Optimal Trading of Algorithmic Orders in a Liquidity Fragmented Market Place

Miles Kumaresan * Nataša Krejić †

May 31, 2013

Abstract

An optimization model for the execution of algorithmic orders at multiple trading venues is herein proposed and analyzed. The optimal trajectory consists of both market and limit orders, and takes advantage of any price or liquidity improvement in a particular market. The complexity of a multi-market environment poses a bi-level nonlinear optimization problem. The lower-level problem admits a unique solution thus enabling the second order conditions to be satisfied under a set of reasonable assumptions. The model is computationally affordable and solvable using standard software packages.

The simulation results presented in the paper show the model's effectiveness using real trade data. From the outset, great effort was made to ensure that this was a challenging practical problem which also had a direct real world application.

To be able to estimate in realtime the probability of fill for tens of thousands of orders at multiple price levels in a liquidity fragmented market place and finally carry out an optimization procedure to find the most optimal order placement solution is a significant computational breakthrough.

*Algonetix LLP, Berkeley Square House, Berkeley Square, London W1J 6BD, United Kingdom, e-mail: miles@algonetix.com

†Department of Mathematics and Informatics, University of Novi Sad, Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia. This author was supported by Ministry of Education and Science, Republic of Serbia (Grant 174030). e-mail: natasak@uns.ac.rs

Key words: nonlinear programming, optimal execution strategy, multiple trading venues, algorithmic trading, price impact, fill probability

AMS Subject Classification: 90C30, 90C90, 90B90.

1 Introduction

Algorithmic Trading, also known as Algorithmic Execution, is the automated process of trading exogenous orders in electronic (stock) exchanges. There are many aspects to algorithmic trading that make it attractive. Algorithmic trading consists of a whole range of standard algorithms to mimic mainstream execution styles such as VWAP (Volume Weighted Average Price), Participation (Volume Participation with VWAP as benchmark), Implementation Shortfall, and others. In addition there is a wide range of variations of these mainstream algorithms and an endless range of customized algorithms to suit each individual user's requirements.

When executing an order, a trader is faced with the option of either instantly transacting by paying a price premium or waiting for a better price. Both options come with a cost component. If one opted to use a passive limit order in order to wait for a better price, one is faced with volatility risk in the event the price drifting away without filling one's order. In this case, the trader will have to transact at an even worse price. Alternatively, the trader could issue a market order and instantly transact the desired quantity, however, at the cost of paying the spread or more in cost.

Additionally, a market order will cause a shock to the orderbook by removing liquidity. This shock, or market impact, can be divided into two categories, a) temporary and b) permanent. A temporary impact dissipates within a relatively short time, whereas a permanent impact will last long after the trade is completed. Therefore, when executing a large order, one should consider the effect one's action has on the orderbook, that will affect subsequent orders. The total cost of market orders therefore becomes the aggressive price across the spread plus the impact costs. Extensive analysis by Almgren [3] of a large number of trades shows that temporary impact is significantly larger than permanent impact but neither of them can be neglected in real trading. A number of different approaches to impact and the cost of trading caused by impact is present in the literature, for example [8, 9, 14, 15, 16, 20, 23, 24]. The problem of impact modeling and its corresponding costs are relevant in portfolio optimization models as well as in determining optimal execution strategies of large portfolios, [21, 13]. The role of market impact in optimal execution strategy is subject of intensive research as understanding the trade execution is a key issue for market practitioners. The optimal portfolio liquidation over a finite horizon in a limit order book with temporary impact is considered in [18]. The high frequency optimal execution is the subject of [2]. The nonlinear impact is incorporated in a trading strategy in [25] while shape function assumption is presented in [1]. A general framework for intraday trading based on the control of trading algorithms is considered in [10]. A model that explains how high frequency trading can be applied to supply liquidity and reduce execution cost is developed and solved in [26].

Limit orders do not have market impact but have a volatility risk. Intuitively, there has to be a middle ground where, the right combination of both order types should yield a more optimal price. The very ability to specify a price limit in limit orders gives rise to a new dilemma, i.e. choice of risk to take.

Historically, most stocks are listed on a single stock exchange. The emergence of alternative trading venues in recent times, ECNs in the US, MTFs in the UK and others, opens additional possibilities for trading as these venues provide a significant amount of liquidity. Due to the choice of multiple trading venues, it would intuitively seem less optimal to send an order to a specific venue for execution, if that would result in a fill price that is worse than if the order had been divided and sent to multiple venues. There is not much mathematical complexity involved in dividing and routing the aggressive component of an order to different venues. However, deciding where to place passive orders require complex mathematics and still remains an unresolved problem, at least in the public domain.

In the presence of multiple alternative trading venues, one is presented with not only the dilemma of a) what price at which to place the order in order to maximize the probability of getting filled while maximizing the gain, but also with the option of b) whether or not to break up the order and place it at multiple venues as well. In the latter case one is facing a stochastic problem in which order placement has to maximize fill rate while minimizing fill price. Intuitively, this requires some mathematical model to determine the fill properties of the different venues for different order quantities (the

Fill Probability Function) at various price levels. Probabilistic fill estimation models are proprietary models used by large market players (investment banks) to determine the fill probability curve. Even with these models, given that prognosis remains an inaccurate science, a further question arises when the actual fill rate is found to be lower than expected at a given venue. We know that an orderbook is a price and time ordered queue. Therefore, removing an order from one venue and placing it at another venue or even simply moving to a better price level within the same venue adds the disadvantage of being placed at the back of the queue, lowering the fill probability compared to it having been placed at that price level (and venue) from the outset.

Although multiple trading venues increase the problem of placing and executing orders, they are becoming an integral part of the trading environment. The existing state of affairs of modern implementation of optimization algorithms does not include, at least in academic publications, a solution for this kind of trading environment.

To be able to estimate in realtime the probability of fill for tens of thousands of orders at multiple price levels in a liquidity fragmented market place and finally carry out an optimization procedure to find the most optimal order placement solution is a significant computational breakthrough. Furthermore, great effort was made to ensure that this was a challenging practical problem which also had a direct real world application.

The problem we are interested in is determined by short time execution windows (measured in minutes) and quantities of up to 15% of the average traded volume within the considered time window. In other words we are mainly interested in execution of atomic orders. The principal aim of this paper is to propose an optimization procedure that yields an optimal trading trajectory for multiple trading venues applicable in live trading for a large universe of instruments. The proposed optimal trajectory consists of both types of orders, market and limit, and takes advantage of any temporary price or liquidity improvements available at a particular venue. Thus it provides a systematic way of employing both passive and aggressive trading strategies in order to minimize risk and maximize gain. Any optimization procedure meant to be deployed in a real trading environment must be computationally affordable and applicable in real time for potentially large portfolios of securities. Hence some simplifications are inevitable in the modelling process. The results obtained in this paper demonstrate that these simplifications did not interfere with the most important properties of the real situation. Although we used the logic used for a single market problem in [19], the proposed generalization is far from trivial. The complex multi-market environment yields a bilevel nonlinear optimization problem. The convexity of the lower level problem allow us to solve it exactly and thus obtain an affordable algorithm for generating an optimal trading strategy.

This paper is organized as follows. Section 2 contains necessary details concerning the trading process and market microstructure. The optimization model is developed in Section 3. Second order optimality conditions are established. Section 4 deals with the multi-period model with which we demonstrate how to perform re-optimization of the initially determined optimal trajectory. Numerical results obtained from simulation with real-trade data are presented in Section 5.

2 Preliminaries

Without any loss of generality we will assume that the trading is done on two markets, say A and B. Therefore all variables and functions that correspond to the markets A and B will be denoted with superscripts A and B respectively. If the same is true for both markets we will not use any subscript.

Until recently, the majority of companies were listed and traded at a single exchange only. In recent times however, the market place has rapidly become crowded with multiple exchanges where a given security (company) is listed and traded in multiple venues. As such, for a given security, each exchange will have an orderbook for that security. Each orderbook consists of a queue of buyers and another queue of sellers. Price and time of arrival determine the place in which a new arrival is inserted. Priority is given the ones who arrive first within a given price. Cancellations can take place without any restrictions. When there is a price overlap between the two queues, the intersection of all orders will be transacted.

A graphical representation of order books at two markets is shown at Figure 1. It can be seen that, the two exchanges or markets do not have the same order quantity or prices, although they are usually arbitrage free (where one could not simultaneously buy and sell at the two markets, securing an instant risk-less profit). It is however typical that one market can have a price improvement on its best bid and/or ask. Another point to be noted from

Figure 1: A typical orderbook in a multi-market environment.

the two orderbooks is that market A has significantly more orders (liquidity) than market B for the corresponding or comparable price levels. Prices in each market can only change in multiples of a defined minimum quantity called Tick Size. The spread, i.e. the difference between the the best bid and ask is usually a single tick size for liquid securities and several ticks for less liquid ones. Two of the main reasons why one market can have better price bids and/or asks than another market is due to different granularity in tick size and amount of liquidity in those markets.

For a given security in a two venue environment, we will have two sets of market conditions $\mathcal{M}^A, \mathcal{M}^B$. Each would have their respective queue of buying price levels $b_i^A(t), b_i^B(t)$, and selling price levels $a_i^A(t), a_i^B(t)$ for any time t. Time dependence will be dropped occasionally if no confusion is implied.

The price difference between the highest bidder and lowest seller is called bid/ask spread or just spread, defined as

$$\varepsilon^A = a_1^A - b_1^A,\ \varepsilon^B = a_1^B - b_1^B$$

for the two orderbooks in markets A and B. Because two venues could work with different price granularity and tick size, the spreads in the two markets most often differ. In any two liquid securities in liquid venues, the spread is expected to be similar. Spread can differ mostly due to tick size differences. Due to the efficiency of the market, risk free arbitrage is a rare event, an opportunity where one could buy at one venue and sell at a higher price at another venue simultaneously. Although completely independent, both markets track each other very closely - hence their volatility is also near identical.

Most securities have a liquidity pattern associated with the time of the day. However, the ratio of liquidity in one market versus another is not constant. At times, there can be disproportionately larger liquidity in the smaller venue. This excess liquidity could last for an extended period. Since this is a seemingly unpredictable process, market participants would gain by moving their orders from the queue in one market to the queue on another in order to maximize the probability of being filled - even if it meant joining at the back of the queue at the same price level.

In this paper we will assume that all prices follow an arithmetic random walk without drift. Denoting by P the mid-price, $P = (a_1 + b_1)/2$, we assume that

$$P(t) = P(0) + \sigma\sqrt{t}\zeta, \tag{1}$$

and consequently,

$$b_i(t) = b_i(0) - \frac{\varepsilon}{2} \tag{2}$$

$$a_i(t) = a_i(0) + \frac{\varepsilon}{2}, \tag{3}$$

where volatility is denoted by σ and the noise is Gaussian, $\zeta : \mathcal{N}(0,1)$, and ε is the spread. Since our time window is small there is no crucial difference between arithmetic random walk and geometrical Brownian motion. Due to a number of well calibrated models for intraday volatility, see[12], the volatility parameter σ in (1) can be estimated in a satisfactory way in normal market conditions.

We adopt Almgren's market impact model, [3]. Impact function depends on two parameters, spread ε and intensity of trade λ. Intensity of trade is defined as a ratio of traded volume and time, taking into account ADV (Average Daily Volume), and the market impact function is given by

$$f(q) = \varepsilon + \bar{\mu}\lambda^b, \ \lambda = \lambda(q),$$

where ε is the spread and $\bar{\mu}$ is a stock-specific parameter, λ is trading intensity, $b \in [0,1]$ and q is the size of market order. Market impact function

f gives the value of impact in money/share units and thus the total impact cost of trading q shares is

$$\pi(q) = f(q)q \tag{4}$$

For more details see [3, 4, 5, 6, 7].

Non-trivial order sizes cannot be executed as a single market order. As such, Almgren assumes a larger order is broken into a sequence of sub orders, executed according to some distribution. One common option is uniform distribution and we will also assume this distribution for our market orders and a temporary market impact function given by (4).

In two-market situation we are dealing with two sequences of gain coefficients for limit orders

$$c_i^A = a_1^A - b_i^A,\ c_i^B = a_1^B - b_i^B,\ i = 1, \dots, n, \tag{5}$$

for bid levels $i = 1, \dots, n$. Obviously gain (5) occurs only if the order is filled within a given time. We will define gain function for limit orders as follows.

At any of the considered venues at $t = 0$ with market conditions $\mathcal{M}$ we define the set of functions $F_i(q)$ that gives the probability that the order of size q placed at the bid level i will be filled within time interval $[0, T]$. These functions will be called Fill Probability functions in this paper. Assuming that T is fixed and the set of market conditions $\mathcal{M}$ is available these functions clearly satisfy $F_i(q) > F_{i+1}(q)$. Furthermore we will assume that F_i are smooth enough. An analytical expression for F_i is not available and the various trading institutions use their own proprietary functions. For detailed comments one can see [19]. Clearly each market has its own set of Fill Probability functions, F_i^A and F_i^B for each bid level $i = 1, \dots, n$. The Fill Probability Functions for each of the markets are calculated independently of each other.

Using the above defined functions we can define the *success functions* of the considered limit order as

$$H_i^A(q) = qF_i^A(q),\ H_i^B(q) = qF_i^B(q),\ i = 1, \dots, n \tag{6}$$

and *gain functions* as

$$G_i^A(q) = c_i^A H_i^A(q),\ G_i^B(q) = c_i^B H_i^B(q),\ i = 1, \dots, n \tag{7}$$

Clearly functions H_i, G_i are smooth if F_i are smooth. Although we have no analytical expression for $F_i(q)$ we are able to use an estimate of reasonable quality as will be demonstrated by numerical examples in Section 4.

3 The optimization model

The problem we will consider is that of executing an order to buy some volume Q within time window $[0, T]$ of a given security. The sell case is clearly opposite so we will not consider it here. Our execution strategy will be a combination of market and limit orders at both venues A and B that minimizes estimated costs in terms of volatility and market impact. We will follow the general idea successfully applied to a single venue in [19] but taking into consideration additional possibilities arising from two venues. The principal aim is to obtain an optimization model that is computationally affordable in real time for a large portfolio of securities. As already mentioned the price process is not deterministic nor is any of the other market micro properties (liquidity arrival, cancellation pattern, changes in spread etc.) that determine the market conditions. The existence of multiple trading venues with mutual dependency makes the trading environment even more complex.

The question we are facing is distribution between market and limit orders and distribution between venues A and B at $t = 0$. Both costs and gains are clearly stochastic values. At $t = T$ we have the residual amount coming from the unfilled limit orders. As we have a fixed trade window, the residual needs to be executed at $t = T$ relatively fast and in an aggressive manner i.e. using only market orders. This will produce large impact and is subject to volatility risk since the prices $P(0)$ and $P(T)$ will very likely be different. Further more that residual volume can be traded at one or both of the venues. Putting all these considerations together, one is facing a two stage stochastic problem with the objective function being impact and volatility costs of market orders and negative gain of limit orders. Such problems are not computationally feasible for real time use and large portfolios of securities. Hence some simplifications in modelling are necessary.

We will adopt the gain and success functions as already defined in Section 2. Thus the distribution of limit orders between two venues and different bid levels will be determined by the corresponding fill probability functions. The impact costs will be modelled using (4) for each of the venues separately. Possible price and liquidity improvements at one of the venues are thus taken into account and will result in different distribution of market orders between A and B. Therefore we are actually treating two venues as a single combined venue with additional bid-ask levels and two impact functions compared to strategy from [19]. The key difference in the two-venue situation is residual, carrying its volatility risk and impact costs.

The residual is clearly an unknown value at $t = 0$. To simplify the problem we will introduce the residual function as a deterministic function available at $t = 0$ following the logic of the success and gain functions. The volatility risk can be simplified by adopting the risk averse attitude and assuming that the price will move away from us for the whole σ. With these assumptions we cover more than 90% of cases under the process (1). The total impact cost of the residual will be the sum of impact costs at both markets assuming that the residual is divided between them. The exact ratio of the split between A and B is obtained minimizing the total impact costs. As the residual impact and volatility costs influence the distribution of Q between limit and market orders as well as the distribution between the venues at $t = 0$, the resulting optimization problem will be a bi-level problem as stated in this Section.

We assume that the volatility parameter σ is available as well as market impact functions defined in [6] and explained with (4). Risk free arbitrage opportunities force the prices in the two venues to be aligned and as such the volatilities of the different venues are virtually identical. Furthermore, given the market conditions $\mathcal{M}^A, \mathcal{M}^B$ we are able to state the Fill Probability functions $F_i^A(q), F_i^B(q)$ for any order size q and any bid level $i = 1, \ldots, n$ for time interval $[0, T]$ at any of the markets A and B.

If $x^A = (x_1^A, \ldots, x_n^A)$ then we initially place limit order x_i^A at ith bid level for $i = 1, \ldots, n$ and trade market orders of size y^A at market A and analogously for $x^B = (x_1^B, \ldots, x_n^B)$ and y^B at market B. We also use the notation $x = (x^A, x^B) \in \mathcal{R}^{2n},\ y = (y^A, y^B) \in \mathcal{R}^2$.

At $t = T$ we are left with the residual that has not been filled

$$\bar{R} = Q - \sum_{i=1}^{n} \gamma_i^A x_i^A - \sum_{i=1}^{n} \gamma_i^B x_i^B - y^A - y^B \tag{8}$$

where $\Gamma = (\gamma_1^A, \ldots, \gamma_n^A, \gamma_1^B, \ldots, \gamma_n^B)$ is a stochastic variable showing the relative value of each limit order that was filled, i.e. $\gamma_i \in [0, 1]$. We will trade that residual as a market order at any of the markets depending on the market conditions at $t = T$. The residual will be executed in a short time afterwards, say within a fraction of T.

Initial market order y^A is causing market impact and therefore its execution cost is

$$\pi^A(y^A) = (\varepsilon^A + \mu^A y^A) y^A, \tag{9}$$

The same is true for market B and y^B,

$$\pi^B(y^B) = (\varepsilon^B + \mu^B y^B) y^B. \tag{10}$$

Here μ^A and μ^B are stock specific parameters.

Limit orders have their gains according to their respective gain coefficients if filled and opportunity cost if unfilled within $[0, T]$. The residual given by (8) is subject to volatility risk and since we need to execute it fast at $t = T$, its execution will cause larger impact due to larger intensity of trade (larger traded volume within that time window). Let $\Pi^A(R), \Pi^B(R)$ denote these impact costs. With G_i defined by (7) as

$$G_i(x_i) = c_i x_i F_i(x_i),\ c_i = a_1(0) - b_i(0)$$

and assumptions made in Section 2, we can formulate the gain of limit orders as

$$G^A(x^A) = \sum_{i=1}^{n} G_i^A(x_i^A),\ G^B(x^B) = \sum_{i=1}^{n} G_i^B(x_i^B). \tag{11}$$

Instead of considering the volatility risk of the residual as stochastic value dependent on price movement we can assume that during the time window $[0, T]$ the price will drift away for one whole volatility σ. In fact the expected price drift is zero under assumption (1) but volatility of price plays a more important role within short time framework and thus we put an additional safeguard with the residual function. Analogously to gain function (7) we define the *residual function*,

$$R(x, y) = Q - H^A(x^A) - H^B(x^B) - y^A - y^B, \tag{12}$$

$$H^A(x^A) = \sum_{i=1}^{n} H_i^A(x_i^A), H^B(x^B) = \sum_{i=1}^{n} H_i^B(x_i^B).$$

The residual has to be executed within a fraction of T in a manner that will minimize the total impact cost. Hence we need to split it to r^A and r^B such that r^A is executed at A and r^B is executed at B. So r^A and r^B are the solutions of

$$\min_r \phi(r)$$

under constraints

$$r^A + r^B = R(x, y),\ r^A, r^B \geq 0,$$

with $r = (r^A, r^B)$. Given that the residuals are executed faster than y they are causing larger impact than stated by f^A and f^B. So we model the impact cost of the residual orders as

$$\Pi^A(q) = (\varepsilon^A + \eta^A q)q,\ \Pi^B(q) = (\varepsilon^B + \eta^B q)q$$

with $\eta^A > \mu^A, \eta^B > \mu^B$, and

$$\phi(r) = \Pi(r^A) + \Pi(r^B).$$

Denoting

$$\begin{aligned} \varphi(x,y) &= \pi^A(y^A) + \pi^B(y^B) - G^A(x^A) - G^B(x^B) + \\ & \sigma\sqrt{T}R(x,y) + \Pi^A(r^A) + \Pi(r^B), \end{aligned} \tag{13}$$

our problem yields the following bi-level optimization problem

$$\min_{x,y} \quad \varphi(x,y) \tag{14}$$

$$\text{s.t.} \quad \textstyle\sum_{i=1}^n x_i^A + \sum_{i=1}^n x_i^B + y^A + y^B - Q = 0 \tag{15}$$

$$r = arg\min_r \phi(r(x,y)) \tag{16}$$

$$r^A + r^B = R(x,y) \tag{17}$$

$$x, y, r \geq 0 \tag{18}$$

Function $\phi(r)$ is quadratic and the lower level problem is a strictly convex quadratic problem with linear and nonnegativity constraints. Therefore it admits a unique solution so we are able to prove the following statement. Let $\mathcal{R}_0$ be the set of nonnegative real numbers.

Theorem 1 *Let $(H_i^A), (H_i^B) \in C^2(\mathcal{R}_0)$ be concave functions for $i = 1, \ldots, n$ and r be the optimal solution of (16)-(17). Then $\nabla^2\varphi(x,y)$ is a positive definite matrix.*

Proof. See Appendix.

Without an analytical expression for the Fill Probability function F_i, one cannot claim that the success function H_i which is defined by F_i, satisfies the concave condition from this theorem. However, the empirical results give us reasons beyond any doubt that for q smaller than the average traded volume, H_i is indeed concave. Atomic orders rarely are more than 33% of the average traded volume.

3.1 Multiperiod model

After placing the orders at t_0, if the market price moved away, one may then want to revise one's initial order placements by finding another optimal execution strategy, taking into consideration these changes to the market conditions.

Let $\tau \in (0, T)$ be the point when we start the re-optimization procedure. The goal of the re-optimization is to improve the performance of the initially planned execution strategy defined by $x^{0,A}, x^{0,B}, y^{0,A}$ and $y^{0,B}$ which are the optimal values obtained by solving (14)-(18) at $t = 0$. Let us denote by superscript 0 the corresponding Fill Probability functions $F_i^{0,A}$ and $F_i^{0,B}$ and gain function $G_i^{0,A}, G_i^{0,B}$. These functions are assumed to be available at $t = 0$ considering time execution window $[0, T]$. At $t = \tau$ several informations are available. Firstly, for all $x_i^{0,A}$ and $x_i^{0,B}$ initially placed at bid levels $i \in B_0$ the unfilled amounts $\tilde{x}_i^A \leq x_i^{0,A}$, $\tilde{x}_i^B \leq x_i^{0,B}$ are known. The amount traded as market orders at both exchanges is also known and therefore the remaining quantity Q^τ is known. The basic idea of re-optimization procedure is to take advantage of new market conditions $\mathcal{M}^{\tau,A}$ and $\mathcal{M}^{\tau,B}$ if they are significantly different from the initial conditions $\mathcal{M}^{0,A}$ and $\mathcal{M}^{0,B}$. Thus starting with Q^τ and the execution window $[\tau, T]$ one can repeat the reasoning which yields (14) - (18) with one important difference. Namely the unfilled part of the limit orders $x^{0,A}$ and $x^{0,B}$ i.e. $\tilde{x}_i^A$ and $\tilde{x}_i^B$ can be either canceled or left at their position in the corresponding queues at $t = \tau$.

The situation is essentially different from $t = 0$ since the orders which are not filled at $t = \tau$ have very likely progressed in their respective queues and hence have different fill probability than new limit orders one might place at $t = \tau$. Furthermore their fill probability functions are different from the initial F_i^0 since the market conditions as well as the execution window are different. So we will have two sets of fill probability functions, $\tilde{F}_i^{\tau,A}$ and $\tilde{F}_i^{\tau,B}$ for the orders placed at $t = 0$ that we keep at their positions and $F_i^{\tau,A}, F_i^{\tau,B}$ for the new limit orders that will be placed at the end of the corresponding queues at $t = \tau$. To distinguish between these two sets of limit orders we introduce a new set of variables $\ell_i^{\tau,A}$, $\ell_i^{\tau,B}$ $i \in B_0$ denoting the volume we are keeping at the initial positions, while $x_i^{\tau,A}$ and $x_i^{\tau,B}$ are the limit orders submitted at $t = \tau$.

Clearly we cannot rule out the possibility of a significant change of the market conditions contrary to our aims which yields a decrease in the fill probability functions if compared with the initial fill probability functions i.e. $\tilde{F}_i^{\tau,A} < F_i^{0,A}$ and $\tilde{F}_i^{\tau,B} < F_i^{0,B}$ nor a significant (although temporary) change in liquidity distribution between A and B. The change of prices could be of such magnitude that the set of available bid levels change at $t = \tau$. So cancellation of the initially posted but unfilled orders has to be taken as a possibility. All these imply the following inequality conditions on the limit orders we will keep as initially placed

$$\ell_i^{\tau,A} \geq 0,\ \ell_i^{\tau,A} \leq \tilde{x}_i^A\ \ell_i^{\tau,B} \geq 0,\ \ell_i^{\tau,B} \leq \tilde{x}_i^B\ i \in B_0. \tag{19}$$

These orders will have success functions

$$\tilde{H}_i^{\tau,A}(\ell_i^{\tau,A}) = \tilde{F}_i^{\tau,A}(\ell_i^{\tau,A})\ell_i^{\tau,A},\ \tilde{H}_i^{\tau,B}(\ell_i^{\tau,B}) = \tilde{F}_i^{\tau,B}(\ell_i^{\tau,B})\ell_i^{\tau,B} \tag{20}$$

and gain functions $\tilde{G}_i^{\tau,A}(\ell_i^A) = c_i^{\tau,A}\tilde{H}_i^{\tau,A}(\ell_i^A),\ \tilde{G}_i^{\tau,B}(\ell_i^B) = c_i^{\tau,B}\tilde{H}_i^{\tau,B}(\ell_i^B)$ with gain coefficients

$$c_i^{\tau,A} = a_1^A(\tau) - b_i^A(\tau),\ c_i^{\tau,B} = a_1^B(\tau) - b_i^B(\tau),\ i \in B_0. \tag{21}$$

The price process might yield a new set of the available bid levels at $t = \tau$, say B_τ. If $x_k^{\tau,A}$ and $x_k^{\tau,B}, k \in B_\tau$ are the new limit orders to be placed at $t = \tau$ at markets A and B then their success functions are

$$H_k^{\tau,A}(x_k^{\tau,A}) = F_k^{\tau,A}(x_k^{\tau,A})x_k^{\tau,A},\ H_k^{\tau,B}(x_k^{\tau,B}) = F_k^{\tau,B}(x_k^{\tau,B})x_k^{\tau,B}, \tag{22}$$

while the gain functions are

$$G_k^{\tau,A}(x_k^{\tau,A}) = c_k^{\tau,A}H_k^{\tau,A}(x_k^{\tau,A}),\ G_k^{\tau,B}(x_k^{\tau,B}) = c_k^{\tau,B}H_k^{\tau,B}(x_k^{\tau,B})$$

with

$$c_k^{\tau,A} = a_1^A(\tau) - b_k^A(\tau),\ c_k^{\tau,B} = a_1^B(\tau) - b_k^B(\tau),\ k \in B_\tau. \tag{23}$$

Clearly $F_k^{\tau,A}(q) \leq \tilde{F}_k^{\tau,A}(q)$ and $F_k^{\tau,B}(q) \leq \tilde{F}_k^{\tau,B}(q)$ due to different positions in the queues for $k \in B_0 \cap B_\tau$. The distribution of the new limit orders will depend on improvement (deterioration) of $\tilde{F}_k^\tau$ compared to F_i^0 as well as the relationship between $\tilde{F}_k^{\tau,A}(q)$ and $\tilde{F}_k^{\tau,B}(q)$.

Finally let $y^{\tau,A}, y^{\tau,B}$ denote the volumes we will trade as market orders in $[\tau, T]$ in both markets. Then the impact costs with the linear impact function are

$$\pi^{\tau,A}(y^{\tau,A}) = (\varepsilon^A + \mu^{\tau,A}y^{\tau,A})y^{\tau,A},\ \pi^{\tau,B}(y^{\tau,B}) = (\varepsilon^B + \mu^{\tau,B}y^{\tau,B})y^{\tau,B}$$

with $\mu^{\tau,A}, \mu^{\tau,B}$ being a stock specific constants dependent on time $T-\tau$. The new residual function is analogously to (12),

$$\rho(l^\tau, x^\tau, y^\tau) = Q^\tau - \tilde{H}^{\tau,A}(\ell^{\tau,A}) - \tilde{H}^{\tau,B}(\ell^{\tau,B}) - H^{\tau,A}(x^{\tau,A}) - H^{\tau,B}(x^{\tau,B}) - y^{\tau,A} - y^{\tau,B}, \tag{24}$$

with

$$\tilde{H}^{\tau,A}(\ell^{\tau,A}) = \sum_{i\in B_0} \tilde{H}_i^{\tau,A}(\ell_i^{\tau,A}),\ \tilde{H}^{\tau,B}(\ell^{\tau,B}) = \sum_{i\in B_0} \tilde{H}_i^{\tau,B}(\ell_i^{\tau,B}),$$

$$H^{\tau,B}(x^{\tau,B}) = \sum_{k\in B_{\tau,B}} H_k^{\tau,B}(x_k^{\tau,B}),\ H^{\tau,B}(x^{\tau,B}) = \sum_{k\in B_{\tau,B}} H_k^{\tau,B}(x_k^{\tau,B}).$$

Denoting $\ell^\tau = (\ell^{\tau,A}, \ell^{\tau,B}), x^\tau = (x^{\tau,A}, x^{\tau,B}), y^\tau = (y^{\tau,A}, y^{\tau,B})$ and splitting the residual $\rho(l^\tau, x^\tau, y^\tau)$ into two parts, $r^{\tau,A}$ and $r^{\tau,B}$ to be executed at A and B, with $r^\tau = (r^{\tau,A}, r^{\tau,B})$ we are again facing the bilevel problem.

The optimization problem now becomes

$$\begin{aligned}
\min_{l^\tau,x^\tau,y^\tau} \quad & \Phi(\ell^\tau, x^\tau, y^\tau) && (25)\\
\text{s.t.} \quad & \ell_i^\tau \in [0, \tilde{x}_i],\ i \in B_0 && (26)\\
& Q^\tau = y^{\tau,A} + y^{\tau,B} + \sum_{i\in B_0}(\ell_i^{\tau,A} + \ell_i^{\tau,B}) + \sum_{k\in B_\tau}(x_k^{\tau,A} + x_k^{\tau,B}) &&\\
r^\tau \in \quad & arg\min \Pi^{\tau,A}(r^{\tau,A}) + \Pi^{\tau,B}(r^{\tau,B}) && (27)\\
\rho^\tau = \quad & r^{\tau,A} + r^{\tau,B} && (28)\\
& x^\tau, y^\tau \geq 0 &&
\end{aligned}$$

with

$$\begin{aligned}
\Phi(\ell^\tau, x^\tau, y^\tau) \quad = \quad & -\tilde{G}^{\tau,A}(\ell^{\tau,A}) - \tilde{G}^{\tau,B}(\ell^{\tau,B}) - G^{\tau,A}(x^{\tau,A}) - G^{\tau,B}(x^{\tau,B}) + \\
& \pi^{\tau,A}(y^{\tau,A}) + \pi^{\tau,B}(y^{\tau,B}) + \sigma\rho(l^\tau, x^\tau, y^\tau)\sqrt{T-\tau} + \\
& +\Pi^{\tau,A}(r^{\tau,A}) + \Pi^{\tau,B}(r^{\tau,B})
\end{aligned}$$

and $G^{\tau,A}, G^{\tau,B}, \tilde{G}^{\tau,A}, \tilde{G}^{\tau,B}$ defined analogously to the success functions H functions, i.e. summing up all components. Due to faster execution of the residual, the impact costs of the residuals are

$$\Pi^{\tau,A}(q) = (\varepsilon^A + \eta^{\tau,A}q)q,\ \Pi^{\tau,B}(q) = (\varepsilon^B + \eta^{\tau,B}q)q$$

with $\eta^{\tau,A} > \mu^{\tau,A}$ and $\eta^{\tau,B} > \mu^{\tau,B}$.

The problem (25)-(28) has the same structure as (14)-(18) except for the box constrains for l^τ and larger dimension. Therefore the objective function again has positive definite Hessian under the conditions stated below.

Theorem 2 *Let $H_k^{\tau,A}, H_k^{\tau,B}, \tilde{H}_i^{\tau,A}, \tilde{H}_i^{\tau,B} \in C^2(\mathcal{R}_0)$ and $H_k^{\tau,A}, H_k^{\tau,B}, \tilde{H}_i^{\tau,A}, \tilde{H}_i^{\tau,B}$ concave for all $k \in B_\tau$ and $i \in B_0$. Then $\nabla^2\Phi(\ell, x, y)$ is a positive definite matrix.*

4 Numerical Results

Throughout our simulations, we have endeavored to be as faithful as possible to the real-time usage of the proposed models. This is of paramount importance to us as the primary objective is to develop a model that can be utilized in live trading. Therefore, there are no assumptions made in the simulation framework nor in the preprocessing of the data that could prevent direct application. The simulator was written in Java and Matlab and the granularity of data used was level 2 tick data. Where one is concerned with the orderbook queue, queue details such as position and quantity were maintained to accurately assess the fills. Where cancellation positions cannot be determined, we made the conservative assumption that the order was cancelled at the back of the queue. The matlab sub-routine $fmincon()$ was used to solve (14) - (18) and (24) - (28).

The results are given in Tables 1-5. We considered 3 months worth of data (August to October 2009) from LSE and Euronext as the primary markets while Chi-X was the secondary market in our simulations. Each day was sliced into 61 time slots of 8 minutes, from 08:16 to 16:24. In all those tables, the first column gives the order size which is defined as a percentage of period average traded quantity. Therefore our atomic order is defined with 8 minutes duration. The first column denotes quantity. The terms $MMSP$ and $MMMP$ are acronyms for Multi-Market Single-Period and Multi-Market Multi-Period optimal execution strategies.

% of ADV	M	B	O_1
1	47	28	12
3	53	30	17
5	61	32	22
8	70	37	29
10	75	39	31
12	77	40	36
15	75	37	33

Table 1: VOD

% of ADV	M	B	O_1
1	53	15	8
3	61	15	9
5	65	15	9
8	76	19	14
10	80	19	16
12	81	19	17
15	82	17	16

Table 2: AAL

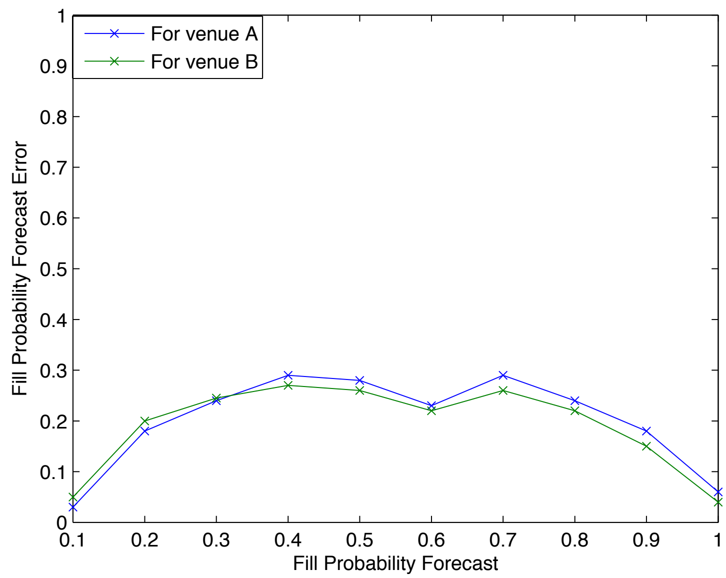

Figure 2: Mean Error of the Fill Probability Model for multiple venues.

% of ADV	M	B	O_1
1	35	25	7
3	58	28	15
5	73	32	20
8	87	35	25
10	88	30	22
12	90	29	20
15	89	25	18

Table3: SASY

% of ADV	M	B	O_1
1	60	18	4
3	66	20	8
5	64	19	9
8	64	20	10
10	61	19	9
12	61	20	11
15	57	17	10

Table 4: KGF

% of ADV	M	B	O_1
1	61	5	3
3	62	5	5
5	57	2	0
8	56	2	-2
10	52	0	-5
12	49	-1	-7
15	41	-5	-10

Table 5: SDR

Columns 2 to 4 show the relative performance of the three basic alternative benchmarks, i.e. Market, All on Bid and Single Period Optimal tra-

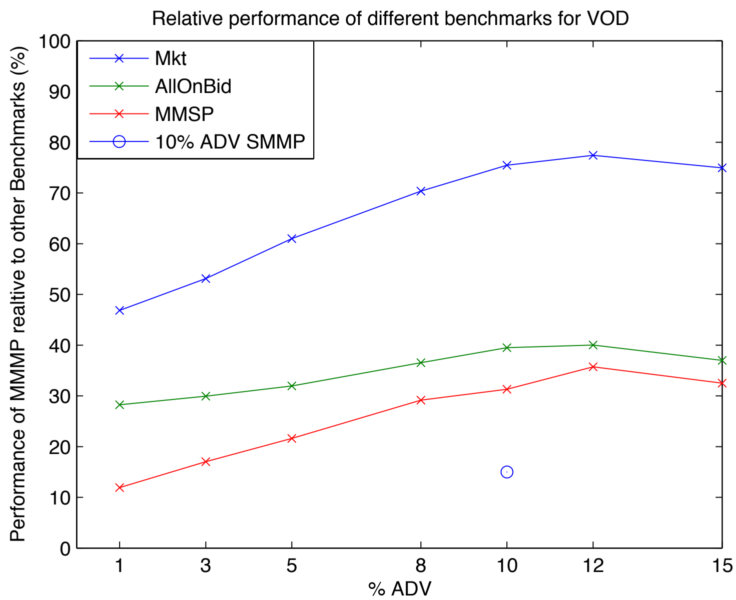

Figure 3: Performance comparison of trading VOD in two venues.

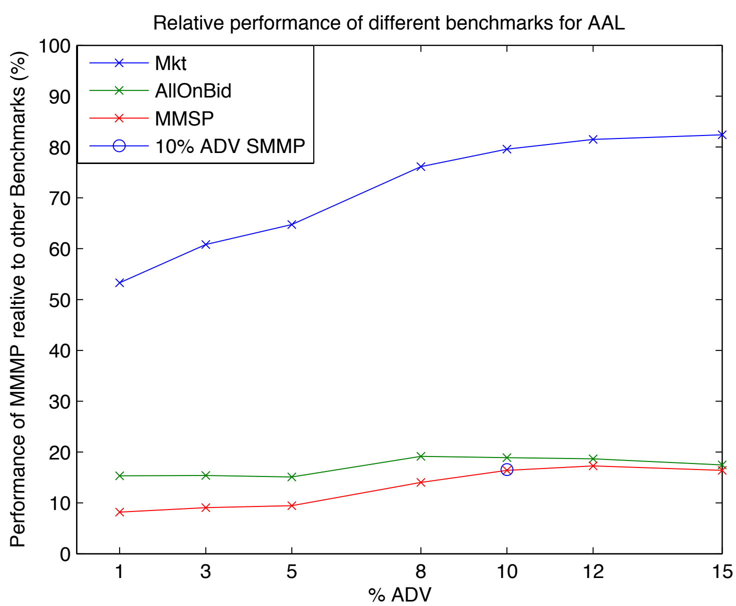

Figure 4: Performance comparison of trading AAL in two venues.

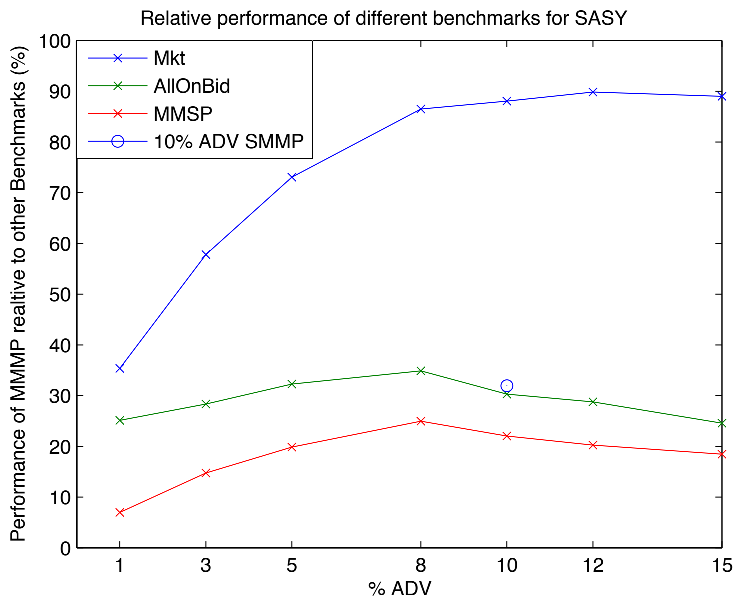

Figure 5: Performance comparison of trading SASY in two venues.

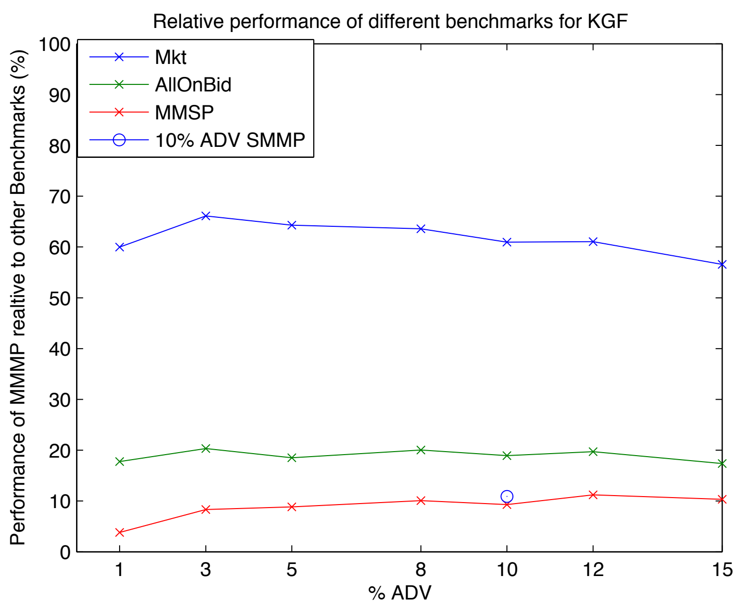

Figure 6: Performance comparison of trading KGF in two venues.

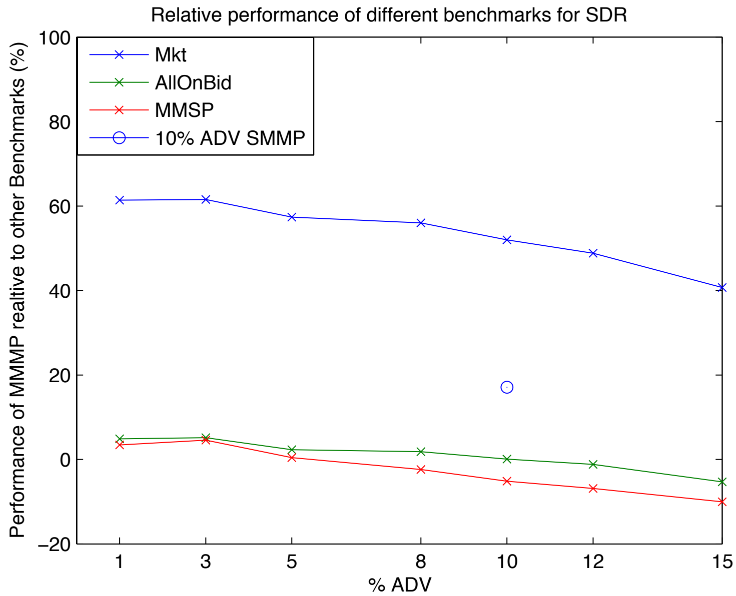

Figure 7: Performance comparison of trading SDR in two venues.

jectory. The performance of these benchmarks are given as the percentage worse than the optimal multi-market multi-period optimization method. The choice of the above three measures as benchmarks is rooted in the fact that the finance industry does not have any valid benchmarks for measuring performance. Furthermore benchmarks can be affected by the way in which one trades. Therefore, we have chosen two benchmarks that are common market practice. In a multi-market environment, the market place could be thought of as an aggregated single market, hence the performance of optimal execution in a single market is included. In addition to these three benchmarks, included also is a single point benchmark of Single Market Multi-Period.

The difference between the single period model and the two-period model has been extensively covered in earlier chapters. The objective of the multi-market execution method was to further improve the execution performance by tapping into the additional liquidity provided by alternative venues to the primary market. We now look in detail at 10% of ADV for VOD order as a typical example. Intuitively, this model resembles the Single Market Multi-Period model in terms of split between market and limit orders. Both models have a mid-point re-balancing opportunity to cancel orders and reconsider better alternatives. In the case of multi-market, one will have the option of not only choosing a better price to place the new orders, but also the option to choose a different venue.

All reported numbers are given as a percentage of the initial order size. Orders are split among two venues, A and B. At $t = 0$, mean values of market orders are $y^{0,A} = 6.8\%$ and $y^{0,B} = 1.6\%$. Limit orders in the two markets are $x_1^{0,A} = 59.2\%$, $x_1^{0,B} = 16.7\%$, $x_2^{0,A} = 10.6\%$, $x_2^{0,B} = 3.3\%$, $x_3^{0,A} = 0.6\%$, $x_3^{0,B} = 0.2\%$, $x_4^{0,A} = 0.1\%$, $x_4^{0,B} = 0.1\%$, $x_5^{0,A} = 0.8\%$, $x_5^{0,B} = 0.1\%$.
At $\tau = T/2$, one half of y^0 is realized while the unrealized limit orders were $\tilde{x}_1^{0,A} = 14.1\%$, $\tilde{x}_1^{0,B} = 4.1\%$, $\tilde{x}_2^{0,A} = 5.3\%$, $\tilde{x}_2^{0,B} = 1.6\%$, $\tilde{x}_3^{0,A} = 0.4\%$, $\tilde{x}_3^{0,B} = 0.1\%$, $\tilde{x}_4^{0,A} = 0.1\%$, $\tilde{x}_4^{0,B} = 0.1\%$, $\tilde{x}_5^{0,A} = 0.8\%$, $\tilde{x}_5^{0,B} = 0.1\%$ with respect to the total order size.
The order size for the second period was $Q^\tau = 30.9\%$ of the initial order and that value was distributed as $y^{1,A} = 2.8\%$ and $y^{1,B} = 0.7\%$ for market orders. New limit orders for the second period were distributed as $x_1^{\tau,A} = 13.6\%$, $x_1^{\tau,B} = 4.5\%$, $x_2^{\tau,A} = 1.0\%$, $x_2^{\tau,B} = 0.4\%$, $x_3^{\tau,A} = 0.0\%$, $x_3^{\tau,B} = 0.0\%$, $x_4^{\tau,A} = 0.0\%$, $x_4^{\tau,B} = 0.0\%$, $x_5^{\tau,A} = 0.0\%$, $x_5^{\tau,B} = 0.0\%$. While we kept at initial bid positions $l_1^{\tau,A} = 5.3\%$, $l_1^{\tau,B} = 0.9\%$, $l_2^{\tau,A} = 1.0\%$, $l_2^{\tau,B} = 0.4\%$.
Therefore the total amount of cancellations was 19% across both markets, given as $s_1^{\tau,A} = 8.8\%$, $s_1^{\tau,B} = 3.2\%$, $s_2^{\tau,A} = 4.3\%$, $s_2^{\tau,B} = 1.2\%$, $s_3^{\tau,A} = 0.4\%$, $s_3^{\tau,B} = 0.1\%$, $s_4^{\tau,A} = 0.1\%$, $s_4^{\tau,B} = 0.1\%$, $s_5^{\tau,A} = 0.8\%$, $s_5^{\tau,B} = 0.1\%$ and new limit orders account for 19.7% of the initial order size Q.

At the end of time window $t = T$, we had average residual size of 7.1% which was executed as a market order within roughly 3 minutes divided among both markets, dictated by price improvement and liquidity.

Figures 6-10 show the performance of the different benchmarks. We can see that the Multi-Market Multi-Period optimization models are not only significantly better than common market practice but are indeed generating distribution of volume between different bid levels and venues. The re-optimization procedure leads to new limit orders as well as preserving some initially posted limit orders as expected.

The share of market orders split among the two venues is relatively small (7.7% within time frame and 7.1% for residual). Another important observation is the high rate of success of limit orders at lower levels of depth as found in single market, multi-period. The gain from the optimal trajectory is increasing with the size of atomic order. That is caused by the quadratic impact cost, so any decrease in cost due to decrease of market orders and increase of limit orders is more significant.

Although the average daily volume traded on the security VOD on B is approximately 25% of that of VOD traded on A, the split of new limit orders between venues A and B at $t = 0$ are $A = 71.3$ and $B = 20.3$ of the total available quantity for execution. Essentially, B is given 28.4% of the order size of A. At $\tau = T/2$, B is given 30.5%. Interestingly, when re-balancing at $\tau = T/2$, a smaller amount of 62.5% was cancelled at A as opposed to 77.6% at B. This difference however in absolute terms is a mere 0.64%. We argue that the general fill properties of A and B as well as the marginally better estimation of fill probability at venue B is the cause of this difference.

Unlike the other securities considered, for SDR, the performance characteristics is somewhat different. In the single market scenario, after a certain order size, the single period performed better than multi-period optimization. Even in a multi-market scenario, single period optimal trajectory has the best performance, for, order size greater than 8% of ADV. The reasons for this are due to high volatility and sparse trading pattern. As a result the Fill Probability model overestimates the real probability for the best bid and at mid point we have large unfilled amounts. By re-optimization we are actually chasing the noise, since 4 minutes is not an optimal reevaluation point for this security. Therefore we end up sending a larger amount as a market order which yields large impact costs. On the other hand, in the single time procedure, we benefit from keeping the initial position at limit orders since the volatility works in the model's favor and the fill rate is significantly better.

References

[1] Alfonsi, A., Fruth, A., Schield, A., Optimal execution strategies in limit order books with general shape functions, Quantitative Finance 10(2), 2010, 143-157.

[2] Avellaneda, M., Stoikov, S., High frequency trading in a limit order books, Quantitative Finance 8 (3), 2008, 217-224.

[3] Almgren, R., Optimal Execution with Nonlinear Impact Functions and Trading-Enchanced Risk, Applied Mathematical Finance 10 (2003), 1-18.

[4] Almgren R., Execution Costs, Encyclopedia of Quantitative Finance, 2008

[5] Almgren, R., Chriss, N., Value Under Liquidation, Risk 12 (1999), 1-18.

[6] Almgren, R., Chriss, N., Optimal Execution of Portfolio Transactions, Journal of Risk 3 (2000), 1-18.

[7] Almgren, R., Thum, C., Hauptmann, E., Li, H., Equity market impact, Risk 18,7, July 2005, 57-62.

[8] Berkowitz, S. A., Logue, D.E., Noser, E.A. Jr., The total cost of transactions on the NYSE, Journal of Finance, 43,1 (1988), 97-112.

[9] Bouchaud, J.P., Gefen, Y., Potters, M., Wyart, M., Fluctuations and response in financial markets: The subtle nature of 'random' price changes, Quantitative Finance 4 (2), 2004, 57-62.

[10] Bouchard, B., Dang, N-M., Lehalle, C-A., Optimal Control of Trading Algorithms: A General Impulse Control Approach, SIAM J. Financial Math., 2 (2011), 404-438.

[11] Birge, J.H., Louveaux, F., Introduction to Stochastic Programming, Springer 1997.

[12] Dacorogna, M.M., Gençay, R., Müller, U., Olsen, R.B., Pictet, O.V., An Introduction to High-Frequency Finance, Academic Press, 2001.

[13] Dravian, T., Coleman, T.F., Li, Y., Dynamic liquidation under market impact, Quantitative Finance 11 (1), 2011, 69-80.

[14] Elsier, Z., Bouchaud, J.P. Kockelkoren, J., The price impact of order book events: market orders, limit orders and cancellations, Quantitative Finance (to appear), DOI:10.1080/14697688.2010.528444

[15] Engle, R., Ferstenberg, R., Execution risk: It's the same as investment risk, J. Portfolio Management, 33(2), (2007), 34-44.

[16] Engle, R.F., Lange, J., Measuring, Forecasting and Explaining Time Varying Liquidity in the Stock Market, Journal of Financial Markets, Vol. 4, No. 2 (2001), 113-142.

[17] Grinold, R., Kahn, R., Active Portfolio Management McGraw-Hill, 2nd edition, 1999.

[18] Kharroubi, I., Pham, H., Optimal Portfolio Liquidation with Execution Cost and Risk, SIAM J. Financial Math. 1, (2010), 897-931.

[19] Kumaresan, M., Krejić, N., A model for optimal execution of atomic orders, Computational Optimization and Applications, 46(2), 2010, 369-389.

[20] Lillo, F., Farmer, J.D., Mantegna, R.N., Master curve for price impact function, Nature 421 (2003), 129-130.

[21] Moazeni, S., Coleman, T.F., Li, Y., Optimal Portfolio Execution Strategies and Sensitivity to Price Impact Parameters, SIAM J. Optimization, 20(3), 2010, 1620-1654.

[22] Nocedal, J., Wright, S., J., Numerical Optimization, Springer, 1999.

[23] Obizhaeva, A., Wang, J. , Optimal Trading strategy and Supply/Demand Dynamics, NBER Working Paper No. 11444, 2005.

[24] Perold, A., The Implementation Shortfall: Paper versus Reality, J. Portfolio Management, 14 (1988), 4-9.

[25] Predoiu, S., Shaikhet, G., Shreve, S., Optimal Execution in a General One-Sided Limit-Order Book, SIAM J. Financial Math., 2 (2011), 183-212.

[26] Sun, E.W., Kruse, T., Yu, M-T., High Frequency Trading, liquidity and execution cost, Ann. Oper. Res. (to appear).

A Appendix

Proof of Theorem 1.

For $\Pi^A(q) = (\varepsilon^A + \eta^A q)q$, $\Pi^B(q) = (\varepsilon^B + \eta^B q)q$ we have

$$\phi(r) = r^T Br + r^T d$$

with $B = diag(\eta^A, \eta^B)$ and $d = (\varepsilon^A, \varepsilon^B)$. As B is positive definite the minimizer of (16)-(18) is given by

$$r = \frac{R + e^T d}{e^T B^{-1} e} B^{-1} e - d, \; e = (1,1)^T \tag{29}$$

with $R = R(x,y) = Q - H^A(x^A) - H^B(x^B) - y^A - y^B$. Plugging (29) back to (14) - (15), after some elementary calculations we can show that for $\eta = \frac{\eta^A \eta^B}{\eta^A + \eta^B}$

$$\frac{\partial^2 \varphi}{\partial (x_i^A)^2} = -(c_i^A + \sigma\sqrt{T} + \frac{\varepsilon^A \eta^B + \varepsilon^B \eta^A}{\eta^A + \eta^B})(H_i^A)''(x_i^A)R + \eta((H_i^A)'(x_i^A))^2$$

$$\frac{\partial^2 \varphi}{\partial (x_i^B)^2} = -(c_i^B + \sigma\sqrt{T} + \frac{\varepsilon^A \eta^B + \varepsilon^B \eta^A}{\eta^A + \eta^B})(H_i^B)''(x_i^B)R + \eta((H_i^B)'(x_i^B))^2$$

$$\frac{\partial^2 \varphi}{\partial (y^A)^2} = \mu^A + \eta, \; \frac{\partial^2 \varphi}{\partial (y^B)^2} = \mu^B + \eta$$

$$\frac{\partial^2 \varphi}{\partial (y^A) \partial (x_i^A)} = \eta (H_i^A)'(x_i^A), \; \frac{\partial^2 \varphi}{\partial (y^B) \partial (x_i^A)} = \eta (H_i^A)'(x_i^A)$$

$$\frac{\partial^2 \varphi}{\partial (y^A) \partial (x_i^B)} = \eta (H_i^B)'(x_i^B), \; \frac{\partial^2 \varphi}{\partial (y^A) \partial (x_i^B)} = \eta (H_i^B)'(x_i^B).$$

Thus $\nabla^2 \varphi(x,y)$ can be expressed as

$$\nabla^2 \varphi = D + uu^T$$

where D is the diagonal matrix with elements

$$d_k = -(c_k^A + \sigma\sqrt{T} + \frac{\varepsilon^A \eta^B + \varepsilon^B \eta^A}{\eta^A + \eta^B})(H_k^A)''(x_k^A)R, \; k = 1, \ldots, n$$

$$d_k = -(c_k^B + \sigma\sqrt{T} + \frac{\varepsilon^A \eta^B + \varepsilon^B \eta^A}{\eta^A + \eta^B})(H_k^B)''(x_k^B)R, \; k = n+1, \ldots, 2n$$

$$d_{2n+1} = \mu^A, \; d_{2n+2} = \mu^B$$

and

$$u = \sqrt{\eta}[(H_1^A)'(x_1^A) \; \ldots (H_n^A)'(x_n^A) \; (H_1^B)'(x_1^B) \; \ldots \; (H_n^B)'(x_n^B) \; 1 \; 1].$$

As $uu^T \geq 0$ the statement follows if all elements of D are positive which is clearly true.