Optimal execution with non-linear transient market impact

GIANBIAGIO CURATO†, JIM GATHERAL∗‡ and FABRIZIO LILLO†§

†Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy
‡Baruch College, One Bernard Baruch Way, New York, NY 10010, USA
§QUANTLab, Via Pietrasantina 123, 56122 Pisa, Italy

(Received 6 October 2015; accepted 13 April 2016; published online 24 May 2016)

We study the problem of the optimal execution of a large trade in the propagator model with non-linear transient impact. From brute force numerical optimization of the cost functional, we find that the optimal solution for a buy programme typically features a few short intense buying periods separated by long periods of weak selling. Indeed, in some cases, we find negative expected cost. We show that this undesirable characteristic of the non-linear transient impact model may be mitigated either by introducing a bid–ask spread cost or by imposing convexity of the instantaneous market impact function for large trading rates; the objective in each case is to robustify the solution in a parsimonious and natural way.

Keywords: Optimal execution; Market impact; Price manipulation; Sequential quadratic programming; Liquidity

JEL Classification: C20, C61, C88, D44, D47, D53

1. Introduction

The optimization of trading strategies has long been an important goal for investors in financial markets. As demonstrated in the context of a linear equilibrium model by Kyle almost 30 years ago (Kyle 1985), the optimal strategy for an investor with insider information on the fundamental price of an asset is to trade incrementally through time. This strategy allows the trader to minimize costs while also minimizing the revelation of information to the rest of the market. The precise way in which it is optimal to split the large order (herein called metaorder) Zarinelli et al. (2015) depends on the objective function and on the market impact model, i.e. the change in price conditioned on signed trade size. In part due to the increasing tendency towards a full automation of exchanges and in part due to the discovery of new statistical regularities of the microstructure of financial markets, the problem of optimal execution is receiving growing attention from the academic and practitioner communities (Abergel et al. 2012, Gatheral and Schied 2013).

As pointed out in Gatheral and Schied (2013), a first generation of market impact models (Bertsimas and Lo 1998, Almgren and Chriss 1999, Almgren and Chriss 2001, Almgren 2003) distinguishes between two impact components. The first component is temporary and only affects the individual trade that has triggered it. The second component is permanent and affects all current and future trades equally. These models can be either in discrete or in continuous time and can assume either linear or non-linear market impact for individual trades.

The second generation of market impact models focuses on the transient nature of market impact (Bouchaud et al. 2004, Bouchaud et al. 2006, Gatheral and Schied 2013, Obizhaeva and Wang 2013).

In this paper, we focus on one such second-generation model, the continuous time propagator model introduced in Gatheral (2010). Let $S_t$ denote the stock price at time $t$ and $x_t$ denote the outstanding position in shares at time $t$ with $x_0 = X$. The model reads

$$S_t = S_0 + \int_0^t f(\dot{x}_s) G(t-s) \, ds + \int_0^t \sigma \, dW_s,$$

(1.1)

where $\dot{x}_s$ is the rate of trading (number of shares per unit of time) at time $s < t$, $f(\dot{x}_s)$ represents the impact of trading at time $s$ and the decay kernel $G(t-s)$ describes how much of the initial impact at time $s < t$ remains at the later time $t$. Finally, $\sigma$ is the volatility of the underlying and $W_t$ is a Wiener process.

The problem of optimal execution in the presence of transient impact has been considered in a series of recent studies. In the case of linear instantaneous market impact (Busseti and
Lillo 2012, Alfonsi and Schied 2013, Gatheral and Schied 2013), where the instantaneous market impact function \( f(v) \) is proportional to the trading rate \( v = \dot{x} \), the problem has been completely solved by showing that the cost minimization problem is equivalent to solving an integral equation. In particular, Gatheral et al. (2012) proved that optimal strategies may be characterized as measure-valued solutions of a Fredholm integral equation of the first kind. They showed that optimal strategies always exist and are non-alternating between buy and sell trades when price impact decays as a convex function of time. This extends the result of Alfonsi et al. (2012) on the non-existence of transaction-triggered price manipulation, i.e. strategies where the expected execution costs of a sell (buy) programme can be decreased by intermediate buy (sell) trades.

However, many empirical studies (Lillo et al. 2003, Bouchaud et al. 2009, Bacry et al. 2014) have clearly shown that instantaneous market impact is a strongly concave function of the volume, well approximated by a power law function, at least for trading rates that are not too high. The resulting optimal execution problem in the presence of non-linear and transient impact is mathematically much more complicated than in the linear case. In this paper, we study this optimization problem and identify some properties of the solution.

Some important results in the non-linear transient case have already been established by Gatheral (2010) who showed that under certain conditions, the model admits price manipulation, i.e. the existence of round trip strategies with positive expected revenues. This money machine should of course be avoided in the modelling of market impact. In particular, Gatheral set some necessary conditions for the absence of price manipulation.

An attempt at a solution of the optimal execution problem under non-linear transient impact has been presented by Dang (2014). Motivated by the solution in the linear case, Dang suggests a way to convert the cost minimization problem into a non-linear integral equation and proposes a numerical fixed point method on a discretization of the trading time interval to solve this equation; the solution of this integral equation using discrete homotopy analysis is presented in Curato et al. (2016). However, the class of admissible strategies is restricted to that of continuous non-vanishing functions of the trading rate.

By minimizing the cost function directly using sequential quadratic programming (SQP), we show that the optimal liquidation strategy is not of the absolutely continuous form assumed by Dang. We find that the cost landscape is rugged, with a very large number of local minima separated by peaks. A significant number of these minima correspond to strategies with similar costs; for a buy programme, the corresponding strategy is an alternation of intense and short bursts of buying periods and long periods of weak selling. In other words, the model admits transaction-triggered price manipulation. More significantly, if the instantaneous market impact function is sufficiently concave at the origin, we find strategies with a negative expected cost, indicating that the propagator model admits price manipulation, even if the necessary conditions for no price manipulation established earlier in Gatheral (2010) are satisfied.

In order to eliminate negative cost solutions, we propose two ways of regularizing the model, one based on the addition of a spread cost and one based on a modification of the instantaneous impact function. In the latter case, the function becomes convex for sufficiently high trading rates. Both methods succeed in avoiding solutions with negative costs and obviously reflect features of real markets.

We then further extend our analysis by minimizing the cost functional with the extra constraint that all trades should have the same sign, so that, for example, selling is disallowed during execution of a buy metaorder. This case requires a derivative-free optimization method. Using a direct search method, namely the Generating Set Search (GSS) method, we find positive expected execution costs and sparse optimal strategies; in this case, it is optimal to trade with a few bursts at a high trading rate interspersed with long periods of no trading. Such strategies are found to substantially decrease the expected cost of execution relative to VWAP. Moreover, we argue that such strategies are both robust to model mis-specification and easy to implement in practice.

The paper is organized as follows. In section 2, we state the problem and explain why it is difficult to solve. In section 4.2, we present our results on the SQP and direct search brute force minimization of the cost function. Section 5 presents two proposed regularization methods. In section 6, we summarize and conclude.

2. The optimal execution problem and its solution

Once again, we assume the following model for the evolution of the underlying price \( S_t \) during the execution of a metaorder starting at time \( t = 0 \):

\[
S_t = S_0 + \int_0^t f(\dot{x}_s) G(t-s) \, ds + \int_0^t \sigma \, dW_s, \tag{2.1}
\]

where \( \dot{x} \) is the rate of trading, \( f \) is the instantaneous market impact function and \( G(t-s) \) is a decay kernel, \( \sigma \) is the volatility and \( W_t \) is a Wiener process. Thus, \( S_t \) follows an arithmetic random walk with whose drift depends on the accumulated impacts of previous trades. In discrete time, this is the propagator model originally developed by Bouchaud et al. (2009), Bouchaud et al. (2004); the above continuous time formulation (2.1) is due to Gatheral (2010). More recently, Bacry et al. (2014) have shown how a market impact model of the form (2.1) may be related to a more fundamental description of order flow using Hawkes processes.

The optimal execution problem consists in finding the trading strategy \( \Pi = \{x_t\}_{t \in [0,T]} \) that minimizes the execution cost for a given total quantity \( X \) of shares to be traded. The expected cost \( C[\Pi] \) associated with a given strategy \( \Pi \) is given by

\[
C[\Pi] = \mathbb{E} \left[ \int_0^T \dot{x}_t (S_t - S_0) \, dt \right] = \int_0^T \dot{x}_t \left( \int_0^t f(\dot{x}_s) G(t-s) \, ds \right) \, dt, \tag{2.2}
\]

and the constraint that all shares should be traded is

\[
\int_0^T \dot{x}_t \, dt = X. \tag{2.3}
\]

Expression (2.2) for the expected cost corresponds to expected implementation shortfall. We search for a statically optimal
strategy. A statically optimal strategy is also dynamically optimal when the cost depends on the stock price only through the term \( \int_0^T S_t \, dt \), with \( S_t \) a martingale (Predoiu et al. 2011). Thus, for the model of (2.1) and the cost function described by (2.2), a statically optimal strategy is also dynamically optimal.

The \( \hat{x}_t \, dt \) shares traded at time \( t \) are traded at an expected price
\[
S_t = S_0 + \int_0^t f (\hat{x}_t) \, G(t-s) \, ds,
\]
which represents the cumulative impact of prior trading up to time \( t \).

The propagator impact model of equation (2.1) is fully specified by the form of the functions \( f \) and \( G \). Assuming this model, a large body of empirical evidence supports specific functional forms for the functions \( f \) and \( G \) as follows. First, the instantaneous impact function \( f (\cdot) \) is strongly concave. For example, based on a large sample of NYSE stocks, Lillo et al. (2003) observed a concave function of the transaction volume. The concave function is well fitted by a power law function \( f (t) \). Indeed, this is one of the major assets of the model (2.1) under the constraint of equation (2.3) in the case of linear market impact, \( f (\hat{x}) \propto \text{sign} (\hat{x}) |\hat{x}|^\delta \), and a power law kernel, \( G(t-s) = (t-s)^{-\gamma} \), the following conditions
\[
\gamma + \delta \geq 1, \quad \gamma \geq \gamma^* = 2 - \log 3 \log 2 \simeq 0.415, (2.4)
\]
are necessary for the absence of price manipulations. In the following, we will refer to these conditions as the no dynamic arbitrage (or NDA) conditions, and we will only consider combinations of \( \delta \) and \( \gamma \) satisfying them. However, even if the NDA conditions are satisfied, there is no guarantee that the resulting model does not admit price manipulation. Later in this paper, we will show that it is indeed the case that the NDA conditions are not sufficient to preclude price manipulation.

2.1. The expected cost of a VWAP

By definition, an ideal VWAP (Volume Weighted Average Price) execution should participate equally in every trade, and assuming that market volume is roughly constant (or equivalently working in volume time), the trading rate \( v = \hat{x} \) is constant. In this case, again with \( G(t) = t^{-\gamma} \), the instantaneous market impact function factors out of the expected total cost (2.2) to give
\[
C_{\text{VWAP}} = \int_0^T \hat{x}_t \int_0^t f (\hat{x}_s) \, G(t-s) \, ds \, dt = v f (v) \int_0^T \int_0^t ds \, (t-s)^{-\gamma} \, dt = v f (v) \frac{T^{2-\gamma}}{(1-\gamma)(2-\gamma)}. (2.5)
\]
Given a large database of VWAP-like executions, equation (2.5) may be used in practice to estimate the instantaneous market impact function \( f \). Indeed, this is one of the major attractions of the propagator model to practitioners: \( f (v) \) is easily estimated on historical execution data, the resulting estimate reflecting the aggregate performance of the various components of the execution algorithm (Zarinelli et al. 2015).

2.2. The case of linear market impact

The optimization problem of minimizing the expected cost of equation (2.2) under the constraint of equation (2.3) in the case of linear impact, \( f (\hat{x}) \propto \hat{x} \), has been solved (Gatheral and Schied 2013). In what follows, we denote the rate of trading \( \hat{x}_t \) by \( v_t \). In particular, proposition 22.9 of Gatheral et al. (2012) states that if \( G \) is positive definite, then \( \hat{x}_t \) minimizes the expected cost if and only if there is a constant \( \lambda \) such that for all \( t \in [0, T] \), \( \hat{x}_t \) solves
\[
\int_0^T G(t-s) \, dx_s = \lambda. \quad (2.6)
\]
As an important example, relevant for this paper, is the case \( G(t-s) = (t-s)^{-\gamma} \) where the integral equation (2.6) becomes the Abel equation with solution...
where \( c \) is uniquely determined by the constraint equation (2.3) as
\[
 c = \frac{\Gamma \left( \frac{2}{1+\gamma} \right) \Gamma \left( \frac{1+\gamma}{2} \right) \Gamma \left( \frac{1+\gamma}{2} \right)}{\sqrt{\pi} \Gamma \left( \frac{1+\gamma}{2} \right)^2}.
\]

where \( \Gamma (\cdot) \) is Euler’s Gamma function. This solution is U-shaped and symmetric under time reversal, i.e. \( v_t = v_{T-t} \), \( t \in [0, T/2] \). In the following, we will refer to this solution as the GSS solution.

2.3. The general case of non-linear market impact

In the general non-linear case, the problem is mathematically much more complicated. A first step in this direction has been presented very recently by Dang (2014) and consists in two contributions.

Dang’s first contribution is to transform the cost minimization problem into an integral equation generalizing (2.6).

Specifically, given \( f \in C^1(\mathbb{R}) \) and \( G \in L^1[0, T] \), for the class of functions \( x \) on \([0, T]\) satisfying
\[
\begin{align*}
\textbullet & \quad x \text{ is absolutely continuous on } (0, T), \\
\textbullet & \quad f \circ v \in L^1[0, T],
\end{align*}
\]

the following necessary condition for the stationarity of the functional of equation (2.2) holds:
\[
\int_0^T f(v_s) G(t-s) \, ds + f'(v_t) \int_t^T v_s G(s-t) \, ds = \lambda, \quad \text{for all } t \in [0, T]
\]

(2.8)

where the trading rate \( v_t = \dot{x}_t \), and again \( \lambda \) is a constant set by the constraint equation (2.3).

When \( f(v) \propto \text{sign}(v) |v|^\delta \) with \( \delta > 1 \), equation (2.8) is well defined for all \( v \in \mathbb{R} \). In contrast, in the concave case, \( \delta < 1 \), equation (2.8) is not defined if the trading velocity \( v \) vanishes at some time \( t \) because the derivative of \( f \) diverges at zero. This observation restricts the class of trajectories that can be considered. Moreover, in the concave case, there is no guarantee that the necessary condition (2.8) is also sufficient because the cost minimization problem could have a large number of extremal points

Equation (2.8) is a weakly singular† Urysohn equation of the first kind (Polyanin and Manzhirov 2008) taking the form
\[
\int_0^T G(|t-s|) F(v_s, t) \, ds = \lambda
\]

(2.9)

where
\[
F(v_s, t) = \begin{cases} 
 f(v_s), & s \leq t \\
 v(s) f'(v_t), & s > t.
\end{cases}
\]

†A kernel is called singular if it becomes infinite at one or more points in the range of integration, such as in Abel’s equation (Wazwaz 2011). A kernel is called weakly singular if its singularity is integrable, i.e. the integral of the function on a range that contains the singularity is finite. In our case, the weakly singular kernel is given by \( G(|t-s|) = |t-s|^{-\gamma} \).

Figure 1. Solution of the Urysohn equation in the weak non-linear case with \( \gamma = 0.5, \delta = 1.02 \) and \( X = 0.1 \). We observe that this solution is not symmetric under time reversal. The dotted line represents the GSS solution, i.e. the solution valid for the linear impact case.

Note that there are two sources of non-linearity in the integral equation (2.9): the non-linear impact function \( f(v) \) and the function \( F \). Moreover, the term involving \( f'(v_t) \) entangles the response of price at time \( t \) with the future trading rates, i.e. \( v_s \) for \( s > t \), i.e. a coupling between present and future values of \( v \).

Dang’s second contribution (Dang 2014) is a numerical scheme to solve the integral equation (2.8) when \( f(v) \propto \text{sign}(v) |v|^\delta \). Dang’s method obviously always converges when \( \delta = 1 \) because in this linear case, the objective function has only one minimum. In addition, his method seems to converge when \( \delta > 1 \). In the weakly non-linear case where \( \delta \) is slightly greater than 1, the optimal strategy is not surprisingly similar to the GSS strategy, except that it is asymmetric, slightly back-loaded with a little less trading at the beginning of the trading period and a little more trading at the end; a typical such optimal strategy is shown in figure 1. However, Dang’s method appears not to converge when \( \delta < 1 \) (the interesting case) and this is no surprise because as we will see later, when \( \delta < 1 \), the objective function has multiple local minima. Equation (2.8) is a necessary but not in general sufficient condition for optimality. Thus, in general, we must resort to more direct numerical techniques.

3. Discretization of the optimization problem and its numerical solution

In order to find an approximate optimal execution strategy, we discretize the interval \([0, T]\) into \( N \) equal subintervals. We suppose that trades in the \( i \)th interval \([t_{i-1}, t_i]\) are executed at the constant rate \( v_i \). The cost function (2.2) is then discretized as
\[
C_N[\Pi] = \sum_{i=1}^N \sum_{j=1}^{i-1} v_j f(v_j) G_{i,j} + \frac{1}{2} \sum_{i=1}^N v_i f(v_i) G_{i,i},
\]

(3.1)

where, for \( j \leq i \),
\[
G_{i,j} = \int_{t_{i-1}}^{t_i} \int_{t_{j-1}}^{t_j} G(|t-s|) \, ds \, dt.
\]
Then, if \( G(\tau) = \tau^{-\gamma} \) and \( i > j \), we have
\[
G_{i,j} = \frac{1}{(1 - \gamma)(2 - \gamma)} \left( \frac{T}{N} \right)^{2 - \gamma} \left\{ (i - j + 1)^{2 - \gamma} - 2(i - j)^{2 - \gamma} + (i - j - 1)^{2 - \gamma} \right\}.
\]
The diagonal terms are given by
\[
G_{i,i} = \frac{2}{(1 - \gamma)(2 - \gamma)} \left( \frac{T}{N} \right)^{2 - \gamma}.
\]
The constraint (2.3) on the traded volume is then given by
\[
\sum_{i=1}^{N} v_i = NX. \tag{3.2}
\]
To find the optimal strategy, we must minimize the discretized cost function \( C_T(\Pi) \) in (3.1) with respect to the \( v_i \). When the instantaneous impact function \( f \) is linear \( (f(v) \propto v) \), the expected cost reduces to a \( N \)-dimensional quadratic form. In the general non-linear case, the optimal strategy corresponds to the global minimum of a complicated non-linear function of \( N \) variables.

Remark 3.1 The solution of this multi-dimensional optimization problem is a piecewise constant approximation of the optimal strategy corresponding to a sequence of interval VWAP executions with trading rates \( v_i \), and thus easy to implement in trading practice.

### 3.1. Numerical cost minimization

In the general non-linear case, the cost function (3.1) is not convex, so we perform a non-convex constrained optimization with respect to the \( N \) variables \( v_i \).

We make use of two incomplete methods, that is methods that are not guaranteed to reach the global minimum. The first method is based on the SQP algorithm. This is one of the most powerful methods for the numerical solution of constrained non-linear optimization problems (NLP). It is an iterative procedure which models the NLP for a given iterate by a quadratic programming (QP) subproblem, and then uses the solution to construct a new iterate. Convergence to a local minimum is then guaranteed. We use the routines implemented in Matlab (The MathWorks Inc. 1990–2012, The MathWorks Inc. 2004–2012a, 2004–2012b).

When we search for strategies with the constraint that all trades should have the same sign, an algorithm based on derivatives, like SQP, can fail on the hyperplanes of the state space defined by \( v_i = 0 \), where the derivatives of the cost function diverge. For this case, we employ a second method based on a direct search approach, the GSS algorithm (Kolda et al. 2003, 2006), implemented in Matlab. This method does not require the computation of the derivatives of the cost function because it searches directly directions of space where the cost decreases. When the search is close to the boundary of the feasible region, the set of search directions must include directions that conform to the geometry of the boundary. As would be expected, the GSS algorithm is more computationally expensive than the SQP algorithm.

We adopt a multiple random start approach, consisting of picking random starting points on the hyperplane defined by equation (3.2) (the constraint on the total quantity traded) and performing local SQP and direct search optimizations starting from these points. We then study the difference between the various extreme points, investigating whether they are local minima, and then among these extrema, we select the solution with the smallest cost.

### 4. A simple model with power law impact

In this section, we study the very following simple model (which we will call the \textit{power law model}) with instantaneous market impact function
\[
f(v) = \alpha \text{sign}(v)|v|^\delta \text{ with } \delta < 1, \tag{4.1}
\]
for some constant \( \alpha > 0 \), and decay kernel of the form
\[
G(\tau) \propto \frac{1}{\tau^{\gamma}}. \tag{4.2}
\]

From equation (2.5), the expected cost of a VWAP execution in the power law model is given by
\[
C_{\text{VWAP}} = \alpha v^{1+\delta} \frac{T^{2-\gamma}}{(1 - \gamma)(2 - \gamma)}. \tag{4.3}
\]

As discussed earlier in Gatheral (2010) and in more detail below, the simple power law form (4.1) of \( f(\cdot) \) is unrealistic. However, the following special case of the model suggests it is nevertheless worthy of further study.

#### 4.1. The square root model

Consider the following special case of (4.1) and (4.2) with
\[
f(v) = \frac{3}{4} \sigma \sqrt{\frac{v}{V}}
\]
where \( \sigma \) denotes (daily) volatility, \( V \) denotes average (daily) volume and
\[
G(\tau) = \frac{1}{\sqrt{\tau}}.
\]

In this case, the stock price process (1.1) becomes
\[
S_t = S_0 + \frac{3}{4} \sigma \int_0^t \frac{v_s}{V} \frac{ds}{\sqrt{t-s}} + \int_0^t \sigma \, dW_s, \tag{4.3}
\]
which we will call the \textit{square root process}.

It follows from equation (2.5) that under the square root process (4.3), the expected cost of a VWAP execution is given by the celebrated square root formula for market impact:
\[
C_{\text{VWAP}} = \sigma \sqrt{\frac{X}{V}}. \tag{4.4}
\]

As noted in Gatheral (2010), the square root formula (4.4) has been widely used in practice for many years to generate a pre-trade estimate of transaction costs. And various studies of market impact costs have found the square root formula to fit transaction costs’ data remarkably well (see Tóth \textit{et al.} 2011 in particular). Of course, this by no means implies that the square root process (4.3) is the true underlying process; see, for example, Donier \textit{et al.} (2015) for an alternative process.
4.2. A two-step approximation

Consider a two-step discretization of the power law model where the metaorder execution consists of two interval VWAPs. Setting $\alpha = 1$ and $T = 1$, the expected liquidation cost (3.1) becomes

$$C[v_1, v_2] = \frac{1}{(1 - \gamma)(2 - \gamma)} \left( \frac{1}{2} \right)^{2 - \gamma} \times \left[ v_1 f(v_1) + v_2 f(v_2) + \left(2^{2 - \gamma} - 2\right)v_2 f(v_1) \right].$$

(4.5)

When $f(v) \propto v$, (4.5) reduces to the formula for a paraboloid in two dimensions. Imposing the constraint $v_1 + v_2 = 2X$, the problem reduces to a minimization with respect to $v_1$.

In figure 2, with $\gamma = 0.5$, $X = 0.1$, we plot $C[v_1, v_2]$ against $v_1$ for various values of $\delta$. We observe that for $\delta < 1$, the cost function has two local minima, one for a positive value of $v_1$ and one for a negative value of $v_1$. When $\delta \gtrsim 0.56$, the global minimum is the one with $v_1 > 0$, while for $\delta \lesssim 0.56$, the global minimum is attained for a negative value of $v_1$. It follows that if the impact function is strongly non-linear, we can decrease the expected cost of a buy programme with an intermediate sell trade, so there is transaction-triggered price manipulation. Note in particular that if we impose $v_1 \geq 0$, we have a boundary solution; it is then optimal not to trade in the first interval, trading the whole order in the second interval.

As we will see in the next section, this effect is accentuated as the number $N$ of trading periods is increased.

4.3. The $N$-step unconstrained case

We first consider the power law model with $\gamma = \delta = 0.5$, $\gamma = 0.45$, $\delta = 0.55$ with no constraint on the trading rate in each of $N = 100$ intervals. We use 1000 starting points distributed uniformly on the hyperplane $\sum_{i=1}^{N} v_i = \text{constant}$ and as discussed in section 3.1, employ the SQP algorithm to find the minimal cost strategy.

In figure 3, we plot the optimal trading profile for a buy programme corresponding to the global minimum of the cost function using our numerical minimization procedure. We observe that the optimal quantity to trade in each subinterval varies very irregularly with time. The cost minimizing solution consists of a series of bursts of intense but short-lived buying, separated by long periods when it is optimal to sell slowly. Evidently, the optimal solution admits transaction-triggered manipulation, as already shown in the $N = 2$ case analysed above in section 4.2.

Strategies that are close-to-optimal are qualitatively similar, but the positions of their spikes can be very different. As an example, in figure 4, we plot the four lowest cost solutions when $\gamma = 0.5$ and $\delta = 0.5$. All are characterized by a few intense positive spikes, separated by periods of slow selling. There is typically a multiplicity of strategies, with bursts at different times, described by similar expected costs. But these solutions differ in the positions of the bursts of buying. In order to compare solutions, taking into account the time distribution of the spikes, we have computed the Euclidean distance of the cumulative volume of trades. In figure 5, we plot six local optimal solutions, ordered according to their costs from the lowest to the highest, against a VWAP strategy for reference. The Euclidean distance confirms the visual intuition, i.e. two SQP solutions can differ between each other more than an SQP solution against a VWAP. However, the intermittent nature of any local SQP strategy allows cost savings respect to a classic VWAP strategy.

In figure 6, we plot the minimal expected cost for different values of $\gamma$ as a function of $\delta$ with $N = 100$ (top panel) and $N = 150$ (bottom panel). Each combination of $\delta$ and $\gamma$ satisfies the dynamic-no-arbitrage constraints. We observe that the minimal expected cost is not a monotonic function of $\delta$. More significantly, for each $\gamma \leq 0.50$, we find a regime of $\delta$ for which the expected cost of the optimal strategy is negative; this effect becomes stronger as we increase the discretization. Thus, our simple power law market impact model not only admits transaction-triggered price manipulation, but worse still exhibits price manipulation in the sense of Huberman and Stanzl; it is clearly mis-specified since it allows arbitrage opportunities.

4.3.1. Characterization of the cost landscape. In our unconstrained SQP minimization, we find a large number of extremal strategies that are very similar in terms of expected cost. As expected, the standard deviation of the cost associated with these local minima decreases as the instantaneous market impact function $f$ becomes more linear. Similarly, the Euclidean distance between local minima decreases as $f$ becomes more linear. A natural question that then arises is whether the cost landscape is sloppy or rugged.

A sloppy (Brown and Sethna 2003, Waterfall et al. 2006) landscape is one in which where there are directions in the $N$-dimensional space of strategies $\{v_i, i = 1, \ldots, N\}$ along which the expected cost varies very little. A well-known example of such a behaviour is given by the Rosenbrock function, in which the global minimum is located inside a long, narrow, parabolic-shaped valley. If the cost landscape is sloppy, it is
Figure 3. Optimal solution given by the SQP algorithm for a buy programme where $X = 0.1$, corresponding to 10% of market volume. We report the share quantities $\Delta x_i = v_i T/N$ to be traded in each time interval $[t_{i-1}, t_i)$.

Figure 4. The four lowest cost solutions given by the SQP algorithm for a buy programme where $X = 0.1$, corresponding to 10% of the market volume, for $\gamma = 0.5$, $\delta = 0.5$. We report the share quantities $\Delta x_i = v_i T/N$ to be traded in each time interval $[t_{i-1}, t_i)$. The costs are reported in the insets.

typically difficult to find the global minimum because there is a manifold where the cost function is almost flat.

Alternatively, the cost landscape may be **rugged**, with many local minima separated by local peaks. Rugged landscapes have attracted attention in physics (Weinberger 1991), evolutionary biology (Weinberger 1990) and computer science. In order to discriminate between these two alternatives and to characterize the cost landscape, we performed two types of analyses.

First, we performed a second derivative test on the local minima found by the SQP algorithm so as to ascertain which of these are actually minima and not, for example, saddle points. In the case of constrained optimization, the second-order sufficient condition for a minimum may be expressed in terms of the determinant of the so-called **bordered Hessian** (Chiang 1974, Magnus and Neudecker 1995). With $N = 100$ and with parameters satisfying the no-dynamic-arbitrage conditions, we found more than 95% of the extremal points to be actual minima. The remaining points were saddle points.

Second, we directly tested the hypothesis that the landscape is sloppy, using the eigenvalues and eigenvectors of the bordered Hessian of the cost function to identify its **stiff** and **sloppy** directions. One can study the sensitivity of the cost function to changes by the eigenvalue spectra of the bordered Hessian computed at a local minimum. Sloppy models are characterized by a constant logarithmic density for eigenvalues over six or more order of magnitude (Brown and Sethna 2003, Waterfall et al. 2006). The sensitivity of the cost function to changes is given by the square root of the eigenvalue. For sloppy models, this means that one should move along the sloppiest eigendirection a thousand times more than along the stiffest eigendirection in order to change the function by the same amount. We computed eigenvalue spectra of the bordered Hessian of the Lagrangian function in the case $\delta < 1$ finding that the spectra are not compatible with a sloppy landscape. The bulk of the eigenvalues has similar small values of the order of $10^{-2}$, implying that the region near a local minimum is not flat. In summary, the above analyses indicate that the cost landscape
We use 1000 starting points for each optimization. We consider a buy programme where $X = 0.1$, corresponding to 10% of the market volume, for $\gamma = 0.5$, $\delta = 0.5$. The costs are reported in the insets. The SQP solutions can differ between each other but ensure cost savings respect to a classic VWAP strategy.

In this section, we consider $\delta < 1$ is rugged rather than sloppy.

### 4.4. N-step monotone strategies

In this section, we consider monotone strategies, where we impose the additional constraint $v_i \geq 0$, $i = 1, \ldots, N$ (for a buy programme). This additional constraint is analogous to the no-short-selling constraint in portfolio optimization (Brodie et al. 2009).

We use a direct search method, specifically the GSS algorithm (Kolda et al. 2003). The computed optimal trading profiles are shown in figure 7 for the case $\gamma = 0.5$; results in the case $\gamma = 0.45$ are similar. In the linear case $\delta = 1$, we recover a discrete approximation to the closed-form solution of Gatheral and Schied (2013). The main feature that we observe is that as $\delta$ decreases, and the instantaneous market impact function $f(\cdot)$ becomes more non-linear, the optimal trading profile becomes more and more sparse. The optimal strategy then consists of a few bursts of buying interspersed with long periods of no trading. The geometrical interpretation of this result is that the solution lies on the boundary of a $(N - 1)$-simplex. This is consistent with our unconstrained SQP results, where optimal strategies lie just beyond the edge of the simplex. The direct search algorithm simply stops at the edge of the feasible region described by the simplex. Finally, notice that with the monotonicity constraint, it is optimal to start trading in the first interval, while without this constraint, it is optimal to slowly push down the price by selling before trading the first burst.

Obviously, when trading rates are constrained to be non-negative, the expected cost is positive. In table 1, we report the expected costs for optimal strategies in the unconstrained and monotone cases, as well as the expected cost of a simple VWAP strategy. We observe that the expected costs in the unconstrained case are sometimes close to zero and even negative, which is of course both consistent with the existence of price manipulation and highly unrealistic. Expected costs with the monotonicity constraint are typically much higher than in the unconstrained case, but can still be significantly lower than VWAP, depending on the value of $\delta$. Note also that in the linear case $\delta = 1$, the saving relative to VWAP from pursuing the optimal GSS strategy (2.7) is negligible.

**Remark 4.1** Although it would be preferable to have a model in which optimal strategies are monotone, rather than impose monotonicity as an explicit constraint, it can be argued that monotone strategies such as those plotted in figure 7 and their expected costs are robust to model mis-specification. For the expected cost savings relative to VWAP to be achievable in practice, the only requirement is that the estimated functions $f$ and $G$ reasonably reflect the characteristics of actual execution data for the range of trading rates in the optimized strategy.

### 5. Potential regularizations of the power law model

In the previous sections, we established that an instantaneous market impact function of the form $f(v) \propto |v|^\delta$ with $\delta < 1$ always gives rise to transaction-triggered price manipulation, and sometimes even to price manipulation proper. In order to have a reasonable model that is useful in practice, we clearly need either an alternative functional form for $f$ or to regularize the model in some other way. We now investigate two different approaches, both of which reflect important features of the market that we have so far neglected. In section 5.1, we modify the shape of the instantaneous market impact function $f$ for high trading rates; the resulting market impact function is then
optimal execution

Table 1. Expected costs of unconstrained and monotone strategies for $\gamma = 0.45$ and $\gamma = 0.5$. The numbers in boldface indicate strategies achieving the lowest expected cost. The difference between expected costs increases as $f$ becomes more non-linear, as does the saving relative to VWAP. We use a discretization of $N = 100$ subintervals and both the SQP and GSS optimizations use 1000 starting points.

| $\delta$ | Unconstrained | Monotone | VWAP | Unconstrained | Monotone | VWAP |
|----------|---------------|----------|------|---------------|----------|------|
| $\gamma = 0.45$ | 0.0115 | 0.0115 | 0.0117 | 0.0131 | 0.0147 | 0.0131 |
| $\gamma = 0.5$  | 0.0136 | 0.0148 | 0.0148 | 0.0158 | 0.0162 | 0.0168 |
| $\gamma = 0.55$ | 0.0112 | 0.0166 | 0.0166 | 0.0176 | 0.0176 | 0.0188 |
| $\gamma = 0.6$  | 0.0132 | 0.0209 | 0.0170 | 0.0202 | 0.0202 | 0.0211 |
| $\gamma = 0.65$ | 0.0117 | 0.0234 | 0.0169 | 0.0220 | 0.0220 | 0.0237 |
| $\gamma = 0.7$  | 0.0117 | 0.0234 | 0.0169 | 0.0220 | 0.0220 | 0.0237 |
| $\gamma = 0.75$ | 0.0092 | 0.0263 | 0.0146 | 0.0238 | 0.0238 | 0.0298 |
| $\gamma = 0.8$  | 0.0047 | 0.0295 | 0.0120 | 0.0245 | 0.0245 | 0.0335 |
| $\gamma = 0.85$ | 0.0029 | 0.0331 | 0.0075 | 0.0262 | 0.0262 | 0.0376 |
| $\gamma = 0.9$  | 0.0047 | 0.0331 | 0.0075 | 0.0262 | 0.0262 | 0.0376 |
| $\gamma = 0.95$ | 0.0047 | 0.0331 | 0.0075 | 0.0262 | 0.0262 | 0.0376 |

Figure 7. Optimal solutions given by the direct search algorithm for a monotone buy programme where $X = 0.1$, corresponding to 10% of market volume, for $\gamma = 0.5$. We report the share quantities $\Delta x_i = v_i T / N$ to be traded in each time interval $[t_{i-1}, t_i)$.

5.1. Concave–convex impact

As previously argued in Gatheral (2010), at some point, the rate of trading is so high that one trades deeply into the order book, where less liquidity is available, and at this point, one can expect that $f(\cdot)$ to become convex. Specifically, it was argued that for very high trading rates, the market impact function should take the form

$$f(v) \sim \frac{1}{(v - v_{\text{max}})^{1/\mu}}$$

where $\mu$ is the tail exponent† of the empirical power law distribution of limit orders as a function of distance to the best quote. It then makes sense to express instantaneous impact in terms of trading rate as a percentage of market volume $V$ per unit time, specifically in terms of the dimensionless participation rate $\rho(v) = |v|/(|v| + V)$ which is the trading rate $v$ as a fraction of total volume. Since one can trade at most the entire market volume, $\rho(v) \leq 1$.

Choosing $\mu = 1/2$ for simplicity, we thus postulate the following concave–convex impact function designed to both penalize excessively high trading rates and mitigate the problem of negative expected liquidation costs:

$$f_G(v) = c \text{sign}(v) \left\{ \rho(v)^{\delta} + d \frac{\rho(v)}{(1 - \rho(v))^2} \right\}$$

$$= c \text{sign}(v) \left\{ \left( \frac{|v|}{|v| + V} \right)^{\delta} + d \frac{|v|(|v| + V)}{V^2} \right\}$$

(5.1)

where $c$ and $d$ are positive constants. The constant $d$ is a measure of the magnitude of the convex term with respect to the concave term. The parameter $d$ also determines the participation rate $\rho^* = \rho(v^*)$ where the function’s convexity changes. Specifically, a well-known trader’s rule of thumb says that maximum participation rate at which it is reasonable to

†Bouchaud et al. (2002) estimate $\mu = 0.6$. 

concave–convex.
trade is around 20%. From equation (5.1), assuming \( v > 0 \), we have that \( f_G'(\rho) = 0 \) when

\[
d = (1 - \delta) \delta \frac{\rho^{\delta-2}}{(1-\rho)^{3}}.
\]

Substituting the participation rate \( \rho = 0.2 \) with \( \delta = 0.50 \) gives \( d \approx 0.26 \). Decreasing the critical participation rate naturally increases \( d \); for example, when \( \rho = 0.15 \) and \( \delta = 0.50 \), \( d \approx 0.52 \). Thus, in practice, we expect the dimensionless constant \( d \) to be of order one.

We now illustrate the results of the numerical minimization of expected cost in the case \( \gamma = 0.45, \delta = 0.55, N = 100 \) for four different values of \( d: 0.1, 0.5, 1, 2 \), using 1000 starting points for each optimization. In order to calibrate the other parameters to realistic values, we notice that for \( 0 < \delta < 1 \), \( f_G(v) = c \text{sign}(v) (\frac{v}{\sqrt{v}})^\delta + O(v) \) as \( v \to 0 \) and \( f_G(v) \sim c \text{sign}(v) d \) as \( v \to \infty \). Comparing the first limit with the expression of the impact for the square root process of section 4.1, we choose \( c = 3/4 \sigma \), where \( \sigma \) is the daily volatility. Moreover, without loss of generality, we set the daily volume at \( V = 1 \) and we consider the case \( X = 0.1 \), corresponding to a trade equal to 10% of daily market volume. Finally, in order to measure costs in relative terms, we normalize the impact cost by the net notional traded \( X S_0 \), where \( S_0 \) is the stock’s initial price. For the numerical computations, we choose \( S_0 = 25 \) and \( \sigma = 0.04 \).

In figure 8, we plot the optimal strategies for the four different choices of \( d \). We observe that an increase in the magnitude of the convex term causes a decrease in maximum trading rates and an increase in the number of periods in which buying is optimal. As expected, the convex component of the market impact function acts as a disincentive to trade at high participation rates.

In table 2, we compare the expected cost of the optimal unconstrained strategy estimated using the SQP algorithm to the cost of the corresponding VWAP strategy. Costs are in basis points of the notional traded.

As the magnitude of the convex term increases, we no longer find negative expected liquidation costs. It is worth noticing that the value of \( \rho^* \) is close to the half of the mean positive participation rate in each case. Thus, just knowing the functional form of \( f \), one has a rough estimate of the optimal positive participation rate.

Finally, in figure 9, we plot the expected cost for different values of \( \gamma \) as a function of \( \delta \), in the cases \( d = 0.1 \) and \( d = 1 \). In the case \( d = 0.1 \), we observe a region of parameter space where the cost is negative, while if \( d \) is sufficiently large, we observe that it is possible to eliminate price manipulation for all combinations of \( \gamma \) and \( \delta \) satisfying the NDA constraints. Further investigation shows that this result does not change as we increase \( N \) (data not shown).

5.2. Adding a spread cost

In this section, following Busseti and Lillo (2012), we add a spread cost to the model of equation (2.1) to penalize wrong way trading. Equation (2.1) could be regarded as describing the evolution of the mid-price. When a market order is executed,† there is an extra cost of half of the bid–ask spread \( 2 \delta S \), and the trading price is given by

\[
S_t = S_0 + \int_0^t \frac{1}{\sqrt{s}} \frac{1}{\sqrt{\delta s}} G(t-s) \, ds \\
+ \int_0^t \rho \, dW_s + \delta S \int_0^t \delta (s-t) \text{sign}(\dot{\xi}_s) \, ds, \tag{5.2}
\]

The spread term is a temporary impact term that can be described by a \( \delta \)-impact function; it affects only the price at which

†Even if execution is with limit orders, one finds in practice that due to adverse selection, there is an extra cost of a portion of the bid–ask spread.
Table 2. Fixing \( \gamma = 0.45, \delta = 0.55 \), for various values of \( d \) (first column), we report the value of \( \rho^* \) for which \( f_G^d(\rho^*) = 0 \). The next columns report the mean and standard deviation of the positive SQP optimal participation rates \( \rho \) with \( N = 100 \). The last two columns compare the expected costs in basis points of the notional traded of SQP and VWAP executions, respectively.

| \( d \) | \( \rho^* \) | \( \langle \rho_{SQP} > 0 \rangle \) | \( \sigma (\rho_{SQP} > 0) \) | Cost SQP (bp) | Cost VWAP (bp) |
|-------|-----------|----------------|-----------------|-------------|-------------|
| 0.1   | 0.2729    | 0.5199         | 0.1078          | -0.294      | 3.92        |
| 0.5   | 0.1459    | 0.3232         | 0.0469          | 2.01        | 4.54        |
| 1     | 0.1045    | 0.2432         | 0.0264          | 3.46        | 5.31        |
| 2     | 0.0722    | 0.1779         | 0.0198          | 5.70        | 6.86        |

![SQP optimization, \( N = 100, d = 0.1 \)](image)

**Figure 9.** Expected cost of the optimal solution given by the SQP algorithm for a buy programme, where \( X = 0.1 \), corresponding to 10% of market volume and with \( N = 100 \), and a concave–convex impact function \( f_G^d(v) \) with \( d = 0.1 \) (top) and \( d = 1 \) (bottom). We use 1,000 starting points for each optimization and the cost is expressed in basis points of the net notional traded. With \( d = 0.1 \), negative expected costs are observed for all values of \( \gamma \), while with \( d = 1 \), we observe no negative expected costs.

Optimal execution

This term can represent also any cost or fee proportional to the absolute volume executed. The expected execution cost is then given by

\[
C[\Pi] = C_G[\Pi] + C_S[\Pi] = \int_0^T \dot{x}_t \int_0^t f(\dot{x}_s) G(t-s) ds \, dt + \delta_S \int_0^T |\dot{x}_t| \, dt. \tag{5.3}
\]

The spread term thus penalizes any strategy that consists of both buy and sell trades; it achieves its minimum \( C_S[\Pi] = \delta_S X \) for monotone strategies where all trades have the same sign. This penalty is a form of \( L_1 \) or LASSO regularization widely used in computer science and used for example by Brodie et al. (2009) to penalize short positions in Markowitz portfolio optimization. Busseti and Lillo (2012) show that in typical real-life VWAP executions, the spread cost and the impact cost are typically of the same order of magnitude.

In order to parameterize the relative cost of spread and impact, we define the dimensionless quantity \( r = C_S/VWAP \) or \( C_G/VWAP \), which is the ratio between the spread cost and the impact cost for a VWAP execution according to the model of equation (5.2). This quantity is given by

\[
\delta_S = r f(X) \frac{T^{1-\gamma}}{(1-\gamma)(2-\gamma)}. \tag{5.4}
\]

We perform a numerical optimization of the discretized cost function

\[
C_N[v] = N \sum_{i=1}^{N} \sum_{j=1}^{N} v_i f(v_j) G_{i,j} + \frac{1}{2} \sum_{i=1}^{N} v_i f(v_i) G_{i,i} + \delta_S \left( N \sum_{i=1}^{N} |v_i| \right), \tag{5.5}
\]

with the same parameters as in section 4.2, \( T = 1, N = 100, X = 0.1 \), with \( \gamma = 0.45 \) and \( \delta = 0.55 \). Moreover, as before, we calibrate the other parameters at \( c = 3/4 \sigma \) (with \( \sigma = 0.04 \)) and \( S_0 = 255 \). We consider two values of the spread: the first is \( \delta_S = 0.005 \), corresponding to a spread of one cent, as in most low price/large tick stocks. The second is \( \delta_S = 0.001 \), a case where the spread is vanishingly small, which is useful to see how the contributions to the total costs change in this regime. These two cases correspond to \( r = 0.5 \) and \( r = 0.1 \), respectively. Finally, as before, we express the costs in basis points of the notional traded, \( S_0 X \).

Figure 10 plots the strategies corresponding to the global minimum in the numerical optimization in the high and low spread cost cases. Despite the fact that transaction-triggered price manipulation is still observed, in both cases, the expected execution cost is now positive. When the spread is large \( (r = 0.5) \), the total cost is 3.12 bp, which is the sum of 0.46 bp of impact cost and 2.66 bp of spread cost. Note that the total absolute volume traded is 0.13 which is slightly larger than the net volume traded \( X = 0.1 \). The reason is that the relatively large spread costs are a disincentive to wrong way trading. On the contrary, when the spread is small \( (r = 0.1) \), the total cost is still positive, 0.71 bp, but the total absolute volume traded is more than twice of \( X (0.26) \). Moreover, the spread cost is 1.05 bp, but the impact cost is negative \((-0.34) \). As before, in both cases, the optimized strategies substantially outperform VWAP, which has \( [\frac{3/4\sigma X}{(1-\gamma)(2-\gamma)}] S_0 X = 3.97 \) bp of impact.
The mechanism for this in the LLOB model can be understood from the share quantities $\Delta x_i = v_i T N$ to be traded in each time interval $[t_{i-1}, t_i)$ with $\gamma = 0.45$ and $\delta = 0.55$. The high spread cost $r = 0.5$ case is on the left, and on the right is the low spread cost case with $r = 0.1$. The expected execution costs are 3.12 bp when $r = 0.5$ and 0.71 bp when $r = 0.1$.

In conclusion, a high spread cost acts as a strong disincentive to wrong way trading analogous to the effect observed by Brodie et al. (2009) for Markowitz portfolio weights. In that case, a penalization term proportional to the sum of absolute values of weights leads to optimal solutions where the resulting portfolio is sparse with few assets and no short positions. Similarly, in the power law model with a high spread cost, optimal solutions for a buy programme are characterized by a few bursts of trading separated by intervals of weak wrong way trading. As we would expect, a high spread cost leads to liquidation costs similar to those corresponding to monotone strategies.

5.3. Comparison with the LLOB model

One of the attractive features of the LLOB model of Donier et al. (2015) is that price manipulation is formally precluded. The mechanism for this in the LLOB model can be understood as follows. Without loss of generality, consider a buy metaorder. As execution proceeds, the slope of the order book on the ask side steepens relative to the slope on the bid side. Consequently, if the trade is reversed, the resulting sell metaorder causes higher price impact. In the context of our propagator model (1.1), this is as if the instantaneous market impact function $f$ were to depend on the order book flow.

Recall from section 2.1 that in our propagator model, the expected cost of a VWAP order is given by

$$C_{\text{VWAP}} = v f(v) \frac{T^{2-\gamma}}{(1-\gamma)(2-\gamma)}.$$  

In practice, most algorithmic metaorder executions can be reasonably well approximated by a VWAP, $f(v)$ is then easy to estimate from historical metaorder execution data for which $T$ is known. This ability to estimate $f$ easily from historical metaorder data is a feature of the propagator model which would be lost if $f$ were to depend on the state of the market or the state of the order book. However, if we were to make $f$ explicitly depend on order flow history, we would be able to mitigate price manipulation or even eliminate it altogether.

Conversely, market impact depends on order flow history in the LLOB model. To calibrate the LLOB model to historical metaorder data, assuming the initial state of the order book and the history of order flow are not known, we would need to make extra assumptions to compute market impact. It is thus unclear how to apply the model in practice to pre-trade estimation of execution costs or to the computation of optimal strategies, especially given that the linear approximation likely breaks down at extreme trading rates.

6. Conclusions

In this paper, we have addressed the problem of finding optimal execution strategies in the propagator model with non-linear transient impact of Gatheral (2010). We considered several different approaches, discussing their ranges of validity and the characteristics of the solutions they generate, and compared their associated expected costs. We also focused our attention on the presence of different forms of possible price manipulation.

From our analysis using an SQP global optimization technique, we found that the discretized cost function exhibits a rugged landscape, with many local minima separated by peaks. The global minimum (as well as other many other suboptimal minima) corresponds to a strategy that alternates buys and sells, and thus admits transaction-triggered price manipulation. In particular, for a buy programme, the optimal solution consists of short bursts of trading at high rates separated by long periods of selling at low trading rates. Even worse, when the non-linearity is strong and when the discretization is sufficiently fine, we find negative execution costs. Coupled with the transient property of market impact in our model, this implies that the non-linear transient impact model with purely power law concave impact admits price manipulation, eliminating such a
model from consideration for practical use. In particular, this shows that the conditions (2.4) of Gatheral (2010) for no price manipulation are necessary but not sufficient.

We also studied optimal monotone strategies where selling (buying) is disallowed in a buy (sell) programme. This optimization is performed by a derivative-free direct search method, specifically the GSS algorithm. In this case, the expected execution cost is positive by construction and the corresponding optimal strategy is sparse. In other words, it is optimal to trade in a few intense bursts separated by long periods of no trading. We argue that though the simple power law market impact model is clearly mis-specified, such constrained monotone strategies are robust to model mis-specification; the savings relative to VWAP predicted by the model should be achievable in practice, provided the instantaneous market impact function $f$ and the decay kernel $G$ are estimated from and consistent with actual metaorder execution data.

Finally, we propose two ways to regularize the transient impact model, both of them reflecting important features of the market that are neglected in the simple version of the model; the objective in each case is to robustify the solution in a parsimonious and natural way. In the first approach, we add to the market impact function a convex component which makes it very costly to trade at high trading rates. We observe that if the convex component is sufficiently large, negative expected execution costs are eliminated. In the second approach, we add a spread cost, effectively performing a regularization, which discourages wrong way trading. When transactions costs are sufficiently high, the optimal execution strategy again no longer exhibits a negative expected cost, and the optimal strategy becomes sparse and bursty, with a few moments of strong trading separated by longer periods of zero or slightly wrong way trading, similar to the constrained monotone case.

Though the regularized versions of the power law model are more or less consistent with empirical observation, we have not shown whether or under what conditions they admit or preclude price manipulation. On the other hand, by inspection, the recent Locally Linear Order Book (LLOB) model of Donier et al. (2015) does preclude price manipulation. It remains to be seen however whether the LLOB model is consistent with observation and how easy it is to compute optimal strategies. A fully satisfactory and practical model of market impact, which is free of price manipulation and consistent with empirical observation, thus seems to be still lacking, leaving a fruitful field for future research.

Acknowledgements

We thank Dario Trevisan for useful discussions regarding calculus of variations. This research was partially supported by HSBC Holdings plc.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Abergel, F., Bouchaud, J.-P., Foucault, T., Lehalle, C.-A. and Rosenbaum, M., Market Microstructure: Confronting Many Viewpoints, 2012 (John Wiley & Sons: Ponsdow).

Alfonsi, A. and Schied, A., Capacitive measures for completely monotone kernels via singular control. SIAM J. Control Optim., 2013, 51(2), 1758–1780.

Alfonsi, A., Schied, A. and Slynko, A., Order book resilience, price manipulation, and the positive portfolio problem. SIAM J. Financ. Math., 2012, 3(1), 511–533.

Almgren, R.F., Optimal execution with nonlinear impact functions and trading-enhanced risk. Appl. Math. Finance, 2003, 10(1), 1–18.

Almgren, R. and Chriss, N., Value under liquidation. Risk, 1999, 12(12), 61–63.

Almgren, R. and Chriss, N., Optimal execution of portfolio transactions. J. Risk, 2001, 3, 5–40.

Bacry, E., Iuga, A., Lasnier, M. and Lehalle, C.-A., Market impacts and the life cycle of investors orders. Market Microstruct. Liquidity, 2014, 1, 1550009.

Bertsimas, D. and Lo, A.W., Optimal control of execution costs. J. Financ. Markets, 1998, 1(1), 1–50.

Bouchaud, J.P., Farmer, J.D. and Lillo, F., How markets slowly digest changes in supply and demand. In Handbook of Financial Markets: Dynamics and Evolution, edited by T. Hens and K. Reiner Schenk-Hoppe, pp. 57–156, 2009 (North Holland: Amsterdam).

Bouchaud, J.-P., Gefen, Y., Potters, M. and Wyart, M., Fluctuations and response in financial markets: The subtle nature of ‘random’ price changes. Quant. Finance, 2004, 4(2), 176–190.

Bouchaud, J.-P., Kockelkoren, J. and Potters, M., Random walks, liquidity molasses and critical response in financial markets. Quant. Finance, 2006, 6(2), 115–123.

Bouchaud, J.-P., Mézard, M., Potters, M., et al., Statistical properties of stock order books: Empirical results and models. Quant. Finance, 2002, 2(4), 251–256.

Brodie, J., Daubechies, I., De Mol, C., Giannone, D. and Loris, I., Sparse and stable Markowitz portfolios. Proc. Nat. Acad. Sci., 2009, 106(30), 12267–12272.

Brown, K.S. and Sethna, J.P., Statistical mechanical approaches to models with many poorly known parameters. Phys. Rev. E, 2003, 68(2), 021904.

Busseti, E. and Lillo, F., Calibration of optimal execution of financial transactions in the presence of transient market impact. J. Stat. Mech.: Theory Exp., 2012, 2012(09), P09010.

Chiang, A.C., Fundamental Methods of Mathematical Economics, 1974 (McGraw-Hill, Kogakusha: Tokyo).

Curato, G., Gatheral, J. and Lillo, F., Discrete homotopy analysis for optimal trading execution with nonlinear transient market impact. Commun. Nonlinear Sci. Numer. Simul., 2016, 39, 332–342.

Dang, N.-M., Optimal execution with transient impact, 2014. Available online at: http://ssrn.com/abstract=2183685.

Donier, J., Bonart, J., Mastromatteo, I. and Bouchaud, J.-P., A fully consistent, minimal model for non-linear market impact. In Handbook on Systemic Risk edited by J.-P. Fouque and J.A. Langsam, pp. 579–599, 2013 (Cambridge University Press: Barcelona).

Gatheral, J., Schied, A. and Slynko, A., Exponential resilience and decay of market impact. In Econophysics of Order-driven Markets, edited by F. Abergel, B.K. Chakrabarti, A. Chakraborti, and M. Mitra, pp. 225–236, 2011 (Springer: Milano).

Gatheral, J., Schied, A. and Slynko, A., Transient linear price impact and Fredholm integral equations. Math. Finance, 2012, 22(3), 445–474.
Huberman, G. and Stanzl, W., Price manipulation and quasi-arbitrage. *Econometrica*, 2004, 72(4), 1247–1275.

Kolda, T.G., Lewis, R.M. and Torczon, V., Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Rev.*, 2003, 45(3), 385–482.

Kolda, T.G., Lewis, R.M. and Torczon, V., Stationarity results for generating set search for linearly constrained optimization. *SIAM J. Optim.*, 2006, 17(4), 943–968.

Kyle, A.S., Continuous auctions and insider trading. *Econometrica*, 1985, 53(6), 1315–1335.

Lillo, F., Farmer, J.D. and Mantegna, R.N., Econophysics: Master curve for price-impact function. *Nature*, 2003, 421(6919), 129–130.

Magnus, J.R. and Neudecker, H., *Matrix Differential Calculus with Applications in Statistics and Econometrics*, 1995 (John Wiley & Sons: Chichester).

Obizhaeva, A.A. and Wang, J., Optimal trading strategy and supply/demand dynamics. *J. Financ. Markets*, 2013, 16(1), 1–32.

Polyanin, A.D. and Manzhirov, A.V., *Handbook of Integral Equations*, 2008 (Chapman & Hall/CRC: Boca Raton).

Predoiu, S., Shaikhet, G. and Shreve, S., Optimal execution in a general one-sided limit-order book. *SIAM J. Financ. Math.*, 2011, 2(1), 183–212.

The MathWorks Inc., *Optimization toolbox user’s guide*, 1990–2012.

The MathWorks Inc., *Global optimization toolbox user’s guide*, 2004–2012a.

The MathWorks Inc., *Parallel computing toolbox user’s guide*, 2004–2012b.

Tibshirani, R., Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. Ser. B (Methodological)*, 1996, 58(1), 267–288.

Tóth, B., Lemperiere, Y., Deremble, C., De Lataillade, J., Kockelkoren, J. and Bouchaud, J.-P., Anomalous price impact and the critical nature of liquidity in financial markets. *Phys. Rev. X*, 2011, 1(2), 021006.

Waterfall, J.J., Casey, F.P., Gutenkunst, R.N., Brown, K.S., Myers, C.R., Brouwer, P.W., Elser, V. and Sethna, J.P., Sloppy-model universality class and the Vandermonde matrix. *Phys. Rev. Lett.*, 2006, 97(15), 150601.

Wazwaz, A.-M., *Linear and Nonlinear Integral Equations*. Vol. 639, 2011 (Springer: Beijing).

Weinberger, E., Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biol. Cybern.*, 1990, 63(5), 325–336.

Weinberger, E.D., Local properties of Kauffman’s n–k model: A tunably rugged energy landscape. *Phys. Rev. A*, 1991, 44(10), 6399–6413.

Zarinelli, E., Treccani, M., Farmer, J.D. and Lillo, F., Beyond the square root: Evidence for logarithmic dependence of market impact on size and participation rate. *Market Microstruct. Liquidity*, 2015, 1, 1550004.