A Multi-class Queueing Model of Limit Order Book Dynamics

Modern equity markets are computerized technological systems, operating as so-called “electronic limit order books” (LOBs). Market participants, including institutional investors, market makers, and opportunistic investors, are faced with a new set of operational trading challenges as they interact within today’s high-frequency marketplace. This has necessitated the development of electronic trade execution algorithms. At a high level, these algorithms dynamically optimize where, how often, and at what price to trade. They seek to optimize their own best execution objectives while taking into account the state of the exchanges, and real-time market information. Growing evidence for the importance of execution algorithms is evident through formation of specialized groups at investment banks and other organizations that offer algorithmic trading services.

This paper describes a multi-class queueing model of a limit order book, drawing a strong connection between the field of high-frequency market microstructure and the ubiquitous service of trade execution to the tools from queueing networks and stochastic control. We subsequently formulate and solve a stylized, yet relevant, short-term optimal execution problem: how to buy or sell a block of shares at the best possible price, taking into account the short term queueing dynamics of the limit order book. Finally, we propose a microstructure model of market impact that captures the expected adverse price movement due to the trader’s own activity, which is an essential ingredient of most trade execution algorithms used in practice. We illustrate our model in an empirically calibrated example.

**Model.** Exchanges typically function as electronic limit order books, operating under a “price-time” priority rule. They can be modeled as a multi-class queueing system, as follows:

*Order arrivals.* There are two types of orders that arrive to trade in a LOB. (i) **Limit orders** that seek to buy a quantity of the stock at a price that is less than or equal to an upper price “limit”, where both the quantity and price are defined by the trader; similarly, limit sell orders seek to sell a quantity of the stock at a price that is higher or equal to a lower price limit. The **bid** is defined to be the highest price at which buyer limit orders are posted and waiting to trade in the LOB, and the **ask** is the lowest limit prices at which limit orders to sell are posted in the LOB. (ii) **Market orders** that seek to buy a quantity at the “best” possible price. Since market orders do not impose a constraint on the execution price, they execute instantaneously. In contrast, limit orders may not be able to execute upon their arrival and instead join the queue associated with their price limit.

The quantities, price limits, and order inter-arrival times of such orders are stochastic, which we
model as non-homogeneous compound Poisson processes of appropriate state-dependent rates and order size distributions. Order sizes and order arrival intensities vary significantly as a function of the distance of the queue from the best bid and best ask, the time of day, and the queue length, but for simplicity we will assume that these dependencies are captured in the order arrival rates.

Order cancellations. Limit orders may be canceled at any point in time due to a variety of reasons, such as a change in the state of the limit order book or constraints imposed by the original intent of the trader, e.g., to follow a particular market benchmark. For purposes of tractability we will assume that each limit order submitted has a “patience” time associated with it, which could be understood as a maximum delay it is willing to incur before executing. Patience times are assumed to be i.i.d. draws from an exponential distribution with rates that may depend on the price level and time-of-day. Order cancellations are extremely prevalent in practice, where 75%–90% of the limit orders posted in the market to trade, get canceled before they execute.

Limit order book dynamics. The queueing system described is driven by the arrival of limit orders, which join the queue associated with a particular price. The arrival of market orders to buy act as “services” that processes the limit orders to sell according to a strict priority rule defined by the prices that correspond to the queues. Orders within a queue (at a given price level) are processed in a FIFO fashion. Orders abandon the respective queues according to exponentially distributed patience clocks. Finally, the evolution of the buy and sell side of the book are coupled through the state dependence of order arrival rates. Viewed as a stochastic network, a limit order book evolves as two coupled multi-class priority queues with abandonments.

Information. Exchanges publish real-time information for each security that allow investors to know or compute the quantities available for trade at each price level (i.e., the queue lengths), and estimate rates at which limit and market orders arrive into the system, and cancel from the system. That is, the state of the exchange and the model primitives can be measured and be known to the trader. This queueing system offers a tractable mathematical model for analysis of the high-frequency market microstructure of today’s limit order books.

Problem formulation. We focus on one side of the book, formulate an optimal execution problem, and subsequently explore how its findings provide the basis of a market microstructure model of market impact. The specific problem can be described as follows: what is the optimal way to buy \( C \) shares of a security using limit orders at the lowest possible price over a given time horizon \( T \). Here, it is natural to think of the time horizon to be of the order of a few minutes. This order placement decision is a type of routing problem in the queueing model of the LOB, where the trader has to decide to which queues and when to submit orders so as to optimize the above objective. In a stochastic environment it is possible that the target quantity may not execute by \( T \), but whatever residual is left over can be completed by a “clean-up” trade via a market order.

This transient optimal control problem motivates a few simplifying assumptions. First, the horizon of that problem is short when compared to the time scale over which historically estimated parameters for the model may change, and as such one can suppress the time dependence of the
order arrival intensities. Second, instead of formulating the above control problem in the context of the queueing model, we will study its fluid model analog, where stochastic and discrete arrival and cancellation processes have been replaced by continuous and deterministic flows. This is a common tool in the analysis of queueing systems that can be justified through an asymptotic analysis in settings where the rate of arrivals of limit and market orders grows proportionally large, and is particularly suitable for transient optimization problems. The asymptotic regime that justifies such a fluid model suggests that this style of analysis is relevant for fairly liquid securities, e.g., such as the securities that comprise the S&P500 index.

**Results.** Our main results are as follows:

**Characterization of the optimal execution strategy.** First, we show that the optimal strategy is to place all limit orders at the beginning of the execution path, giving the orders more time to queue up, gain priority by moving to the head of their respective queues, and therefore managing to execute in time at good prices. Next, we adopt an equivalent formulation of the optimal execution problem that strives to minimize the time it would take to trade \( C \) shares subject to a constraint that the price at which they would be acquired is at least as good as some upper bound. The latter problem is one of minimizing a concave function over a polyhedron, that admits a extreme point solution with at most two non-zero entries. In other words, it is optimal to post the target quantity in at most two queues, i.e., place orders in at most two price levels. Moreover, we show that these price levels need to be adjacent. Using the latter observations we construct explicitly the optimal trading strategy and specify how much quantity to place and at what price levels as a function of the state of the system at the time the trader arrives to the market and its real-time conditions. The above analysis allows us to study two practical extensions that impose constraints on the rate of trading and allows for the use of market orders, especially when the quantity to trade over the specified horizon is too large.

**Efficient frontier.** The above characterization allows us to see that there is a trade off between the trading rate and the cost of trading: trading faster necessitates trading at disadvantageous prices. We crisply characterize the efficient frontier associated with the optimal execution problem that relates the trading rate to the best achievable price, under the optimal strategy.

**Structural microstructure model of market impact.** Apart from providing guidance towards the design of practical execution algorithms over short time durations, the above analysis characterizes the execution cost as a function of the quantity and time duration and other LOB variables and market microstructure primitives. The latter are observable in real-time. This cost function offers an intuitive structure of a model that one could try to calibrate against market data and trading cost estimates. This is a completely alternative way of modeling the market impact cost that increases the accuracy of its predictions five fold over the prevailing models used in practice for trades with short time durations. We illustrate our model by calibrating an example to empirical data.