner, N. Touzi, and N. Victoir. Second order backward stochastic differential equations and fully non-linear parabolic pde's. *Comm. on Pure and Applied Math*, 60(7):1081–1110, 2007.
13. D. Cuoco and J. Cvitanic. Optimal consumption choices for a large investor. *Journal of Economic Dynamics and Control*, 22:401–436, 1998.
14. J. Cvitanic and J. Ma. Hedging options for a large investor and forward backward sdes. *Annals of Applied Probability*, 6:370–398, 1996.
15. R. Frey. Perfect option hedging for a large trader. *Finance and Stochastics*, 2:115–141, 1998.
16. R. Frey and A. Stremme. Market volatility and feedback effects from dynamic hedging. *Mathematical Finance*, 7:351–374, 1997.
17. J. Gatheral. No-dynamic-arbitrage and market impact. To appear in Quantitative Finance, 2008.
18. S. Gökay and H.M. Soner. Çetin-Jarrow-Protter model of liquidity in a binomial market. 2009.
19. G. Huberman and W. Stanzl. Price manipulation and quasi-arbitrage. *Econometrica*, 72(4):1247–1275, 2004.
20. G. Huberman and W. Stanzl. Optimal liquidity trading. *Review of Finance*, 9:165–200, 2005.
21. R. Jarrow. Market manipulation, bubbles, corners and short squeezes. *The Journal of Financial and Quantitative Analysis*, 27(3):311–336, 1992.
22. R. Jarrow. Derivative securities markets, market manipulation and option pricing theory. *The Journal of Financial and Quantitative Analysis*, 29(2):241–261, 1994.
23. R. Jarrow, P. Protter, and A. Roch. A liquidity based model for asset price bubbles. Preprint, 2010.

24. V. Ly Vath, M. Minf, and H. Pham. A model of optimal portfolio selection under liquidity risk and price impact. *Finance and Stochastics*, 11:51–90, 2007.
25. J. Ma, P. Protter, and J. Yong. Solving forward-backward stochastic differential equations explicitely–a four step scheme. *Probab. Theory Related Fields*, 98:339–359, 1994.
26. A. Obizhaeva and J. Wang. Optimal trading strategy and supply/demand dynamics. Preprint, 2005.
27. E. Platen and M. Schweizer. On feedback effects from hedging derivatives. *Mathematical Finance*, 8:67–84, 1998.
28. A. Roch. Liquidity risk, price impacts and the replication problem. Preprint, 2009.
29. L.C.G. Rogers and S. Singh. The cost of illiquidity and its effects on hedging. Preprint, 2007.
30. 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.
31. R. Sircar and G. Papanicolaou. General black-scholes models accounting for increased market volatility from hedging strategies. *Applied Mathematical Finance*, 5:45–82, 1998.
32. H.M. Soner and N. Touzi. Superreplication under gamma constraints. *SIAM J. Control and Opt.*, 39(1):73–96, 2000.
33. H.M. Soner and N. Touzi. The dynamic programming equation for second order stochastic target problems. *SIAM J. Control and Opt., 48(4), 2344-2365*, 2009.