UNIVERSITY OF NOVI SAD
FACULTY OF SCIENCE
DEPARTMENT OF MATHEMATICS
AND INFORMATICS

Miles Kumaresan

Optimization of Conditional Trajectories in a Market Place of Multiple Liquidity Pools

- Doctor of Philosophy thesis -

Novi Sad, 2010.

This thesis is dedicated to my family.

Acknowledgement

My profound gratitude to prof Dr Nataša Krejić, for all the help and support during these three years. I could not have chosen a more enthusiastic and committed supervisor.

I would also like to thank prof Marko Nedeljkov and prof Zorana Lužanin for their feedback on the first draft of this thesis.

Thanks are due to my parents for their excellent guidance, in both education and in life, and also to Periamma.

Sincere thanks to my dearest wife Elizabeth, for her unflinching support of all my obscure endeavors over the years and in helping me maintain focus.

Finally, to my son Dominic who is too young to understand why all of my already limited spare time could not be dedicated to playing.

§ § §

Abstract

Algorithmic Trading, also known as Algorithmic Execution, is the automated process of trading exogenous orders in electronic (stock) exchanges. It became widely available to all market participants over the last decade. The execution efficiency was noted early on by market participants and algorithmic trading rapidly grew to become a widely utilized product globally within Equites markets.

There are many aspects to algorithmic trading that make it attractive. As a user, one has the ability to accurately specify the desired execution profile with an indicative notion of the associated risk. Algorithmic trading consists of a whole range of standard algorithms to mimic mainstream execution styles as well customizable algorithms to to suit individual's needs.

Execution has become a sensitive area where fund managers are consciously looking to reduce cost. In capital markets where even marginal competitive edge by one institution is rewarded with disproportionately large profits, this is an effort worth pursuing. Algorithmic Trading being one of the largest invested technological arms races in Wall Street today is yet another evidence of its importance. Furthermore, having had the privilege of working for a range of large investment banks, the insight into current common practice legitimizes our quest for finding a mathematical solution for Optimal Execution for Atomic Orders.

Execution itself is an exceptionally complex problem, consisting of many uncertain factors. Therefore, modeling most of these uncertain factors is a fundamental part of algorithmic trading. At present, there exists no known formalism for the optimal execution of atomic orders - the fundamental building blocks of all algorithmic orders. At its core, like most problems in capital markets, the problem is that of finding the right balance between desired reward and associated risk.

On one hand, one could choose a low risk option where one does not want to miss the prevailing market price and is willing to pay a premium to lock in this price by trading at market. Alternatively, one could wait a short period in the hope of getting a better price than the currently prevailing price. This latter option runs the price-volatility risk of not getting filled - partially or completely. If one does not get the entire order filled, the residual order has to trade at a worse price. This seemingly simple trade-off in finding the optimal weighting of the gain of waiting, the price volatility during the waiting period and the premium to pay for the instantaneous trading opportunity

is riddled with intrinsic complexities surrounding primitive order properties. The emergence of multiple trading venues has exacerbated this complexity significantly. In the latter case, a security's liquidity is fragmented over multiple liquidity pools or venues. Therefore, one is no longer faced with whether to trade instantly or wait for a better price, but has to consider which venue to place passive orders in. Furthermore, if the market conditions change adversely after placing an order in a given venue or move more favorably in another venue, should one move this order to another venue and if so, to which one and at what price level.

The primary objective in execution is to achieve the most efficient price. We propose two optimal strategies for the execution of atomic orders based on minimization of impact and volatility costs, in both single and multiple market environments. 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. Finally, we consider how the model that allows re-optimization perform in a multiple trading venue environment. In all cases, a combination of market and limit orders are 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 four optimization models we are discussing in this thesis. Under certain conditions the objective functions of all 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.

We first approached the simplest of problems, namely single market and single period optimization. We were able prove that the problem at hand could be optimized with an SQP variant procedure. This was followed by formalizing the objective function. Next, we extended the model to deal with multiple re-optimization. In our example, we re-optimized once only, however the procedure is general can be re-optimized more frequently. Finally, we extended the general multi period problem to an environment consisting of multiple trading venues. This last mentioned problem is a bi-level nonlinear optimization problem for modeling risks and gains of the decision variables and the residuals.

The notation adopted is giving in chapter 1 together with definitions and theorems. Chapter 2 introduces the key concepts related to financial markets. The market structures are explained in the context of the research. Chapter 3 introduces the whole notion of algorithmic trading and the key problems associated with order execution. The key innovation, Fill Probability, is explored in detail with respect to its functional properties. Constrained non-linear problems are covered in Chapter 4, with particular emphasis on SQP method as it was used in our optimizations. In particular, fmincon() subroutine in Matlab was used in our optimization.

The nucleus of this thesis are presented in Chapters 5 and 6, single-market and multiple-market respectively. These two chapters contain the original work presented in this thesis - a novel solutions for hitherto non-existing formulation for optimal execution of orders in the financial market. The findings detailed in chapter 5 is published in the Journal of Computational Optimization and Applications (Kumaresan and Krejić [38]). The findings in chapter 6 are presented in Kumaresan, Krejic [39].

Chapter 7 defines the parameters and methodology of the optimization conducted to empirically prove our model. Because the models presented here are intended for live-trading, there are no assumptions made that would prevent it from trading. In fact, the models developed in this thesis was deployed to live trading at a global proprietary trading house. Chapter 8 contains the numerical results of our models clearly showing the performance characteristics of the different models. The last chapter contains the list of references.

Key words: nonlinear programming, convex programming, optimal execution strategy, multiple trading venues, algorithmic trading

Apstrakt

Algoritamsko trgovanje, poznato i kao automatsko izvršenje, je automatizovani proces izvršavanja naloga na elektronskim berzama (berzama akcija). Tokom poslednje decenije ovaj vid trgovanja je postao dostupan širokom krugu učesnika na tržištu. Efikasnost ovog načina izvršenja je primećena i ranije od strane učesnika na tržištu i algoritamsko trgovanje se brzo raširilo do globalno rasprostranjenog proizvoda na trštima kapitala.

Algoritamsko trgovanje je atraktivno po mnogim osobinama. Korisnik ima mogućnost da precizno definivsse željeni način izvršenja uz inidikativni pokazatelj rizika. Algoritamsko trgovanje sadrži čitav niz standardizovanih algoritama koji imitiraju glavne tipove izvršenja kao i mogućnost prilagodjavanja algoritama pojedinačnim potrebama korisnika.

Izvršenje je postalo osetljiva oblast u kojoj fond menadžeri neprekidno pokušavaju da smanje troškove. Na tržištima kapitala, na kojima čak i marginalna prednost u načinu izvršenja dovodi do nesrazmerno velikog profita, trud uložen u smanjenje troškova je veoma isplativ. Algoritamsko trgovanje je trenutno jedna od najve¸ih tehnoloških trka na Wall Street-u što govori o njegovom značaju. Imajući privilegiju rada za nekoliko velikih investicionih banaka i stečeno znanje o sadašnjoj standardnoj praksi u izvršenju naloga, čini pokušaj nalaženja matematičkog rešenja za problem optimalnog izvršenja atomskih naloga, legitimnim i značajnim.

Izvršenje zavisi od mnoštva slučajnih faktora i predstavlja ekstremno složen problem. Zbog toga je modeliranje slučajnih faktora fundamentalni deo algoritamskog trgovanja. U ovom momentu ne postoji formalni okvir za optimalno izvršenje atomskih naloga - koji su fundamentalni element svih algoritamskih naloga. U suštini, problem se svodi na nalaženje ravnoteže izmedju željenog prinosa i pridruženog rizika, , kao i većina problema na tržištu kapitala.

Sa jedne strane, moguće je izabarti opciju niskog rizika u kojoj ne želimo da propustimo preovladjujuću cenu i spremni smo da platimo premiju da bi osigurali tu cenu trgujući market nalogom. Druga alternativa je da se sačeka kratko vreme, davanjem limit naloga, u nadi da će se postići povoljnija cena. Ova opcija nosi rizik cena-volatilnost koji može dovesti do delimičnog ili potpunog neizvršenja naloga. Ako ceo nalog nije izvršen onda se rezidualni nalog izvršava po lošijoj ceni. Ova naizgled jednostavna odluka, koja se donosi pri odredjivanja optimalnog perioda čekanja radi dobitka, je zapravo veoma složena zbog volatilnosti cene tokom perioda čekanja i premije koja

se plaća za trenutno izvršenje naloga. Razvoj trgovanja na više tražišta istovremeno je značajno povećao kompleksnost problema. U ovom slučaju je likvidnost hartije rasporedjena na više tržišta istovremeno. Usled toga, dilema nije samo da li nalog izvršiti odmah ili čekati bolju cenu već se mora odlučiti i na kom tržištu će se plasirati pasivni nalozi. Sem toga, ukoliko se tržišni uslovi promene u nepovoljnom pravcu nakon što je nalog dat na jednom tržištu, ili se pak na drugom tržištu uslovi promene u povoljnom pravcu, postavlja se pitanje da li premestiti nalog, i ako se vrši promena, kako je izvršiti - aktivno ili pasivno i na kom cenovnom nivou.

Osnovni cilj u izvršenju je postizanje najefikasnije cene. Ovde su predložene dve optimalne strategije za izvršenje atomskih naloga zasnovane na minimizaciji troškova impakta i volatilnosti u slučaju jednog tržišta i više tržišta. Prva posmatrana strategija je zasnovana na relativno jednostavnom nelinearnom optimizacionom modelu, dok druga dozvoljava reoptimizaciju u nekom trenutku unutar zadatog vremenskog intervala izvršenja. Konačno, posmatran je model koji dozvoljava reoptimizaciju u okruženju sa više tržišta. U svim slučajevima koristi se kombinacija market i limit naloga.

Glavna inovacija u našem pristupu je uvodjenje Fill Probability funkcije koja omogućava kombinaciju market i limit naloga u sva četiri modela diskutovana u ovoj tezi. Pod odredjenim uslovima funkcije cilja svih posmatranih problema su konveksne te se mogu primeniti standardni metodi optimizacije. Efikasnost predloženih strategija je testirana u odnosu na dve representativne strategije, koje predstavljaju uobičajenu praksu, na realnom uzorku podataka sa tržišta.

Prvo je posmatran najjednostavniji model - jedno tržište i jedna vremenski period. Pokazano je da se ovakav problem može rešiti jednom varijantom standardnog SQP metoda. Zatim je model proširen na model sa više perioda tako da dozvoljava reoptimizaciju. U našim primerima reoptimizacija se radi samo jednom u svakom vremenskom intervalu ali je procedura generalna i može se primeniti proizvoljan broj puta. Na kraju je model za više perioda proširen na okruženje sa više tržišta. Ovaj poslednji model zahteva rešavanje bi-level problema nelinearne optimizacije pri modeliranju rizika i dobiti za osnovne i rezidualne promenljive.

U prvom poglavlju je dat pregled oznaka, definicja i teorema koje su korišćene u radu. U glavi 2 su uvedeni osnovni pojmovi finansijskih tržišta od značaja za posmatrani problem. Koncept algoritamskog trgovanja i opis osnovnih pojmova su dati u glavi 3. Glavna inovacija u našem pristupu, Fill Probability funkcija, je detaljno objašnjena. Nelinearni problemi sa

ograničenjima su razmatrani u glavi 4, sa posebnim naglaskom na SQP metode koje su korišćene u daljem radu. U numeričkim eksperimentima je korišćena Matlab funkcija fmincon().

Suštinski doprinos teze je dat u poglavljima 5 i 6, za jedno tržište i za više tržišta redom. Ova dva poglavlja sadrže originalne rezultate ove teze - formulaciju (ranije nepostojeću) problema optimalnog izvršenja na finansijskim tržištima i njegovo rešenje. Rezultati iz poglavlja 5 su publikovani u Journal of Computational Optimizationa nd Applications (Kumaresan and Krejić, [38]). Rezultati iz poglavlja 6 su dati u Kumaresan, Krejić [39].

Poglavlje 7 definiše parametre i metodologiju koja je primenjena u optimizaciji pri empirijskom testiranju modela. Kako su predloženi modeli namenjeni stvarnom trgovanju nisu uvedene nikakve pretpostavke koje bi onemogućavale direktnu implementaciju u trgovanju. Zapravo, jedan od modela u tezi iz ove teze se primenjuje u jednoj berzanskoj kući koja trguje globalno. Poglavlje 8 sadrži numeričke rezultate za posmatrane modele koji jasno pokazuju njihove karakteristike. Poslednje poglavlje sadrži pregled korišćene literature.

Klučne reči: nelinearno programiranje, konveksno programiranje, optimalna strategija izvršenja, više tržišta, algoritamsko trgovanje.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

Mathematical approaches to finance are evident in early research from the 50s by Harry Markowitz in formulating the Modern Portfolio Theory, Bill Sharpe and others with their work on Capital Asset Pricing Model - to become one of the foundations of financial economics. Significant contributions were made by Black and Scholes in the early 70s with their ground breaking pricing model for financial options. While, for instance, complex derivatives pricing is today a well explored area, a large proportion of the International Capital Markets are still in their infancy. As such, the use of complex mathematical methods in these areas is a relatively rare phenomenon. Most real world quantitative finance problems involve working with high levels of uncertainty - hence requiring complex mathematical models. However, given the prerequisites of a diverse and in-depth domain knowledge a priori by the Financial Engineers, a somewhat rare combination of skillset, these domains in general remain less exploited. The outcome of this shortcoming is, unsurprisingly, the wide usage of rudimentary rules of thumb based on simple observations. One of the important areas in capital markets that fall into this category is that of Automated Execution of orders of securities at electronic exchanges.

Historically, transactions at listed securities exchanges, like stock or futures exchanges, were conducted by brokers. On receipt of an instruction to transact on behalf of a client, the broker would execute the order manually using a dedicated computer terminal, in a piece-wise manner over a longer period. With the major modernisation of exchanges in the late 90s, member organisations of these exchanges were provided electronic pipes, open interfaces, thus empowering the members to develop their own automatons to execute some of the simpler manual orders. Early to capitalise on this emerging technology were the Quantitative Trading Groups within investment banks. These groups made their investment decisions based on quantitative models that recommended buying and selling timing of financial securities. Being very quantitative in their approach to investment decisions, the automation allowed them to expand employing their trading models to a large number of securities and markets. Beside removing the manual execution which was a logistic overhead, the automation proved to be significantly more cost efficient, facilitating a new breed of trading models with, small profit margin strategies, to be employed.

During the years 2000-2003, the rudimentary automation rapidly got adopted across the wider community in the equities market. By 2008, the early-stage electronic assistants to human operators had become a major discipline on their own right. Better known as Algorithmic Trading, this is one of the largest invested technological arms races in Wall Street today. The growth of electronic execution of orders in the major stock exchanges in Europe and the US has been tremendous in the recent years. Stock exchanges across the globe have had to change their business models to facilitate Algorithmic Trading. In some exchanges, 80% of the transactions are originated by Algorithms. It is estimated that 40% of the volume traded in 2008 across Europe and the US were algorithmic trades. With such a demand for electronic trading, new execution venues mushroomed to accommodate the diverse requirements of execution. These range from alternative exchanges to traditional exchanges, to new dark liquidity pools, block trading systems, anonymous crossing engines, liquidity aggregators, and so forth.

There are many reasons for the rapid growth in Algorithmic Trading. Its emergence coincided with the arrival of smart technology to facilitate routing orders to multiple venues. Execution is fundamental to most sell side institutions as this is usually agency business and hence revenue is based on commission / fee as opposed to revenue generated from taking trading risk. By using mathematical models to formalise the execution problem, the quality of execution improved significantly. Automation allowed algorithms to be customizable, to meet the specific needs of the client and also allowed them to assess their risks prior to entering into a transaction. The latter effectively gave the client better control over their orders. Of the dozen or so classes of execution algorithms, 90% of the algorithmic market share is attributed to three types of algorithms only (Volume Weighted Average Price - VWAP, Participation and Implementation Shortfall).

The fabric of algorithms used in algorithmic trading is deeply entrenched in multiple disciplines, such as trading, market micro structures, mathematics, high frequency data, forecasting, stochastic problems, etc. Therefore, creating advanced algorithms is a task too difficult even for most of the current breed of financial engineers. This is observable from the share performance of the algorithms offered by the different financial institutions (sell side).

Algorithms remain relatively basic. Though the behavioural properties of securities in the equites market, which consists of thousands of securities, cannot be characterised by mere standard deviation of price and average daily volume, they often are. Similarly, the Probability of Fill of an order placed at the best bid/ask price, for a vaguely defined short period, is not 60% for all stocks across all times of the day, contrary to popular belief. The practice of using static properties of stocks to estimate expected Market Impact will always either over- or under estimate the true market impact as market impact is a dynamic factor. At present, there exists no known formalism for the Optimal Execution of Atomic Orders - the fundamental building blocks of all algorithmic orders.

In the absence of truly innovative differences among the top tier institutional competitors in the Algorithmic Trading space, the ranking of the different providers of algorithms is somewhat ambiguous. This is party due to the inherent difficulties in accurately measuring performance of the different algorithms by the buy-side users. The recent emphasis and investment by sell side in algorithmic trading has been on the infrastructure technology to reduce the latency to the market - the time it takes to send an order to an execution venue. Only a few years ago, reducing the latency from 100+ milliseconds down to single digit milli-seconds delay was much welcomed. However, the reduction in latency does not necessarily translate into proportionate betterment of execution quality.

With the performance and latency of most serious algorithm providers being comparable, once again the focus is starting to look towards innovation in developing more complex variations of the currently popular models. We are of the belief that this innovation should start at the heart of an algorithm. Regardless of the objectives of the algorithm used to execute an order, all orders are decomposed into a sequence of atomic orders. How well these atomic orders are executed will directly translate into how well the overall objectives of the algorithms in question are meet. Therefore, a poor atomic order execution is guaranteed to have poor overall performance.

In capital markets where even marginal competitive edge by one institution is rewarded with disproportionately large profits, this is an effort worth pursuing. Furthermore, having had the privilege of working for a range of large investment banks, the insight into current common practice legitimizes our quest for finding a mathematical solution for Optimal Execution for Atomic Orders.

1.2 Notation

$\mathcal{N}$ - set of natural numbers

$\mathcal{R}$ - set of real numbers

$\mathcal{R}_+$ - set of non-negative real numbers

$\mathcal{R}^n$ - set of real n-dimentional vectors

$\mathcal{R}^{m \times n}$ - the space of all m-by-n real matrices

$x, y \ldots$ n-dimensional vectors with components $x_i, i = 1, \ldots, n$ i.e. $x = (x_1, x_2, \ldots, x_n)^T$

$A, B \ldots$ - set of matrices

$A = \left[a_{ij}\right]_{nxm}$ - an m x n matrix with elements $a_{ij}, i = 1, \ldots, n,\ j = 1, \ldots, m$.

A^T - the transpose of the matrix A

A^{-1} - the inverse of the matrix A

I - the identity matrix

$e^1, ..., e^n$ - the coordinate vector of $\mathcal{R}^n$

$A = diag(a_1, a_2, ..., a_n) -$ diagonal matrix, *i.e.* $a_{ii} = a_i, \quad i = 1, .., n$

$a_{ij} = 0, \ i \neq j, \ i, j = 1, .., n$

$\|.\|$ - vector norm

$\|.\|$ - matrix norm

$\|x\|_1 = \sum_{i=1}^n |x_i| - l_1$ norm

$\|x\|_2 = \left(\sum_{i=1}^n |x_i|^2 \right)^{\frac{1}{2}} - l_2$ norm

$\|x\|_\infty = \max_{1 \le i \le n} |x_i| - l_\infty$ norm

$\|A\|_F = \left(\sum_{i=1}^n \sum_{j=1}^n |a_{ij}|^2 \right)^{\frac{1}{2}}$ - Frobenius norm

1.3 Overview of Definitions and Theorems

Definition 1 *Matrix* $A \in R^{m \times n}$ *is called:*

- *symmetric if* $A^T = A$;

- *positive definite if* $x^T Ax > 0$ *for all* $x \in \mathcal{R}^n$;

Definition 2 *We say that mapping* $F : D \subset \mathcal{R}^n \to \mathcal{R}^n$ *satisfies a Lipschitz condition at* $x \in D$ *with constant* $L > 0$ *if*

$\|F(x) - F(y)\| \le L\|x - y\|$, *for all* $y \in D$

Definition 3 *A mapping $F : D \subset \mathcal{R}^n \rightarrow \mathcal{R}^n$ is Frechet-differentiable at $x \in int(D)$ if there is a linear operator $A \in \mathcal{R}^{n \times m}$ such that*

$$\lim_{h \to 0} \frac{\|F(x+h) - Fx - Ah\|}{\|h\|} = 0. \tag{1.1}$$

The unique linear operator A for which (1.1) holds is denoted by $F'(x)$, and is called the F-derivative of F at x. The limit in (1.1) is independent of the particular norm on $\mathcal{R}^n$. The matrix representation of $F'(x)$ is given by the Jacobian matrix, $\mathcal{J}(x) = [\frac{\partial f_i}{\partial x_j}]_{ij}$

Lemma 1 *[61] Assume that $F : D \subset R^n \rightarrow R^m$ is differentiable on a convex set $D_0 \subset D$. Then, for any $x, y \in D_0$,*

$\|F(y) - F(x)\| \leq \sup_{0 \leq t \leq 1} \|\mathcal{J}(x + t(y-x)\|.\|y - x\|.$

Lemma 2 *[61] Let $F : D \subset R^n \rightarrow R^m$ be continuously differentiable on a convex set $D_0 \subset D$. Suppose that*

$\|\mathcal{J}(y) - \mathcal{J}(x)\| \leq \gamma(x)\|y - x\|$ *for all* $y \in D_0$.

Then,

$\|F(y) - F(x) - \mathcal{J}(x)(y - x)\| \leq \frac{1}{2}\gamma(x)\|y - x\|^2$, *for all* $y \in D_0$.

Lemma 3 *[61] Let a mapping $F : D \subset R^{n \times n} \rightarrow R^{n \times n}$ be continuously differentiable in an open convex set D and F' satisfies a Lipschitz condition at $z \in D$. Then*

Then,

$\|F(x) - F(y) - \mathcal{J}(z)(x - y)\| \leq L \; max\{\|x - z\|, \|y - z\|\}\|x - y\|$, *for all x and y in D.*

Definition 4 *Let $\{x^k\}_{k=0}^{\infty} \subset R^n$ and $\lim_{x \to \infty} k^k = x^*$. Then,*

- $x^k \to x^*$ *q-linearly if exists* $\sigma \in (0,1)$ *such that*

  $\|x^{k+1} - x^*\| \leq \sigma \|x^k - x^*\|,$

- $x^k \to x^*$ *q-superlinearly if*

  $\lim_{k\to\infty} \frac{\|x^{k+1}-x^*\|}{\|x^k-x^*\|} = 0,$

- *q-superlinearly with q-order* $\alpha > 1$ *if exists* $K > 0$ *such that*

  $\|x^{k+1} - x^*\| \leq K \|x^k - x^*\|^\alpha,$

- *q-quadratically (q-superlinearly with q-order 2) if exists* $K > 0$ *such that*

  $\|x^{k+1} - x^*\| \leq K \|x^k - x^*\|^2,$

for k sufficiently large.

Definition 5 *For* $f : R^n \to R$ *and* $x \in R^n$ *we define the gradient as*

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \cdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \tag{1.2}$$

with

$\frac{\partial f}{\partial x_i} = \lim_{\varepsilon \to 0} \frac{f(x+\varepsilon c_i)-f(x)}{\varepsilon}.$

The Hessian is defined as

$\nabla^2 f(x) = [\frac{\partial^2 f}{\partial x_i \partial x_j}]$

Definition 6 *For* $f : R^n \to R$ *and* $p \in R^n$ *the directional derivative of* f *in the direction of* p *is given by*

$D(f(x);p) = \lim_{\varepsilon \to 0} \frac{f(x+\varepsilon p)-f(x)}{\varepsilon}$

If f is continuously differentiable then

$D(f(x);p) = \nabla f(x)^T p.$

1.4 Chapter outline

The notation adopted is giving in Chapter 1 together with definitions and theorems. Chapter 2 introduces the key concepts related to financial markets, in particular relating to market structures and constructs. Chapter 3 introduces the whole notion of algorithmic trading and the key problems associated with order execution. Constrained non-linear problems are covered in Chapter 4, with particular emphasis on SQP method as it was used in our optimizations.

Chapters 5 and 6 essentially constitute the nucleus of this thesis. These contain the original work that propose novel solutions for hitherto non-existing formulation for optimal execution of orders in the financial market (in the public domain). They cover Single Market Model and Multi Market model respectively. The findings detailed in chapter 5 is published in the Journal of Computational Optimization and Applications (Kumaresan and Krejić [38]). The findings in chapter 6 are presented in Kumaresan, Krejic [39].

Chapter 7 defines the parameters and methodology of the optimization conducted to empirically prove our model. Chapter 8 contains the numerical results of our models clearly showing the performance characteristics of the different models.

Chapter 2

Market Structure

Transactions in securities can be carried out in many different ways. Traders at institutional investment houses, regulated by the financial services regulator, could trade with each other via telephone, through exchanges, through regulated liquidity pools - internal or external, and others. Discussed here are three key venues that account for significant proportion of market share.

2.1 Evolution of Exchanges

In the context of stock markets, an exchange is a corporation which provides a trading facility for member organizations to trade shares in its listed universe of stocks. The non-member organisations access an exchange via intermediaries, such as stockbrokers. The exchange being a meeting place of parties interested in buying and selling securities, transactions are the result of supply and demand.

The first major change was the move away from the pits (floor trading) with the introduction of computerised order entry terminals and electronically matching buyers and sellers. The majority of the floor traders could not adjust to the new rules of the market place when the competitive edge of the floor vanished, and was replaced with a fairer playing rule applying to all participants. The order matching process in a stock exchange is based on price priority and for where there are many participants at a given price, the priority will be in the order of arrival. As a result, the floor traders mostly faded away and a new breed of traders/executioners emerged to occupy the vacuum left by the floor traders. A decade ago, the stock exchange went through yet another major transformation and offered all member organisations electronic pipes - providing them the opportunity to develop their own order entry systems. The first usage of this open interface was in the form of simple automation of simple repetitive execution tasks usually carried out by junior traders. Seing the advantage of this automation, the usage spread with a speed beyond most market practitioners' expectations.

In the early part of the millennium, member organisations started developing more complex trading algorithms to execute complete orders autonomously from start to finish. In general, there were only advantages to be gained with automation, in the form of lower slippage to benchmark and lower staff requirement. The execution performance improved significantly and became more consistent than achievable by human operators on average for small to medium sized orders. The transition from pit trading to electronic execution has also significantly affected the nature of liquidity and the market microstructure in general.

An exchange has other roles than merely the mechanical facilitation of transaction among buyers and sellers. Among other roles, corporate governance is perhaps one of the most important roles of an exchange.

2.2 Special Venues

Popularly called Dark Pools, these are essentially a stock exchange like matching engine without visibility of other participants. The primary purpose of dark pools of liquidity is to have anonymity of the participant's intention in terms of the price and quantity they want to transact and whether they are a buyer or a seller. This information is important when a trader has a large order size to execute as this would usually impact the market cost. Therefore, in order to not give away sensitive infirmation to the marketplace, anonymity is favored by large players.

Dark pools are becoming common among all major investment banks. In the interest of avoiding market impact and also to reduce the transaction fee payable to exchanges, any orders worked on behalf of a client (internal or external) will be matched off in a dark pool at a price no worse than the current market price. In a large volume business, the small savings result in profits of tens of millions of dollars.

The success rate of dark pools and that of crossing engines are dependent on the careful balance of liquidity makers (providers) and liquidity takers.

Yet another class of anonymous (and very valuable) form of trading venue are crossing engines. Crossing engines match off larger market participants (typically) - wanting to transact without causing excess market impact. As such, anonymous orders are submitted by traders and crossing usually takes place at pre-scheduled intervals (typically hourly). Each order may or may not have a price limit and the matching algorithm employed will match-off orders such that maximum amount is transacted while adhering the constraints imposed at the individual order level. As a result, at the end of the matching process, there may be many unfilled orders. This can also be the result of large imbalances in supply and demand. Examples of major crossing engines are ITG's Possit, NYFix Millennium, Liquidnet and Pipeline Trading Systems.

2.3 Successful Venues

A successful venue is one in which there is a good balance between providers of liquidity and takers of liquidity. If there were fewer liquidity takers (aggressive orders) than liquidity makers (passive orders), then we will have a situation where the majority of orders are dormant and the market will remain static in this particular venue. If the roles were reversed, the venue would not be of interest to liquidity takers as only a small quantity is available to trade without causing excess market impact. Venues with visible orderbook (visibility of all orders at the venue) will allow traders to react accordingly to realise their order execution aims. However, if the lined up buyers and sellers are not visible, then the hit rate could be very low. This could potentially lead to the death of such venues.

The above dilemma of balancing providers and takers of liquidity is the primary cause for the formation of "aggregators", ie. umbrella organisations that work as liquidity pool gateways by amalgamating liquidity in diverse pools into one portal. Although aggregators are attractive compared to the alternative of not having any, the entire arena of multiple liquidity pools is in the early stages of evolution. The well known Virt-X exchange, recently renamed SWX, was the first cross border trading platform for pan-European blue chip stocks. Formed in 2001, it was not a big success until recently. This is partly due to the lack of Smart Order Routing (SOR) systems until recently. It is also partly due to the lack of end user sophistication at the time. In a similar light, the complexity of the liquidity pool cannot be significantly more complex than the technology commonly available and the state of the end user's ability to utilize the complex functionality that is offered.

An attractive venue is not all about being visible or dark. The visibility should be a function of one's intention of engagement. For instance, one of the key concerns with dark liquidity pools is the use of "pinging", a liquidity spotting method utilized by various market participants, by sending very small share quantities to many venues as market orders. The fills from the different dark venues would provide some insight into where there may be a large availability of liquidity. However, traders placing orders in dark pools do not want to be pinged because they have large orders to transact and are looking for traders with similar intentions. As such, pinging will make the intentions of dark pool participants visible, hence they can be exploited.

It is our belief that an attractive venue is one in which there is a continuity in the level of visibility to suit a wide range of market participants. However, each order entering dark pools will have addition attributes, specifying factors such as clip size for instance, the minimum quantity the trader is willing to trade with anyone, effectively shielding one from being pinged by small orders. Other attributes could be laddering of clip sizes, where the trader may have different limit prices attached to different clip sizes. This would allow one to transact even very large order sizes with differing prices. Therefore, a composition of liquidity makers with large and varied orders can co-exist in a single venue and could potentially facilitate larger hit rates. As in the case of Virt-X, unless there are enough user wanting to take advantage of the sophistication of complex venues, there will not be much business in these venues.

2.4 Market Micro-structure

Whether or not research in quantitative finance is about market microstructure, this topic will enter the scene directly or indirectly. Market microstructure has many definitions without being any one specific thing. Actually, the use of the term "Market microstructure" in today's finance is somewhat ambiguous. The National Bureau of Economic Research's market microstructure research group [34] defines it as being "devoted to theoretical, empirical, and experimental research on the economics of securities markets, including the role of information in the price discovery process, the definition, measurement, control, and determinants of liquidity and transactions costs, and their implications for the efficiency, welfare, and regulation of alternative trading mechanisms and market structures". A better definition is provided by O'Hara [47], "Market microstructure studies the effects of market structure and individual behaviour on the process of price formation". Madhaven [41] nicely defines market microstructure as "the area of finance that studies the process by which investors' latent demands are ultimately translated into price formation and volumes".

The term Market Microstructure, coined by Garman [26], writes:

"[W]e depart from the usual approach of the theory of exchange by (1) making the assumption of asynchronous, temporally discrete market activities on the part of market agents and (2) adopting a viewpoint which treats the temporal microstructure, i.e., moment-to-moment aggregate exchange behaviour, as an important descriptive aspect of such markets."

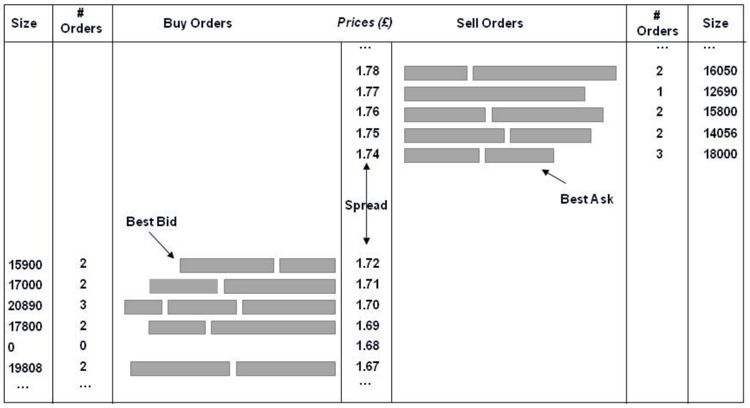
Although our focus is not on market microstructure per se, the factors affecting price formation is important to our later discussions.

2.4.1 Orderbook

At the core of any securities exchange is the notion of an *Orderbook*. A market place consists of two sets of traders, buys and sells of securities. The two camps of traders have similar but opposing objectives for a given security. The collective representation of all buyers and sellers of a given security is the orderbook. An orderbook is therefore a price-ordered list of buyers and sellers on two opposing sides. The basic building block of an orderbook is a Price-Quantity pair. Orders in an orderbook in the stock market, follow a FIFO queuing convention. That is, there an be multiple orders in a given price level. Priority is assigned by price and then by arrival time of the order. Figure 2.1 illustartes an orderbook. There are instruments such as Euro Dollar or Corn futures where a pro-rated order filling rule is used in combination with FIFO. A trade takes place when there is an overlap in supply and demand.

2.4.2 Multiple Orderbooks

In a market place of multiple trading venues, there are multiple orderbooks for one and the same security as shown in figure 2.2. As can be seen, the two exchanges or markets do not have the same order quantity or prices, although



Size	# Orders	Buy Orders	Prices (£)	Sell Orders	# Orders	Size
			...		...	...
			1.78		2	16050
			1.77		1	12690
			1.76		2	15800
			1.75		2	14056
			1.74		3	18000
15900	2		1.72			
17000	2		1.71			
20890	3		1.70			
17800	2		1.69			
0	0		1.68			
19808	2		1.67			
...	...		...			

Best Ask
Spread
Best Bid

5

Figure 2.1: A typical orderbook in a single-market. Tipična knjiga naloga na jednom tržištu

they are usually arbitrage free (where one could 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 or ask. Another point to be noted from 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.

2.4.3 Bid/Ask Spread

The highest buying price and lowest selling price is called the *touch price*. This is the major divide between buyers and sellers. The most aggressive buying price is often referred to as best bid (denoted in this work as b_1), and similarly for ask. The difference between the best bid and best ask, ie. the spread, is quoted in basis points (bps) in Europe and Asia (1 basis point is 1/100 of a percent) and typically in cents/share or fractional cents/share in the US.

The magnitude of spread has been well studied in Demsetz [18], Tinic

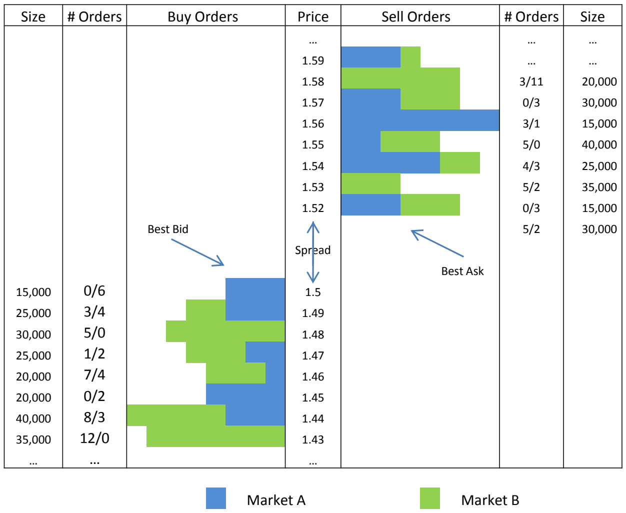


Figure 2.2: A typical orderbook in a multi-market environment. Tipična knjiga naloga u okruženju sa više tržišta

[54], O'Hara and Oldfield [48], Stoll [52], Amihud and Mendelson [6], Roll [50], Ho and Stoll [31] and others. However, it should also be noted that price formation in today's market place is significantly different to the time when the above studies were made. The actual cause of different spread size is not central to this research. Suffice to say, the spread is dependent on supply and demand of a stock and the granularity of price increments (tick size, see below).

From a behavioural perspective, the size of the spread will dramatically affect the nature of the price process of a stock and its liquidity. A good example of this effect is clearly visible in Vodafone, VOD.L, listed on the LSE. Prior to March 2007, the typical spread of VOD.L was a single price increment. However, the size of the price increment changed in March 2007. There were many side effects due to this change. The otherwise relatively static orderbook now became noisy and traded in a small noisy channel. The individual trade sizes diminished, but, the number of trades increased. Similarly, the order arrival and cancellation patterns also changed. Therefore, a seemingly small change to a security can have drastic effects on its price process.

The spread in exceptionally liquid securities is typically a single tick wide. However, many large market capitalisation securities often have spreads varying from 1 to 3 ticks. Less liquid stocks tend to have spreads of 10 or more ticks.

2.4.4 Depth and Detail

Depth is the term commonly used to describe orders queued in the orderbook. The levels of depth are numbered $1 \dots n$, where 1 is the best bid or best ask. A typical characteristic of the orderbook is that the liquidity is higher the further down the depth, as there are more traders willing to buy / sell at prices more favorable to themselves. At each level of a buy side depth, there is typically an attribute that denotes the number of buyers at that level making up the liquidity amount. One of the shortcomings of a consolidated depth orderbook market view is felt in the event of a cancellation of an order occurring at touch. When the order quantity at touch changes to a smaller quantity, it is often useful to know whether this was caused by a trade or a cancellation. In the absence of high precision market data, it is not always easy to determine the cause of an order size reduction at touch. In short, some trades can be reported as a sequence of multiple smaller trades, where their sum is equal to the actual trade and also the changes in the orderbook following the trade. Further complexities arise due to the ordering of the reporting of the trades and changes to the orderbook. In the absence of accurate data, various matching techniques are sought to match changes to trades.

In some markets, a more granular market feed exists, providing the composition of the consolidated quantity (all of the individual orders). In this case, when there is a cancellation of an order, the event is tagged with the exact detail of the order being cancelled. This rich level of information is often referred to as orderbook *detail*.

Although the overall effect on the orderbook is the same regardless of whether the cause of the change was a trade or a cancellation, the information content of this state transition varies. For instance, whether the first order or the last one in the Bid1 queue got cancelled has different meaning / intention. One could possibly argue that some one who has waited in the queue for a while to advance to the first place will only cancel the order if he/she has come into some information advantage, thus making this price level less attractive. The same reasoning may not apply to the one at the back of the queue who cancelled. Orders at the back of the queue have a lower chance of getting filled, hence, cancellation could be the result of being risk averse and wanting to secure the prevailing asking price by cancelling the order and making it a market order.

Another usefulness of market detail data is that you are able to track the amount of orders ahead of yourself. With market depth, although you could record the quantity ahead at the time of placement of an order, a cancellation can happen either in-front or behind your own order. Therefore, after a cancellation, the certainty of your own position will be lost.

As discussed under Fill probability below, a forecasting model that is modelled with both market depth and detail data show that the forecasting accuracy is significantly higher with market detail as opposed to using market depth.

2.4.5 Tick Size rule

Price granularity, the rule governing determination of tick sizes of securities varies from stock exchange to stock exchange. In Europe, tick size is a form of step function of a security's price. Interestingly, in some markets, the granularity changes intraday when the price crosses the tick size defining boundary. For instance, a security may have a tick size of 0.25 when its price range is in a given price range and a tick size of 0.5 when the price range is in the subsequent range. Technically, the orderbook can be in a state where all bids are in a given tick size band while all offers are in another.

2.4.6 Orders

Each stock exchange provides a set of valid order types. Each order type specifies how an order is executed. There are two main classes of orders as discussed below, namely, *limit orders* and *market orders*. In addition to the exchange provided orders, due to popular demand by traders, Order Management Systems (OMS) provide a standard set of widely used order types that are executed at the exchanges as a series of primitive order types, where the exchanges do not support them. These are usually some form of contingent orders that are composed of market and limit orders, with dependance on some market related event.

2.4.7 Market Orders

The simplest form of orders are market orders, used to instantly complete the transaction. It requires only the quantity Q one wants to execute in a given security. The instruction will cause the matching engine at the exchange or venue to swipe through the orderbook until Q is filled. The transaction price, or execution price, is the weighted average price of all quantities traded at the different price levels (if any). For instance, in an order to buy Q=1000 shares, if there are only 500 shares available at the first ask level with $a_1 = 140$ and 5000 at the second ask level $a_2 = 142$, then the fill price will be 141.

A trader chooses to use an aggressive order type such as market order when he/she is a liquidity taker and wants to avoid any risk of the price drifting. The fundamental aim in all financial transaction is to buy low and sell high (the reverse for yield). If a risk neutral trader is crossing the spread and removing multiple levels of liquidity without any information edge, he or she will be demonstrating lack of sophistication. *Information edge* is the term used to describe traders who have some information advantage on short term price movement. The key side-effect of market orders is the market impact ones transactions cause. If the traders have a large Q to transact, market order will not suffice as there may not be sufficient visible liquidity. Therefore, the order has to be divided into a number of smaller market orders

q_i. Even then, the temporal effects of poorly constructed multiple market order trajectories can cause excess market impact resulting in a very poor execution price. Furthermore, such an execution strategy will also become very visible to other market participants, thus providing them with an edge. The longer one takes to execute an order using market order, the lesser the market impact would be.

2.4.8 Limit Orders

Complementing the market orders are limit orders. These passive orders are used to provide liquidity to the market. They consist of a Price/Quantity attribute pair for a given security. Effectively, a limit order shows our intention to buy or sell a certain quantity at a specified price. In the case of a buy order, this price is smaller than or the best ask. Technically, an aggressive limit price will have a similar effect to a market order. Indeed, many practitioners use aggressive limit orders to place market orders - providing a price cap on the price exposure. This is partly due to liquidity being removed by someone else at the same instant an order is placed, hence the fill price may be very poor. It is also to restrict exposure to be gamed (described below) - a technique used by some market participants to trick other traders with false prices.

The primary incentive to use limit orders is to capture a better price than the currently available price on the other side of touch. As such, placing an order with the hope of being filled within a specific time window T, will result in an attractive fill price if it is filled. However, if not filled, the price may have drifted away after time T. This would result in either a missed opportunity if the trader decided not to trade, and a higher cost if he/she chases the market and trades at the prevailing market price. Generally, it is safe to assume that limit orders do not have visible ”negative” market impact. An example of indirect effect limit orders have on the activity of the orderbook is when a typical and relatively large order is placed on a very illiquid stock. This action can lure out liquidity takers - as less liquid stocks are very sensitive to order visibility. Arrival of a limit order can also have the opposite effect of scaring sellers away.

2.4.9 Market vs Limit Orders

The choice of whether to use market or limit orders is rather complex, with no simple answer. Indeed, the core research presented in this thesis addresses how to deal with this dilemma in a formal manner. The choice of execution is decomposed into a series of market and limit orders, by optimising a set of key features affecting the quality of execution. The problem at hand crudely depends on whether (1) an information edge is available to the trader, and (2) his or her aversion to risk.

As also addressed by O'Hara [48], traders with and without information edge have significantly different execution strategies. Uninformed traders tend to place limit orders early in the execution period and rapidly convert to market orders in order to complete their orders. Informed traders trade to capture a certain value they know about the security. Therefore, they demonstrate a very complex pattern of execution consisting of both market and limit orders. Intuitively, there seem to be a complex relationship between gain and risk associated with the choice of orders in order to capture the expected value of the security.

2.4.10 Smart Order Routing

In a multiple venue market place, Smart Order Routing is a relatively new technology that has emerged to take away some of the technological complexity involved in sending market orders to the venue with the most attractive price. The best price can be defined in terms of the actual price only, or a combination of factors. These factors can range, from broker commission fee payable to that exchange, exchange fees etc, to latency to the market. This is a transparent routing problem defined by a number of clearly defined parameters.

SORs are critical to the success of multi-venue market places. Although very useful, SORs are somewhat restricted in their value. When a trader wants to trade a quantity Q, he or she may use SOR to take the liquidity *up to* a quantity q_1 within a given price limit and place the remaining quantity q_2 as a limit order. If there are multiple venues to place an order, one may wonder, at which venue should q_2 be placed? Customarily, this has been with the primary exchange at which the order is listed. With the example of Vodafone, LSE will be the destination chosen above Chi-X for instance. Although LSE may have a higher liquidity than Chi-X on average, the intraday fluctuations in liquidity may be such that Chi-X is at times a better choice. Furthermore, Chi-X credits one for providing liquidity, provided the order gets filled. Today's SOR technology cannot deal with these situations. Therefore, Smart Order Routing's usefulness is restricted to market orders only.

2.4.11 Need for better Smart Order Routing

The dilemma of where to place the residual order in the case of SOR above, poses the problem of *dynamic evaluation* of the attractiveness of venues. If the SOR was to dynamically pick Chi-X as being the better choice and place the limit order there, the question is when does it have to re-evaluate the situation and perhaps move the order to LSE? Any modification to an order would put the order at the back of the orderbook queue at that particular price level. Therefore, one is disadvantaged by any change to a limit order unless the probability of being filled improves with the change. In order to assess the fill probability, if such a measure existed, one would need to know the duration T a trader is willing to let the order work passively to get filled. Since the fill probability is rarely 1, one is faced with yet another dilemma of what to do after T. That is, if at T, the order is not completed, what action needs to be taken? In fact, the price would most likely have drifted away. Therefore, when determining the duration T, the price drift risk should be one of the key drivers.

Therefore, the so-called intelligent order routers would require a much more complex drivers and most users of such a hypothetical technology would have to provide most of this information. Without key information, such as Fill Probability for instance, the router would no longer be intelligent.

2.5 Visible and Hidden Liquidity

There is a range of loose definitions of liquidity and widely accepted usage of the terms. All of these are vaguely defined terms to describe the temporal supply (and demand) of an asset. Essentially, liquidity is a key factor in completing a transaction with minimal price impact. Therefore, it is a problem of three dimensions, namely, Price, Quantity and Time-frame. Given that supply is finite at a given price level, there will be some form of market impact as a result of executing a non-trivial order. The larger the liquidity, the greater availability of counterparties to transact with over time, the smaller the impact the transaction will have on the price. Hasbrouck [30] summarises liquidity as being:

"Liquidity is a summary quality or attribute of a security or asset market. There are no formal definitions, except those that are very context-specific."

Liquidity, a core concept in our research, is believed to be affected by many factors - including, in particular ticksize and bid/ask spread. The size of the spread is critical to the stationarity property of an orderbook, which will again affect the fill probability. Also, important to orderbook liquidity are the liquidity arrival time and quantity distributions. Put differently, if one were to remove liquidity by placing a market order, how much time will elapse before a certain amount of new liquidity arrives? A high refill rate will imply less market impact, allowing one to follow with more liquidity taking orders, if so wished, to complete the entire quantity.

With electronic trading gaining significance in exchanges, there are structural changes in the liquidity arrival and removal patterns. The average arrival order size today is approximately a fifth the typical arrival size of 6-8 years ago. This is despite a significantly larger turnover in securities today. Prior to electronic trading, most orders were entered manually. As such, for practical reasons, a natural and typical order size for each stock was established to balance manual convenience versus cost. With electronic trading, where one has to disguise one's intent, the amount shown on the orderbook is sufficient to attract interest, but small enough not to betray your intention. When liquidity is removed, it will be quickly replaced by new liquidity. This type of hidden liquidity is very difficult to quantify.

The structural changes to the arrival pattern, size and price of liquidity will invariably affect the price formation process of the security in ways quite dissimilar to before. Once again, the mere change in tick size of Vodafone transformed the intraday price formation process instantly - going from being piecewise static to a noisy price process.

2.6 Market Impact

Market Impact (MI) consideration is of fundamental importance to all financial transactions conducted in exchanges, whether executed manually or electronically. In general, all orders placed in an orderbook will have some level of impact on the price process. A very large passive limit order, a provider of liquidity can also has negative price impact by the share presence. Such a large buy limit order sends out the signal to the market place that one has a need to complete a large order and this information could make the sellers move their limit orders further away from the touch, with all other factors assumed to be constant. However, we will in this work assume that limit orders will not have any indirect price impact of this nature as our order quantity is cautiously constructed. Instead, we focus on the impact caused by aggressive market orders.

All market orders remove liquidity from the orderbook. This causes a temporary and permanent market impact of varying magnitude to the price. Intuitively, if a significantly large order is to be executed, it will be divided into a number of atomic orders and worked over a period. Execution of each of the atomic orders will cause an impact on the market, ie. it moves the price further away. This is partly due to the visible liquidity being exhausted. But, liquidity replenishes itself and the price will revert somewhat and find new price equilibrium. This impact and price movement is often referred to as temporary impact. Permanent impact is much less intuitive and is the impact on price that is persistent - perhaps even for days. The magnitude of permanent versus temporary impact is approximately an order of magnitude, with temporary impact being the larger one.

The magnitude of the impact is dependent on the intensity of a trade. Intensity λ, in its static form is defined as being a relative measure of quantity traded versus the Average Daily Volume (ADV). Therefore, if a large order is excuted as a trajectory of smaller market orders over time, there will be a temporal effect of market impact on the price process arising from each of the smaller orders.

When building an algorithmic trading model, market impact too has to be included. The difficulty is in how one quantifies the market impact or rather the expected market impact. Market impact is a widely used term in algorithmic trading circles, at times with obscure meaning. This is a particularly difficult problem to comprehend despite its apparent simplistic nature. It is not exactly measurable since the decomposition of noise and impact is unclear. Each institution has its own flavor of market impact model, most being very rudimentary. There are many models in the literature studied and published mostly by academics, where most of them base the analysis on publically available trades and quotes information. These works include Breen et al. [56] with their work on net market move and buy/sell imbalance over a time window. Kissel and Glantz [37] work in similar line while Rydberg and Shephard [51] propose an econometric framework for changes in price affected by market impact. Others include Dufour and Engle [19] with their study of waiting time between successive trades and a statistical physics approach by Lillo, Farmer and Mantegna in [40]. Bouchaud et al [11] study the serial correlation in volume and price data - an interesting work successfully used by some banks in an altogether different context to spot the presence of undefined algorithmic orders in the market.

Trying to quantify market impact without identifying the source of the orders makes it very difficult to distinguish market impact from noise. Therefore, in order to better understand market impact, Chan and Lakonishok [13], Holthausen et al. [32], Madhavan [36] and others have studied trades with limited information edge over other academic research on public data. In these cases, trades conducted by specific asset managers with date stamp but no other additional information, were studied.

2.6.1 Impact Measure

Execution of any order is subject to two costs - volatility and market impact. We adopt Almgren's market impact model, Almgren [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 a relatively short time according to the liquidity pattern while permanent impact can stay well after the trade is executed. Temporary impact, according to Almgren [1], is larger than permanent by an order of magnitude and hence is significantly more important for our model. Impact function depends on two parameters, spread ε and intensity of trade λ. Intensity of trade is defined with respect to ADV (Average Daily Volume) and the market impact function is given by

$$f(q) = \varepsilon + \bar{\mu}\lambda^b, \ \lambda = \lambda(q), \tag{2.1}$$

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

$$\pi(q) = f(q)q \tag{2.2}$$

For more details see Almgren [1], Almgren and Chriss [4].

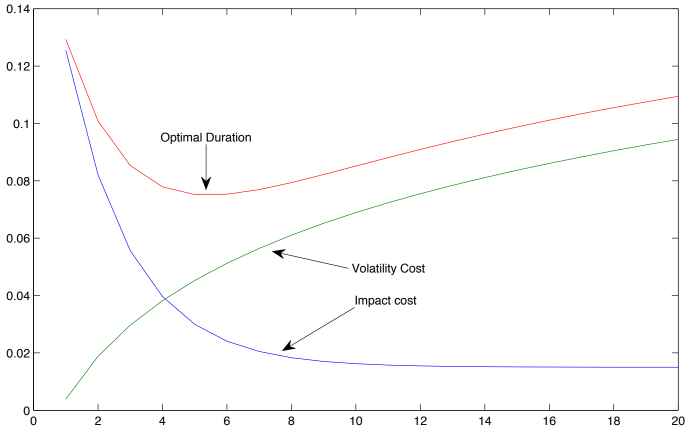


Figure 2.3: Impact and volatility costs versus time. Troškovi impakta i volatilnosti u vremenu

A 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. Another 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 by Figure 2.3. We will assume uniform execution of market orders and use the temporary market impact function (2.2) as suggested in Almgren and Chriss [4].

2.7 Price Process

Let us denote the bid prices as $b_1(t), \ldots, b_n(t)$ and ask prices $a_1(t), \ldots a_n(t)$ for any given time t. 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, cancellation 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 thesis we will assume that all prices follow an arithmetic random walk without drift,

$$b_i(t) = b_i(0) + \sigma\sqrt{t}\xi_i, \tag{2.3}$$

$$a_i(t) = a_i(0) + \sigma\sqrt{t}\gamma_i, \tag{2.4}$$

$$P(t) = P(0) + \sigma\sqrt{t}\zeta, \tag{2.5}$$

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, $\xi_i, \gamma_i : \mathcal{N}(0,1),\ i = 1, \ldots, n$ 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 Dacorogna et al. [17], the volatility parameter σ in (2.3)-(2.5) can be estimated in a satisfactory way in normal market conditions.

2.8 High Frequency data

Main properties of high frequency financial data (HFFD)are irregular temporal spacing, discreteness, diurnal patterns and temporal dependence. Furthermore multiple transactions occur within a second with different transaction prices and transaction volumes. These properties make HFFD more difficult to analyze and built reliable forecasting models. On the other hand HFFD are an extremely rich source of information that has been widely used in recent years for understanding market dynamics and building models.

Fundamental questions that arise which we can take for granted in the homogenous data world are, how does one calculate return between two points? In homogenous data, the time interval between all consecutive data points are equally space. This interval may have a large number of non-homogenous ticks or there may be no ticks. If one does not have a return number, then how can volatility or correlation be calculated?

In order to answer these questions, one needs to perform an analysis of HFFD that will provide deep insight into market dynamics. This is crucial for building reliable models for price process, calculating returns at arbitrary time interval and volatility. Yet once again, this is beyond the scope of this research. However, extensive analysis of this form went into building the continuous fill probability model as discussed below.

The first property of HFFD which makes it different from classical time series is irregular temporal spacing. Market ticks arrive at random times and transactions occur at different frequencies. Therefore each stock has its own temporal pattern. In general temporal pattern depends heavily on the liquidity of a stock. There are two basic approaches to resolve the irregular temporal spacing. The first one is homogenization i.e. transformation of nonhomogeneous time series into homogenous ones. In this transformation, some average information is extracted for each time interval of fixed length. The key issues are how to choose time intervals of fixed length and then how to define "average" information that will represent whole intervals. The time interval should be small enough to capture all micro-changes but it should not be too small. The homogenous time series will be the subject of analysis and modeling that include serious computational effort and its size must be manageable in a computational sense. Clearly the type of observation we are transforming into homogeneous series determines the choice of "average" value that will serve for a whole time interval.

Another possibility is to use stochastic time intervals when dealing with non-homogeneous time series. In that case we transform the raw time series into a new one (also non-homogeneous) but with new variables. Such a procedure has its advantages as discussed in Engle and Russell [21] and Dacorogna et al. [17] but it is not possible to judge which one is better.

An additional complication is the fact that irregular spacing is even more complex when dealing with multiple time series where each one has its own transaction frequency.

All financial data is discrete. The trading prices, depending on a particular market, take only discrete values. In general, transaction prices take only a couple of pre-specified values. Bid and ask prices obey the same restrictions. Volume is always given as an integer number. It is well known that discreteness also introduces a high degree of kurtosis in the data, see Engle and Russell [21].

HFFD exhibit strong diurnal patterns. For most stock markets volatility, frequency of trades, volumes typically have a U-shaped pattern over the course of a day. Volatility is systematically higher near opening and close. On the other hand duration (time between two consecutive transactions) has an inverse pattern since the time between two trades are shortest near open and just prior to close. Therefore all analysis must include some deseasonalization. It should be noted that the U-shaped pattern of trading activity is more typical in the US than it is in Europe or Asia.

One more important property of HFFD is temporal dependence. High frequency financial returns typically display strong dependence. This dependence is caused by price discreteness and spread in price paid by buyer and seller initiated trades - the so-called bid-ask bounce. There is also the common practice of breaking larger orders up into a sequence of smaller orders in hopes of transacting a better overall price. These sequential buys (or sells) might lead to a sequence of transactions that move the price in the same direction with the data exhibiting positive autocorrelations. Such autocorrelation is present both at transaction rates and volumes.

Chapter 3

Algorithmic Trading

Algorithmic trading is the automated process of the execution of orders in an electronic market place. The orders are typically originated by portfolio managers, traders or automated trading systems. These orders are either sent electronically or verbally over the telephone. The figure 3.1 describes a typical algorithmic trading architecture.

3.1 Algorithms

The algo marketplace is flooded with a large variety of algorithms to choose from. One of the current challenges faced by users of algorithms is to understand the nuances of all of the exotically named algorithms such as Stealth Hunter, Guerilla, Wait & Pounce, Night Hawke, etc. to mention a few. In an extremely competitive environment, it is difficult for the user to differentiate between what is hype and what is real. In general, for liquid stocks, algorithms can be divided into two main variants, VWAP and Arrival Price.

VWAP and Arrival Price are merely two main classes of benchmarking. In the first case, the execution quality is measured relative to the volume weighted average price of the execution window, while the latter is measured relative to the market price at the start of the execution.

The performance of all mainstream algorithms are measured first by their mean slippage to benchmark, followed closely by variation in slippage.

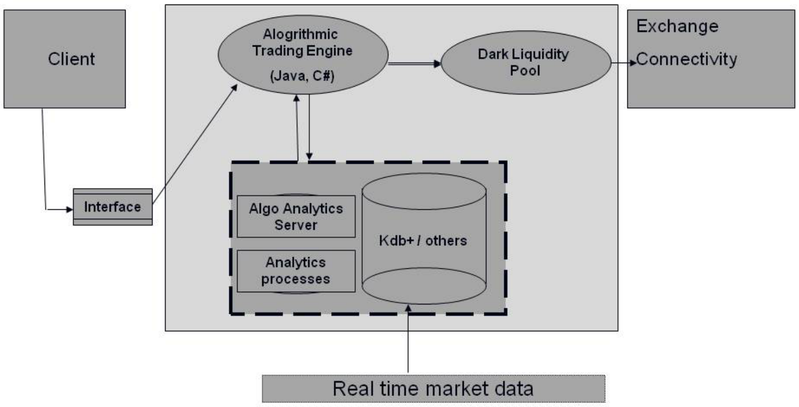


Figure 3.1: High level algorithmic trading architecture. Arhitektura algoritamskog trgovanja.

3.1.1 VWAP

There are a number of flavours of VWAP algorithm. In general, execution is defined over a fixed time window and the aim is to trade proportionately (relative to the traded volume) throughout the trading window. Typically, a static / historic volume profile for the time window in question is used in most VWAP algorithms as a proxy for expected volume pattern in the currently working window. Although there is sufficient empirical evidence of U-shaped historical volume profiles over a trading day Madhavan [42], this approach is far too simplistic. Furthermore, despite the historic volume profile, the daily variations can be very large. This is particularly so in less liquid stocks. The variation around historic distribution also differs considerably across stocks.

The more advanced VWAP algorithms tend to use some form of dynamic volume forecasting models. Performance benchmarking of VWAP algorithms with and without volume forecasting shows a significant difference in variation in performance although their mean slippage is often very similar. It is worth noting that VWAP algorithms are not sensitive to drift in price relative to the start price of the execution - their sensitivity is mostly to the

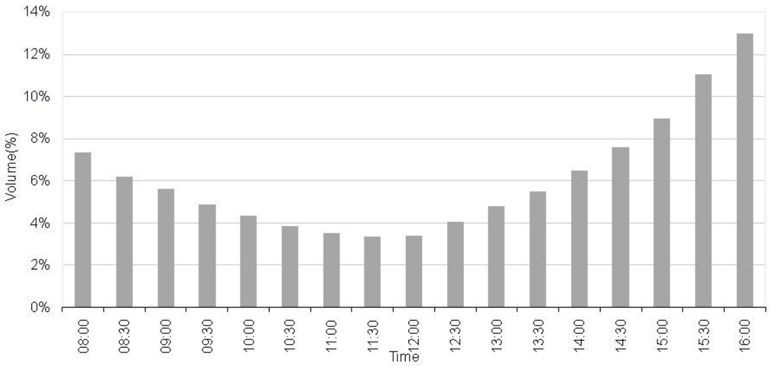


Figure 3.2: VWAP execution profile. VWAP profil izvršenja.

evolving volume profile.

Regardless of how one determines the expected volume profile, a time window is divided into a number of bins, with bin size dependent on the volume profile estimation method. Given a bin, all other logic of efficient execution is still valid. In fact, the order / quantity to be executed in a given bin is an *an atomic order*. The primary aim of the execution within a bin is to obtain the best possible price, followed closely by minimize the market impact caused. As such, we are faced once again with the dilemma of providing liquidity and capturing spread at the potential risk of increasing tracking error around VWAP in the case of execution not taking place, versus causing market impact and better track volume pattern. Intuitively, a combination of these two types of orders are required. Shown in figure 3.2 is a volume profile and typical bins for a VWAP order.

Interestingly, VWAP is a moving benchmark. The benchmark itself is affected by our own trading action. Therefore, a large order traded with excess market impact would not show up as a bad execution because the traded prices of this order would have significant impact on the benchmark. This is perhaps one of the biggest shortcomings of VWAP benchmark.

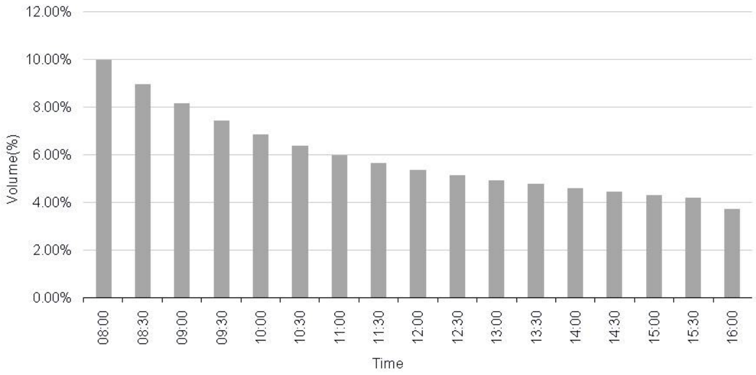


Figure 3.3: Implementation Shortfall execution profile

3.1.2 Implementation Shortfall

The main contender to VWAP type algorithms are arrival price based algorithms. The primary objective of this algorithm is to minimize the slippage relative to the arrival price / the prevailing price when the order started.

While volume profile and consequently market impact are still important, Implementation Shortfall (IS or arrival price) is mostly concerned with the price process during the execution window. Intuitively, given a stock's volatility, the longer an order is worked, the further the price could move away from the reference price, although this could work in one's favor as well. This potential variation in price will translate itself into variation in slippage as well, although it does not imply any worse mean slippage.

The general shape of the execution profile of an IS is shown in figure 3.3. It aims to trade more intensely at the start to reduce the effect of price drift.

Not unlike VWAP algorithms, IS also slices the trading window into a number of bins, ie. a number of atomic order-windows where the primary objective is to obtain the lowest price in the window. In contrast to VWAP, a great deal of care is needed to ensure that no excess market impact is caused - thus keeping the price as favourable as possible for the subsequent bin. The last mentioned is important to all algorithms. However, the effect of poor execution in an atomic order window will have a larger impact on arrival price type algorithms in their benchmarking, although the effect on the actual price process and its quality is the same.

Arrival price type strategies are considered particularly difficult - particularly for large orders. While with VWAP, very large orders help define the benchmark ie. effectively reducing the influence of one's own market impact, the opposite is the case in IS. Larger orders will have larger market impact and larger impacts will have larger price slippage to reference price. In the case of larger orders, the optimal execution window of the IS order will also be larger.

3.1.3 Exotic Algos

There are a whole range of smaller algorithms that are popular among traders. These are specialist algorithms whose performance is difficult to quantity. In fact, some brokers do not even provide or define the benchmark for this class of algos. To understand why this is so, one needs to look at the objective of the algorithm in question. In the case of VWAP or IS, there was a clearly defined objective, namely slippage relative to a benchmark price. In the case of exotic algos, the primary objective of the best price is indirectly achieved since the same rules of execution may not apply. For instance, trading illiquid stocks is very difficult due to the inherent thinness of the visible orderbook. In these stocks, showing one's hand by displaying an order will quickly be observed by market participants with one of two main outcomes: (1) arrival of new liquidity takers, (2) liquidity on the opposite side being removed. The reasoning is as follows: By providing liquidity, one may be able to attract a trade waiting to take a significant amount of liquidity to enter into the market. If there was no such significant liquidity taker awaiting, the sellers will look at the newly arrived liquidity on the bid as a large buy order and remove the better offers.

Therefore, when trading these stocks, clever methods to disguise one's intention is considered as critical to success as best price, and benchmark price becomes less relevant.

3.2 Pre and Post Trade Analysis

Transaction cost analysis (TCA) has been very well researched during the past few decades. It is well explained in Kissell and Glantz [37]. The primary purpose of this type of analysis is to quantify and decompose the relationships between expected execution price, the realized execution price and the desired benchmark price.

Prior to executing an order, a trader has a vast amount of options as to how best to execute the order. He also will have some expectation of what it will cost to transact the order in each case with the help of pre-trade analysis tools. In short, a pre-trade analysis allows a trader to find the optimal trading window for the different algorithms, and estimate the expected market impact cost for the various risks he/she is willing to take.

Figure 2.3 illustrates an optimal execution duration for a Implementation Shortfall Algorithm. It depicts the relationship between the market impact as a function of time for a fixed order quantity and the volatility risk curve under the same conditions. The market impact curve here is the price retardation caused by the entire order being traded instantly vs stretching it over a longer period. Similarly, the volatility curve is the potential price risk over the same period. Although the price could move in the trader's favor, only the worst case is considered. The optimal execution duration is the minima of the combined curves, defining the balance between risk and gain.

Another important part of pre-trade analysis is to look at the history of execution of the chosen class of execution algorithms and interpolate the historic performance to obtain the expected mean transaction cost and standard deviation of slippage for the order one wishes to execute. Suffice it to say, although this approach is very widely used, the regression methods used yield a very low R^2. In order to get reasonable results from such a method, one needs a large trade history in order to have a good statistical sampling of the wide order range. Even then, the market conditions under which the orders were executed can be so vast that the problem requires a far more sophisticated solution.

The average executed price could be compared with the associated benchmark price in order to determine the slippage. For most traders this simple number would suffice, as this is the outcome of the execution. However, from a knowledgable trader's or a corporate wide perspective, it is important to probe into the execution details. One may want to decompose the resultant cost into a number of cost contributing factors.

Kissell [37] identifies nine key components of transaction cost. They are:

- Commission: payment to the executing broker
- Fees: exchange fee, etc.
- Taxes: various government taxes
- Spreads: percentage of orders that crossed the spread (market order)
- Delay Cost: delay between the investment decision and start of execution
- Volatility Cost: price trend, drift, momentum or alpha.
- Market Impact Cost: change in price caused by these orders
- Timing Risk: uncertainty created with executing the order over many days
- Missed Opportunity Cost: the true cost of un-completed orders

For a portfolio of orders, these basic categories of cost attributes can be aggregated across *industry groups, sectors, exchanges and country.* By performing this type of cost decomposition, one may be able to focus on where improvements can be made, or devise more suitable algorithms.

3.3 Fill Probability

When a trader decides to place an order in the market, he or she has some level of urgency to complete that order. Depending on their sensitivity to price fluctuations, the length of the window will also vary - a risk aversion factor. This risk aversion will directly translate into the price level at which to place the order into the orderbook in order to complete within the expected time. Experienced traders develop, over the years, an intuitive feel for estimating the probability of getting their orders filled, which aids their manual execution task well enough. However, in the realm of algorithmic trading, since all decision making has to be formal, estimation of many of the factors needed to make decisions, including probability of fill, becomes a major problem.

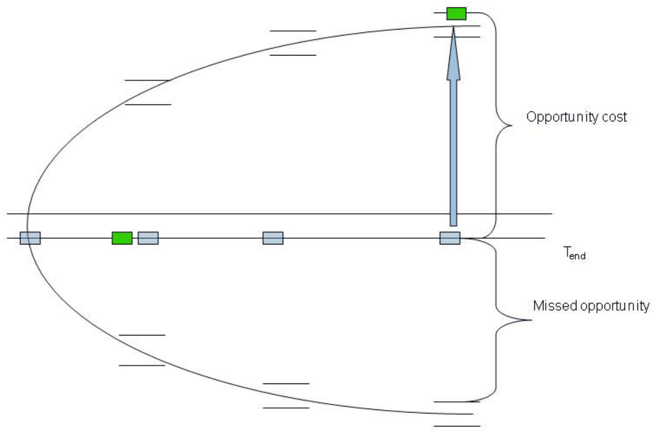


Figure 3.4: Negative Selection caused by a asymmetry in the payoff. Negativna selekcija uzrokovana nesimetrijom isplativosti.

Accurately estimating of probability of fill is an exceptionally difficult problem. As a result, the overwhelming majority of quantitative trading professionals devise crude models, which can be summaried as:

"The probability of getting filled on an order placed at the best bid/ask is approximately 60% within a reasonable time."

Even in the absence of an information edge, like a model to forecast fill probability, the current practice is far from logical even if the vague factors were rectified. The primary aim of all execution is to maximise price quality. Generally, given that in any small arbitrary window of time, the price moving up or down is virtually random, the execution algorithm should also utilize this property of the price process. The logical flaw with the above rule of execution is the asymmetry in payout resulting from always staying at touch. For instance, a buy order at the best bid will get filled instantly at the specified price if the market were to drift downwards. However, if it moved up the same amount, one would chase the market and get filled at a price much higher up. Therefore, we are exposed to a skew in payoff.

In essence, we have a classical case of *negative selection* as shown in figure 3.4, where we get filled when we least want to - with no gain if the price moves down as we were filled too early and lose out when the price rallies and taking a fill price that is significantly away from the initial order price. This type of asymmetry is clearly evident when analyzing the quality of past fills when orders were submitted to brokers to execute manually. Typically, a broker with a basket of stocks to execute will determine a suitable execution schedule (unless otherwise specified by the client). Next, small chunks of orders in all the stocks will be placed passively in the market. If orders get filled, the broker will be swift in replenishing with new passive supply at the prevailing bid. If the market moves away, they will wait in the anticipation that the price will come back.

The net effect of the "constant 60% expect fill mode" as described above and that of a typical broker execution are that they both subscribe to negative selection.

The quality of an order's fill price is very much dependent on maximizing the opportunity arising from price symmetry. Intuitively, in order to do this, extending the above approaches by breaking up an order into many sub-orders and placing them at many different price levels may seem logical. This should result in a better average fill price when the price moves in our favour. In the event of price moving away, all orders could be shifted upwards, much like in the original case.

The core of our research is centred around how one can decompose an atomic order, the fundamental building block of a large order, into multiple price levels across multiple markets in order to achieve optimal execution. Before embarking on the quest to construct an optimal execution framework, it is essential to understand the notion of fill probability, as this is singularly the most important factor determining the quality of fill in an optimisation problem.

The fill probability model used in this research is a proprietary mathematical model and its inner workings cannot be disclosed. However, we will address all key properties of the model as required for the analysis. It should be noted that the optimisation framework we propose herein is not dependent on this particular implementation of a fill probability model.

3.3.1 Definition of Fill Probability

Unlike 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 opportunity cost, and since we we want the order executed, it has to be placed at a higher price. On the other hand, if the order got filled there is a clear gain in price compared with market order. Therefore, for any bid level we define gain coefficients as

$$c_i = a_1 - b_i,\ i = 1, \ldots, n. \tag{3.1}$$

Obviously the gain c_i (3.1) 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

$$\beta_i(q) : \begin{pmatrix} 1 & 0 \\ p_i(q) & 1 - p_i(q) \end{pmatrix}. \tag{3.2}$$

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 (3.2) we define the set of functions $F_i(q)$ for all $i = 1, \ldots, n$ as

$$F_i(q) = p_i(q), \tag{3.3}$$

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 document. 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

$$H_i(q) = qF_i(q) \tag{3.4}$$

and *gain functions* as

$$G_i(q) = c_i H_i(q). \tag{3.5}$$

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 chapter 8. The empirical data also gives us reason to believe that F_i are convex functions (see Figure 3.5.)

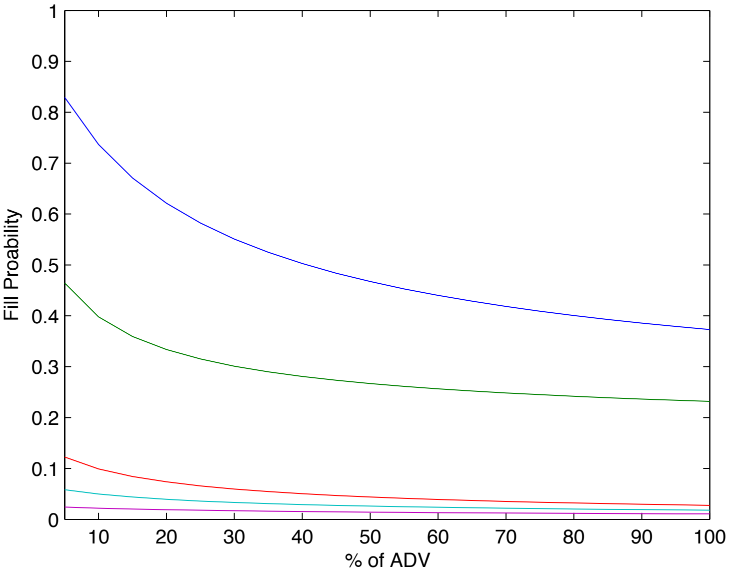


Figure 3.5: Fill Probability functions for five bid levels. Fill Probability funkcije za pet bid nivoa

Chapter 4

Constrained Non-linear Problems

4.1 The Problem and Optimality Conditions

The general formulation of constrainted optimization on the variables is given as

$$\begin{aligned} \min_{x \in R^n} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) = 0, \quad i \in \varepsilon \\ & g_i(x) \geq 0, \quad i \in \mathcal{I} \end{aligned} \tag{4.1}$$

where f and the function g_i are all smooth, real-valued functions on a subset of $\mathcal{R}^n$, and $\mathcal{I}$ and ε are two finite sets of indices. We call f the objective function, while, g_i, $i \in \varepsilon$ are the equality constraints and $g_i, i \in \mathcal{I}$ are the inequality constraints. We define the feasibile set Ω to be the set of points x that satisfy the constraints; that is,

$$\Omega = \{x \mid g_i(x) = 0, \quad i \in \varepsilon; \quad g_i(x) \geq 0, \quad i \in \mathcal{I}\} \tag{4.2}$$

so that we can rewrite the equation (4.1) more compactly as

$$\min_{x \in \Omega} \quad f(x) \tag{4.3}$$

There are two types of optimality conditions, namely, *necessary* and *sufficient* conditions. All solution points must satisfy necessary conditions (under certain assumptions). Sufficient conditions on other hand guarantee that x^* is a solution, provided the conditions were satisfied at a certain point x^*.

Local solutions in constrained cases are restricted to the feasible points in the neighborhood of x^*. It should be noted that although isolated local solutions are strict, the reverse does not hold. We have the following definition types of local solutions

A vector x^* is a local solution of the problem (4.3) if $x^* \in \Omega$ and there is a neighborhood $\mathcal{N}$ of x^* such that $f(x) \geq f(x^*)$ for $x \in \mathcal{N} \cap \Omega$.

A vector x^* is a strict local solution if $x^* \in \Omega$ and there is a neighborhood $\mathcal{N}$ of x^* such that $f(x) > f(x^*)$ for $x \in \mathcal{N} \cap \Omega$ with $x \neq x^*$.

A point x^* is an isolated local solution if $x^* \in \Omega$ and there is a neighborhood $\mathcal{N}$ of x^* such that x^* is the only local solution in $\mathcal{N} \cap \Omega$.

A fundamental concept that provides a great deal of insight as well as simplifying the required theoretical development is that of an *active constraint*. An inequality constraint $g_i(x) \leq 0$ is said to be *active* at a feasible point x if $g_i(x) = 0$ and *inactive* at x if $g_i(x) < 0$. By convention, we refer to any equality constraint $g_i(x) = 0$ as active at any feasible point. The constraints active at a feasible point x restrict the domain of feasibility in neighborhoods of x, while the other, inactive constraints, have no influence in neighborhoods of x. Therefore, in studying the properties of a local minimum point, it is clear that attention can be restricted to the active constraints. The active set is formally defined as

Definition 7 *The active set $\mathcal{A}(x)$ at any feasible x consists of the equality constraint indices from ε together with the indices of the inequality constraints i for which $g_i(x) = 0$; that is,*

$$\mathcal{A}(x) = \varepsilon \cup \{i \in \mathcal{I} \mid g_i(x) = 0\}. \tag{4.4}$$

In order to proceed, we will need to define the tangent cone. Given a feasible point x, we call $\{z_k\}$ a feasible sequence approaching x if $z_k \in \Omega$ for all k sufficiently large and $z_k \rightarrow x$. A tangent is a limit in the direction of a feasible sequence.

Definition 8 *The vector d is said to be a tangent (or tangent vector) to Ω at a point x if there are a feasible sequence $\{z_k\}$ approaching x and a sequence of positive scalars $\{t_k\}$ with $t_k \to 0$ such that*

$$\lim_{k\to\infty} \frac{z_k - x}{t_k} = d$$

The set of all tangents to Ω at x^* is called the tangent cone and is denoted by $T_\Omega(x^*)$.

We will also use a linearized feasible direction set which we define as follows:

Definition 9 *Given a feasible point x and the active constraint set $\mathcal{A}(x)$ of Definition 7, the set of linearized feasible directions $\mathcal{F}(x)$ is*

$$\mathcal{F}(x) = \left\{ d : \begin{array}{ll} d^T \nabla g_i(x) = 0, & \textit{for all } i \in \varepsilon, \\ d^T \nabla g_i(x) \geq 0, & \textit{for all } \ i \in \mathcal{A}(x) \ \cap \ \mathcal{I} \end{array} \right\} \tag{4.5}$$

Tangent cone definition relies only on the geometry of the feasible set while the linearized feasible direction set depend on the definition of the constraint functions.

Constraint qualification are conditions under which the linearized feasible set $\mathcal{F}(x)$ is similar to the tangent cone $T_\Omega(x)$. Most constraints actually ensure that these two sets are identical. These conditions ensure that $\mathcal{F}(x)$, which is constructed by linearizing the algebraic description of the set Ω at x, captures the essential geometric features of the set Ω in the vicinity of x, as represented by $T_\Omega(x)$.

The constraint quantification most often used in the design of algorithms is the subject of the following definition

Definition 10 *Given a point x and the active set $\mathcal{A}(x)$ defined in Definition 7, we say that the linear independence constraint qualification (LICQ) holds if the set of active constraint gradients $\nabla g_i(x), i \in \mathcal{A}(x)$ is linearly independent.*

We define the Lagrangian function for the general problem (4.1),

$$\mathcal{L}(x, \lambda) = f(x) - \sum_{i \in \varepsilon \cup \mathcal{I}} \lambda_i g_i(x), \tag{4.6}$$

The necessary conditions defined in the following theorem are called *first-order-conditions* because they are concerned with properties of the gradients (first-derivative vectors) of the objective and constraint functions.

Theorem 1 *First-Order Necessary Conditions.*
Suppose that x^* is a local solution of (4.1), that the functions f and g_i in (4.1) are continuously differentiable, and that the LICQ holds at x^*. Then there is a Lagrange multiplier vector λ^*, with components $\lambda_i^*, i \in \varepsilon \cup \mathcal{I}$, such that the following conditions are satisfied at (x^*, λ^*)

$$\begin{aligned}
\nabla_x \mathcal{L}(x^*, \lambda^*) &= 0, && (4.7)\\
g_i(x^*) &= 0, \quad \textit{for all } \ i \in \varepsilon, && (4.8)\\
g_i(x^*) &\geq 0, \quad \textit{for all } \ i \in \mathcal{I}, && (4.9)\\
\lambda_i^* &\geq 0, \quad \textit{for all } \ i \in \mathcal{I}, && (4.10)\\
\lambda_i^* g_i(x^*) &= 0, \quad \textit{for all } \ i \in \varepsilon \ \cup \ \mathcal{I}, && (4.11)
\end{aligned}$$

The conditions (4.7 - 4.11) are often known as the *Karush-Kuhn-Tucker conditions*, or *KKT conditions* for short. The conditions 4.11 are *complementary conditions*; they imply that either constraint i is active or $\lambda_i^* = 0$, or possibly both. In particular, the Lagrange multipliers corresponding to inactive inequality constraints are zero, we can omit the terms for indices $i \notin \mathcal{A}(x^*)$ from (4.7) and rewrite this condition as

$$0 = \nabla_x \mathcal{L}(x^*, \lambda^*) = \nabla f(x^*) - \sum_{i \in \mathcal{A}(x^*)} \lambda_i^* \nabla g_i(x^*). \qquad (4.12)$$

Satisfaction of the strict complementarity property usually makes it easier for algorithms to determine the $\mathcal{A}(x^*)$ and converge rapidly to the solution x^*. For a given problem (4.1) and solution point x^*, there may be many vectors λ^* for which the conditions (4.7-4.11) are satisfied. When the LICQ holds, however, the optimal λ^* is unique.

Definition 11 *Strict Complementarity.*
Given a local solution x^* of (4.1) and a vector λ^* satisfying (4.7 - 4.11), we say that the strict complementarity condition holds if exactly one of λ_i^* and $g_i(x^*)$ is zero for each index $i \in \mathcal{I}$. In other words, we have that $\lambda_i^* > 0$ for each $i \in \mathcal{I} \cap \mathcal{A}(x^*)$.

The KKT conditions tell us how the first derivative of f and the active constraints g_i are related to each other at a solution x^*. When these conditions are satisfied, a move along any vector w from $\mathcal{F}(x^*)$ either increases the first-order approximation to the objective function (that is, $w^T \nabla f(x^*) > 0$, or else keeps this value the same (that is, $w^T \nabla f(x^*) = 0$).

The second derivative plays a "tiebreaking" role. For the direction $w \in \mathcal{F}(x^*)$ for which $w^T \nabla f(x^*) = 0$, first derivative alone is not sufficient to determine whether a move along this direction will increase or decrease the objective function f. Second-order conditions examine the second order derivative terms in the Taylor series expansions of f and g_i, to see whether this extra information resolves the issue of increase or decrease in f.The second-order conditions are essentially concerned with the curvature of the Lagrangian in the "undecided" directions - the directions $w \in \mathcal{F}(x^*)$ for which $w^T \nabla f(x^*) = 0$. Stronger smoothness assumptions are needed since we are now discussing second-order derivatives. Therefore, f and g_i, $i \in \varepsilon \cup \mathcal{I}$, are all assumed to be twice as continuously differentiable.

Given $\mathcal{F}(x^*)$ from definition 4.3 and some Lagrange multiplier vector λ^* satisfying the KKT conditions (4.7 - 4.11), we define the *critical cone* $\mathcal{C}(x^*, \lambda^*)$ as follows:

$$\mathcal{C}(x^*, \lambda^*) = \{w \in \mathcal{F}(x^*) \mid \nabla g_i(x^*)^T w = 0, \ \textit{all } i \in \mathcal{A}(x^*) \cap \mathcal{I} \textit{ with } \lambda_i^* > 0\}$$

Equivalently, $w \ \in \mathcal{C}(x^*, \lambda^*$ if and only if

$$\begin{cases} \nabla g_i(x^*)^T w = 0, & \text{for all } i \in \varepsilon, \\ \nabla g_i(x^*)^T w = 0, & \text{for all } i \in \mathcal{A}(x^*) \cap \mathcal{I} \text{ with } \lambda_i^* > 0, \\ \nabla g_i(x^*)^T w \geq 0, & \text{for all } i \in \mathcal{A}(x^*) \cap \mathcal{I} \text{ with } \lambda_i^* = 0. \end{cases} \tag{4.13}$$

The critical cone $g(x^*, \lambda^*)$ contains direction w with the property

$$w \in g(x^*, \lambda^*) \ \Rightarrow \ \lambda_i^* \nabla g_i(x^*)^T w = 0 \quad \text{for all } \ i \in \varepsilon \cup \mathcal{I} \tag{4.14}$$

and the KKT conditions imply

$$w \in g(x^*, \lambda^*) \ \Rightarrow \ w^T \nabla f(x^*) = \sum_{i \in \varepsilon \cup \mathcal{I}} \lambda_i^* w^T \nabla g_i(x^*) = 0. \tag{4.15}$$

Hence the critical cone $g(x^*, \lambda^*)$ contains directions for which it is not clear whether f will increase or decrease. Therefore, second-order information is necessary.

Theorem 2 *Second-Order Necessary Conditions.*
Suppose that x^* is a local solution of (4.1) and that the LICQ condition is satisfied. Let λ^* be the Lagrange multiplier vector for which the KKT conditions (4.7 - 4.11), are satisfied. Then

$$w^T \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*) w \ \geq 0, \ for \ all \ w \in \mathcal{C}(x^*, \lambda^*). \tag{4.16}$$

Unlike necessary conditions, which assume that x^* is a local solution and deduce properties of f and g_i, for the active indices i, *Sufficient Conditions* are conditions on f and $g_i, i \in \varepsilon \cup \mathcal{I}$, that ensure that x^* is a local solution of the problem (4.1). The second-order sufficient condition below looks very much like the necessary condition discussed above, however, it differs in that the constraint qualification is not required, and the inequality of (4.16) is replaced by a strict inequality.

Theorem 3 *Second-Order Sufficient Conditions.*
Suppose that for some feasible point $x^* \ \in \mathcal{R}^n$ there is a Lagrange multiplier vector λ^* such that the KKT conditions (4.7 - 4.11) are satisfied. Suppose also that

$$w^T \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*) w \ > 0, \ for \ all \ w \in \mathcal{C}(x^*, \lambda^*), \ w \neq 0. \tag{4.17}$$

Then x^* is a strict local solution for (4.9).

One situation in which the linearized feasible direction set $\mathcal{F}(x^*)$ is obviously an adequate representation of the actual feasible set occurs when all the active constraints are linear. Since the model we are considering in the following chapters is of that kind, we state the lemma 4 below.

Lemma 4 *Suppose that at some $x^* \ \in \Omega$, all active constraints $g_i(.), i \in \mathcal{A}(x^*)$, are linear functions. Then $\mathcal{F}(x^*) = T_\Omega(x^*)$.*

4.2 SQP Method

We consider the equality-constrained problem

$$\begin{aligned} &\min f(x) \\ &\text{subject to} \quad g(x) = 0, \end{aligned} \tag{4.18}$$

where $f : \mathcal{R}^n \rightarrow \mathcal{R}$ and $c : \mathcal{R}^n \rightarrow \mathcal{R}^m$ are smooth functions. The idea behind the SQP approach is to model (4.18) at the current iterate x_k by a quadratic programming subproblem, then use the minimizer of this subproblem to define a new iterate x_{k+1}. Yielding a good step for the nonlinear optimization problem is the main challenge in the design of the quadratic problem. The simplest derivation of SQP methods is to consider an application of Newton's method to the KKT optimality conditions for (4.18).

From (4.6), we know that the Lagrangian function for this problem is $\mathcal{L}(x, \lambda) = f(x) - \lambda^T g(x)$. We use $A(x)$ to denote the Jacobian matrix of the constraints, that is,

$$A(x)^T = [\nabla g_1(x),\ \nabla g_2(x), ...,\ \nabla g_m(x)], \tag{4.19}$$

where $g_i(x)$ is the ith component of the vector $g(x)$. The first-order (KKT) conditions (4.7 - 4.11) of the equality-constrained problem (4.18) can be written as a system of $n + m$ equations with the $n + m$ unknowns x and λ:

$$F(x, \lambda) = \begin{bmatrix} \nabla f(x) - A(x)^T \lambda \\ \\ g(x) \end{bmatrix} = 0. \tag{4.20}$$

Any solution (x^*, λ^*) of the equality-constrained problem (4.18) for which $A(x^*)$ has full rank satisfies (4.20). One approach that suggests itself is to solve the nonlinear equations (4.20) by using Newton's method.
The Jacobian of (4.20) with respect to x and λ is given by

$$F'(x, \lambda) = \begin{bmatrix} \nabla^2_{xx} \mathcal{L}(x, \lambda) & -A(x)^T \\ \\ A(x) & 0 \end{bmatrix} \tag{4.21}$$

The Newton step from the iterate (x_k, λ_k) is thus given by

$$\begin{bmatrix} x_{k+1} \\ \lambda_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ \lambda_k \end{bmatrix} + \begin{bmatrix} p_k \\ p_\lambda \end{bmatrix}, \tag{4.22}$$

where p_k and p_λ solve the Newton-KKT system

$$\begin{bmatrix} \nabla^2_{xx} \mathcal{L}_k & -A_k^T \\ \\ A_k & 0 \end{bmatrix} \begin{bmatrix} p_k \\ \\ p_\lambda \end{bmatrix} = \begin{bmatrix} -\nabla f_k + A_k^T \lambda_k \\ \\ -g_k \end{bmatrix} \tag{4.23}$$

This Newton iteration is well defined when the KKT matrix in (4.23) is nonsingular.

Assumptions 1 *(a) The constraint Jacobian $A(x)$ has full row rank;*
(b) The matrix $\nabla^2_{xx}\mathcal{L}(x,\lambda)$ is positive definite on the tangent space of the constraints, that is, $d^T\nabla^2_{xx}\mathcal{L}(x,\lambda)d > 0$ for all $d \neq 0$ such that $A(x)d = 0$.

There is an alternative way to view the iteration (4.22) - (4.23). Suppose that at the iterate (x_k, λ_k) we model problem (4.18) using the quadratic program

$$\begin{array}{ll} \min\limits_{p} & f_k + \nabla f_k^T\, p + \tfrac{1}{2}p^T\nabla^2_{xx}\mathcal{L}_k p \\ \text{s.t.} & A_k p + g_k = 0. \end{array} \tag{4.24}$$

If Assumptions 1 hold, then this problem has a unique solution (p_k, l_k) that satisfies

$$\nabla^2_{xx}\mathcal{L}_k p_k + \nabla f_k - A_k^T l_k = 0, \tag{4.25}$$

$$A_k p + g_k = 0. \tag{4.26}$$

The vectors p_k and l_k can be identified with the solution of the Newton equations (4.23). If we subtract $A_k^T\lambda_k$ from both sides of the first equation in (4.23), we obtain

$$\begin{bmatrix} \nabla^2_{xx}\mathcal{L}_k & -A_k^T \\ & \\ A_k & 0 \end{bmatrix} \begin{bmatrix} p_k \\ \\ \lambda_{k+1} \end{bmatrix} = \begin{bmatrix} -\nabla f_k \\ \\ -g_k \end{bmatrix} \tag{4.27}$$

Hence, by nonsingularity of the coefficient matrix, we have that $\lambda_{k+1} = l_k$ and that p_k solves (4.24) and (4.23). Therefore, (x_{k+1}, λ_{k+1}) can be seen either as the solution of (4.24) or as the Newton iteration (4.22) - (4.23). The latter approach facilitates the convergence analysis while the SQP framework allows us to define the following practical algorithm.

The SQP method in its simplest form is given below.

Algorithm 1 *Local SQP Algorithm for solving (4.18)*

```
Choose an initial pair $(x_0, \lambda_0)$; set $k \leftarrow 0$;
repeat until a convergence test is satisfied
        Evaluate $f_k, \nabla f_k, \nabla^2_{xx}\mathcal{L}_k, g_k$, and $A_k$;
        Solve (4.24) to obtain $p_k$ and $l_k$;
        Set $x_{k+1} \leftarrow x_k + p_k$ and $\lambda_{k+1} \leftarrow l_k$;
end(repeat)
```

4.3 Inequality Constraints

The SQP framework can be extended easily to the general nonlinear programming problem

$$\min f(x) \tag{4.28}$$

$$\text{s.t.} \quad g_i(x) = 0, \quad i \in \varepsilon \tag{4.29}$$

$$g_i(x) \geq 0, \quad i \in \mathcal{I} \tag{4.30}$$

Linearizing both the inequality and equality constraints, we get

$$\min_p f_k \ + \ \nabla f_k^T \ p \ + \ \frac{1}{2} p^T \nabla_{xx}^2 \mathcal{L}_k p \tag{4.31}$$

$$\text{s.t.} \quad \nabla g_i(x_k)^T p + g_i(x_k) = 0, i \in \varepsilon \tag{4.32}$$

$$\nabla g_i(x_k)^T p + g_i(x_k) \geq 0, i \in \mathcal{I} \tag{4.33}$$

The following theorem connects this problem with the previously considered equality constrained problem.

Theorem 4 *Suppose that x^* is a local solution of (4.28) at which the KKT conditions are satisfied for some λ^*. Suppose, too, that the linear independence constraint qualification (LICQ), the strict complementarity condition, and the second-order sufficient conditions hold at (x^*, λ^*). Then if (x_k, λ_k) is sufficiently close to (x^*, λ^*), there is a local solution of the subproblem (4.31)-(4.34) whose active set $\mathcal{A}_k$ is the same as the active set $\mathcal{A}^*(x)$ of the nonlinear program (4.28)-(4.30) at x^*.*

The problem we are going to model and solve in this thesis has equality, box and non-negativity constraints. Box constraints are relatively easy to solve, so we will concentrate on the problem

$$\min f(x) \tag{4.34}$$

$$s.t. \quad Ax = b \tag{4.35}$$

In SQP framework, our problem will be approximated by the quadratic model as usual but A_k and g_k appearing in (4.24) are the same through the whole process, i.e. $A_k = A, g_k = -b$.

Algorithm 1 is stated using $\nabla^2_{xx}\mathcal{L}_k$ in the quadratic model and assumes that (4.24) is solved to obtain the direction p_k (and the Lagrange multiplier l_k). We will look at these issues in more detail. First of all, step computation should be performed in a computationally affordable way and steps should yield global convergence. Derivation or calculation of exact $\nabla^2 xx\mathcal{L}_k$ can be cumbersome so some approximation of the Hessian is often used. Step computation is determined by KKT system (4.25)-(4.26). One obvious alternative is to solve the full $(n+m)\times(n+m)$ system with symmetric indefinite factorization or using some iterative method.

The matrix $\nabla^2_{xx}\mathcal{L}_k$ in the quadratic model (4.24) (being the exact Hessian) implies equivalence between SQP and Newton method applied to the first order optimality conditions. However, this matrix does not need to be easy to compute. So, alternative choices of the quadratic model can be quite useful. We will consider here Quasi-Newton approximations.

4.4 Full Quasi-Newton Approximations

The quadratic model (4.24) can be replaced by

$$m_k(p) = f_k + \nabla f_k^T \beta + \frac{1}{2} p^T B_k p \tag{4.36}$$

where B_k is some approximation of $\nabla^2_{xx}\mathcal{L}_k$. Quasi-Newton methods are quite a popular approach since they require only the gradient of the objective function - in our case the Lagrange function $\mathcal{L}$, to be supplied at each iteration. By measuring the changes in gradients, they construct a model of the objective function that is good enough to produce superlinear convergence. The most popular quasi-Newton method is the BFGS method while SR1 is a simpler rank-1 update. We will describe theoretical properties and implementational issues for these two quasi-Newton methods here.

Let B_k be a symmetric positive definite matrix that will be updated at every iteration. The minimizer p_k of (4.36) is clearly

$$p_k = -B_k^{-1} \nabla f_k^T \tag{4.37}$$

so using it as a search direction, the new iterate is

$$x_{k+1} = x_k + \alpha_k p_k \tag{4.38}$$

for the step length α_k approximately chosen. The new quadratic model, constructed at x_{k+1}, is of the same form

$$m_{k+1}(p) = f_{k+1} + \nabla f_{k+1}^T p + \frac{1}{2} p B_{k+1} p \tag{4.39}$$

Reasonable conditions to be imposed on B_{k+1} are that the gradient of m_{k+1} should match the gradient of the objective function f at the latest two iterates x_{k+1} and x_k. Since $\nabla m_{k+1}(0) = \nabla f_{k+1}$ and

$$\nabla m_{k+1}(-\alpha_k p_k) = \nabla f_{k+1} - \alpha_k B_{k+1} p_k = \nabla f_k. \tag{4.40}$$

we obtain

$$B_{k+1} \alpha_k p_k = \nabla f_{k+1} - \nabla f_k. \tag{4.41}$$

With notation

$$s_k = x_{k+1} - x_k, \quad y_k = \nabla f_{k+1} - \nabla f_k, \tag{4.42}$$

(4.40) becomes

$$B_{k+1} s_k = y_k, \tag{4.43}$$

what is known as the secant equation.

Imposing the condition of B_{k+1} being a symmetric and positive definite matrix, we end up with the curvature condition

$$s_k^T y_k > 0. \tag{4.44}$$

Therefore we can deduce that if the curvature condition is satisfied, the secant equation (4.43) always has a solution B_{k+1}, but it is not unique. In fact, (4.43) still admits an infinite number of solutions. To determine B_{k+1} uniquely, we impose the additional condition that among all symmetric matrices satisfying the secant equation, B_{k+1} is closest to the current matrix B_k. In other words we are solving the problem

$$\begin{aligned} &\min_B \|B - B_k\| \\ &s.t. \quad B = B^T, \quad B s_k = y_k. \end{aligned} \tag{4.45}$$

The solution of (4.45) depends on the norm one uses. Taking the weighted Frobenius norm defined by the average Hessian, the unique solution to (4.45) is

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}. \tag{4.46}$$

Clearly, the new update update differs from B_k by a rank-2 matrix.

There is a simpler rank-1 update that maintains symmetry of the matrix and allows it to satisfy the secant equation. But, this symmetric rank-1, update SR1 does not necessarily maintain positive definitiveness of the matrices B_{k+1}.

Starting from the general form

$$B_{k+1} = B_k + \sigma v v^T, \quad \sigma \in \{1, -1\} \tag{4.47}$$

and choosing σ and v such that B_{k+1} satisfies the secant equation we get

$$y_k = B_k s_k + [\sigma v^T s_k] v \tag{4.48}$$

so v must be a multiple of $y_k - B_k s_k$. Therefore, the only symmetric rank-1 updating formula that satisfies the secant equation is given by

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{(y_k - B_k s_k)^T y_k} \tag{4.49}$$

Both Quasi-Newton updates, BFGS and SR1 have advantages and disadvantages. For a detailed overview one can see Nocedal and Wright [44].

In SQP context, we are interested in minimizing the Lagrange function at each iteration i.e. our quadratic Quasi-Newton method should correspond to (4.24). Therefore, we will update B_k using (4.43) or (4.49) with

$$s_k = x_{k+1} - x_k, \quad y_k = \nabla_x \mathcal{L}(x_{k+1}, \lambda_{k+1}) - \nabla_x \mathcal{L}(x_k, \lambda_{k+1}) \tag{4.50}$$

We can view this process as the application of quasi-Newton updating to the case in which the objective function is given by the Lagrangian $\mathcal{L}(x, \lambda)$ (with λ fixed). This viewpoint immediately reveals the strengths and weaknesses of this approach.

If $\nabla^2_{xx}\mathcal{L}$ is positive definite in the region where the minimization takes place, then the BFGS quasi-Newton approximation B_k will reflect some of the curvature information of the problem, just as in the unconstrained BGFS method, and the iteration will converge robustly and rapidly. If, however, $\nabla^2_{xx}\mathcal{L}$ contains negative eigenvalues, then the BFGS approach of approximating it with a positive definitive matrix may be problematic. A requirement of BFGS updating is that s_k and y_k satisfy the curvature condition $s_k^T y_k > 0$, which may not hold when s_k and y_k are defined by (4.50), even when the iterates are close to the solution.

To overcome this difficulty, we could skip the BFGS update if the condition

$$s_k^T y_k \geq \theta s_k^T B_k s_k \tag{4.51}$$

is not satisfied, where θ is a positive parameter (10^{-2}, say). A more effective modification ensures that the update is always well defined by modifying the definition of y_k.

Procedure 1 *Damped BFGS Updating*
Given: symmetric and positive definite matrix B_k;
Define s_k and y_k as in (4.50) and set

$$r_k = \theta_k y_k + (1 - \theta_k) B_k s_k, \tag{4.52}$$

where the scalar θ_k is defined as

$$\theta_k = \begin{cases} 1, & \textit{if} \quad s_k^T y_k \geq 0.2 s_k^T B_k s_k, \\ (0.8 s_k^T B_k s_k)/(s_k^T B_k s_k - s_k^T y_k), & \textit{if} \quad s_k^T y_k < 0.2 s_k^T B_k s_k, \end{cases} \tag{4.53}$$

Update B_k as follows:

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{r_k r_k^T}{s_k^T r_k} \tag{4.54}$$

The formula (4.54) is simply the standard BFGS update formula, with y_k replaced by r_k. It guarantees that B_{k+1} is positive definite, since it is easy to show that when $\theta_k \neq 1$ we have

$$s_k^T r_k = 0.2 s_k^T B_k s_k > 0. \tag{4.55}$$

4.5 Merit Functions

To determine whether or not a trial should be accepted by SQP methods, a merit function is often used to make this decision. The merit function controls size of the step in the case of line search methods. In trust region methods it determines whether the step is accepted or rejected and whether the trust-region radius should be adjusted. A whole range of merit functions have been used in SQP methods, including augmented Lagrangians and nonsmooth penalty functions. We focus on exact, nonsmooth merit functions.

For the purpose of step computation and evaluation of a merit function, inequality constraints $g(x) \geq 0$ are often converted to the form

$$\tilde{g}(x, s) = g(x) - s = 0.$$

where $s \geq 0$ is a vector of slacks. (The condition $s \geq 0$ is typically not monitored by the merit function.) Therefore, in the discussion that follows we assume that all constraints are in the form of equalities, and we focus our attention on problem (4.18).

The l_1 merit function for (4.18) takes the form

$$\phi_1(x; \mu) = f(x) + \mu \|g(x)\|_1. \tag{4.56}$$

In a line search method, a step $\alpha_k p_k$ will be accepted if the following sufficient decrease condition holds:

$$\phi_1(x_k + \alpha_k p_k; \mu_k) \leq \phi_1(x_k; \mu_k) + \eta \alpha_k D(\phi_1(x_k; \mu); p_k), \quad \eta \in (0, 1), \tag{4.57}$$

where $D(\phi_1(x_k; \mu); p_k)$ denotes the directional derivative of ϕ_1 in the direction p_k. This condition is analogous to the Armijo conditions for unconstrained problems assuming that p_k is a descent direction, $D(\phi_1(x_k; \mu); p_k) < 0$. The condition holds for the penalty parameter μ large enough, as stated below.

Theorem 5 . *Let p_k and λ_{k+1} be generated by the SQP iteration (4.27). Then directional derivative of ϕ_1 in the direction p_k satisfies*

$$D(\phi_1(x_k; \mu); p_k) = \nabla f_k^T p_k - \mu \|g_k\|_1. \tag{4.58}$$

Moreover, we have that

$$D(\phi_1(x_k; \mu); p_k) \leq -p_k^T \nabla_{xx}^2 \mathcal{L}_k p_k - (\mu - \|\lambda_{k+1}\|_\infty) \|g_k\|_1 \tag{4.59}$$

.

4.6 Line Search and Trust-Region SQP Methods

There is a large variety of SQP methods depending on the way the Hessian approximation is computed, the choice of the merit function and the step-computation procedure. The algorithm presented below is a practical quasi-Newton algorithm for solving the problem (4.1) with equality and inequality constraints.

Algorithm 2 *Line Search SQP algorithm*

```
Choose parameters $\eta \in (0,\ 0.5)$, $\tau \in (0,\ 1)$, and an initial pair $(x_0, \lambda_0)$;
Evaluate $f_0, \nabla f_0, g_0, A_0$;
If a quasi-Newton approximation is used, choose an initial $n \times n$ symmetric
positive definite Hessian approximation $B_0$, otherwise compute $\nabla_x^2 x\mathcal{L}_0$;
repeat until a convergence is satisfied
    Compute $p_k$ by solving (4.13); let $\hat{\lambda}$ be the corresponding multiplier;
    Set $p_\lambda \leftarrow \hat{\lambda} - \lambda$;
    Choose $\mu_k$ to satisfy
        $\mu \geq \frac{\nabla f_k^T p_k + (\sigma/2) p_k^T \nabla_{xx}^2 \mathcal{L}_k p_k}{(1-p)\|g_k\|_1}$ with $\sigma = 1$;
    Set $\alpha_k = 1$;
    while $\phi_1(x_k + \alpha_k p_k; \mu_k) > \phi_1(x_k; \mu_k) + \eta \alpha_k D_1(\phi(x_k; \mu_k) p_k)$
        reset $\alpha_k \leftarrow \tau_\alpha \alpha_k$ for some $\tau_\alpha \in (0, \tau]$;
    end (while)
    Set $x_{k+1} \leftarrow x_k + \alpha_k p_k$ and $\lambda_{k+1} \leftarrow \lambda_k + \alpha_k p_\lambda$;
    Evaluate $f_{k+1}, \nabla f_{k+1}, g_{k+1}, A_{k+1}$, (and possibly $\nabla_{xx}^2 \mathcal{L}_{k+1}$);
    If a quasi-Newton approximation is used, set
        $s_k \leftarrow \alpha_k p_k$ and $y_k \leftarrow \nabla_x \mathcal{L}(x_{k+1}, \lambda_{k+1}) - \nabla_x \mathcal{L}(x_k, \lambda_{k+1})$
    and obtain $B_{k+1}$ by updating $B_k$ usign a quasi-Newton formula;
end (repeat)
```

In the algorithm presented above, the choice of quasi-Newton method for updating B_k as well as the choice of the merit function are left unspecified.

Another possibility for globalization of a local SQP method is to use a trust-region approach. This approach allows direct use of second-derivative information and can treat the case where active constraint gradients are linearly dependent.

Starting from the equality constrained problem and adding a trust-region constraint, we obtain the new model

$$\begin{aligned} \min_p \quad & f_k + \nabla_k^T p + \frac{1}{2} p^T \nabla_{xx}^2 \mathcal{L}_k p && (4.60) \\ \text{s. t. } & \nabla g_i(x_k)^T p + g_i(x_k) = 0, \quad i \in \varepsilon && (4.61) \\ & \nabla g_i(x_k)^T p + g_i(x_k) \geq 0, \quad i \in \mathcal{I} && (4.62) \\ & \|p\| \leq \Delta_k, && (4.63) \end{aligned}$$

The trust-region constraint (4.63) might imply that the problem does not have a solution. To resolve the possible conflict between the linear constraints (4.61)-(4.62)and the trust region constraint (4.63) one might argue that there is no reason to satisfy the linearized constraints at every step. We should aim to improve the feasibility of these constraints and satisfy them exactly only if the trust region constraints permit it. This argument leads to the three classes of methods: relaxation methods, penalty methods and filter methods. Details can be seen in Nocedal and Wright [44]. Here we will state a relaxation method for equality constrained optimization.

At the iterate x_k we compute the SQP step by solving the subproblem

$$\begin{aligned} \min_p \quad & f_k + \nabla_k^T p + \frac{1}{2} p^T \nabla_{xx}^2 \mathcal{L}_k p && (4.64) \\ \text{s. t. } & A_k p + g_k = r_k, && (4.65) \\ & \|p\|_2 \leq \Delta_k, && (4.66) \end{aligned}$$

One should try to choose r_k as the smallest vector such that (4.65)-(4.66) are consistent for some reduced value of trust region radius A_k. Therefore, we first solve the subproblem

$$\begin{aligned} \min_v \ & \|A_k v + g_k\|_2^2 && (4.67) \\ \text{s. t. } & \|v\|_2 \leq 0.8\Delta_k. && (4.68) \end{aligned}$$

If v_k is the solution of this subproblem, we define

$$r_k = A_k v_k + g_k \tag{4.69}$$

A merit function that fits well with this approach is the nonsmooth l_2 function $\phi_2(x; \mu) = f(x) + \mu \|g(x)\|_2$. We model it by

$$q_\mu(p) = f_k + \nabla f_k^T p + \frac{1}{2} p^T \nabla_{xx}^2 \mathcal{L}_k p + \mu m(p) \tag{4.70}$$

where

$$m(p) = \|g_k + A_k p\|_2. \tag{4.71}$$

We define the ratio to be monitored in the usual way,

$$p_k = \frac{ared_k}{pred_k} = \frac{\phi_2(x_k, \mu) - \phi_2(x_k + p_k, \mu)}{q_\mu(0) - q_\mu(p_k)} \tag{4.72}$$

the trust region due to Byrd-Omojokun is defined as follows.

Algorithm 3 *Byrd-Omojokun Trust-Region SQP Method*

```
 Choose constants $\epsilon > 0$ and $\eta, \gamma \in (0, 1]$;
Choose starting point $x_0$, initial trust region $\Delta_0 > 0$;
for $k = 0, 1, 2, ...$
    Compute $f_k, g_k, \nabla f_k, A_k$;
    Compute multiplier estimates $\hat{\lambda}_k = A(_{k-1}A_{k-1}^T)^{-1}A_{k-1}\nabla f_{k-1}$
    if $\|\nabla f_k - A_k^T\hat{\lambda}_k\|_\infty < \epsilon$ and $\|g_k\|_\infty < \epsilon$
        stop with approximate solution $x_k$;
    Compute normal subproblem (4.67)-(4.68) for $v_k$ and $r_k$ from (4.69)
    Compute $\nabla^2_{xx}\mathcal{L}_k$ or a quasi-Newton approximation;
    Compute $p_k$ by applying the projected CG method to

        $$\begin{aligned} \min_p \quad & f_k + \nabla_k^T p + \frac{1}{2}p^T\nabla^2_{xx}\mathcal{L}_k p \\ & s.\ t.\ A_k p + g_k = r_k, \\ & \|p\|_2 \le \Delta_k, \end{aligned}$$

    Choose $\mu_k$ to satisfy

        $$q_\mu(0) - q_\mu(p_k) \ge \rho\mu[m(0) - m(p_k)], \quad \text{for } \rho \in (0, 1).$$

    Choose $p_k = ared_k/pred_k$;
    if $\rho_k > \eta$
        Set $x_{k+1} = x_k + p_k$;
        Choose $\Delta_{k+1}$ to satisfy $\Delta_{k+1} \ge \Delta_k$;
    else
        Set $x_{k+1} = x_k$;
        Choose $\Delta_{k+1}$ to satisfy $\Delta_{k+1} \le \gamma\|p_k\|$;
end (for).
```

4.7 Convergence Analysis

Consider an SQP method that computes a search direction p_k by solving the quadratic program (4.31). We assume that the Hessian $\nabla^2_{xx}\mathcal{L}_k$ is replaced in (4.31) by some symmetric and positive definite approximation B_k. The new iterate is defined as $x_{k+1} + \alpha_k p_k$, where α_k is computed by a backtracking line search, starting from the unit step length, and terminating when

$$\phi_1(x_k + \alpha_k p_k; \mu) \leq \phi_1(x_k; \mu) - \eta\alpha_k(q_\mu(0) - q_\mu(p_k)),$$

where $\eta \in (0, 1)$, with ϕ_1 as defined in

$$\phi_1(x; \mu) = f(x) + \mu \sum_{i \in \varepsilon} |g_i(x)| + \mu \sum_{i \in \mathcal{I}} [g_i(x)]^-$$

and q_μ as defined in

$$\begin{aligned} \min_p \quad q_\mu(P) \equiv f_k + \nabla_k^T p + \frac{1}{2} p^T \nabla^2_{xx}\mathcal{L}_k p + \mu \sum_{i \in \varepsilon} |g_i(x) + \\ \nabla g_i(x_k)^T p| + \mu \sum_{i \in \mathcal{I}} [g_i(x) + \nabla g_i(x_k)^T p]^- \\ \text{s. t.} \quad \|p\|_\infty \leq \Delta_k, \end{aligned} \tag{4.73}$$

for some penalty parameter μ, where we use the notation $[y]^- = max\{0, -y\}$. To establish the convergence result, we assume that each quadratic program (4.31) is feasible and determines a bounded solution p_k. We also assume that the penalty parameter μ is fixed for all k and is sufficiently large.

Theorem 6 *Suppose that the SQP algorithm just described is applied to the nonlinear program (4.28). Suppose that the sequence $\{x_k\}$ and $\{x_k + p_k\}$ are contained in a closed, convex region of $\mathcal{R}^n$ in which f and g_i have continuous first derivatives. Suppose that the matrix B_k and multipliers are bounded and that μ satisfies $\mu \geq \|\lambda_k\|_\infty + \rho$ for all k, where ρ is a positive constant. Then all limit points of the sequence $\{x_k\}$ are KKT points of the nonlinear program (4.28)-(4.30).*

In order to prove superlinear and quadratic convergence we need to assume the following.

Assumptions 2 *The point x^* is a local solution of problem (4.18) in which the following conditions hold.*

(a) The function f and c are twice differentiable in a neighborhood of x^* with Lipschitz continuous second derivatives.

(b) The linear independence constraint qualification holds at x^*. This condition implies that the KKT conditions (4.7) - (4.11) are satisfied for some vector of multipliers λ^*.

(c) The second-order sufficient conditions hold at (x^*, λ^*).

Theorem 7
Suppose that Assumptions 2 hold. Then, if (x_0, λ_0) is sufficiently close to (x^*, λ^*) the pairs $\{(x_k, \lambda_k)\}$ generated by Algorithm 1 converge quadratically to (x^*, λ^*).

Theorem 8
Suppose that Assumptions 2 hold and that the iterates $\{x_k\}$ generated by Algorithm 1 with quasi-Newton approximate Hessian B_k converge to x^*. The $\{x_k\}$ converges superlinearly if and only if the Hessian approximation B_k satisfies

$$\lim_{k\to\infty} \frac{\|P_k(B_k - \nabla^2_{xx}\mathcal{L}_*)(x_{k+1} - x_k)\|}{\|x_{k+1} - x_k\|} = 0 \tag{4.74}$$

Theorem 9 *Suppose that Assumption 2 hold. Assume also that $\nabla^2_{xx}\mathcal{L}_*$ and B_0 are symmetric and positive definite. If $\|x_0 - x_*\| and \|B_0 - \nabla^2_{xx}\mathcal{L}_*\|$ are sufficiently small, the iterates $\{x_k\}$ generated by Algorithm 1 with BFGS Hessian approximation B_k defined by (4.50) and (4.54) (with $r_k = s_k$)satisfy the limit (4.74). Therefore, the iterates $\{x_k\}$ converge superlinearly to x^*.*

$$\lim_{k\to\infty} \left[\tfrac{P_k(B_k - \nabla^2_{xx}\mathcal{L}_*)P_k(x_{k+1} - x_k)}{\|x_{k+1} - x_k\|} + \tfrac{P_k(B_k - \nabla^2_{xx}\mathcal{L}_*)(I - P_k)(x_{k+1} - x_k)}{\|x_{k+1} - x_k\|}\right] = 0,$$

Theorem 10 *Suppose that Assumption 2(a) holds and that the matrices B_k are bounded. Assume also that the iterate $\{x_k\}$ generated by Algorithm 1* *with approximate Hessian B_k converge to x^*, and that*

$$\lim_{k\to\infty} \frac{P_k(B_k - \nabla^2_{xx}\mathcal{L}_*)P_k(x_{k+1} - x_k)}{\|x_{k+1} - x_k\|} = 0,$$

Then the sequence x_k converges to x^* two-step superlinearly, that is,

$$\lim_{k\to\infty} \frac{\|x_{k+2} - x^*\|}{\|x_k - x^*\|} = 0$$

Chapter 5

Optimal Execution: Single-Market Orders

5.1 Preliminaries

The main objective in execution is to achieve the most efficient price. We propose two optimal strategies for the execution of atomic orders based on the 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. 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.

The model we present herein 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 a Fill Probability function that gives the probability of being filled (executed) for limit orders. Such a function is not available analytically but it can be reasonably well estimated given a set of market conditions. 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.

5.2 Single-period 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 a 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 a given time. We will assume that the order book has n visible levels with price trajectories following the arithmetic random walk given by (2.3)-(2.5). 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 Almgren and Chriss [4] and explained with (2.2). 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 $[0, T]$, see (3.3).

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

$$y + \sum_{i=1}^{n} x_i = Q. \tag{5.1}$$

The execution of limit orders is an uncertain event. Let $\Gamma = (\Gamma_1, \dots, \Gamma_n)$ be a stochastic variable which denotes the filled quantity (in relative terms) at each bid level during $[0, T]$ and let $\gamma = (\gamma_1, \dots, \gamma_n)$ be a realization of Γ. At the end of time window, $t = T$, we are left with the unfilled residual

$$R = Q - \sum_{i=1}^{n} \gamma_i x_i - y \tag{5.2}$$

and we will trade that residual as 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

$$\pi(y) = f(y)y. \tag{5.3}$$

Limit orders have gains according to their respective gain coefficients if they are filled, and an opportunity cost if unfilled within $[0, T]$. The residual given by (5.2) 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 a larger impact due to the larger intensity of trade (smaller average traded volume within that time window). Let $\Pi(R)$ denote that impact. With $G_i(q)$ defined by (3.5) as $G_i(x_i) = c_i x_i F_i(x_i)$, $c_i = a_i(0) - b_i(0)$ and with the assumptions made in the previous section, we can formulate the gain of limit orders as

$$\sum_{i=1}^{n} G_i(x_i). \tag{5.4}$$

Residual R is clearly a stochastic variable depending on Γ. Volatility risk is dependent on price trajectories (2.3)-(2.5) and we will denote it by $V(R), V(R) = (P(T) - P(0))R$. 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 the 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 Birge and Louveaux [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 a two-stage stochastic problem is not an affordable option.

Because of all these reasons we will define a deterministic model which has good theoretical properties and agrees with intuitive risk averse behaviour of traders.

Instead of considering the volatility risk of the residual as a 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 (2.5) but volatility of price plays a more important role within a 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 5.3 since $\Phi(1) > 0.9$, with Φ cumulative distribution function for ζ.

Similarly to the gain function (3.5), instead of considering the residual as a stochastic variable, we define the *residual function* as a deterministic function,

$$r(x, y) = Q - \sum_{i=1}^{n} H_i(x_i) - y \tag{5.5}$$

with $H_i(x_i) = x_i F_i(x_i)$. These simplifications and taking the linear impact function we are able to state the impact and volatility costs as follows,

$$V(r(x, y)) = \sigma\sqrt{T} r(x, y) \tag{5.6}$$

and

$$\pi(y) = (\varepsilon + \mu y)y,\ \Pi(r) = (\varepsilon + \eta r)r. \tag{5.7}$$

The constants μ and η are dependent on the time duration for execution of the corresponding market orders and the average traded volumes within these time windows. Therefore a larger intensity of trade (shorter execution time) of the residual implies $\eta > \mu$, while ε is the average[1] historical spread value. Putting together all analyzed costs and gains with

$$\phi(x, y) = -\sum_{i=1}^{n} G_i(x_i) + \pi(y) + \sigma\sqrt{T} r(x, y) + \Pi(r(x, y)), \tag{5.8}$$

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

our problem is

$$\min_{x,y} \quad \phi(x,y) \tag{5.9}$$

$$\text{s.t.} \quad Q = \sum_{i=1}^{n} x_i + y \tag{5.10}$$

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

Problem (5.9)-(5.10) is a nonlinear optimization problem with a single linear constraint and nonnegativity constraints. 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 implies the application of second order conditions and every KKT point is a minimizer of (5.9)-(5.10). Let $\mathcal{R}_0$ be the set of nonnegative real numbers.

Theorem 11 *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

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

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

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

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

Therefore we can write

$$\nabla^2\phi(x,y) = D + uu^T, \ D = \operatorname{diag}(-A_1H_1''(x_1),\ldots,-A_nH_n''(x_n), 2\mu)$$

with

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

Since uu^T is positive semi-definite, it is sufficient to prove that D is positive definite. As D is diagonal we must infer 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. □

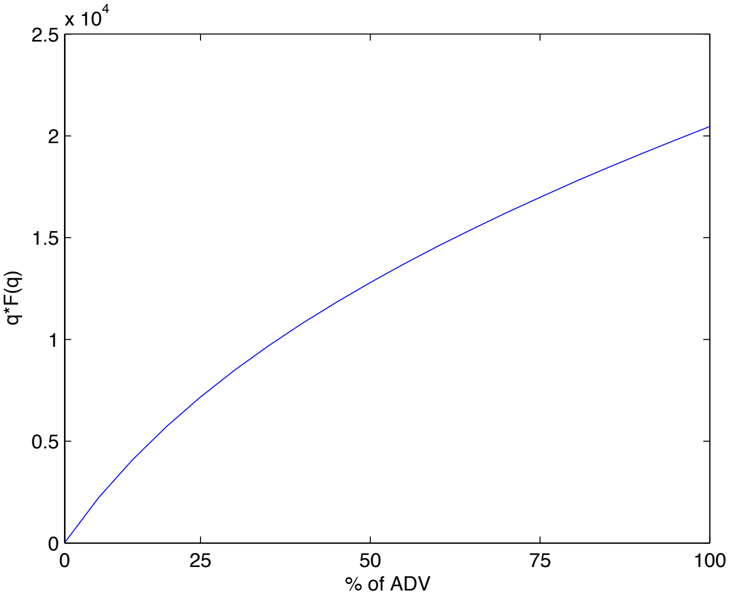


Figure 5.1: Empirical Success function. Empirijska funkcija uspešnosti

We can not claim that the concave condition from this theorem is satisfied for success functions H_i as 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'' cannot be determined from this information. But empirical results give 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 5.1.

5.3 Multi-period model

The time window for execution of an atomic order is generally small, say around 10 minutes. However if we are buying a liquid but volatile stock we might find that too long a time window to be waiting to see if orders will be filled according to our expectations. The market conditions can change significantly and the strategy obtained from (5.9)-(5.10) might be subject to re-optimization at a 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 requires 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 the two-period model could be easily translated into a multi-period model with as many re-optimizations as appropriate.

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, cancellations, new liquidity arrivals, 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 (5.9)-(5.10) at $t=0$ are denoted by y^0 and x^0, while the Fill Probability functions for $[0,T]$ are F_i.

At $t=\tau$ we know the volume that is already traded so we have to trade some $Q^\tau,\ Q^\tau \leq Q$, within $[\tau,T]$. Also, for all x_i^0 initially posted at bid levels $i \in B_0$, the unfilled amount $\tilde{x}_i,\ \tilde{x} \leq x_i^0$ is known. Reasoning the same way 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 have progressed in their queues. We can also consider cancellation 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 no longer visible. When canceling unfilled orders we are losing our place in the queue. Placing a new limit order means that we would be going to the end of the existing queue. Clearly, unfilled orders that are orders originally 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 transition 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

$$\ell_i^\tau \geq 0,\ \ell_i^\tau \leq \tilde{x}_i\ \ i \in B_0. \tag{5.11}$$

These orders will have success rate functions

$$\tilde{H}_i^\tau(\ell_i^\tau) = \tilde{F}_i^\tau(\ell_i^\tau)\ell_i^\tau \tag{5.12}$$

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

$$c_i^\tau = a_1(\tau) - b_i(\tau),\ i \in B_0. \tag{5.13}$$

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

$$H_k^\tau(x_k^\tau) = F_k^\tau(x_k^\tau) x_k^\tau, \tag{5.14}$$

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

$$c_k^\tau = a_1(\tau) - b_k(\tau),\ k \in B_\tau. \tag{5.15}$$

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 analogously to (5.5) is:

$$\rho(l^\tau, x^\tau, y^\tau) = Q^\tau - \sum_{i \in B_0} \tilde{H}_i^\tau(\ell_i^\tau) - \sum_{k \in B_\tau} H_k^\tau(x_k^\tau) - y^\tau. \tag{5.16}$$

The optimization problem now becomes

$$\begin{aligned} \min_{l^\tau, x^\tau, y^\tau} \quad & \Phi(\ell^\tau, x^\tau, y^\tau) && (5.17) \\ \text{s.t.} \quad & \ell_i^\tau \in [0, \tilde{x}_i],\ i \in B_0 && (5.18) \\ & Q^\tau = y^\tau + \sum_{i \in B_0} \ell_i^\tau + \sum_{k \in B_\tau} x_k^\tau \\ & x^\tau, y^\tau \geq 0 \end{aligned}$$

with

$$\begin{aligned} \Phi(\ell^\tau, x^\tau, y^\tau) \quad = \quad & -\sum_{i \in B_0} \tilde{G}_i^\tau(\ell_i^\tau) - \sum_{k \in B_\tau} G_k^\tau(x_k^\tau) + \pi^\tau(y^\tau) + \qquad (5.19) \\ & \sigma \rho(l^\tau, x^\tau, y^\tau) \sqrt{T - \tau} + \Pi^\tau(\rho(l^\tau, x^\tau, y^\tau)) \end{aligned}$$

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., smaller average traded volume within shorter execution window for the residual ρ.

The problem (5.16)-(5.19) has the same structure as (5.9)-(5.10) 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 12 *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 (5.9)-(5.10) on consecutive time intervals $[0, \tau]$ and $[\tau, T]$. Re-optimization of the execution trajectory according to (5.16)-(5.19) allows an important advantage due to the fact that we can keep initially placed orders in the queue if the 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 (5.9)-(5.10) at $t = 0$ and then (5.16)-(5.19) at $t = \tau$ is better than applying (5.9)-(5.10) 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 to disjointed $[0, \tau]$ and $[\tau, T]$. On the other hand, market orders y^0 and y^τ are always realized according to some predefined schedule, (see Almgren and Chriss [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 the re-optimization procedure as many times as we want within $[0, T]$. We report numerical results for $\tau = T/2$ in a later section. We have also tested 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 actually chases high-frequency noise and thus loses the main advantage of this approach: Fill Probability function and a combination of market and limit orders.

Chapter 6

Optimal Execution: Multi-Market Orders

6.1 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.

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 spreads are defined as

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

for the two orderbooks in market A and B. Because two venues could work with different price granularity, tick size, the spreads in the two markets most often differ. However, in any two liquid securities in liquid venues, the spread is expected to be similar. But there are clear examples of tick size being significantly different, as in the case of for instance Deutsche Telecom (ticker: DTEGn.DE), where the tick size on the small venue Chi-X is 0.001 with a typical quote like 9.594/9.595 for $1500X1700$ shares. The same security which trades at the primary exchange, Xetra, with 0.005 in tick size, may have a quote of say 9.590/9.595 for $18000X21000$ shares. As can be seen, although the Chi-X spread is much tighter, if one wishes to trade non-trivial size, one would have to trade through many price levels to complete the order in Chi-X, while it could be done at the same level on Xetra.

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 instantly. 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 two-market situation we are dealing with two sequences of gain coefficients

$$c_i^A = a_1^A - b_i^A,\ c_i^B = a_1^B - b_i^B,\ i = 1, \ldots, n. \tag{6.1}$$

Obviously gain (6.1) occurs only if the order is filled within given time. We will define gain function for limit orders as follows.

Clearly each market has its own set of Fill Probability functions, F_i^A and F_i^B. The fill probability function for each of the markets are calculated independently of the others.

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

$$H_i^A(q) = qF_i^A(q),\ H_i^B(q) = qF_i^B(q),\ i = 1, \ldots, n \tag{6.2}$$

and *gain functions* as

$$G_i^A(q) = c_i^A H_i^A(q),\ G_i^B(q) = c_i^B H_i^B(q),\ i = 1, \ldots, n \tag{6.3}$$

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 results chapter.

6.2 Single-period model

The problem we 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 Kumaresan and Krejić [38] 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 strategy we want to pursue consists of distribution of orders into market and limit orders at both venues within $[0, T]$ with the aim of buying Q. Market orders are causing the costs represented as market impact and limit orders are facing uncertainty of being filled within $[0, T]$ and providing possible gain governed by their gain coefficients. Thus the question we are facing is distribution between market and limit orders and distribution between venues A and B. 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 aggressive manner i.e. using only market orders. This will produce larger impact and is subject to volatility risk since the prices $P(0)$ and $P(T)$ will very likely be different. Furthermore, the 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 portfolio of securities. Hence the same simplifications in modeling as in the previous chapter are necessary.

We will adopt the gain and success functions as already defined in the previous chapter. 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 modeled as the deterministic functions (2.2) 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 if compared with strategy from Kumaresan and Krejić [38]. The key difference in the two venue situation is coming from the residual carrying its volatility risk and impact costs. The residual is clearly an unknown stochastic 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 adopting the risk averse attitude and assuming that the price will move away from us for the whole σ. With such assumption we are covering more than 90% of cases under the process (2.5). 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, 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 Almgren and Chriss [4] and explained with (2.2). Risk free arbitrage opportunities will force the prices in the two venues to be aligned and as such the volatilities of the different venues will be 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 will 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. We will 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

$$\bar{R} = Q - \sum_{i=1}^{n} \gamma_i^A x_i^A - \sum_{i=1}^{n} \gamma_i^B x_i^B - y^A - y^B \tag{6.4}$$

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

$$\pi^A(y^A) = (\varepsilon^A + \mu^A y^A)y^A, \tag{6.5}$$

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

$$\pi^B(y^B) = (\varepsilon^B + \mu^B y^B)y^B. \tag{6.6}$$

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 (6.4) 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 $G_i(q)$ defined by (6.3) 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 the previous section, we can formulate the gain of limit orders as

$$G^A(x^A) = \sum_{i=1}^{n} G_i^A(x_i^A),\ G^B(x^B) = \sum_{i=1}^{n} G_i^B(x_i^B). \tag{6.7}$$

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 (2.5) but volatility of price plays a more important role within short time framework and therefore we adopt this risk averse attitude.

Analogously to gain function (6.3), instead of considering the residual as a stochastic variable, we define the *residual function* as deterministic function,

$$R(x, y) = Q - H^A(x^A) - H^B(x^B) - y^A - y^B, \tag{6.8}$$

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

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 of the residual orders as

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

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

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

Denoting

$$\begin{aligned} \varphi(x, y) \quad = \quad & \pi^A(y^A) + \pi^B(y^B) - G^A(x^A) - G^B(x^B) + \\ & \sigma\sqrt{T}R(x, y) + \Pi^A(r^A) + \Pi(r^B), \end{aligned} \tag{6.9}$$

our problem yields the following bi-level optimization problem

$$\begin{aligned} \min_{x,y} \quad & \varphi(x, y) && (6.10) \\ \text{s.t.} \quad & H^A(x^A) - H^B(x^B) - y^A - y^B - Q \ = 0 && (6.11) \\ & r = arg\min\nolimits_r \phi(r(x, y)) && (6.12) \\ & r^A + r^B = R(x, y) && (6.13) \\ & x, y, r \geq 0 && (6.14) \end{aligned}$$

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 13 *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 6.12-6.13. Then $\nabla^2 \varphi(x, y)$ is a positive definite matrix.*

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

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

with $B = diag(\eta^A, \eta^B)$ and $d = (\varepsilon^A, \varepsilon^B)$. As B is positive definite the minimizer of (6.12)-(6.14) is given by

$$r = \frac{R + e^T d}{e^T B^{-1} e} B^{-1} e - d, \; e = (1,1)^T \tag{6.15}$$

with $R = R(x, y) = Q - H^A(x^A) - H^B(x^B) - y^A - y^B$. Plugging (6.15) back to (6.10) - (6.11), 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

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

where D is the diagonal matrix with elements

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

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

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

and

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

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

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.

6.3 Multi-period model

After placing the orders at t_0, if the market price moved away, one may then want to revise ones' initial order placements by finding another optimal placement 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 (6.10)-(6.14) 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 (6.10) - (6.14) 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 can not 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 implies the following inequality conditions on the limit orders we will keep as initially placed

$$\ell_i^{\tau,A} \geq 0,\ \ell_i^{\tau,A} \leq \tilde{x}_i^A\ \ell_i^{\tau,B} \geq 0,\ \ell_i^{\tau,B} \leq \tilde{x}_i^B\ i \in B_0. \tag{6.16}$$

These orders will have success functions

$$\tilde{H}_i^{\tau,A}(\ell_i^{\tau,A}) = \tilde{F}_i^{\tau,A}(\ell_i^{\tau,A})\ell_i^{\tau,A},\ \tilde{H}_i^{\tau,B}(\ell_i^{\tau,B}) = \tilde{F}_i^{\tau,B}(\ell_i^{\tau,B})\ell_i^{\tau,B} \tag{6.17}$$

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

$$c_i^{\tau,A} = a_1^A(\tau) - b_i^A(\tau),\ c_i^{\tau,B} = a_1^B(\tau) - b_i^B(\tau),\ i \in B_0. \tag{6.18}$$

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

$$H_k^{\tau,A}(x_k^{\tau,A}) = F_k^{\tau,A}(x_k^{\tau,A})x_k^{\tau,A},\ H_k^{\tau,B}(x_k^{\tau,B}) = F_k^{\tau,B}(x_k^{\tau,B})x_k^{\tau,B}, \tag{6.19}$$

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

$$c_k^{\tau,A} = a_1^A(\tau) - b_k^A(\tau),\ c_k^{\tau,B} = a_1^B(\tau) - b_k^B(\tau),\ k \in B_\tau. \tag{6.20}$$

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 (6.8),

$$\rho(l^\tau, x^\tau, y^\tau) = Q^\tau - \tilde{H}^{\tau,A}(\ell^{\tau,A}) - \tilde{H}^{\tau,B}(\ell^{\tau,B}) - H^{\tau,A}(x^{\tau,A}) - H^{\tau,B}(x^{\tau,B}) - y^{\tau,A} - y^{\tau,B}, \tag{6.21}$$

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 two level bilevel problem.

The optimization problem now becomes

$$\begin{aligned}
\min_{l^\tau, x^\tau, y^\tau} \quad & \Phi(\ell^\tau, x^\tau, y^\tau) && (6.22)\\
\text{s.t.} \quad & \ell_i^\tau \in [0, \tilde{x}_i],\ i \in B_0 && (6.23)\\
& Q^\tau = y^{\tau,A} + y^{\tau,B} + \sum_{i \in B_0} (\ell_i^{\tau,A} + \ell_i^{\tau,B}) + \sum_{k \in B_\tau} (x_k^{\tau,A} + x_k^{\tau,B}) && \\
r^\tau \in \quad & arg \min \Pi^{\tau,A}(r^{\tau,A}) + \Pi^{\tau,B}(r^{\tau,B}) && (6.24)\\
\rho^\tau = \quad & r^{\tau,A} + r^{\tau,B} && (6.25)\\
& x^\tau, y^\tau \geq 0 &&
\end{aligned}$$

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 (6.22)-(6.25) has the same structure as (6.10)-(6.14) 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 14 *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.*

Chapter 7

Simulation Framework

All numerical results presented here are derived from simulations. A simulator was written in Java and MATLAB for this purpose. Since our research topic originated from the urgent 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 to be used in actual trading. [1]

7.1 Tick Data

The data used in the simulation is European level-1 and level-2 tick data provided by Reuters. This consists of 5 levels of orderbook depth with consolidated volume on each price level as well as traded price and quantity. Securities considered are the following five: VOD, AAL, KGF, SDR and SASY. The period in question is January - March 2008 and August - October 2009. Simulations were run from 08:15 until 16:30 everyday with continuous tick steam - every single tick was considered.

7.2 Simulated Orderbook

The simulator requires the state of the market in the form of an orderbook providing a snapshot of the market with tick level granularity. The simulator

[1] In fact, the models formulated here are actually in use at TransMarket Group globally.

then decomposes the orderbook into a stream of tick changes. These ticks are then used 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 at a given price level. This level of distinction is particularly important for the fill probability model.

When the trading models within the simulator harness place an order, the order is added to the back 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 cancellation 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, and assume that all canceled orders were behind ours if there were any, hence not change ahead quantity. Otherwise, if the cancellation took place ahead of us, naturally the ahead quantity will be reduced.

7.3 Fill Assumptions

In our simulated orderbook, an order x_i at price level i is filled when the quantity traded at that price level exceeds the ahead quantity (X_i). If the volume behind is less than x_i, naturally, $X_i{+}x_i$ will not take place. Therefore, we will look to the price level j right below i (i.e. $i+1$) for the residual trades to take place.

In the event of multiple orders working in the same price level or at different price levels, the same logic is applied. However, our order quantity will be included in the criteria for how much must trade before orders start filling.

7.4 Re-Streaming

The data streamed from different exchanges contain some noise. In essence, orderbook changes are not disseminated chronologically. At time, the time stamping of the events are also somewhat arbitrary within a small window of time from when the event in question took place. For instance, a trade of 100 shares on the bid should reduce the volume on best bid by 100 shares.

However, this is not always the case.

Since the only way to tell whether a change to the orderbook is a trade or a cancellation is from whether there was trade for the same quantity at the same price level or not, the task of determining changes to the touch is due to a cancellation or a trade becomes very difficult. The interleaving of changes to the orderbook and a potential corresponding trade with other changes makes the problem even worse. This is particularly problematic when one consider that there can be tens or even hundreds of trades and changes to orderbook taking place within a one second duration at busy times.

To further complicate the matter, some exchanges such as Euronext sometimes reports a single trade broken into two or three smaller trades. However, the change to the orderbook will be a single change equivalent to the sum of the smaller trades. These are merely some of the many intricate peculiarities of the exchanges.

We use a proprietary data cleaning 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%.

It should be stressed that re-streaming of this type is critical to the success of most high frequency models.

7.5 Static Variables

The optimisation uses a number of key static variables. There are defined as:

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

- Intraday Volatility
  The objective functions in our models use the intraday volatility to estimate the short term volatility risk. We calculate this from 90 days of historic data for non-overlapping 15 minute intervals. 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-minute time slice.

- Market Impact Model Coefficients
  Based on thousands of actual trades of 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.

7.6 Benchmarking

Accurately measuring the performance of trading algorithms is a very difficult task due to a number of random variables involved in the execution task. In the case of implementation shortfall, the arrival price is the reference benchmark price. If the price moved away during the execution period, we would expect the fill price to be worse than the reference price. In the event the price came in the order's favor, we would naturally expect an average fill price better than that of the reference price. The overall performance of the algorithm can therefore only be obtained by calculating the mean from a large sample of execution. Any single user of an algorithm will usually not have a sufficient number of trade samples to calculate the mean. The performance measurement is further complicated by structural changes in the market and the continuous evolution of the algorithms.

We are faced with two key obstacles when looking to benchmark our proposed optimal framework. Firstly, the proposed framework has to be better than market practice. This is a near impossibility since the trading algorithms are proprietary. Even the claimed performance numbers are not valid as they are marketing material as opposed to being factual. Secondly, unlike real life, the simulated algorithm does not distort the orderbook, it instead quantifies the disturbance caused by the order as additional impact costs.

Herein, we have only considered order sizes 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 orders will look attractive, not incorporating the significant effects of our trades into the simulation will make the results depart from our aim to be consistent 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 we could have achieved if we 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 measures, 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 a market order at the end, while P_M is obtained from the uniform trajectory of market orders only.

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

Chapter 8

Numerical results

We tested all the previously mentioned algorithms for 5 stocks chosen to cover the whole spectrum of liquidity with VOD being very liquid, SDR very illiquid, AAL and KGF medium liquid and SASY 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 being the most volatile one. The mean spreads are also quite different, varying from 23bp for SDR with standard deviation of 18 bp to 8 bp for VOD with standard deviation of 4 bp. The size of spread and its deviation directly influences the gain coefficients in our models.

8.1 Single-Market

The results are given in Tables 1-5. We considered 3 months worth of data (January to March 2008), with each day sliced into 61 time slots of 8 minutes, from 08:16 to 16:24 hrs. In all these tables, the first column gives the order size expressed as a percentage of average traded quantity of the execution period. 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 $B = (P_B - P_{perfect})/P_{perfect}$, where P_B is the best bid. The cost of optimal strategy coming from (5.9)-(5.10) is $SMSP$ (Single Market Single Period) given in column four and the cost of two-period optimal strategy (5.16)-(5.19), $SMMP$ (Single Market Multi period), 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 Prob-

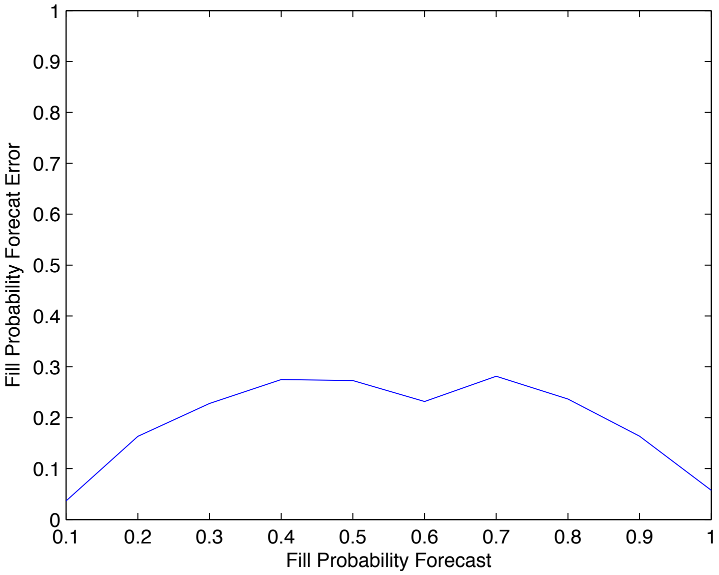


Figure 8.1: Mean error of the Fill Probability model. Srednja greška Fill Probability modela.

ability we are using is illustrated in Figure 8.1. For 10% of ADV of VOD we plot the mean error between forecasted F_i by our model and the realized fill rate for the whole tested range. The cumulative results for the whole considered period are illustrated graphically at Figure 8.2 for 10% of ADV for VOD

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 common market practice (algorithms M and B) for these criteria.

The difference between the performance of the single period model and two-period model is evident when we looks in Tables 1-5. We give more details taking the example of 10% ADV for SASY 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 was realized while the unrealized limit orders were $\tilde{x}_1 = 14.8\%$, $\tilde{x}_2 =$

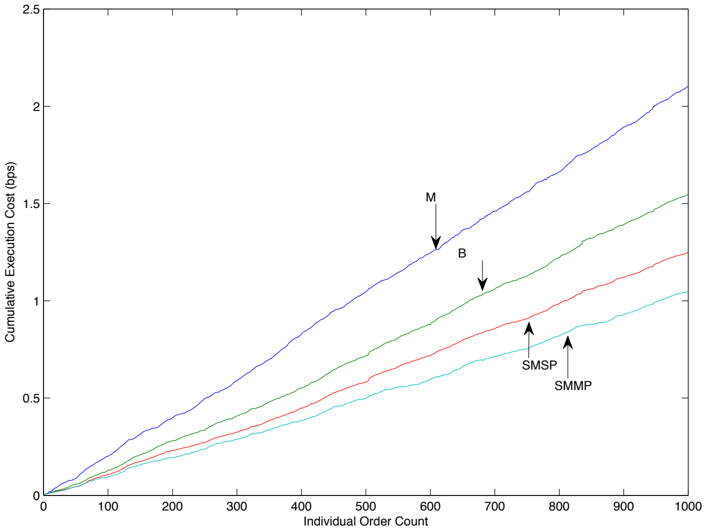


Figure 8.2: Performance comparison for VOD for trading 10% of ADV. Poredjenje za VOD za trgovanje 10% ADV.

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 cancellations 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 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 the 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 significantly better than common market practice. In addition, they 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. This is caused by the quadratic impact cost, so any decrease in cost due to the decrease of market orders and increase of limit orders is more significant.

The same behaviour can be seen if we consider the gain achieved by the two-period procedure against single period for four stocks but not for SDR. 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. This behaviour is 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 a large unfilled amount. By re-optimization we are actually chasing the noise since 4 minutes is not an optimal point for re-evaluation of the market conditions. Therefore we end up sending a 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 than the rate within 4 minutes.

% of ADV	M	B	$SMSP$	$SMMP$	$B - SMSP$	$SMSP - SMMP$
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

% of ADV	M	B	$SMSP$	$SMMP$	$B - SMSP$	$SMSP - SMMP$
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

% of ADV	M	B	$SMSP$	$SMMP$	$B - SMSP$	$SMSP - SMMP$
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

% of ADV	M	B	$SMSP$	$SMMP$	$B - SMSP$	$SMSP - SMMP$
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

% of ADV	M	B	$SMSP$	$SMMP$	$B - SMSP$	$SMSP - SMMP$
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

*	M	B	$SMSP$	$SMMP$
VOD	14.4	12.8	10.3	9.9
AAL	18.7	15.4	15.6	11.4
SASY	15.4	14.8	11.5	11.4
KGF	21.7	18.1	15.6	15.8
SDR	21.3	16.9	12.5	13.5

Table 6: Standard deviation of execution costs

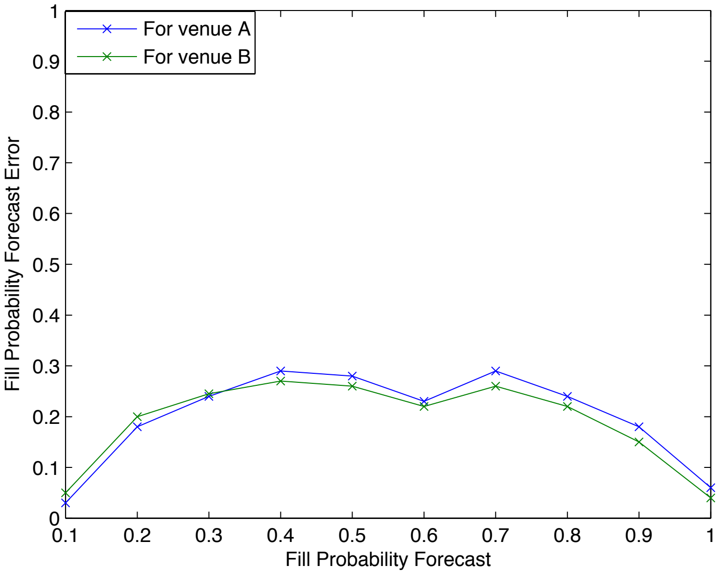


Figure 8.3: Mean Error of the Fill Probability Model for multiple venues. Srednja greška Fill Probability modela za više tržišta

8.2 Multi-Market

The results are given in Tables 7-11. We considered 3 months worth of data (August to October 2009) from LSE and Euronext. Each day is 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 terms $MMSP$ and $MMMP$ are acronyms for Multi-Market Single-Period and Multi-Market Multi-Period optimal execution strategies strategies.

% of ADV	M	B	$MMMP$
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 7: VOD

% of ADV	M	B	$MMMP$
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 8: AAL

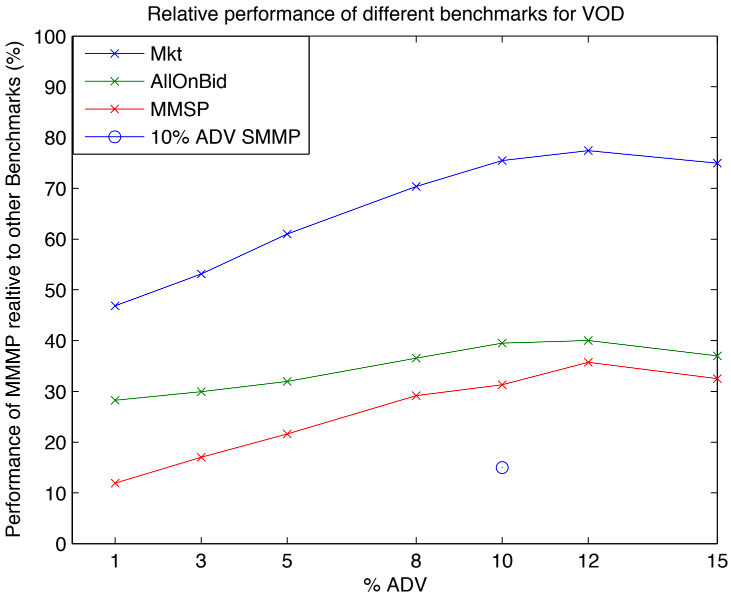


Figure 8.4: Performance comparison of trading VOD in two venues. Poredjenje za trgovanje VOD na dva tržišta.

% of ADV	M	B	$MMMP$
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

Table 9: SASY

% of ADV	M	B	$MMMP$
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 10: KGF

% of ADV	M	B	$MMMP$
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 11: SDR

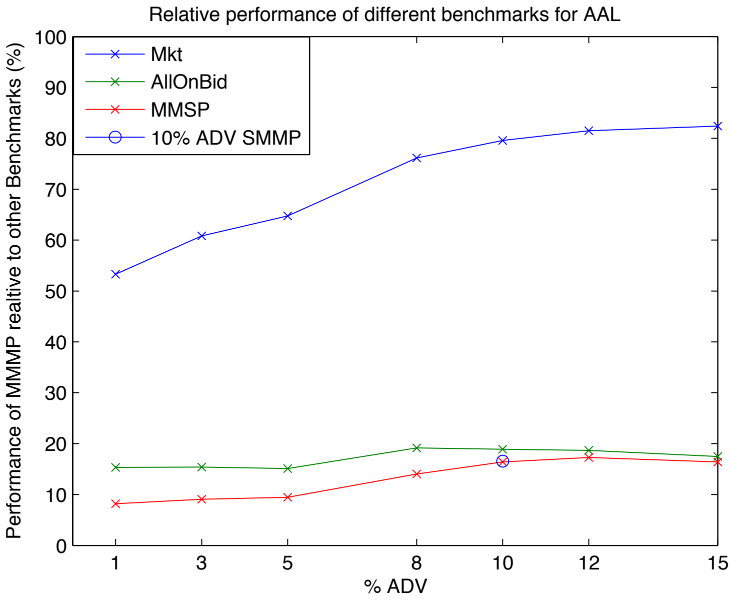


Figure 8.5: Performance comparison of trading AAL in two venues. Poredjenje za trgovanje AAL na dva tržišta.

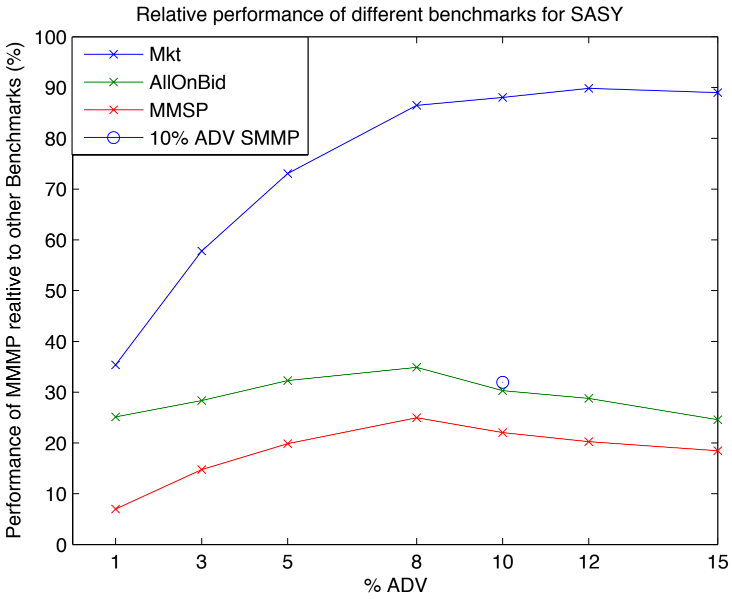


Figure 8.6: Performance comparison of trading SASY in two venues. Poredjenje za trgovanje SASY na dva tržišta.

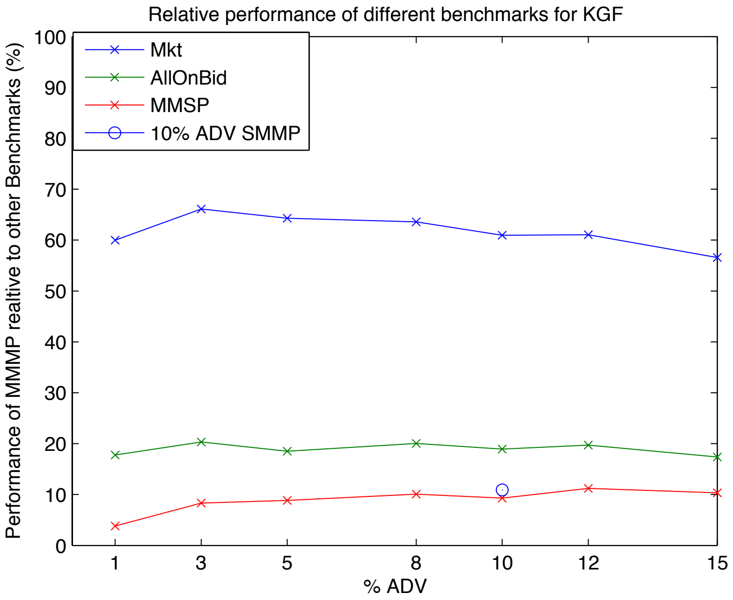


Figure 8.7: Performance comparison of trading KGF in two venues. Poredjenje za trgovanje KGF na dva tržišta.

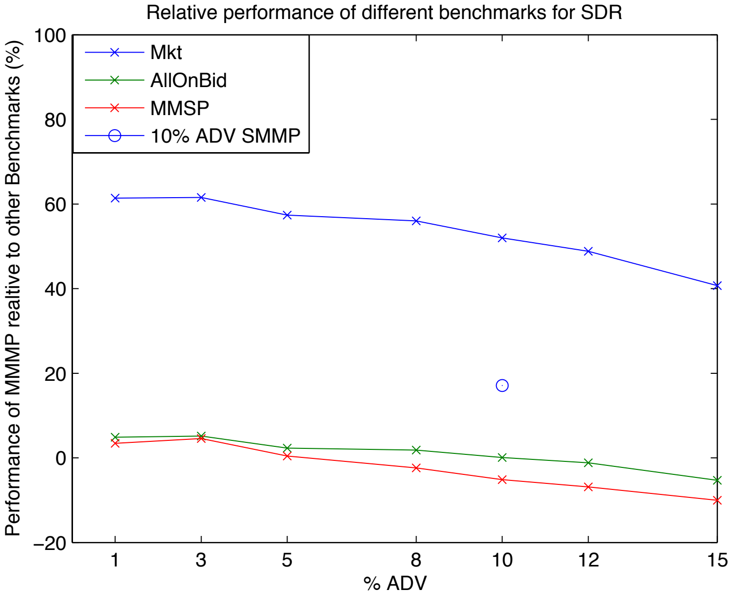


Figure 8.8: Performance comparison of trading SDR in two venues. Poredjenje za trgovanje SDR na dva tržišta.

The columns 2 to 4 are the relative performance of the three basic alternative benchmarks, namely, Market, All on Bid and single period optimal trajectory. The performance of these benchmarks are given as 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 the the finance industry does not have any valid benchmarks for measuring performance. Furthermore, as discussed earlier, benchmarks can be affected by the way in which once trades. Therefore, we have chosen two benchmarks that are common market practise. In a multi-market environment, the market place could be thought of as an aggregated single market, hence 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 single period model and two-period model is well addressed in earlier chapters. The objective of multi-market execution was to further improve the execution performance by tapping into the additional liquidity provided by the alternative venues to the primary market. We give more details taking the example of 10% 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, 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 cancelations 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 venue 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 say order size greater than 8% of ADV. The reasons for this as explained before, 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 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.

Chapter 9

Conclusions

The research project we undertook was to find an optimization procedure for algorithmic trading and to this end we conclude that our research has meet it's aims.

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, in both single and multiple market environments. 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. Finally, we consider how the model that allows re-optimization perform in a multiple trading venue environment. In all cases, a combination of market and limit orders are 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 three optimization models we are discussing in this thesis. Under certain conditions the objective functions of all 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.

We first approached the simplest of problems, namely single market and single period optimization. We were able prove that the problem at hand could be optimized with an SQP variant procedure. This was followed by formalizing the objective function. Next, we extended the model to deal with multiple re-optimization. In our example, we re-optimized once only, however the procedure is general can be re-optimized more frequently. Finally, we extended the general multi period problem to an environment consisting of multiple trading venues. This last mentioned problem required a bi-level optimization method for the estimation of residuals.

In all of this research, the most significant problem we ran into was that the fill probability model could not be calibrated well enough for less liquid stocks. This is an inherent short-coming of the Fill Probability model and not an obstacle in our research. Beside that, the calibration of market impact coefficients from real traded data was quite tedious and time consuming. In the early part of the research, we had minor problems with finding the correct scale for variables used in fmincon(). Further, we had numerous small problems with bad prints in the tick data used - a common problem in high frequency data. Finally, the most difficult part of the simulation work was to validate whether all variables used in the optimization and evaluation were correct. The simulation output consisted of 150+ columns of numbers for each window and it was a painful process to validate the correctness of it.

Overall, with the exception of very illiquid stocks for which the Fill Probability model did not perform too well, our results were in line with initial expectation. The models performed significantly better than the common market practice. Among models themselves, the multi-market multi-period worked best as expected than single-market multi-period. In similar way, among the single market model, the multi-period performed better than single period.

We embarked on this research due to lack for formalism in this discipline. To that end, we have achieved our objectives in formalizing a general optimization framework that enhances performance relative to market practice. Therefore, any bank or broker should be able to combine their fill probability model with the rest of the framework proposed herein. It is our sincere view that the algorithmic trading community will gain from this work. The only recommendations we can make to potential user is to build a solid back-testing framework. Without that, wrong fill assumptions would affect the results to the extend of making the results invalid.

Findings from the first part of this thesis, single market model, is already published. The multi-market results are presented in Kumaresan and Krejic [39]. Further, we are currently adding to our ongoing research ways in which we could quantify the quality of order placement.

Algorithmic trading is a complex problem and there are endless amounts of improvements one could make. One aspect in particular that we think is important and have not yet looked into in greater detail is that the window size should be dynamic. It is our belief that by dynamically estimating this variable one could improve the performance even more. Each stock will have a different execution window reflecting fundamental properties of the security.

In our work, we proposed a general framework for multiple re-optimizations. However, too frequent re-optimization can give poor results as limit orders won't get to mature. Therefore, it would be an interesting challenge to link the security level properties with how and when to perform re-optimization.

Due to lack of a formal framework of optimizing execution of algorithmic orders, we embarked on a challenging task for creating different models. We have successfully shown that the most sophisticated of the models, multi-period-multi-market model is indeed significantly better than common market practice. Because the whole model was created to be faithful to live trading in terms of assumptions used, deploying this model into production is straight forward.

Above all else, this has been a frightfully enjoyable project.

Bibliography

[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] Amihud, Y., Mendelson, H., Dealership markets: market making with inventory, Journal of Financial Economics, 8, 31-53, 1980.

[7] Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P., Exchange rate returns standardized by realised volatility are (nearly) Gaussian, NBER Working Paper No 7488, 2000.

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

[9] Bollerslev, T., Litvinova, J., Tauchen, G., Leverage and Volatility Feedback Effects in High-Frequency Data, Journal of Financial Econometrics, Vol. 4 No. 3 (2006), 353-384

[10] Bollerslev, Zhoub, H., Volatility puzzles: a simple framework for gauging return-volatility regressions, Journal of Econometrics 131 (2006), 123-150

[11] 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.

[12] Brownlees, C.T., Giampiero, M. Gallo, Financial Econometric Analysis at Ultra-High Frequency: Data Handling Concerns, 2006, Technical Report

[13] Chan, Louis K C, Lakonishok, Josef, The Behavior of Stock Prices around Institutional Trades, Journal of Finance, American Finance Association, vol. 50(4), pages 1147-74, September, 1995.

[14] Chanda, A., Engle, R.F., Sokalska, M.E., High Frequency Multiplicative Component GARCH, Technical Report, Departmnet of Finance, Stern School of Business, New York University, 2005.

[15] Chernov, M., Gallant, A.R., Ghysels, E., Tauchen, G., Alternative Models for Stock Price Dynamics, Journal of Econometrics 116, 225-258, 2003.

[16] Corsi, F., Zumbach, G.O., Muller, U.A., Dacorogna, M.M., consistent high- precision volatility from high frequency data, Economic Notes 2001.

[17] Dacorogna, M.M., Gencay, R., Muller, U.A., Olsen, R., An Introduction to High Frequency Finance, Academic Press, 2001.

[18] Demsetz, H., The cost of transacting, Quarterly Journal of Economics, 82: 33-53, 1968.

[19] Dufour, A., Engle, R.F., Time and the Price Impact of a Trade., The Journal of Finance 55 (6): 246798, 2000.

[20] 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.

[21] Engle, R.F., Russell, J.R., Analysis of High Frequency Financial Data, Technical Report 2004.

[22] Engle, R.F., Lange, J., Measuring, Forecasting and Explaining Time Varying Liquidity in the Stock Market, 1997.

[23] Engle, R.F., Gallo, G.M., A multiple indicators model for volatility using intra-day data, Journal of Econometrics, 131 (2006), 3-27.

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

[25] Falkenberry, T.,N., High frequency data filtering, Technical report, Tick Data.

[26] Garman, Mark B., Market microstructure, Journal of Financial Economics, 3(3), 257275, 1976.

[27] Gallant, A.R., Long, J.R., Estimating stochastic differential equations efficiently by minimum chi-squared, Biometrika 84 (1), 12-141, 1997.

[28] Ghysels, E., Santa-Clara P., Valkanov, R., Predicting volatility: getting the most out of return data sampled at different frequencies, Journal of Econometrics, 131 (2006), 59-95.

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

[30] Hasbrouck, Joel, Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading. New York: Oxford University Press, 2007.

[31] Ho, T., Stoll, H. R., Optimal dealer pricing under transactions and return uncertainty, Journal of Financial Economics, Elsevier, vol. 9(1), pages 47-73, March, 1981.

[32] Holthausen, R.W., Leftwich, R.W.,Mayers, D., Large-Block Transactions, the Speed of Response, and Temporary and Permanent Stock-Price Effects., Journal of Financial Economics, 26(1), pp. 71-95, 1990.

[33] Hong, Y., Chung, J., Are Directions of Stock Price Changes Predictable? Statistical Theory and Evidence, Technical report, Cornell University, 2003.

[34] http://www.nber.org/workinggroups/mm/mm.html

[35] Iori, G., Daniels, M.G., Farmer, J.D., Gillemot, L., Krishnamurty, S., Smith, E., An analysis of price impact function in order-driven markets, Physica A, 324 (2003), 146-151.

[36] Keim, D.B., Madhavan, A., The Upstairs Market for Large-Block Transactions: Analysis and Measurement of Price Effects, Review of Financial Studies, Oxford University Press for Society for Financial Studies, vol. 9(1), pages 1-36, 1996.

[37] Kissell, R., Glantz, M., Optimal Trading Strategies: Quantitative Approaches for Managing Market Impact and Trading Risk, AMACOM, 2003.

[38] Kumaresan, M., Krejić, N., A model for optimal execution of atomic orders, Computational Optimization and Applications, Volume 46, Issue 2 (2010), Page 369.

[39] Kumaresan, M., Krejić, N., Optimal Trading of Algorithmic Orders in a Liquidity Fragmented Market Place. (submitted).

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

[41] Madhavan, A., Market microstructure: A survey, Journal of Financial Markets, 2000.

[42] Madhavan, A., VWAP Strategies, Investment Guides, Transaction Performance: The Changing Face of Trading, 32-38, edited by R. Bruce, Institutional Investor Journal, Inc., New York, NY., Spring 2002.

[43] Murge, M.G., Pachpatte, B.G., Explosion and asymptotic Behaviour of Nonlinear Ito type Stochastic Integrodifferential equations, Kodai Math. J., 9 (1986), 1-18.

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

[45] Oberguggenberger, M. Russo, F., Nonlinear SPDEs: Colombeau solutions and pathwise limits. Stochastic analysis and related topics, VI (Geilo, 1996), 319–332, Progr. Probab., 42, Birkhuser Boston, Boston, MA, 1998.

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

[47] O'Hara, M., Market Microstructure Theory, Wiley, 1998.

[48] O'Hara, M., Oldfield, G., The microeconomics of market making, Journal of Financial and Quantitative Analysis, 21, 361-376, 1986.

[49] Plerou, V., Gopikrishnan, P., Gabaix, X., Stanley, H.E., Quantifying stock-price responce to demand fluctuations, Physical Review E 66 027104 (2002).

[50] Roll, R., A simple implicit measure of the effective bid-ask spread in an efficient market, Journal of Finance, 39, 1127-1139, 1984.

[51] Rydberg, T., H., Shephard, N., Dynamics of Trade-by-Trade Price Movements: Decomposition and Models, Journal of Financial Econometrics, Oxford University Press, vol. 1(1), pages 2-25, 2003.

[52] Stoll, Hans R, The Pricing of Security Dealer Services: An Empirical Study of NASDAQ Stocks, Journal of Finance, American Finance Association, vol. 33(4), pages 1153-72, September, 1978.

[53] Tauchen, G., Recent Development in Stochastic Volatility: Statistical Modelling and General Equilibrium Analysis, 2004.

[54] Tinic, S.M., The Economics of Liquidity Services, The Quarterly Journal of Economics, 86: 79-93, 1972.

[55] Torben G. Andersen, Tim Bollerslev, Peter F. Christoffersen, Francis X. Diebold, VOLATILITY FORECASTING, Working paper 11188.

[56] William J. Breen, Laurie Simon Hodrick, Robert A. Korajczyk, Predicting Equity Liquidity, Management Science, Vol. 48, No. 4, pp. 470-483, April 2002.

[57] Wright, J.H., Bollerslev, T., High frequency data, frequency domain inference and volatility forecasting, International Finance Discussion Papers, 64, 1999.

Biography

I was born in Jaffna, Tamil Eelam. My family emigrated to Norway, where I received my primary and secondary educations. I obtained a BSc in Computer Systems Engineering from the University of Warwick (England) and an MPhil in Computer Science from the University of Exeter (England).

Shortly after completing my national service with the Royal Norwegian Navy, I relocated to Australia.

I joined Deutsche Bank Australia in 1993 where I was a quantitative trader until 1999, when I returned to London to join Credit Suisse First Boston at their Statistical and Index arbitrage desk. I moved to New York in 2000, to join a start-up hedge fund as a partner.

In 2006 I joined Dresdner Kleinwort as its Global Head of Algorithmic Trading and moved to TMG (TransMarket Group) in early 2008 as a Managing Director and Global Head of Quantitative Trading.

In 2007, I commenced my research leading to a PhD in Mathematics at the University of Novi Sad (Serbia).

In 2010, I founded Algonetix LLP, a new high frequency trading entity.

London, June 17, 2010. Miles Kumaresan

UNIVERZITET U NOVOM SADU
PRIRODNO - MATEMATIČKI FAKULTET
KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj:
Identifikacioni broj:
IBR
Tip dokumentacije: Monografska dokumentacija
TD
Tip zapisa: Tekstualni štampani materijal
TZ
Vrsta rada: doktorska disertacija
VR
Autor: Miles Kumaresan
AU
Mentor: Prof. dr Nataša Krejić
MN
Naslov rada: Optimizacija uslovnih trajektorija trgovanja na više tržišta sa različitim likvidnošću
MR
Jezik publikacije:Engleski (latinica)
JP
Jezik izvoda: s / e
JI
Zemlja publikovanja: Republika Srbija
ZP
Uže geografsko područje: Vojvodina
UGP
Godina: 2010
GO
Izdava v c: Autorski reprint
IZ
Mesto i adresa: Novi Sad, Departman za matematiku i informatiku, PMF, Trg Dositeja Obradovića 4
MA
Fizički opis rada: (9,130,57,11, ,17 ,)
(Broj poglavlja, strana, citata, tabela, slika, grafika, priloga)
FO

Naučna oblast: Matematika
NO
Naučna disciplina: Numerička matematika
ND
Ključne reči: nelinearno programiranje, konveksno programiranje, optimalna strategija izvršenja, više tržišta, algoritamsko trgovanje
PO
UDK:
Čuva se: u biblioteci Departmana za matematiku i informatiku, Novi Sad
ČU
Važna napomena:
VN

Izvod: Algoritamsko trgovanje je automatizovani proces izvršavanja naloga na elektronskim berzama (berzama akcija). Osnovni cilj u izvršenju je postizanje najefikasnije cene. Ovde su predložene dve optimalne strategije za izvršenje atomskih naloga zasnovane na minimizaciji troškova impakta i volatilnosti u slučaju jednog tržišta i više tržišta. Prva posmatrana strategija je zasnovana na relativno jednostavnom nelinearnom optimizacionom modelu, dok druga dozvoljava reoptimizaciju u nekom trenutku unutar zadatog vremenskog intervala izvršenja. Konačno, posmatran je model koji dozvoljava reoptimizaciju u okruženju sa više tržišta. U svim slučajevima koristi se kombinacija market i limit naloga. Glavna inovacija u našem pristupu je uvodjenje Fill Probability funkcije koja omogućava kombinaciju market i limit naloga u sva četiri modela diskutovana u ovoj tezi. Pod odredjenim uslovima funkcije cilja svih posmatranih problema su konveksne te se mogu primeniti standardni metodi optimizacije. Efikasnost predloženih strategija je testirana u odnosu na dve representativne strategije, koje predstavljaju uobičajenu praksu, na realnom uzorku podataka sa tržišta.
IZ

Datum prihvatanja teme od strane NN veća: 20.5.2008.
DP
Datum odbrane:
DO
Članovi komisije:
Predsednik:
Član:
Član:
KO

UNIVERSITY OF NOVI SAD
FACULTY OF NATURAL SCIENCES AND MATHEMATICS
KEY WORDS DOCUMENTATION

Accession number:
ANO
Identification number:
INO
Document type: Monograph type DT Type of record: Printed text
TR
Contents Code:
CC
Author: Miles Kumaresan
AU
Mentor: Prof. Dr. Nataša Krejić
MN
Title: Optimization of Conditional Trajectories in a Market Place of Multiple Liquidity
XI
Language of text: English
LT
Language of abstract: Serbian. English
LA
Country of publication: Republic of Serbia
CP
Locality of publication: Vojvodina
LP
Publication year: 2010
PY
Publisher: Author's reprint
PU
Publ. place: Novi Sad, Faculty of Natural Sciences and Mathematics, Trg Dositeja Obradoviča 4 PP
Physical description: (9,130,57,11, ,17 ,)
(Chapters, pages, literature, tables, pictures, graphics, appendices)
PD
Scientific field: Mathematics SF
Scientific discipline: Numerical Mathematics

SD
Key words: nonlinear programming, convex programming, optimal execution strategy, multiple trading venues, algorithmic trading
KW
Holding data: Library of the Department of Mathematics and Informatics, Novi Sad HD
Note:
N Abstract: Algorithmic Trading is the automated process of trading exogenous orders in electronic (stock) exchanges.The primary objective in execution of orders is to achieve the most efficient price. We propose two optimal strategies for the execution of atomic orders based on minimization of impact and volatility costs, in both single and multiple market environments. 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. Finally, we consider how the model that allows re-optimization perform in a multiple trading venue environment. In all cases, a combination of market and limit orders are 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 four optimization models we are discussing in this thesis. Under certain conditions the objective functions of all 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.
AB

Accepted by the Scientific Board on: May 20, 2008.
Defended:
DE Thesis defend board:
President:
Member:
Member:
Member:
DB