A Model for Optimal Execution of Atomic Orders

Miles Kumaresan * Nataša Krejić †

March 23, 2009

Dedicated to José Mario Martínez, on the occasion of his 60th birthday

Abstract

Atomic Orders are the basic elements of any algorithm for automated trading in electronic stock exchanges. The main concern in their execution is achieving the most efficient price. We propose two optimal strategies for the execution of atomic orders based on minimization of impact and volatility costs. The first considered strategy is based on a relatively simple nonlinear optimization model while the second allows re-optimization at some time point within a given execution time. In both cases a combination of market and limit orders is used. The key innovation in our approach is the introduction of a Fill Probability function which allows a combination of market and limit orders in the two optimization models we are discussing in this paper. Under certain conditions the objective functions of both considered problems are convex and therefore standard optimization tools can be applied. The efficiency of the resulting strategies is tested against two benchmarks representing common market practice on a representative sample of real trading data.
Key words: nonlinear programming, convex programming, optimal execution strategy, algorithmic trading

MSC: 90C30, 90C90, 90B90.

1 Introduction

Algorithmic Trading is a relatively new way of executing orders at stock exchanges that began in the 1990's and is now used extensively. This kind of trading relies heavily on smart technologies and mathematical methods in order

*TransMarket Group Ltd., 68 Upper Thames Street, London EC4 U3BJ, e-mail: miles.kumaresan@transmarketgroup.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 Science, Republic of Serbia (Grant 144006). e-mail: natasak@uns.ac.rs

to provide efficient execution of orders. There are different models to suit the different needs of end-users such as VWAP (Volume Weighted Average Price), implementation shortfall and participation. (See [6],[15].) Of all the asset classes, equities has benefitted most from this kind of trading activity.

The main objective of algorithmic trading is the efficient execution of a given order (buy or sell) with specified trading quantity and price conditions. In other words, algorithmic trading is a mechanical execution of the investment decision made by an investor. The price and quantity are specified by the issuer of the order. The execution of an order inevitably yields execution costs and the main concern in this work is an execution strategy which minimizes this execution cost.

Execution costs are the difference between an ideal and actual trade. Direct and predictable costs such as commissions and fees are in general proportional to the transaction value and therefore not relevant to any optimization procedure (although they may be large and significant). Indirect costs, which depend heavily on an execution algorithm, come from limited liquidity and price motion due to the volatility. These costs are difficult to measure and predict and they will be the main concern in our optimization model.

The model will be built for the so-called atomic orders. Regardless of the objectives of the algorithm used to execute an order, all orders are decomposed into a sequence of atomic orders. The execution of these atomic orders directly translate into the overall behavior of the algorithms. The main characteristics of atomic orders is relatively short time span (measured in minutes) and relatively small size (quantity that we are buying/selling) compared to average daily volume (ADV). Without any loss of generality we will consider only buy orders in this paper.

Indirect costs, such as price impact and volatility, are dependent on the order type. There are two main order types - limit and market order. Limit order is considered passive since one specifies the price and quantity he/she is willing to pay and waits until/if the order is filled. Such orders are considered liquidity providers and essentially do not make any market impact i.e., they do not move the price[1]. The primary objective of limit orders is to capture a better price than the currently available ask-price (the price acceptable for sellers) but due to their passive nature they are subject to volatility costs and execution risk. Since the price can move away in the opposite direction the order might be left unfilled during a given time span and since execution must be fulfilled, a new order has to be issued afterwards at a higher price. Market orders are used to execute the transaction immediately and are considered aggressive. Execution of market orders bears no execution risk but they assume a number of costs. First of all, the order will be executed at the best available ask price (the smallest price asked by the sellers) which is certainly bigger than all bid prices (prices offered by buyers). Second, market orders are liquidity takers and therefore cause price

[1] In less liquid stocks arrival of limit order could either attract liquidity on the opposite side or scare of participants. In that sense limit order has an impact but we will disregard that aspect here since this kind of impact is significantly smaller than the impact made by trading market orders

movements that is called market impact. Intuitively it is quite clear that if one is buying aggressively then the price is going up but actually measuring market impact is quite difficult. Finally, if the order is large then it is divided into a sequence of smaller orders that are executed within a given time window so the volatility risk is again present. Since we are dealing with atomic orders, time span is short and the volatility risk of suborders will not have a significant influence.

There are many models of market impact available in the literature. Trading institutions use their own models to measure market impact, their models are based on academic work but are not available publicly. One of the most detailed studies is done by Almgren and Chriss, [4] and Almgren et al., [5]. Another approaches are presented in Bouchard et al. [7] and Lillo et al. [13]. In this paper we will adopt the market impact model from Almgren and Chriss [4].

Models for optimal execution strategy in the sense of minimal execution costs that are available in the literature deal with volatility and impact cost of market orders only. Optimal strategy based on implementation shortfall is derived in Almgren and Chriss [4]. A number of models focus on the risk and the mean cost of execution of a single trade or in some cases a sequence of trades, for example see Almgren and Chriss [3], Grinold and Kahn [12], Almgren [1], Obizhaeva and Wang [14]. An optimization model that integrates the portfolio decision and the execution strategy is developed in Engle and Fersteberg [10]. But in practice no main stream algorithm consists of market orders alone as they can not be competitive in the execution costs of those using a combination of market and limit orders. Therefore developing an optimization model for execution of atomic orders as a combination of market and limit orders is of considerable practical importance. In order to be applicable, a model must be relatively simple to allow real time solution for a large portfolio of stocks that are typically traded.

The model we present in this paper is based on minimization of execution costs of atomic orders consisting of limit and market orders. The key innovation in our model is the introduction of Fill Probability function that gives the probability of being filled (executed) for limit orders. Such function is not available analytically but it can be reasonably well estimated given the set of market conditions. The Fill Probability model used in this research is a proprietary mathematical model and its inner working cannot be disclosed. However we will address all the key properties of the model as required for the analysis. It should be noted that the optimization framework we propose herein is not dependent on this particular implementation of a Fill Probability model. Fill Probability function is incorporated into the objective function together with volatility and impact costs. We explain the necessary simplifications of trading process and reasoning that yields a deterministic nonlinear optimization problem. The strategy obtained from the model is risk-averse and the model is solvable by standard optimization tools in real time due to its simplicity. Given the differences in market properties of a large universe of stocks (mainly differences in volatility and liquidity) we also introduce a two-period optimization model that allows re-optimization of the strategy at mid (or some other appropriately chosen) point in time interval. This procedure appears to be particularly useful for liquid and volatile stocks.

Both presented models are tested on real trade data from the London Stock Exchange and Euronext. Comparison of trading strategies is dependent on the choice of benchmark. There are a couple of benchmarks available in the literature, (see Almgren [2].) The two most popular are *Arrival Price* and VWAP. The cost measured using arrival price as benchmark is called implementation shortfall. In the case of VWAP the benchmark is the evolving VWAP of the market. We are dealing with short time span using a combination of limit and market orders. Furthermore we are simulating and not affecting the real market, so we will measure the execution cost as a difference from an *ideal trade*. The ideal trade is defined here as a combination of limit and market orders that would yield the smallest execution costs i.e., the trade that would be possible if we were able to predict all trades in a considered time window with certainty. Using that benchmark, named *perfect* in this paper, we compare the optimal strategies developed here with the two strategies that represent common market practice.

This paper is organized as follows. In Section 2 we introduce details of the problem describing order book, possible risks and gains, market impact model and relevant market parameters as well as Fill Probability function. The model is developed in Section 3 and the two-period model is presented in Section 4. Numerical results are given in Section 5 while some conclusions are drawn in Section 6.

2 Preliminaries

Let us start explaining the structure of orderbook in details that will be used in optimization models. An orderbook at any moment contains buy and sell orders for a given security as shown in Figure 1.

Each order is placed on the corresponding bid (ask) price level according to the arrival time. Therefore orders form a queue of different sizes and filling (execution) process is governed by price process and arrival time priority. In other words, transactions take place when there is an agreement in price between buy and sell orders and it is done respecting the arrival queue. At any given time t, we will denote by $b_i(t)$ the price at ith bid level and by $a_i(t)$ the price at ith ask level. If t is fixed, we might drop it from price expressions but the meaning will be clear. The number of visible price levels varies at different exchanges and in numerical experiments we will assume that 5 levels are visible. If one is placing a limit order with price $b_i(t)$ and volume Q_i then the order is placed at the end of the existing queue at ith bid level. The order can be filled only after the whole queue ahead of it is filled or cancelled. Filling and cancellation distributions are very complex issues, (see [11]) but we will not need any details of these processes here.

The smallest possible difference in prices is called tick size and it is determined by the rules of stock exchange. Therefore prices are discrete. For very

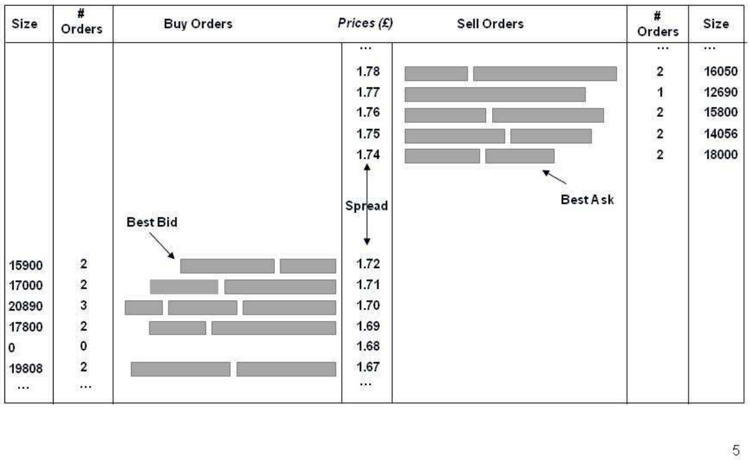

Figure 1: Orderbook.

liquid stocks the difference between bid levels is 1 tick - the smallest possible, while less liquid stocks can have multiple ticks in difference. This property of a particular stock will significantly influence the optimal execution trajectory. The difference

$$\varepsilon = a_1(t) - b_1(t)$$

is called the spread. The size of spread is again dependent on stock liquidity. Placing a market order actually means crossing the spread and buying at $a_1(t)$ or greater price, depending on order size and available ask volume.

A number of additional properties is available from the orderbook. If $\mathcal{M}$ denotes the current orderbook (current market conditions) then one can determine bid and ask prices, quantities, number of participants at each price level, volatility, VWAP price, cancelation pattern and so on. In further consideration we will denote by $\mathcal{M}$ an unspecified number of these properties since traders differ in their choice of relevant parameters and these differences will not influence our model.

In this paper we will assume that all prices follow an arithmetic random walk without drift,

$$b_i(t) = b_i(0) + \sigma\sqrt{t}\xi_i, \tag{1}$$

$$a_i(t) = a_i(0) + \sigma\sqrt{t}\gamma_i, \tag{2}$$

$$P(t) = P(0) + \sigma\sqrt{t}\zeta, \tag{3}$$

where P denotes the mid-price, $P = (a_1 + b_1)/2$, volatility is denoted by σ and the noise is Gaussian for bid and ask prices, $\xi_i, \gamma_i : \mathcal{N}(0,1),\ i = 1,\dots,n$ and consequently $\zeta : \mathcal{N}(0, (\sqrt{1/2})^2)$. 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[9], the volatility parameter σ in (1)-(3) can be estimated in a satisfactory way in normal market conditions.

Execution of any order is subject to two costs - volatility and market impact. We adopt Almgren's market impact model, [1]. Market impact is any deviation (even a fractional one) from the equilibrium price due to one's own trading activity. It can be divided into permanent and temporary impact. Temporary impact disappears in relatively short time according to liquidity pattern while permanent impact can stay well after the trade is executed. Temporary impact, according to [1], is larger than permanent by the order of magnitude and hence significantly more important for our model. Impact function depends on two parameters, spread ε and intensity of trade λ. Intensity of trade is defined as 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 [1], [4].

Market order of a reasonable size, meaning of non negligible volume, is never executed as a single trade. So the Almgren model is assuming some kind of optimal execution of market orders in a given time frame. Dividing an order into a sequence of small suborders we have several possibilities for their time schedule. One obvious possibility is uniform schedule within the time window. Other possibility is optimization of schedule with respect to implementation shortfall i.e. taking into account market impact and volatility. The relationship between volatility and impact which yields optimal duration for market orders is shown at Figure 2. We will assume uniform execution of market orders and use the temporary market impact cost function (4) as suggested in [4].

Contrary to market orders, limit orders do not produce market impact but face uncertainty of execution. Placing an order of size q at any bid level is thus subject to volatility risk: If the price drifts away before the order is filled we have the opportunity cost and since execution of an order is a must in our case, a new order has to be placed at a higher price. On the other hand, if the order is filled there is a clear gain in price compared with market order. Therefore, for any bid level we define gain coefficients as

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

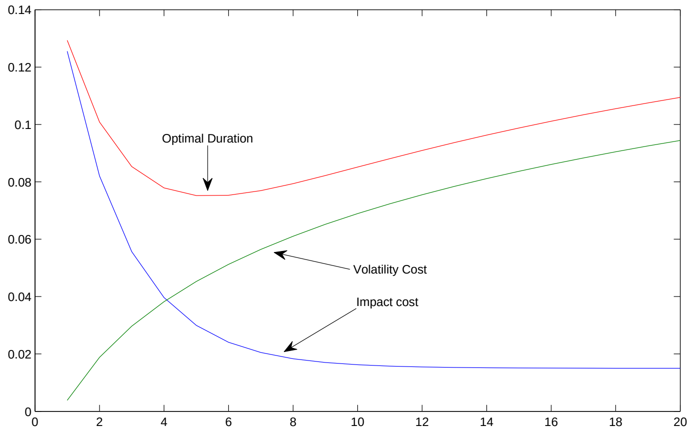

Figure 2: Impact and volatility costs versus time.

Obviously gain (5) occurs only if the order is filled within given time. We will define gain function for limit orders as follows. For any fixed bid level i and order of size q we define $\beta_i(q)$ as a random variable of Bernoulli type which takes value 1 if the order is filled within time interval $[0, t]$. Then

$$\beta_i(q) : \begin{pmatrix} 1 & 0 \\ p_i(q) & 1 - p_i(q) \end{pmatrix} . \tag{6}$$

Clearly p_i is the probability that the order will be filled and it is dependent on $\mathcal{M}$ and T. Keeping T fixed and placing an order at $t = 0$ with the price $b_i = b_i(0)$ we therefore expect that the filled amount will be qp_i. Using (6) we define the set of functions $F_i(q)$ for all $i = 1, \ldots, n$ as

$$F_i(q) = p_i(q),$$

assuming that T is fixed and $\mathcal{M}$ is available when we place the order at the ith bid level. Functions F_i will be called Fill Probability functions in this paper. In further considerations we will assume that given T and $\mathcal{M}$, all Fill Probability functions $F_i(q)$ are smooth enough for $q \geq 0$. If q_0 denotes the volume ahead of us at bid levels $k = 1, \ldots, i$ then

$$\lim_{q_0+q \to 0} F_i(q) = 1, \quad \text{and} \quad \lim_{q_0+q \to \infty} F_i(q) = 0.$$

Also $F_i(q) > F_{i+1}(q)$. Using the above defined functions we can define the *success functions* of the considered limit order as

$$H_i(q) = qF_i(q) \tag{7}$$

and *gain functions* as

$$G_i(q) = c_i H_i(q). \tag{8}$$

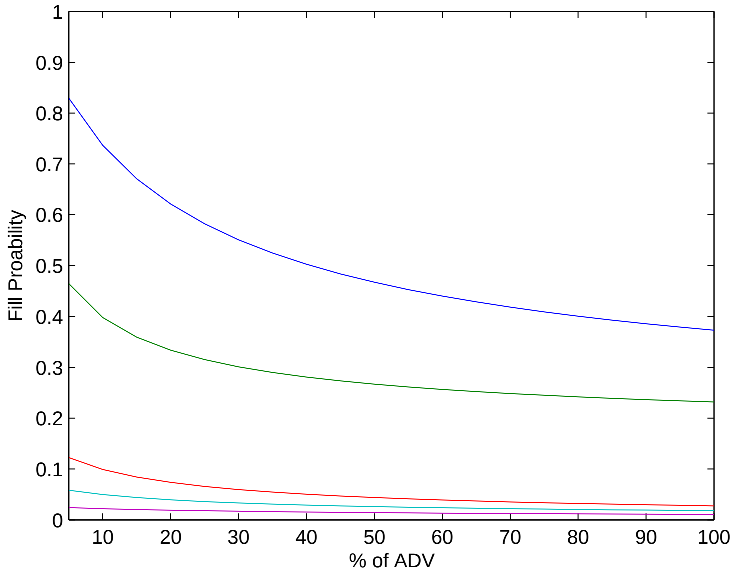

Figure 3: Fill Probability functions for five bid levels.

Clearly functions H_i, G_i are smooth if the 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 5. The empirical data also give us reason to believe that F_i are convex functions, (see Figure 3.)

3 The optimization model

Let us consider an atomic buy order with given size Q and execution time within $[0, T]$. In this context atomic means that Q is up to certain percentage of the average traded quantity within time window $[0, T]$ and T is small, say 10 minutes or similar. We want to formulate and solve an optimization problem which yields an optimal combination of market and limit orders for buying Q within given time. We will assume that the order book has n visible levels with price trajectories given by (1)-(3). Our execution strategy will be a combination of market and limit orders that minimizes expected costs in terms of volatility and market impact.

We assume that the volatility parameter σ is available as well as market impact functions defined in [4] and explained with (4). Furthermore, given the market conditions $\mathcal{M}$, we are able to state the Fill Probability functions $F_i(q)$ for any order size q and any bid level $i = 1, \dots, n$ for time interval $[0, T]$.

If $x = (x_1, \dots, x_n)$ then we will initially place limit order x_i at ith bid level for $i = 1, \dots, n$ and trade market orders of size y. Since the order size is Q we naturally have

$$y + \sum_{i=1}^{n} x_i = Q. \tag{9}$$

The execution of limit orders is an uncertain event. Let $\Gamma = (\Gamma_1, \ldots, \Gamma_n)$ be a stochastic variable which denotes the filled quantity (in relative terms) at each bid level during $[0, T]$ and let $\gamma = (\gamma_1, \ldots, \gamma_n)$ be a realization of Γ. At the end of time window, $t = T$, we are left with the residual that has not been filled

$$R = Q - \sum_{i=1}^{n} \gamma_i x_i - y \tag{10}$$

and we will trade that residual as a market order in a short time afterwards, say within a fraction of T.

Our objective is to minimize the execution cost of the above strategy, so let us describe all possible costs. Initial market order y is causing market impact and therefore its execution cost is

$$\pi(y) = f(y)y. \tag{11}$$

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 (10) is subject to volatility risk and since we need to execute it fast at $t = T$ (usually within a fraction of T) its execution will cause larger impact due to larger intensity of trade (larger traded volume within that time window). Let $\Pi(R)$ denotes that impact costs. With $G_i(q)$ defined by (8) 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

$$\sum_{i=1}^{n} G_i(x_i). \tag{12}$$

Residual R is clearly a stochastic variable depending on Γ. Volatility risk is depending on price trajectories (1)-(3) and we will denote it with $V(R), V(R) = (P(T) - P(0))R$. Putting together all these costs we are facing a two-stage stochastic problem - decision variables x, y are determined at $t = 0$ taking into account expected value of the residual R and the costs that will be caused by fast execution of the residual. Two-stage stochastic problems are solvable under additional assumptions for Γ and the price trajectory P, (see [8].) The distribution of Γ is not known. Furthermore Γ and P are not independent since the fill rate depends directly on P but Γ also depends on the whole set of variables in $\mathcal{M}$. Solving the above problem is not a realistic task without further simplifications and assumptions that are questionable in real life. Furthermore, one needs to determine an optimal strategy in real time and for a large universe of different stocks so solving two-stage stochastic problem is not an affordable option. Due to all these reasons we will define a deterministic model which has good theoretical properties and agrees with intuitive risk averse behavior of traders.

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 (3) but volatility of price plays a more important role within short time framework. Assuming that the price will move away from us for σ we are actually being risk-averse in more than 90% of cases under the assumption (4) since $\Phi(1) > 0.9$, with Φ cumulative distribution function for ζ.

Analogously to gain function (8), instead of considering the residual as a stochastic variable, we define the *residual function* as deterministic function,

$$r(x,y) = Q - \sum_{i=1}^{n} H_i(x_i) - y. \tag{13}$$

With these simplifications and taking the linear impact function we are able to state the volatility and impact costs as follows

$$V(r(x,y)) = \sigma\sqrt{T}r(x,y) \tag{14}$$

and

$$\pi(y) = (\varepsilon + \mu y)y,\ \Pi(r) = (\varepsilon + \eta r)r. \tag{15}$$

The constants μ and η are depending on time duration for execution of the corresponding market orders and the average traded volumes within these time windows. Therefore larger intensity of trade (shorter execution time) of the residual implies $\eta > \mu$, while ε is the average[2] historical spread value. Putting together all analyzed costs and gains with

$$\phi(x,y) = -\sum_{i=1}^{n} G_i(x_i) + (\varepsilon + \mu y)y + \sigma\sqrt{T}r(x,y) + (\varepsilon + \eta r(x,y))r(x,y), \tag{16}$$

our problem is

$$\min_{x,y} \quad \phi(x,y) \tag{17}$$

$$\text{s.t.} \quad Q = \sum_{i=1}^{n} x_i + y \tag{18}$$

$$x \geq 0,\ y \geq 0$$

Problem (17)-(18) is a nonlinear optimization problem with a single nonnegativity constraint on the variables. It can be solved by standard optimization tools. We will show that the Hessian matrix of the objective function is positive definite under some conditions. The simple structure of the problem and positive definitness of the Hessian then imply application of second order conditions and every KKT point is a minimizer of (17)-(18). Let $\mathcal{R}_0$ be the set of nonnegative real numbers.

[2]Using the average historical spread value is slightly less precise than the actual spread in function π but in the line with already introduced simplifications since $\varepsilon(T)$ is not known at $t = 0$.

Theorem 1 *Let $H_i \in C^2(\mathcal{R}_0)$ and concave ($H_i'' < 0$) for all i. Then $\nabla^2\phi(x,y)$ is a positive definite matrix.*

Proof. Let f_{ij} denote the elements of $\nabla^2\phi(x,y)$. Elementary calculations give us

$$f_{n+1,n+1} = 2\mu + 2\eta,$$

$$f_{ii} = 2\eta(H_i'(x_i))^2 - A_i H_i''(x_i),\;\; A_i = \sigma\sqrt{T} + c_i + \varepsilon + 2\eta r(x,y),\; i = 1,\dots,n,$$

$$f_{n+1,j} = f_{j,n+1} = 2\eta H_j'(x_j), j = 1,\dots,n,$$

$$f_{ij} = 2\eta H_i'(x_i)H_j'(x_j),\; i \neq j.$$

Therefore we can write

$$\nabla^2\phi(x,y) = D + uu^T,\;\; D = \mathrm{diag}(-A_1 H_1''(x_1),\dots,-A_n H_n''(x_n), 2\mu)$$

with

$$u = [\sqrt{2\eta}H_1'(x_1),\dots,\sqrt{2\eta}H_n'(x_n),\sqrt{2\eta}]^T.$$

Since uu^T is positive semi definite it is sufficient to prove that D is positive definite. As D is diagonal we must have that each entry of the diagonal is positive but that is clear since $A_i > 0$ and $H_i''(x_i) < 0$. So, we can conclude that $\nabla^2\phi(y,x)$ is a positive definite matrix. □

We can not claim that the concave condition from this theorem is satisfied for success functions H_i defined by Fill Probability functions F_i without analytical expression for F_i. By definition, $H_i''(q) = qF_i''(q) + 2F_i'(q)$ and F_i is decreasing and convex for $q \in \mathcal{R}_0$. Clearly, the sign of H_i'' can not be determined from these information only. But empirical results gives us good reasons to believe that the functions H_i are indeed concave, at least for q smaller than the average traded volume. Atomic orders are always significantly smaller than the average traded volume (up to one third of that volume) so, it seems reasonable to assume that H_i satisfy the conditions from the previous theorem. One typical empirical example is shown in Figure 4.

4 Two-period model

Time window for execution of atomic order is generally small, say 10 minutes or similar. However if we are buying a liquid but volatile stock we might find that time too long to wait and see if orders will be filled according to our expectations. The market conditions can change significantly and the strategy obtained from (17)-(18) might be subject to re-optimization at certain time point τ within $(0,T)$. On the other hand re-optimization cannot be performed too often because the passive nature of limit orders require some time for them to be realized. Taking into account both possibilities we will present a two-period model without any loss of generality since it could be easily translated into a multi-period model with as many re-optimizations as appropriate.

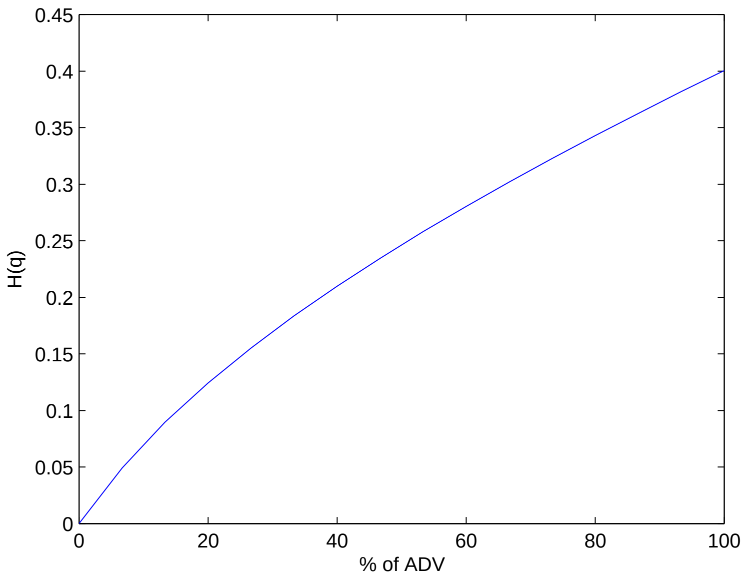

Figure 4: Empirical Success function.

Let $\tau \in (0, T)$ be the point when we start the re-optimization procedure. Clearly market conditions $\mathcal{M}_0$ at $t = 0$ and $\mathcal{M}_\tau$ at $t = \tau$ can differ significantly due to price changes, cancelations, new liquidity arrival, trading activity, announcement of important news etc.

Let $B_0 = \{i_1, \ldots, i_n\}$ be the set of visible bid levels at $t = 0$. The optimal market and limit orders obtained from (17)-(18) at $t = 0$ are denoted by y^0 and x^0, while the initial gain functions for $[0, T]$ are G_i^0.

At $t = \tau$ we know the volume that is already traded so we have to trade some Q^τ, $Q^\tau \leq Q$, within $[\tau, T]$. Also, for all x_i^0 initially posted at bid levels $i \in B_0$, the unfilled amount $\tilde{x}_i$, $\tilde{x}_i \leq x_i^0$ is known. Reasoning the same way as at $t = 0$ we can distribute Q^τ between market and limit orders taking into account the existing limit orders that are still unfilled but potentially progressed in their queues. We can also consider cancelation of initially posted limit orders x_i^0 if $\mathcal{M}_\tau$ is significantly different from $\mathcal{M}_0$ or if the price has moved so the level i is not visible anymore. When canceling unfilled orders we are losing the place in the queue. Placing a new limit order means that we are going to the end of the existing queue. Clearly unfilled order placed at $t = 0$ and a new order placed at $t = \tau$ at the same bid level will have different Fill Probability functions for the same time interval $[\tau, T]$. For the existing but unfilled $\tilde{x}_i$, the Fill Probability function has changed due to the change from $\mathcal{M}_0$ to $\mathcal{M}_\tau$. Therefore, we will have two sets of Fill Probability functions, $F_i^\tau(q)$ for orders placed at $t = \tau$ and $\tilde{F}_i^\tau(q)$, for unfilled orders posted at $t = 0$, both of them depending on $\mathcal{M}_\tau$ and considering time $[\tau, T]$ but depending on the order's queue position.

Furthermore, $\tilde{F}_i^\tau$ will be different from the initial function F_i^0.

Let $\ell_i^\tau,\ i \in B_0$ denote the volume we are keeping at the initial position. Then clearly

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

These orders will have success rate functions

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

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

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

Due to price movement the set of visible bid levels might have changed so let

$$B_\tau = \{k_1, \ldots, k_n\}$$

be the set of visible bid levels at $t = \tau$. If $x_k^\tau, k \in B_\tau$ are new limit orders to be placed at $t = \tau$ then their success functions are

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

while the gain functions are $G_k^\tau(x_k^\tau) = c_k^\tau H_k^\tau(x_k^\tau)$ with

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

Finally let y^τ denote the volume we will trade as market orders in $[\tau, T]$. Then the impact cost with the linear impact function is

$$\pi^\tau(y^\tau) = (\varepsilon + \mu_\tau y^\tau)y^\tau$$

with μ_τ being a stock specific constant dependent on time $T - \tau$. The new residual function is analogously to (13),

$$\rho(l^\tau, x^\tau, y^\tau) = Q^\tau - \sum_{i \in B_0} \tilde{H}_i^\tau(\ell_i^\tau) - \sum_{k \in B_\tau} H_k^\tau(x_k^\tau) - y^\tau. \tag{24}$$

The optimization problem now becomes

$$\min_{l^\tau, x^\tau, y^\tau} \quad \Phi(\ell^\tau, x^\tau, y^\tau) \tag{25}$$

$$\begin{aligned} \text{s.t.} \quad & \ell_i^\tau \in [0, \tilde{x}_i],\ i \in B_0 \qquad (26) \\ & Q^\tau = y^\tau + \sum_{i \in B_0} \ell_i^\tau + \sum_{k \in B_\tau} x_k^\tau \\ & x^\tau, y^\tau \geq 0 \end{aligned}$$

with

$$\begin{aligned} \Phi(\ell^\tau, x^\tau, y^\tau) \quad = \quad & -\sum_{i \in B_0} \tilde{G}_i^\tau(\ell_i^\tau) - \sum_{k \in B_\tau} G_k^\tau(x_k^\tau) + \pi^\tau(y^\tau) + \qquad (27) \\ & \sigma\rho(l^\tau, x^\tau, y^\tau)\sqrt{T - \tau} + \Pi^\tau(\rho(l^\tau, x^\tau, y^\tau)) \end{aligned}$$

and

$$\Pi^\tau(\rho) = (\varepsilon + \eta_\tau \rho)\rho$$

with $\eta_\tau > \mu_\tau$ due to faster execution of the residual at the end of time window i.e., larger traded volume within shorter execution window for the residual ρ.

The problem (24)-(27) has the same structure as (17)-(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 $H_k^\tau, \tilde{H}_i^\tau \in C^2(\mathcal{R}_0)$ and $H_k^\tau, \tilde{H}_i^\tau$ concave for all $k \in B_\tau$ and $i \in B_0$. Then $\nabla^2\Phi(\ell, x, y)$ is a positive definite matrix.*

One important issue deserves additional clarification here. The proposed two-period model is not equivalent to the application of (17) - (18) on consecutive time intervals $[0, \tau]$ and $[\tau, T]$. Re-optimization of the execution trajectory according to (24)-(27) allows an important advantage by the fact that we can keep initially placed orders in the queue if chances of being filled are good enough. Since

$$\tilde{F}_i^\tau(q) > F_i^\tau(q)$$

due to different positions in the corresponding queue it is clear that solving (17)-(18) at $t = 0$ and then (24)-(27) at $t = \tau$ is better than applying (17)-(18) twice due to the passive nature of limit orders and queue positions. Furthermore the fill probability is an increasing function of time. Therefore, overlapping time windows $[0, T]$ and $[\tau, T]$ is preferable over disjoint $[0, \tau]$ and $[\tau, T]$. On the other hand, market orders y^0 and y^τ are always realized according to some predefined schedule, (see [4]), and their executions bear no time risk. So any change between initially planned y^0 and second period y^τ is actually capturing market movements.

As already mentioned, it is quite easy to perform re-optimization procedure as many times as we want within $[0, T]$. We report numerical results for $\tau = T/2$ in the next section. We also tried three-period models but the results made us stick to the initial idea of one re-optimization at $\tau = T/2$. It appears that more frequent re-optimization is actually chasing the high-frequency noise and thus losing the main advantage of this approach: Fill Probability function and combination of market and limit orders.

5 Numerical results

All numerical results presented here are derived from simulations. A simulator was written in MATLAB for this purpose. Within that simulator *fmincon* subroutine was used to solve (17) - (18) and (24) - (27). Since our research topic originated from the dire need of a framework for optimal execution, we have endeavored to be as faithful as possible to the real-time usage of the proposed model. There are no assumptions made in the simulation framework that would prevent deployment to production from being used in actual trading.

Data used in the simulation is European level-2 tick data provided by Reuters. This consists of 5 levels of orderbook depth with consolidated volume on each price level. Data used are for the following five securities: VOD.L, AAL.L, KGF.L, SDR.L and SASY.PA. The period in question is January - March 2008. Simulations are run everyday with continuous tick, i.e. every single tick is considered.

The historic tick database being used provides snapshots of the market every time a change takes place. The simulation process changes between subsequent snapshots to recreate the orderbook. When recreating the orderbook, we maintain the changes to a given price level as a sequence of individual orders. This will effectively evolve into reflecting the size of the individual orders in a given price level.

When the trading models within the simulator harness places an order, the order is added to the end of the queue and tagged. The tag will record the position and quantity ahead. For all subsequent trades on that price level, the quantity ahead is reduced by the traded amount. However, a cancelation may or may not change the ahead quantity as one does not know whether the canceled order was in front or behind our order in question. We choose the worst case scenario to assume that all canceled orders were behind ours if there were any, hence not changing the ahead quantity.

In our simulated orderbook, an order q_i at price level i is filled only when the quantity traded at that price level exceed the ahead quantity and q_i. The task of determining changes to the best bid price due to a cancelation or a trade is very difficult. We use a proprietary data matching filter to re-stream the data in real-time in the correct chronological order and change attribution. The success rate of this filter varies from exchange to exchange. For LSE, 98% of the tick changes are correctly identified and re-streamed. With Euronext for instance, this number is approximately 90%.

The optimization models use a number of different static variables. Three variables of particular importance are:

- Average Daily Volume (ADV)
  ADV is used by the Market Impact Model to measure the relative size of an order. A simple 90 days average is used in this calculation.

- Intraday Volatility
  We use the intraday volatility to estimate the short term volatility risk σ. We calculate this from 90 days of historic data for non-overlapping 15 minutes. The sample of 15-minutely time-of-day sensitive volatility estimates are further interpolated to cater for arbitrary time of day. Return numbers in the volatility calculation are calculated between two mid-prices at the start and end of the 15-minutes time slice.

- Market Impact Model Coefficients
  Based on thousands of actual trades on the stocks in question, we use a method similar to Almgren as discussed above, to estimate the model's

coefficients with a proprietary modification. Nevertheless, following Almgren's algorithm for calibration will also work.

We have only considered order size up to 15% of ADV. Order quantities larger than this will cause significant market impact. The effect of this impact is difficult to quantify. The market impact itself will become non-linear. The excess impact will affect the liquidity arrival pattern in the orderbook. This will further affect other quantitative models such as fill probability, etc. Therefore, although simulated results for larger ADV order will look attractive, not incorporating the significant effects of our trades into the simulation will make the results depart from our aim to be in line with real trading.

We propose a benchmarking scheme that makes a fairer measure taking into consideration the price process, when estimating slippage to the benchmark price. The primary aim of all atomic orders is to get the best possible price within a small window. As such, we define the universal reference price $P_{perfect}$. This reference price is theoretically the best possible that one could have achieved if one had complete foresight of where the market was to trade during the window. With this foresight, the quantity that would not have been filled will be traded using a uniform profile over the entire window.

We introduce two measure, P_B and P_M, to closely reflect market practice. P_B is achieved through an algorithm that always places the entire order on the first bid level and trades the residual as market order at the end, while P_M is obtained from the uniform trajectory of market orders only.

All execution costs are calculated as relative difference between $P_{perfect}$ and the individual algorithm's performance, expressed in basis points ($1bp = 10^{-4}$).

We tested all mentioned algorithms for 5 stocks which cover whole spectrum of liquidity with VOD.L being very liquid, SDR.L very illiquid, AAL.L and KGF.L medium liquid and SASY.PA fluctuating between quite liquid to medium liquid. In terms of volatility, less liquid is usually more volatile so these 5 stocks cover the whole range with SDR.L being the most volatile one. The mean spreads are also quite different, varying from 23bp for SDR.L with standard deviation of 18 bp to 8 bp for VOD.L with standard deviation of 4 bp. The size of spread and its deviation directly influence the gain coefficients in our models.

The results are given in Tables 1-5. We considered 3 months worth of data (January to March 2008), each day sliced into 61 time slots of 8 minutes, from 8.16 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 and first column quantity. The second column gives the mean execution costs of uniform trajectory of market orders, i.e $M = (P_M - P_{perfect})/P_{perfect}$, while in column 3 we have $B = (P_B - P_{perfect})/P_{perfect}$. The cost of optimal strategy coming from (17)-(18) is O_1 given in column four and the cost of two-period optimal strategy (24)-(27), O_2, is reported in column 5. All values are expressed in basis points. The last two columns give the differences between the corresponding strategies. The quality of Fill Probability we are using is illustrated in Figure 5. For 15% of ADV of VOD.L we plot the mean error between forecasted F_i by our model and the

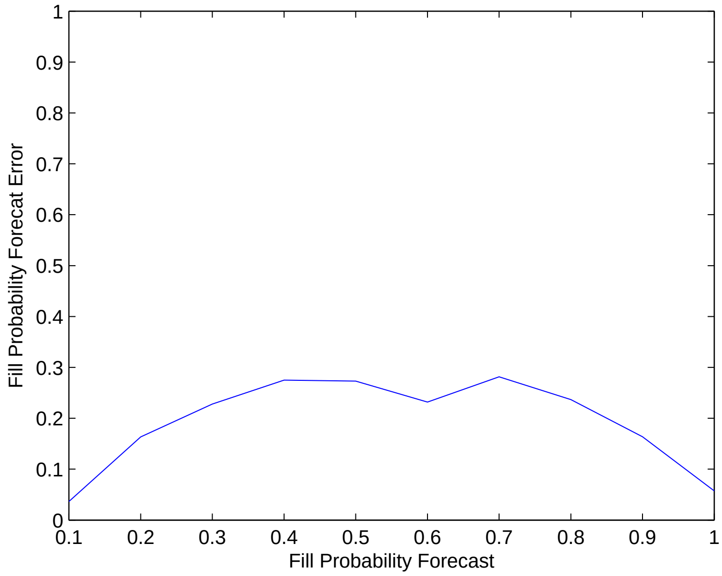

Figure 5: Mean error of the Fill Probability model.

realized fill rate for the whole tested range. The cumulative results for the whole considered period are illustrated graphically at Figure 6 for 10% of ADV for VOD.L.

In addition to the mean execution costs, one is naturally interested in the standard deviation of execution costs. We report these numbers in Table 6 for all considered stocks and 10% of ADV as a representative example of all simulations again comparing all four algorithms. The strategies proposed in this paper have smaller variance numbers and are preferable to the common market practice (algorithms M and B) by these criteria.

The difference between single period model and two-period model is obvious in Tables 1-5. We give more details taking the example of 10% ADV for SASY.PA order as a typical example. All reported numbers are given as a percentage of the initial order size. At $t = 0$ mean values of market and limit orders are $y^0 = 3.7\%$ and $x_1^0 = 64.1\%$, $x_2^0 = 21.5\%$, $x_3^0 = 5.9\%$, $x_4^0 = 0.9\%$, $x_5^0 = 0.4\%$. At $\tau = T/2$ one half of y^0 is realized while the unrealized limit orders were $\tilde{x}_1 = 14.8\%$, $\tilde{x}_2 = 11.9\%$, $\tilde{x}_3 = 3.6\%$, $\tilde{x}_4 = 0.4\%$ and $\tilde{x}_5 = 0.1\%$ with respect to the total order size. The order size for the second period was $Q^\tau = 34.5\%$ of the initial order and that value was distributed as $y^1 = 3.4\%$, $x_1^\tau = 23.1\%$, $x_2^\tau = 5.3\%$, $x_3^\tau = 0.2\%$, while we kept at initial bid positions $l_1^\tau = 2.0\%$, $l_2^\tau = 0.4\%$ and $l_3^\tau = 0.2\%$. Therefore the total amount of cancelations was 28% and new orders account for 28.5% of the initial order size Q with $y^\tau = 3.4\%$. At the end of time window $t = T$, we had average residual size of 8.8% which was executed as a market order within roughly 3

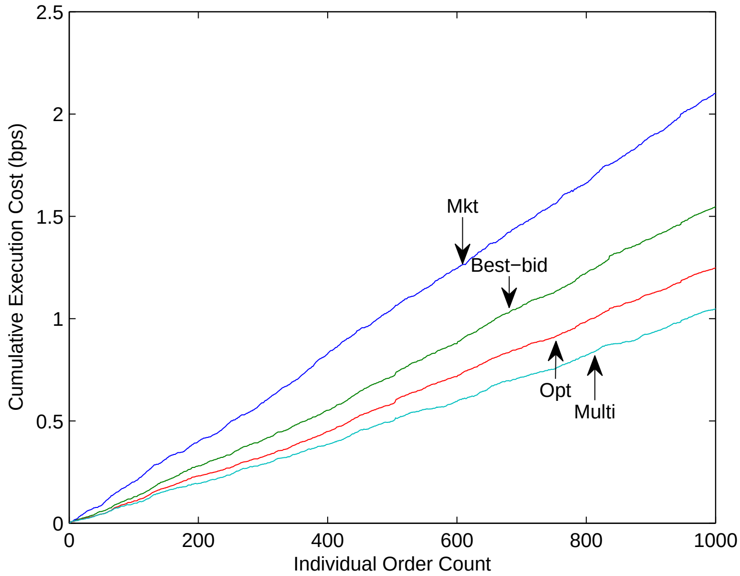

Figure 6: Costs for 10% of ADV through whole tested period, VOD.L.

minutes. Looking at the same example with the single period model we get the same values initially with $y = 7.4\%$. The realized quantities during whole time period $[0, T]$ are different - filled quantity at bid level 1 is 59.9% and then 6.8%, $1.3\%, 0.2\%$ and 0.2% at the lower bid levels. The residual is 24.2%

We can see that both optimization models are not only significantly better than common market practice but are indeed generating distribution of volume between different bid levels and re-optimization procedure leads to new limit orders as well as preserving some initially posted limit orders as expected. The share of market orders is relatively small (7.3% within time frame and 8.8% for residual) in the two period optimization procedure against 31.6% for single period and that is the key reason for small execution costs. Another important observation is the high rate of success of limit orders at lower levels of depth which yields significantly higher gain than putting everything at best bid position. 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.

The same behavior can be seen if we consider the gain achieved with two-period procedure against single period for four stocks but not for SDR.L. For this stock the single period trajectory has the best performance of all considered stocks. However the two-period model performs worse than the single period one for larger orders. The reasons for this behavior are due to high volatility and sparse trading pattern. As a results the Fill Probability model overestimates the real probability for the best bid and at mid point we have large unfilled amount. By re-optimization we are actually chasing the noise since 4 minutes is not an optimal reevaluation point of the market conditions. Therefore we end up sending large 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 effect of volatility disappears and the fill rate is significantly better that the rate within 4 minutes.

% of ADV	M	B	O_1	O_2	$B - O_1$	$O_1 - O_2$
1	13.8	11.4	10.0	9.6	1.4	0.5
2	14.6	11.7	10.3	9.6	1.4	0.7
3	15.4	12.1	10.6	9.7	1.5	0.9
4	16.2	12.5	10.9	9.8	1.7	1.1
5	17.0	13.0	11.1	9.8	1.9	1.3
6	17.8	13.5	11.4	9.9	2.1	1.5
7	18.6	13.9	11.6	10.0	2.3	1.6
8	19.4	14.4	11.9	10.1	2.6	1.7
9	20.2	14.9	12.1	10.3	2.8	1.8
10	21	15.4	12.4	10.5	3.0	1.9
11	21.8	15.9	12.7	10.7	3.2	2.0
12	22.6	16.5	13.1	11.0	3.4	2.1
13	23.4	17.0	13.4	11.2	3.6	2.2
14	24.2	17.5	13.8	11.5	3.8	2.3
15	24.9	18.1	14.1	11.8	4.0	2.4

Table 1: VOD.L

% of ADV	M	B	O_1	O_2	$B - O_1$	$O_1 - O_2$
1	14.9	11.9	10.7	9.2	1.2	1.5
2	16.3	12.4	11.1	9.0	1.3	2.1
3	17.9	12.9	11.6	9.1	1.3	2.5
4	19.5	13.5	12.1	9.1	1.4	3.0
5	21	14.1	12.7	9.2	1.4	3.5
6	22.5	14.7	13.2	9.2	1.5	4.0
7	24.1	15.3	13.7	9.3	1.5	4.4
8	25.7	15.9	14.2	9.4	1.6	4.8
9	27.2	16.5	14.8	9.5	1.7	5.2
10	28.8	17.1	15.3	9.7	1.8	5.7
11	30.4	17.8	15.9	9.9	1.8	6.1
12	31.9	18.4	16.5	10.1	1.9	6.4
13	33.5	19.1	17.0	10.3	2.0	6.8
14	35.1	19.8	17.6	10.5	2.1	7.1
15	36.6	20.5	18.3	10.8	2.2	7.5

Table 2: AAL.L

% of ADV	M	B	O_1	O_2	$B - O_1$	$O_1 - O_2$
1	21.5	16.2	14.7	13.4	1.5	1.3
2	22.1	16.6	15	13.5	1.6	1.6
3	22.8	17	15.3	13.5	1.8	1.8
4	23.5	17.5	15.5	13.6	1.9	1.9
5	24.2	17.9	15.8	13.7	2.2	2.0
6	24.9	18.4	16.0	13.9	2.4	2.1
7	25.6	18.8	16.2	14.0	2.6	2.2
8	26.3	19.3	16.5	14.2	2.8	2.2
9	27	19.8	16.7	14.4	3.0	2.3
10	27.7	20.3	17.0	14.6	3.3	2.4
11	28.4	20.8	17.3	14.9	3.5	2.4
12	28.4	21.3	17.6	15.1	3.7	2.5
13	29.8	21.9	18	15.4	3.9	2.6
14	30.5	22.5	18.4	15.7	4.1	2.7
15	31.2	23.1	18.8	16.0	4.3	2.8

Table 3: KGF.L

% of ADV	M	B	O_1	O_2	$B - O_1$	$O_1 - O_2$
1	16.7	12.0	11.1	9.9	0.9	1.2
2	17.0	12.4	11.0	9.9	1.4	1.1
3	17.4	13.0	11.1	10.1	1.9	1.0
4	17.9	13.6	11.2	10.4	2.4	0.8
5	18.4	14.3	11.4	10.8	2.9	0.6
6	18.9	15.0	11.6	11.3	3.4	0.4
7	19.5	15.8	11.9	11.7	3.9	0.2
8	20.0	16.5	12.2	12.2	4.3	0.0
9	20.6	17.3	12.5	12.8	4.8	-0.3
10	21.2	18.1	12.9	13.4	5.2	-0.5
11	21.7	18.9	13.2	14.0	5.7	-0.7
12	22.3	19.7	13.6	14.6	6.1	-1.0
13	22.9	20.6	14.0	15.3	6.6	-1.2
14	23.4	21.4	14.4	15.9	7.0	-1.5
15	24.0	22.3	14.8	16.6	7.4	-1.7

Table 4: SDR.L

% of ADV	M	B	O_1	O_2	$B - O_1$	$O_1 - O_2$
1	11.7	9.0	8.3	7.2	0.8	1.0
2	13.1	9.7	8.7	7.2	1.0	1.5
3	14.6	10.4	9.2	7.4	1.2	1.8
4	16.0	11.1	9.7	7.5	1.4	2.1
5	17.3	11.8	10.2	7.7	1.6	2.5
6	18.8	12.6	10.8	8.0	1.9	2.8
7	20.2	13.5	11.3	8.2	2.1	3.1
8	21.6	14.3	11.9	8.6	2.4	3.4
9	23.0	15.1	12.5	8.9	2.6	3.6
10	24.4	16.0	13.2	9.3	2.8	3.9
11	25.8	16.8	13.8	9.7	3.1	4.1
12	27.2	17.7	14.4	10.0	3.3	4.3
13	28.6	18.6	15	10.4	3.6	4.6
14	29.9	19.5	15.7	10.9	3.9	4.8
15	31.3	20.5	16.4	11.4	4.1	5.0

Table 5: SASY.PA

*	M	B	O_1	O_2
VOD.L	14.4	12.8	10.3	9.9
AAL.L	18.7	15.4	15.6	11.4
SASY.PA	15.4	14.8	11.5	11.4
KGF.L	21.7	18.1	15.6	15.8
SDR.L	21.3	16.9	12.5	13.5

Table 6: Standard deviation of execution costs

6 Conclusions

Execution of atomic orders is the core element of any algorithm for automated trading in electronic stock exchanges. Given the specification of a trade order (buy or sell, quantity, price and time window for execution) a sequence of atomic orders is generated and executed. The execution of an (atomic) order is causing execution costs and the principal objective in efficient trading is the minimization of these costs. The main characteristic of atomic orders, short time window and relatively small size, allow significant simplifications in modeling complex market conditions such as price trajectory, volatility, market impact etc. Execution costs can be divided into two main categories. Direct costs are predictable and proportional to the transaction value so they are not considered in the models we present. Indirect costs, mainly price impact and volatility, are depending on order type (market or limit) and market conditions as well as execution strategy. Their minimization was the main topic of this research.

We considered an execution strategy for atomic orders consisting of limit and market orders. The key innovation in our approach was the introduction of Fill Probability function for limit orders. A deterministic model with good theoretical properties (positive definite hessian of the objective function) was derived. The optimal execution strategy, obtained from the model, is risk averse and the model is solvable in real time for a large universe of stocks.

Different properties of typically traded stocks (volatility and liquidity) call for re-optimization of the strategy within atomic order execution time. Therefore a two period model was also considered. The model has the same structure as the first one with larger number of variables and is again solvable in real time.

Both proposed models are tested on a representative sample of real trade data from London stock Exchange and Euronext. A simulator is build in Matlab environment and the optimization problems were solved by fmincon subroutine. The optimal strategies obtained from the proposed models are tested against two trading strategies that represent common market practice. The execution costs are calculated as a deviation from an ideal trade - the trade that would be optimal if all market conditions were known in advance. Five stocks with different liquidity and volatility are tested. The optimal strategies are significantly better than common market practice. Further more, they generated distribution of volume between different levels and thus employed fully the advantage of Fill Probability function. The second period execution path, arising from the two-period model, yields new limit orders and preserves some initially posted limit orders. Therefore the reasoning in formulating these models proves empirically good although significant simplifications of the real environment are introduced. Another important property of the considered models is that they generate execution strategy with reasonably small standard deviation of execution costs. This property is highly desirable in real trading.

Acknowledgements. We are grateful to the anonymous referees whose insightful comments and remarks improved the quality of this paper significantly.

References

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

[2] Almgren R., Execution Costs, submitted to Encyclopedia of Quantitative Finance, 2008

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

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

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

[6] 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.

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

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

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

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

[11] 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.

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

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

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

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