



## **Transaction Cost Analytics in DROP**

**v7.50** 27 December 2025



## Exchange Order

### Overview

1. Definition of an Exchange Order: An *order* is an instruction to buy or sell on a trading venue such as a stock market, bond market, commodity market, financial derivative market, or crypto-currency exchange (Wikipedia (2023)).
2. Instructions for Executing an Order: These instructions can be simple or complicated, and can be sent to either a broker or directly to a trading venue in direct market access. There are some standard instructions for such orders.

### Market Order

1. Definition of a Market Order: A *market order* is a buy or sell order to be executed immediately at the *current market* prices. As long as there are willing sellers and buyers, market orders are filled.
2. Motivation behind Using Market Orders: Market orders are used when certainty of execution is a priority over the price of execution.
3. Lack of Control over Price: A market order is the simplest of the order types. This order type does not allow any control over the price received.
4. Filling up of Market Order: The order is filled at the best price available at the relevant time.
5. Ability to execute at Quoted Price: In fast-moving markets, the price paid or received may be quite different from the last price quoted before the order was entered.



6. Slicing up the Market Order: A market order may be split across multiple participants on the other side of the transaction, resulting in different prices for some of the shares.
7. Commissions for the Market Order: It is the most basic of all order and therefore, they incur the lowest of commissions, from both online and traditional brokers.

## Limit Order

1. Definition of the Limit Order: A *limit order* is an order to buy a security at no more than a specific price, or to sell a security at no less than a specific price – called “better” for either direction.
2. Uncertainty over the Order Execution: This gives the trader/customer control over the price at which the trade is executed; however, the order may never be executed/filled.
3. Trader Control over the Price: Limit orders are used when the trader wishes to control price rather than certainty of execution.
4. Definition of Buy Limit Order: A *buy limit order* can only be executed at the limit price or lower. For example, if an investor wants to buy a stock, but doesn’t want to pay more than \$30 for it, the investor can place a limit order to buy the stock at \$30.
5. Purpose of the Buy Limit Order: By entering the limit order rather than a market order, the investor will not buy the stock at a higher price, but, may get fewer shares than he wants or not get the stock at all.
6. Definition of Sell Limit Order: A *sell limit order* is analogous; it can only be executed at the limit price or higher.
7. Marketability of the Limit Order: A limit order that can be satisfied by orders in the limit book when it is received is *marketable*.
8. Filling up of Buy Order: For example, if a stock is asked for \$86.41 – a large size, a buy order with a limit of \$90 can be filled right away.
9. Filling up of Sell Order: Similarly, if a stock is bid \$86.40, a sell order with a limit of \$80 will be filled right away.



10. Partial Fill of Limit Order: A limit order may be partially filled from the book and the rest added to the book.
11. Constrained Buy and Sell Orders: Both buy and sell orders can be additionally constrained. Two of the most common additional constraints are fill or kill FOK, and all or none AON.
12. Definition of FOK and AON Orders: FOK orders are either filled completely on the first attempt or canceled outright, while AON orders stipulate that the order must be filled with the entire number of shares specified, or not filled at all.
13. Scheduling the Unfilled Orders: If it is not filled, it is still held on the order book for later execution.

## Time in Force

1. Good for Day Order GFD: A *day-order* or *good-for-day-order* GFD – the most common – is a market or a limit order that is in force from the time the order is submitted to the end of the day's trading session.
2. Closing Time for Stock Markets: For stock markets, the closing time is defined by the exchange.
3. Closing Time for FX Markets: For the foreign exchange market, this is until 5 PM EST/EDT for all currencies except the New Zealand Dollar.
4. Good-till-Canceled GTC Orders: GTC require a specific canceling order, which can persist indefinitely – although brokers may set some limits, for example, 90 days.
5. Immediate or Cancel IOC Orders: IOC orders are immediately executed or canceled by the exchange. Unlike FOK orders, IOC orders allow for partial fills.
6. Fill or Kill FOK Orders: FOK orders are usually limit orders that must be executed or canceled immediately. Unlike IOC orders, FOK orders require full quantity to be executed.



7. Open/Close Single-Price Auctions: Most markets have single-price auctions at the beginning – or “open” – and the end – or “close” – of regular trading. Some markets may also have before-lunch and after-lunch orders.
8. On the Close/Open: An order may be specified *on the close* or *on the open*, then it is entered in an auction but has no effect otherwise. There is often some deadline, for example, orders must be in 20 minutes before the auction.
9. Definition of Single-Price Orders: They are single-price because all orders, if they transact at all, transact at the same price, the open price and the close price, respectively.
10. MOC/MOO/LOC/LOO Orders: Combined with price instructions, this gives *market-on-close* MOC, *market-on-open* MOO, *limit-on-close* LOC, and *limit-on-open* LOO.
11. Example - Market-on-Open: For example, a market-on-open order is guaranteed to get the open price, whatever that may be.
12. Buy-Limit or Buy-Open Order: A buy limit-on-open order is filled if the open price is lower, not filled if the open price is higher, and may or may not be filled if the open price is the same.
13. Regulation NMS Compliant Order Routing: Regulation NMS – Reg NMS – which applies to US stock exchanges, supports two types of IOC orders, one of which is Reg NMS compliant and will not be routed during an exchange sweep, and one that can be routed to other exchanges.
14. Liquidity Impact on Order Routing: Optimal order routing is a difficult problem that cannot be addressed with the usual perfect market paradigm. Liquidity needs to be modeled in a realistic way (Polimenis (2005a)) if one is to understand such issues as optimal order routing and placement (Polimenis (2005b)).
15. Adopting the Order Protection Rule: The Order Protection or Trade Through Rule – Rule 611 – was designed to improve intermarket price priority for quotations that are immediately and automatically accessible, but its role in predatory trading behavior has faced mounted controversy in the recent years.



## Conditional Orders

A conditional order is any order other than a limit order which is executed only when a specific condition is met.

## Stop Orders

1. Definition of Stop-loss Orders: A *stop-order* or *stop-loss order* is an order to buy or sell a stock once the price of the stock reaches a specified price, known as the *stop-price*. When the stop price is reached, a stop order becomes a market order.
2. Definition of Buy-stop Order: A buy-stop order is entered at a stop price above the current market price. Investors generally use a buy-stop order to limit a loss or to protect a profit on a stock that they have sold short.
3. Definition of Sell-stop Order: A sell-stop order is entered at a stop price below the current market price. Investors generally use a sell-stop order to limit a loss or to protect a profit on a stock that they own.
4. Impact of Using Stop-Orders: When a sell-stop order is reached, the stop order becomes a market order. This means that the trade will definitely be executed, but not necessarily at or near the stop price, particularly when the order is placed into a fast-moving market, or if there is insufficient liquidity available relative to the size of the market.
5. Markets where Stop-orders are Used: The use of stop orders is much more frequent for stocks and futures that trade on an exchange than those that trade in the over-the-counter OTC market.

## Stop-Sell Order



1. Purpose of the Sell-stop Order: A *sell-stop order* is an instruction to sell at the best available price after the price goes below the stop price. A sell-stop price is always below the current market price.
2. Sell-stop Order in Action: For example, if an investor holds a stock currently valued at \$50 and is worried that the value may drop, he/she can place a sell-stop order at \$40. If the share price drops to \$40, the broker sells the stock at the next available price.
3. Advantages of Sell-stop Order: This can limit the investor's losses or lock in some of the investor's profits – if the stop price is at or above the purchase price.

## Buy-stop Order

1. Purpose of Buy-stop Order: A *buy-stop order* is typically used to limit a loss – or to protect an existing profit – on a short sale. A buy-stop price is always above the current market price.
2. Buy-stop Order in Action: For example, if an investor sells a stock short – hoping for the stock price to go down so they can return the borrowed shares at a lower price, i.e., *covering* – the investor may use a buy-stop order to protect against losses if the prices go too high.
3. Advantages of Buy-stop Order: It can also be used to advantage in a declining market when an investor decides to enter a long position at when he perceives to be prices close to the bottom after a market sell-off.

## Stop-limit Order

1. Definition of Stop-Limit Order: A *stop-limit order* is an order to buy or sell a stock that combines the features of a stop order and a limit order.



2. Stop Order to a Limit Order: Once the stop price is reached, a stop-limit order becomes a limit order that will be executed at a specified price – or better.
3. Lack of Execution Certainty: As with all limit orders, a stop-limit order doesn't get filled if the security's price never reaches the specified limit price.

## Conditional Trailing Stop Order

1. Definition of Trailing Stop Order: A *trailing stop order* is entered with a stop parameter that creates a moving or *trailing* activation price, hence the name.
2. Specification of Trailing Activation Price: This parameter is entered as a percentage change or actual specific amount of rise – or fall – in the security price.
3. Using Trailing Stop-sell Orders: Trailing stop-sell orders are used to maximize and protect profit as a stock's price rises and limit losses when its price falls.
4. Trailing Stop-sell Example #1: As an example, a trader has bought stock ABC at \$10.00 and immediately placing a trailing stop sell order to sell ABC with a \$1.00 trailing stop, i.e., 10% of its current price. This sets the stop price to \$9.00.
5. Trailing Stop-sell Example #2: After placing the order, ABC does not exceed \$10.00 and falls to a low of \$9.01. The trailing stop order is not executed because ABC has not fallen \$1.00 from \$10.00.
6. Trailing Stop-sell Example #3: Later, the stock rises to a high of \$15.00 which resets the stop price to \$13.50. It then falls to \$13.50 - \$1.50 or 10% from its high of \$15.00 – and the trailing stop sell order is entered as a market order.

## Trailing Stop-limit Order

A *trailing stop-limit order* is similar to a trailing stop order. Instead of selling at market price when triggered, the order becomes a limit order.



## Peg Orders

To behave like a market maker, it is possible to use what are called peg orders.

### Peg Best Orders

1. Market Making Stepper Step #1: Like a real market maker, first, the stepper uses the other side of the spread.
2. Market Making Stepper Step #2: Second, the stepper always jumps over the competitors order to be the best one, the first one in the line.
3. Conditions that Limit Pegging:
  - a. Price limitation, no more jumping over, unless the price moves back to its area
  - b. Step value

### Mid-Price Peg Order

1. Definition of Mid-Price Order: A mid-price is an order whose limit price is continually set at the average of the “best bid” and “best offer” prices in the market.
2. Definition of Peg-to-Midpoint: The values of the bid and the offer prices used in the calculation may be either a local or a national best bid and offer. They are also called peg-to-midpoint.
3. Mid-price Peg Order Type: Mid-price peg order types are commonly supported on alternative trading systems and dark pools, where they enable market participants to trade whereby each pays half of the bid-offer spread, often without revealing their trading intentions to others beforehand.



## Market-if-Touched Order

1. But Market-if-Touched Order: A buy *market-if-touched order* is an order to buy at the best available price, if the market price goes down to the “if touched” level. As soon as this trigger price is touched the order becomes a market buy order.
2. Sell Market-if-touched Order: A sell *market-if-touched order* is an order to sell at the best available price, if the market price goes up to the “if touched” level. As soon as this trigger price is touched the order becomes a market sell order.

## One Cancels Other Orders

1. Definition of the OCO Order: One-cancels-other OCO orders are used when the trader wishes to capitalize on only one of two or more trading possibilities. For instance, the trader may wish to trade stock ABC at \$10.00 or XYZ at \$20.00.
2. A Sample OCO Order: In this case, he would execute an OCO order composed of two parts: A limit order for ABC at \$10.00 and a limit order for XYZ at \$20.00.
3. Triggering of an OCO Order: If ABC reaches \$10.00, ABC’s limit order would be executed, and XYZ limit order would be canceled.

## One Sends Other Orders

1. Definition of OSO Order: One sends other – OSO – orders are used when the trader wishes to send a new order only when another one has been executed.
2. Example of an OSO Order: For instance, the trader may wish to buy stock ABC at \$10.00 and then immediately try to sell it at \$10.05 to gain the spread.
3. Components of an OSO Order: In this case, they would execute an OSO order composed of two parts: A limit buy order for ABC at \$10.00, and a limit sell order for



the same stock at \$10.05. If ABC reaches \$10.00, ABC's limit order would be executed, and the sell limit order would be sent.

4. Sequential Execution of OSO Trades: In short, multiple orders are attached to a main order and the orders are executed sequentially.

## Tick-sensitive Orders

1. Uptick and Downtick Orders: An uptick is when the last non-zero price change is positive, and a downtick is when the last non-zero price change is negative.
2. Buy-on-downtick: Any tick-sensitive instruction can be entered at the trader's option, for example *buy-on downtick*, although these orders are rare.
3. Tick Sensitive Short-sell Order: In markets where short sales may only be executed on an uptick, a short-sell order is inherently tick-sensitive.

## Opening Orders

1. At-the-opening Order: *At-the-opening* order is an order type to be executed at the very opening of the stock market trading day.
2. Cancel on Failure to Execute: If it wouldn't be possible to execute it as part of the first trade for the day, it would instead be canceled (Weiss (2006)).

## Discretionary Order

1. Definition of a Discretionary Order: A *discretionary order* is an order that allows the broker to delay the execution at its discretion to try to get a better price; these are sometimes called not-held orders.



2. Use in Electronic/Voice Trades: It is commonly added to stop-loss orders and limit orders. They can be placed via a broker or an electronic trading system.

## Bracket

1. Pair of Two Orders: Puts to the market a pair of two orders, for the same title and direction, e.g., both to sell, as below:
2. Sell Order #1: One sell order is to realize the profit.
3. Sell Order #2: The second is to lock the loss, not to get even deeper.

## Quantity and Display Instructions

1. Prevention of the Order Display: A broker may be instructed not to display the order to the market. Examples are provided below.
2. Case of an AON Order: An “all-or-none” buy limit is an order to buy at the specified price if another trader is offering to sell the full amount of the order, but otherwise not display the order.
3. Case of Iceberg/Hidden Order: A hidden – or “iceberg” – order requires the broker to display only a small part of the order, leaving a large undisplayed quantity “below the surface”.

## Electronic Markets

1. Availability of the Order Type: All of the above order types are usually available in modern electronic markets, but order priority rules encourage simple market and limit orders.



2. Priority of Market and Limit Orders: Market orders receive the highest priority, followed by limit orders. If a limit order has priority, it is the next trade executed at the limit price.
3. Order Priority Determined by Time/Complexity: Simple limit orders generally get the highest priority, based on a first-come-first-served rule.
4. Priority of Conditional Orders: Conditional orders generally get priority based on the instant the condition is met.
5. Priority of Iceberg/Dark Pool Orders: Iceberg and dark pool orders – which are not displayed – are given lower priority.

## References

- Polimenis, V. (2005a): A Realistic Model of Market Liquidity and Depth *Journal of Futures Markets* **25 (5)** 443-464
- Polimenis, V. (2005b): Slow and Fast Markets *Journal of Economics and Business* **57 (6)** 576-593
- Weiss, D. (2006): *After the Trade is Made: Processing Securities Transactions* **Portfolio Publishing** London UK
- Wikipedia (2023): [Order \(Exchange\)](#)



## Time-in-Force: Definition, Types, and Examples

### Abstract

1. Motivation behind Time-in-Force: Time-in-force indicates how long an order will remain active before it expires with the broker (Chen (2021)).
2. Time-in-Force Option Settings: Time-in-force settings option is accomplished through different order types.
3. Time-in-force Order Examples: Common examples of time-in-force specifications include day order, immediate-or-cancel IOC, fill-or-kill FOK, or good-till-canceled GTC.

### What is Time-in-Force?

Time-in-force TIF is a special instruction used when placing a trade to indicate how long an order will remain active before it is executed or expires. These options are especially important for active traders and allow them to be more specific about the time parameters.

### Basics of Time-in-Force

1. Focus of the TIF Concept: TIF orders are a useful way for active traders to keep from accidentally executing trades. By setting time parameters, traders don't have to remember to cancel old trades.



2. Impact of Unintended Trade Executions: Unintended trade executions can be very costly, if they occur during volatile market conditions when prices are rapidly changing.
3. Price Control Using Limit Orders: Most active traders use limit order to limit the price they pay for a stock, which means that they set a TIF option to control how long the order stays open.
4. Most Common TIF Order Type: While day orders are the most common type of order, there are many circumstances when it makes sense to use other order types.
5. TIF Options Offered by Brokers: Some brokers only offer a limited set of order types, but active traders are often given more options.
6. TIF Type Acronyms in Use: Many brokers use acronyms like DAY, GTC, OPC, IOC, GTD, DTC, etc. to refer to these orders. The next section looks a little more closely at these order types.

## Types of TIF Orders

1. DAY Type TIF Order: DAY orders are a popular type of TIF order. They are canceled if the trade does not execute by the close of the trading day. These are often the default order types for brokerage accounts.
2. Good-Till-Canceled GTC Orders: Another type of TIF order is the Good-Till-Canceled GTC order, which is effective until the trade is executed or canceled.
3. Exceptions when GTC Does not apply: Some common exceptions include stock splits, distributions, account inactivity, modified orders, and quarterly sweeps.
4. Wait Duration inside the GTC: GTC can be a useful option for a long-term investor who is willing to wait for a stock to reach their desired price point before pulling the trigger. Sometimes, traders might want to wait for several days or even weeks for a trade to execute at their desired price.
5. Fill-or-Kill FOK Orders: Fill-or-kill – FOK – orders are a third type of order. They are canceled if the entire order does not execute as soon as it is available.



6. Usage of the FOK Orders: Often, FOK orders are used to avoid purchasing shares in multiple blocks at different prices and to ensure that an entire order executes at a single price.
7. Effectives of the FOK Orders: FOK orders can be popular during fast-moving markets when day traders want to ensure that they get a good price on their trade.
8. Remaining TIF Order Types: A few other order types include Market-On-Open MOO and Limit-On-Open LOO orders, which execute as soon as a market opens; Immediate-Or-Cancel IOC orders which must be filled immediately or are canceled; Day-Till-Canceled DTC orders that are deactivated at the end of the day instead of canceled, making it easier to re-transmit the order.

## Example of Time-in-Force

1. John's ABC Alpha Projection View: John believes that the price of stock ABC, currently trading at \$10, will rise, but it will take time, approximately three months.
2. \$15 Call Plus GTC Order: John purchases ABC call option with a strike of \$15 and places a GTC order. To avoid having the order remain on hold indefinitely, he places a limit of three months on the order.
3. Response to Realized ABC Price: After 3 months, ABC's price is still struggling to break past the \$12 mark. John's order is canceled immediately.

## References:

- Chen, J. (2021): [Time in Force: Definition, Types, and Examples](#)



## Order State Change Matrices

### Introduction

1. Sequence of Messages/Order Status: The following matrices are included to clarify the sequence of messages and the status of orders involved in the submission and the processing of new orders, executions, cancel requests, cancel/replace requests, and order processing requests. The matrices have been arranged in groups as follows.
2. Order Processing Scenarios and Groups:

Ref	Group	Description
D1	Vanilla	Filled order
D2	Vanilla	Part-filled day order, done for day
D3	Cancel	Cancel request issued for a zero-filled order
D4	Cancel	Cancel request issued for a part-filled order – executions occur whilst cancel request is active
D5	Cancel	Cancel request issued for an order that becomes filled before cancel request can be accepted
D6	Replace to increase quantity	Zero-filled order, cancel/replace request issued to increase order quantity
D7	Replace to increase quantity	Part-filled order, followed by cancel/replace request to increase order quantity, execution occurs whilst order is pending replace
D8	Replace to increase quantity	Filled-order followed by cancel/replace request to increase order quantity
D9	Replace not for qty change	Cancel/replace request (not for quantity change) is rejected as a fill has occurred



D10	Replace to decrease quantity	Cancel/replace request sent whilst execution is being reported – the requested order quantity exceeds the cumulative quantity. Order is replaced then filled
D11	Replace to decrease quantity	Cancel/replace request sent whilst execution is being reported – the requested order quantity equals the cumulative quantity – order quantity is amended to cumulative quantity
D12	Replace to decrease quantity	Cancel/replace request sent whilst execution is being reported – the requested order quantity is below cumulative quantity – order quantity is amended to cumulative quantity
D13	Replace - sequence	One cancel/replace request is issued which is accepted – another one is issued which is also accepted
D14	Replace - sequence	One cancel/replace request is issued which is rejected before order becomes pending replace – then another one is issued which is accepted
D15	Replace - sequence	One cancel/replace request is issued which is rejected after it is in pending replace – then another one is issued which is accepted
D16	Replace - chaining	One cancel/replace request is issued followed immediately by another – broker processes sequentially
D17	Replace - chaining	One cancel/replace request is issued followed immediately by another – broker rejects the second as order is pending replace
D18	Unsolicited reports	Telephoned order
D19	Unsolicited reports	Unsolicited cancellation of a part-filled order
D20	Unsolicited reports	Unsolicited replacement of a part-filled order
D21	Unsolicited reports	Unsolicited reduction of order quantity by sell side
D22	Order reject	Order rejected due to duplicate ClOrdID



D23	Order reject	Order rejected because the order has already been verbally submitted
D24	Status	Order status request rejected for unknown order
D25	Status	Status request followed by "Nothing done".
D26	Status	Order sent, immediately followed by a status request. Subsequent status requests sent
D27	GT	GTC order partially filled, restated (renewed) and partially filled the following day
D28	GT	GTC order with partial fill, a 2:1 stock split then a partial fill and fill the following day
D29	GT	GTC order partially filled, restated (renewed) and canceled the following day
D30	GT	GTC order partially filled, restated (renewed) followed by replace request to increase quantity
D31	Resend	Possible resend
D32	TIF	Fill or kill order that cannot be filled
D33	TIF	Immediate or Cancel order that cannot be immediately hit
D34	Execution correct/cancel	Filled order, followed by correction and cancellation of executions
D35	Execution correct/cancel	A cancel of a partially filled order followed by an execution cancel (bust) and new execution.
D36	Execution correct/cancel	GTC order partially filled, restated (renewed) and partially filled the following day, with corrections of quantity on both executions.
D37	Stopped/Guarantee	A stopped (execution price guarantee) report followed by execution.

3. Order State Transition – Matrix Representation: The grid below shows which state transitions have been illustrated, and they are marked with an asterisk. The row represents the current status of the OrdStatus and the column represents the next



value as reported back to the buy-side via an execution report or order cancel reject message.

4. Precedence of the Order Status Value: Next to each OrdStatus value is its precedence – this should be used when an order exists in a number states simultaneously to determine the value that should be reported back.
5. Absent Scenario in the Grid: Note that the absence of a scenario should not necessarily be interpreted as meaning that the state transition is not allowed.
6. Order State Transition Matrix:

<i>OrdStatus (precedence value)</i>	New (2)	Partially Filled (4)	Filled (8)	Done For Day (10)	Pending Cancel (12)	Pending Replace (11)	Replaced (3)	Canceled (5)	Rejected (2)	Stopped (7)
Pending New (2)	*								*	
New (2)	*	*	*	*	*	*	*		*	*
Partially Filled (4)		*	*	*	*	*		*		
Filled (8)		*	*			*				
Done for Day (10)			*							
Pending Cancel (12)	*	*	*		*			*		
Pending Replace (11)	*	*	*			*	*	*		
Replaced (3)		*								
Canceled (5)										



Rejected (2)										
Stopped (7)		*								

## Scenario Order State Change Matrices

1. Reading Order State Change Matrices: The mechanics of reading the order state change matrix is as follows.
2. Execution Report: The ‘Execution Report’ message is referred to as simply ‘Execution’.
3. Order State Change Request Messages: The ‘Order Cancel/Replace Request’ and the ‘Order Cancel Request’ messages are referred to as ‘Replace Request’ and ‘Cancel Request’ respectively.
4. Messages from the Buy-Side: The shaded rows represent the messages sent from the buy-side to the sell-side.
5. Times Common across Matrix Lines: In general, when two lines of the matrix share the same time, this means one of the following:
6. Request is Accepted or Rejected: There are two possible paths, e.g., a request is either accepted or rejected. In this case, the first row of the two possible paths is the reject case, which is italicized. The non-italicized row is the path that is continued by the remainder of the matrix.
7. Messages Sent in different Directions: Two messages are being sent at the same time but in different directions such that the messages cross on the connection, e.g., a cancel request is sent at the same time a sell-side sends an execution. In this case, both lines have bold text.
8. Marking Original/Replacing Orders: For scenarios involving cancel requests or cancel/replace requests, ‘X’ refers to the original order and ‘Y’ to the cancel/replacing order. A similar convention is used for corrections or cancels to executions.



## D1 Filled Order

D1 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0
3		Execution (X)	Rejected	Rejected	New	10000	0
3		Execution (X)	Partial Fill	Partially Filled	New	10000	2000
4		Execution (X)	Partial Fill	Partially Filled	New	10000	3000
5		Execution (X)	Fill	Filled	New	10000	10000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	0	0	<i>If order is rejected by sales</i>
2	10000	0	
3	0	0	<i>If order is rejected by trader/exchange</i>
3	8000	2000	Execution of 2000



4	7000	1000	Execution of 1000			
5	0	7000	Execution of 7000			

## D2 – Part-Filled Day Order, Done for Day

D2 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0
3		Execution (X)	Partial Fill	Partially Filled	New	10000	2000
4		Execution (X)	Partial Fill	Partially Filled	New	10000	3000
5		Execution (X)	Done for Day	Done for Day	New	10000	3000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	0	0	<i>If order is rejected</i>
2	10000	0	



3	8000	2000	Execution of 2000				
4	7000	1000	Execution of 1000				
5	0	0	Assuming day order. See other examples which cover GT orders				

### D3 – Cancel Request Issued for a Zero-filled Order

D3 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0
3	Cancel Request (Y, X)					10000	
4		Cancel Reject (Y, X)		New		10000	
4		Execution (Y, X)	Pending Cancel	Pending Cancel	New	10000	0
5		Cancel Reject (Y, X)		New		10000	
5		Execution (Y, X)	Canceled	Canceled	New	10000	0

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>



1			
2	0	0	<i>If order is rejected</i>
2	10000	0	
3			
4			<i>If rejected by salesperson</i>
4	10000	0	
5			<i>If rejected by trader/exchange</i>
5	0	0	

#### D4 – Cancel Request Issued for a Part-filled Order – Executions occur when Cancel Request is Active

D4 – Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0
3		Execution (X)	Partial Fill	Partially Filled	New	10000	2000
4	<b>Cancel Request (Y, X)</b>					<b>10000</b>	
4		Execution (X)	Partial Fill	Partially Filled	New	10000	5000
5		Cancel Reject (Y, X)		Partially Filled		10000	



5		Execution (Y, X)	Pending Cancel	Pending Cancel	New	10000	5000
6		Execution (X)	Partial Fill	Pending Cancel	New	10000	6000
7		<i>Cancel Reject (Y, X)</i>		<i>Partially Filled</i>		<i>10000</i>	
7		Execution (Y, X)	Canceled	Canceled	New	10000	6000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	0	0	<i>If order is rejected</i>
2	10000	0	
3	8000	2000	Execution for 2000
4			
4	<b>5000</b>	<b>3000</b>	<b>Execution for 3000. This execution passes the cancel request on the connection</b>
5			<i>If request is rejected</i>
5	5000	0	'Pending cancel' order status takes precedence over 'partially filled' order status
6	4000	1000	Execution for 1000 whilst order is pending cancel – 'pending cancel' order status takes precedence over 'partially filled' order status
7			<i>If request is rejected</i>
7	0	0	'Canceled' order status takes precedence over 'partially filled' order status

## D5 – Cancel Request issued for an Order that becomes Filled before Cancel Request can be Accepted

D5 – Order Flow Sequence:



<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0
3		Execution (X)	Partial Fill	Partially Filled	New	10000	2000
4	<b>Cancel Request (Y, X)</b>					<b>10000</b>	
4		Execution (X)	Partial Fill	Partially Filled	New	10000	5000
5		Cancel Reject (Y, X)		Partially Filled		10000	
5		Execution (Y, X)	Pending Cancel	Pending Cancel	New	10000	5000
6		Execution (X)	Fill	Pending Cancel	New	10000	10000
7		Cancel Reject (Y, X)		Filled		10000	

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	0	0	If order is rejected
2	10000	0	
3	8000	2000	Execution for 2000
4			
4	<b>5000</b>	<b>3000</b>	<b>Execution for 3000. This execution passes the cancel request on the connection</b>



5			<i>If request is rejected</i>			
5	5000	0	'Pending cancel' order status takes precedence over 'partially filled' order status			
6	0	5000	Execution for 5000 whilst order is pending cancel. 'Pending cancel' order status takes precedence over 'filled' order status			
7			Cancel request rejected  CxlRejectReason = 0  i.e., too late to cancel			

## D6 – Zero-filled Order, Cancel/Replace Request issued to increase Order Quantity

D6 – Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0
3	Replace Request (Y, X)					11000	
4		Cancel Reject (Y, X)		New		10000	
4		Execution (Y, X)	Pending Replace	Pending Replace	New	10000	0



5		<i>Cancel Reject (Y, X)</i>		<i>New</i>		10000	
5		Execution (Y, X)	Replace	Replaced	New	11000	0
6		Execution (Y)	Partial Fill	Partially Filled	New	11000	1000
7		Execution (Y)	Partial Fill	Partially Filled	New	11000	3000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	0	0	<i>If order is rejected by broker</i>
2	10000	0	
3			Request to increase order qty to 11000
4			<i>If request is rejected by salesperson</i>
4	10000	0	
5			<i>If rejected by trader/exchange</i>
5	11000	0	‘Replaced’ order status takes precedence over ‘new’ order status
6	10000	1000	Execution for 1000
7	8000	2000	Execution for 2000

**D7 – Part-filled Order, followed by Cancel/Replace Request to increase Order Quantity, Execution occurs while Order is Pending Replace**

D7 – Order Flow Sequence:



<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0
3		Execution (X)	Partial Fill	Partially Filled	New	10000	1000
4	Replace Request (Y, X)					12000	
5		Cancel Reject (Y, X)		Partially Filled		10000	
5		Execution (Y, X)	Pending Replace	Pending Replace	New	10000	1000
6		Execution (X)	Partial Fill	Pending Replace	New	10000	1100
7		Cancel Reject (Y, X)		Partially Filled		10000	
7		Execution (Y, X)	Replace	Partially Filled	New	12000	1100
8		Execution (Y)	Fill	Filled	New	12000	12000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	0	0	If order is rejected by broker
2	10000	0	
3	9000	1000	Execution for 1000
4			Request increase in order quantity to 12000



5			<i>If request is rejected</i>
5	9000	0	‘Pending replace’ order status takes precedence over ‘partially filled’ order status
6	8900	100	Execution for 100 before cancel/replace request is responded to
7			<i>If request is rejected</i>
7	10900	0	‘Partially filled’ order status takes precedence over ‘replaced’ order status
8	0	10900	Execution for 10900

## D8 – Filled Order Followed by Cancel/Replace to increase Order Quantity

D8 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0
3		Execution (X)	Fill	Filled	New	10000	10000
4	Replace Request (Y, X)					12000	
5		Cancel Reject (Y, X)		Filled		10000	
5		Execution (Y, X)	Pending Replace	Pending Replace	New	10000	10000
6		Cancel Reject (Y, X)		Filled		10000	
6		Execution (Y, X)	Replace	Partially Filled	New	12000	10000



7		Execution (Y)	Fill	Filled	New	12000	12000
---	--	---------------	------	--------	-----	-------	-------

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	0	0	<i>If order is rejected by broker</i>
2	10000	0	
3	0	10000	Execution for 10000
4			Request increase in order quantity to 12000
5			<i>If request is rejected</i>
5	0	0	‘Pending replace’ order status takes precedence over ‘partially filled’ order status
6			<i>If request is rejected</i>
6	2000	0	‘Partially filled’ order status takes precedence over ‘replaced’ order status.
7	0	2000	Execution for 2000

**D9 – Cancel/Replace Request – Not for Quantity Change – is rejected as a Fill has Occurred**

D9 – Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0



3		Execution (X)	Partial Fill	Partially Filled	New	10000	1000
4	<b>Replace Request (Y, X)</b>					<b>10000</b>	
4		Execution (X)	Fill	Filled	New	<b>10000</b>	<b>10000</b>
5		Cancel Reject (Y, X)		Filled		10000	

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	0	0	<i>If order is rejected by broker</i>
2	10000	0	
3	9000	1000	Execution for 1000
4			<b>Assume in this scenario that client does not wish to increase qty (e.g. client wants to amend limit price)</b>
4	0	9000	<b>Execution for 9000 – the replace request message and this execution report pass each other on the connection</b>
5			CxlRejectReason = 0  i.e., too late to cancel

**D10 – Cancel/Replace Request Sent while Execution is being Reported.  
The Requested Order Quantity exceeds the Cumulative Quantity. The Order is replaced then Filled**

D10 Order Flow Sequence:



<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0
3		Execution (X)	Partial Fill	Partially Filled	New	10000	1000
4	<b>Replace Request (Y, X)</b>					<b>8000</b>	
4		Execution (X)	Partial Fill	Partially Filled	New	10000	1500
5		Cancel Reject (Y, X)		Partially Filled		10000	
5		Execution (Y, X)	Pending Replace	Pending Replace	New	10000	1500
6		Execution (X)	Partial Fill	Pending Replace	New	10000	1600
7		Cancel Reject (Y, X)		Partially Filled		10000	
7		Execution (Y, X)	Replace	Partially Filled	New	8000	1600
8		Execution (Y)	Fill	Filled	New	8000	8000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	0	0	<i>If order is rejected</i>



2	10000	0	
3	9000	1000	Execution for 1000
4			<b>Request a decrease order quantity to 8000 (leaving 7000 open)</b>
4	<b>8500</b>	<b>500</b>	<b>Execution for 500 sent. Replace request and this execution report pass each other on the connection</b>
5			<i>If request is rejected by salesperson</i>
5	8500	0	'Pending replace' order status takes precedence over 'partially filled' order status
6	8400	100	Execution for 100 occurs before cancel/replace request is accepted
7			<i>If request is rejected by trader/exchange</i>
7	6400	0	'Partially filled' order status takes precedence over 'replaced' order status. Replace is accepted as requested order qty exceeds cum qty
8	0	6400	Execution for 6400.

**D11 – Cancel/Replace Request Sent while Order Execution is being Reported. The Requested Order Quantity equals the Cumulative Quantity. The Order Quantity is Amended to the Cumulative Quantity**

D11 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0
3	<b>Replace Request (Y, X)</b>					<b>7000</b>	
3		<b>Execution (X)</b>	Partial Fill	Partially Filled	New	<b>10000</b>	<b>7000</b>



4		Execution (Y, X)	Replace	Filled	New	7000	7000
---	--	------------------	---------	--------	-----	------	------

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	0	0	<i>If order is rejected by broker</i>
2	10000	0	
3			Client wishes to amend order qty to 7000 shares
3	3000	7000	Execution for 7000 - the replace message and this execution report pass each other on the connection
4	0	0	The replace request is interpreted as requiring the balance of the order to be canceled – the ‘filled’ order status takes precedence over ‘canceled’ or ‘replaced’

**D12 – Cancel/Replace Request Sent while Order Execution is being Reported. The Requested Order Quantity is below the Cumulative Quantity. The Order Quantity is Amended to the Cumulative Quantity**

D12 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0
3	Replace Request (Y, X)					7000	



3		Execution (X)	Partial Fill	Partially Filled	New	10000	8000
4		Execution (Y, X)	Replace	Filled	New	8000	8000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	0	0	<i>If order is rejected by broker</i>
2	10000	0	
3			Client wishes to amend order qty to 7000 shares
3	2000	8000	Execution for 8000 - the replace message and this execution report pass each other on the connection
4	0	0	The replace request is interpreted as requiring the balance of the order to be canceled – the ‘filled’ order status takes precedence over ‘canceled’ or ‘replaced’

### D13 – One Cancel/Replace Request is Issued which is Accepted.

Another One is Issued which is also Accepted

D13 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0



3		Execution (X)	Partial Fill	Partially Filled	New	10000	1000
4	Replace Request (Y, X)					8000	
5		Execution (Y, X)	Pending Replace	Pending Replace	New	10000	1000
6		Execution (X)	Partial Fill	Pending Replace	New	10000	1500
7		Execution (Y, X)	Replace	Partially Filled	New	8000	1500
8		Execution (Y)	Partial Fill	Partially Filled	New	8000	3500
9	Replace Request (Z, Y)					6000	
10		Execution (Z, Y)	Pending Replace	Pending Replace	New	8000	3500
11		Execution (Z, Y)	Replace	Partially Filled	New	6000	3500
12		Execution (Z)	Fill	Filled	New	6000	6000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	0	0	<i>If order is rejected by broker</i>
2	10000	0	
3	9000	1000	Execution for 1000
4			Request decrease in order quantity to 8000, leaving 7000 open
5	9000	0	‘Pending replace’ order status takes precedence over ‘partially filled’ order status
6	8500	500	Execution for 500
7	6500	0	‘Partially filled’ order status takes precedence over ‘replaced’ order status



8	4500	2000	Execution for 2000			
9			Request decrease in order quantity to 6000, leaving 2500 open			
10	4500	0				
11	2500	0	'Partially filled' order status takes precedence over 'replaced' order status			
12	0	2500	Execution for 2500			

**D14 – One Cancel/Replace Request is issued which is Rejected before Order becomes Pending Replace. Then Another is issued which is Accepted**

D14 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0
3		Execution (X)	Partial Fill	Partially Filled	New	10000	1000
4	Replace Request (Y, X)					8000	
5		Cancel Reject (Y, X)		Partially Filled		10000	
6		Execution (X)	Partial Fill	Partially Filled	New	10000	1500
7		Execution (X)	Partial Fill	Partially Filled	New	10000	3500
8	Replace Request (Z, X)					6000	



9		Execution (Z, X)	Pending Replace	Pending Replace	New	10000	3500
10		Execution (Z, X)	Replace	Partially Filled	New	6000	3500
11		Execution (Z)	Partial Fill	Partially Filled	New	6000	5000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	0	0	<i>If order is rejected by broker</i>
2	10000	0	
3	9000	1000	Execution for 1000
4			Request decrease in order quantity to 8000, leaving 7000 open
5			Request is rejected
6	8500	500	Execution for 500
7	6500	2000	Execution for 2000
8			Request decrease in order quantity to 6000, leaving 2500 open. Note that OrigClOrdID = X
9	6500	0	Note that OrigClOrdID = X



			Note that  OrigClOrdID = X
10	2500	0	
11	1000	1500	Execution for 1500

**D15 – One Cancel/Replace Request is Issued which is Rejected after it is in Pending Replace. Then Another One is Issued which is Accepted**

D15 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0
3		Execution (X)	Partial Fill	Partially Filled	New	10000	1000
4	Replace Request (Y, X)					8000	
5		Execution (Y, X)	Pending Replace	Pending Replace		10000	1000
6		Execution (X)	Partial Fill	Pending Replace	New	10000	1500
7		Cancel Reject (Y, X)		Partially Filled		10000	
8		Execution (X)	Partial Fill	Partially Filled	New	10000	3500



9	Replace Request (Z, X)					6000	
10		Execution (Z, X)	Pending Replace	Pending Replace	New	10000	3500
11		Execution (Z, X)	Replace	Partially Filled	New	6000	3500
12		Execution(Z)	Partial Fill	Partially Filled	New	6000	5000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	0	0	<i>If order is rejected by broker</i>
2	10000	0	
3	9000	1000	Execution for 1000
4			Request decrease in order quantity to 8000, leaving 7000 open
5	9000	0	
6	8500	500	Execution for 500. ‘Pending replace’ order status takes precedence over ‘partially filled’ order status
7			Request is rejected (e.g. by trader/exchange)
8	6500	2000	Execution for 2000
9			Request decrease in order quantity to 6000, leaving 2500 open. Note that OrigClOrdID = X
10	6500	0	
11	2500	0	
12	1000	1500	Execution for 1500



## D16 – One Cancel/Replace Request is Issued Followed immediately by Another. Broker Processes Sequentially

D16 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	New	New	New	10000	0
3		Execution (X)	Partial Fill	Partially Filled	New	10000	1000
4	Replace Request (Y, X)					8000	
5	Replace Request (Z, Y)					7000	
6		Execution (Y, X)	Pending Replace	Pending Replace	New	10000	1000
7		Execution (Y, X)	Replace	Partially Filled	New	8000	1000
8		Execution (Z, Y)	Pending Replace	Pending Replace	New	8000	1000
9		Execution (Z, Y)	Replace	Partially Filled	New	7000	1000
10		Execution (Z)	Fill	Filled	New	7000	7000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	10000	0	



3	9000	1000	Execution for 1000
4			Request decrease in order quantity to 8000, leaving 7000 open
5			Request decrease in order quantity to 7000, leaving 6000 open
6	9000	0	Broker processes Replace (Y, X) first
7	7000	0	Broker processes Replace (Y, X) first
8	7000	0	Broker then processes Replace (Z, Y)
9	6000	0	Broker then processes Replace (Z, Y)
10	0	6000	Execution for 6000

**D17 – One Cancel/Replace Request is issued followed immediately by Another. Broker rejects the Second as Order is Pending Replace**

1. Scenario where D17 is Invoked: This matrix illustrates the case where the broker does not support multiple outstanding order cancel or order cancel/replace requests.
2. D17 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	New	New	New	10000	0
3		Execution (X)	Partial Fill	Partially Filled	New	10000	1000
4	Replace Request (Y, X)					8000	
5	Replace Request (Z, Y)					7000	
6		Execution (Y, X)	Pending Replace	Pending Replace	New	10000	1000
7		Cancel Reject (Z, Y)		Pending Replace		10000	



8		Execution (Y, X)	Replace	Partially Filled	New	8000	1000
9		Execution (Y)	Partial Fill	Partially Filled	New	8000	3000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	10000	0	
3	9000	1000	Execution for 1000
4			Request decrease in order quantity to 8000, leaving 7000 open
5			Request decrease in order quantity to 7000, leaving 6000 open
6	9000	0	
7			Rejected because broker does not support processing of order cancel replace request whilst order is pending cancel.  CxlRejReason = 'Order already in pending cancel or pending replace status'
8	7000	0	'Partially filled' order status takes precedence over 'replaced' order status
9	5000	2000	Execution for 2000

## D18 – Telephoned Order

D18 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1							



2		Execution	New	New	New	10000	0
3		Execution	Partial Fill	Partially Filled	New	10000	2000
4		Execution	Partial Fill	Partially Filled	New	10000	3000
5		Execution	Fill	Filled	New	10000	10000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			Order for 10000 shares phoned to broker
2	0	0	Confirm that the broker has accepted the order – note that broker does not need to capture a ClOrdID
3	8000	2000	Execution of 2000
4	7000	1000	Execution of 1000
5	0	7000	Execution of 7000

## D19 – Unsolicited Cancel of a Part-timed Order

1. Scenario when D19 is Invoked: This scenario might occur if the Buy-side has not implemented Order Cancel requests or alternatively there is an electronic communication problem at the point that the buy-side wishes to send a cancel request.
2. D19 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	



2		<i>Execution (X)</i>	<i>Rejected</i>	<i>Rejected</i>	New	10000	0
2		Execution (X)	New	New	New	10000	0
3		Execution (X)	Partial Fill	Partially Filled	New	10000	1000
4							
5		Execution (X)	Canceled	Canceled	New	10000	1000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	0	0	<i>If order is rejected by broker</i>
2	10000	0	
3	9000	1000	Execution for 1000
4			Broker verbally agrees to cancel order
5	0	0	Broker signifies that order has been canceled.  ExecRestatementReason = Verbal change

## D20 – Unsolicited Replacement of a Part-filled Order

1. Scenario where D20 is Invoked: This scenario would occur if the buy-side has not implemented order cancel/replace requests or alternatively there is an electronic communication problem at the point where the buy-side wishes to send a cancel replace request.
2. D20 Order Flow Sequence:



<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0
3							
4		Execution (X)	Restated	New	New	11000	0
5		Execution (X)	Partial Fill	Partially Filled	New	11000	1000
6							
7		Execution (X)	Restated	Partially Filled	New	12000	1000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	0	0	<i>If order is rejected by broker</i>
2	10000	0	
3			Broker verbally agrees to increase order quantity to 11000
4	0	0	Broker signifies that order has been replaced.  ExecRestatementReason = Verbal change
5	10000	1000	Execution for 1000
6			Broker verbally agrees to increase order quantity to 12000
7	11000	0	Broker signifies that order has been replaced.  ExecRestatementReason = Verbal change



**D21 – Unsolicited Reduction of Order Quantity by Sell-side. For example, the US ECNs Communication NASDAQ SelectNet Declines**

D21 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0
3		Execution (X)	Restated	New	New	9000	0
4		Execution (X)	Fill	Filled	New	9000	9000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	0	0	<i>If order is rejected by broker</i>
2	10000	0	
3	9000	0	ExecRestatementReason = Partial Decline of OrderQty
4	0	9000	

**D22 – Order Rejected due to Duplicate ClOrdID**

D22 Order Flow Sequence:



<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	New	New	New	10000	0
3		Execution (X)	Partial Fill	Partially Filled	New	10000	1000
4	New Order (X)					10000	
5		Execution (X)	Rejected	Partially Filled	New	10000	1000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	10000	0	
3	9000	1000	Execution for 1000
4			Order submitted with the same order id
5	9000	0	OrdRejReason = duplicate order

## D23 – Order Rejected because Order has already been Verbally Submitted

1. Mechanism for Detecting Duplicate Orders: The sell-side may employ a number of mechanisms to detect that the electronic order is potentially a duplicate of a verbally passed order.
2. Mechanism #1: Check the *possdup* flag on the order message header.
3. Mechanism #2: Check the incoming order details against others from the client-side, i.e., side, quantity



4. Mechanism #3: Looking at the transaction time on the order as a guide to “staleness”
5. D23 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2							
3		Execution (X)	Rejected	Rejected	New	10000	0

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			Order for 10000 sent electronically
2			Order passed verbally as there is communication problem and order does not arrive. The verbally passed order starts getting executed
3	0	0	Order finally arrives and is detected as a duplicate of a verbal order and is therefore rejected.  OrdRejReason = duplicate of a verbal order

## **D24 – Order Status Request Rejected for Unknown Order**

D24 Order Flow Sequence:



<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	New	New	New	10000	0
3		Execution (X)	Partial Fill	Partially Filled	New	10000	1000
4	Status Request (Y)						
5		Execution (Y)	Rejected	Rejected	Status	0	0

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	10000	0	
3	9000	1000	Execution for 1000
4			
5	0		<p>OrdRejReason = unknown order</p> <p>LastShares not required when</p> <p>ExecTransType = Status</p>

## D25 – Transmitting a CMS-style “Nothing Done” in Response to a Status Request

D25 Order Flow Sequence:



<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0
3	Status Request (X)						
4		Execution (X)	New	New	Status	10000	0

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	0	0	<i>If order is rejected by broker</i>
2	10000	0	
3			
4	10000	0	Text = "Nothing Done"

## D26 - Order sent, immediately followed by a status request. Subsequent status requests sent during life of order

D26 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	



2	Status Request (X)						
3		Execution (X)	Pending New	Pending New	Status	10000	0
4		<i>Execution (X)</i>	<i>Rejected</i>	<i>Rejected</i>	<i>New</i>	<i>10000</i>	<i>0</i>
4		Execution (X)	New	New	New	10000	0
5	Status Request (X)						
6		Execution (X)	New	New	Status	10000	0
7		Execution (X)	Partial Fill	Partially Filled	New	10000	2000
8	Status Request (X)						
9		Execution (X)	Partial Fill	Partially Filled	Status	10000	2000
10		Execution (X)	Fill	Filled	New	10000	10000
11	Status Request (X)						
12		Execution (X)	Fill	Filled	Status	10000	10000
13	Replace Request (Y, X)					12000	
14		Execution (Y, X)	Pending Replace	Pending Replace	New	10000	10000
15		Execution (Y, X)	Replace	Partially Filled	New	12000	10000
16	Status Request (X)						
17		Execution (Y, X)	Partial Fill	Partially Filled	Status	12000	10000
18	Status Request (Y)						
19		Execution (Y)	Partial Fill	Partially Filled	Status	12000	10000

<u>Time</u>	<u>Leaves</u> <u>Quantity</u>	<u>Last</u> <u>Shares</u>	<u>Comment</u>
1			
2			



3	10000		Sent in response to status request. LastShares not required when ExecTransType = status
4	0	0	<i>If order is rejected</i>
4	10000	0	
5			
6	10000		Sent in response to status request
7	8000	2000	Execution for 2000
8			
9	8000		Sent in response to status request
10	0	8000	Execution for 8000
11			
12	0		Sent in response to status request
13			Request to increase order quantity
14	0	0	
15	2000	0	
16			
17	2000		Sent in response to status request. Note reference to X to allow tie back of execution report to status request
18			
19	2000		Sent in response to status request

## **D27 – GTC Order Partially Filled, Restated/Renewed and Partially Filled the Following Day**

D27 Order Flow Sequence:



<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>Ord Status</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
Day 1,1	New Order (X)					10000	
Day 1,2		Execution (X)	New	New	New	10000	0
Day 1,3		Execution (X)	Partial Fill	Partially Filled	New	10000	2000
Day 1,4		Execution (X)	Done for Day	Done for Day	New	10000	2000
Day 2,1		Execution (X)	Restated	Partially Filled	New	10000	2000
Day 2,2		Execution (X)	Partial Fill	Partially Filled	New	10000	3000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Day Order Quantity</u>	<u>Day Cumulative Quantity</u>	<u>Comment</u>
Day 1,1					
Day 1,2	10000	0			
Day 1,3	8000	2000			Execution for 2000
Day 1,4	8000	0			Optional at end of trading day
Day 2,1	8000	0	8000	0	ExecRestatementReason = GTC renewal/restatement  No change – optionally sent the following morning
Day 2,2	7000	1000	8000	1000	Execution for 1000

## D28 – GTC Order with a Partial Fill, a 2:1 Stock Split, then a Partial Fill and Fill the Following Day

D28 Order Flow Sequence:



<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>Ord Status</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
Day 1,1	New Order (X)					10000	
Day 1,2		Execution (X)	New	New	New	10000	0
Day 1,3		Execution (X)	Partial Fill	Partially Filled	New	10000	2000
Day 1,4		Execution (X)	Done for Day	Done for Day	New	10000	2000
Day 2,1		Execution (X)	Restated	Partially Filled	New	20000	4000
Day 2,2		Execution (X)	Partial Fill	Partially Filled	New	20000	9000
Day 2,3		Execution (X)	Fill	Filled	New	20000	20000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Day Order Quantity</u>	<u>Day Cumulative Quantity</u>	<u>Comment</u>
Day 1,1					
Day 1,2	10000	0			
Day 1,3	8000	2000			Execution for 2000 @ 50
Day 1,4	8000	0			Optional at end of trading day
Day 2,1	16000	0	16000	0	<p>Sent the following morning after the split</p> <p>ExecRestatementReason = GTC corporate action</p> <p>AvgPx = 25</p> <p>DayAvgPx = 0</p>
Day 2,2	11000	5000	16000	5000	Execution for 5000



Day 2,3	0	11000	16000	16000	Execution for 11000
---------	---	-------	-------	-------	---------------------

## D29 – GTC Order Partially Filled, Restated/Renewed and Canceled the Following Day

D29 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>Ord Status</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
Day 1,1	New Order (X)					10000	
Day 1,2		Execution (X)	New	New	New	10000	0
Day 1,3		Execution (X)	Partial Fill	Partially Filled	New	10000	2000
Day 1,4		Execution (X)	Done for Day	Done for Day	New	10000	2000
Day 2,1		Execution (X)	Restated	Partially Filled	New	10000	2000
Day 2,2	Cancel Request (Y, X)					10000	
Day 2,3		<i>Cancel Reject (Y, X)</i>		<i>Partially Filled</i>		<i>10000</i>	
Day 2,3		Execution (Y, X)	Pending Cancel	Pending Cancel		10000	2000
Day 2,4		Cancel Reject (Y, X)		Partially Filled		10000	
Day 2,4		Execution (Y, X)	Canceled	Canceled		10000	2000



<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Day Order Quantity</u>	<u>Day Cumulative Quantity</u>	<u>Comment</u>
Day 1,1					
Day 1,2	10000	0			
Day 1,3	8000	2000			Execution for 2000
Day 1,4	8000	0			Optional at end of trading day
Day 2,1	8000	0	8000	0	<p>ExecRestatementReason = GTC renewal/restatement (no change)</p> <p>Optionally sent the following morning</p>
Day 2,2					
Day 2,3					<i>If rejected by salesperson</i>
Day 2,3	8000	0	8000	0	
Day 2,4					If rejected by trader/exchange
Day 2,4	0	0	8000	0	

### D30 – GTC Order Partially Filled, Restated/Renewed Followed by Replace Request to increase Quantity

D30 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>Ord Status</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
Day 1,1	New Order (X)					10000	
Day 1,2		Execution (X)	New	New	New	10000	0



Day 1,3		Execution (X)	Partial Fill	Partially Filled	New	10000	2000
Day 1,4		Execution (X)	Done for Day	Done for Day	New	10000	2000
Day 2,1		Execution (X)	Restated	Partially Filled	New	10000	2000
Day 2,2	Replace Request (Y, X)					15000	
Day 2,3		<i>Cancel Reject (Y, X)</i>		<i>Partially Filled</i>		10000	
Day 2,3		Execution (Y, X)	Pending Replace	Pending Replace		10000	2000
Day 2,4		Execution (X)	Partial Fill	Pending Replace		10000	3000
Day 2,5		<i>Cancel Reject (Y, X)</i>		Partially Filled		10000	
Day 2,5		Execution (Y, X)	Replace	Partially Filled		15000	3000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Day Order Quantity</u>	<u>Day Cumulative Quantity</u>	<u>Comment</u>
Day 1,1					
Day 1,2	10000	0			
Day 1,3	8000	2000			Execution for 2000
Day 1,4	8000	0			Optional at end of trading day
Day 2,1	8000	0	8000	0	<p>ExecRestatementReason = GTC renewal/restatement (no change)</p> <p>Optionally sent the following morning</p>
Day 2,2					Increasing quantity



Day 2,3					<i>If rejected by salesperson</i>
Day 2,3	8000	0	8000	0	
Day 2,4	7000	1000	8000	1000	Execution for 1000
Day 2,5					If rejected by trader/exchange
Day 2,5	12000	0	13000	1000	

## D31 – Possible Resend Order

D31 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	New	New	New	10000	0
3	New Order (X)					10000	
4		Execution (X)	New	New	Status	10000	0
5	New Order (Y)					15000	
6		Execution (Y)	New	New	New	15000	0

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	10000	0	
3			PossResend = Y



Because order X has already been received, confirm back the current state of the order. Last shares not required when

ExecTransType = Status

PossResend = Y

Because order Y has not been received before, confirm back as a new order.

### D32 – Fill or Kill Order cannot be Filled

D32 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec</u> <u>Type</u>	<u>OrdStatus</u>	<u>Exec</u> <u>Trans</u> <u>Type</u>	<u>Order</u> <u>Quantity</u>	<u>Cumulative</u> <u>Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0
3		Execution (X)	Canceled	Canceled	New	10000	0

<u>Time</u>	<u>Leaves</u> <u>Quantity</u>	<u>Last</u> <u>Shares</u>	<u>Comment</u>
1			Order is FOK
2	0	0	If order is rejected by broker
2	10000	0	
3	0	0	If order cannot be immediately filled



## D33 – Immediate-Or-Cancel Order that cannot be immediately Hit

D33 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0
3		Execution (X)	Partial Fill	Partially Filled	New	10000	1000
4		Execution (X)	Canceled	Canceled	New	10000	1000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			Order is IOC
2	0	0	If order is rejected by broker
2	10000	0	
3	9000	1000	Execution for 1000
4	0	0	If order cannot be immediately hit

## D34 – Filled Order, Followed by Correction and Cancelation of Executions

D34 Order Flow Sequence:



<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0
3		Execution (X)	Partial Fill	Partially Filled	New	10000	1000
4		Execution (X)	Fill	Filled	New	10000	10000
5		Execution (X)	Partial Fill	Partially Filled	Cancel	10000	9000
6		Execution (X)	Partial Fill	Partially Filled	Correct	10000	9000
7		Execution (X)	Fill	Filled	New	10000	10000
8		Execution (X)	Fill	Filled	Correct	10000	10000
9	Replace Request (Y, X)					12000	
10		Execution (Y, X)	Pending Replace	Pending Replace	New	10000	10000
11		Execution (Y, X)	Replace	Partially Filled	New	12000	10000
12		Execution (Y)	Partial Fill	Partially Filled	Correct	12000	10500

<u>Time</u>	<u>Leaves Quantity</u>	<u>AvgPx</u>	<u>Last Shares</u>	<u>Last Px</u>	<u>ExecId (ExecRefID)</u>	<u>Comment</u>
1						
2	0		0		A	If order is rejected by broker
2	10000	0	0		B	
3	9000	100	1000	100	C	Execution for 1000 @ 100
4	0	109	9000	110	D	Execution for 9000 @ 110
5	1000	110	0	0	E (C)	Cancel execution for 1000
6	1000	100	9000	100	F (D)	Correct price on execution for 9000 to 100



7	0	102	1000	120	G	Execution for 1000 @ 120
8	0	120	9000	120	H (F)	Correct price on execution for 9000 to 120
9						Request to increase order quantity
10	0	120	0	0	I	
11	2000	120	0	0	J	
12	1500	120	9500	120	K (H)	Correct execution of 9000 @ 120 to 9500 @ 120

## D35 – A Canceled Order Followed by a Busted Execution and a new Execution

D35 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>Ord Status</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	New	New	New	10000	0
3		Execution (X)	Partial Fill	Partially Filled	New	10000	5000
4	Cancel Request (Y, X)					10000	
5		Execution (Y, X)	Pending Cancel	Pending Cancel	New	10000	5000
6		Execution (Y, X)	Canceled	Canceled	New	10000	5000
7		Execution (Y)	Partial Fill	Canceled	Cancel	10000	0
8		Execution (Y)	Partial Fill	Canceled	New	10000	4000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>ExecID (ExecRefID)</u>	<u>Comment</u>



1						
2	10000	0	A			
3	5000	5000	B		LastPx = 50	
4						
5	5000	0	C			
6	0	0	D			
7	0	0	E (B)		Cancel of the execution. ‘Canceled’ order status takes precedence over ‘New’	
8	0	4000	F		Fill for 4000  LastPx = 51	

### **D36 – GTC Order Partially Filled, Restated/Renewed, and Partially Filled the Next Day, with Corrections of Quantities on both Executions**

D36 Order Flow Sequence:

<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>Ord Status</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
Day 1,1	New Order (X)					10000	
Day 1,2		Execution (X)	New	New	New	10000	0
Day 1,3		Execution (X)	Partial Fill	Partially Filled	New	10000	2000
Day 1,4		Execution (X)	Done for Day	Done for Day	New	10000	2000
Day 2,1		Execution (X)	Restated	Partially Filled	New	10000	2000
Day 2,2		Execution (X)	Partial Fill	Partially Filled	New	10000	3000
Day 2,3		Execution (X)	Partial Fill	Partially Filled	Correct	10000	2500



Day 2,4		Execution (X)	Partial Fill	Partially Filled	Correct	10000	2000
---------	--	---------------	--------------	------------------	---------	-------	------

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Day Order Quantity</u>	<u>Day Cumulative Quantity</u>	<u>ExecID (ExecRefID)</u>	<u>Comment</u>
Day 1,1						
Day 1,2	10000	0			A	
Day 1,3	8000	2000			B	Execution for 2000
Day 1,4	8000	0			C	Optional at end of trading day
Day 2,1	8000	0	8000	0	D	<p>ExecRestatementReason = GTC renewal /restatement (no change)</p> <p>Optionally sent the following morning</p>
Day 2,2	7000	1000	8000	1000	E	Execution for 1000
Day 2,3	7500	1500	8500	1000	F (B)	Correct quantity on previous day's execution from 2000 to 1500
Day 2,4	8000	500	8500	500	G (E)	Correct quantity on today's execution from 1000 to 500

### **D37 – Transmitting a Guarantee of Execution Prior to Execution**

D37 Order Flow Sequence:



<u>Time</u>	<u>Message Received</u> (ClOrdID, OrigClOrdID)	<u>Message Sent</u> (ClOrdID, OrigClOrdID)	<u>Exec Type</u>	<u>OrdStatus</u>	<u>Exec Trans Type</u>	<u>Order Quantity</u>	<u>Cumulative Quantity</u>
1	New Order (X)					10000	
2		Execution (X)	Rejected	Rejected	New	10000	0
2		Execution (X)	New	New	New	10000	0
3		Execution (X)	Partial Fill	Stopped	New	10000	0
4		Execution (X)	Partial Fill	Stopped	New	10000	1000

<u>Time</u>	<u>Leaves Quantity</u>	<u>Last Shares</u>	<u>Comment</u>
1			
2	0	0	<i>If order is rejected by broker</i>
2	10000	0	
3	10000	1000	Text = "You are guaranteed to buy 1000 at 50.10"  LastPx = 50.10  This is similar to the concept of a 'protected' trade
4	9000	1000	LastPx = 50  Executed price is better than guaranteed



## Central Limit Order Book

### Overview

1. Definition of a Central LOB: A *central limit order book – CLOB* – is a centralized database of limit orders proposed by the SEC in 2000. However, the concept was opposed by securities companies (Wikipedia (2023)).
2. CLOB as an Order Matching System: A central limit order book or “CLOB” is a trading method used by most exchanges globally. It is a transparent system that matches customer orders, e.g., bids and offers, on a “price time priority” basis.
3. Best Market or “Touch”: The highest/best bid order and the lowest/offer order constitutes the best market or “the touch” in a given security or swap context.
4. Crossing the Bid/Ask Gap: Customers can routinely cross the bid/ask spread to effect immediate execution.
5. CLOB Market Depth or “Stack”: Customers can also see the market depth or the “stack” in which customers can view bid orders for various sizes and prices on one side vs. viewing offer orders at various sizes and prices on the other side.
6. Fully Transparent, Real-time, or Anonymous: The CLOB is by definition fully transparent, real-time, anonymous, and low cost in execution.
7. D2D, D2C, and C2C Trades: In the CLOB model, customers can trade directly with dealers, dealers can trade with other dealers, and importantly, customers can trade directly with customers anonymously.
8. The Request-for-Quote Method: In contrast to the CLOB approach is the Request-for-Quote – “RFQ” trading method.
9. RFQ as an Asymmetric Trade Execution Model: In this method, a customer queries a finite set of participant market makers who quote a bid/offer – a market – to the customer.



10. Trading Options available to Customers: The customer may only “hit the bid”, i.e., sell to the highest bidder, or “lift the offer”, i.e., buy from the cheapest seller.
11. Stepping inside Bid/Ask Group: The customer is prohibited from stepping inside the bid/ask spread and thereby reducing its execution fees.
12. Only D2C Allowed: Contrary to the CLOB model, customers can only trade with dealers. They cannot trade with other customers, and importantly, they cannot make markets themselves.

## References

- Wikipedia (2023): [Central Limit Order Book](#)



## How Storing Supply and Demand affects Price Diffusion

### Abstract

1. The LOB as a Capacitor: The limit order book is a device for storing supply and demand in financial markets, somewhat like a capacitor is a device for storage charge.
2. Flow-oriented LOB Model: This chapter develops a microscopic statistical model of the limit order book under random order flow using simulation, dimensional analysis, and an analytic treatment based off of a master equation.
3. Testable Predictions of Order Book Properties: It makes testable predictions of the price diffusion rate, the depth of stored demand vs. price, the bid-ask spread, and the price impact function, and shows that even under completely random order flow, the process of storing supply and demand induces anomalous diffusion and temporal structure in prices.

### Main

1. Bachelier Price Process: The random walk model was originally introduced by Bachelier (1900) to describe prices, 5 years before Einstein used it to model Brownian motion.
2. Modeling the Price Formation Process: This chapter takes the Bachelier model to a deeper level by modeling the microscopic mechanism of price formation (Daniels, Farmer, Iori, and Smith (2001)).
3. Order Flow induced LOB Adjustments: It shows how the need to store supply and demand in and of itself induces structure in price changes.



4. Price Response to Supply/Demand Fluctuations: The model enables the study of price diffusion rates, and gives a prediction for the universal functional form for the response of prices for small fluctuations in supply and demand.
5. Price Properties from Order Flow: It shows how the most basic properties of the market, such as spread, liquidity, and volatility emerge, emerge naturally from properties of order flow.
6. Parameter free, Dimensionless Model: The model makes falsifiable predictions with no free parameters.
7. Zero-intelligence, Random Agents: It differs from the standard models in economics in that the agents are assumed to have zero-intelligence and that their behavior is random (Gode and Sunder (1993)).
8. Mismatch between Buyers and Sellers: Most modern markets operate continuously. The mismatch between buyers and sellers that typically exists at any given time is solved via an order-based market with 2 basic kinds of orders.
9. Market Orders: Impatient traders submit *market orders*, which are requests to buy or sell a given number of shares at the best available price.
10. Limit Orders: More patient traders submit *limit orders*, which also state a limit price, corresponding to the worst allowable price for the transaction.
11. Limit Order Book: Limit orders often fail to result in an immediate transaction, and are stored in a queue called the *limit order book*.
12. Bids vs. Offers/Asks: Buy limit orders are called *bids*, and sell limit orders are called *offers* or *asks*. The best – or lowest – offer is labeled  $a(t)$  and the best – highest bid is  $b(t)$ .
13. Spread: There is typically a non-zero price gap between them called *spread*

$$s(t) = a(t) - b(t)$$

14. Matching of Market Orders: As market orders arrive, they are matched against limit orders of the opposite sign in order of price and arrival time.



15. Matching over Multiple Price Levels: Because orders are placed for varying numbers of shares, matching is not one-to-one.
16. Multi-level Matching - Example: For example, suppose the best offer is for 200 shares at \$60.00, and the next best is for 300 shares at \$60.25; a buy market order for 250 shares buys 200 shares at \$60 and 50 shares at \$60.25, moving the best offer  $a(t)$  from \$60 to \$60.25.
17. Dense LOB means High Liquidity: A high density of limit orders results in high *liquidity* for market orders, i.e.,  $t$  decreases the average price movement when a market order is placed.
18. Order Flows as Poisson Process: A simple random order placement model shown in the figure below is proposed. All order flows are modeled as Poisson processes.

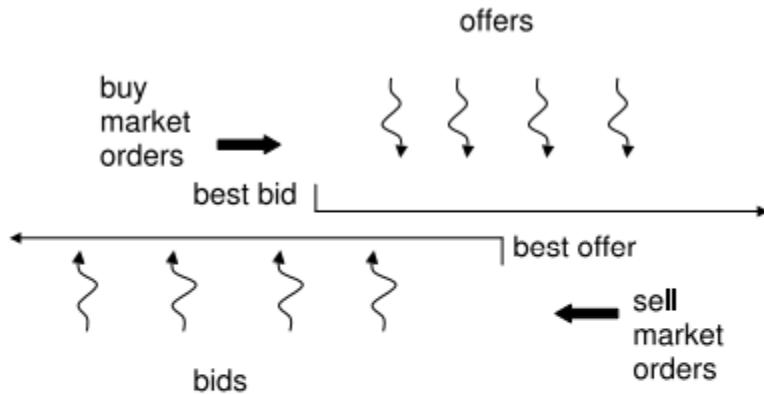


FIG. 1: Schematic of the order-placement and clearing process. New offers (sell limit orders) can be placed at any price greater than the best bid, and new bids (buy limit orders) can be placed at any price less than the best offer.

19. Market Order Arrivals: Market orders in chunks of  $\sigma$  shares arrive at the rate of  $\mu$  shares per unit time, with an equal probability for buy and sell orders.
20. Limit Order Arrivals: Similarly, limit orders in chunks of  $\sigma$  shares arrive at the rate of  $\alpha$  shares per unit price and unit time.



21. Uniform Placement of Limit Orders: Offers are placed at a uniform probability at integer multiples of a tick size  $p_0$  in the range

$$b(t) < p < \infty$$

and similarly for bids on

$$-\infty < p < a(t)$$

22. Impact of a Market Order: When a market order arrives, it causes a transaction; under the assumption of constant order size, a buy market order removes an offer at price  $a(t)$  and a sell market order removes a bid at price  $b(t)$
23. Explicit Limit Order Removal: Alternatively, limit orders can also be removed spontaneously by being canceled or by expiring.
24. Explicit Limit Order Removal Rate: This is modeled by letting them be removed randomly with a constant probability  $\delta$  per unit time.
25. Analytic Solution: The order placement setup above is designed to permit an analytic solution.
26. Previous Work: This chapter builds on previous work on limit order modeling (Domowitz and Wang (1994), Bollerslev, Domowitz, and Wang (1997), Eliezer and Kogan (1998), Maslov (2000), Matassini and Franci (2001), Slanina (2001), Challet and Stinchcombe (2001), Iori and Chiarella (2002)).
27. Order Placement over Infinite Interval: While the assumption of limit order placement over an infinite interval is clearly unrealistic, it provides a tractable boundary condition for modeling the behavior of the limit order book in the region of interest, near the midpoint price

$$m(t) = \frac{a(t) + b(t)}{2}$$



28. Choice of Representation: For convenience, the model has been formulated in terms of price differences instead in terms of percentage price changes. This leads to the problem that prices can become negative, and that the volatility of the price returns depends on the scale of the prices. This can be fixed by letting

$$p \rightarrow \log p$$

but this has the disadvantage that real order books do not have logarithmic price ticks. Alternatively, one could require that the volatility of price returns be independent of price, which leads to

$$\alpha(p) = \frac{\alpha_0}{p}$$

since  $\alpha$  is the only parameter that depends on prices. From this point of view, the solution presented here is an approximation over time and price scales where  $\alpha(p)$  can be considered constant.

29. Limit Order Far from Mid: It is also justified because the limit order placed far from mid usually expire or they are canceled before they are executed.
30. Constant Probability of Cancelation: Assuming a constant probability of cancelation is clearly *ad hoc*, but in simulations, one finds that other assumptions – such as constant duration time – give similar results.
31. Constant Order Size: The analytic model here uses a constant order size  $\sigma$ . Variable order sizes have been used in the simulations, e.g., half-normal distributions with standard deviation  $\sqrt{\frac{2}{\pi}}\sigma$ . The differences do not affect any of the results reported here.
32. Definition of the Limit Order: For simplicity, a limit order here is defined as any order that is not executed immediately.
33. Limit Order High Execution Likelihood: In reality, a limit order with an aggressive limit price can result in – perhaps partial – immediate execution. The part that is



executed is treated as an effective market order, and the part that remains is an effective limit order.

34. Boundary Conditions for Order Placement: This automatically determines the boundary conditions of the order placement process, since an offer with

$$p \leq b(t)$$

or a bid with

$$p \geq a(t)$$

would result in an immediate transaction, and this is effectively the same as a market order.

35. Price Range of Order Placement: Note that these boundary conditions realistically allow limit orders to be placed at prices anywhere inside the current spread.
36. Depth Profile at  $p, t$ : The distribution over the values of *depth profile*  $N(p, t)$  is sought, i.e., the density of shares in the order book at price  $p$  and time  $t$ .
37. Depth Profile Sign Convention: For convenience, the depth is positive for bids and negative for offers.
38. Random Walk of Mid-price: From the symmetry of the order process, mid-point prices make a random walk, with a non-stationary distribution.
39. Use of Co-moving Coordinates: The key to finding a stationary analytic solution for the average depth is to use co-moving coordinates.
40. Analysis using Centered Prices: Without loss of generality, one studies the depth of offers using price coordinates centered at the midpoint  $m(t)$  so that

$$b(t) \equiv -a(t)$$

41. Independence of Adjacent Fluctuations: Fluctuations about the mean depth at adjacent prices are treated as independent.



42. Probability Density for  $N$  and  $p$  at  $t$ : This allows replacing the distribution over depth profile with a simpler density over occupation number  $N$  at each  $p$  and  $t$ .
43. Infinitesimal Bin Size: The bin size

$$p_0 \rightarrow dp$$

is let to go infinitesimal. With finite order flow rates, this will give vanishing probability for the existence of more than one order in any bin as

$$dp \rightarrow 0$$

44.  $\pi$ ,  $P_+$ , and  $P_-$ : Let  $\pi(N, p, t)$  be the probability that an interval  $dp$  centered around price  $p$  has  $N$  shares at time  $t$ , and let  $P_+(\Delta p, t)$  be the probability that  $m(t)$  increases by  $\Delta p$ , and  $P_-(\Delta p, t)$  be the probability that it decreases by  $\Delta p$ .
45. Ignoring Joint Event Occurrence: Assume that  $\alpha$ ,  $N\delta$ ,  $P_+$ , and  $P_-$  are small enough so that higher order terms corresponding to simultaneous events can be neglected.
46. General Master Equation for  $\pi$ : Going to continuous time, a general master equation for  $\pi$  can be written as

$$\begin{aligned} \frac{\partial \pi(N, p)}{\partial t} = & \frac{\alpha(p)dp}{\sigma} [\pi(N - \sigma, p) - \pi(N, p)] + \frac{\delta}{\sigma} [(N + \sigma)\pi(N + \sigma, p) - N\pi(N, p)] \\ & + \frac{\mu(p)}{2\sigma} [\pi(N + \sigma, p) - \pi(N, p)] \\ & + \sum_{\Delta p} P_+(\Delta p)[\pi(N, p - \Delta p) - \pi(N, p)] \\ & + \sum_{\Delta p} P_-(\Delta p)[\pi(N, p + \Delta p) - \pi(N, p)] \end{aligned}$$

The  $t$  variable appears in every term.



47. Component Contributions to Master Density: The  $\alpha$  term corresponds to the receipt of a limit order, the  $\delta$  term to spontaneous removal of a limit order, the  $\mu$  term to the receipt of the market order, the  $P_+$  term to an increase in the midpoint

$$m(t) \rightarrow m(t) + \Delta p$$

and the  $P_-$  term to a decrease

$$m(t) \rightarrow m(t) - \Delta p$$

48. Co-moving Reference Frame for Mid-price Diffusion: The last two terms in the master equation are particularly important because they imply a diffusion process for the depth when viewed in the co-moving reference point of the price coordinate when the midpoint moves.
49. Definition of the Best Offer: The definition of the best offer gives the boundary condition

$$\pi(N, p) = 0$$

for

$$p < 0$$

50. Mean Field Solution: The next step is to seek a mean-field solution. In the co-moving reference frame,  $\mu(p)$  becomes a function corresponding to the rate at which market orders are executed at a price  $p$ .
51. Analogy to a Moving Particle: The market order can be thought of as a *particle* that is *created* at  $b(t)$  and moves to the right until it is *absorbed* at price  $p$ .
52. Approximate Solution to the Master Equation: An approximate solution to



$$\begin{aligned}
\frac{\partial \pi(N, p)}{\partial t} = & \frac{\alpha(p) dp}{\sigma} [\pi(N - \sigma, p) - \pi(N, p)] + \frac{\delta}{\sigma} [(N + \sigma)\pi(N + \sigma, p) - N\pi(N, p)] \\
& + \frac{\mu(p)}{2\sigma} [\pi(N + \sigma, p) - \pi(N, p)] \\
& + \sum_{\Delta p} P_+(\Delta p) [\pi(N, p - \Delta p) - \pi(N, p)] \\
& + \sum_{\Delta p} P_-(\Delta p) [\pi(N, p + \Delta p) - \pi(N, p)]
\end{aligned}$$

is found by relating the distributions  $P_+(\Delta p)$  and  $P_-(\Delta p)$  self-consistently to mean-functional forms  $\alpha(p)$  and  $\mu(p)$  for which the parameters  $\alpha$  and  $\mu$  furnish boundary conditions.

- 53. [Steady State Value for  \$N\(p\)\$](#) : The steady-state mean value  $\langle N(p) \rangle$  is then obtained from the generating functional for the moments of  $\pi$ .
- 54. [Analytic Solutions vs. Simulation Results](#): The figure below compares the analytic solutions in the non-dimensional form to simulation results for the cumulative distribution of the spread and the average depth  $\langle N(p) \rangle$ .

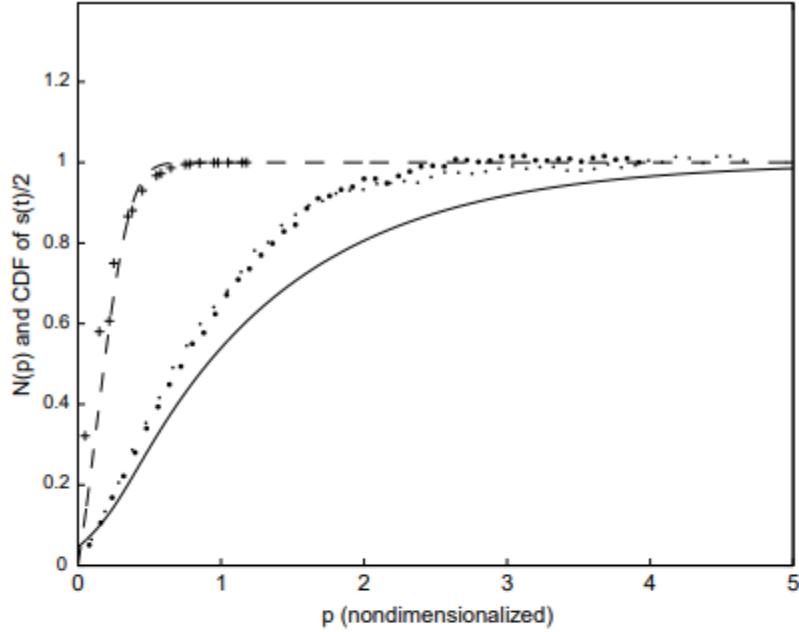


FIG. 2: A comparison of analytic and simulation results. Simulation results for the cumulative distribution of the spread/2 are shown as crosses, and the mean depth of offers  $\langle N(p) \rangle$  as circles. We plot the nondimensionalized depth  $\hat{N} = \delta N/\alpha$  versus the nondimensionalized price  $\hat{p} = 2\alpha p/\mu$ , measured relative to the midpoint. In units with the mean order size  $\sigma$  and the tick size  $p_0$  set to one,  $\alpha = 0.002$ . Two simulation results are shown, with  $\delta = 10^{-3}$  and  $\mu = 0.1$  in one case, and  $\delta = 10^{-4}$  and  $\mu = 0.01$  in the other. The curves (dashed for spread, solid for depth) are from the self-consistent solution of Eq. (1)

55. Qualitative Similarity among Curves: For a wide range of parameters, non-dimensionalization collapses all the results into qualitatively similar curves.
56. Effectiveness of the Mean-field Approach: The simulation results approximately match the mean-field solution; thus, the mean-field analysis does a good job of capturing the leading order properties.



57. Asymptotic Depth and Spread: In the parameter range shown, the asymptotic depth varied by a factor of ten and the width of the transition from the midpoint to the asymptotic region varied by three orders of magnitude.
58. Market Price Impact Function: The liquidity for executing a market order can be characterized by a price impact function

$$\Delta p = \phi(\omega, \tau, t)$$

59. Definition of Market Impact:  $\Delta p$  is the price shift at time  $t + \tau$  caused by a market order of  $\omega$  orders.
60. Deriving the Price Impact Function: One of the main results is the derivation of the average price impact function. This is important for practical reasons such as minimizing transaction costs, and also because it is in a sense inverse demand function, providing a natural starting point for theories on statistical or dynamical properties of markets (Bouchaud and Cont (2002), Farmer (2002)).
61. Naïve Expectation of Convex Impact: Naïve arguments predict that  $\phi$  should be a convex function – with increasing derivative – for positive  $\omega$ .
62. Reasoning behind the Naïve Expectation: The naïve argument goes as follows:  
Fractional price changes should not depend on the scale of the price.
63. Buying  $\omega$  Shares Back-to-back: Suppose buying a single share raises the price by a factor

$$a > 1$$

If  $a$  is a constant, buying  $\omega$  shares in succession should raise it by  $a^\omega$ .

64. Implicit Convex Nature of the Impact: Thus, if buying  $\omega$  shares all at once affects the price as much as buying them one at a time, price impact should be strongly convex.
65. Consequence of Concave Market Impact: It is common practice to break up orders in order to reduce losses due to market impact. With a sufficiently concave market impact function, in contrast, it is cheaper to execute an order all at once.



66. Empirical Evidence for Concavity: In contrast, from empirical studies,  $\phi(\omega)$  for

$$\omega > 0$$

appears to be concave (Huasman and Lo (1992), Farmer (1996), Torre and Ferrari (1997), Kempf and Korn (1998), Plerou, Gopikrishnan, Gabaix, and Stanley (2002)).

At least for small  $\omega$ , these studies suggest a function of the form

$$\phi(\omega) \sim \omega^\beta$$

67. Estimates for  $\beta$ : Estimates for  $\beta$  are poor, and some of the results are contradictory, but on balance the literature appears to indicate that

$$\beta \sim 0.5$$

68. Instantaneous Price Impact: The model here predicts this result. The instantaneous price impact  $\phi(\omega, 0, t)$  depends on the depth of orders as a function of time.

69. Average Depth Fluctuation: Although the depth at any given time is a discontinuous random function, the average depth  $\langle N(p) \rangle$  can be approximated as a smooth function that vanishes at the midpoint.

70. Taylor Expansion for Average Depth: Providing its derivative exists, it can be expanded in a Taylor series, and the leading term can be written as

$$\langle N(p) \rangle \approx 2\lambda p$$

71. Time-varying Liquidity: We let the instantaneous depth  $N(p, t)$  near the midpoint as a function of this form, with time varying liquidity  $\lambda(t)$

72. Walking the Order Book Depth: The shift in the price caused by a market order

$$\omega > 0$$



can be approximated using the continuum transaction condition

$$\omega = \int_0^{\Delta p} N(p, t) dp$$

which says that the size of the market order equals the number of shares removed from the book.

73. Linear Approximation over Time Averages: Plugging in the linear approximation and taking time averages yields the expected price impact conditioned on  $\omega$  as

$$\langle \Delta p(\omega) \rangle = \langle \sqrt{\frac{\omega}{\lambda(t)}} \rangle$$

74. Generality Inherent in the Derivation: This is confirmed in the numerical experiments. Note that the power  $\frac{1}{2}$  only depends on the assumption of non-zero derivative at

$$p = 0$$

so the result is generic.

75. Time Average of  $\langle \sqrt{\frac{1}{\lambda(t)}} \rangle$ : In general, the time average

$$\langle \sqrt{\frac{1}{\lambda(t)}} \rangle \neq \langle \frac{1}{\sqrt{\lambda(t)}} \rangle$$



so the magnitude of the price impact cannot be predicted exactly from the stationary solution for the depth.

76. Alternate Explanation for Concave Function: Some arguments for

$$\beta = \frac{1}{2}$$

involving a completely different mechanism have been offered by Zhang (1999).

77. Scaling of Liquidity and Average Spread: Ignoring the effects caused by finite bin size  $p_0$  and finite order size  $\sigma$ , the scaling behavior of the liquidity and the average spread can be derived from dimensional analysis.
78. Fundamental Dimensional Quantities: The fundamental dimensional quantities are shares, price, and time.
79. Fundamental Non-dimensional Quantities: In the continuum limit

$$p_0 \rightarrow 0$$

and

$$\sigma \rightarrow 0$$

these are uniquely represented by  $\alpha$ , with dimensions of  $\frac{\text{shares}}{\text{price} \times \text{time}}$ ,  $\mu$ , with dimensions of  $\frac{\text{shares}}{\text{time}}$ , and  $\delta$ , with dimensions of  $\frac{1}{\text{time}}$ .

80. Non dimensional Spread: The average spread has dimensions of *price* and is proportional to  $\frac{\mu}{\sigma}$ ; this comes from a balance between the total order placement rate inside the spread  $\alpha s$  and the order removal rate  $\mu$ .
81. Asymptotic Depth Definition: The asymptotic depth is the density of shares far away from the midpoint, where market orders are unimportant.



82. Non-dimensional Asymptotic Depth: It has dimensions of  $\frac{\text{shares}}{\text{price}}$  and is Poisson distributed with a mean  $\frac{\alpha}{\delta}$ .
83. Dimensional Analysis of Liquidity: The liquidity  $\lambda$  depends on the average slope of the depth profiles near the mid-point, and has the dimensions  $\frac{\text{shares}}{\text{price}^2}$ .
84. Non dimensional Liquidity: The liquidity is proportional to the ratio of the asymptotic depth to the spread, which implies that it scales as  $\frac{\alpha^2}{\mu\delta}$
85. Market Properties from Order Flow:

Quantity	Dimensions	Scaling relation
Asymptotic depth	$\text{shares}/\text{price}$	$d \sim \alpha/\delta$
Spread	$\text{price}$	$s \sim \mu/\alpha$
Slope of depth profile	$\text{shares}/\text{price}^2$	$\lambda \sim \alpha^2/\mu\delta = d/s$
Volatility ( $\tau \rightarrow 0$ )	$\text{price}^2/\text{time}$	$D_0 \sim \mu^2\delta/\alpha(\delta\sigma/\mu)^{-0.5}$
Volatility ( $\tau \rightarrow \infty$ )	$\text{price}^2/\text{time}$	$D_\infty \sim \mu^2\delta/\alpha(\delta\sigma/\mu)^{0.5}$

TABLE I: Predictions of scaling of market properties as a function of properties of order flow.  $\alpha$  is the limit order rate,  $\mu$  is the market order rate,  $\delta$  is the spontaneous limit order removal rate, and  $\sigma$  is the order size.

86. Impact of Discreteness on Dimensional Analysis: When discreteness is important, unique deviations from continuum dimensional analysis are lost and scaling can deviate from the predictions above.
87. Fundamental Discreteness Dimensional Quantities: In this case the behavior also depends on  $\sigma$ , which has the dimensions of *shares*, and  $p_0$ , which has the dimensions of *price*.
88. Non-dimensional Order Granularity and Tick: Effects due to granularity of orders depend on the non-dimensional order size

$$\hat{\sigma} = \frac{\sigma\delta}{\mu}$$



and effects due to finite tick size depend on the non-dimensional price

$$\hat{p} = \frac{p_0 \alpha}{\mu}$$

89. Continuum Limit for Tick Size: Simulation results show that there is a well-defined continuum limit with respect to the non-dimensional tick  $\frac{p_0 \alpha}{\mu}$ .
90. Continuum Limit for Order Granularity: The same is not true of non-dimensional granularity parameter  $\frac{\sigma \delta}{\mu}$  which does not affect the slope, spread, or asymptotic depth, but does affect price diffusion.
91. Non-dimensional Volatility: The price diffusion rate, or volatility, is a property of central interest. From continuum analysis, the volatility should scale as  $\frac{\delta \mu^2}{\sigma^2}$ .
92. Derivation of Non-dimensional Volatility: This occurs from squaring

$$\langle \Delta p(\omega) \rangle = \langle \sqrt{\frac{\omega}{\lambda(t)}} \rangle$$

before averaging and substituting

$$\omega = \mu$$

and

$$\lambda = \frac{\sigma^2}{\mu \delta}$$

However, this scaling is violated because discreteness is inherent to price diffusion.

93. Variance of Change in Mid-price:

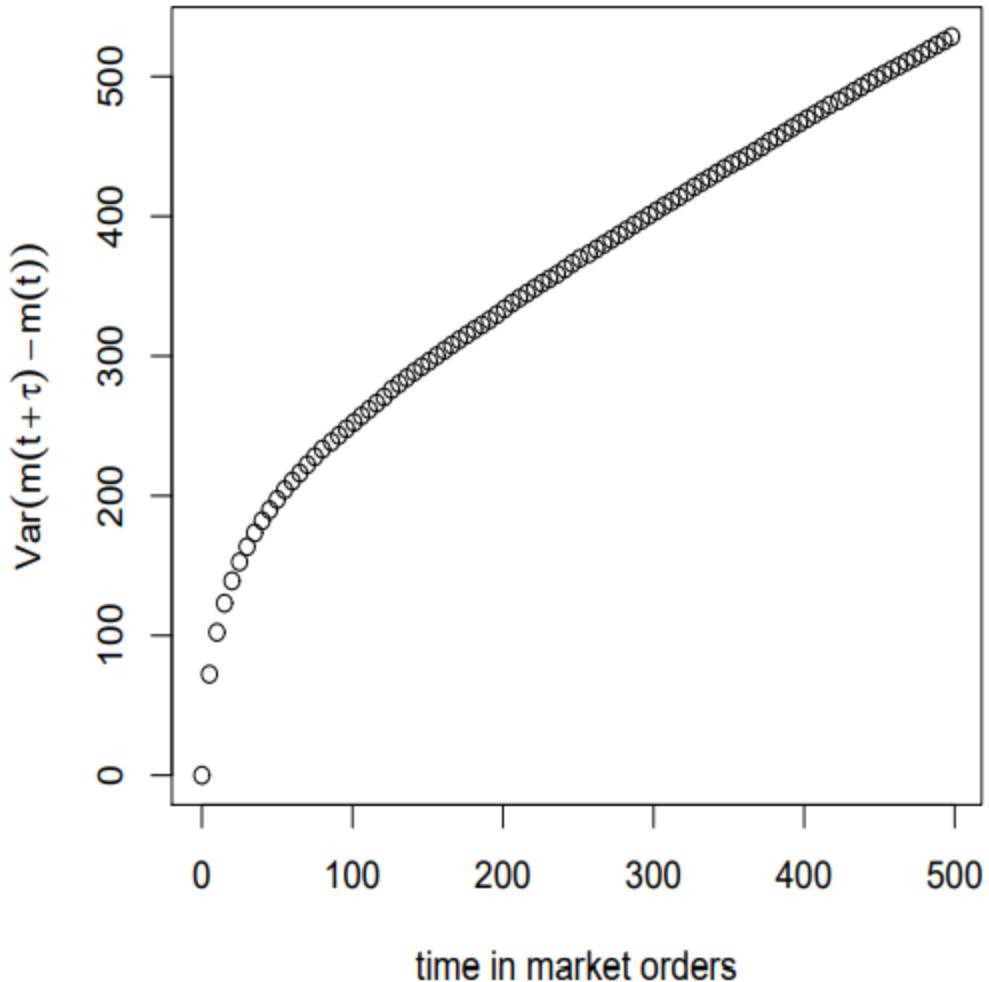


FIG. 3: The variance of the change in the midpoint price on timescale  $\tau$ . For a pure random walk this would be a line whose slope is the diffusion rate. The fact that the slope is steeper for short times comes from the nontrivial temporal persistence of the order book. Time is measured in terms of the characteristic interval between market orders,  $\sigma/\mu$ , with  $\alpha = 1$ ,  $\delta = 5 \times 10^{-4}$ ,  $\mu = 0.1$ , and  $\sigma = 1$ .

94. Price Variance over Time: The figure above plots the simulation results for the variance of the change in the mid-price at timescale  $\tau$   $V[m(t + \tau) - m(t)]$ . The slope is the diffusion rate.



95. Multiple Diffusion Timescales: It appears that there are at least two different timescales involved, with a faster diffusion rate for shorter timescales and a slower diffusion rate for long timescales. Such anomalous diffusion is not predicted by mid-field analysis.
96. Granularity Dependence on Diffusion: Simulation results show that the diffusion rate is correctly described by the product of the continuum diffusion rate  $\frac{\delta\mu^2}{\sigma^2}$  and a  $\tau$ -dependent power of the non-dimensional granularity parameter  $\frac{\delta\sigma}{\mu}$  as summarized in the Table above.
97. Explicit Form for Diffusion Granularity Dependence: Why this power is apparently  $-\frac{1}{2}$  for short term diffusion and  $\frac{1}{2}$  for long-term diffusion remains a mystery at this stage.
98. Autocorrelation between Mid-prices: Note that the temporal structure in the diffusion process also implies non-zero autocorrelations of  $m(t)$ . This corresponds to weak autocorrelations in price differences  $m(t) - m(t - 1)$  that persist for timescales until the variance vs  $\tau$  becomes a straight line.
99. Decay of Mid-price Autocorrelation: The timescale depends on the parameters, but is typically of the order of 50 market order arrival times.
100. Horizon Dependence of Market Impact: Another consequence of this is that the magnitude of the average price impact

$$\langle \Delta p \rangle = \langle \phi(\omega, \tau) \rangle$$

varies with the time horizon  $\tau$ . The simulations make it clear that the functional form is

$$\langle \phi(\omega, \tau) \rangle = f(\tau)\sqrt{\omega}$$

101. Temporal Market-Impact Coefficient: The function  $f(\tau)$  appears to decrease from its initial value, reaching a non-zero  $f(\infty)$  related to the asymptotic diffusion rate.



102. Price-Dependent  $\alpha$ : This model contains several unrealistic assumptions. For real markets, order flow rates vary in time, and  $\alpha(p)$  – in non-comoving coordinates – is not uniform.
103. Time-Dependent  $\delta$ : Of course, a reasonable expectation is that the order cancelation rate  $\delta$  depends on the time elapsed since the order placement.
104.  $\mu$  Dependence on Market Condition: Market participants place orders in response to varying market conditions, and real order flow processes are not unconditionally random.
105. Nature of Diffusion and Autocorrelation: Real markets display super-diffusive behavior of prices at short timescales, and autocorrelations of absolute price changes that decay as a power law, neither of which occur in this model.

## References

- Bachelier, L. (1900): [Theorie de la Speculation](#)
- Bollerslev, T., I. Domowitz, and J. Wang (1997): Order-flow and the Bid-ask Spread: An Empirical Probability Model of Screen-based Trading *Journal of Economic Dynamics and Control* **21** (8-9) 1471-1491
- Bouchaud, J. P., and R. Cont (2002): A Langevin Approach to Stock-market Fluctuations and Crashes *European Physical Journal B* **6** (4) 543-550
- Challet, D., and R. Stinchcombe (2001): Analyzing and Modeling 1+1d Markets *Physica A* **300** (1-2) 285-299
- Daniels, M. G., J. D. Farmer, G. Iori, and E. Smith (2001): *How storing Supply and Demand affects Price Diffusion* [arXiV](#)
- Domowitz, I., and J. Wang (1994): Auctions as Algorithms *Journal of Economic Dynamics and Control* **18** (1) 29-60
- Eliezer, D., and I. I. Kogan (1998): *Scaling Laws for the Market Microstructure of the Interdealer Broker Markets* [arXiV](#)
- Farmer, J. D. (1996): [Slippage March 1996](#)



- Farmer, J. D. (2002): Market Force, Ecology, and Evolution *Industrial and Corporate Change* **11 (5)** 895-953
- Gode, D. K., and S. Sunder (1993): Allocative Efficiency of Markets with Zero-intelligence Traders: Markets as a Partial Substitute for Individual Rationality *Journal of Political Economy* **101 (1)** 119-137
- Huasman, J. A., A. W. Lo, and A. C. MacKinlay (1992): An Ordered Probit Analysis of Transaction Stock Prices *Journal of Financial Economics* **31 (3)** 319-379
- Iori, G., and C. Chiarella (2002): [A Simple Microstructure Model of Double Auction Markets](#)
- Kempf, A., and O. Korn (1998): [Market Depth and Order Size](#)
- Maslov, S. (2000): Simple Model of a Limit-order Driven Market *Physica A* **278 (3-4)** 571-578
- Matassini, L., and F. Franci (2001): *How Traders enter the Market through the Book* arXiv
- Plerou, V., P. Gopikrishnan, X. Gabaix, and H. E. Stanley (2002): Quantifying Stock-price Response to Demand Fluctuations *Physical Review E* **66** 027104
- Slanina, F. (2001): Mean-field Approximation for a Limit-order Driven Market Model *Physical Review E* **64** 056136
- Torre, N. G., and M. Ferrari (1997): *The Market Impact Model* MSCI BARRA New York NY
- Zhang, Y. C. (1999): Toward a Theory of Marginally Efficient Markets *Physica A* **269 (1)** 30-44



## Statistical Theory of Continuous Double Auction

### Abstract

1. Continuous Double Auction Mechanism: Most modern financial markets use a continuous double auction mechanism to store and match orders and facilitate trading (Smith, Farmer, Gillemot, and Krishnamurthy (2003)).
2. Microscopic Dynamical Statistical Model: This develops a microscopic dynamic statistical model for the continuous double auction under the assumption of IID random order flow, and analyzes it using simulation, dimensional analysis, and theoretical tools based on mean-field approximations.
3. Teasable Predictions for Market Properties: The model makes testable predictions for basic properties of markets, such as price volatility, depth of stored supply vs. demand for price, the bid-ask spread, the [price-impact function, and the time and the probability of filling orders.
4. Order Flow and LOB Properties: These predictions are based on the order flow and the LOB, such as share volume of limit and market orders, cancellations, typical order sizes, and tick sizes.
5. Direct Measurement of Market Quantities: Because all quantities can be directly measured, there are no free parameters.
6. Order Size as Marker Driver: It is shown that the order size, which can be cast as the non-dimensional granularity parameter, is in most cases a more significant determinant of market behavior than tick size.
7. Concave Nature of Market Impact: It also provides an explanation for the observed highly concave nature of market impact functions.
8. Applicability of Zero-intelligence Models: On a broader level, this work suggests how stochastic models based on zero-intelligence agents may be useful to probe the structure of market institutions.



9. Assumptions Underlying the Predictions: Like the model of perfect rationality, the stochastic zero-intelligence models may be used to make strong predictions based on a compact set of assumptions, even if these assumptions are not fully believable.

## Introduction

1. Model Description and Context: This section provides background and motivation, a description of the model, and some historical context to work in this model.
2. Model Phenomenology and Dimensional Analysis: The next section gives an overview of the phenomenology of the model, explaining how dimensional analysis applies in this context, and presenting a summary of numerical results.
3. Analytical Treatment: The section after that develops an analytical treatment of the model, explaining some of the numerical findings of the previous section.
4. Model Enhancements and Approach Comparison: The final section concludes with a discussion of how the model may be enhanced to bring it closer to real-life markets, and some comments comparing the approach taken here to standard models based on information arrival and valuation.

## Introduction – Motivation

1. Importance of Financial Institutions in Setting Prices: The model in this chapter demonstrates the importance of financial institutions in setting prices, and how solving a necessary economic function such as providing liquidity can have unanticipated side-effects.
2. Demand Storage Causes Persistence: In a world of imperfect rationality and imperfect information, the task of demand storage necessarily causes persistence.



3. Imperfect Rationality in Real Markets: Under perfect rationality, all traders would instantly update their orders with the arrival of each piece of new information, but this is clearly not true for real markets.
4. The LOB: The LOB, which is used for storing unexecuted orders, has long memory when these are persistent orders.
5. IID Random Order Flow: It is shown here that even under completely random IID order flow, the price process displays anomalous diffusion and interesting temporal structure.
6. Random Price Process: The converse is also interesting; for prices to be effectively random, incoming order flow must be non-random, in just the right way to compensate for the persistence.
7. Alternative Approach to doing Economics: The work is also of interest from a fundamental point of view because it suggests an alternative approach to doing economics.
8. Assumption of Perfect Rationality: The assumption of perfect rationality has been popular in economics because it provides a parsimonious model to make strong predictions.
9. Zero-intelligence Random Behavior: In the spirit of Gode and Sunder (1993), it is shown that the opposite extreme of zero-intelligence random behavior also provides another reference model that makes strong predictions.
10. Zero intelligence as a Simplification: Like perfect rationality, zero-intelligence is an extreme simplification that is obviously not literally true. But as shown here, it provides a useful tool for probing the behavior of financial institutions.
11. Rational, Bounded Behaviors: The resulting model may be easily extended by introducing simple, boundedly rational behaviors.
12. Fundamental Assumptions about Utility: It also differs from standard treatments in that it does not attempt to understand prices from fundamental assumptions about utility.
13. Price Dependence on Order Flow: Rather, it focuses on how prices depend on order flow rates, leaving the problem of what determines these rates for the future.



14. Approaches used in this Chapter: The approach makes extensive use of dimensional analysis, the solution to a master equation through a generating functional, and a mean-field approach that is commonly used to analyze non-equilibrium reaction-diffusion systems and evaporation-deposition problems.

## Introduction – Background: The Continuous Double Auction

1. Definition of a “Quote”: Patient traders submit *limit orders*, or *quotes* which also state a linear price, corresponding to the worst allowable price for the transaction.
2. Inside Quotes: The logarithmic price  $a(t)$  is used to denote the position of the best/lowest offer and  $b(t)$  for the position best/highest bid. These are also called the *inside quotes*.
3. Spread: There is typically a non-zero price gap between them, called the *spread*

$$s(t) = a(t) - b(t)$$

4. Ticks: Prices are not continuous, but rather have discrete quanta called *ticks*.
5. Log of the Price: Throughout this chapter, all prices will be expressed as logarithmic, and the word *price* will be used to mean the log of the price.
6. Tick Size: The minimum interval prices change is on the *tick size*  $\varepsilon_p$  – also defined on a logarithmic scale; note that this is not true for real markets. Further,  $\varepsilon_p$  is *not* necessarily infinitesimal.
7. Clearing of the Market Orders: As market orders arrive, they are matched against the limit orders of the opposite sign in the order of price and then arrival time, as shown in the figure below.
8. Schematic Illustration of the Continuous Double Auction Mechanism:

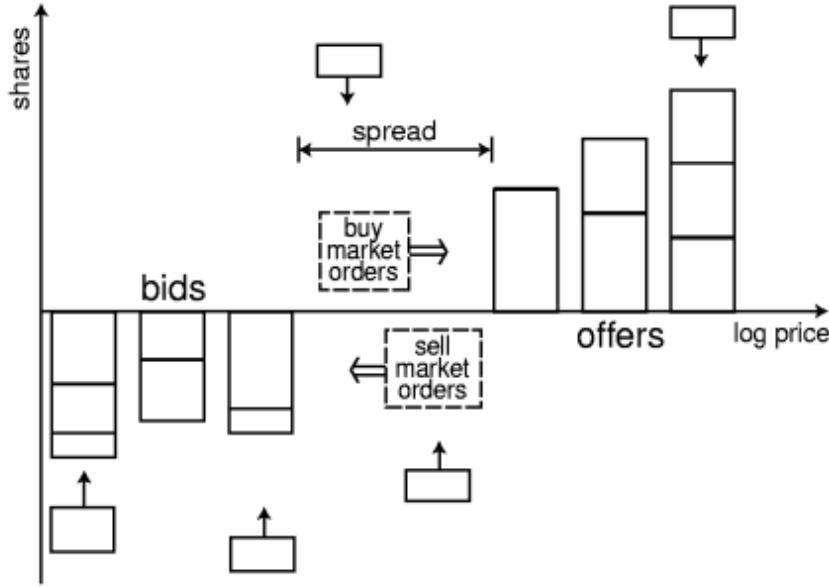


FIG. 1: A schematic illustration of the continuous double auction mechanism and our model of it. Limit orders are stored in the limit order book. We adopt the arbitrary convention that buy orders are negative and sell orders are positive. As a market order arrives, it has transactions with limit orders of the opposite sign, in order of price (first) and time of arrival (second). The best quotes at prices  $a(t)$  or  $b(t)$  move whenever an incoming market order has sufficient size to fully deplete the stored volume at  $a(t)$  or  $b(t)$ . Our model assumes that market order arrival, limit order arrival, and limit order cancellation follow a Poisson process. New offers (sell limit orders) can be placed at any price greater than the best bid, and are shown here as “raining down” on the price axis. Similarly, new bids (buy limit orders) can be placed at any price less than the best offer. Bids and offers that fall inside the spread become the new best bids and offers. All prices in this model are logarithmic.

9. Depth Profile at Price  $p$ : Let  $n(p, t)$  be the stored density of the limit order volume at price  $p$ , which is called the *depth profile* of the LOB at any time  $t$ .



10. Total Price at Price Level: The total stored limit order volume at a price level  $p$  is  $n(p, t)\varepsilon_p$ .
11. Ask Prices after Executing Market Orders: For unit order size, the shift in the best ask  $a(t)$  produced by a buy market order is given by solving the equation

$$\sum_{p=a(t)}^{p'} n(p, t)\varepsilon_p$$

for  $p'$ .

12. Buy Market Impact: The shift in the best-ask  $p' - a(t)$  is the instantaneous price impact for buy market orders.
13. Sell Market Impact: A similar statement applies for sell market orders, where the price impact can be defined in terms of the shift in the best bid.
14. Impact in Terms of the Mid-price Shift: Alternatively, it is also possible to define the price impact in terms of the change in the midpoint price.
15. Crossing or Marketable Limit Order: The buy limit order whose limit price is greater than the best ask, or the sell limit order whose limit price less than the best bid, is referred to as a *crossing limit order* or a *marketable limit order*.
16. Immediate Execution of Marketable Limit Orders: Such limit orders result in immediate transactions, with at least part of the order immediately executed.

## Introduction – The Model

1. Buy/Sell Market Order Arrival Rate: The rate at which buy orders or sell orders arrive individually is  $\frac{\mu}{2}$ .
2. Implications of an Infinite Interval: While the assumption of limit order placement over an infinite interval is clearly unrealistic, it provides a tractable boundary condition for modeling the behavior of the LOB near the midpoint price



$$m(t) = \frac{a(t) + b(t)}{2}$$

which is the region of interest since it is where the transactions occur.

3. Poisson Nature far from Midpoint: Limit orders far from midpoint are usually canceled before they are executed – as demonstrated later – and so far from the midpoint, limit order arrivals and cancelations have a steady state distribution characterized by a simple Poisson distribution.
4. Order Placement per Unit Time: Although under the limit order placement process the total number of orders placed per unit time is infinite, the order placement per unit price interval is bounded and thus the assumption of an infinite interval creates no problem.
5. Order Book never Empties: Indeed, it guarantees that there are always an infinite number of limit orders of both signs stored in the book, so that the bid and the ask are always well-defined and the book never empties.
6. Alternate LOB Placement Assumptions: Under other assumptions about limit order placement, this is not necessarily true, as will be shown later.
7. More Realistic Order Placement Functions: Also considered in a later section are more realistic order placement functions.
8. Effective Market/Limit Orders: In this model, to keep things simple, the conceptual simplification of *effective market orders* or *effective limit orders* is used.
9. Crossing Limit Order Component: When a crossing limit order is placed, part of it may be executed immediately. The effect of this part on price is indistinguishable from that of the market order of the same size.
10. Non-crossing Limit Order Component: Similarly, given that this market order has been placed, the remaining order is equivalent to a non-crossing limit order of the same size.
11. Decomposing a Crossing Limit Order: Thus, a crossing limit-order can be modeled as an effective market order followed by an effective non-crossing limit order.



12. Correlation between Market and Crossing Limit Components: In assigning independently random distributions for the two events, the market neglects the correlation between the market order and the limit order arrivals induced by crossing limit orders.
13. Effective Market/Limit Order Arrival Rate: Working in terms of effective market and limit orders effects the analysis: the effective market order arrival rate  $\mu$  contains both pure market orders and the immediately executed limit orders, and similarly the limit order arrival rate  $\alpha$  corresponds only to the components of the limit orders that are not executed immediately.
14. Order Placement Boundary Conditions: This is consistent with the boundary conditions for the order placement process, since an offer with

$$p \leq a(t)$$

or a bid with

$$p \geq b(t)$$

would result in an immediate transaction, and thus would be effectively same as a market order.

15. Limit Orders inside the Spread: Defining the order placement processes with these boundary conditions realistically allows limit orders to be placed anywhere within the spread.
16. Use of Logarithmic Prices: Another simplification of this model is to use logarithmic prices, both for order placement and the tick size  $\varepsilon_p$ . This has the important advantage that the tick prices are always positive.
17. Consistency across Calculations and Simulations: In real markets price ticks are linear, and the use of logarithmic price ticks is an approximation that makes both the calculation and the simulation more convenient.
18. Limit of  $\varepsilon_p \rightarrow 0$ : The limit



$$\varepsilon_p \rightarrow 0$$

where the tick size is irrelevant, is a good approximation for many cases.

19. Importance of Tick Size Choice: It is found that the tick size is less important than the other parameters of the problem, which provides some justification for the approximation of the logarithmic price ticks.
20. Constant Probability for Cancellation: Assuming a constant probability for cancellation is clearly *ad hoc*, but in simulations it is found that other assumptions such as well-defined timescales, such as constant duration time, give similar results.
21. Simulations using Variable Order Sizes: For the analytic models here a constant order size  $\sigma$  is used. The simulations also use variable order sizes, e.g., half-normal distributions with standard deviation  $\sqrt{\frac{\pi}{2}}\sigma$ , which ensures that the mean value remains  $\sigma$ .
22. Distributions with Thin Tails: As long as these distributions have thin tails, the differences do not qualitatively affect most of the results reported here, except in a trivial way.
23. Improperly Defined Characteristic Decay Times/Sizes: As discussed later, decay processes without well-defined characteristic size and time distributions with power law tails give qualitatively different results and will be discussed elsewhere.
24. Complexity in the Dynamics: Even though the model is simply defined, the time evolution is not trivial.
25. Phenomenological Breakdown of LOB Dynamics: One can think of the dynamics as being composed of three parts:
  - a. The buy market order/sell limit order interaction, which determines the best ask.
  - b. The sell market order/buy limit order interaction, which determines the best bid.
  - c. The random cancellation process.



26. Determining the Cross Boundary Conditions: Process a and b determine each other's boundary condition. That is, a determines the best ask, which sets the boundary condition in the limit order placement process in b, and b determines the best bid, which determines the boundary condition for limit order placement in a.
27. Strong Coupling of a and b: Thus, processes a and b are strongly coupled. It is this coupling that causes the bid and the ask to remain close to each other, and guarantees that the spread

$$s(t) = a(t) - b(t)$$

is a stationary variable, even though the bid and the ask are not.

28. The Resulting LOB Dynamics Complexity: It is the coupling of these processes through their boundary conditions that provides the non-linear feedback that makes the price process complex.

## Introduction – Summary of Prior Work

1. Earlier Approaches: There are two independent lines of prior work, one in the financial economics literature, and the other in the physics literature.
2. Economics Literature: The models in the economic literature are directed towards empirical analysis, and treat the order process as static.
3. Physics Literature: In contrast, the models in the physics literature are conceptual toy models, but they allow the order process to react to changes in prices, and thus are fully dynamic.
4. Details of the Models: The chapter bridges this gap. More details on the literature are as follows.
5. Mendelson (1982): The first model of this type appears to be due to Mendelson (1982), who modeled random order placement with periodic clearing.



6. Cohen, Conroy, and Maier (1985): This was developed along different directions by Cohen, Conroy, and Maier (1985), who use techniques from the queuing theory, but assumed only one price level and addresses the issue of time priority at that level – motivated by the existence of a specialist who pinned prices to make them effectively stationary.
7. Domowitz and Wang (1994), Bollerslev, Domowitz, and Wang (1997): Domowitz and Wang (1994) and Bollerslev, Domowitz, and Wang (1997) further develop this to allow more general order placement process that depend on prices, but without solving the full dynamical problem. This allows them to get a stationary solution for the price.
8. Origin of the Price's Randomness: In contrast, in the prices in this chapter, the prices that emerge make a random walk, and so are much more realistic.
9. Co-moving Price Coordinates: In order to get the solution to the depth of the order book, one needs to go into the price coordinates that co-moves with the random walk.
10. Prices/Order Placement Interplay: Dealing with the interaction between prices and the order placement makes the problem much more difficult, but is key to getting reasonable results.
11. Models in the Physics Literature: The models in the physics literature incorporate price dynamics, but have tended to be conceptual toy models designed to understand the anomalous properties of price diffusion.
12. Earlier Work on Physics Modeling: This line of work begins with Bak, Paczuski, and Shubik (1997), and was further developed by Eliezer and Kogan (1998) and Tang and Tian (1999).
13. Hypothesis on Price Diffusion: They assume that the limit orders are placed at a fixed distance from the midpoint, and then the limit prices of these orders are randomly shuffled until they result in transactions. It is this random shuffling that causes price diffusion.
14. Use of Reaction-Diffusion Model: This unrealistic assumption was made to take advantage of the analogy to the standard reaction-diffusion model in the physics literature.



15. Maslov (2000) and Slanina (2001): Maslov (2000) introduced an alternative model that was solved analytically in the mean-field limit by Slanina (2001).
16. Random Buy/Sell Market/Limit Orders: Each order is randomly chosen to be a buy or a sell, and either a limit order or a market order.
17. Price Placement of Limit Orders: If it is a limit order, it is randomly placed at a fixed distance from the current price; this again gives rise to anomalous price diffusion.
18. Challet and Stinchcombe (2001): A model allowing limit orders with Poisson cancellation was proposed by Challet and Stinchcombe (2001).
19. Iori and Chiarella (2002): Iori and Chiarella (2001) have numerically studied a model including fundamentalists and technical traders.
20. Daniels, Farmer, Iori, and Smith (2001): The model studied in this chapter was introduced by Daniels, Farmer, Iori, and Smith (2001). It treats the feedback between order placement and price movement, while having enough realism so that the parameters can be tested against real data.
21. Earlier focus on Anomalous Diffusion: The previous models in the physics literature have tended to focus primarily on the anomalous diffusion of prices. While interesting and important for refining risk calculations, this is a second-order effect.
22. Current Behaviors of Interest: In contrast, this chapter focuses on the first order effects of primary interest to market participants, such as bid-ask spread, volatility, price impact, depth profile, and the probability and time to fill an order.
23. Dimensional Analysis and Mean field Formulation: It also demonstrates how dimensional analysis becomes a useful tool in an economic setting, and develops mean-field theories in a context that is more challenging than that of the previous works.
24. Bouchaud, Mezard, and Potters (2002): Subsequent to Daniels, Farmer, Iori, and Smith (2001), Bouchaud, Mezard, and Potters (2002) demonstrated that, under the assumption that prices execute a random walk, and by introducing an additional free parameter, they can derive a simple equation for the depth profile.
25. Enhancement in this Chapter: This chapter shows how to do this from first principles without introducing a free parameter.



## Overview of the Predictions of the Model

1. The Model Parameters: This section gives an overview of the phenomenology of the model. Because this model has 5 parameters, understanding all their effects would generally be a complicated task in itself.
2. Reduction to Two Parameters: This task is generally simplified by the use of dimensional analysis, which reduces the number of independent parameters from 5 to 2.
3. Approach behind Dimensional Analysis: Thus, before the results can be reviewed, one needs to explain first explain how dimensional analysis applies in this setting. One of the surprising aspects of this model is that one can derive several powerful results using the simple technique of dimensional analysis alone.
4. Analysis of the Simulations: Unless otherwise mentioned, results in this section are based on simulations. A later section compares them to theoretical predictions.

## Overview of the Model Predictions – Dimensional Analysis

1. Review of Dimensional Analysis: Because dimensional analysis is not commonly used in economics, a review is presented first. Bridgeman (1922) contains more details.
2. What is Dimensional Analysis: Dimensional analysis is a technique that is commonly used in physics and engineering to reduce the number of independent degrees of freedom by taking advantage of the constraints improved by dimensionality.
3. Usage in Constrained Settings: For sufficiently constrained settings, it can be used as a starting point for a solution before a full analysis.



4. Factors Underpinning the Driving Phenomenon: The idea is to write down all the factors that a given phenomenon depends on, and then find the combination that has the correct dimensions.
5. Example - Period of a Pendulum: As an example, consider the problem of estimating the period of the pendulum. The period  $T$  has dimensions of *time*.
6. Mass Length, and Gravity: Obvious candidates that  $T$  might depend on are the mass  $m$  of the bob – which has the unit *mass*; the length  $l$  – Which has the unit of *distance*, and the acceleration of gravity  $g$  – which has the unit  $\frac{\text{distance}}{\text{time}^2}$
7. Combination to Produce Time: There is only one way to combine these to produce dimensions of *time*, i.e.

$$T \sim \sqrt{\frac{l}{g}}$$

This determines the correct formula for the period of a pendulum up to a constant.

8. Independence from Mass: Note that it makes it clear that the period does not depend on mass, a result that is not obvious *a priori*.
9. Reduction in Free Parameter Count: This problem has 3 parameters and 3 dimensions, and the unique combinations of the parameters contain the right dimensions; in general, dimensional analysis can only be reduce the number of free parameters through the constraint imposed by the dimensions.
10. Fundamental Model Dimensions: For the LOB problem, the fundamental model dimensions are *shares*, *price*, and *time*.
11. Logarithm of Price: Note that *price* means logarithm of price; as long as it is consistent, it does not create problems with the dimensional analysis.
12. Three Rates and Two Discreteness Parameters: There are five parameters, three rates and two discreteness parameters.
13. Dimensions of the Flow Parameters: The *order flow rates* are  $\mu$ , the market order arrival rate, with dimensions of *shares per time*;  $\alpha$ , the limit order arrival rate per



price, with the dimensions of *shares per price per time*, and  $\delta$ , the rate of limit order decays, with dimensions of  $\frac{1}{time}$ . These play roles similar to constants in physical problems.

14. Dimensions of the Discreteness Parameters: The two discreteness parameters are the price tick size  $\varepsilon_p$ , with dimensions of *price*, and the order size  $\sigma$ , with dimensions of *shares*.

15. Parameters Characterizing the Model:

<b>Parameter</b>	<b>Description</b>	<b>Dimensions</b>
$\alpha$	limit order rate	<i>shares/(price time)</i>
$\mu$	market order rate	<i>shares/time</i>
$\delta$	order cancellation rate	$1/time$
$dp$	tick size	<i>price</i>
$\sigma$	characteristic order size	<i>shares</i>

TABLE I: The five parameters that characterize this model.  $\alpha$ ,  $\mu$ , and  $\delta$  are order flow rates, and  $dp$  and  $\sigma$  are discreteness parameters.

16. Process Parameters Exceeding the Dimensions: Because there are five parameters and three dimensions – *price*, *shares*, and *time* – and because in this case the dimensionality of the parameters is sufficiently rich, the dimensional relationships reduce the degrees of freedom, so that all properties of the LOB can be described by the two parameters.
17. Constructing the Analysis Parameters: It is useful to construct these two parameters so that they are non-dimensional.
18. Dimensional Reduction of the Model: A dimensional reduction of the model is performed by guessing that the effect of the order flow rates is primary to that of the discreteness parameters.
19. Non dimensionalization Based on Order Flow: This leads to the construction of non-dimensional units based on the order flow alone, and take the nondimensional version



of the discreteness parameters as independent parameters whose effects need to be understood.

20. Weak Dependence on Discreteness Properties: As will be seen, this is justified by the fact that many of the properties of the model can thus be understood based on dimensional analysis alone.
21. Phenomenon Examination Using Dimensional Analysis: Much of the richness of the phenomenology can thus be understood based on dimensional analysis alone.
22. Unique Combination using Order Flow: There are three order flow rates and three fundamental dimensions. If one temporarily ignores the discreteness parameters, there are unique combinations of order flow rates with units of shares, price, and time.
23.  $N_c$ ,  $p_c$ , and  $t_c$ : These define a characteristic number of shares

$$N_c = \frac{\mu}{2\delta}$$

a characteristic price interval

$$p_c = \frac{\mu}{2\alpha}$$

and a characteristic timescale

$$t_c = \frac{1}{\delta}$$

24. Important Characteristic Scales and Nondimensional Quantities:



Parameter	Description	Expression
$N_c$	characteristic number of shares	$\mu/2\delta$
$p_c$	characteristic price interval	$\mu/2\alpha$
$t_c$	characteristic time	$1/\delta$
$dp/p_c$	nondimensional tick size	$2\alpha dp/\mu$
$\epsilon$	nondimensional order size	$2\delta\sigma/\mu$

TABLE II: Important characteristic scales and nondimensional quantities. We summarize the characteristic share size, price and times defined by the order flow rates, as well as the two nondimensional scale parameters  $dp/p_c$  and  $\epsilon$  that characterize the effect of finite tick size and order size. Dimensional analysis makes it clear that all the properties of the limit order book can be characterized in terms of functions of these two parameters.

25. Factor of 2 in  $\mu$ : The factor of 2 has occurred because one has defined the market order rate for either a buy or a sell to be  $\frac{\mu}{2}$ .

26. Nondimensional Units as Basis: One can thus express everything in the model in non-dimensional terms by dividing  $N_c$ ,  $p_c$ , or  $t_c$  as appropriate, e.g., to measure shares in non-dimensional units

$$\hat{N} = \frac{N}{N_c}$$

or to measure the price in nondimensional units

$$\hat{p} = \frac{p}{p_c}$$

27. Usefulness of Nondimensional Units:

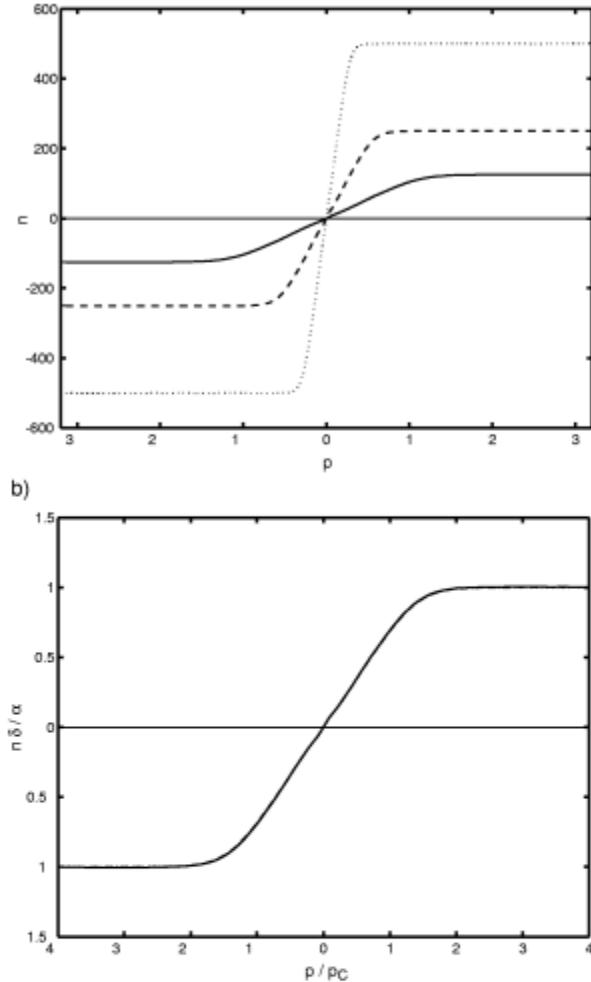


FIG. 2: The usefulness of nondimensional units. (a) We show the average depth profile for three different parameter sets. The parameters  $\alpha = 0.5$ ,  $\sigma = 1$ , and  $dp = 0$  are held constant, while  $\delta$  and  $\mu$  are varied. The line types are: (dotted)  $\delta = 0.001$ ,  $\mu = 0.2$ ; (dashed)  $\delta = 0.002$ ,  $\mu = 0.4$  and (solid)  $\delta = 0.004$ ,  $\mu = 0.8$ . (b) is the same, but plotted in nondimensional units. The horizontal axis has units of *price*, and so has nondimensional units  $\hat{p} = p/pc = 2\alpha p/\mu$ . The vertical axis has units of *n shares/price*, and so has nondimensional units  $\hat{n} = np_c/N_c = n\delta/\alpha$ . Because we have chosen the parameters to keep the nondimensional order size  $\epsilon$  constant, the collapse is perfect. Varying the tick size has little effect on the results other than making them discrete.



28. Depth Profile for  $\frac{\mu}{\delta}$  Variations: The value of using nondimensional units is illustrated in the figure above. Part (a) of the figure shows the average depth profile for three different values of  $\mu$  and  $\delta$  with the other parameters held fixed.
29. Plot using Nondimensional Units: On plotting these results in dimensional units, the results look quite different. However, plotting them in nondimensional units as in (b), the results are indistinguishable.
30. Invariance due to Fixed, Nondimensional Order Size: As explained below, because the nondimensional order size has been kept constant, the collapse is perfect.
31. Profile Plots for Tick/Order Sizes: Thus, the problem of understanding the behavior of the model is reduced to studying the effect of tick size and order size.
32. Nondimensional Tick Size: To understand the effect of tick size, and order size, it is useful to do them in nondimensional terms. The nondimensional scale parameter based on tick size is constructed by scaling the characteristic price, i.e.

$$\frac{\varepsilon_p}{p_c} = \frac{2\alpha\varepsilon_p}{\mu}$$

33. Continuum around  $\varepsilon_p \rightarrow 0$ : The theoretical analysis and the simulations show that there is a sensible continuum limit as the tick size

$$\varepsilon_p \rightarrow 0$$

in the sense that there is a non-zero price diffusion and a finite spread.

34. Weak Dependence on Tick Size: Furthermore, the dependence on the tick size is weak, and for many purposes, the limit

$$\varepsilon_p \rightarrow 0$$

approximates the finite tick size very well.



35.  $\varepsilon_p \rightarrow 0$  for Analytic Tractability: As will be seen, working in this limit is essential for getting tractable analytical results.

36. Order Size Nondimensional Scale Parameters: A nondimensional scale parameter based on order size is constructed by scaling the typical order size – measured in shares – by the characteristic number of shares, i.e.

$$\epsilon = \frac{\sigma}{N_c} = \frac{2\delta\sigma}{\mu}$$

37.  $\epsilon$  as Order Quanta Indicator:  $\epsilon$  characterizes the chunkiness of the order stored in the LOB.

38. Importance of  $\epsilon$  for Liquidity/Volatility: As will be seen,  $\epsilon$  is an important determinant of liquidity, and a particularly important determinant of volatility.

39. No Diffusion as  $\epsilon \rightarrow 0$ : In the continuum limit

$$\epsilon \rightarrow 0$$

there is no price diffusion. This is because price diffusion can only occur if there is a finite probability for price levels outside the spread to be non-empty, thus allowing the best bid or the best ask to make a persistent shift.

40. Consequence of  $\epsilon \rightarrow 0$ : If one lets

$$\epsilon \rightarrow 0$$

as the average depth is held fixed, the number of individual orders becomes infinite, and the probability that spontaneous decays or market orders can create gaps becomes zero.

41. Ineffectiveness of the  $\epsilon \rightarrow 0$  Approximation: This is verified in simulations, Thus, the limit



$$\epsilon \rightarrow 0$$

is always a poor approximation to a real market.

42. Letting  $\frac{\varepsilon_p}{p_c} \rightarrow 0$ :  $\epsilon$  is a more important parameter than the tick size  $\frac{\varepsilon_p}{p_c}$ . In the mean-field analysis later, one lets

$$\frac{\varepsilon_p}{p_c} \rightarrow 0$$

thereby reducing the number of independent parameters from 2 to 1. One finds that in many ways this is a good approximation.

43. Scale for Order Size:  $N_c$  provides the scale against which the order size is measured,  $\epsilon$  characterizes the granularity in relative terms.
44. Annihilation Rate in Market Orders: Alternatively,  $\frac{1}{\epsilon}$  can be thought of as the annihilation rate from market orders expressed in units of the size of spontaneous decays.
45. Nondimensionalization of Share Count: Note that in nondimensional units, the number of shares can be written as

$$\hat{N} = \frac{N}{N_c} = \frac{N\epsilon}{\sigma}$$

46. Independence of Spontaneous Decay Process: The construction of the nondimensional granularity parameter illustrates the importance of including a spontaneous decay process in the model.

47. Impact of Zero Spontaneous Decay: If

$$\delta = 0$$

implying



$$\epsilon = 0$$

there is no spontaneous decay of orders, and depending on the relative value of  $\mu$  and  $\alpha$ , generally either the depth of the orders will accumulate without bound or the spread will become infinite. As long as

$$\delta > 0$$

in contrast, this is not a problem.

48. Low Tick Size/Order Size Impact: For certain purposes, the effects of varying tick sizes and order sizes are fairly small, and one can derive approximate formulas based only on order flow rates using dimensional analysis.
49. Order Flow Dimensional Analysis Estimates:

Quantity	Dimensions	Scaling relation
Asymptotic depth	$\text{shares}/\text{price}$	$d \sim \alpha/\delta$
Spread	$\text{price}$	$s \sim \mu/\alpha$
Slope of depth profile	$\text{shares}/\text{price}^2$	$\lambda \sim \alpha^2/\mu\delta = d/s$
Price diffusion rate	$\text{price}^2/\text{time}$	$D_0 \sim \mu^2\delta/\alpha^2$

TABLE III: Estimates from dimensional analysis for the scaling of a few market properties based on order flow rates alone.  $\alpha$  is the limit order density rate,  $\mu$  is the market order rate, and  $\delta$  is the spontaneous limit order removal rate. These estimates are constructed by taking the combinations of these three rates that have the proper units. They neglect the dependence on the order granularity  $\epsilon$  and the nondimensional tick size  $dp/p_c$ . More accurate relations from simulation and theory are given in table IV.

50. Dimensional Analysis from Order Flow: For example, the above table provides the dimensional scaling formulas for the average spread, market order liquidity – as



measured by the average slope of the depth profile near the midpoint – the volatility, and the asymptotic depth.

51. Ignoring the Impact of Discreteness: Because these estimates neglect the effect of discreteness, they are only approximations to the true behavior of the model, which do a better job of explaining some properties more than the others.
52. Dependence on  $\epsilon$  and  $\frac{\epsilon_p}{p_c}$ : The numerical and the analytical results show that some quantities also depend on the granularity parameters  $\epsilon$  and to a weaker extent on the tick size  $\frac{\epsilon_p}{p_c}$ .
53. Flow Scaling as a Starting Point: Nonetheless, dimensionless estimates based on order flow alone provide a good starting point for understanding the market behavior.
54. Granularity Adjustments: A comparison to more precise formulas derived from theory and simulations is given below.
55. Market Properties Dependence on Model Parameters:



Quantity	Scaling relation	Figure
Asymptotic depth	$d = \alpha/\delta$	3
Spread	$s = (\mu/\alpha)f(\epsilon, dp/p_c)$	10, 24
Slope of depth profile	$\lambda = (\alpha^2/\mu\delta)g(\epsilon, dp/p_c)$	3, 20 - 21
Price diffusion ( $\tau \rightarrow 0$ )	$D_0 = (\mu^2\delta/\alpha^2)\epsilon^{-0.5}$	11, 14(c)
Price diffusion ( $\tau \rightarrow \infty$ )	$D_\infty = (\mu^2\delta/\alpha^2)\epsilon^{0.5}$	11, 14(c)

TABLE IV: The dependence of market properties on model parameters based on simulation and theory, with the relevant figure numbers. These formulas include corrections for order granularity  $\epsilon$  and finite tick size  $dp/p_c$ . The formula for asymptotic depth from dimensional analysis in table III is exact with zero tick size. The expression for the mean spread is modified by a function of  $\epsilon$  and  $dp/p_c$ , though the dependence on them is fairly weak. For the liquidity  $\lambda$ , corresponding to the slope of the depth profile near the origin, the dimensional estimate must be modified because the depth profile is no longer linear (mainly depending on  $\epsilon$ ) and so the slope depends on price. The formulas for the volatility are empirical estimates from simulations. The dimensional estimate for the volatility from Table III is modified by a factor of  $\epsilon^{-0.5}$  for the early time price diffusion rate and a factor of  $\epsilon^{0.5}$  for the late time price diffusion rate.

56. Approximate Formula for the Mean Spread: An approximate formula for the mean spread can be derived by noting that it has the dimensions of *price*, and the unique combination of order flow rates with these dimensions is  $\frac{\mu}{\alpha}$ .
57. Scaling Factor for Spread: While dimensions indicate scaling of spread, they cannot determine multiplicative factors order unity.
58. Order Removal inside the Spread: A more intuitive argument can be made by noting that, inside the spread, removal due to cancellations is dominated by removal due to market orders.



59. Factor of  $\frac{1}{2}$ : Thus, the total limit order placement inside the spread, for either the buy or the sell limit orders,  $\alpha s$ , must equal the removal rate  $\frac{\mu}{2}$ , which implies that the spread is

$$s = \frac{\mu}{2\alpha}$$

60. Generalization of the above Approach: As will be seen later, this argument will be generalized and made more precise within the mean-field analysis which then also predicts the observed dependence on the granularity parameter  $\epsilon$ .

61. Weak Spread Dependence on  $\epsilon$ : However, this dependence is rather weak and only causes a variation of a factor of 2 for

$$\epsilon < 1$$

and a factor of  $\frac{1}{2}$  derived above is a good first approximation.

62. Predicted Mean Spread: Note that the predicted mean spread is just the characteristic price  $p_c$ .

63. Mean Asymptotic Depth: It is also easy to derive the *mean asymptotic depth*, which is the density of shares far away from the midpoint.

64. Motivation behind the Asymptotic Depth: The asymptotic depth is an artificial construct of the assumption of the order placement on an infinite interval. It should be regarded as providing a boundary condition for the behavior near the midpoint.

65. Expression for the Asymptotic Depth: The mean asymptotic depth has the dimensions of  $\frac{\text{shares}}{\text{price}}$  and is therefore given by  $\frac{\alpha}{\delta}$ .

66. Relation between  $\alpha$  and  $\delta$ : Furthermore, because removal by market orders is insignificant in this regime, it is determined by the balance between order placement and decay, and far from the midpoint the depth at any given price is Poisson distributed. The result is exact.



67. Depth Slope near the Mid: The average shape of the depth slope near the midpoint is an important determinant of the liquidity, since it affects the expected price response when a market order arrives.
68. Expression for the Depth Slope: The slope has dimensions of  $\frac{\text{shares}}{\text{price}^2}$  which implies that, in terms of order flow rates, it scales roughly as  $\frac{\alpha^2}{\mu\delta}$ .
69. Ratio of Asymptotic Depth to Spread: The above expression is also the ratio of asymptotic depth to the spread.
70. Behavior around  $\epsilon \sim 0.01$ : As will be seen later, this is a good approximation when

$$\epsilon \sim 0.01$$

but for smaller values of  $\epsilon$  the depth profile is not linear near the midpoint, and this approximation fails.

71. Empirical Estimates for Price Diffusion: The last two entries in the previous table are the empirical estimates for the price diffusion rate  $D$ , which is proportional to the square of the volatility.
72. Linear Price Variance in Time: That is, for normal distribution, starting from

$$t = 0$$

the variance  $v$  after time  $t$  is

$$v = Dt$$

The volatility at any given timescale is the square root of the variance at timescale  $t$ .

73. Diffusion Rate Estimate: The estimate for the diffusion rate based on dimensional analysis in terms of the order flow alone is  $\frac{\mu^2\delta}{\alpha^2}$



74. Short- vs Long-Term Diffusion: However, simulations show that short-term diffusion is much faster than long-term diffusion, due to negative autocorrelations in the price process.
75. Initial vs Asymptotic Diffusion Rates: The initial and the asymptotic diffusion rates appear to obey the scaling relationships given on the table above.
76. Explaining the Diffusion Rate Differences: Though the mean-field theory is not able to predict this functional form, the fact that the early and the late term diffusion rates are different can be understood within the framework of the analysis, as described later.
77. Negative Autocorrelations in Mid-price: Anomalous diffusion of this type implies negative autocorrelations at midpoint prices.
78. Anomalous Diffusion: Note that the term *anomalous diffusion* is used here to imply that the diffusion rate is different on the short- and the long- timescales.
79. Anomalous Diffusion in Physics Literature: This term is not used in the sense that is normally used in the physics literature, i.e., the long-term diffusion rate is proportional to  $t^\gamma$  with

$$\gamma \neq 1$$

– for long timescales, here

$$\gamma = 1$$

## Overview of the Model Predictions – Varying the Granularity

### Parameter $\epsilon$

1. Limit of  $\varepsilon_p \rightarrow 0$ : The effect of varying the order granularity  $\epsilon$  in the limit

$$\varepsilon_p \rightarrow 0$$



is first investigated.

2. Impact of the Granularity Parameter: As will be seen, granularity has an important effect on most of the properties of the model, particularly on depth, price impact, and price diffusion.
3. Regimes of the Granularity Parameter: The behavior can be divided into three regimes, roughly as follows.
4. Large  $\epsilon$ :

$$\epsilon \gtrsim 0.01$$

5. Large  $\epsilon$  Analysis #1: This corresponds to a large accumulation of the orders near the best bid and the best ask, nearly linear market impact, and roughly equal short- and long-term diffusion rates.
6. Large  $\epsilon$  Analysis #2: This is the regime where the mean-field approximation used works best.
7. Medium  $\epsilon$ :

$$\epsilon \sim 0.01$$

8. Medium  $\epsilon$  Analysis #1: In this case, the accumulation of orders at the best bid and the best ask is small, and near the midpoint the depth profile increases linearly with price.
9. Medium  $\epsilon$  Analysis #2: As a result, as a crude approximation, the price increases roughly as square root of the order size.
10. Small  $\epsilon$ :

$$\epsilon \lesssim 0.001$$



11. Small  $\epsilon$  Analysis #1: The accumulation of orders at the best bid and the best ask is very small, and near the midpoint the depth profile is a convex function of the price. The price impact is very concave.
12. Small  $\epsilon$  Analysis #2: The short-term price diffusion rate is much smaller than the long-term price diffusion rate.
13. Symmetry among the Bid/Offer Treatment: Since the results of the bids are similar to those of the offers at

$$p = 0$$

results only for offers, i.e., buy market orders and sell limit orders, are shown.

14. Prices Relative to Midpoint: In this sub-section, prices are measured relative to the midpoint, and simulations are in the continuum limit where the tick size

$$\varepsilon_p \rightarrow 0$$

x

15. Results far from the Midpoint: Also, the predictions of the model are not valid far from the midpoint due to the unrealistic assumption of an order placement process with an infinite domain.
16. Range of Price Validity: Thus, the results are potentially relevant to real markets only when the price  $p$  is at most a few times large as the characteristic price  $p_c$ .

## Overview of the Mode Predictions – Varying the Granularity Parameter $\epsilon$ ; Depth Profile

1. Mean and Cumulative Depth vs  $\hat{p}$ :

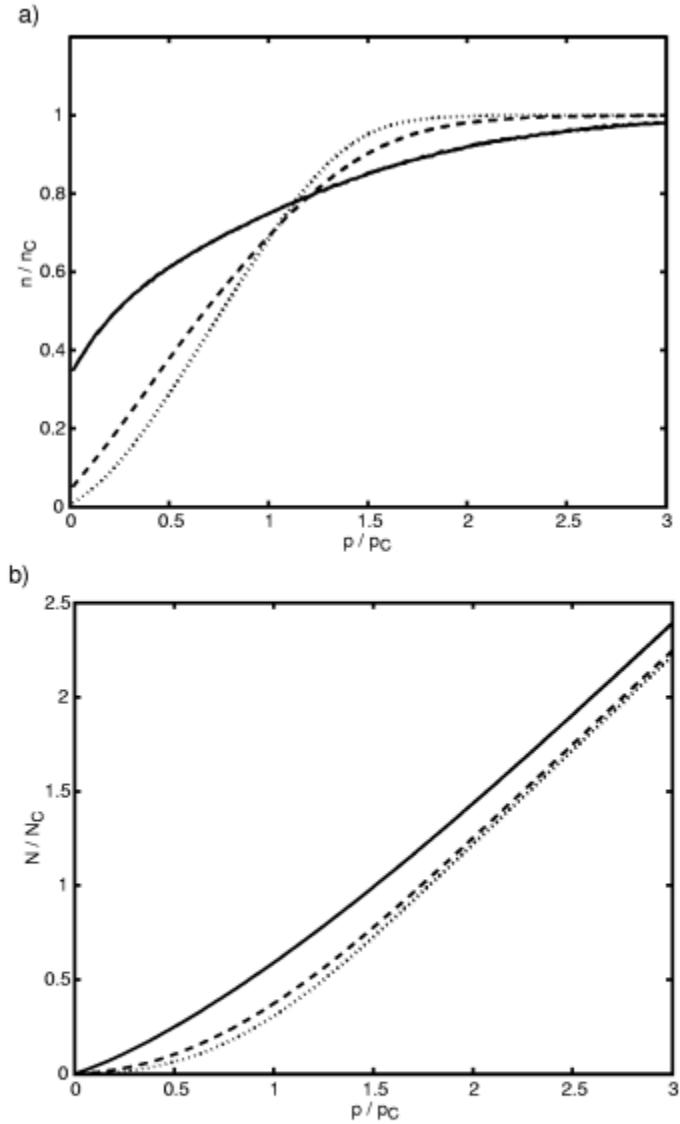


FIG. 3: The mean depth profile and cumulative depth versus  $\hat{p} = p/p_c = 2\alpha p/\mu$ . The origin  $p/p_c = 0$  corresponds to the midpoint. (a) is the average depth profile  $n$  in nondimensional coordinates  $\hat{n} = np_c/N_c = n\delta/\alpha$ . (b) is nondimensional cumulative depth  $N(p)/N_c$ . We show three different values of the nondimensional granularity parameter:  $\epsilon = 0.2$  (solid),  $\epsilon = 0.02$  (dash),  $\epsilon = 0.002$  (dot), all with tick size  $dp = 0$ .

2. Mean Depth Profile: The *mean depth profile*, i.e., the average number of shares per price interval, and the mean cumulative depth profile are shown in the figure above.



3. Units of the Depth Profile: Since the depth profile has units  $\frac{\text{shares}}{\text{price}}$ , non-dimensional units of the depth profile are

$$\hat{n} = \frac{np_c}{N_c} = \frac{n\delta}{\alpha}$$

4. Cumulative Depth Profile: The cumulative depth profile at any time  $t$  is given as

$$N(p, t) = \sum_{\tilde{p}=0}^p n(\tilde{p}, t) \varepsilon_p$$

5. Units of the Cumulative Depth Profile: This has units of shares, and so in non-dimensional terms it is

$$\hat{N}(p) = \frac{N(p)}{N_c} = \frac{2\delta N(p)}{\mu} = \frac{N(p)\epsilon}{\mu}$$

6. High  $\epsilon$  Regime Depth Profile: In the high- $\epsilon$  regime, the annihilation rate due to market orders is low relative to  $\delta\sigma$  and there is a significant accumulation of the orders at the best ask, so that the average depth is much greater than zero at its mid-point. The mean depth profile is a concave function of price.
7. Medium  $\epsilon$  Regime Depth Profile: In the median  $\epsilon$  regime, the market order removal rate increases, depleting the average depth near the best ask, and the profile is nearly linear over the range

$$\frac{p}{p_c} \leq 1$$



8. Low  $\epsilon$  Regime Depth Profile: In the small  $\epsilon$  regime, the market order removal rate increases further, making the average depth near the ask very close to zero, and the profile is a convex function over the range

$$\frac{p}{p_c} \leq 1$$

9. Standard Deviation of the Depth:

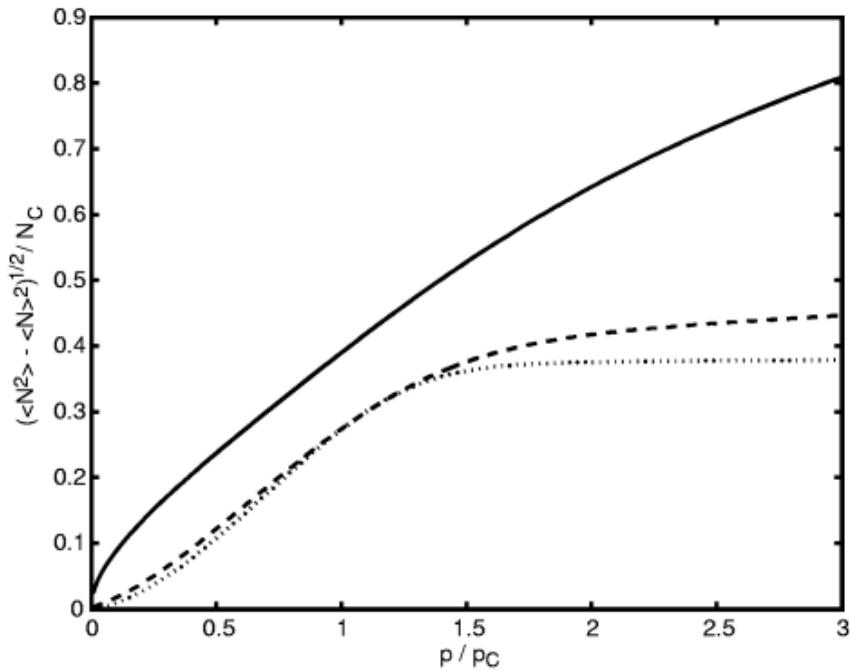


FIG. 4: Standard deviation of the nondimensionalized cumulative depth versus nondimensional price, corresponding to Fig. (3).

10. Standard Deviation of the Cumulative Depth: It can be seen that the standard deviation of the cumulative depth is comparable to the mean depth, and as  $\epsilon$  increases, near the midpoint there is a similar transition from convex to concave behavior.



11. Uniform Order Placement Process: The uniform order placement process seems at the first glance to be one of the most unrealistic assumptions of the model, leading to depth profile with a finite asymptotic depth – which also implies that there are an infinite number of orders in the book.
12. Orders far away from the Spread: However, orders far away from the spread in the asymptotic region almost never get executed and thus do not affect the market dynamics.
13. Depth vs Executed Profiles:

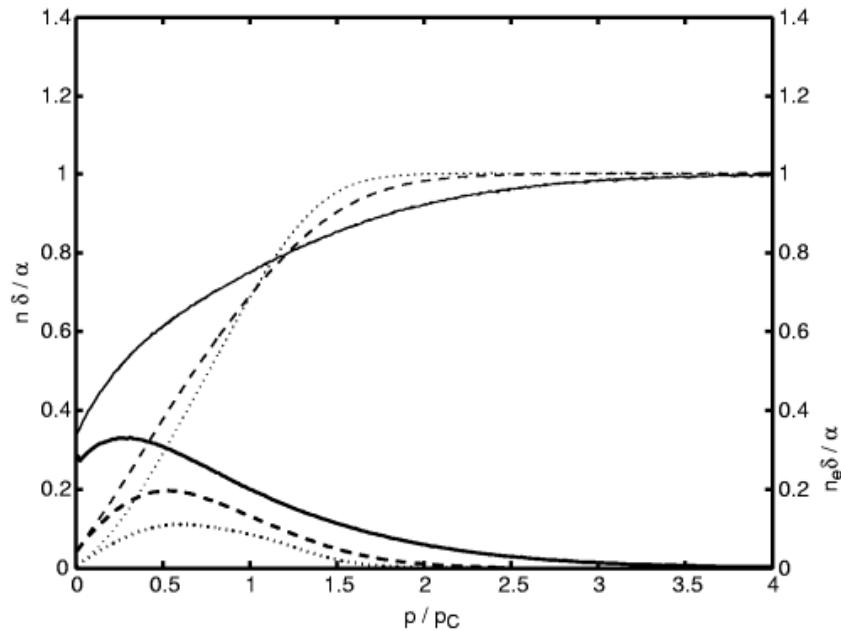


FIG. 5: A comparison between the depth profiles and the effective depth profiles as defined in the text, for different values of  $\epsilon$ . Heavy lines refer to the effective depth profiles  $n_e$  and the light lines correspond to the depth profiles.

14. Limit vs Executed Orders Depth: To demonstrate this, the figure above shows the comparison between the limit order dept and the depth  $n_e$  of only those orders which eventually get executed.



15. Probability of Execution vs Order Filling: Note that the ratio  $\frac{n_e}{n}$  is not the same as the probability of filling orders, because in that case the price  $\frac{p}{p_c}$  refers to the distance of the order from the midpoint at the time when it was placed.
16. Density of Executed Orders: The density  $n_e$  of executed orders decreases rapidly as a function of the distance from the midpoint.
17. Alternative Order Placement Processes: Therefore, one expects that near the midpoint the results should be similar to alternative order placement processes, as long as they lead to an exponentially decaying profile of executed orders – which is what is observed.
18. More Realistic Order Placement Processes: However, to understand the behavior further away from the midpoint, a later section summarizes the enhancements that includes more realistic order placement processes grounded on empirical measurements of market data.

## **Overview of the Model Predictions – Varying the Granularity Parameter $\epsilon$ ; Liquidity for Market Orders: The Price Impact Function**

1. Instantaneous Price Impact: The subsection studies instantaneous price impact function  $\phi(t, \omega, \tau \rightarrow 0)$
2. Definition of Temporary Price Impact: This is defined as the logarithm of the mid-price shift immediately after the arrival of a market order in the absence of any other events.
3. Permanent Price Impact: This should be distinguished from the asymptotic price impact  $\phi(t, \omega, \tau \rightarrow \infty)$  which describes the permanent price shift. While the permanent price impact is clearly important, it is not considered here.
4. Liquidity for Market Orders: The price impact function provides a measure of the liquidity for executing market orders.



5. Liquidity for Limit Orders: The liquidity for limit orders, in contrast, is given by the probability of execution.
6. Inverse of Cumulative Depth Profile: At any time  $t$ , the instantaneous

$$\tau = 0$$

price impact function is the inverse of the cumulative depth profile.

7. Price Impact Continuum over Depth: This follows immediately from

$$\omega = \sum_{p=a(t)}^{p'} n(p, t) \varepsilon_p$$

and

$$N(p, t) = \sum_{\tilde{p}=0}^p n(\tilde{p}, t) \varepsilon_p$$

which in the limit

$$\varepsilon_p \rightarrow 0$$

can be replaced by the continuum transaction equation:

$$\omega = N(p, t) = \int_0^p n(\tilde{p}, t) d\tilde{p}$$

8. Explicit Form for Price Impact: The equation makes it clear that for any fixed  $t$  the price impact can be regarded as the inverse of the cumulative depth profile  $N(p, t)$ .



9. When Depth Fluctuations are Small: When fluctuations are sufficiently small, one can replace  $n(p, t)$  by its mean value

$$n(p) = n(p, t)$$

10. Simplification for Average of Inverse Depth: In general, however, fluctuations can be large, and the average of the inverse is not equal to the inverse of the average.

11. Moments based on Depth Profile Corrections: There are corrections based on higher order moments of the depth profile, as given in the moment expansion derived in the next few sections.

12. Insight using Simplified Depth Distribution: Nonetheless, the inverse of the mean cumulative depth provides a qualitative approximation that gives insight into the behavior of the price impact function.

13. Advantage of Using Median Equivalents: Note that everything becomes much simpler using the medians, since the median of the cumulative price impact is exactly the opposite of the median price impact, as derived later.

14. Average Price Impact:

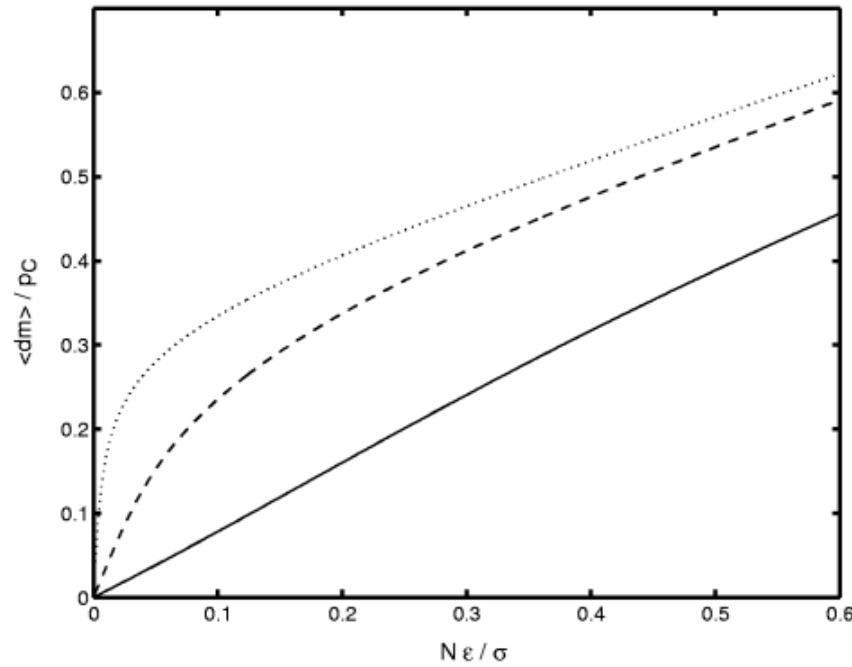


FIG. 6: The average price impact corresponding to the results in Fig. (3). The average instantaneous movement of the nondimensional mid-price,  $\langle dm \rangle / p_c$  caused by an order of size  $N/N_c = N\epsilon/\sigma$ .  $\epsilon = 0.2$  (solid),  $\epsilon = 0.02$  (dash),  $\epsilon = 0.002$  (dot).

15. Standard Deviation of Price Impact:

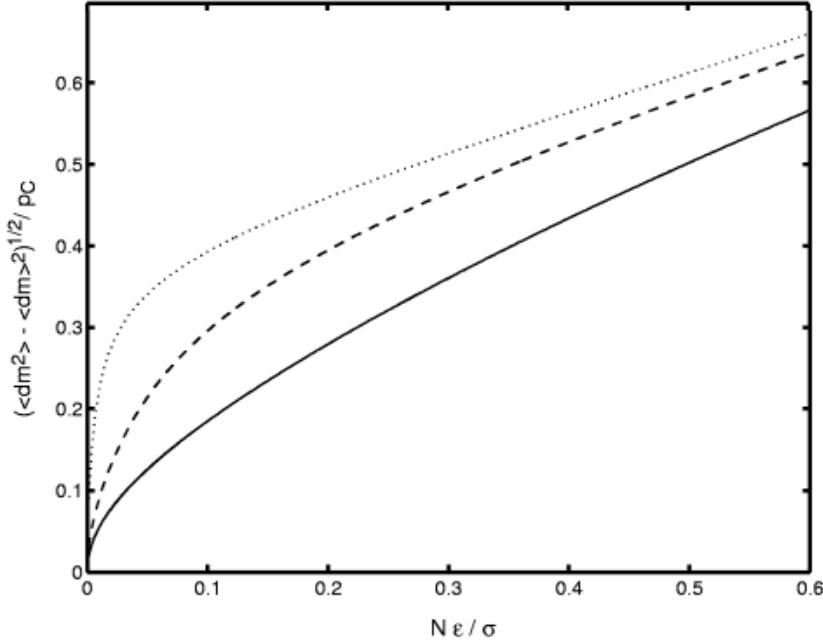


FIG. 7: The standard deviation of the instantaneous price impact  $dm/p_c$  corresponding to the means in Fig. 6, as a function of normalized order size  $\epsilon N/\sigma$ .  $\epsilon = 0.2$  (solid),  $\epsilon = 0.02$  (dash),  $\epsilon = 0.002$  (dot).

16. Mean/Standard Deviation of Price Impact: The above two pictures show the mean price impact function and the standard deviation of the price impact.
17. Large Fluctuations of the Price Impact: The price impact exhibits very large fluctuations for all values of  $\epsilon$ . The standard deviation has the same order of magnitude as the mean or even greater for small  $\frac{N\epsilon}{\sigma}$  values.
18. Virtualness of the Price Impact: Note that these are *virtual impact functions*. That is, to explore the behavior of the instantaneous price impact over a wide range of order sizes, one periodically computes the price impact that the order of a give size would have caused at that instant, if it had been submitted.
19. Comparison to Real-world Impact: Checks that real price impact curve has been the same have been carried out, but they require a much longer time to accumulate statistics.



20. Scale of the Price Impact: One of the interesting results of the above figures is the scale of the price impact. The price impact is measured relative to the characteristic price scale  $p_c$ , which is roughly equal to the mean spread.
21. Impact over Wide  $\epsilon$  Range: As will be argued later, the range of nondimensional shares shown in the horizontal axis spans the range of reasonable order sizes.
22. Price Impact Compared to Spread: Thus, it will be shown that throughout the range, the price impact is typically of the order of magnitude, and typically less, than the mean spread size.
23. Price Impact in Large  $\epsilon$  Regime: Due to accumulation of orders in the large  $\epsilon$  regime, the small  $p$  mean price impact is roughly linear.
24. Large  $\epsilon$  Price Impact Derivation: This follows from

$$\omega = N(p, t) = \int_0^p n(\tilde{p}, t) d\tilde{p}$$

under the assumption that  $n(p)$  is constant.

25. Median  $\epsilon$  Price Impact: In the median  $\epsilon$  regime, under the assumption that the variance in depth can be neglected, the mean price should increase roughly as  $\sqrt{\omega}$ .
26. Median  $\epsilon$  Price Impact Derivation: This follows from

$$\omega = N(p, t) = \int_0^p n(\tilde{p}, t) d\tilde{p}$$

under the assumption that  $n(p)$  is linearly increasing and

$$n(0) \approx 0$$

27. Corrections to Price Impact Analysis: Note that this is a crude approximation, but there can be substantial corrections introduced by the variance of the depth profile.



28. Small  $\epsilon$  Price Impact: Finally, in the small  $\epsilon$  regime, the price impact is highly concave, increasing much slower than  $\sqrt{\omega}$ . This follows because

$$n(0) \approx 0$$

and the depth profile  $n(p)$  is convex.

29. Nondimensional Mid-price Movement:

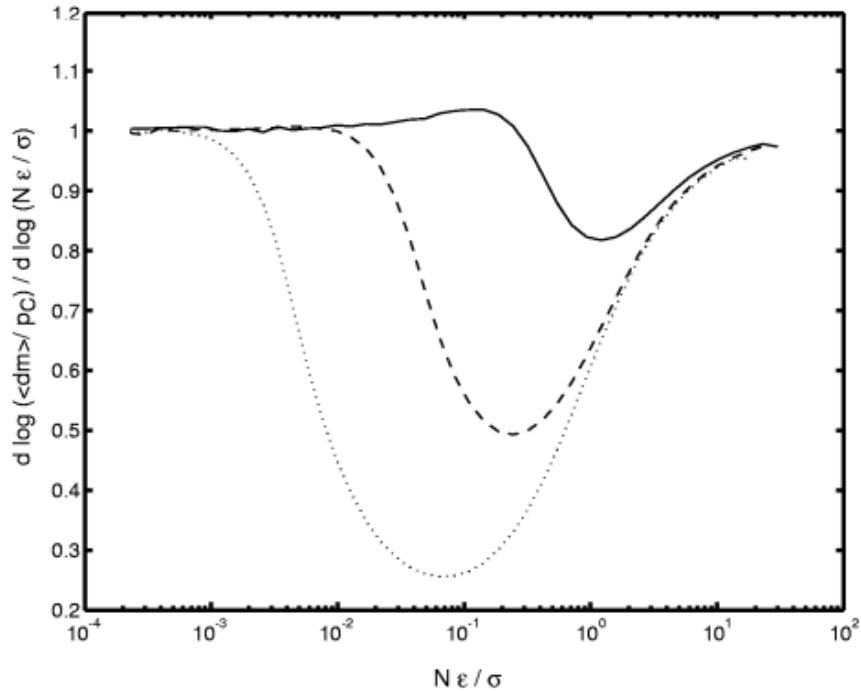


FIG. 8: Derivative of the nondimensional mean mid-price movement, with respect to logarithm of the nondimensional order size  $N/N_c = N\epsilon/\sigma$ , obtained from the price impact curves in Fig. 6.



30. Price Impact vs Order Size: To get a better feel for the functional form of the price impact function, the figure above illustrates numerical differentiation vs. log order size, and the result is plotted as a function of the appropriately scaled order size.
31. Automatic Application of Log Price: Note that because the prices are logarithmic, the vertical axis already incorporates the logarithm.
32. Locally Implied Power Law: If one were to fit a local power law approximation to the function at each price, this corresponds to the exponent of the power law near that price.
33. Exponent always Less than One: Notice that the exponent is always less than one, so that the price impact is almost always concave.
34. Assuming Limited Depth Variance: Making the assumption that the effect of the variance of the depth is not too large, so that

$$\omega = N(p, t) = \int_0^p n(\tilde{p}, t) d\tilde{p}$$

is a good assumption, the behavior can be understood as follows.

35. Near  $\frac{N}{N_c} \approx 0$ : For

$$\frac{N}{N_c} \approx 0$$

the price impact is dominated by the constant term in the average depth profile  $n(0)$ , and so the logarithmic slope of the price impact is always near one.

36. Increasing  $\frac{N}{N_c}$ : As  $\frac{N}{N_c}$  increases, the logarithmic is driven by the shape of the average depth profile, which is liner or convex for smaller  $\epsilon$ , resulting in concave price impact.
37. Large  $\frac{N}{N_c}$ : For large values of  $\frac{N}{N_c}$ , one reaches the asymptotic region where the depth profile is flat – and where the model is invalid by design.



38. Impact of the Mean Approximation: Of course, there can be deviations to this caused by the fact that the mean of the inverse depth profile is not in general the inverse of the mean, i.e.

$$\langle N^{-1}(p) \rangle \neq \langle N(p) \rangle^{-1}$$

39. Interpreting  $\frac{N}{\sigma}$ : To compare real data, note that

$$\frac{N}{N_c} = \frac{N\epsilon}{\sigma}$$

$\frac{N}{\sigma}$  is just the order size in shares in relation to the average order size, so, by definition, it has a typical value of one.

40. London Stock Exchange Example: For the London Stock Exchange, typical values of  $\epsilon$  are in the range  $0.001 - 0.1$

41. Corresponding  $\frac{N}{N_c}$  Range: For a typical range of order sizes from  $100 - 100,000$  shares, the meaningful range for  $\frac{N}{N_c}$  is therefore roughly  $10^{-5}$  to 1

42. Corresponding  $\epsilon$  Range: In this range, for small values of  $\epsilon$ , the exponent can reach values as low as 0.2

43. The Price Impact Concavity: This offers a possible explanation for the previously mysterious nature of the price impact function, and contradicts the linear increase in price impact based on the naïve argument presented in the introduction.

## Relationship of Price Impact to Cumulative Depth

1. Definition of Immediate Liquidity: An important aspect of the markets is immediate liquidity, by which one means immediate response of prices to incoming market orders.



2. Execution Range of a Market Order: When a market order enters, its execution depends both on the spread and on the depth of the orders in the book.
3. Sequence of Transacted Prices: These determine the sequence of transacted prices produced by that order, as well as the instantaneous market impact.
4. Longer Horizon Liquidity: Long term liquidity depends on the longer-term response of the LOB, and is characterized by the price impact function  $\phi(\omega, \tau)$  for values

$$\tau > 0$$

5. Relation between Liquidity and Volatility: Immediate liquidity affects short-term volatility and long-term liquidity affects volatility measured over longer timescales.
6. Focus on Short-term Liquidity: This section addresses only short-term liquidity. Volatility on longer timescales is addressed later.
7. Liquidity Using Depth/Market Impact: Liquidity is characterized in terms of either depth profile or price impact.
8. Recap - Depth Profile: The *depth profile*  $n(p, t)$  is the number of shares  $n$  at price  $p$  at time  $t$ .
9. Cumulative Depth Profile  $N$ : For many purposes, it is convenient to think in terms of the *cumulative depth profile  $N$* , which is the sum of  $n$  values up – or down – to some price.
10. Book Center as Reference: For convenience, the reference point is established at the center of the book, where one defines

$$p \equiv 0$$

and

$$N(0) \equiv 0$$



11. What Constitutes Price/Depth Reference: The reference point can either be the midpoint quote, the best bid, or the best ask.
12. Explicit Price Impact Function: Also studied is the price impact function

$$\Delta p = \phi(\omega, \tau, t)$$

where  $\Delta p$  is the shift in price at  $t + \tau$  caused by an order size  $\omega$  placed at time  $t$ .

13. Shift in the Mid-price: Typically, one defines  $\Delta p$  as a shift in the mid-price though it is possible to use best bid or best ask

$$\omega = \sum_{p=a(t)}^{p'} n(p, t) \varepsilon_p$$

14. Relation between Impact and Depth: The price impact function and the depth profile are closely related, but the relationship is not trivial.
15.  $N(\Delta p)$ :  $N(\Delta p)$  gives the average total number of orders up to a distance  $\Delta p$  from the origin.
16. Price Shift Caused by Orders: Whereas, in order to calculate the price impact, what one needs is the average shift caused by a *fixed* number of orders.
17. Instantaneous Price Impact: Making the identifications

$$p = \Delta p$$

and

$$N = \omega$$

and choosing a common reference point, the instantaneous price impact is the inverse of the instantaneous cumulative depth, i.e.



$$\phi(0, \omega, t) = N^{-1}(\omega, t)$$

18. Average Price Impact: This relationship is clearly true instant by instant. However, it is not true for averages, since the mean of the inverse is in general not equal to the inverse of the mean, i.e.

$$\langle \phi \rangle = \langle N \rangle^{-1}$$

19. Estimating the Moment Dependence: This is highly relevant here, since because the fluctuations in these functions are huge, the interest is primarily in their statistical properties, and in particular the first few moments.
20. Price Moments vs Depth Moments Relationship: The relationship between the moments is derived in the following section.

## Relationship of Price Impact to Cumulative Depth – Moment Expansion

1. Market Impact from Cumulative Count: There is some subtlety in how one relates the market impact to the cumulative order count.
2. Market Impact as Midpoint Shift: One eligible definition of the market impact  $\Delta p$  is the movement of the midpoint, following the placement of an order of size  $\omega$ .
3. Market Impact from Ask Price: If one defines the reference point so that

$$N(a, t) \equiv 0$$

and the market order is a buy, the definition puts

$$\omega(\Delta p, t) = N(a + 2\Delta p, t) - N(a, t)$$

In other words, the midpoint is half the shift in the best offer.



4. Midpoint Centered Price Reference: An alternative choice would be to let

$$\omega(\Delta p, t) = N(\Delta p, t)$$

which would include part of the instantaneous spread in the definition of impact in midpoint-centered coordinates, or none of it in ask coordinates.

- 5. Importance of Reference Point Choice: The issue of how impact is related of  $N(p, t)$  is separate from whether the best ask is equal to the reference point of the prices, and may be chosen differently to answer different questions.
- 6. Monotonicity of  $\Delta p$  vs  $\omega$ : Under any such definition, however, the impact  $\Delta p$  is a monotonic function of  $\omega$  in ever instance, so either may be taken as the independent variable along with the index  $t$  that labels the instance.
- 7. Relationship between  $\omega/\Delta p$  Averages: One wishes to account for the differences in instance averages of  $\omega$  and  $\Delta p$ , regarded respectively as the dependent variables, in terms of the fluctuations of the other.
- 8. Monotonicity of the Cumulative  $N(p, t)$ : In spite of the fact that the density  $n(p, t)$  is a highly discontinuous variable in general, monotonicity of the cumulative  $N(p, t)$  enables us to picture a power series expansion for  $\omega(p, t)$  in  $p$  with coefficients that fluctuate in time.
- 9. Postulates for  $\omega(p, t)$  and  $p(\omega, t)$ : The simplest such expression that captures the behavior of the simulated output is

$$\omega(p, t) = a(t) + b(t) \cdot p + \frac{c(t)^2}{2} \cdot p^2$$

if  $p$  is regarded as the independent variable, or

$$p(\omega, t) = \frac{-b(t) + \sqrt{b^2(t) + 2c(t)[\omega - a(t)]}}{c(t)}$$



if  $\omega$  is.

10. Need for  $a(t)$ : While the variable  $a(t)$  would seem unnecessary since  $\omega$  is 0 at

$$p = 0$$

empirically one finds that the simultaneous fits for both  $\omega$  and  $\omega^2$  at the lower order can be made better by incorporating the additional freedom of fluctuations in  $a$ .

11. Explicit Expression for the Time Fluctuation: One imagines splitting each  $t$ -dependent coefficient into its mean, and a zero-mean fluctuation component, as

$$a(t) = \bar{a} + \delta_a(t)$$

$$b(t) = \bar{b} + \delta_b(t)$$

and

$$c(t) = \bar{c} + \delta_c(t)$$

The fluctuation components will, in general, depend on  $\epsilon$ .

12.  $a/b/c$  from  $\omega$ 's Moments: The values of the mean and the second moment of the fluctuations can be extracted from the mean distributions of  $\langle \omega \rangle$  and  $\langle \omega^2 \rangle$ .  
 13. Expectation of  $\omega$  at  $p = 0$ : The mean values come from the linear expectation

$$\langle \omega(0) \rangle = \bar{a}$$

$$\left. \frac{\partial \langle \omega(p) \rangle}{\partial p} \right|_{p=0} = \bar{b}$$

and



$$\left. \frac{\partial^2 \langle \omega(p) \rangle}{\partial p^2} \right|_{p=0} = \bar{c}$$

14. Expectation of  $\omega^2$  at  $p = 0$ : Given these, the fluctuations then come from the quadratic expectation as

$$\left. \frac{\partial^2 \langle \omega^2(p) \rangle}{\partial p^2} \right|_{p=0} = 2(\bar{b}^2 + \bar{a}\bar{c}) + 2\langle \delta_b^2 + \delta_a\delta_c \rangle$$

and

$$\left. \frac{\partial^3 \langle \omega^2(p) \rangle}{\partial p^3} \right|_{p=0} = 6(\bar{b}\bar{c} + \langle \delta_b\delta_c \rangle)$$

and

$$\left. \frac{\partial^4 \langle \omega^2(p) \rangle}{\partial p^4} \right|_{p=0} = 6(\bar{c}^2 + \langle \delta_c^2 \rangle)$$

15.  $\omega$  Density from Occupation Number Density: When  $\omega$  is given a specific definition in terms of the cumulative distribution, its averages become averages over the density in the order book.

16. Taylor's Expansion of  $p$  off of  $a$ ,  $b$ , and  $c$ : The values of the moments as obtained above may then be used in a derivative expansion of the inverse function

$$p(\omega, t) = \frac{-b(t) + \sqrt{b^2(t) + 2c(t)[\omega - a(t)]}}{c(t)}$$

making the prediction of the average impact



$$\begin{aligned} \langle p(\omega) \rangle &= \bar{p} + \frac{1}{2} \overline{\frac{\partial^2 p}{\partial a^2}} \Big|_{a(t)=\bar{a}} \langle \delta_a^2 \rangle + \frac{1}{2} \overline{\frac{\partial^2 p}{\partial b^2}} \Big|_{b(t)=\bar{b}} \langle \delta_b^2 \rangle + \frac{1}{2} \overline{\frac{\partial^2 p}{\partial c^2}} \Big|_{c(t)=\bar{c}} \langle \delta_c^2 \rangle \\ &\quad + \overline{\frac{\partial^2 p}{\partial a \partial b}} \Big|_{a(t)=\bar{a}; b(t)=\bar{b}} \langle \delta_a \delta_b \rangle + \overline{\frac{\partial^2 p}{\partial a \partial c}} \Big|_{a(t)=\bar{a}; c(t)=\bar{c}} \langle \delta_a \delta_c \rangle \\ &\quad + \overline{\frac{\partial^2 p}{\partial b \partial c}} \Big|_{b(t)=\bar{b}; c(t)=\bar{c}} \langle \delta_b \delta_c \rangle \end{aligned}$$

where the overbar denotes the evaluation of

$$p(\omega, t) = \frac{-b(t) + \sqrt{b^2(t) + 2c(t)[\omega - a(t)]}}{c(t)}$$

or its indicated derivative at

$$a(t) = \bar{a}$$

$$b(t) = \bar{b}$$

$$c(t) = \bar{c}$$

and  $\omega$ .

17.  $\langle \delta_b^2 \rangle$  and  $\langle \delta_a \delta_c \rangle$ : The fluctuations  $\langle \delta_b^2 \rangle$  and  $\langle \delta_a \delta_c \rangle$  cannot be determined independently from

$$\left. \frac{\partial^2 \langle \omega^2(p) \rangle}{\partial p^2} \right|_{p=0} = 2(\bar{b}^2 + \bar{a}\bar{c}) + 2\langle \delta_b^2 + \delta_a \delta_c \rangle$$

18. Determining  $\langle \delta_b^2 \rangle$  and  $\langle \delta_a \delta_c \rangle$ : However, in keeping with this fact, their coefficient functions in



$$\begin{aligned}\langle p(\omega) \rangle &= \bar{p} + \frac{1}{2} \overline{\frac{\partial^2 p}{\partial a^2}} \Big|_{a(t)=\bar{a}} \langle \delta_a^2 \rangle + \frac{1}{2} \overline{\frac{\partial^2 p}{\partial b^2}} \Big|_{b(t)=\bar{b}} \langle \delta_b^2 \rangle + \frac{1}{2} \overline{\frac{\partial^2 p}{\partial c^2}} \Big|_{c(t)=\bar{c}} \langle \delta_c^2 \rangle \\ &\quad + \overline{\frac{\partial^2 p}{\partial a \partial b}} \Big|_{a(t)=\bar{a}; b(t)=\bar{b}} \langle \delta_a \delta_b \rangle + \overline{\frac{\partial^2 p}{\partial a \partial c}} \Big|_{a(t)=\bar{a}; c(t)=\bar{c}} \langle \delta_a \delta_c \rangle \\ &\quad + \overline{\frac{\partial^2 p}{\partial b \partial c}} \Big|_{b(t)=\bar{b}; c(t)=\bar{c}} \langle \delta_b \delta_c \rangle\end{aligned}$$

are identical, so the inversion remains fully specified.

19. Components of  $\langle p(\omega) \rangle$  Taylor Expansion: Denoting by  $\bar{Z}$  the radical

$$\bar{Z} = \sqrt{b^2(t) + 2c(t)[\omega - a(t)]}$$

the various partial derivative functions in

$$\begin{aligned}\langle p(\omega) \rangle &= \bar{p} + \frac{1}{2} \overline{\frac{\partial^2 p}{\partial a^2}} \Big|_{a(t)=\bar{a}} \langle \delta_a^2 \rangle + \frac{1}{2} \overline{\frac{\partial^2 p}{\partial b^2}} \Big|_{b(t)=\bar{b}} \langle \delta_b^2 \rangle + \frac{1}{2} \overline{\frac{\partial^2 p}{\partial c^2}} \Big|_{c(t)=\bar{c}} \langle \delta_c^2 \rangle \\ &\quad + \overline{\frac{\partial^2 p}{\partial a \partial b}} \Big|_{a(t)=\bar{a}; b(t)=\bar{b}} \langle \delta_a \delta_b \rangle + \overline{\frac{\partial^2 p}{\partial a \partial c}} \Big|_{a(t)=\bar{a}; c(t)=\bar{c}} \langle \delta_a \delta_c \rangle \\ &\quad + \overline{\frac{\partial^2 p}{\partial b \partial c}} \Big|_{b(t)=\bar{b}; c(t)=\bar{c}} \langle \delta_b \delta_c \rangle\end{aligned}$$

evaluate to

$$\frac{1}{2} \overline{\frac{\partial^2 p}{\partial a^2}} \Big|_{a(t)=\bar{a}} = -\frac{\bar{c}}{2\bar{Z}^3}$$

$$\overline{\frac{\partial^2 p}{\partial a \partial b}} \Big|_{a(t)=\bar{a}; b(t)=\bar{b}} = \frac{\bar{b}}{\bar{Z}^3}$$



$$\frac{1}{2} \overline{\frac{\partial^2 p}{\partial b^2}} \Big|_{b(t)=\bar{b}} = \overline{\frac{\partial^2 p}{\partial a \partial b}} \Big|_{a(t)=\bar{a}; b(t)=\bar{b}} = \frac{\omega - \bar{a}}{\bar{Z}^3}$$

$$\overline{\frac{\partial^2 p}{\partial b \partial c}} \Big|_{b(t)=\bar{b}; c(t)=\bar{c}} = \frac{1}{\bar{c}^2} - \frac{\bar{b}}{\bar{Z} \bar{c}^2} - \frac{(\omega - \bar{a})^2}{2 \bar{c} \bar{Z}^3}$$

20. The Mean-price Impact: Plugging these into

$$\begin{aligned} \langle p(\omega) \rangle &= \bar{p} + \frac{1}{2} \overline{\frac{\partial^2 p}{\partial a^2}} \Big|_{a(t)=\bar{a}} \langle \delta_a^2 \rangle + \frac{1}{2} \overline{\frac{\partial^2 p}{\partial b^2}} \Big|_{b(t)=\bar{b}} \langle \delta_b^2 \rangle + \frac{1}{2} \overline{\frac{\partial^2 p}{\partial c^2}} \Big|_{c(t)=\bar{c}} \langle \delta_c^2 \rangle \\ &\quad + \overline{\frac{\partial^2 p}{\partial a \partial b}} \Big|_{a(t)=\bar{a}; b(t)=\bar{b}} \langle \delta_a \delta_b \rangle + \overline{\frac{\partial^2 p}{\partial a \partial c}} \Big|_{a(t)=\bar{a}; c(t)=\bar{c}} \langle \delta_a \delta_c \rangle \\ &\quad + \overline{\frac{\partial^2 p}{\partial b \partial c}} \Big|_{b(t)=\bar{b}; c(t)=\bar{c}} \langle \delta_b \delta_c \rangle \end{aligned}$$

gives the predicted mean price impact, compared to the actual mean in the figure below.

21. Inverse Mean Cumulative Order Comparison:

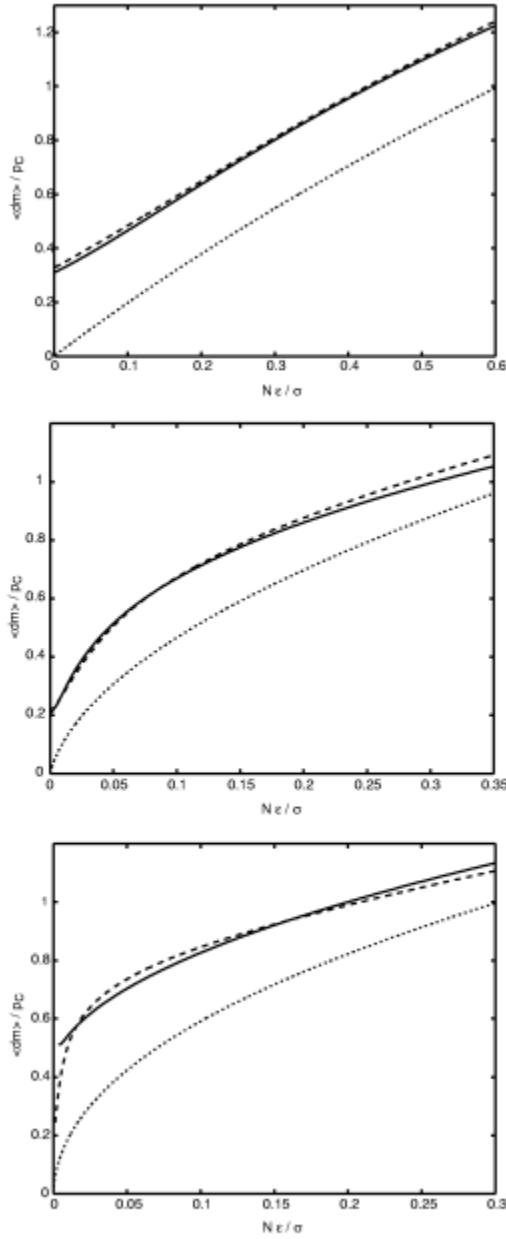


FIG. 31: Comparison of the inverse mean cumulative order distribution  $\bar{p}$  (dot), to the actual mean impact (solid), and the second-order fluctuation expansion (A14, dash). (a) :  $\epsilon = 0.2$ . (b) :  $\epsilon = 0.02$ . (c) :  $\epsilon = 0.002$ .

22. Order Distribution in Ask-Centered Coordinates: The cumulative order distribution is computed in ask-centered coordinates, eliminating the contribution from the half-spread in the  $p$  coordinates.



23. Comparison without Second-Order Fluctuations: The inverse of the mean cumulative distribution which correspond to  $\bar{p}$  in

$$\begin{aligned}\langle p(\omega) \rangle = \bar{p} + & \frac{1}{2} \overline{\frac{\partial^2 p}{\partial a^2}} \Big|_{a(t)=\bar{a}} \langle \delta_a^2 \rangle + \frac{1}{2} \overline{\frac{\partial^2 p}{\partial b^2}} \Big|_{b(t)=\bar{b}} \langle \delta_b^2 \rangle + \frac{1}{2} \overline{\frac{\partial^2 p}{\partial c^2}} \Big|_{c(t)=\bar{c}} \langle \delta_c^2 \rangle \\ & + \overline{\frac{\partial^2 p}{\partial a \partial b}} \Big|_{a(t)=\bar{a}; b(t)=\bar{b}} \langle \delta_a \delta_b \rangle + \overline{\frac{\partial^2 p}{\partial a \partial c}} \Big|_{a(t)=\bar{a}; c(t)=\bar{c}} \langle \delta_a \delta_c \rangle \\ & + \overline{\frac{\partial^2 p}{\partial b \partial c}} \Big|_{b(t)=\bar{b}; c(t)=\bar{c}} \langle \delta_b \delta_c \rangle\end{aligned}$$

clearly underestimated the actual mean impact.

24. Corrections from Second Order Fluctuations: However, the corrections from only the second-order fluctuations in  $a$ ,  $b$ , and  $c$  account for much of the difference at all values of  $\epsilon$ .

## Relationship of Price Impact to Cumulative Depth – Quantiles

- Quantile Relationship between Impact/Depth: Another way to characterize the relationship between the depth profile and market impact is in terms of their quantiles – fraction greater than a given value, for example the median is the 0.5 quantile.
- Explicit Form of Relationship: Interestingly, the relationship between the quantiles is trivial. Letting  $Q_r(x)$  be the  $r^{th}$  quantile of  $x$ , because the cumulative depth  $N(p)$  is a non-decreasing function with inverse

$$p = \phi(N)$$

one has the relation

$$Q_r(\phi) = [Q_{1-r}(N)]^{-1}$$



3. Case of Fine Price Ticks: This provides an easy and accurate way to compare depth and price impact when the tick size is sufficiently small.
4. Case of Coarse Price Ticks: However, when tick size is very coarse, the quantiles in general are not very useful, because unlike the mean, the quantiles do not vary continuously, and only take a few discrete values.

## **Overview of the Mode Parameters – Varying the Granularity Parameter $\epsilon$ ; Spread**

1. Probability Density of the Spread:

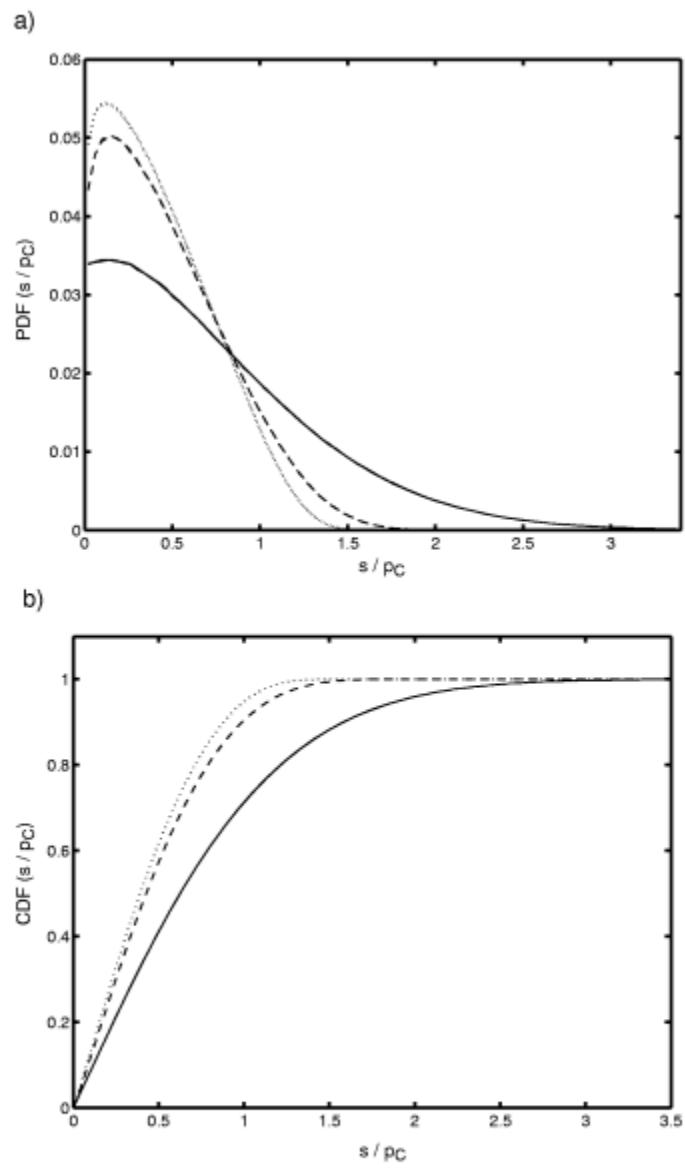


FIG. 9: The probability density function (a), and cumulative distribution function (b) of the nondimensionalized bid-ask spread  $s/p_c$ , corresponding to the results in Fig. (3).  $\epsilon = 0.2$  (solid),  $\epsilon = 0.02$  (dash),  $\epsilon = 0.002$  (dot).

2. Spread Probability Density at  $\frac{s}{p_c} = 0$ : This shows that the spread probability density is substantial at



$$\frac{s}{p_c} = 0$$

which is the limit

$$\varepsilon_p \rightarrow 0$$

3. Peak of the Spread Density: The spread probability density reaches a maximum value of the spread approximately  $0.2p_c$ , and then decays.
4. Spread Decay Speed with  $\epsilon$ : It might seem surprising that at first it decays more slowly for large  $\epsilon$  where there is large accumulation of orders at the ask.
5. Characteristic Price Dependence on  $\epsilon$ : However, it should be borne in mind that the characteristic price

$$p_c = \frac{\mu}{\alpha}$$

depends on  $\epsilon$ .

6. Recasting  $p_c$  using  $\epsilon$ : Since

$$\epsilon = \frac{2\delta\sigma}{\mu}$$

by eliminating  $\mu$  this can be written

$$p_c = \frac{2\delta\sigma}{\alpha\epsilon}$$

Thus, holding the other parameters fixed, large  $\epsilon$  corresponds to small  $p_c$ , and vice versa.



7. Spread Dependence on  $\epsilon$ : So, in fact, spread is very small for large  $\epsilon$ , and large for small  $\epsilon$ , as expected.
8. Corrections to the Dimensional Scaling Limits: The figure above shows small corrections to the large effects predicted by the dimensional scaling relations.
9. Spread Decay for Large  $\epsilon$ : For large  $\epsilon$ , the probability density of the spread decays roughly exponentially away from the midpoint.
10. Depth Fluctuations for Large  $\epsilon$ : This is because for large  $\epsilon$  the fluctuations around the mean depth are roughly independent.
11. Probability for Market Order Penetration: Thus, the probability for a market order to penetrate at a given price level is roughly the probability that all the ticks smaller than the price level contain no orders, which gives rise to an exponential decay.
12. Behavior for Small  $\epsilon$ : This is no longer true for small  $\epsilon$ . Note that for small  $\epsilon$ , the probability distribution of the spread becomes insensitive to  $\epsilon$ , i.e. the nondimensionalized distribution for

$$\epsilon = 0.02$$

is nearly the same as that for

$$\epsilon = 0.002$$

13. Mean Spread increases with  $\epsilon$ : It is apparent from the previous figure that, in nondimensional units, the mean spread increases with  $\epsilon$ . This is confirmed in the next figure, which displays the mean value of the spread as a function of  $\epsilon$ . The mean spread increases monotonically with  $\epsilon$ .
14. Mean Spread as a Function of  $\epsilon$ :

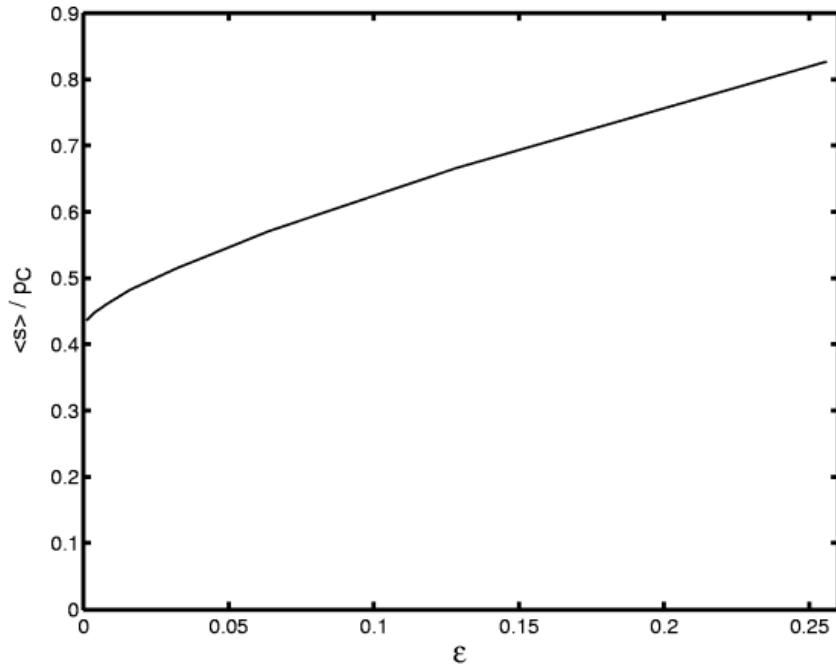


FIG. 10: The mean value of the spread in nondimensional units  $\hat{s} = s/p_c$  as a function of  $\epsilon$ . This demonstrates that the spread only depends weakly on  $\epsilon$ , indicating that the prediction from dimensional analysis given in table (III) is a reasonable approximation. .

15. Linear Spread vs  $\epsilon$ : It depends on  $\epsilon$  as roughly a constant – equal to approximately 0.45 in nondimensional coordinates – plus a linear term whose slope is rather small.

16. Spread Variation in a Realistic Range: It is believed that for most financial instruments

$$\epsilon < 0.3$$

Thus, the variation in spread caused by varying  $\epsilon$  is the range

$$0 < \epsilon < 0.3$$



is not large, and the dimensional analysis based on the rate parameters given in previous table is a good approximation.

17. Spread over the Full  $\epsilon$  Range: One gets an accurate prediction of  $\epsilon$  dependence across the full range of  $\epsilon$  from the Independent Interval Approximation technique derived later.

## **Overview of the Mode Parameters – Varying the Granularity Parameter $\epsilon$ ; Volatility and Price Diffusion**

1. Price Diffusion Rate: The price diffusion rate, which is proportional to the square of the volatility, is important for determining risk, and is a property of central interest.
2. Diffusion Rate from Order Flow: From dimensional analysis in terms of the order flow rates, the price diffusion has units of  $\frac{\text{price}^2}{\text{time}}$  and so must scale as  $\frac{\mu^2 \delta}{\alpha^2}$
3. Characteristic Step for Random Walk: A crude argument for this is as follows: the dimensional estimate for spread is  $\frac{\mu}{2\alpha}$ .
4. Characteristic Time for Random Walk: Let this be the characteristic step size of a random walk, and let the step frequency be the characteristic time  $\frac{1}{\delta}$ , which is the average lifetime for a share to be canceled. This argument also gives the above estimate for the diffusion rate.
5. Impact of Autocorrelation: However, this is not correct in the presence of negative autocorrelations in the step sizes.
6.  $\epsilon$ -Based Adjustments to Diffusion: The numerical simulations make it clear that there are  $\epsilon$ -dependent corrections to this model, as demonstrated below.
7. Price Variance over Time:

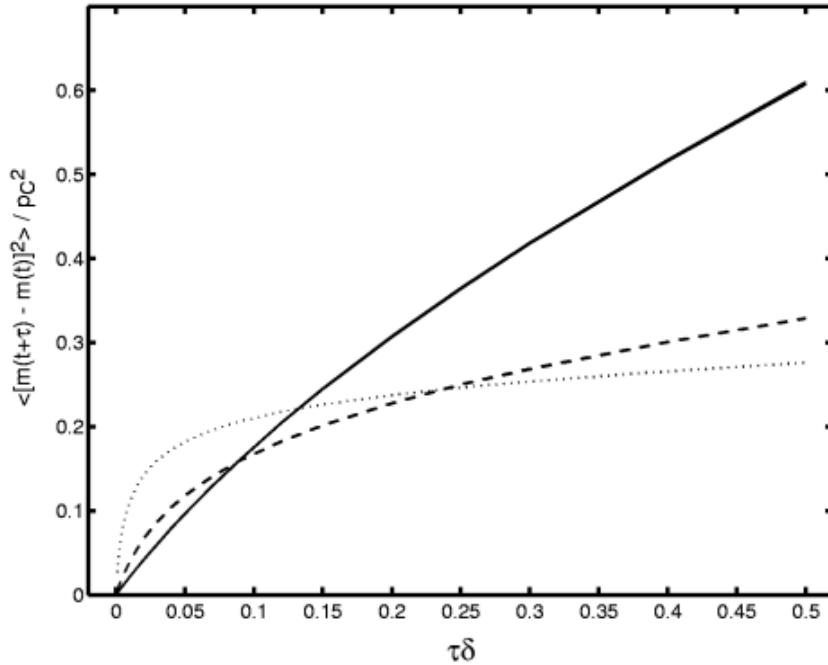


FIG. 11: The variance of the change in the nondimensionalized midpoint price versus the nondimensional time delay interval  $\tau\delta$ . For a pure random walk this would be a straight line whose slope is the diffusion rate, which is proportional to the square of the volatility. The fact that the slope is steeper for short times comes from the nontrivial temporal persistence of the order book. The three cases correspond to Fig. 3:  $\epsilon = 0.2$  (solid),  $\epsilon = 0.02$  (dash),  $\epsilon = 0.002$  (dot).

8. Variance in Midpoint Process: The figure above shows simulation results for the variance of the change in the midpoint price at timescale  $\tau$ ,  $\mathbb{V}[m(t + \tau) - m(t)]$
9. Diffusion Rate from Variance Plot: The slope is the diffusion rate, which at a fixed timescale is proportional to the square of the volatility.
10. Multiple Diffusion Timescales: It appears there are at least two timescales involved, with a faster diffusion rate for short timescales and a slower diffusion rate for long timescales.



11. Diffusion Corrections from Order Flow: Simulation results show that the diffusion rate is correctly described by the product of the estimate from dimensional analysis based on order flow parameters alone,  $\frac{\mu^2 \delta}{\alpha^2}$ , and a  $\tau$ -dependency parameter of the nondimensional granularity

$$\epsilon = \frac{2\delta\sigma}{\mu}$$

12. Qualitative Understanding of Diffusion Adjustments: However, a qualitative understanding can be gained based on the conservation law derived later.
13. Conservation Law Impact on Price Diffusion: A discussion on how this relates to price diffusion is shown later.
14. Evolution Behavior Implications for Price Autocorrelation: Not that the temporal structure in the diffusion process also implies non-zero autocorrelations of the mid-price  $m(t)$ .
15. Persistence of Weak Autocorrelation: This corresponds to weak negative autocorrelations in price differences  $m(t) - m(t - 1)$  that persist for timescales until the variance vs.  $\tau$  becomes a straight line.
16. Timescales for the Persistence: The timescale depends on parameters, but is typically of the order of 50 market order arrival times.

## Overview of the Mode Parameters – Varying the Granularity Parameter $\epsilon$ ; Liquidity for Limit Orders: Probability and Time-to-Fill

1. Liquidity for Limit Orders: The liquidity for limit orders depends on the probability that they will be filled, and the time to be filled.
2. Limit Orders Close to Spread: This obviously depends on price; limit orders close to current transaction prices are more likely to be filled quickly, while those far away have a lower likelihood of being filled.



3. Probability of Filling a Limit Order vs Price:

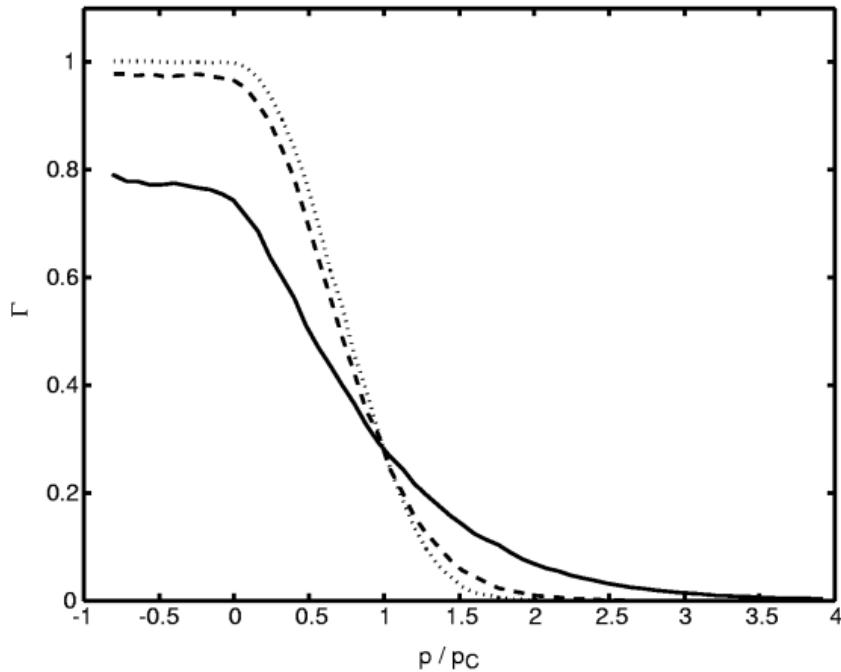


FIG. 12: The probability  $\Gamma$  for filling a limit order placed at a price  $p/p_c$  where  $p$  is calculated from the instantaneous mid-price at the time of placement. The three cases correspond to Fig. 3:  $\epsilon = 0.2$  (solid),  $\epsilon = 0.02$  (dash),  $\epsilon = 0.002$  (dot).

4. Limit Order Fill Probability: The figure above plots the probability  $\Gamma$  of a limit order being filled versus the nondimensionalized price at which it was placed. As with all figures in this section, this is based off a mid-centered frame.
5. Fill Probability versus  $\epsilon$ : The figure shows that, in nondimensional coordinates, the probability of filling close to the bid for sell limit orders – or ask for buy limit orders – decreases as  $\epsilon$  increases.
6. Effect of High  $\epsilon$  on Fill: For large  $\epsilon$ , this is less than 1 even for negative prices. This says that even for sell orders that are placed close to the best bid, there is a significant chance that the offer is deleted before being executed.
7. Effect of Low  $\epsilon$  on Fill: This is not true for smaller values of  $\epsilon$ , where



$$\Gamma(0) \approx 1$$

8. Fill Probability far from Spread: Far from the spread, the fill probabilities as a function of  $\epsilon$  are reversed, i.e., the probability of filling limit orders increases as  $\epsilon$  increases.
9. Fill Probability  $\epsilon$  Crossover: The crossover point, where the fill probabilities are roughly the same, occurs at

$$p \approx p_c$$

10. Comparison to Depth Profiles: This is consistent with the earlier illustration on depth profile that also shows that the depth profile for different values of  $\epsilon$  cross at about

$$p \sim p_c$$

11. Time for Filling a Limit Order:

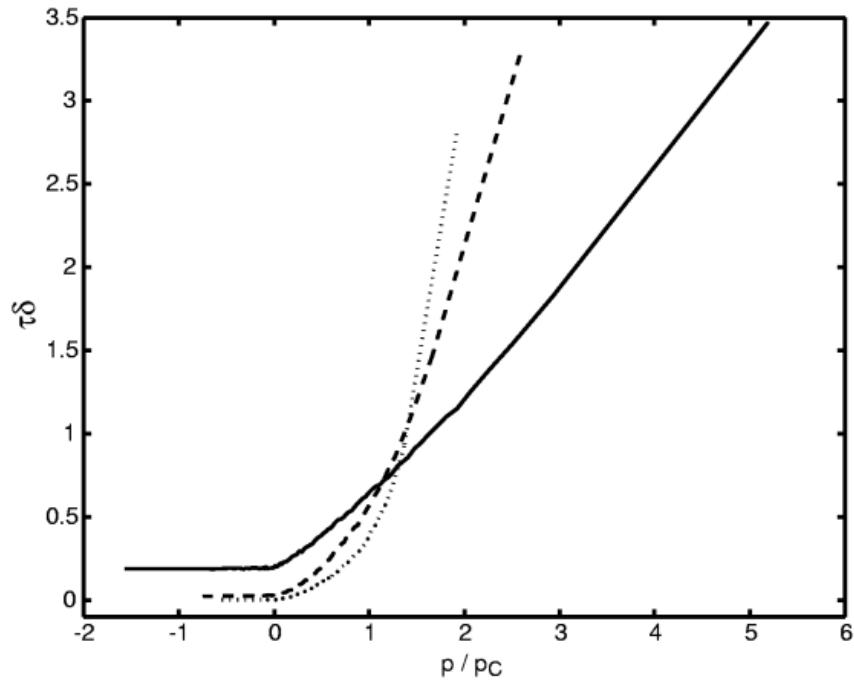


FIG. 13: The average time  $\tau$  nondimensionalized by the rate  $\delta$ , to fill a limit order placed at a distance  $p/p_c$  from the instantaneous mid-price.

12. Fill Time vs Price: Similarly, the figure above shows the average time  $\tau$  taken to fill an order placed at a distance  $p$  from the instantaneous mid-price.

13. Reappearance of the Crossover: Again, one sees that the average time is larger at larger values of  $\epsilon$  for small  $\frac{p}{p_c}$ , and that this behavior reverses at

$$p \sim p_c$$

## Overview of the Mode Parameters – Varying the Tick Size $\frac{\epsilon_p}{p_c}$

1. Dependence of Market Properties on Tick Size:

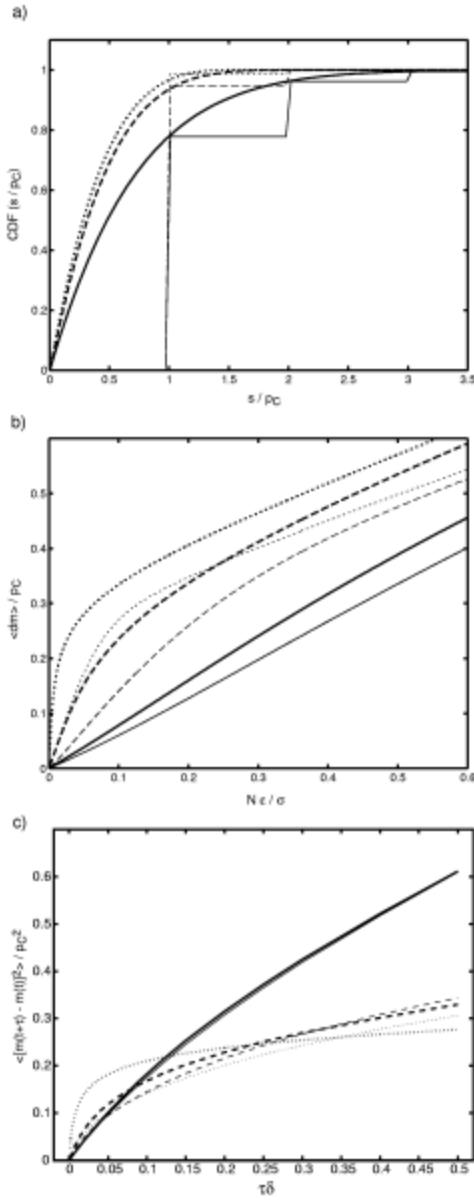


FIG. 14: Dependence of market properties on tick size. Heavy lines are  $dp/p_c \rightarrow 0$ ; light lines are  $dp/p_c = 1$ . Cases correspond to Fig. 3, with  $\epsilon = 0.2$  (solid),  $\epsilon = 0.02$  (dash),  $\epsilon = 0.002$  (dot). (a) is the cumulative distribution function for the nondimensionalized spread. (b) is instantaneous nondimensionalized price impact, (c) is diffusion of the nondimensionalized midpoint shift, corresponding to Fig. 11.



2. Tick Size Dependence: The dependence on discrete tick size  $\frac{\varepsilon_p}{p_c}$ , of the cumulative distribution function for the spread, instantaneous price impact, and mid-price diffusion, are shown above.
3. Choice of  $\frac{\varepsilon_p}{p_c} = 1$ : An unrealistically large value of the tick size, with

$$\frac{\varepsilon_p}{p_c} = 1$$

has been chosen to show that even with very coarse ticks, the qualitative changes in behavior are relatively minor.

4. CDF of the Spread: Part (a) of the figure shows the cumulative density function of the spread, comparing

$$\frac{\varepsilon_p}{p_c} = 0$$

and

$$\frac{\varepsilon_p}{p_c} = 1$$

5. Spread Distribution for Coarse Ticks: It is apparent from this figure that the spread distribution for coarse ticks *effectively integrates* the distribution in the limit

$$\varepsilon_p \rightarrow 0$$

6. Cumulative Depth at Integer Ticks: That is, at integer tick values, the mean cumulative depth profiles roughly match, and in between the integer tick values, the probability for coarse ticks is smaller.



7. Quantization Resulting from Coarse Ticks: This happens for the obvious reason that the coarse ticks quantize the possible values of the spread, and place a lower value of one tick for spread.
8. Shift in Mean Spread: The shift in mean spread from this effect is not shown, but this is consistent with this result; there is a constant offset of roughly  $\frac{1}{2}$  tick.
9. Price Impact of  $\frac{\varepsilon_p}{p_c}$ : The alteration in the price impact is shown in part (b). Unlike the spread distribution, the average price impact varies continuously.
10. Averaging over many Ticks: Even though the tick size is quantized, one is averaging over many events and the probability of a price impact of each tick size is a continuous function of the order size.
11. Price Impact Function of  $p$ : Large tick size consistently lowers the price impact. The price impact rises more slowly for small  $p$ , but is then similar except for a downward translation.
12. Effects of Coarse Ticks on Diffusion: The effect of coarse ticks is less trivial for the mid-price diffusion, as shown in figure (c).
13. Coarse Ticks at Small  $\epsilon$ : At

$$\epsilon = 0.002$$

coarse ticks remove most of the rapid short-term volatility at the mid-point, which in the continuous price case arises from price fluctuations smaller than

$$\frac{\varepsilon_p}{p_c} = 1$$

14. Reduction in Anomalous Diffusion: This lowers the negative autocorrelation of mid-point price returns, and reduces the anomalous diffusion.
15. Coarse Ticks at  $\epsilon = 0.02$ : At

$$\epsilon = 0.02$$



where both early volatility and late auto-correlation are similar, coarse ticks have less effect.

16. Sensitivity of Diffusion to  $\epsilon$  with Coarseness: The next result is that the mid-price diffusion becomes less sensitive to the value of  $\epsilon$  as tick size increases, and there is less anomalous price diffusion.

## Theoretical Analysis – Summary of Analytic Methods

1. Twin Analytical Approach: The model has been investigated analytically using two approaches.
2. Approach Based on Master Equation: The first one is based on a master equation. This works best in the mid-point centered frame.
3. Shares at each Price Tick: The attempt here is to solve correctly for the average number of shares at each price tick as a function of price.
4. Mid-price Process: The midpoint price makes a random walk with a nonstationary distribution.
5. Analytic Solution for the Average Depth: Thus, the key to finding a stationary analytic solution for the average depth is to use co-moving coordinates, which are centered at a reference point near the center of the book, such as the midpoint or the best bid.
6. Independence of Adjacent Price Fluctuations: In the first approximation, fluctuations about the main depth at adjacent prices are treated as independent.
7. Probability Density over Occupation Numbers: This allows replacing the depth profile over a simpler probability density over occupation numbers  $n$  at each  $p$  and  $t$ .
8. Continuum Limit in Tick Price: The continuum limit can be taken by letting the tick size  $\epsilon_p$  become infinitesimal.
9. Multiple Order at a Tick: With finite order flow rates, this gives a vanishing probability for the existence of more than one order at any tick as



$$\varepsilon_p \rightarrow 0$$

10. Correlations as a Function of  $\epsilon$ : With this approach, one is able to test the relevance of the correlation of function of parameter  $\epsilon$  as well as predict the functional dependence of the cumulative distribution of the spread on the depth profile.
11. Correlation Dependence on  $\epsilon$ : It is seen that correlations are negligible for large values of  $\epsilon$

$$\epsilon \sim 0.2$$

while they are very important for small values, i.e.

$$\epsilon \sim 0.002$$

12. Independent Interval Approximation IIA: The second analytic approach, called the *Independent Interval Approximation*, is most easily carried out in the bid-centered frame.
13. Representation Used in the IID Approach: This uses a different representation, in which the solution is expressed in terms of the empty intervals between non-empty price ticks.
14. Evolution of the Interval Set: The system is characterized at any instant of time by a set of intervals  $\{\dots, x_{-1}, x_0, x_1, x_2, \dots\}$  where, for example,  $x_0$  is the spread distribution between the bid and the ask,  $x_{-1}$  is the distance between the second buy limit order and the bid, and so on.
15. Price Space and Order Profile:

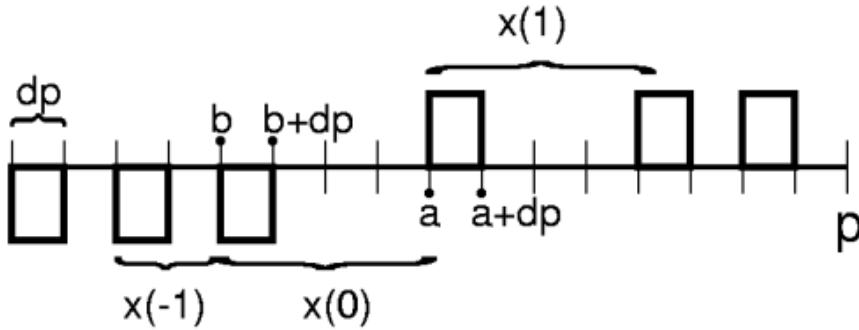


FIG. 15: The price space and order profile.  $n(p, t)$  has been chosen to be 0 or  $\pm 1$ , a restriction that will be convenient later. Price bins are labeled by their lower boundary price, and intervals  $x(N)$  will be defined below.

16. Master Equation for Interval Evolution: Equations are written for how a given interval varies with time.
17. System of Coupled Non-linear Equations: Changes to adjacent interval are related, resulting in an infinite set of coupled non-linear equations.
18. Solution using Mean-field Approximations: However, using a mean-field approximation, one is able to solve the equations, albeit only numerically.
19. Evolution of Spread and Depth: Besides predicting how the various intervals, e.g., the spread, vary with the parameters, this approach also predicts the depth profile as a function of these parameters.
20. Comparisons against Numerical Simulations: These predictions are compared against numerical simulations.
21. Comparison over the Range of  $\epsilon$ : They match very well for large  $\epsilon$  and less well for smaller values of  $\epsilon$ .
22. Extensions to the IIA Approach: The IIA can also be modified to incorporate various extensions to the model, as seen later.
23. Deployment of the Mean-field Approach: On both approaches, one uses a mean-field approximation to get a solution.



24. Independence among Adjacent Fluctuations: The approximation basically lies in assuming that fluctuations in adjacent intervals – which might be adjacent price ranges in the master equation approach or empty levels in the IIA – are independent.
25. Use in Continuum Tick Limit: Also, both approaches are most easily tractable only in the continuum limit

$$\varepsilon_p \rightarrow 0$$

when every tick has at most only one order.

26. Generalizations across the Tick Levels: They may, however, be extended to general tick sizes as well.
27. Effectiveness in the Large- $\epsilon$  Limit: Because correlations are important for small  $\epsilon$ , both methods work well mostly in the large- $\epsilon$  limit, although qualitative aspects of small  $\epsilon$  may also be gleaned from them.
28. Empirical Evidence for  $\epsilon$  Levels: Unfortunately, at least based on preliminary investigation of LSE data, it seems that it is this small  $\epsilon$  limit that real markets may tend more towards.
29. Conceptual Insights into Small  $\epsilon$ : Thus, these approximate solutions may not be as useful as one would like. Nonetheless, they do provide some conceptual insights into what determines depth and price impact.
30. Shape of the Mean Depth: In particular, one finds that the shape of the mean depth profile depends on a single parameter  $\epsilon$ , and that the relative sizes of its first few derivatives account for both the order-size dependence of the market impact, and the re-normalization of the midpoint diffusivity.
31. Market vs Limit Order Comparison: A higher relative rate of market vs limit order depletes the center of the book, though less than the classical estimate predicts. This leads to more concave impact and faster short-term diffusivity.
32. Profile Far from the Mid: However, orders pile up more quickly – versus classically nondimensionalized price – with the distance from the midpoint, causing rapid early diffusion to suffer large mean reversion.



33. Impact Dependence on Price Autocorrelation: The following sections will remark on the above, however, the qualitative relation to midpoint autocorrelation supplies a potential interpretation of the data, which may be more robust than details of the model assumptions or its quantitative results.
34. Applying Global Conservation Laws: Both of the treatments described above are approximations. One can derive the exact global law of order placement and removal later, whose consequences are elaborated there.
35. Conservation Laws as Solution Constraint: The conservation law must be respected in any sensible analysis of the model, giving one a check on the approximation.
36. Explaining Anomalous Diffusion: It also provides insight into the anomalous diffusion properties of the model.

## Characterizing Limit-order Books: Dual Coordinates

1. Definition of the Price Space: Price is a dimensional quantity, and the space is divided into bins of length  $\varepsilon_p$  representing the ticks, which may be finite or infinitesimal.
2. Discrete or Continuous Valued Prices: Prices are discrete or continuous valued, respectively.
3. LOB Configurations: Statistical properties of interest are computed from the temporal sequences or ensembles of LOB configurations.
4.  $n(p, t)$  corresponding to the Configuration: If  $n$  is the variable used to denote the number of shares in some bonds  $(p, p + \varepsilon_p)$  at the beginning of  $t$  of an elementary interval, a configuration is specified by a function  $n(p, t)$ .
5. Sign Convention for  $n(p, t)$ : It is convenient to take  $n$  positive for sell limit orders, and negative for buy limit orders.
6. Corresponding Bid and Ask Prices: Because the model dynamics preclude crossing limit orders, there is in general a highest instantaneous buy limit-order price, called the bid  $b(t)$  and a lowest sell limit-order price, the ask  $a(t)$ , with



$$b(t) < a(t)$$

always.

7. Midpoint Price: The *midpoint price*, defined as

$$m(t) = \frac{a(t) + b(t)}{2}$$

may or may not be the price of any actual bin if prices are discrete -  $m(t)$  may be a half-integer of  $\varepsilon_p$ .

8. Illustration of the Price Space: These quantities are diagrammed in the figure above.  
 9. Cumulative Order Count LOB Representation: An equivalent specification of an LOB configuration is given by the cumulative order count

$$N(p, t) = \sum_{-\infty}^{p-\varepsilon_p} |n(p, t)| - \sum_{-\infty}^{a-\varepsilon_p} |n(p, t)|$$

where  $-\infty$  denotes the lower boundary of the price space, whose exact value must not affect the results.

10. Equivalent Use of Bid as Origin: Because, by definition, there are no orders between the bid and the ask, the bid could have equivalently been used as the origin of the summation.  
 11. Price Bins Indexed by Lower Boundaries: Because price bins will be indexed here by their lower boundaries, though, it is convenient here to use the ask.  
 12. Sign Convention for  $N$ : The absolute values have been placed so that  $N$ , like  $n$ , is negative in the range of buy orders and positive in the range of sells.  
 13. Construction of  $N(p, t)$ :

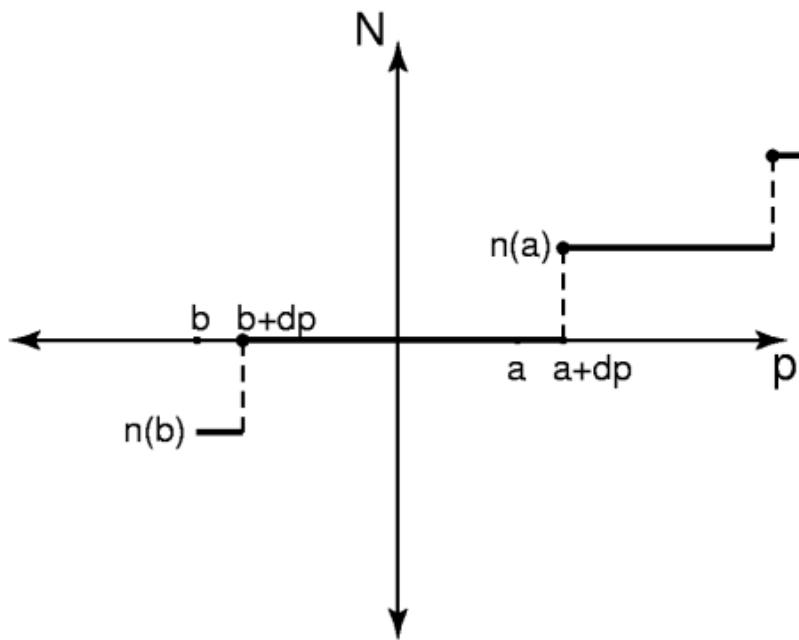


FIG. 16: The accumulated order number  $N(p, t)$ .  $N(a, t) \equiv 0$ , because contributions from all bins cancel in the two sums.  $N$  remains zero down to  $b(t) + dp$ , because there are no uncanceled, nonzero terms.  $N(b, t)$  becomes negative, because the second sum in Eq. (4) now contains  $n(b, t)$ , not canceled by the first.

14.  $N \leftrightarrow$  Price Inverse Relation: In many cases of either sparse orders or infinitesimal  $\varepsilon_p$ , with fixed order size – which one may well define to a single share – there will be zero or one share in any bin, and the above specification of the LOB configuration

$$p(N, t) = \max\{p \mid N(p, t) = N\}$$

as shown below.

15. Inverse Function  $p(N, t)$ :

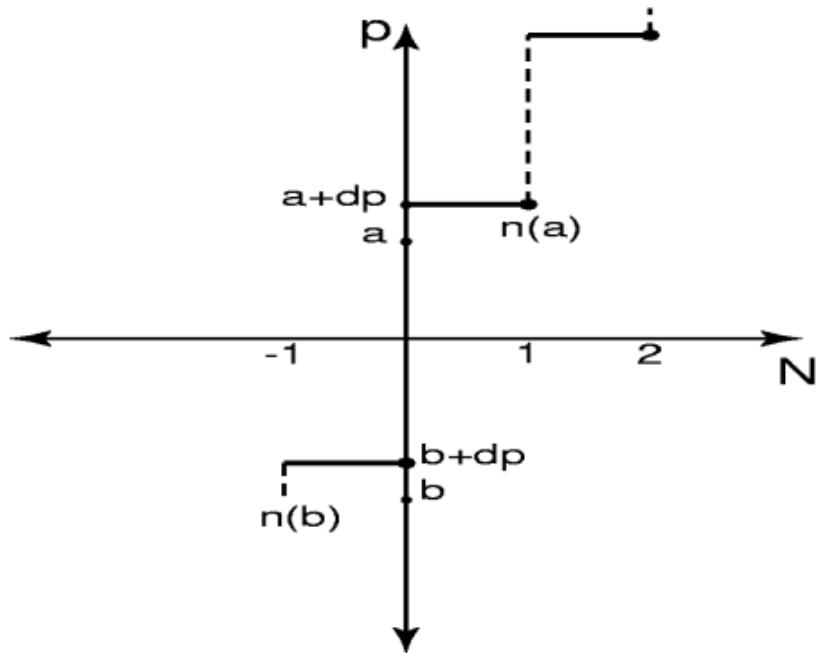


FIG. 17: The inverse function  $p(N, t)$ . The function is in general defined only on discrete values of  $N$ , so this domain is only invariant when order size is fixed, a convenience that will be assumed below. Between the discrete domain, and the definition of  $p$  as a maximum, the inverse function effectively interpolates between vertices of the reflected image of  $N(p, t)$ , as shown by the dotted line.

16. Discreteness Inherent in the Inversion: Strictly, the inversion may be performed for any distribution of order sizes, but resulting function is intrinsically discrete, so its domain is only invariant when order size is fixed.
17. Well-definedness for  $p(N, t)$ : The above assumption will therefore be made to give  $p(N, t)$  the convenient properties of a well-defined function on an invariant domain.
18. Defining the Intervals between Orders: Using the definition above

$$p(0, t) \equiv a(t)$$

$$p(-1, t) \equiv b(t)$$



and one can define the intervals between orders as

$$x(N, t) = p(N, t) - p(N - 1, t)$$

19. Instantaneous Bid-Ask Spread: Thus

$$x(0, t) = a(t) - b(t)$$

the instantaneous bid-ask spread.

20. For Symmetric Order Placement Rules: Here the probability distributions over configurations will be symmetric under either

$$n(p, t) \rightarrow -n(-p, t)$$

or

$$x(N, t) \rightarrow x(-N, t)$$

21. Dual Configuration Descriptor Using  $n/x$ : Coordinates  $N$  and  $p$  furnish a dual description of configurations, and  $n$  and  $x$  are their associated differences.

22. Independent Fluctuations in either Representation: The Master Equation approach assumes independent fluctuations in  $n$  while IIA assumes independent fluctuation in  $x$ .

23.  $x_N$ : In the next sections, it is convenient to abbreviate

$$x(N, t) \equiv x_N(t)$$

## Frames and Martingales



1. Stationary Nature of  $x(N, t)$  Specification: The  $x(N, t)$  specification of the LOB configurations has the property that its distribution is stationary under the dynamics considered here.
2. Nonstationary Nature of  $p(N, t)$  or  $n(p, t)$ : The same is true for  $p(N, t)$  or  $n(p, t)$  directly, because bid, mid, and ask prices undergo a random walk, with a renormalized diffusion coefficient.
3. Stationary Distributions in Comoving Frames: Stationary distributions for  $n$ -variables can be obtained in *co-moving frames*, of which there are several natural choices.
4. Bid-centered Configuration: The *bid-centered configuration* is defined as

$$n_b(p, t) = n(p - b(t), t)$$

5. Mid-centered Configuration: If an appropriate rounding convention is adopted in the case of discrete prices, the *mid-centered configuration* can also be defined as

$$n_m(p, t) = n(p - m(t), t)$$

6. Differences between the above Centerings: The mid-centered configuration has qualitative differences from bid-centered configurations, which will be explored below.
7. Order Distributions and Diffusion Processes: Both give insights to the order distribution and diffusion processes.
8. Ask-centered Configuration: The ask-centered configuration  $n_a(p, t)$  need not be considered if the order placement and removal are symmetric, because it is a mirror image of  $n_b(p, t)$ .
9. Bid-centered Ask: The *spread* is defined as the difference

$$s(t) \equiv a(t) - b(t)$$



and is a value of ask in bid-centered coordinates.

10. Mid-centered Ask: In mid-centered coordinates, the ask appears at  $\frac{s(t)}{2}$ .
11. Relation between  $n_b$  and  $n_m$ : The configurations  $n_b$  and  $n_m$  are dynamically correlated over short time intervals, but evolve ergodically in periods longer than finite characteristic correlation times.
12. Probability Distributions as Time Averages: Marginal probability distributions for these can therefore be computed as time averages, either as functions on the while price space, or at discrete sets of prices.
13. Corresponding Means at  $p$ : Their marginal mean values at a single price  $p$  will be denoted  $\langle n_b(p) \rangle$  and  $\langle n_m(p) \rangle$ , respectively.
14. Global Balance Constraints: The means are subject to global balance constraints, between total order placement and removal in the price space.
15. Bid centered Balance Constraint: Because all limit orders are placed above the bid, the bid-centered configuration obeys a simple balance relation

$$\frac{\mu}{2} = \sum_{b+\varepsilon_p}^{\infty} [\alpha - \delta \langle n_b(p) \rangle]$$

16. Explanation of the above Constraint: The above equation says that market orders must account, on average, for the difference between all limit orders placed and all decays.
17. Conservation Constraint Impact on Diffusivity: After passing to nondimensional coordinates below, thus implies an inverse relation between corrections to the classical estimate for diffusivity at early and late times.
18. Conservation Laws determining  $x$ : In addition, this conservation law plays an important role in the analysis and determination of the  $x(N, t)$ 's.
19. Mid-centered Flow Configuration Constraint: The midpoint-centered averages satisfy a different constraint:



$$\frac{\mu}{2} = \alpha \frac{\langle s \rangle}{2} + \sum_{b+\varepsilon_p}^{\infty} [\alpha - \delta \langle n_m(p) \rangle]$$

20. Explain of the Above Constraint: Market orders in the equation above account not only for the excess of limit order placement over evaporation at prices above the midpoint, but also the *excess* order placed between  $b(t)$  and  $m(t)$ .
21. Shape of  $\langle n_m(p) \rangle$  relative to  $\langle n_b(p) \rangle$ : Since these always lead to midpoint shifts, they ultimately appear at positive co-moving coordinates, altering the shape of  $\langle n_m(p) \rangle$  relative  $\langle n_b(p) \rangle$ .
22. Corresponding Rate of Arrival: Their rate of arrival is

$$\alpha \langle m - b \rangle = \frac{\alpha \langle s \rangle}{2}$$

These results are also confirmed in the simulations.

## Factorization Tests

1. Solving the PDF for  $n(p)$ : Whether in the bid- and the mid-centered frame, the PDF for the entire configuration  $n(p)$  is too difficult a problem to solve in its entirety.
2. Marginal Equation using Independent Marginals: However, an approximate master equation can be found for  $n$  independently at each  $p$  if all joint probabilities factor into independent marginals, as

$$\mathbb{P}[\{n(p_i)\}_i] = \prod_i \mathbb{P}[n(p_i)]$$

where  $\mathbb{P}[\cdot]$  denotes, for instance, the probability density for  $n$  orders in some interval around  $p$ .



3. Expected Number of a Bin: Whenever orders are sufficiently sparse that the expected number in any price bin is simply the probability that the bins is occupied – up to a constant of proportionality – the independence assumption implies a relationship between the cumulative distribution for the spread of the ask and the mean density profile.
4. Explicit Relation for Spread Probability: In units where the order size is 1, the relation is

$$\mathbb{P}\left[\frac{s}{2} < p\right] = 1 - e^{-\sum_{p'=b+\varepsilon_p}^{b+\varepsilon_p} \langle n_m(p') \rangle}$$

5. Comparison against Simulations: This relation is tested against simulation results in the figure below. One can observe there are three regimes.
6.  $\mathbb{P}\left[\frac{s}{2} < p\right]$  CDF:

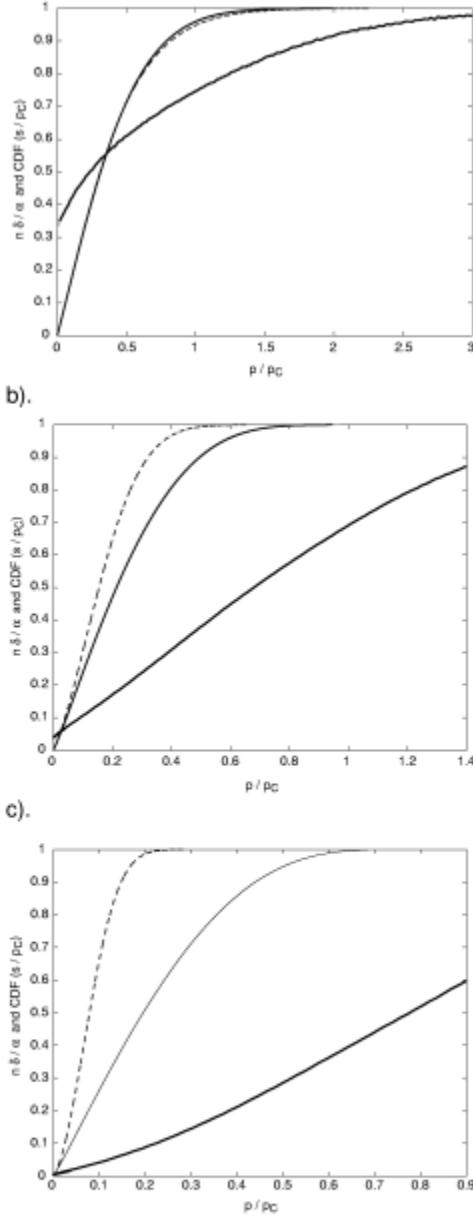


FIG. 18: CDFs  $\Pr(s/2 < p)$  from simulations (thin solid), mean density profile  $\langle n_m(p) \rangle$  from simulations (thick solid), and computed CDF of spread (thin dashed) from  $\langle n_m(p) \rangle$ , under the assumption of uncorrelated fluctuations, at three values of  $\epsilon$ . (a):  $\epsilon = 0.2$  (low market order rate); approximation is very good. (b):  $\epsilon = 0.02$  (intermediate market order rate); approximation is marginal. (c):  $\epsilon = 0.002$  (high market order rate); approximation is very poor.

7. Validity at High  $\epsilon$  #1: The high  $\epsilon$  regime is defined when the mean-density profile at the midpoint



$$\langle n_m(0) \rangle \lesssim 1$$

and strongly concave demand.

8. Validity at High  $\epsilon$  #2: In this regime, the approximation of independent fluctuations is excellent, and a master equation is expected to be useful.
9. Validity at Intermediate  $\epsilon$ : Intermediate  $\epsilon$  regime is defined by

$$\langle n_m(0) \rangle \ll 1$$

and nearly linear, and the approximation of independence is marginal.

10. Validity at Small  $\epsilon$ : Small  $\epsilon$  is defined by

$$\langle n_m(0) \rangle \ll 1$$

and concave upward, and the approximation of independent fluctuations is completely invalid. These regimes of validity also correspond to the qualitative ranges already noted.

11. Validity in the Bid-Centered Frame: In a bid-centered frame, however, that above equation never seems to be valid for any range of parameters. This will be discussed later why.
12. Mid centered Master Equation: For the present, therefore, the master equation approach is carried out in the mid frame.
13. MFT in Bid centered Frame: Alternatively, the mean-field theory of separations is most convenient in the bid-centered frame, so that the frame will be studied in the dual basis.
14. Comparison across Frames and Coordinates: The relation of results in the two frames, and via two methods of treatment, will provide a good qualitative, and for some properties quantitative, understanding of the depth profile and its effect on impacts.



15. Enforced Matching of Correlations: It is possible, in a modified treatment, to match certain features of the simulation at any  $\epsilon$ , by limited incorporation of correlated fluctuations.
16. Master Equation for Large  $\epsilon$ : However, the general master equation will be developed independent of these, and tested against simulations results at large  $\epsilon$ , where its defining assumptions are well met.

## Comments on Renormalized Diffusion

1. Diffusion Dependence on Horizon/ $\epsilon$ : A qualitative understanding of why the diffusivity is different over short and long-time scales, as well as why it may be dependent on  $\epsilon$ , may be gleaned from the following observations.
2. Constraints from Global Order Conservation: First, global order conservation places a strong constraint on classically nondimensionalized density profile in the bid-centered frame.
3. Profile Concavity at  $\epsilon \ll 1$ : It is seen that

$$\epsilon \ll 1$$

the density profile becomes concave upward near the bid, accounting for an increasing fraction of the allowed *remainder area*.

4. Density Profile for Order Conservation: Since this remainder area is fixed at unity, it can be conserved only if the density profile approaches unity more quickly with increasing pace.
5. The Resulting Short-Term Diffusivity: Low density at low price appears to lead to more persistent steps in the effective short-term random walk, and hence large short-term diffusivity.



6. Long-Term Diffusivity: However, increased density far from the bid indicates less impact from the market orders relative to the relaxation time of the Poisson distribution, and thus a lower long-time diffusivity.
7. Qualitative Behavior of the Bid-centered Density: The qualitative behavior of the bid-centered profile is the same as that of the mid-centered profile, and this is expected because the spread distribution is stationary rather than diffusive.
8. Consequences of Stationary Spread: In other words, the only way that the diffusion of the bid or the ask can differ from that of the mid is for the spread to either increase or decrease for several succeeding steps.
9. Autocorrelations of the Spread: Such autocorrelations of the spread cannot accumulate with time if the spread itself is to have a stationary distribution.
10. Gap between Mid and Bid/Ask: Thus, the shift in the mid over some time interval can only differ from that of the bid or the ask by at most a constant, as a result of a few correlated changes in the spread.
11. Diffusivity at Long Times: This result cannot grow with time, however, and so does not affect the diffusivity at long times.
12. Comparisons with Simulations: Indeed, both these predicted corrections to the classical estimate for the diffusivity are seen in the simulation results from midpoint diffusion.
13. Autocorrelation Impact on Diffusivity: The simulation results, however, show that the implied autocorrelations change the diffusivity by factors of  $\sqrt{\epsilon}$ , suggesting that these corrections require a more subtle derivation than the one attempted here.
14. Source Term in Density Coordinates: This will be evidence by the difficulty of obtaining a source term  $S$  in density coordinates that satisfies both the global order conservation law and the proper zero price boundary conditions in the mid-centered frame.
15. Bid-centered Mean Field Approximation: An interesting speculation is that the subtlety of these correlations also causes the density in bi-centered coordinates not to approximate the mean-field conditions at any of the parameters studied here.



16. Relation between the Term Diffusivities: Since short-term and long-term diffusivities are related by a hard constraint, the complexity in producing a late-term density profile should match that of the early term profile.
17. Convenience of Mid-centered Representation: The mid-centered profile is potentially easier, in that the late-time complexity must be matched with that of the early time density profile and scaling of the expected spread.
18. Complexity in Spread Scaling: It appears that the complex scaling is absorbed in the spread, leaving a density that can approximately be calculated with the methods used here.

## Master Equations and Mean Field Approximations

1. Limits Guiding Configuration Simplicity: There are two natural limits in which functional configurations become simple enough to be tractable probabilistically with analytical methods.
2. Independence of  $N(p, t)/p(N, t)$  Fluctuations: They correspond to the mean-field theories in which the fluctuations of the dual differentials of either  $N(p, t)$  and  $p(N, t)$  are important.
3. Consequence of the Fluctuations' Independence: In the first case, probabilities may be defined for any density  $n(p, t)$  independently at each  $p$ , and in the second for the separation intervals  $x(N, t)$  at each  $N$ .
4. MFT for  $n(p, t)$ : The MFT from these approximations is treated in these subsequent sections.
5. Mid-centering for  $n(p, t)$ : Because the fluctuation independence approximation is only usable in a mid-centered frame,  $n(p, t)$  will always refer to this frame.
6. MFT for  $x(N, t)$ :  $x(N, t)$  is well-defined without reference to any frame.



## Master Equations and Mean Field Approximations – A Number Density Equation

1. Density of the Distribution  $\pi(n, p, t)$ : If share-number fluctuations are independent at different  $p$ , a density  $\pi(n, p, t)$  may be defined, which gives the probabilities of finding  $n$  orders in the bin  $(p, p + \varepsilon_p)$  at time  $t$ .
2. Normalization across Price Bins/Times: The normalization condition defining  $\pi$  as a probability density

$$\sum_n \pi(n, p, t) = 1$$

for each bin index  $p$  and at every  $t$ .

3. The Density Evolution Master Equation: Supposing an arbitrary density of the order-book configurations  $\pi(n, p, t)$  at time  $t$ , the stochastic dynamics of the configurations causes probability to be redistributed according to the master equation

$$\begin{aligned} \frac{\partial \pi(n, p, t)}{\partial t} = & \frac{\alpha(p)\varepsilon_p}{\sigma} [\pi(n - \sigma, p, t) - \pi(n, p, t)] \\ & + \frac{\delta}{\sigma} [(n + \sigma)\pi(n + \sigma, p, t) - n\pi(n, p, t)] \\ & + \frac{\mu(p)}{2\sigma} [\pi(n + \sigma, p, t) - \pi(n, p, t)] \\ & + \sum_{\Delta p} P_+(\Delta p) [\pi(n, p - \Delta p, t) - \pi(n, p, t)] \\ & + \sum_{\Delta p} P_-(\Delta p) [\pi(n, p + \Delta p, t) - \pi(n, p, t)] \end{aligned}$$



4. Discretization of the Density Evolution: Here  $\frac{\partial \pi(n,p,t)}{\partial t}$  is a continuum for  $\frac{\pi(n,p,t+\Delta t) - \pi(n,p,t)}{\Delta t}$  where  $\Delta t$  is an elementary timestep chosen short enough that at most one event alters any typical configuration.
5. Balance between Additions and Removals: The master equation represents a general balance between additions and removals without regard to the meaning of  $n$ .
6. Frame Induced Definition of  $\alpha(p)$ : Thus,  $\alpha(p)$  is a function that must be determined self-consistently with the choice of frame.
7.  $\alpha(p)$  in a Bid-centered Frame: As an example of how this works, in a bid-centered frame,  $\alpha(p)$  takes a fixed value  $\alpha(\infty)$  at all  $p$ , because the deposition rate is independent of position and frame shift.
8.  $\alpha(p)$  in Mid-centered Frame: The mid-centered frame is more complicated, because depositions below the midpoint can cause shifts that deposit orders above the midpoint. The specific consequence for  $\alpha(p)$  in this case will be considered below.
9.  $\mu(p)$  in Mid-centered Frame:  $\frac{\mu(p)}{2}$  is, similarly, the rate of market orders surviving to cancel limit orders at price  $p$ .
10. Price Range of  $\mu(p)$  Values:  $\frac{\mu(p)}{2}$  decreases from  $\frac{\mu(0)}{2}$  at the ask – for buy orders, and because  $\mu(p)$  total orders are divided evenly between buys and sells – to zero as

$$p \rightarrow \infty$$

are screened probabilistically by intervening limit orders.

11. Time Parameters for the Simulation:  $\alpha(\infty)$  and  $\mu(0)$  are, thus, the parameters  $\alpha$  and  $\mu$  of the simulation.
12. Components of the Master Equation: The lines of



$$\begin{aligned}
\frac{\partial \pi(n, p, t)}{\partial t} = & \frac{\alpha(p)\varepsilon_p}{\sigma} [\pi(n - \sigma, p, t) - \pi(n, p, t)] \\
& + \frac{\delta}{\sigma} [(n + \sigma)\pi(n + \sigma, p, t) - n\pi(n, p, t)] \\
& + \frac{\mu(p)}{2\sigma} [\pi(n + \sigma, p, t) - \pi(n, p, t)] \\
& + \sum_{\Delta p} P_+(\Delta p) [\pi(n, p - \Delta p, t) - \pi(n, p, t)] \\
& + \sum_{\Delta p} P_-(\Delta p) [\pi(n, p + \Delta p, t) - \pi(n, p, t)]
\end{aligned}$$

correspond to the following events.

13. Limit Order Arrival Component: The term proportional to  $\frac{\alpha(p)\varepsilon_p}{\sigma}$  describes depositions of discrete orders at that rate – because  $\alpha(p)$  is expressed in *shares per price per unit time*, which raise configurations from  $n - \sigma$  to  $n$  shares at price  $p$ .
14. Order Cancelation/Removal Component: The term proportional to  $\delta$  comes from deletions and has the opposite effect, and is proportional to  $\frac{n}{\sigma}$ , the number of *orders* that can independently decay.
15. Market Order Arrival Component: The term proportional to  $\frac{\mu(p)}{2\sigma}$  describes market order annihilations.
16. Up/Down Frame Price Shifts: For general configurations, the preceding three effects may lead to shifts of the origin by arbitrary intervals  $\Delta p$ , and  $P_{\pm}$  for the moment unknown distributions over the frequency of these shifts.
17. Determination of the Price Shifts: They must be determined self-consistently with the configuration of the book which emerges from any solution to



$$\begin{aligned}
\frac{\partial \pi(n, p, t)}{\partial t} = & \frac{\alpha(p)\varepsilon_p}{\sigma} [\pi(n - \sigma, p, t) - \pi(n, p, t)] \\
& + \frac{\delta}{\sigma} [(n + \sigma)\pi(n + \sigma, p, t) - n\pi(n, p, t)] \\
& + \frac{\mu(p)}{2\sigma} [\pi(n + \sigma, p, t) - \pi(n, p, t)] \\
& + \sum_{\Delta p} P_+(\Delta p) [\pi(n, p - \Delta p, t) - \pi(n, p, t)] \\
& + \sum_{\Delta p} P_-(\Delta p) [\pi(n, p + \Delta p, t) - \pi(n, p, t)]
\end{aligned}$$

18. Price Shift Independence from Depth: A limitation of the simple product representation of frame shifts is that it assumes that whole order-book configurations are transported under

$$p \pm \Delta p \rightarrow p$$

independently of the value of  $n(p)$ .

19. Assumption Validity Outside the Spread: As long as the fluctuations are independent, this is a good approximation for orders at all  $p$  which are either the bid or the ask, either before or after the event that causes the shift.
20. Assumption Validity Inside the Spread: The correlations are never ignorable for the bins which are the bid and the ask, though, there is some distribution of instances in which any  $p$  of interest plays those parts.
21. Rigorous Inside-Spread Density Handling: Approximate methods to incorporate those correlations will require replacing the product with a sum of products conditioned on states of the order book, as will be derived below.
22. Order flow vs Consistent Parameterization: The important point is that the order-flow dependence of



$$\begin{aligned}
\frac{\partial \pi(n, p, t)}{\partial t} = & \frac{\alpha(p)\varepsilon_p}{\sigma} [\pi(n - \sigma, p, t) - \pi(n, p, t)] \\
& + \frac{\delta}{\sigma} [(n + \sigma)\pi(n + \sigma, p, t) - n\pi(n, p, t)] \\
& + \frac{\mu(p)}{2\sigma} [\pi(n + \sigma, p, t) - \pi(n, p, t)] \\
& + \sum_{\Delta p} P_+(\Delta p) [\pi(n, p - \Delta p, t) - \pi(n, p, t)] \\
& + \sum_{\Delta p} P_-(\Delta p) [\pi(n, p + \Delta p, t) - \pi(n, p, t)]
\end{aligned}$$

is independent of these self-consistency requirements, and may be solved by the use of generating functionals at general  $\alpha(p)$ ,  $\mu(p)$ , and  $P_{\pm}$ .

23. Formulation of the Exact Solution: The solution, exact but not analytically tractable at general  $\varepsilon_p$ , will be derived in closed form in the next subsection.
24. Special Case of  $\varepsilon_p \rightarrow 0$ : It has a well-behaved continuum limit at

$$\varepsilon_p \rightarrow 0$$

however, which analytically tractable, and so that special case will be considered in the following subsection.

## Theoretical Analysis – Master Equations and Mean-Field Approximations: Solution by Generating Functional

1. MGF Functional for  $\pi(n, p)$ : The MGF for  $\pi(n, p)$  is defined for a parameter

$$\lambda \in [0, 1]$$

as



$$\Pi(\lambda, p) = \sum_{\substack{n=0 \\ \sigma}}^{\infty} \lambda^{\sigma} \pi(n, p)$$

2. MGF at  $\lambda = 0$  and  $\lambda = 1$ : Introducing a shorthand for its value at

$$\lambda = 0$$

$$\Pi(0, p) = \pi(0, p) = \pi_0(p)$$

while the normalization condition

$$\sum_n \pi(n, p, t) = 1$$

for probabilities gives

$$\Pi(1, p) = 1 \quad \forall p$$

3. Linear Approximation of MGF across  $\lambda$ : By definition of the average  $n(p)$  in the distribution  $\pi$  denoted  $\langle n(p) \rangle$

$$\left| \frac{\partial \Pi(\lambda, p)}{\partial \lambda} \right|_{\lambda=1} = \frac{\langle n(p) \rangle}{\sigma}$$

and because  $\Pi(\lambda, p)$  will be regular in sufficiently small neighborhood of

$$\lambda = 1$$

one can expand



$$\Pi(\lambda, p) = 1 + (\lambda - 1) \frac{\langle n(p) \rangle}{\sigma} + \mathcal{O}([\lambda - 1]^2)$$

4. Stationary Master Equation for the MGF: Multiplying

$$\begin{aligned} \frac{\partial \pi(n, p, t)}{\partial t} &= \frac{\alpha(p)\varepsilon_p}{\sigma} [\pi(n - \sigma, p, t) - \pi(n, p, t)] \\ &\quad + \frac{\delta}{\sigma} [(n + \sigma)\pi(n + \sigma, p, t) - n\pi(n, p, t)] \\ &\quad + \frac{\mu(p)}{2\sigma} [\pi(n + \sigma, p, t) - \pi(n, p, t)] \\ &\quad + \sum_{\Delta p} P_+(\Delta p) [\pi(n, p - \Delta p, t) - \pi(n, p, t)] \\ &\quad + \sum_{\Delta p} P_-(\Delta p) [\pi(n, p + \Delta p, t) - \pi(n, p, t)] \end{aligned}$$

by  $\lambda^{\frac{n}{\sigma}}$  and summing over  $n$ , the stationary solution for  $\Pi(\lambda, p)$  must satisfy

$$\begin{aligned} 0 &= \frac{\lambda - 1}{p} \left| \alpha(p)\varepsilon_p \Pi(\lambda, p) - \delta\sigma \frac{\partial \Pi(\lambda, p)}{\partial \lambda} - \frac{\mu(p)}{2\lambda} [\Pi(\lambda, p) - \pi_0(p)] \right|_{\lambda, p} \\ &\quad + \sum_{\Delta p} P_+(\Delta p) [\Pi(\lambda, p - \Delta p) - \Pi(\lambda, p)] \\ &\quad + \sum_{\Delta p} P_-(\Delta p) [\Pi(\lambda, p + \Delta p) - \Pi(\lambda, p)] \end{aligned}$$

5. Symmetric Shift with No Drift: Only the symmetric case with no net drift will be considered here for simplicity, which requires

$$P_+(\Delta p) = P_-(\Delta p) = P(\Delta p)$$



6. Expression for  $\Delta p$  Based Diffusivity: In a Fokker-Planck expression, the unrenormalized diffusivity of whatever reference price is used as coordinate origin, is related to the distribution  $P$  by

$$D = \sum_{\Delta p} P(\Delta p) \Delta p^2$$

7. Rate of Frame Density Shift: The rate at which the shift events happen is

$$R = \sum_{\Delta p} P(\Delta p)$$

and the mean shift amount appearing at linear rate in derivatives – relevant at

$$p \rightarrow 0$$

is

$$\langle \Delta p \rangle = \frac{\sum_{\Delta p} P(\Delta p) \Delta p}{\sum_{\Delta p} P(\Delta p)}$$

8. Frame Diffusivity Stationary Master Equation: Anywhere in the interior if the price range – where  $p$  is not at any stage in the bid, the ask, or a point inside the spread

$$\begin{aligned} 0 = & \frac{\lambda - 1}{p} \left| \alpha(p) \varepsilon_p \Pi(\lambda, p) - \delta \sigma \frac{\partial \Pi(\lambda, p)}{\partial \lambda} - \frac{\mu(p)}{2\lambda} [\Pi(\lambda, p) - \pi_0(p)] \right|_{\lambda, p} \\ & + \sum_{\Delta p} P_+(\Delta p) [\Pi(\lambda, p - \Delta p) - \Pi(\lambda, p)] \\ & + \sum_{\Delta p} P_-(\Delta p) [\Pi(\lambda, p + \Delta p) - \Pi(\lambda, p)] \end{aligned}$$



may be written

$$\frac{\partial \Pi(\lambda, p)}{\partial \lambda} - \frac{D}{\delta(\lambda - 1)} \frac{\partial^2 \Pi(\lambda, p)}{\partial p^2} - \frac{\alpha(p)\varepsilon_p - \frac{\mu(p)}{2\lambda}}{\delta\sigma} \Pi(\lambda, p) = \frac{\mu(p)}{2\delta\sigma\lambda} \pi_0(p)$$

9. Linear MGF Approximation at  $\lambda \rightarrow 1$ : Evaluated at

$$\lambda \rightarrow 1$$

with the use of the expression

$$\Pi(\lambda, p) = 1 + (\lambda - 1) \frac{\langle n(p) \rangle}{\sigma} + \mathcal{O}([\lambda - 1]^2)$$

this becomes

$$\langle n(p) \rangle - \frac{D}{\delta} \frac{\partial^2 \langle n(p) \rangle}{\partial p^2} = \frac{\alpha(p)\varepsilon_p}{\delta} - \frac{\mu(p)}{2\delta} [1 - \pi_0(p)]$$

10. Simplification under  $\varepsilon_p \rightarrow 0$ : At this point, it is convenient to specialize to the case

$$\varepsilon_p \rightarrow 0$$

wherein the eligible values of any  $\langle n(p) \rangle$  become just  $\sigma$  and 0.

11. Elimination of  $\pi_0(p)$  using  $\langle n(p) \rangle$ : The expectation is then related to the probability of zero occupancy at each  $p$  is

$$\langle n(p) \rangle = \sigma[1 - \pi_0(p)]$$



yielding immediately

$$\frac{\alpha(p)\varepsilon_p}{\delta} = \frac{\mu(p)}{2\delta} \langle n(p) \rangle + \langle n(p) \rangle - \frac{D}{\delta} \frac{\partial^2 \langle n(p) \rangle}{\partial p^2}$$

12. Stationary Solution  $\langle n(p) \rangle$  at  $\varepsilon_p \rightarrow 0$ : The above equation defines the general equation for  $\langle n(p) \rangle$  from the master equation

$$\begin{aligned} \frac{\partial \pi(n, p, t)}{\partial t} &= \frac{\alpha(p)\varepsilon_p}{\sigma} [\pi(n - \sigma, p, t) - \pi(n, p, t)] \\ &\quad + \frac{\delta}{\sigma} [(n + \sigma)\pi(n + \sigma, p, t) - n\pi(n, p, t)] \\ &\quad + \frac{\mu(p)}{2\sigma} [\pi(n + \sigma, p, t) - \pi(n, p, t)] \\ &\quad + \sum_{\Delta p} P_+(\Delta p) [\pi(n, p - \Delta p, t) - \pi(n, p, t)] \\ &\quad + \sum_{\Delta p} P_-(\Delta p) [\pi(n, p + \Delta p, t) - \pi(n, p, t)] \end{aligned}$$

in the continuum limit

$$\frac{2\alpha(p)\varepsilon_p}{\mu(p)} \rightarrow 0$$

13. Diffusive Term for Shift Distribution: The shift distribution  $p(\Delta p)$  appears only through the diffusion  $D$ , which must be solved self-consistently, along with otherwise arbitrary functions  $\alpha$  and  $\mu$ .
14. Generalized Solution for  $\varepsilon_p$ : A more general solution for larger  $\varepsilon_p$  is carried out in the next two sections.
15. Nondimensional Form of Stationary Master: A first step toward nondimensionalization may be taken by recasting



$$\frac{\alpha(p)\varepsilon_p}{\delta} = \frac{\mu(p)}{2\delta} \langle n(p) \rangle + \langle n(p) \rangle - \frac{D}{\delta} \frac{\partial^2 \langle n(p) \rangle}{\partial p^2}$$

to the form

$$\frac{\alpha(p)}{\alpha(\infty)} = \left[ \frac{\mu(p)}{\mu(0)} + \epsilon \left( 1 - \frac{D}{\delta} \frac{\partial^2}{\partial p^2} \right) \right] \frac{1}{\epsilon} \frac{\delta \langle n(p) \rangle}{\alpha(p) \varepsilon_p}$$

16.  $\langle n(p) \rangle$  Far from the Midpoint #1: For far from the midpoint, where only depositions and cancellations take place, orders in bins of width  $\varepsilon_p$  are Poisson-distributed with mean  $\frac{\alpha(\infty)\varepsilon_p}{\delta}$ .

17.  $\langle n(p) \rangle$  Far from the Midpoint #2: Thus, the asymptotic value of  $\frac{\delta \langle n(p) \rangle}{\alpha(\infty)\varepsilon_p}$  at large  $p$  is unity.

18. Asymptotic  $\alpha(p)$  and  $\mu(p)$  at  $p \rightarrow \infty$ : This is consistent with a limit  $\frac{\alpha(p)}{\alpha(\infty)}$  of unity, and a limit for the second  $\frac{\mu(p)}{\mu(0)}$  to zero.

19. Rationale for the  $\frac{\alpha(p)}{\alpha(\infty)}$  Expression above: The reason for grouping the nondimensionalized number density with  $\frac{1}{\epsilon}$ , together with the proper normalization of the characteristic price scale, will come from examining the decay of the dimensionless function  $\frac{\mu(p)}{\mu(0)}$ .

## Supporting Calculations in Density Coordinates

1. Master Solution in Density Coordinates: The following two subsections provide details for the master equation solution in density coordinates.



2. Generalized MGF and Calculation Sources: The first provides the generating functional solution for the density functional at general  $\varepsilon_p$ , and the second the approximate source from correlated fluctuations.

## Supporting Calculations in Density Coordinates – Generating Functional at General Bin Width

1.  $\alpha/\mu$  as Functions of  $p$ : As in the treatment above,  $\alpha$  and  $\mu$  represent the functions of  $p$  everywhere in this section, because the boundary values do not propagate globally.
2.  $\frac{D}{\delta}$  Power Series for  $\Pi$ :

$$\frac{\partial \Pi(\lambda, p)}{\partial \lambda} - \frac{D}{\delta(\lambda - 1)} \frac{\partial^2 \Pi(\lambda, p)}{\partial p^2} - \frac{\alpha(p)\varepsilon_p - \frac{\mu(p)}{2\lambda}}{\delta\sigma} \Pi(\lambda, p) = \frac{\mu(p)}{2\delta\sigma\lambda} \pi_0(p)$$

can be solved by assuming there is a convergent expansion in  $\frac{D}{\delta}$

$$\Pi(\lambda, p) = \sum_j \left(\frac{D}{\delta}\right)^j \Pi_j(\lambda, p)$$

and it is convenient to embellish the short-hand notation as well, with

$$\Pi_j(0, p) = \pi_{0j}(p)$$

3.  $\frac{D}{\delta}$  Power Series for  $\langle n(p) \rangle$ : It follows that the expected number expands as

$$\langle n(p) \rangle = \sum_j \left(\frac{D}{\delta}\right)^j \langle n(p) \rangle_j$$



4. Collecting the Power Series Coefficients: Order by order in  $\frac{D}{\delta}$

$$\frac{\partial \Pi(\lambda, p)}{\partial \lambda} - \frac{D}{\delta(\lambda - 1)} \frac{\partial^2 \Pi(\lambda, p)}{\partial p^2} - \frac{\alpha(p)\varepsilon_p - \frac{\mu(p)}{2\lambda}}{\delta\sigma} \Pi(\lambda, p) = \frac{\mu(p)}{2\delta\sigma\lambda} \pi_0(p)$$

requires

$$\left[ \frac{\partial}{\partial \lambda} - \frac{\alpha(p)\varepsilon_p - \frac{\mu(p)}{2\lambda}}{\delta\sigma} \right] \Pi_j(\lambda, p) = \frac{\mu(p)}{2\delta\sigma\lambda} \pi_{0j}(p) + \frac{\partial^2}{\partial p^2} \frac{\Pi_{j-1}(\lambda, p)}{(\lambda - 1)}$$

5. Normalization Constraints on  $\Pi_j(1, p)$ : Because  $\Pi_j(\lambda, p)$  have been introduced in order to be chosen homogeneous of degree zero in  $\frac{D}{\delta}$ , the normalization condition requires that

$$\Pi_0(1, p) = 1$$

$$\Pi_{j \neq 0}(\lambda, p) = 0 \quad \forall p$$

6. Recursion Relation for  $\langle n(p) \rangle_j$ : The implied recursion relations for expected occupation numbers are

$$\langle n(p) \rangle_0 = \frac{\alpha(p)\varepsilon_p}{\delta} - \frac{\mu(p)}{2\delta} [1 - \pi_{00}]$$

at

$$j = 0$$



and

$$\langle n(p) \rangle_j = \frac{\mu(p)}{2\delta} \pi_{0j} + \frac{\partial^2}{\partial p^2} \langle n(p) \rangle_{j-1}$$

otherwise.

#### 7. Integral Solution to the $\Pi_j(\lambda, p)$ Series:

$$\left[ \frac{\partial}{\partial \lambda} - \frac{\alpha(p)\varepsilon_p - \frac{\mu(p)}{2\lambda}}{\delta\sigma} \right] \Pi_j(\lambda, p) = \frac{\mu(p)}{2\delta\sigma\lambda} \pi_{0j}(p) + \frac{\partial^2}{\partial p^2} \frac{\Pi_{j-1}(\lambda, p)}{(\lambda - 1)}$$

is solved by the use of an integrating factor, to give the recursive integral relation

$$\Pi_j(\lambda, p) = \pi_{0j}(p) \left[ 1 + \frac{\alpha(p)\varepsilon_p}{\delta\sigma} \mathcal{I}(\lambda) \right] + \mathcal{I}(\lambda) \left[ \frac{\partial^2}{\partial p^2} \frac{\Pi_{j-1}(\lambda, p)}{(\lambda - 1)} \right]_\lambda$$

where

$$\mathcal{I}(\lambda) = \lambda \int_0^1 e^{\lambda \frac{\alpha(p)\varepsilon_p}{\delta\sigma} (1-z)} z^{\frac{\mu(p)}{2\delta\sigma}} dz$$

and

$$\left[ \frac{\partial^2}{\partial p^2} \frac{\Pi_{j-1}(\lambda, p)}{(\lambda - 1)} \right]_\lambda \equiv \frac{\lambda}{\mathcal{I}(\lambda)} \cdot \int_0^1 e^{\lambda \frac{\alpha(p)\varepsilon_p}{\delta\sigma} (1-z)} z^{\frac{\mu(p)}{2\delta\sigma}} \frac{\partial^2}{\partial p^2} \frac{\Pi_{j-1}(\lambda z, p)}{(\lambda z - 1)} dz$$

#### 8. $\Pi(\lambda, p)$ Normalizer as a Starting Point: The surface condition

$$\Pi_0(1, p) = 1$$



$$\Pi_{j \neq 0}(\lambda, p) = 0 \quad \forall p$$

provided the starting point for this recursion, by giving at

$$j = 0$$

$$\pi_{00} = \frac{1}{1 + \frac{\alpha(p)\varepsilon_p}{\delta\sigma} J(1)}$$

9. Recursion Sequence for  $\langle n(p) \rangle_j$ : Given forms for  $\alpha(p)$  and  $\mu(p)$

$$\langle n(p) \rangle_0 = \frac{\alpha(p)\varepsilon_p}{\delta} - \frac{\mu(p)}{2\delta} [1 - \pi_{00}]$$

may be solved directly for

$$\pi_{00} = \frac{1}{1 + \frac{\alpha(p)\varepsilon_p}{\delta\sigma} J(1)}$$

and extended by

$$\langle n(p) \rangle_j = \frac{\mu(p)}{2\delta} \pi_{0j} + \frac{\partial^2}{\partial p^2} \langle n(p) \rangle_{j-1}$$

to solve for  $\langle n(p) \rangle$

10. Recursion Sequence for  $\Pi_j(\lambda, p)$ : More generally



$$\mathcal{I}(\lambda) = \lambda \int_0^1 e^{\lambda \frac{\alpha(p)\varepsilon_p}{\delta\sigma}(1-z)} z^{\frac{\mu(p)}{2\delta\sigma}} dz$$

$$\left[ \left[ \frac{\partial^2}{\partial p^2} \frac{\Pi_{j-1}(\lambda, p)}{(\lambda - 1)} \right] \right]_\lambda \equiv \frac{\lambda}{\mathcal{I}(\lambda)} \cdot \int_0^1 e^{\lambda \frac{\alpha(p)\varepsilon_p}{\delta\sigma}(1-z)} z^{\frac{\mu(p)}{2\delta\sigma}} \frac{\partial^2}{\partial p^2} \frac{\Pi_{j-1}(\lambda z, p)}{(\lambda z - 1)} dz$$

and

$$\pi_{00} = \frac{1}{1 + \frac{\alpha(p)\varepsilon_p}{\delta\sigma} \mathcal{I}(1)}$$

may be solved to any desired order numerically, to obtain the fluctuation characteristics of  $n(p)$ .

11. Incorporating Self consistent  $\alpha(p)/\mu(p)$  Representations: Finding a solution, however, becomes difficult when  $\alpha(p)$  and  $\mu(p)$  must be self-consistently related to the solutions for  $\Pi(\lambda, p)$ .
12. Simplification available when  $\varepsilon_p \rightarrow 0$ : The special case

$$\varepsilon_p \rightarrow 0$$

admits a drastic simplification, in which the whole expansion for  $\langle n(p) \rangle$  may be directly summed, to recover the result in the main treatment.

13. ODE for  $\langle n(p) \rangle$ : In this limit, one gets a single differential equation in  $p$  which is solvable by numerical integration.
14. Price Continuum for  $\langle n(p) \rangle$ : The existence and the regularity of this solution demonstrates the presence of a continuum limit on the price space, and can be simulated directly by allowing orders to be placed at arbitrary real-valued prices.



## Supporting Calculations in Density Coordinates – Generating Functional at General Bin Width: Recovering the Continuum Limit for Prices

1.  $\mathcal{I}(\lambda)$  and  $\langle n(p) \rangle_0$  as  $\varepsilon_p \rightarrow 0$ : In the limit that the dimensionless quantity

$$\frac{\alpha(p)\varepsilon_p}{\delta\sigma} \rightarrow 0$$

$$\mathcal{I}(\lambda) = \lambda \int_0^1 e^{\lambda \frac{\alpha(p)\varepsilon_p}{\delta\sigma}(1-z)} z^{\frac{\mu(p)}{2\delta\sigma}} dz$$

simplifies to

$$\mathcal{I}(\lambda) \rightarrow \frac{\lambda}{1 + \frac{\alpha(p)}{2\delta\sigma}} + \mathcal{O}(\varepsilon_p)$$

from which it follows that

$$\langle n(p) \rangle_0 \rightarrow \frac{\frac{\alpha(p)\varepsilon_p}{\delta\sigma}}{1 + \frac{\mu(p)}{2\delta\sigma}} + \mathcal{O}(\varepsilon_p^2)$$

2. Linear MGF Postulate for  $\Pi_0(\lambda, p)$ : The important simplification given vanishing  $\varepsilon_p$ , as will be seen below, is that the expansion

$$\Pi(\lambda, p) = 1 + (\lambda - 1) \frac{\langle n(p) \rangle}{\sigma} + \mathcal{O}([\lambda - 1]^2)$$

collapses, at leading order in  $\varepsilon_p$  to



$$\Pi_0(\lambda, p) = 1 + (\lambda - 1) \frac{\langle n(p) \rangle_0}{\sigma} + \mathcal{O}(\varepsilon_p^2)$$

3. Inductive Recursive Relation for  $\Pi_j(\lambda, p)$ : The above equation is used as an input to an inductive hypothesis

$$\Pi_{j-1}(\lambda, p) = \Pi_{j-1}(1, p) + (\lambda - 1) \frac{\langle n(p) \rangle_{j-1}}{\sigma} + \mathcal{O}(\varepsilon_p^2)$$

n.b.

$$\langle n(p) \rangle_{j-1} \sim \mathcal{O}(\varepsilon_p)$$

$$\Pi_{j-1}(1, p) = \text{either } 1 \text{ or } 0$$

which, with

$$\left[ \left[ \frac{\partial^2}{\partial p^2} \frac{\Pi_{j-1}(\lambda, p)}{(\lambda - 1)} \right] \right]_\lambda \equiv \frac{\lambda}{J(\lambda)} \cdot \int_0^1 e^{\lambda \frac{\alpha(p)\varepsilon_p}{\delta\sigma}(1-z)} z^{\frac{\mu(p)}{2\delta\sigma}} \frac{\partial^2}{\partial p^2} \frac{\Pi_{j-1}(\lambda z, p)}{(\lambda z - 1)} dz$$

then recovers the condition at  $j$ :

$$\Pi_j(\lambda, p) = (\lambda - 1) J(1) \frac{\partial^2}{\partial p^2} \frac{\langle n(p) \rangle_{j-1}}{\sigma} + \mathcal{O}(\varepsilon_p^2)$$

4.  $\langle n(p) \rangle_j$  in Terms of  $\langle n(p) \rangle_{j-1}$ : Using

$$\Pi(\lambda, p) = 1 + (\lambda - 1) \frac{\langle n(p) \rangle}{\sigma} + \mathcal{O}([\lambda - 1]^2)$$



at

$$\lambda \rightarrow 1$$

and

$$\mathcal{I}(\lambda) \rightarrow \frac{\lambda}{1 + \frac{\alpha(p)}{2\delta\sigma}} + \mathcal{O}(\varepsilon_p)$$

for  $\mathcal{I}(\lambda)$  gives the recursion for the number density

$$\langle n(p) \rangle_{j \neq 0} \rightarrow \frac{1}{1 + \frac{\mu(p)}{2\delta\sigma}} \frac{\partial^2}{\partial p^2} \frac{\langle n(p) \rangle_{j-1}}{\sigma} + \mathcal{O}(\varepsilon_p^2)$$

5. Series Summation for  $\langle n(p) \rangle$  #1: The sum

$$\langle n(p) \rangle = \sum_j \left( \frac{D}{\delta} \right)^j \langle n(p) \rangle_j$$

is then

$$\langle n(p) \rangle = \sum_j \left[ \frac{D}{\delta} \frac{1}{1 + \frac{\mu(p)}{2\delta\sigma}} \frac{\partial^2}{\partial p^2} \right]^j \langle n(p) \rangle_0$$

6. Series Summation for  $\langle n(p) \rangle$  #2: Using



$$\langle n(p) \rangle_0 \rightarrow \frac{\frac{\alpha(p)\varepsilon_p}{\delta\sigma}}{1 + \frac{\mu(p)}{2\delta\sigma}} + \mathcal{O}(\varepsilon_p^2)$$

and re-arranging terms

$$\langle n(p) \rangle = \frac{1}{1 + \frac{\mu(p)}{2\delta\sigma}} \sum_j \left[ \frac{D}{\delta} \frac{\partial^2}{\partial p^2} + \frac{1}{1 + \frac{\mu(p)}{2\delta\sigma}} \right]^j \frac{\alpha(p)\varepsilon_p}{\delta}$$

7. Series Expansion in Price Laplacian: The series expansion in the price Laplacian is formally the geometric sum

$$\left[ 1 + \frac{\mu(p)}{2\delta\sigma} \right] \langle n(p) \rangle = \left[ 1 - \frac{D}{\delta} \frac{\partial^2}{\partial p^2} \frac{1}{1 + \frac{\mu(p)}{2\delta\sigma}} \right]^{-1} \frac{\alpha(p)\varepsilon_p}{\delta}$$

which can be inverted to give

$$\frac{\alpha(p)\varepsilon_p}{\delta} = \left[ \frac{\mu(p)}{2\delta\sigma} + \left( 1 - \frac{D}{\delta} \frac{\partial^2}{\partial p^2} \right) \right] \langle n(p) \rangle$$

a relation that is local in derivatives.

## Theoretical Analysis - Master Equation and Mean-field Approximations: Screening of the Market Rate

1. Market Orders Screening from  $\langle n(p) \rangle$ : In the context of independent fluctuations

$$\langle n(p) \rangle = \sigma [1 - \pi_0(p)]$$



implies a relation between the mean density and the rate at which the market orders are screened as price increases.

2. Exhausting Limit Orders before  $p$ : The effect of a limit order, resident in the price bin  $p$  when a market order survives to reach that bin, is to prevent the order arriving at  $p + \varepsilon_p$ .
3. Effective  $\mu(p)$  Increment at  $p$ : Though the nature of the shift induced, when such annihilation occurs, depends on the co-moving frame being modeled, the change in the number of orders surviving is independent of the frame, and is given by

$$\Delta\mu(p) = -\mu(p)[1 - \pi_0(p)] = -\frac{\mu(p)\langle n(p) \rangle}{\sigma}$$

4. Differential Form for  $\mu(p)$ : The above expression may be re-written

$$\frac{d \log \frac{\mu(p)}{\mu(0)}}{dp} = -\frac{1}{\epsilon} \left[ \frac{2\alpha(\infty)}{\mu(0)} \right] \left[ \frac{\delta\langle n(p) \rangle}{\alpha(\infty)\varepsilon_p} \right]$$

identifying the characteristic scale for prices as

$$p_c = \frac{\mu(0)}{2\alpha(\infty)}$$

5. Market Orders Screening Rate Function: Writing

$$\hat{p} \equiv \frac{p}{p_c}$$

the function that screens market orders is the same as the argument of



$$\frac{\alpha(p)}{\alpha(\infty)} = \left[ \frac{\mu(p)}{\mu(0)} + \epsilon \left( 1 - \frac{D}{\delta} \frac{\partial^2}{\partial p^2} \right) \right] \frac{1}{\epsilon} \frac{\delta \langle n(p) \rangle}{\alpha(p) \varepsilon_p}$$

and will be denoted

$$\frac{1}{\epsilon} \frac{\delta \langle n(p) \rangle}{\alpha(p) \varepsilon_p} = \psi(\hat{p})$$

#### 6. $p$ -dependent Forms for $\mu/\alpha$ : Defining a nondimensional diffusivity

$$\beta = \frac{D}{\delta p_c^2}$$

$$\frac{\alpha(p) \varepsilon_p}{\delta} = \left[ \frac{\mu(p)}{2\delta\sigma} + \epsilon \left( 1 - \frac{D}{\delta} \frac{\partial^2}{\partial p^2} \right) \right] \langle n(p) \rangle$$

can be put in the form

$$\frac{\alpha(p)}{\alpha(\infty)} = \left[ \frac{\mu(p)}{\mu(0)} + \epsilon \left( 1 - \beta \frac{\partial^2}{\partial p^2} \right) \right] \psi(\hat{p})$$

with

$$\frac{\mu(p)}{\mu(0)} = \varphi(\hat{p}) = e^{- \int_0^{\hat{p}} \psi(\hat{p}') d\hat{p}'}$$

### Theoretical Analysis - Master Equation and Mean-field Approximations: Verifying the Conservation Laws



1. Built-in Adherence to Conservation Laws: Since nothing about the derivation so far has made explicit use of the frame in which  $n(p)$  is averaged, the combination of

$$\frac{\alpha(p)}{\alpha(\infty)} = \left[ \frac{\mu(p)}{\mu(0)} + \epsilon \left( 1 - \beta \frac{\partial^2}{\partial p^2} \right) \right] \psi(\hat{p})$$

with

$$\frac{\mu(p)}{\mu(0)} = \varphi(\hat{p}) = e^{-\int_0^{\hat{p}} \psi(\hat{p}') d\hat{p}'}$$

respects the conservation laws of

$$\frac{\mu(p)}{2} = \sum_{b+\varepsilon_p}^{\infty} [\alpha(p) - \delta\langle n_b(p) \rangle]$$

and

$$\frac{\mu(p)}{2} = \alpha \frac{\langle s \rangle}{2} + \sum_{b+\varepsilon_p}^{\infty} [\alpha(p) - \delta\langle n_m(p) \rangle]$$

if appropriate forms are chosen for the deposition rate  $\alpha(p)$ .

2.  $\alpha(p)$  in Bid-centered Frame: For example, in the bid-centered frame

$$\frac{\alpha(p)}{\alpha(\infty)} = 1$$

everywhere.

3. Nondimensional Bid-centered Flow Conservation: Multiplying



$$\frac{\alpha(p)}{\alpha(\infty)} = \left[ \frac{\mu(p)}{\mu(0)} + \epsilon \left( 1 - \beta \frac{\partial^2}{\partial p^2} \right) \right] \psi(\hat{p})$$

by  $d\hat{p}$  and integrating over the whole range from the bid to  $\pm\infty$ , one recovers the nondimensionalized form of

$$\frac{\mu(p)}{2} = \sum_{b+\varepsilon_p}^{\infty} [\alpha(p) - \delta \langle n_b(p) \rangle]$$

$$\int_0^{\infty} [1 - \epsilon \psi(\hat{p})] d\hat{p} = 1$$

if one is carful with a specific convention.

4. Caveat on the Boundary Condition  $\frac{d\psi(\hat{p})}{d\hat{p}} \Big|_0^\infty$  #1: The integral of the diffusion term formally produces the first derivative  $\frac{d\psi(\hat{p})}{d\hat{p}} \Big|_0^\infty$
5. Caveat on the Boundary Condition  $\frac{d\psi(\hat{p})}{d\hat{p}} \Big|_0^\infty$  #2: This must be regarded as the true first derivative, and its evaluation considered at zero continued far enough below the bid to capture the identically zero fist derivative of the sell order depth profile.
6.  $\alpha(p)$  in the Mid-centered Frame: In the midpoint-centered frame, the correct form for the source term should be

$$\frac{\alpha(\hat{p})}{\alpha(\infty)} = 1 + \mathbb{P} \left[ \frac{\hat{s}}{2} \leq \hat{p} \right]$$

whatever the expression for the CDF of  $\hat{p}$ .

7. Nondimensionalized Mid-centered Frame: Recognizing that the integral of the CDF is, by parts, the mean value of  $\frac{\hat{s}}{2}$ , the same integration of



$$\frac{\alpha(p)}{\alpha(\infty)} = \left[ \frac{\mu(p)}{\mu(0)} + \epsilon \left( 1 - \beta \frac{\partial^2}{\partial p^2} \right) \right] \psi(\hat{p})$$

gives the nondimensionalized form of

$$\frac{\mu(p)}{2} = \alpha \frac{\langle s \rangle}{2} + \sum_{b+\varepsilon_p}^{\infty} [\alpha(p) - \delta \langle n_m(p) \rangle]$$

$$\int_0^{\infty} [1 - \epsilon \psi(\hat{p})] d\hat{p} = 1 - \frac{\langle \hat{s} \rangle}{2}$$

8. Caveat on the Boundary Condition  $\frac{d\psi(\hat{p})}{d\hat{p}} \Big|_0^\infty$  #3: Again, this works only if the surface contribution from integrating the diffusion term vanishes.
9. Assumption of Independent Fluctuations: Neither of these the assumption of independent fluctuations, though that will be used below to give a simple approximate form for

$$\mathbb{P} \left[ \frac{\hat{s}}{2} \leq \hat{p} \right] \approx \varphi(\hat{p})$$

10. Correctness of the Extinction Term: Therefore, they provide a check that the extinction form

$$\frac{\mu(p)}{\mu(0)} = \varphi(\hat{p}) = e^{- \int_0^{\hat{p}} \psi(\hat{p}') d\hat{p}'}$$

propagates market orders correctly into the interior of the order-book distribution to respect global conservation.



11. Plausible Form for  $\alpha(p)$ : They also check the consistency for the intuitively plausible form for  $\alpha(p)$  in the midpoint-centered frame.
12. Validity under Independent Fluctuations Assumption: The detailed form is then justified whenever the assumption of independent fluctuations is checked to be valid.

## Theoretical Analysis - Master Equation and Mean-field Approximations: Self-consistent Parametrization

1. CDF Implied by the Independent Fluctuations: The assumption if independent fluctuations of  $n(p)$  used above to derive the screening of the market orders is equivalent to the specification of the CDF of the ask.
2. Mid centered Market Order Removal: Market orders are only removed between  $p$  and  $p + \varepsilon_p$  in those instances where the ask is  $p$ .
3. Continuum Form for  $\mathbb{P} \left[ \frac{\hat{s}}{2} \leq \hat{p} \right] \approx \varphi(\hat{p})$ : Therefore

$$\mathbb{P} \left[ \frac{\hat{s}}{2} \leq \hat{p} \right] \approx \varphi(\hat{p})$$

is the continuum limit of

$$\mathbb{P} \left[ \frac{\hat{s}}{2} > \hat{p} \right] = 1 - e^{-\sum_{p'=b+\varepsilon_p}^{p-\varepsilon_p} \langle n_m(p') \rangle}$$

4. Mid centered Form for  $\varphi(\hat{p})$ : Together with the form

$$\frac{\alpha(\hat{p})}{\alpha(\infty)} = 1 + \mathbb{P} \left[ \frac{\hat{s}}{2} \leq \hat{p} \right]$$



$$\frac{\alpha(p)}{\alpha(\infty)} = \left[ \frac{\mu(p)}{\mu(0)} + \epsilon \left( 1 - \beta \frac{\partial^2}{\partial p^2} \right) \right] \psi(\hat{p})$$

becomes

$$1 + \varphi(\hat{p}) = - \left[ \frac{d\varphi(\hat{p})}{d\hat{p}} + \epsilon \left( 1 - \beta \frac{\partial^2}{\partial p^2} \right) \frac{d \log \varphi(\hat{p})}{d\hat{p}} \right]$$

5. Bid centered Form for  $\varphi(\hat{p})$ : If the assumption of independent fluctuations were valid in the bid-centered frame, it would take the same form, but with  $\varphi(\hat{p})$  removed on the LHS.
6. Density and  $\psi(\hat{p})$  for Negative Price: To consistently use the diffusion approximation, with the realization that for

$$p = 0$$

$$n\pi(n, p - \Delta p) = 0$$

essentially for all of  $\Delta p$  in

$$\begin{aligned} \frac{\partial \pi(n, p, t)}{\partial t} &= \frac{\alpha(p)\varepsilon_p}{\sigma} [\pi(n - \sigma, p, t) - \pi(n, p, t)] \\ &\quad + \frac{\delta}{\sigma} [(n + \sigma)\pi(n + \sigma, p, t) - n\pi(n, p, t)] \\ &\quad + \frac{\mu(p)}{2\sigma} [\pi(n + \sigma, p, t) - \pi(n, p, t)] \\ &\quad + \sum_{\Delta p} P_+(\Delta p) [\pi(n, p - \Delta p, t) - \pi(n, p, t)] \\ &\quad + \sum_{\Delta p} P_-(\Delta p) [\pi(n, p + \Delta p, t) - \pi(n, p, t)] \end{aligned}$$



it is necessary to set the Fokker-Planck approximation to

$$\psi(0 - \langle \Delta p \rangle) = 0$$

as a boundary condition.

7. Approximating  $\langle \Delta \hat{p} \rangle \frac{d\psi(\hat{p})}{d\hat{p}} \Big|_0$  #1: Nondimensionalized, this gives

$$\frac{\beta}{2} \frac{\partial^2 \psi(\hat{p})}{\partial \hat{p}^2} = \frac{R}{\delta} \left( \langle \Delta \hat{p} \rangle \frac{d}{d\hat{p}} - 1 \right) \psi(\hat{p}) \Big|_0$$

where  $R$  is the rate at which shifts occur

$$R \equiv \sum_{\Delta p} P(\Delta p)$$

8. Approximating  $\langle \Delta \hat{p} \rangle \frac{d\psi(\hat{p})}{d\hat{p}} \Big|_0$  #2: In the solutions below, the curvature will typically by much smaller than

$$\psi(0) \sim 1$$

as it will be convenient to enforce the simple condition

$$\langle \Delta \hat{p} \rangle \frac{d\psi(\hat{p})}{d\hat{p}} \Big|_0 - \psi(0) \approx 0$$

and verify that it is consistent once solutions have been evaluated.

9. Self-consistent Expressions for  $\beta$  and  $\langle \Delta p \rangle$ : Self-consistent expressions for  $\beta$  and  $\langle \Delta p \rangle$  are then constructed as follows.



10. Price Shifts in Mid-centered Frame: Given as ask at some position  $a$  - in the mid-centered frame – there is a range from  $+a$  to  $-a$  in which the sell limit may be placed, which will induce positive shifts.
11. Corresponding Price Shift Probability: The shift amount is half as great as the distance from the bid, so the measure shifts  $dP_+(\Delta p)$  from sell limit-order addition inherits a term  $2\alpha(\infty)(d\Delta p)\mathbb{P}[a \geq \Delta p]$  where the last factor counts for instances with asks large enough to admit shifts by  $\Delta p$ .
12. Symmetric Contribution from Buy Orders: There is an equal contribution  $dP_-$  from the addition of buy limit orders.
13. Symmetric Contribution from Order Removal: Symmetry requires that for every positive shift due to an addition, there is a negative shift due to evaporation with equal measure, so the contribution from buy limit order should equal that for sell limit order distribution.
14. Consolidated Expressions for  $P_+/P_-$ : When these contributions are summed, the measures for positive and negative shifts both equal

$$P_{\pm}(\Delta p) = 4\alpha(\infty)(d\Delta p)\mathbb{P}[a \geq \Delta p]$$

15. Expression for  $\beta$  from  $\psi(\Delta \hat{p})$ : The above equation may be inserted into the continuum limit of the definition

$$D = \sum_{\Delta p} P(\Delta p)\Delta p^2$$

for  $D$ , and then nondimensionalized to give

$$\beta = \frac{4}{\epsilon} \int_0^\infty (\Delta \hat{p})^2 \varphi(\Delta \hat{p}) d\Delta \hat{p}$$

where the mean-field substitution for  $\varphi(\Delta \hat{p})$  for  $\mathbb{P}[a \geq \Delta p]$  has been used.



16. Corresponding Expression for  $\langle \Delta\hat{p} \rangle$ : Similarly, the mean shift amount used in

$$\langle \Delta\hat{p} \rangle \frac{d\psi(\hat{p})}{d\hat{p}} \Big|_0 - \psi(0) \approx 0$$

is

$$\langle \Delta\hat{p} \rangle = \frac{\int_0^\infty (\Delta\hat{p}) \varphi(\Delta\hat{p}) d\Delta\hat{p}}{\int_0^\infty \varphi(\Delta\hat{p}) d\Delta\hat{p}}$$

17. Corresponding Expression for  $\varphi(\hat{p})$ : A fit of

$$1 + \varphi(\hat{p}) = - \left[ \frac{d\varphi(\hat{p})}{d\hat{p}} + \epsilon \left( 1 - \beta \frac{\partial^2}{\partial p^2} \right) \frac{d \log \varphi(\hat{p})}{d\hat{p}} \right]$$

to simulations using these self-consistent measures for shifts, is shown in the figure below.

18. Comparisons to Simulations in a Mid-centered Frame:

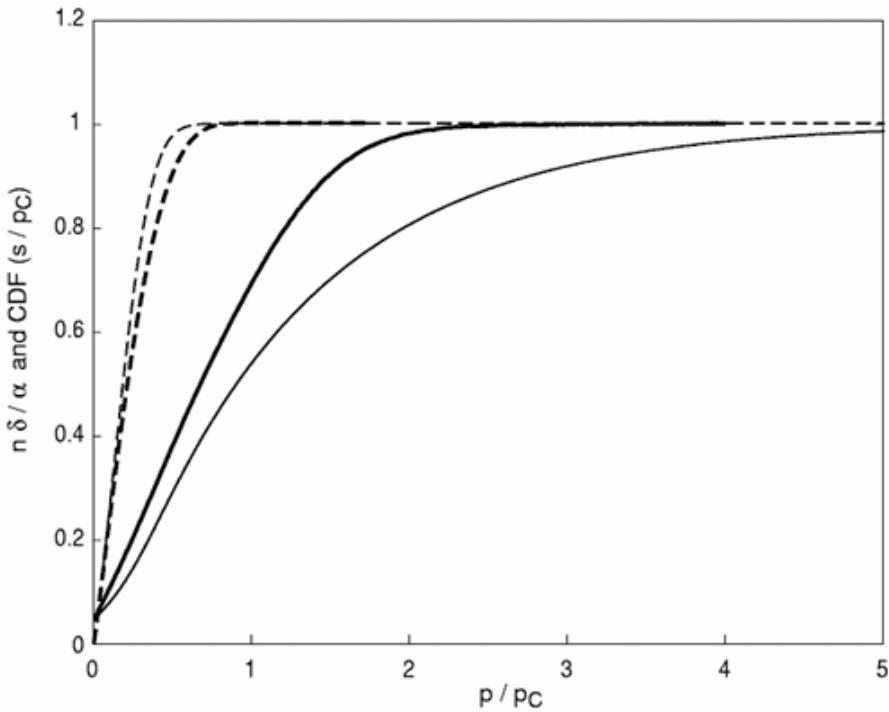


FIG. 19: Fit of the self-consistent solution with diffusivity term to simulation results for the midpoint-centered frame. Thin solid line is the analytic solution for the mean number density, and thick solid line is simulation result, at  $\epsilon = 0.02$ . Thin dashed line is the analytic prediction for the cumulative distribution function  $\Pr(\hat{s}/2 \leq \hat{p})$ , and thick dashed line is simulation result.

19. Factors Contributing to the Differences: This solution is a compromise between approximations with opposing ranges of validity.
20. Nonzero Transport through the Midpoint: The diffusion equation through the mean order depth describes nonzero transport of limit orders through the midpoint, an approximation inconsistent with the correlation of shifts with the states of the order book. This approximation is a small error only at

$$\epsilon \rightarrow 0$$



21. Range of Validity for  $\epsilon$ : On the other hand, both the form for  $\alpha(p)$  and the self-consistent solutions for  $\langle \Delta \hat{p} \rangle$  and  $\beta$  made use of the mean-field approximation, which is only valid for

$$\epsilon \lesssim 1$$

22. Compensation from Opposite Contributing Factors: The two approximations appear to create roughly compensating errors in the intermediate range of

$$\epsilon \sim 0.02$$

## Theoretical Analysis - Master Equation and Mean-field Approximations: Accounting for Correlations

1. Violation of Global Conservation Laws: The numerical integral implementing the diffusion solution doesn't actually satisfy the global conservation condition that the diffusion term integrate to zero over the whole price range.
2. Incorrect  $\hat{p} = 0$  Boundary Condition: Thus, it describes the diffusive transport of order through the midpoint, and as such does not have the correct

$$\hat{p} = 0$$

boundary condition.

3. Comparison with Bouchaud, Mezard, and Potters (2002): The effective absorbing boundary condition represented by the pure diffusion solution corresponds roughly to the approximation made by Bouchaud, Mezard, and Potters (2002).
4. Spread Reduced to a Point: It differs from this, though, in that their method of images effectively approximates the region of the spread as a point, whereas



$$\frac{\alpha(p)}{\alpha(\infty)} = \left[ \frac{\mu(p)}{\mu(0)} + \epsilon \left( 1 - \beta \frac{\partial^2}{\partial p^2} \right) \right] \psi(\hat{p})$$

actually resolves the screening of market orders as the spread fluctuates.

5. Spread as Market Order Screening Range: Treating the spread region – roughly defined as the range over which the market orders are screened – as a point is consistent with treating the coarse-grained *midpoint* as an absorbing boundary.
6. Order Transfer through the Midpoint: If the spread is resolved, however, it is not consistent for the diffusion to transport any finite number through the midpoint, because the midpoint is always in the center of an open set in a continuous price space.
7. Order Flow Accounting using Correlations: The correct behavior in the neighborhood of a *fine-grained midpoint* can be obtained by explicitly accounting for the correlation of the state of orders, with the shifts that are produced when market or limit order additions occur.
8. Dual Challenge: Conservation Laws + Boundary Condition: One expects the problem of recovering both the conservation law and the correct

$$\hat{p} = 0$$

boundary condition to be difficult, as it should be responsible for non-trivial corrections to short-term and long-term diffusion indicated earlier.

9. Prioritizing Boundary Conditions over Conservation: It is however, found that by explicitly sacrificing the global conservation law, one can incorporate the dependence of shifts on the position of the ask, in an interesting range around the midpoint.
10. CDF over the Spread Range: At general  $\epsilon$ , the corrections to diffusion reproduce the mean density over the main support of the CDF of the spread.



11. Comparing Predicted CDF vs. Simulations: While the resulting density does not predict the CDF due to correlated fluctuations, it closely resembles the real density that the independent CDFs of the two are similar.

## Theoretical Analysis - Master Equation and Mean-field Approximations: Generalizing the Self-induced Source Terms

1. Nondimensionalization of the MGF Approximator: Nondimensionalizing the master equation MGF

$$0 = \frac{\lambda - 1}{p} \left| \alpha(p) \varepsilon_p \Pi(\lambda, p) - \delta \sigma \frac{\partial \Pi(\lambda, p)}{\partial \lambda} - \frac{\mu(p)}{2\lambda} [\Pi(\lambda, p) - \pi_0(p)] \right|_{\lambda, p} \\ + \sum_{\Delta p} P_+(\Delta p) [\Pi(\lambda, p - \Delta p) - \Pi(\lambda, p)] \\ + \sum_{\Delta p} P_-(\Delta p) [\Pi(\lambda, p + \Delta p) - \Pi(\lambda, p)]$$

and keeping the leading terms in  $\varepsilon_p$  at

$$\lambda \rightarrow 0$$

one gets

$$\frac{\alpha(\hat{p})}{\alpha(\infty)} = \left[ \frac{\mu(\hat{p})}{\mu(0)} + \epsilon \right] \psi(\hat{p}) + \int dP_+(\Delta \hat{p}) [\psi(\hat{p} - \Delta \hat{p}) - \psi(\hat{p})] \\ + \int dP_-(\Delta \hat{p}) [\psi(\hat{p} + \Delta \hat{p}) - \psi(\hat{p})]$$



where  $dP_{\pm}(\Delta\hat{p})$  is the nondimensionalized measure that results from taking the continuum limit  $P_{\pm}$  in the variable  $\Delta\hat{p}$ .

2. Order Passage through a Price: The above equation is inaccurate because the number of orders shifted into or out of a price bin  $\hat{p}$ , at a given spread, must be identically zero, rather than the unconditional mean value  $\psi(\hat{p})$ .
3. Inserting Source-dependent Price Terms: One takes into the last two lines of

$$\begin{aligned}\frac{\alpha(\hat{p})}{\alpha(\infty)} = & \left[ \frac{\mu(\hat{p})}{\mu(0)} + \epsilon \right] \psi(\hat{p}) + \int dP_+(\Delta\hat{p}) [\psi(\hat{p} - \Delta\hat{p}) - \psi(\hat{p})] \\ & + \int dP_-(\Delta\hat{p}) [\psi(\hat{p} + \Delta\hat{p}) - \psi(\hat{p})]\end{aligned}$$

with lists of source terms, whose forms depend on the position of the ask, weighted by the probability for that ask.

4. Use of the Spread Measure: Measure fluctuations are assumed by using

$$\mathbb{P}\left[\frac{\hat{s}}{2} \leq \hat{p}\right] \approx \varphi(\hat{p})$$

5. Source Terms Depending on  $P_{\pm}$ : It is convenient at this point to denote the replacement of the last two lines of

$$\begin{aligned}\frac{\alpha(\hat{p})}{\alpha(\infty)} = & \left[ \frac{\mu(\hat{p})}{\mu(0)} + \epsilon \right] \psi(\hat{p}) + \int dP_+(\Delta\hat{p}) [\psi(\hat{p} - \Delta\hat{p}) - \psi(\hat{p})] \\ & + \int dP_-(\Delta\hat{p}) [\psi(\hat{p} + \Delta\hat{p}) - \psi(\hat{p})]\end{aligned}$$

with the notation  $\mathcal{S}$ , yielding

$$\frac{\alpha(\hat{p})}{\alpha(\infty)} = \left[ \frac{\mu(\hat{p})}{\mu(0)} + \epsilon \right] \psi(\hat{p}) + \mathcal{S}$$



6. Criterion for Global Conservation: The global conservation laws for orders would be satisfied if

$$\int \mathcal{S} d\hat{p} = 0$$

7. Derivation of the Source Term: The source term  $\mathcal{S}$  is derived approximately in the next few sections.
8. Comparison with Simulation #1:

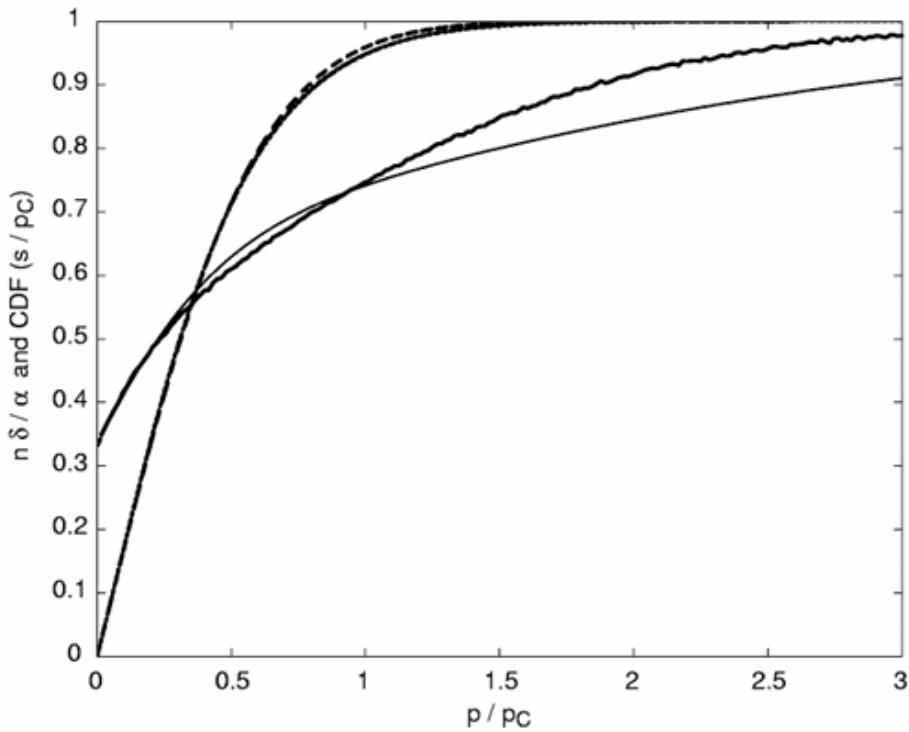


FIG. 20: Reconstruction with source terms  $\mathcal{S}$  that approximately account for correlated fluctuations near the midpoint.  $\epsilon = 0.2$ . Thick solid line is averaged order book depth from simulations, and thin solid is the mean field result. Thin dotted line is the simulated CDF for  $\hat{s}/2$ , and thick dotted line is the mean field result. Thick dashed line is the CDF that would be produced from the simulated depth, if the mean-field approximation were exact.

9. Comparison with Simulation #2: The solution to

$$\frac{\alpha(\hat{p})}{\alpha(\infty)} = \left[ \frac{\mu(\hat{p})}{\mu(0)} + \epsilon \right] \psi(\hat{p}) + \mathcal{S}$$

at



$$\epsilon = 0.2$$

with the simple diffusive source terms replaced by evaluations in the next sections, is compared to the simulated order book depth and spread distribution shown in the above figure.

10. Error Area for  $\langle n(p) \rangle$ : The simulated  $\langle n(p) \rangle$  satisfies

$$\int_0^{\infty} [1 - \epsilon\psi(\hat{p})]d\hat{p} = 1 - \frac{\langle \hat{s} \rangle}{2}$$

showing what is correct *remainder area* below the line

$$\langle n(p) \rangle \equiv 1$$

11. Global Conservation Mismatch Metric: The numerical integral deviates from that value by the incorrect integral

$$\int \mathcal{S} d\hat{p} \neq 0$$

12. Efficacy of the Source Term Approximation: However, most of the probability for the spread lies within the range where the source term  $\mathcal{S}$  is approximately correct, and as a result the distribution for  $\frac{\langle \hat{s} \rangle}{2}$  is predicted fairly well.

13. Capturing the Order Book Distribution: Even where the mean-field approximation is known to be inadequate, the source terms defined here capture most of the behavior of the order-book distribution in the region that affects the spread distribution.

14. Comparison for  $\epsilon = 0.02$ :

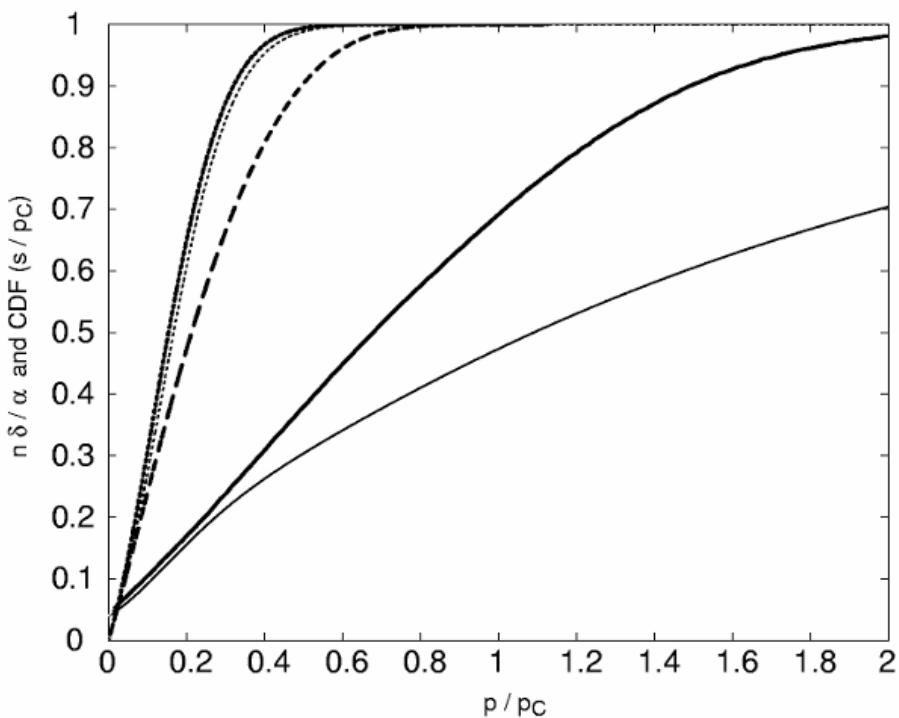


FIG. 21: Reconstruction with correlated source terms for  $\epsilon = 0.02$ . Line style and thickness are the same as in Fig. 20.

15. Comparison for  $\epsilon = 0.002$ :

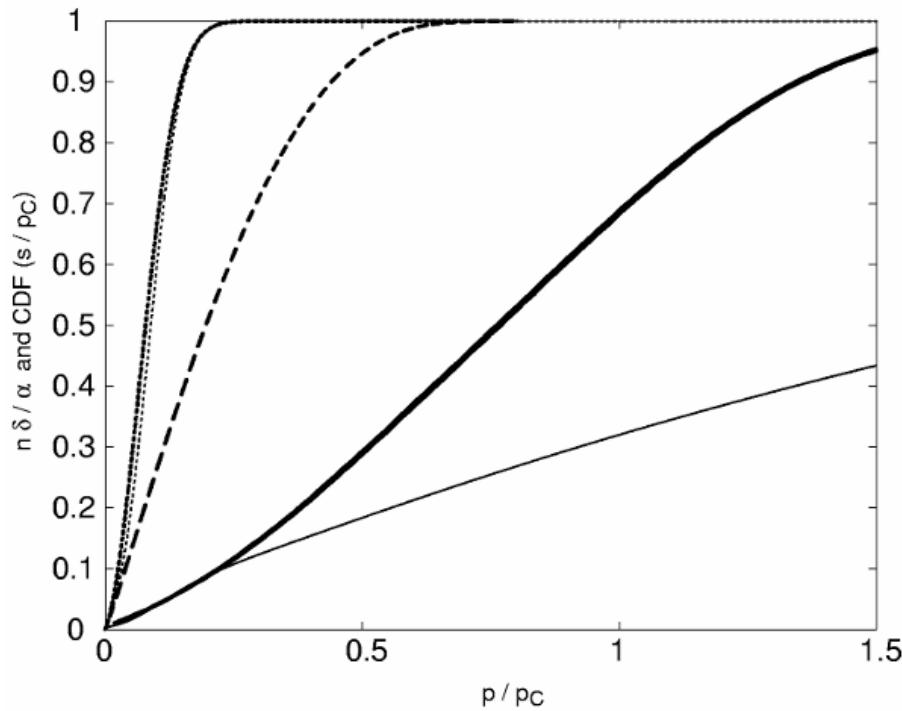


FIG. 22: Reconstruction with correlated source terms for  $\epsilon = 0.002$ . Line style and thickness are the same as in Fig. 20.

16. Comparison across Additional  $\epsilon$ 's: The above figures show the comparison to simulations for

$$\epsilon = 0.02$$

and for

$$\epsilon = 0.002$$

17. Capturing the Spread Distribution Errors: Both cases fail to reproduce the distribution for the spread, and also fail to capture large- $\hat{p}$  behavior of  $\psi(\hat{p})$ .



18. Approximating  $\psi(\hat{p})$  at small  $\hat{p}$ : However, they approximate  $\psi(\hat{p})$  at small  $\hat{p}$  well enough that the resulting distribution for the spread is close to what would be produced by the simulated  $\psi(\hat{p})$  if fluctuations were independent.

## Supporting Calculations in Density Coordinates – Cataloging Correlations

1. Events Producing the Occupation Shifts: The correct source term  $\mathcal{S}$  must correlate the incidences of zero occupation with events producing shifts.
2. Decomposing the Source Term: It is convenient to separate these into four independent types of additional and removal.
3. Removal of Buy Limit Orders: First, one considers removal of buy limit orders, which generates a negative shift of the midpoint.
4. Contributing Source Terms: Let  $\hat{a}'$  denote the position of the ask after the shift. Then, all possible shifts  $\Delta\hat{p}$  are related to a given price bin  $\hat{p}$  and  $\hat{a}'$  in one of the three ordering cases, shown in the table below.
5. Buy Order Removal Catalog:

case	source	prob
$\Delta\hat{p} \leq \hat{a}' < \hat{p}$	$\psi(\hat{p} - \Delta\hat{p}) - \psi(\hat{p})$	$\varphi(\Delta\hat{p}) - \varphi(\hat{p})$
$\Delta\hat{p} \leq \hat{p} < \hat{a}' \leq \hat{p} + \Delta\hat{p}$	$0 - \psi(\hat{p})$	$\varphi(\hat{p}) - \varphi(\hat{p} + \Delta\hat{p})$
$\hat{p} < \Delta\hat{p} < \hat{a}' \leq \hat{p} + \Delta\hat{p}$	$0 - \psi(\hat{p})$	$\varphi(\Delta\hat{p}) - \varphi(\hat{p} + \Delta\hat{p})$

TABLE VII: Contributions to “effective  $P_-$ ” from removal of a buy limit order, conditioned on the position of the ask relative to  $p$ .

6. Buy Order Removal Scenarios: For each one, the source term corresponding to  $\psi(\hat{p} - \Delta\hat{p}) - \psi(\hat{p})$  in



$$\begin{aligned}\frac{\alpha(\hat{p})}{\alpha(\infty)} = & \left[ \frac{\mu(\hat{p})}{\mu(0)} + \epsilon \right] \psi(\hat{p}) + \int dP_+(\Delta\hat{p}) [\psi(\hat{p} - \Delta\hat{p}) - \psi(\hat{p})] \\ & + \int dP_-(\Delta\hat{p}) [\psi(\hat{p} + \Delta\hat{p}) - \psi(\hat{p})]\end{aligned}$$

is given, together with the measure of order-book configurations for which that case occurs.

7. Using Mean-field Measure: The mean-field assumption

$$\mathbb{P}\left[\frac{\hat{s}}{2} \leq \hat{p}\right] \approx \varphi(\hat{p})$$

is used to estimate the measures.

8. Estimating the Shifts: As argued when defining  $\beta$  in the simpler diffusion approximation for the source terms, the measure of shifts from removal of either buy or sell limit orders should be symmetric with that of their addition within the spread, which is  $2d\Delta\hat{p}$  for either type, in cases when the shift  $\pm\Delta\hat{p}$  is consistent with the value of the spread.
9. Measures corresponding to the Shifts: The only change in these more detailed source terms is replacement of the simple  $\mathbb{P}[a \leq \Delta\hat{p}]$  with the entries in the last column in the above table.
10. Buy Order Removal Diffusion Integral: When the  $\Delta\hat{p}$  cases are integrated over this range as specified in the first column and summed, the result is a contribution of  $\mathcal{S}$  to

$$\int_0^{\hat{p}} 2\psi(\hat{p} - \Delta\hat{p})[\varphi(\Delta\hat{p}) - \varphi(\hat{p})]d\Delta\hat{p} - \int_0^{\infty} 2\psi(\hat{p})[\varphi(\Delta\hat{p}) - \varphi(\hat{p} + \Delta\hat{p})]d\Delta\hat{p}$$

11. Sell Order Removals: Sell limit-order removals generate another sequence of cases, symmetric with the buys, but inducing positive shifts.



12. Sell Order Removal Scenarios: The cases, source terms, and frequencies are given in the table below.

13. Sell Order Removal Catalog:

case	source	prob
$\Delta\hat{p} \leq \hat{a}' < \hat{p}$	$\psi(\hat{p} + \Delta\hat{p}) - \psi(\hat{p})$	$\varphi(\Delta\hat{p}) - \varphi(\hat{p})$
$\Delta\hat{p} \leq \hat{p} < \hat{a}' \leq \hat{p} + \Delta\hat{p}$	$\psi(\hat{p} + \Delta\hat{p}) - 0$	$\varphi(\hat{p}) - \varphi(\hat{p} + \Delta\hat{p})$
$\hat{p} < \Delta\hat{p} < \hat{a}' \leq \hat{p} + \Delta\hat{p}$	$\psi(\hat{p} + \Delta\hat{p}) - 0$	$\varphi(\Delta\hat{p}) - \varphi(\hat{p} + \Delta\hat{p})$

TABLE VIII: Contributions to “effective  $P_+$ ” from removal of a sell limit order, conditioned on the position of the ask relative to  $p$ .

14. Sell Order Removal Diffusion Integral: Their contribution to  $\mathcal{S}$ , after integration over  $\Delta\hat{p}$ , is then

$$\int_0^\infty 2\psi(\hat{p} + \Delta\hat{p})[\varphi(\Delta\hat{p}) - \varphi(\hat{p} + \Delta\hat{p})]d\Delta\hat{p} - \int_0^{\hat{p}} 2\psi(\hat{p})[\varphi(\Delta\hat{p}) - \varphi(\hat{p})]d\Delta\hat{p}$$

15. Sell Order Addition: Order addition is treated similarly, except that  $\hat{a}$  denotes the position of the ask before the event.

16. Sell Order Addition Scenario: Sell limit-order additions generate negative shifts, with the cases shown below.

17. Sell Order Addition Catalog:



case	source	prob
$\Delta\hat{p} \leq \hat{a} < \hat{p} - \Delta\hat{p}$	$\psi(\hat{p} - \Delta\hat{p}) - \psi(\hat{p})$	$\varphi(\Delta\hat{p}) - \varphi(\hat{p} - \Delta\hat{p})$
$\Delta\hat{p} \leq \hat{p} - \Delta\hat{p} < \hat{a} \leq \hat{p}$	$0 - \psi(\hat{p})$	$\varphi(\hat{p} - \Delta\hat{p}) - \varphi(\hat{p})$
$\hat{p} - \Delta\hat{p} < \Delta\hat{p} < \hat{a} \leq \hat{p}$	$0 - \psi(\hat{p})$	$\varphi(\Delta\hat{p}) - \varphi(\hat{p})$

TABLE IX: Contributions to “effective  $P_-$ ” from addition of a sell limit order, conditioned on the position of the ask relative to  $p$ .

18. Sell Order Addition Diffusion Integral: Integration over  $\Delta\hat{p}$  consistent with these cases gives the negative-shift contribution to  $\mathcal{S}$

$$\int_0^{\frac{\hat{p}}{2}} 2\psi(\hat{p} + \Delta\hat{p})[\varphi(\Delta\hat{p}) - \varphi(\hat{p} - \Delta\hat{p})]d\Delta\hat{p} - \int_0^{\hat{p}} 2\psi(\hat{p})[\varphi(\Delta\hat{p}) - \varphi(\hat{p} + \Delta\hat{p})]d\Delta\hat{p}$$

19. Buy Order Addition: The corresponding buy limit order addition cases are given in the table below, and their positive-shift contribution to  $\mathcal{S}$  turns out to be the same as that from removal of sell limit orders

$$\int_0^{\infty} 2\psi(\hat{p} + \Delta\hat{p})[\varphi(\Delta\hat{p}) - \varphi(\hat{p} + \Delta\hat{p})]d\Delta\hat{p} - \int_0^{\hat{p}} 2\psi(\hat{p})[\varphi(\Delta\hat{p}) - \varphi(\hat{p})]d\Delta\hat{p}$$

20. Buy Order Addition Catalog:



case	source	prob
$\Delta\hat{p} \leq \hat{a}' < \hat{p}$	$\psi(\hat{p} + \Delta\hat{p}) - \psi(\hat{p})$	$\varphi(\Delta\hat{p}) - \varphi(\hat{p})$
$\Delta\hat{p} \leq \hat{p} < \hat{a}' \leq \hat{p} + \Delta\hat{p}$	$\psi(\hat{p} + \Delta\hat{p}) - 0$	$\varphi(\hat{p}) - \varphi(\hat{p} + \Delta\hat{p})$
$\hat{p} < \Delta\hat{p} < \hat{a}' \leq \hat{p} + \Delta\hat{p}$	$\psi(\hat{p} + \Delta\hat{p}) - 0$	$\varphi(\Delta\hat{p}) - \varphi(\hat{p} + \Delta\hat{p})$

TABLE X: Contributions to “effective  $P_+$ ” from addition of a buy limit order, conditioned on the position of the ask relative to  $p$ .

21. Combined Buy Additions and Removals: Writing the source as a sum of two terms

$$\mathcal{S}(\hat{p}) = \mathcal{S}_{buy}(\hat{p}) + \mathcal{S}_{sell}(\hat{p})$$

the combined contribution from buy limit-order additions and removals is

$$\begin{aligned} \mathcal{S}_{buy}(\hat{p}) &= \int_0^{\hat{p}} 2[\psi(\hat{p} - \Delta\hat{p}) - \psi(\hat{p})][\varphi(\Delta\hat{p}) - \varphi(\hat{p})]d\Delta\hat{p} \\ &\quad - \int_0^{\infty} 2[\psi(\hat{p} + \Delta\hat{p}) - \psi(\hat{p})][\varphi(\Delta\hat{p}) - \varphi(\hat{p} + \Delta\hat{p})]d\Delta\hat{p} \end{aligned}$$

22. Combined Sell Additions and Removals: The corresponding source term from sell order addition and removal is



$$\begin{aligned}\mathcal{S}_{sell}(\hat{p}) &= \int_0^{\frac{\hat{p}}{2}} 2\psi(\hat{p} - \Delta\hat{p})[\varphi(\Delta\hat{p}) - \varphi(\hat{p} - \Delta\hat{p})]d\Delta\hat{p} \\ &\quad - 2 \int_0^{\hat{p}} 2\psi(\hat{p})[\varphi(\Delta\hat{p}) - \varphi(\hat{p})]d\Delta\hat{p} \\ &\quad + \int_0^{\infty} 2\psi(\hat{p} + \Delta\hat{p})[\varphi(\Delta\hat{p}) - \varphi(\hat{p} + \Delta\hat{p})]d\Delta\hat{p}\end{aligned}$$

23. Violation of Global Constraint: The forms

$$\begin{aligned}\mathcal{S}_{buy}(\hat{p}) &= \int_0^{\hat{p}} 2[\psi(\hat{p} - \Delta\hat{p}) - \psi(\hat{p})][\varphi(\Delta\hat{p}) - \varphi(\hat{p})]d\Delta\hat{p} \\ &\quad - \int_0^{\infty} 2[\psi(\hat{p} + \Delta\hat{p}) - \psi(\hat{p})][\varphi(\Delta\hat{p}) - \varphi(\hat{p} + \Delta\hat{p})]d\Delta\hat{p}\end{aligned}$$

and

$$\begin{aligned}\mathcal{S}_{sell}(\hat{p}) &= \int_0^{\frac{\hat{p}}{2}} 2\psi(\hat{p} - \Delta\hat{p})[\varphi(\Delta\hat{p}) - \varphi(\hat{p} - \Delta\hat{p})]d\Delta\hat{p} \\ &\quad - 2 \int_0^{\hat{p}} 2\psi(\hat{p})[\varphi(\Delta\hat{p}) - \varphi(\hat{p})]d\Delta\hat{p} \\ &\quad + \int_0^{\infty} 2\psi(\hat{p} + \Delta\hat{p})[\varphi(\Delta\hat{p}) - \varphi(\hat{p} + \Delta\hat{p})]d\Delta\hat{p}\end{aligned}$$

do not lead to



$$\int \mathcal{S} d\hat{p} = 0$$

and correcting this presumably requires distributing the orders erroneously transported through the midpoint by the diffusion term, to interior locations where they then influence long-term diffusion autocorrelation.

24. Validation of Mid-point Constraint: The source terms manifestly satisfy

$$\mathcal{S}(0) = 0$$

though, and that determines the intercept of the average order depth.

## Supporting Calculations in Density Coordinates – Cataloging Correlations; Getting the Intercept Right

1.  $\frac{\alpha(\hat{p})}{\alpha(\infty)}$  at  $\hat{p} = 0$ : Evaluating

$$\frac{\alpha(\hat{p})}{\alpha(\infty)} = \left[ \frac{\mu(\hat{p})}{\mu(0)} + \epsilon \right] \psi(\hat{p}) - \mathcal{S}$$

with

$$\frac{\alpha(\hat{p})}{\alpha(\infty)} = 1 + \mathbb{P} \left[ \frac{\hat{S}}{2} \geq \hat{p} \right]$$

at

$$\hat{p} = 0$$



gives the boundary value of the nondimensionalized, midpoint-centered, mean-order density

$$\psi(\hat{0}) = \frac{2}{1 + \epsilon}$$

which dimensionalizes to

$$\frac{\langle n(0) \rangle}{\sigma \varepsilon_p} = \frac{\frac{2\alpha(\infty)}{\sigma}}{\frac{\mu(0)}{2\sigma} + \delta}$$

2. Comparison  $\langle n(0) \rangle$  with  $\langle n \rangle_0$ : The above equation for *total* density, is the same as the form

$$\langle n(p) \rangle_0 = \frac{\alpha(p)\varepsilon_p}{\delta} - \frac{\mu(p)}{2\delta} [1 - \pi_{00}]$$

produced by the diffusion solution for the *zeroth order* density, as should be the case if diffusion no longer transports orders through the midpoint.

3. Verification of  $\langle n(0) \rangle$  from Simulations: This form is verified in simulations, with midpoint-centered averaging.
4.  $\psi(\hat{0})$  in Bid-centered Frame: Interestingly, the same argument for the bid-centered frame would simply omit from  $\frac{\alpha(\hat{0})}{\alpha(\infty)}$  predicting, that

$$\psi(\hat{0}) = \frac{1}{1 + \epsilon}$$

a result which is *not* confirmed in simulations.



5. Bid-centered Violation of Mean-field Approximations: This, in addition to not satisfying the mean-field approximations, the bid-centered density average appears to receive some diffusive transport of orders all the way down to the bid.

## Supporting Calculations in Density Coordinates – Cataloging Correlations; Fokker-Planck Expanding Correlations

1. Challenges computing  $\mathcal{S}_{buy}(\hat{p})$  and  $\mathcal{S}_{sell}(\hat{p})$ :

$$\begin{aligned}\mathcal{S}_{buy}(\hat{p}) &= \int_0^{\hat{p}} 2[\psi(\hat{p} - \Delta\hat{p}) - \psi(\hat{p})][\varphi(\Delta\hat{p}) - \varphi(\hat{p})]d\Delta\hat{p} \\ &\quad - \int_0^{\infty} 2[\psi(\hat{p} + \Delta\hat{p}) - \psi(\hat{p})][\varphi(\Delta\hat{p}) - \varphi(\hat{p} + \Delta\hat{p})]d\Delta\hat{p}\end{aligned}$$

and

$$\begin{aligned}\mathcal{S}_{sell}(\hat{p}) &= \int_0^{\frac{\hat{p}}{2}} 2\psi(\hat{p} - \Delta\hat{p})[\varphi(\Delta\hat{p}) - \varphi(\hat{p} - \Delta\hat{p})]d\Delta\hat{p} \\ &\quad - 2 \int_0^{\hat{p}} 2\psi(\hat{p})[\varphi(\Delta\hat{p}) - \varphi(\hat{p})]d\Delta\hat{p} \\ &\quad + \int_0^{\infty} 2\psi(\hat{p} + \Delta\hat{p})[\varphi(\Delta\hat{p}) - \varphi(\hat{p} + \Delta\hat{p})]d\Delta\hat{p}\end{aligned}$$

are not directly easy to use in a numerical integral.



2. Fokker-Planck Expansion of  $\mathcal{S}$ : However, they can be Fokker-Planck expanded to terms with behavior comparable to the diffusion equation, and the correct behavior near the midpoint.
3.  $\frac{d\psi(\hat{p})}{d\hat{p}}$  Derivative Components in  $\mathcal{S}$ : Doing so gives the non-dimensional expansion of the source term  $\mathcal{S}$  corresponding in

$$\frac{\alpha(p)}{\alpha(\infty)} = \left[ \frac{\mu(p)}{\mu(0)} + \epsilon \left( 1 - \beta \frac{d^2}{d\hat{p}^2} \right) \right] \psi(\hat{p})$$

$$\mathcal{S} = \mathcal{R}(\hat{p})\psi(\hat{p}) + \mathcal{P}(\hat{p}) \frac{d\psi(\hat{p})}{d\hat{p}} + \epsilon\beta(\hat{p}) \frac{d^2\psi(\hat{p})}{d\hat{p}^2}$$

4.  $\mathcal{R}(\hat{p})$ ,  $\mathcal{P}(\hat{p})$ , and  $\epsilon\beta(\hat{p})$ : The rate terms above are integrals defined as

$$\mathcal{R}(\hat{p}) = \int_0^\infty 2[\varphi(\Delta\hat{p}) - \varphi(\hat{p} + \Delta\hat{p})]d\Delta\hat{p} - 2 \int_0^{\hat{p}} 2[\varphi(\Delta\hat{p}) - \varphi(\hat{p})]d\Delta\hat{p}$$

$$+ \int_0^{\frac{\hat{p}}{2}} 2[\varphi(\Delta\hat{p}) - \varphi(\hat{p} - \Delta\hat{p})]d\Delta\hat{p}$$

$$\begin{aligned} \mathcal{P}(\hat{p}) &= 2 \int_0^\infty 2\Delta\hat{p}[\varphi(\Delta\hat{p}) - \varphi(\hat{p} + \Delta\hat{p})]d\Delta\hat{p} - \int_0^{\hat{p}} 2\hat{p}[\varphi(\Delta\hat{p}) - \varphi(\hat{p})]d\Delta\hat{p} \\ &\quad - \int_0^{\frac{\hat{p}}{2}} 2\Delta\hat{p}[\varphi(\Delta\hat{p}) - \varphi(\hat{p} - \Delta\hat{p})]d\Delta\hat{p} \end{aligned}$$



$$\begin{aligned}\epsilon\beta(\hat{p}) &= \int_0^\infty 2(\Delta\hat{p})^2[\varphi(\Delta\hat{p}) - \varphi(\hat{p} + \Delta\hat{p})]d\Delta\hat{p} + \frac{1}{2}\int_0^{\hat{p}} 2(\Delta\hat{p})^2[\varphi(\Delta\hat{p}) - \varphi(\hat{p})]d\Delta\hat{p} \\ &\quad + \frac{1}{2}\int_0^{\frac{\hat{p}}{2}} 2(\Delta\hat{p})^2[\varphi(\Delta\hat{p}) - \varphi(\hat{p} - \Delta\hat{p})]d\Delta\hat{p}\end{aligned}$$

5. Recovering the Simplification Diffusion Constant: All of the coefficients in

$$\begin{aligned}\mathcal{R}(\hat{p}) &= \int_0^\infty 2[\varphi(\Delta\hat{p}) - \varphi(\hat{p} + \Delta\hat{p})]d\Delta\hat{p} - 2\int_0^{\hat{p}} 2[\varphi(\Delta\hat{p}) - \varphi(\hat{p})]d\Delta\hat{p} \\ &\quad + \int_0^{\frac{\hat{p}}{2}} 2[\varphi(\Delta\hat{p}) - \varphi(\hat{p} - \Delta\hat{p})]d\Delta\hat{p}\end{aligned}$$

and

$$\begin{aligned}\mathcal{P}(\hat{p}) &= 2\int_0^\infty 2\Delta\hat{p}[\varphi(\Delta\hat{p}) - \varphi(\hat{p} + \Delta\hat{p})]d\Delta\hat{p} - \int_0^{\hat{p}} 2\hat{p}[\varphi(\Delta\hat{p}) - \varphi(\hat{p})]d\Delta\hat{p} \\ &\quad - \int_0^{\frac{\hat{p}}{2}} 2\Delta\hat{p}[\varphi(\Delta\hat{p}) - \varphi(\hat{p} - \Delta\hat{p})]d\Delta\hat{p}\end{aligned}$$

manifestly vanish as

$$\hat{p} \rightarrow 0$$

and at large  $\hat{p}$

$$\mathcal{R}(\hat{p}) \rightarrow 0$$



$$\mathcal{P}(\hat{p}) \rightarrow 0$$

while

$$\epsilon\beta(\hat{p}) \rightarrow 4 \int_0^\infty (\Delta\hat{p})^2 \varphi(\Delta\hat{p}) d\Delta\hat{p}$$

recovering the diffusion constant  $\beta(\hat{p})$  from  $\varphi(\Delta\hat{p})$  seen earlier, of the simplified source term.

6. Nonlocal Nature of  $\varphi(\hat{p})$ : However, they are still not convenient for numerical integration, being non-local in  $\varphi(\hat{p})$ .
7. Approximating  $\varphi(\hat{p} \pm \Delta\hat{p})$  from  $\hat{p}, \Delta\hat{p}$ : The exponential form

$$\frac{\mu(\hat{p})}{\mu(0)} = \varphi(\hat{p}) = e^{-\int_0^{\hat{p}} \psi(\hat{p}') d\hat{p}'}$$

is therefore exploited to approximate  $\varphi(\hat{p})$ , in the region where its value is largest, with the expansion

$$\varphi(\hat{p} \pm \Delta\hat{p}) \approx \varphi(\hat{p}) \varphi(\Delta\hat{p}) e^{\pm \hat{p} \Delta\hat{p} \frac{d\psi(\hat{p})}{d\hat{p}}|_0}$$

8. Additional Simplification Applied on  $\varphi(\hat{p} \pm \Delta\hat{p})$ : In the range where the mean-field approximation is valid,  $\varphi(\hat{p} \pm \Delta\hat{p})$  is dominated by the constant term  $\psi(0)$ , and even the factors  $e^{\pm \hat{p} \Delta\hat{p} \frac{d\psi(\hat{p})}{d\hat{p}}|_0}$  can be approximated as unity.
9. Simplified  $\mathcal{R}(\hat{p})$ ,  $\mathcal{P}(\hat{p})$ , and  $\epsilon\beta(\hat{p})$ : This leaves the much-simplified expansions

$$\mathcal{R}(\hat{p}) = [1 - \varphi(\hat{p})] \mathcal{I}_0(\infty) - 2[\mathcal{I}_0(\hat{p}) - 2\hat{p}\varphi(\hat{p})] + [1 - \varphi(\hat{p})] \mathcal{I}_0\left(\frac{\hat{p}}{2}\right)$$



$$\mathcal{P}(\hat{p}) = 2[1 - \varphi(\hat{p})]\mathcal{I}_1(\infty) - [\mathcal{I}_1(\hat{p}) - \hat{p}^2\varphi(\hat{p})] + [1 - \varphi(\hat{p})]\mathcal{I}_1\left(\frac{\hat{p}}{2}\right)$$

and

$$\epsilon\beta(\hat{p}) = [1 - \varphi(\hat{p})]\mathcal{I}_2(\infty) - \frac{1}{2}\left[\mathcal{I}_2(\hat{p}) - \frac{2}{3}\hat{p}^3\varphi(\hat{p})\right] + [1 - \varphi(\hat{p})]\mathcal{I}_2\left(\frac{\hat{p}}{2}\right)$$

10. Explicit Expressions for  $\mathcal{I}_j(\hat{p})$ : In

$$\mathcal{R}(\hat{p}) = [1 - \varphi(\hat{p})]\mathcal{I}_0(\infty) - 2[\mathcal{I}_0(\hat{p}) - 2\hat{p}\varphi(\hat{p})] + [1 - \varphi(\hat{p})]\mathcal{I}_0\left(\frac{\hat{p}}{2}\right)$$

$$\mathcal{P}(\hat{p}) = 2[1 - \varphi(\hat{p})]\mathcal{I}_1(\infty) - [\mathcal{I}_1(\hat{p}) - \hat{p}^2\varphi(\hat{p})] + [1 - \varphi(\hat{p})]\mathcal{I}_1\left(\frac{\hat{p}}{2}\right)$$

and

$$\epsilon\beta(\hat{p}) = [1 - \varphi(\hat{p})]\mathcal{I}_2(\infty) - \frac{1}{2}\left[\mathcal{I}_2(\hat{p}) - \frac{2}{3}\hat{p}^3\varphi(\hat{p})\right] + [1 - \varphi(\hat{p})]\mathcal{I}_2\left(\frac{\hat{p}}{2}\right)$$

above

$$\mathcal{I}_j(\hat{p}) \equiv \int_0^{\hat{p}} 2(\Delta\hat{p})^j \varphi(\Delta\hat{p}) d\Delta\hat{p}$$

for

$$j = 0, 1, 2$$



## 11. Plugging Back into $\mathcal{S}$ : The forms

$$\mathcal{R}(\hat{p}) = [1 - \varphi(\hat{p})]\mathcal{I}_0(\infty) - 2[\mathcal{I}_0(\hat{p}) - 2\hat{p}\varphi(\hat{p})] + [1 - \varphi(\hat{p})]\mathcal{I}_0\left(\frac{\hat{p}}{2}\right)$$

$$\mathcal{P}(\hat{p}) = 2[1 - \varphi(\hat{p})]\mathcal{I}_1(\infty) - [\mathcal{I}_1(\hat{p}) - \hat{p}^2\varphi(\hat{p})] + [1 - \varphi(\hat{p})]\mathcal{I}_1\left(\frac{\hat{p}}{2}\right)$$

and

$$\epsilon\beta(\hat{p}) = [1 - \varphi(\hat{p})]\mathcal{I}_2(\infty) - \frac{1}{2}\left[\mathcal{I}_2(\hat{p}) - \frac{2}{3}\hat{p}^3\varphi(\hat{p})\right] + [1 - \varphi(\hat{p})]\mathcal{I}_2\left(\frac{\hat{p}}{2}\right)$$

are inserted into

$$\mathcal{S} = \mathcal{R}(\hat{p})\psi(\hat{p}) + \mathcal{P}(\hat{p})\frac{d\psi(\hat{p})}{d\hat{p}} + \epsilon\beta(\hat{p})\frac{d^2\psi(\hat{p})}{d\hat{p}^2}$$

for  $\mathcal{S}$  to produce the mean-field results compared to simulations, as illustrated earlier.

## **Theoretical Analysis – A Mean-Field Theory of Order Separation Intervals: The Independent Interval Approximation**

1. Independent Fluctuations in  $x(N)$ : A simplifying assumption that is in some sense a dual to independent fluctuations of  $n(p)$ , in independent fluctuations in the intervals  $x(N)$  at different  $N$ .
2. MFT for Order Separation Intervals: This section develops a MFT for the order separation intervals in this model.



3. Estimate of the Depth Profiles: From this, one may also be able to make an estimate of the depth profiles for any values of the parameters.
4. Limit Orders on Unoccupied Sites: Limit order placements are considered to occur strictly on sites which are not occupied. This is the same level of approximation as made earlier.
5. Time Step Normalization: The time step is normalized to unity, as above, so that rates are equal to probabilities after one update of the whole configuration.
6. Values for  $\alpha$  and  $\mu$ : The rates  $\alpha$  and  $\mu$  used in this section correspond to  $\alpha(\infty)$  and  $\mu(0)$  as defined earlier.
7. Instantaneous Order Separation Interval Configuration: As shown in an earlier figure, the configuration is entirely specified instant by instant if the instantaneous values of the order separation intervals are known.
8. Order Separation Interval Dynamics: Consider now how these intervals might change due to various processes.
9. Change at a Specified Spread Locals: For the spread  $x_0$ , these processes and the corresponding change in  $x_0$ , are listed below.
10. Removal of an Ask:

$$x_0 \rightarrow x_0 + x_1$$

with rate  $\delta + \frac{\mu}{2\alpha}$  when the ask either evaporates or is deleted by a market order.

11. Removal of a Bid:

$$x_0 \rightarrow x_0 + x_{-1}$$

with rate  $\delta + \frac{\mu}{2\alpha}$  when the bid either evaporates or is deleted by a market order.

12. Adding a Single Sell:

$$x_0 \rightarrow x'$$



for any value

$$1 \leq x' \leq x_0 - 1$$

when a sell limit order is deposited anywhere in the spread.

13. Cumulative Sell Additions: The rate for any single deposition is  $\frac{\alpha \varepsilon_p}{\sigma}$  to the cumulative rate for *some* deposition is  $\frac{\alpha \varepsilon_p(x_0-1)}{\sigma}$ . The  $-1$  comes from the prohibition against depositing on occupied sites.

14. Adding a Single Buy: Similarly

$$x_0 \rightarrow x_0 - x'$$

for any value

$$1 \leq x' \leq x_0 - 1$$

when a buy limit order is deposited in the spread, also with cumulative rate  $\frac{\alpha \varepsilon_p(x_0-1)}{\sigma}$ .

15. Likelihood of Order Staying Untouched: Since the above processes describe all possible single-event changes to the configuration, the probability that it remains unchanged in a single time step  $1 - 2\delta - \frac{\mu}{\sigma} - \frac{2\alpha \varepsilon_p(x_0-1)}{\sigma}$

16. Setting  $\sigma = 1$ : In all that follows, one sets

$$\sigma = 1$$

without loss of generality.

17. Time Evolution of  $x_0$  #1: If one knows  $x_0$ ,  $x_1$ , and  $x_{-1}$  at  $t$ , the expected value at  $t + \Delta t$  is



$$x_0(t + \Delta t) = x_0(t)[1 - 2\delta - \mu_0 - 2\alpha(x_0 - 1)] + (x_0 + x_1)\left(\delta + \frac{\mu}{2}\right) \\ + (x_0 + x_{-1})\left(\delta + \frac{\mu}{2}\right) + \alpha_0 \varepsilon_p x_0(x_0 - 1)$$

18.  $x_i(t)$  as an Ensemble Average: Here,  $x_i(t)$  represents the value of the integral averaged over many realizations of the process evolved up to time  $t$ .
19. Time Evolution of  $x_0$  #1: Again, representing the finite difference as a time derivative, and change in the expected value, given  $x_0$ ,  $x_1$ , and  $x_{-1}$ , is
- $$\frac{dx_0(t)}{dt} = (x_1 + x_{-1})\left(\delta + \frac{\mu}{2}\right) - \alpha \varepsilon_p x_0(x_0 - 1)$$
20. Quadratic Dependence on  $x_0$ : Were it not for the quadratic term arising from deposition, the above equation would be a linear function of  $x_0$ ,  $x_1$ , and  $x_{-1}$ .
21. Approximation for  $\langle x_0^2 \rangle$  #1: However, one needs an approximation for  $\langle x_0^2 \rangle$ , where the angle brackets represent on average over realizations as before or equivalently a prime average in the steady state.
22. Approximation for  $\langle x_0^2 \rangle$  #2: Assuming for a moment that  $\langle x_0^2 \rangle$  can be approximated by  $a\langle x_0^2 \rangle$  where  $a$  is some as yet undetermined constant to be set self-consistently.
23. Approximation for  $\langle x_k^2 \rangle$ : This assumption will be made for all  $x_k$ 's. The assumption is inaccurate because the PDF of  $x_k$  could depend on  $k$ .
24. Validity of the Approximation: However, as will be seen, this is still a very good approximation.
25. Steady-State Recurrence Relation: It then follows that

$$(x_1 + x_{-1})\left(\delta + \frac{\mu}{2}\right) = \alpha \varepsilon_p x_0(x_0 - 1)$$

26. Inverse Distance from the Bid: The interval  $x_k$  may be thought of as the inverse of the density at a distance  $\sum_{j=0}^{k-1} x_j$  from the bid.
27. Explicit Form for  $x_i$ : That is



$$x_i \approx \frac{1}{\langle n(\sum_{j=0}^{k-1} x_j \varepsilon_p) \rangle}$$

the dual to the mean depth, at least at large  $i$ .

28. The Normalized Order Separation Interval: It therefore makes sense to introduce a normalized interval

$$\hat{x}_i \equiv \epsilon \frac{\alpha}{\delta} x_i \varepsilon_p = \frac{x_i \varepsilon_p}{p_c} \approx \frac{1}{\psi(\sum_{j=0}^{k-1} \hat{x}_j)}$$

the mean-field inverse of the normalized depth  $\psi$ .

29. Nondimensionalization of the Recurrence Relation: In this nondimensionalized form

$$(\hat{x}_1 + \hat{x}_{-1}) \left( \delta + \frac{\mu}{2} \right) = \alpha \varepsilon_p x_0 (x_0 - 1)$$

becomes

$$(\hat{x}_1 + \hat{x}_{-1})(1 + \epsilon) = a \hat{x}_0 (\hat{x}_0 - \hat{\varepsilon}_p)$$

where

$$\hat{\varepsilon}_p = \frac{\varepsilon_p}{p_c}$$

30. Symmetry of Depth about Origin: Since the depth profile is symmetric about the origin

$$\hat{x}_1 = \hat{x}_{-1}$$



31. Nondimensionalized, Symmetrized Recurrence Relation: From the equation, it can be seen that this ansatz is self-consistent and extends to all higher  $\hat{x}_i$ . Substituting this in

$$(\hat{x}_1 + \hat{x}_{-1})(1 + \epsilon) = a\hat{x}_0(\hat{x}_0 - \hat{\varepsilon}_p)$$

one gets

$$\hat{x}_1(1 + \epsilon) = \frac{a}{2}\hat{x}_0(\hat{x}_0 - \hat{\varepsilon}_p) = \hat{x}_{-1}(1 + \epsilon)$$

32.  $x_1$  Transitions: Proceeding to the change of  $x_1$ , the events that can occur, with their probabilities, are shown in Table V, with the remaining probability that  $x_1$  remains unchanged.

33.  $x_1$  Transitions Table:

case	rate	range
$x_1 \rightarrow x_2$	$(\delta + \mu/2)$	
$x_1 \rightarrow (x_1 + x_2)$	$\delta$	
$x_1 \rightarrow x'$	$\alpha dp$	$x' \in (1, x_0 - 1)$
$x_1 \rightarrow x_1 - x'$	$\alpha dp$	$x' \in (1, x_1 - 1)$

TABLE V: Events that can change the value of  $x_1$ , with their rates of occurrence.

34. PDE for the  $x_1$  Evolution: The differential equation for the mean change of  $x_1$  can be derived along the previous lines and becomes

$$\frac{dx_1(t)}{dt} = \left(2\delta + \frac{\mu}{2}\right)x_2 - \left(\delta + \frac{\mu}{2}\right)x_1 + \alpha\varepsilon_p \left[ \frac{x_0(x_0 - 1)}{2} - \frac{x_1(x_1 - 1)}{2} - x_1(x_0 - 1) \right]$$



35. Applying Independent Interval Approximation: Note that in the above equations, the mean-field approximation consists of assuming that terms like  $\langle x_0 x_1 \rangle$  are approximated by the product  $\langle x_0 \rangle \langle x_1 \rangle$ . This is thus an *independent interval* approximation.
36. Nondimensional Recurrence Relation for  $\hat{x}_2$ : Nondimensionalizing  $\frac{dx_1(t)}{dt}$  and combining the result with

$$\hat{x}_1(1 + \epsilon) = \frac{a}{2} \hat{x}_0 (\hat{x}_0 - \hat{\varepsilon}_p) = \hat{x}_{-1}(1 + \epsilon)$$

gives the stationary value for  $\hat{x}_2$  from  $\hat{x}_0$  and  $\hat{x}_1$ :

$$\hat{x}_2(1 + 2\epsilon) = \frac{a}{2} \hat{x}_1 (\hat{x}_1 - \hat{\varepsilon}_p) + \hat{x}_1 (\hat{x}_0 - \hat{\varepsilon}_p)$$

37. Nondimensional Recurrence Relation for  $\hat{x}_k$ : Following the same procedure for general  $k$ , the nondimensionalized recurrence relation is:

$$\hat{x}_k(1 + k\epsilon) = \frac{a}{2} \hat{x}_{k-1} (\hat{x}_{k-1} - \hat{\varepsilon}_p) + \hat{x}_{k-1} \sum_{i=0}^{k-2} (\hat{x}_i - \hat{\varepsilon}_p)$$

## Theoretical Analysis – A Mean-field Theory of Order Separation Intervals; The Independent Interval Approximations: Asymptotes and Conservation Rules

1.  $x_\infty$  Asymptote #1: Far from the bid or the ask,  $\hat{x}_k$  must go to a constant value, which is denoted  $\hat{x}_\infty$ . In other words, for large  $k$

$$\hat{x}_{k+1} \rightarrow \hat{x}_k$$



2.  $x_\infty$  Asymptote #2: Taking the difference of

$$\hat{x}_k(1 + k\epsilon) = \frac{a}{2} \hat{x}_{k-1}(\hat{x}_{k-1} - \hat{\varepsilon}_p) + \hat{x}_{k-1} \sum_{i=0}^{k-2} (\hat{x}_i - \hat{\varepsilon}_p)$$

for  $k + 1$  and  $k$  in this limit gives the identification

$$\epsilon \hat{x}_\infty = \hat{x}_\infty (\hat{x}_\infty - \hat{\varepsilon}_p)$$

or

$$\hat{x}_\infty = \epsilon + \hat{\varepsilon}_p$$

3. Comparison with  $\psi(\infty)$ : Apart from the factor of  $\hat{\varepsilon}_p$ , arising from the exclusion of deposition on already-occupied sites, this agrees with the limit

$$\psi(\infty) \rightarrow \frac{1}{\epsilon}$$

found earlier.

4.  $\psi(\infty)$  as  $\hat{\varepsilon}_p \rightarrow 0$ : In the continuum limit

$$\hat{\varepsilon}_p \rightarrow 0$$

at fixed  $\epsilon$ , these are the same.

5. Cumulative Departures from the Asymptote: From the large- $k$  limit of



$$\hat{x}_k(1 + k\epsilon) = \frac{a}{2}\hat{x}_{k-1}(\hat{x}_{k-1} - \hat{\varepsilon}_p) + \hat{x}_{k-1} \sum_{i=0}^{k-2} (\hat{x}_i - \hat{\varepsilon}_p)$$

one can also solve easily for the quantity

$$S_\infty = \sum_{i=0}^{k-2} (\hat{x}_i - \hat{x}_\infty)$$

which is related to the bid-centered order conversation law mentioned earlier.

6. Explicit Expression for Cumulative Asymptote Departure: Dividing by a factor  $\hat{x}_\infty$  at large  $k$

$$(1 + k\epsilon) = \frac{a}{2}(\hat{x}_{k-1} - \hat{\varepsilon}_p) + \sum_{i=0}^{k-2} (\hat{x}_i - \hat{\varepsilon}_p)$$

using

$$\epsilon \hat{x}_\infty = \hat{x}_\infty (\hat{x}_\infty - \hat{\varepsilon}_p)$$

and re-writing the sum on the right-hand side as  $\sum_{i=0}^{k-2} (\hat{x}_i - \hat{x}_\infty) + \sum_{i=0}^{k-2} (\hat{x}_\infty - \hat{\varepsilon}_p)$  one gets

$$S_\infty = 1 + \left(1 - \frac{a}{2}\right)\epsilon$$

7. Intuition behind  $S_\infty$ : The interpretation of  $S_\infty$  is straightforward, First, recognize that there are  $k + 1$  order in the price range  $\sum_{i=0}^k x_i$
8. Net Order Removal Rate: Their decay rate is  $\delta(k + 1)$  and the rate of annihilation of market orders is  $\frac{\mu}{2}$ .



9. Net Order Addition Rate: The rate of additions, up to an uncertainty about what should be considered the center of the interval, is  $\alpha \varepsilon_p \sum_{i=0}^k (x_i - 1)$  in the bid-centered frame where the effective  $\alpha$  is constant and additions on top of previously occupied sites is forbidden.
10. Applying Bid-centered Conservation Law: Equality of addition and removal is the bid-centered order conservation law, in the form

$$\frac{\mu}{2} + \delta(k+1) = \alpha \varepsilon_p \sum_{i=0}^k (x_i - 1)$$

11. Corresponding Value for  $S_\infty$ : Taking large  $k$ , nondimensionalizing, and using

$$\epsilon \hat{x}_\infty = \hat{x}_\infty (\hat{x}_\infty - \hat{\varepsilon}_p)$$

$$\frac{\mu}{2} + \delta(k+1) = \alpha \varepsilon_p \sum_{i=0}^k (x_i - 1)$$

becomes

$$S_\infty = 1$$

12. Comparison with Monte Carlo #1: This conservation law is indeed respected to a remarkable accuracy in Monte-Carlo simulations in the model as indicated in the table below.
13. Comparison with Monte Carlo #2:



$\epsilon$	$S_\infty$ from theory	$S_\infty$ from MCS
0.66	1	1.000
0.2	1	1.000
0.04	1	0.998
0.02	1	1.000

TABLE VI: Theoretical vs. results from simulations for  $S_\infty$ .

14.  $\langle x_k^2 \rangle$  from  $\langle x_k \rangle^2$ : The value of  $a$  implies that one now sets

$$\langle x_k^2 \rangle = 2\langle x_k \rangle^2$$

15. Exponential PDF of  $x_k$ : This would be strictly true if the PDF of the interval  $x_k$  were exponential across  $k$ .

16. Exponential PDF Validity across  $\epsilon$ : This is generally a good approximation for large  $k$  for any  $\epsilon$ .

17. PDF from Monte Carlo #1:

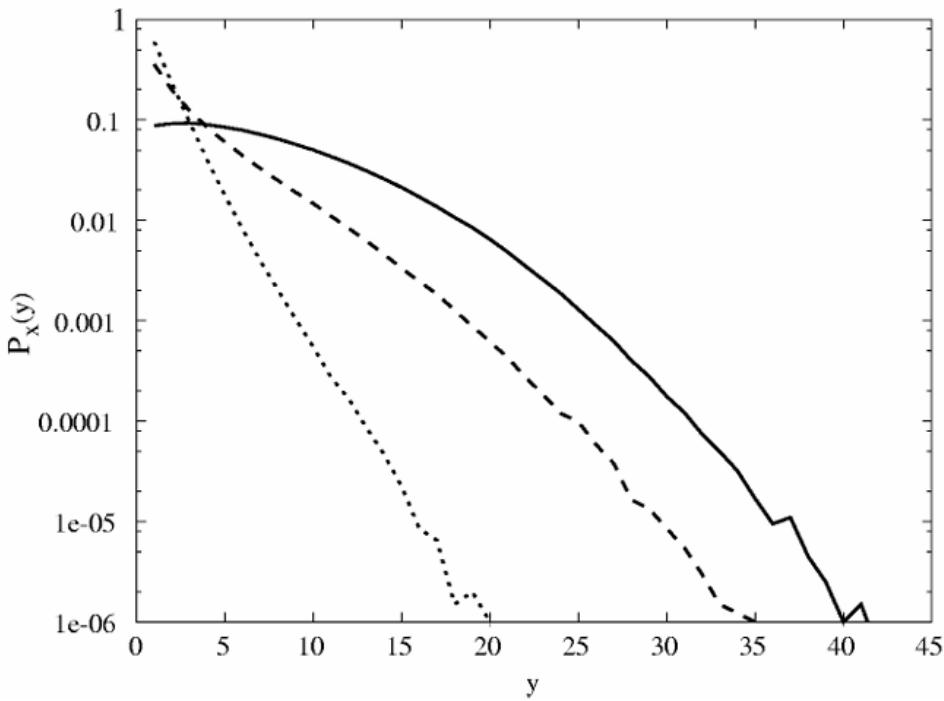


FIG. 23: The probability distribution functions  $P_x(y)$  vs.  $y$  for the intervals  $x = x_0, x_1$  and  $x_5$  at  $\epsilon = 0.1$ , on a semi-log scale. Solid curve is for  $x_0$ , dashed for  $x_1$ , and dot for  $x_5$ . The functional form of the distribution changes from a Gaussian to an exponential.

18. [PDF from Monte Carlo #2](#): The figure above shows the numerical results from Monte Carlo simulation of the model, for the PDF for three intervals  $x_0, x_1$ , and  $x_5$  at

$$\epsilon = 0.1$$

19. [PDF Simulation from Gaussian to Exponential](#): The functional form for  $P(x_0)$  and  $P(x_1)$  are better approximated by a Gaussian than an exponential. However,  $P(x_5)$  is clearly an exponential.

20. [Short- vs. Long-Term Diffusion](#):

$$S_\infty = 1$$



has an important consequence for the short-term and the long-term diffusivities, which can also be seen in simulations, as mentioned in earlier sections.

21. Nondimensionalized Diffusivity: The nondimensionalization of the diffusivity  $D$  with the rate parameters suggests a classical scaling of the diffusivity

$$D \sim p_c^2 \delta = \frac{\mu^2}{4\alpha^2} \delta$$

22. Short- vs. Long-Term Diffusion Scaling: As mentioned earlier, it is observed from simulations that the locally best short-term fit to the actual diffusivity of the midpoint is  $\sim \sqrt{\frac{1}{\epsilon}}$  times the  $D$  estimate above, and the long-term diffusivity is  $\sim \sqrt{\epsilon}$  times the classical estimate.

23. Scaling from the Conservation Law: While analytical derivation for this relation is not yet known, the fact that early and late-time renormalizations must have this qualitative relation can be argued from the conservation law

$$S_\infty = 1$$

24. Area between Actual/Asymptotic Densities:  $S_\infty$  is the area enclosed between the actual density and the asymptotic value.

25. Order Depletion Rate vs  $\epsilon$ : Increases in  $\frac{1}{\epsilon}$  – the descaled market-order rate – depletes orders near the spread, diminishing the mean depth at small  $\hat{p}$ , and induce the upward curvature.

26. Impact on the Scaled Diffusivity: As noted above, they cause more frequent shifts – more than compensating for the slight decrease in average step size – and increase the classically descaled diffusivity  $\beta$ .



27. Increase in  $S_\infty$  near Spread: However, as a result, this increases the fraction of the area in  $S_\infty$  accumulation near the spread, requiring that the mean depth at larger  $\hat{p}$  increase to compensate.
28. Consequence Diffusion Decrease with  $\frac{1}{\epsilon}$ : The resulting steeper approach to asymptotic depth at prices greater than the mean spread, and the large negative curvature of the distribution, are fit by an effective diffusivity that *decreases* with increasing  $\frac{1}{\epsilon}$ .
29. Long-term Diffusivity of Distribution: Since the distribution further from the midpoint represents the imprint of the market order activity further in the past, this effective diffusivity describes the long-term evolution of the distribution.
30. Impact on Scaling with  $\epsilon$ : The resulting anticorrelation of the small  $\hat{p}$  and the large  $p$  effective diffusion constants implied by the conservation of the area  $S_\infty$  is exactly consistent with their respective  $\sim \sqrt{\frac{1}{\epsilon}}$  and  $\sim \sqrt{\epsilon}$  setting.
31. Diffusivity Relation across Timescales #1: The general idea here is to connect diffusivities at short and long timescales to the depth profile near the spread and far away from the spread respectively.
32. Diffusivity Relation across Timescales #2: The conservation law for the depth profile then implies a connection between these two diffusivities.

## Theoretical Analysis – A Mean-Field Theory of Order Separation Intervals; The Independent Interval Approximation: Direct Simulation in Interval Coordinates

1. Parametrization of the Recurrence Relation: The set of equations determined by the general form



$$\hat{x}_k(1 + k\epsilon) = \frac{a}{2} \hat{x}_{k-1}(\hat{x}_{k-1} - \hat{\varepsilon}_p) + \hat{x}_{k-1} \sum_{i=0}^{k-2} (\hat{x}_i - \hat{\varepsilon}_p)$$

is ultimately parametrized by the single point  $\hat{x}_0$ .

2. Recursive Solution Converging to  $\hat{x}_\infty$ : The correct value for  $\hat{x}_0$  is determined when the  $\hat{x}_k$  are solved recursively, by requiring convergence to  $\hat{x}_\infty$ .
3. Comparison of Solution to  $\psi(\hat{p})$ : This recursion is done numerically, in the same manner as was done to solve the differential equation for the normalized mean density  $\psi(\hat{p})$ .
4.  $\hat{x}_0$  Numerical Simulation #1:

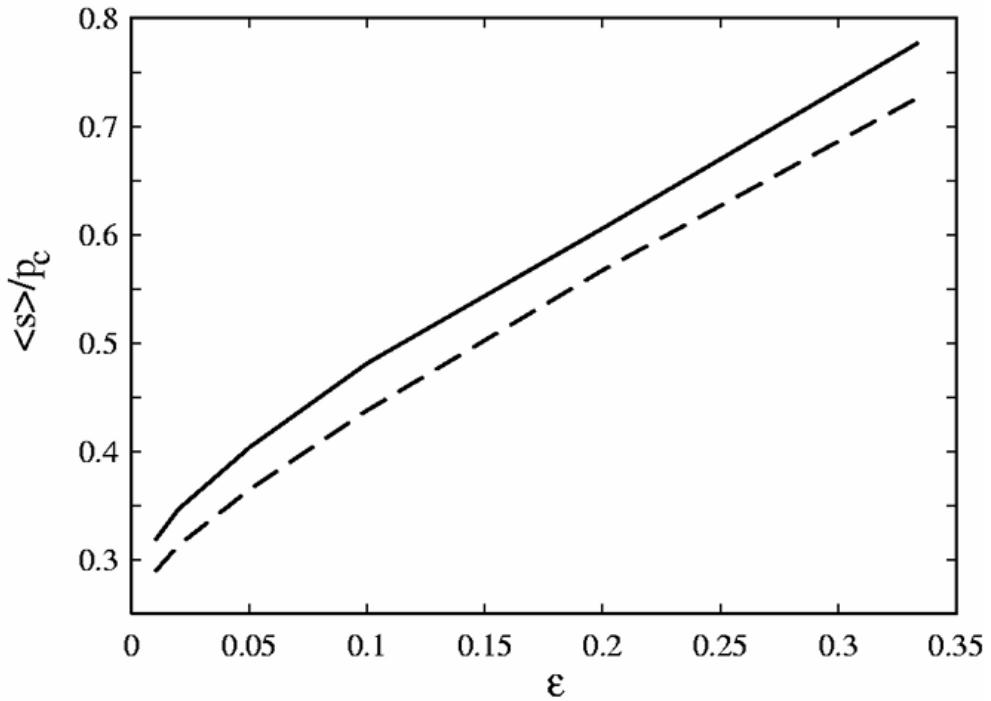


FIG. 24: The mean value of the spread in nondimensional units  $\hat{s} = s/p_c$  as a function of  $\epsilon$ . The numerical value above (solid) is compared with the theoretical estimate below (dash). .



5.  $\hat{x}_0$  Numerical Simulation #2: The figure above compares the numerical result for  $\hat{x}_0$  with the analytical estimate generated as explained above.
6. Match across  $\epsilon$  Range #1: The results are surprisingly good throughout the entire range.
7. Match across  $\epsilon$  Range #2: Though the theoretical value consistently undermines the numerical value, yet the functional form is captured accurately.
8.  $x_k$  - Numerical Simulation #1:

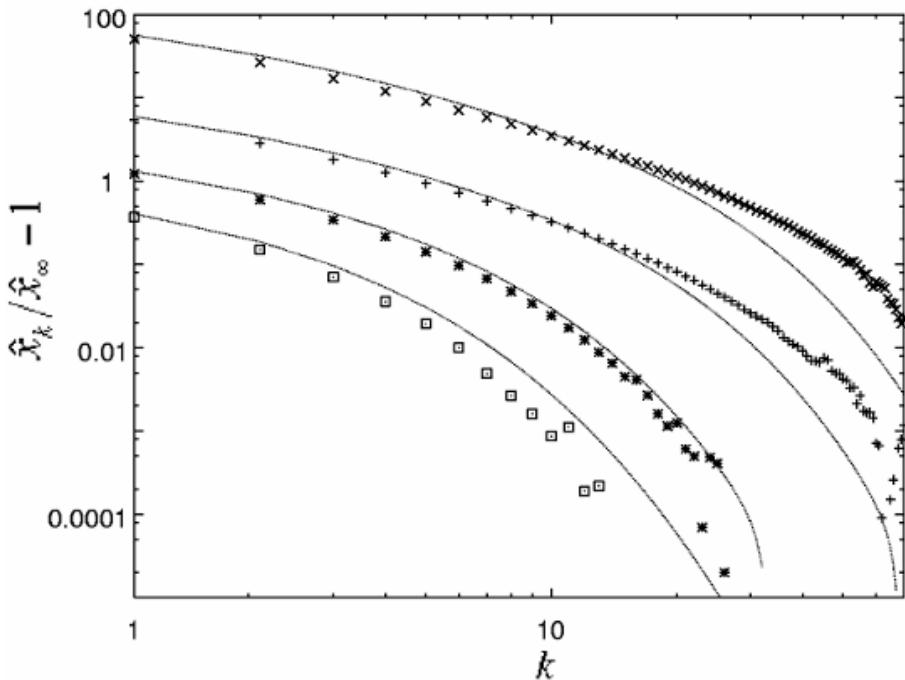


FIG. 25: Four pairs of curves for the quantity  $\hat{x}_k / \hat{x}_\infty - 1$  vs.  $k$ . The value of  $\epsilon$  increases from top to bottom ( $\epsilon = 0.02, 0.04, 0.2, 0.66$ ). In each pair of curves, the markers are obtained from simulations while the solid curve is the prediction of Eq. 53 evaluated numerically. The difference between numerics and mean-field increases as  $\epsilon$  decreases, especially for large  $k$ .



9.  $x_k$  - Numerical Simulation #2: The values of  $x_k$  are compared to the values directly with simulations across all  $k$ .
10. Semi-log Comparison for  $\hat{x}_k$  #1:

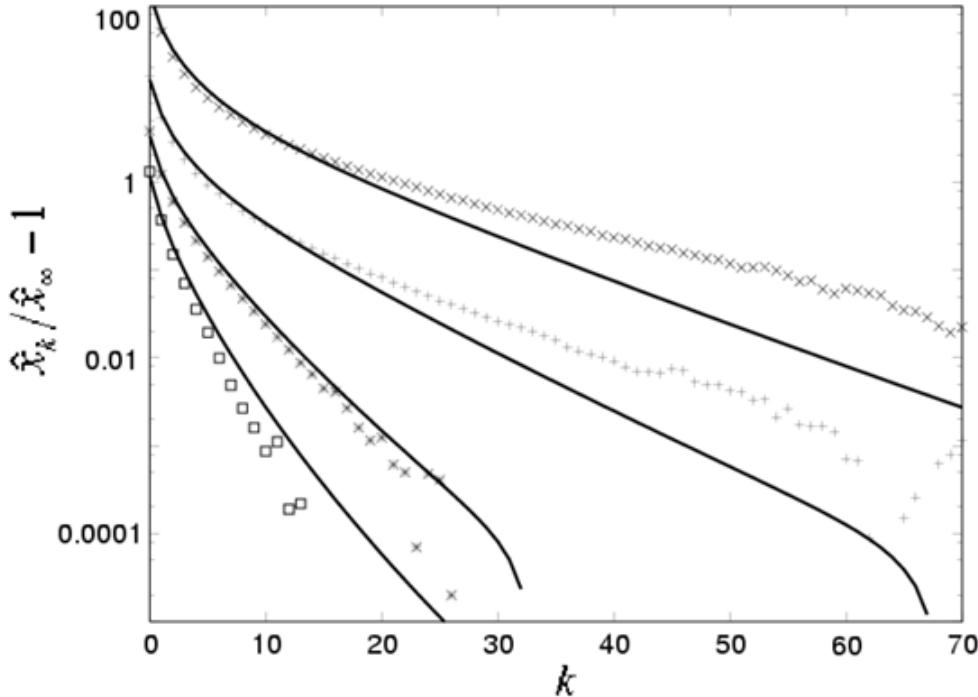


FIG. 26: Same plot as Fig. 25 but on a semi-log scale to show exponential decay at large  $k$ .

11. Semi-log Comparison for  $\hat{x}_k$  #2: The figure above shows the same data on a semi-log scale for  $\frac{\hat{x}_k}{\hat{x}_\infty} - 1$  showing the exponential decay at large arguments characteristic of a simple diffusion equation.
12. Validity of IIA - Large  $\epsilon$ : The IIA is clearly a good approximation for large  $\epsilon$ .
13. Validity of IIA - Small  $\epsilon$ : However, for small  $\epsilon$ , it starts deviating significantly from the simulations, especially for large  $k$ .
14. Price Impact from  $x_k$ : The values of  $x_k$  computed from the IIA can be very directly used to get an estimation of the price impact.



15. Price Impact Definition - Recap: The price impact, as defined in earlier sections, can be thought of as the change in the position of the midpoint – or the bid – consecutive to a certain number of orders being filled.
16. Price Impact after  $k$  Orders: Within the framework of the simplified model studied here, the quantity is

$$\langle \Delta m \rangle = \frac{1}{2} \sum_{k'=1}^k x_{k'}$$

after  $k$  orders.

17. Midpoint Impact Scaler: The factor of  $\frac{1}{2}$  comes from considering the change in the position of the midpoint and not the bid.
18. Nondimensional Impact vs  $\epsilon$  #1:

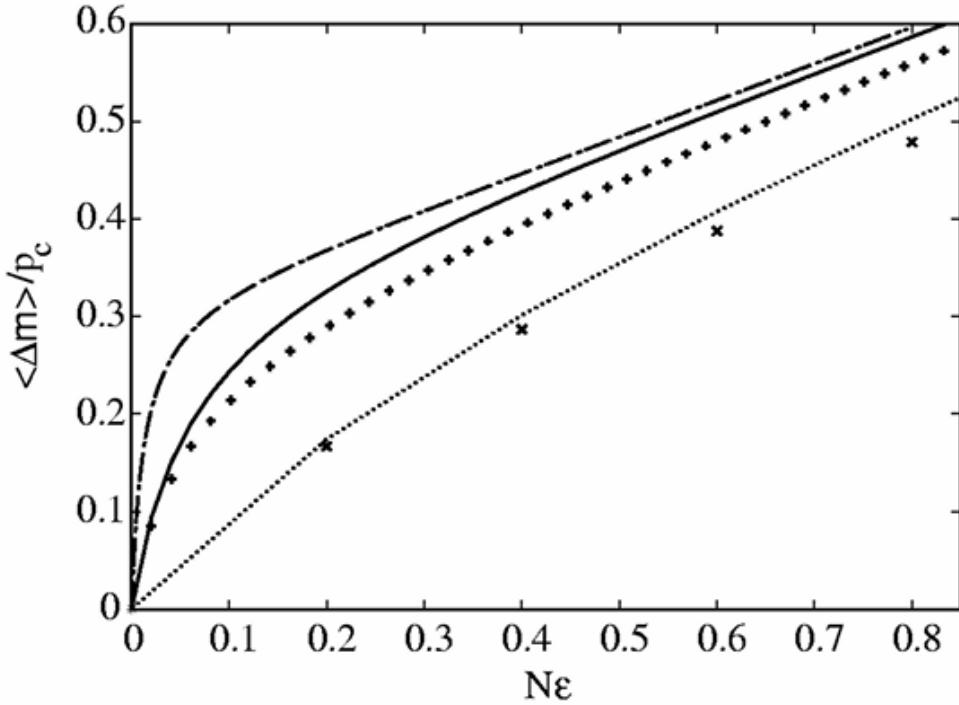


FIG. 27: Three pairs of curves for the quantity  $\langle \Delta m \rangle / p_c$  vs.  $N\epsilon$  where  $\langle \Delta m \rangle = 1/2 \sum_{k=1}^N x_k$ . The value of  $\epsilon$  increases from top to bottom ( $\epsilon = 0.002, 0.02, 0.2$ ). In each pair of curves, the markers are obtained from simulations while the solid curve is the prediction of the IIA. For  $\epsilon = 0.002$ , we show only the theoretical prediction. The theory captures the functional form of the price impact curves for different  $\epsilon$ . Quantitatively, its better for larger epsilon, as remarked earlier.

19. [Nondimensional Impact vs  \$\epsilon\$  #2](#): The figure shows  $\langle \Delta m \rangle$  nondimensionalized by  $p_c$  plotted as a function of the number of orders – by multiplied by  $\epsilon$  - for three different values of  $\epsilon$ .
20. [Comparison with Simulations](#): Again, the theory matches quiet well with the numerical simulations, qualitatively. For large  $\epsilon$ , the match is quantitative as well.



21. Order Interval Evaluation for  $x_k$ : The simplest approximation to the density profiles in the midpoint centered frame is to continue to approximate the mean density as  $\frac{1}{x_k}$ , but to regard that density as being evaluated at the position  $\frac{x_0}{2} + \sum_{k=1}^i x_k$
22. Evaluation Validity inside the Spread: This clearly is not an adequate treatment in the range of the spread, both because the intervals are discrete, whereas the mean  $\psi$  is continuous, and because the density profiles satisfy different global conservation laws associated with non-constancy of  $\alpha$ .
23. Evaluation Validity at Large  $k$ : For large  $k$ , however, this approximation may be valid.
24. Density vs Nondimensionalized Price #1:

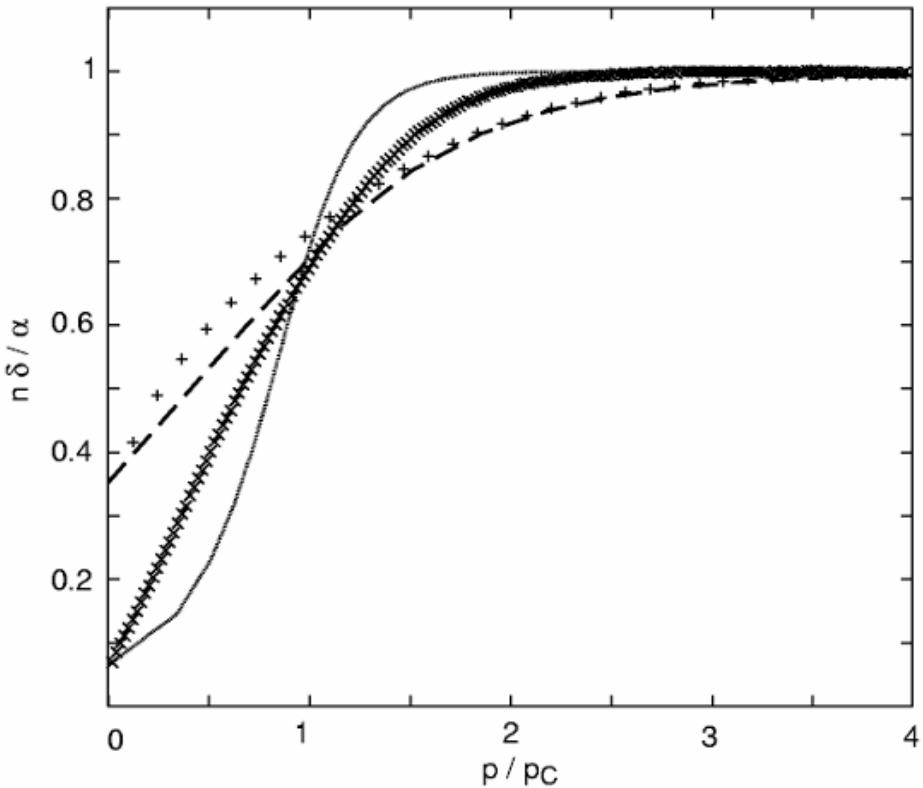


FIG. 29: Density profiles from Monte Carlo simulation (markers) and the Independent Interval Approximation (lines). Pluses and dash line are for  $\epsilon = 0.2$ , while crosses and dotted line are for  $\epsilon = 0.02$ .

25. Density vs Nondimensionalized Price #2: The mean-field values corresponding to a plot of  $\epsilon\psi(\hat{p})$  versus  $\hat{p}$  is shown above.
26. Numerical Density vs Monte Carlo #1:

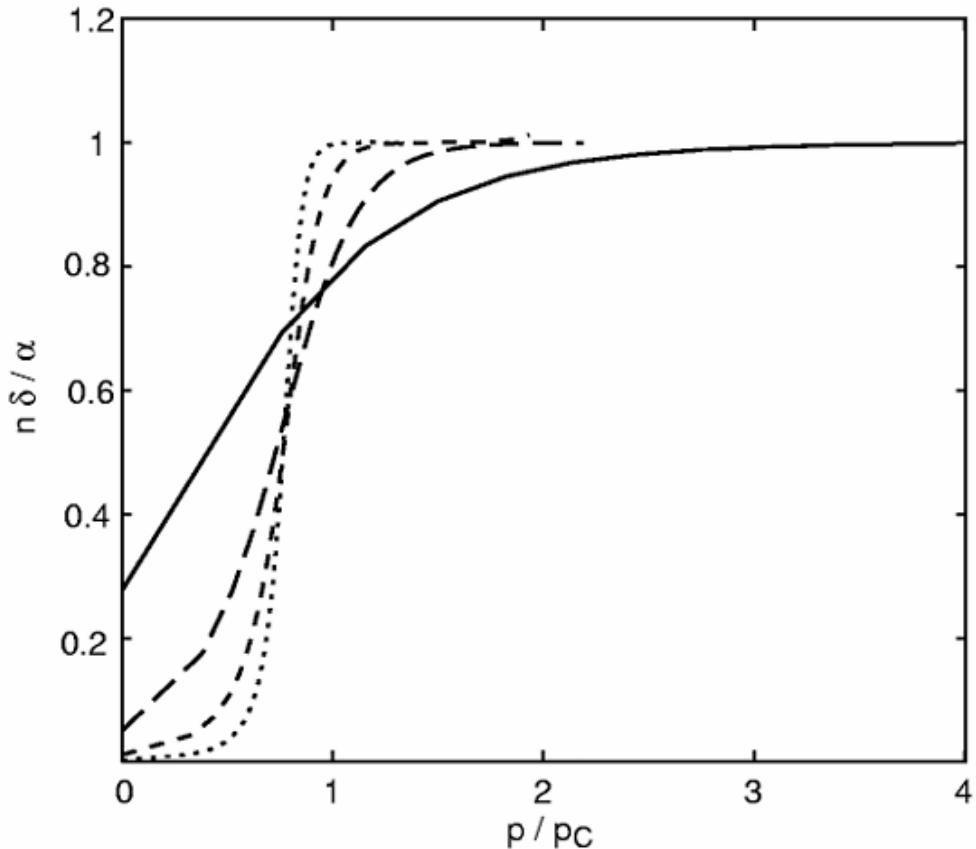


FIG. 28: Density profiles for different values of  $\epsilon$  ranging over the values 0.2, 0.02, 0.004, 0.001, obtained from the Independent Interval Approximation.

27. Numerical Density vs Monte Carlo #2: A comparison of the theoretically estimated profile against results from Monte-Carlo simulations is shown above.
28. Better Match at Large  $\epsilon$ : As is evident, the theoretical estimate for the density profile is better for large  $\epsilon$  rather than for small  $\epsilon$ .
29. Non-uniform Order Placement Process: The above analysis can also be generalized to when the order placement process is no longer uniform.
30. Power-law Order Placement Process: In particular, it has been found that a power-law order placement process is relevant (Bouchaud, Mezard, and Potters (2002), Zovko and Farmer (2002)).



31. Power law Explicit Form: The above analysis is carried out for when

$$\alpha = \frac{\Delta_0^w}{(\Delta + \Delta_0)^w}$$

where  $\Delta$  is the distance from the current bid and  $\Delta_0$  determines the *shoulder* of the power-law.

32. Existence of Solution across  $w$ : An interesting dependence on the presence of solutions exists across  $w$ .

33. Case of  $w > 1$ : In particular, for

$$w > 1$$

$\Delta_0$  needs to be larger than a certain value – which depends on  $w$  as well as other parameters of the model such as  $\mu$  and  $\delta$  - for solutions to IIA to exist.

34. Interpretation - Order Book Wipeout: This might be interpreted as the market order wiping out the entire book if  $w$  is too large.

35. Peak in the Depth Profile: When solutions exist, the depth profile has a peak, consistent with the findings of Bouchaud, Mezard, and Potters (2002).

36. Depth Profile vs Shoulder #1:

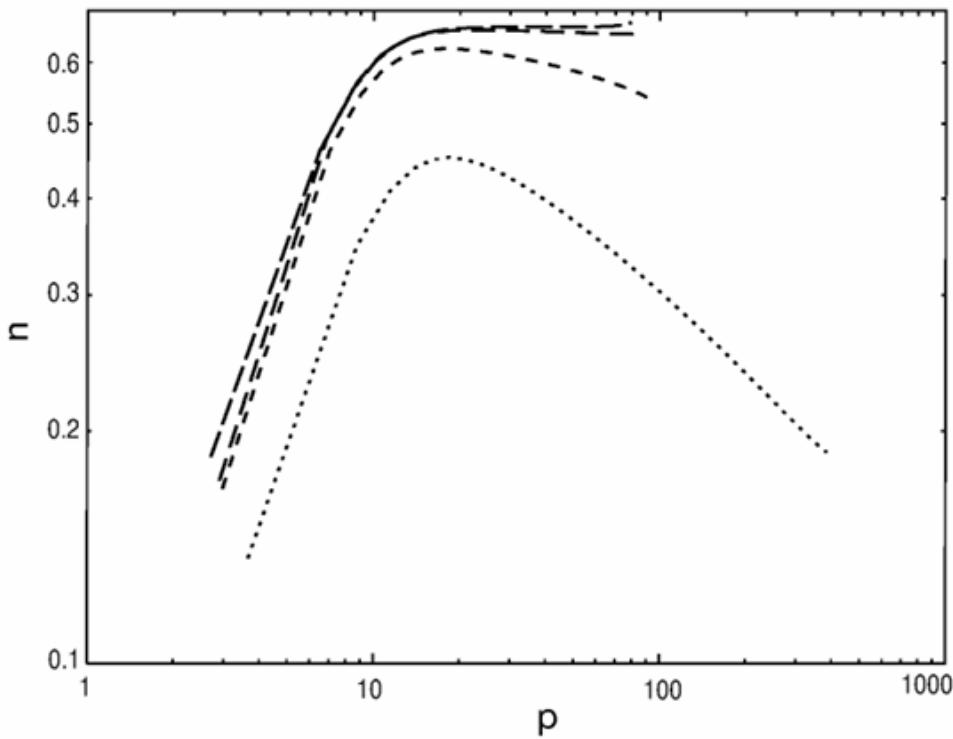


FIG. 30: Density profiles for a power-law order placement process for different values of  $\Delta_0$ .

37. Depth Profile vs Shoulder #2: The figure above shows the depth profile for three different  $\Delta_0$ .

## Concluding Remarks – Future Enhancements

1. Trending of Order Flow #1: It has been demonstrated that IID order flow necessarily leads to non-IID prices. The converse is also true: non-IID order flow is necessary for IID prices.
2. Trending of Order Flow #2: In particular, the order flow must contain trends, i.e., if the order flow has recently been skewed toward buying, it is more likely to be skewed toward buying.



3. Trending of Order Flow #3: If perfect market efficiency is assumed, in the sense that prices are a random walk, this implies that there must be trends in order flow.
4. Power-law Placement of Limit Prices #1: For both the LSE and the Paris Bourse, the distribution of the limit price relative to the best bid or the best ask appears to decay as a power-law (Bouchaud, Mezard, and Potters (2002), Zovko and Farmer (2002)).
5. Power-law Placement of Limit Prices #2: The investigations above show that this can have an important effect. Exponents larger than one result in order books with a finite number of orders.
6. Power-law Placement of Limit Prices #3: In this case, depending on other parameters, there is a finite probability that a single market order can clear the entire book.
7. Power-law or Log-normal Order Size Distribution #1: Real order placement processes have order size distributions that appear to be roughly like a log-normal distribution with a power-law tail (Maslov and Mills (2001)).
8. Power-law or Log-normal Order Size Distribution #2: This has important effects on the fluctuations in liquidity.
9. Near Poisson Order Cancelation Process #1: When considered in real-time, order placement cancelation does not appear to be Poisson (Challet and Stinchcombe (2001)).
10. Near Poisson Order Cancelation Process #2: However, this may be a bad approximation in event time rather than real time.
11. Conditional Order Placement: Agents may conditionally place large market orders when the book is deeper, causing the market impact function to grow more slowly.
12. Feedback between Order Flow and Prices #1: In reality there are feedbacks between order flow and price movements, beyond the feedback in the reference point for limit order placement built into this model.
13. Feedback between Order Flow and Prices #2: This can induce bursts of trading, causing order flow rates to speed up or slow down, and give rise to clustered volatility.



14. Feedback between Order Flow and Prices #3: This is one of many examples of how one can improve the model by making order flow conditional on available information.

## References

- Bak, P., M. Paczuski, and M. Shubik (1997): Price Variations in a Stock Market with many Agents *Physica A* **246** (3-4) 430-453
- Bollerslev, T., I. Domowitz, and J. Wang (1997): Order-flow and the Bid-ask Spread: An Empirical Probability Model of Screen-based Trading *Journal of Economic Dynamics and Control* **21** (8-9) 1471-1491
- Bouchaud, J. P., M. Mezard, and M. Potters (2002): Statistical Properties of Stock Order Books: Empirical Results and Models *Quantitative Finance* **2** (4) 251-256
- Bridgeman, R. W. (1922): *Dimensional Analysis* Yale University New Haven CT
- Challet, D., and R. Stinchcombe (2001): Analyzing and Modeling 1+1d Markets *Physica A* **300** (1-2) 285-299
- Cohen, K. J., R. M. Conroy, and S. F. Maier (1985): Order Flow and the Quality of the Market, in: *Market Making and the Changing Structure of the Securities Industry* (editors: Y. Amihud, T. Ho, and R. Schwartz) Lexington Books Lexington MA
- Daniels, M. G., J. D. Farmer, G. Iori, and E. Smith (2001): *How storing Supply and Demand affects Price Diffusion* arXiv
- Domowitz, I., and J. Wang (1994): Auctions as Algorithms *Journal of Economic Dynamics and Control* **18** (1) 29-60
- Eliezer, D., and I. I. Kogan (1998): *Scaling Laws for the Market Microstructure of the Interdealer Broker Markets* arXiv
- Gode, D. K., and S. Sunder (1993): Allocative Efficiency of Markets with Zero-intelligence Traders: Markets as a Partial Substitute for Individual Rationality *Journal of Political Economy* **101** (1) 119-137



- Iori, G., and C. Chiarella (2002): [A Simple Microstructure Model of Double Auction Markets](#)
- Maslov, S. (2000): Simple Model of a Limit-order Driven Market *Physica A* **278 (3-4)** 571-578
- Maslov, S., and M. Mills (2001): Price Fluctuations from the Order Book Perspective – Empirical Facts and a Simple Model *Physics A* **299 (1-2)** 234-246
- Mendelson, H. (1982): Market Behavior in a Clearing House *Econometrica* **50 (6)** 1505-1524
- Slanina, F. (2001): Mean-field Approximation for a Limit-order Driven Market Model *Physical Review E* **64** 056136
- Smith, E., J. D. Farmer, L. Gillemot, and S. Krishnamurthy (2003): Statistical Theory of Continuous Double Action *Quantitative Finance* **3 (6)** 481-514
- Tang, L. H., and G. S. Tian (1999): Reaction-diffusion-branching Models of Stock Price Fluctuations *Physica A* **264 (3-4)** 543-550
- Zovko, I., and J. D. Farmer (2002): [The Power of Patience: A Behavioral Regularity in Limit Order Placement](#)



## Limit Order Book Simulation: A Review

### Abstract

1. What are Limit Order Books: Limit Order Books – LOBs – serve as a mechanism for buyers and sellers to interact with each other in the financial markets.
2. Modeling and Simulation of the LOBs: Modeling and simulating LOBs is quite often necessary for calibrating and fine-tuning the automated trading strategies developed in algorithmic trading research (Jain, Firoozye, Kochems, and Treleaven (2024)).
3. Compute Power and AI Technologies: The recent AI revolution and availability of faster and cheaper compute power have enabled the modeling and the simulations to grow richer and even use modern AI technologies.
4. State of the Art LOB Simulation Models: This review examines the various kinds of LOB simulation models present in the current state of the art.
5. Model Classifications and Stylized Facts: It provides a classification of the models based on their methodology and provides an aggregate view of the popular stylized facts used in the literature to test the models.
6. Price Impact on the Models: It additionally provides a focused study on the price impact's presence in the models since it is one of more crucial phenomena to model in algorithmic trading.
7. Analysis of Fit Quality: Finally, it also carries a comparative analysis of the qualities of fits of these models and how they perform when tested against empirical data.

### Overview

1. Popularity of the Limit Order Books: The popularity of Limit Order Books in contemporary markets has been ever rising.



2. Real-time Data and Algorithmic Trading: With real-time data access provided by most exchanges and the rise of algorithmic trading, the order book and its history have become one of the most utilized forms of financial data.
3. Complex Nature of the Order Book Dynamics: The reasons for this are plentiful; some of them include the complex nature of supply and demand, which is captured in the time evolution of the order book.
4. Price Formation of the Security: The price formation of the security at the most granular level can be observed in the LOB, and the order book indicates the liquidity of the market, albeit not entirely.
5. Order Book and Trade History: Finally, the order book's history, in addition to the trade's history, enables practitioners to effectively replay history and perform simulations and back tests.
6. Review of LOB Properties: The review of Limit Order Books by Gould, Porter, Williams, McDonald, Fenn, and Howison (2013) focuses on studying the properties of the LOB but also showcases a number of models for LOB simulation.
7. Review of Empirical Tests: In another survey, Cont (2011) showed the utility of zero-intelligence models in LOB modeling and outlined several empirical observations as tests for the model's output.
8. Focus of this Chapter: This chapter focuses on the task of simulating order books using historical data. It surveys the recent developments across all major types of simulators and discusses each category's features and pitfalls.
9. Simulation Methodology and Summary Statistics: It also focuses on providing a brief summary of the methodology used in each simulation technique and provides a comparative study based on the stylized facts used to add priors to the simulator, test the simulator against empirical data, or both.
10. Responsiveness to Exogenous Trades and Market Impact: An in-depth analysis is also performed of one noteworthy aspect of LOB models – the responsiveness to exogenous trades or Market/Price Impact.



## Overview – Motivation

1. Challenges in Simulating the Order Book: There are a number of challenges in simulating the order book from issues related to model complexity, difficulty in replicating the statistic properties of the empirical data, and several mechanical issues stemming from the internal workings of the exchanges as halts in trading, open, intra-day, and close auctions, hidden orders, queue priority, and dark pools. For an in-depth analysis of the challenges faced in LOB modeling, refer to Gould, Porter, Williams, McDonald, Fenn, and Howison (2013).
2. Importance of Simulating LOBs: Despite, or due to, these challenges, modeling and simulating LOBs is of quite high importance for researchers and practitioners alike.
3. Use in Training/Back-testing: An especially noteworthy use case in the case of an LOB simulator is for back-testing – or training – algorithmic trading strategies. The reason being that having a simulator enables the availability of a richer set of data for the strategy to run on and be refined upon.
4. Overfitting and Poor OOS Performance: Since each security's price has had just one realization of the various possible time evolutions of its LOB dynamics, if the trading strategy were to be fitted on just this one trajectory, issues of over-fitting and thereafter true lack of true out-of-sample performance will be apparent (Sullivan, Timmermann, and White (1999), White (2000)).
5. Generation of Additional Training Data: Avoiding over-fitting can be done ideally by adding more data to the training set though generating more data is not trivial without knowing the generating process of the time series.
6. Use of Synthetic Data Simulators: One possible solution to this problem could be synthetic data using simulators. Training the strategy purely on simulators might introduce biases on the trading strategies since the simulator can never be perfectly representative of the real data.
7. Simulated Data Representative of Empiricals: This challenge can potentially be solved in two ways. The first and foremost is making sure that the simulator is



representative enough of the statistical properties of real world observed phenomena, i.e., ‘stylized facts’.

8. Parsimony behind Characterization of Empiricals: This in itself encompasses the entire field of order book simulations – can a simulator be built which can replicate the distributions of stylized facts observed in nature and at the same time be parsimonious?
9. Simulated/Real-world Data: The second is the usage of real-world data to do true out-of-sample testing of the trading strategy, or alternatively, combining the simulated data with the real-world in the training of the trading strategy.
10. Advantage of the Combined Simulations: This will enable the practitioner to effectively ground their strategies in reality and to avoid the pitfalls of back-testing purely on the historical data as well as avoid inducing biases because of the simulator’s lack of realism.

## Overview – Contributions

1. Models Based on their Techniques: This chapter breaks down the types of models using their core modeling technique: Point Processes, Agent-based Models, Deep Learning, and models using Stochastic Differential Equations.
2. Generative Modeling Techniques: In particular, there has been a recent rise of novel simulators with the onset of new generative modeling techniques such as Generative Adversarial Networks and its variants (Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville, and Bengio (2014), Mirza, Osindero (2014), Arjovsky, Chintala, and Bottou (2017)).
3. Empirically Observed LOB Metrics: Also studied are a variety of so-called ‘stylized facts’, or empirically observed metrics of LOB, that were used by the researchers as priors to develop their models and formulate a list of stylized facts that are the most important ones for applications in algorithmic trading.



4. Quality of Fit Tests: Also highlighted are the various quality of fit tests done in each simulator and how they compare with each other.
5. Simulator's Responsiveness to Exogenous Trades: Another important point on which the literature is critiqued is on the simulator's responsiveness to exogenous trades.
6. The Market Impact: This feature's importance stems from the fact that any practically applicable LOB simulator needs to be Market Impact aware as a zero Market Impact approximation may lead to a poor out-of-sample performance (Biais, Hillion, and Spatt (1999), Foucault, Kadan, and Kandel (2005), Cont, Kukanov, and Stoikov (2014)).

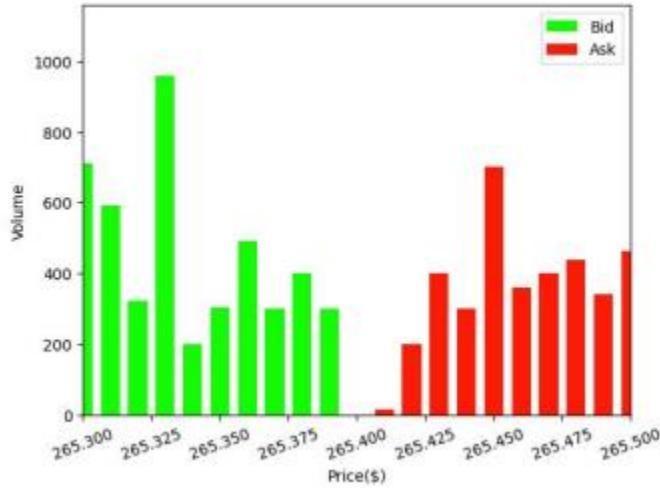
## Limit Order Book

1. Mathematical Description of Limit Order Books: This section provides a brief mathematical description of Limit Order Books and specifically how the time evolution of the order book dynamics can be described in a mathematical fashion. Abergel, Anane, Chakraborti, Jedidi, and Toke (2016) contain a detailed overview.
2. Price Levels of Order Book: The order book consists of discrete price levels at one *tick* difference from each other. The minimum difference in the price levels is known as the *tick-size* and is often specified by the regulators for each exchange.
3. Orders Associated with each Level: Each price level can have a non-negative integer number of ‘orders’ resting there with different ‘sizes’ and different ‘sides’. Orders here refer to the unexecuted – or ‘unmatched’ – limit orders at that price level.
4. What is an Order? Each order shows the intention of the market participant to trade at the specified price level, a quantity equaling the order size, and the direction of the trades – buy or sell – specified by their order sizes.
5. Unmatched Order in the Book: Since the order book consists of unmatched orders, the orders with intentions to buy are always at lower price levels than the orders with intention to sell.



6. Buy/Sell Order Book Sections: The buy section is known as the ‘bid’ and the sell section is known as the ‘ask’. The figure below shows aggregate order sizes at each price level for AAPL at a random time of the day.

Limit Order Book Volume for AAPL at 10:45:01.922806 2019-11-25



7. Best Bid and Ask Prices: The price level at the best bid is known as the ‘bid-price’  $P_B$ , and similarly at the best ask the ‘ask-price’ is  $P_A$ .
8. Spread: The distance between the bid and the ask in price units is called ‘spread’

$$S := P_A - P_B$$

9. Mid-price: The ‘mid-price’ is a theoretical price that signifies the average between  $P_A$  and  $P_B$

$$P = \frac{P_A + P_B}{2}$$

10. Categories of Orders in the Book: There are three categories of orders that can be placed in an order book: Limit orders which are a buy – respectively sell – orders which has a price level equal to lower – respectively higher – than the lowest ask –



respectively highest bid, market orders which are orders that match the bid/ask in the order book and remove the pre-existing limit orders, and finally cancel orders which are cancels of limit orders without any execution.

11. Hidden Liquidity, Orders, and Trades: Note that in many markets, not all of the liquidity, i.e., unexecuted orders, is displayed in the LOB, and we have the possibility of hidden orders, hidden executions, or hidden trades.
12. Non-stationary Nature of the Order Book: The order book is not stationary in time – it evolves with the three order types mentioned previously arriving randomly across the day.
13. Price Process of the Security: These orders evolve the order book which in turn evolves the price process of the security. Hence, the LOB dynamics is the most granular level of the price process formation.
14. Illustration of the LOB Evolution: The figure below portrays the top 10 levels of the LOB on each side and their time evolution for 5 minutes. The darkest hue corresponds to best bid/ask and the lighter hues reflect the deeper levels. Marks for the trades – both visible and hidden – are also seen.





## Limit Order Book – Dynamics

1. Horst and Paulsen (2017) Order Book State: Adopting from Horst and Paulsen (2017) in their treatment of the order book's time evolution, the order book state is defined as

$$S_t(x) := [B_t, A_t, v_{b,t}(x), v_{a,t}(x)]$$

where  $B_t$  and  $A_t$  are the best/ask prices respectively, and  $v_{b/a,t}(x)$  is the order book volume at  $x$  - in units of the respective currency – distance away from the mid-price  $P_t$ .  $t$  here is the discrete time index.

2. Types of Order Book Events: They define 8 types of events – bid/ask – which can change the state of an order book corresponding to Market Orders – bid/ask (A/E); in-spread Limit Orders – bid/ask (B/F); Cancel Order – bid/ask (C/G); and not-in-spread Limit Orders – bid/ask (D/H).
3. Level Depletion of Market Orders: Horst and Paulsen (2017) assume that Market Orders which do not deplete a price level are the same as Cancel Orders. This chapter does not make that assumption.
4. Evolution of the Order Book: The evolution of the order book at the event level can be described by the following:

$$S_{t+1} = S_t + \mathcal{D}_t(S_t)$$

where  $\mathcal{D}_t(\cdot)$  is a random operator which depends which depends on the dynamics that each 8 kinds of events induce on  $S_t$ .

5. List of the Dynamics Induced: The dynamics are induced, for example, if a queue-clearing buy/sell market order arrives, the ask/bid price moves up/down by one tick – or multiple ticks if the price levels near the best are empty; if a smaller market order arrives, it will change  $v_{b/a,t}$  but not the prices; if an in-spread bid/ask limit order arrives, the bid/ask price moves up/down ny one tick – or multiple ticks is the in-spread order is placed farther from bid/ask – and so on for the 8 types of events.



6. Order Book as a Markovian System: The order book state is partially observable and is often modeled as a Markovian system.
7. Statistical Properties of Order Books: Several statistical properties of the order books have been studied in the literature, and the following section provides a brief description of some of them relevant to the order book modeling.

## Stylized Facts

1. Definition of ‘Stylized Facts’: Stylized facts are the various observed statistical properties of the order book or one of the order book’s features such as mid-price, spread, etc.
2. Use of these Statistical Properties: Since the order book itself is partially observable, studying the properties of the order book dynamics is quite useful as well as the empirical ground truth to test the goodness of fit of the simulator.
3. Bouchaud, Mezard, and Potters (2002): In their seminal paper, Bouchaud, Mezard, and Potters (2002) outline two of the most important stylized facts in the literature – order flow statistics and average order book shape.
4. Cont (2011) Stylized Facts: Cont (2011)’s expansive survey lays down a number of other observations of the market.
5. Auto-correlation of Price Changes: Price changes, on a small enough timescale, are auto-correlated negatively at the first lag and then uncorrelated in further lags.
6. Trading Volumes Heterogeneity/Auto-correlation: Trading volumes are heterogenous and strongly auto-correlated.
7. Trading Volumes Exhibit Seasonality: Trading volumes exhibit strong intra-day seasonality.
8. Order Flow is Clustered in Time: This implies that the deviations between the orders are auto-correlated and there is a positive cross-correlation among arrivals of different order types.



9. **LOB Dynamics Stylized Facts**: These observations can be detected by measuring certain statistical properties of the LOB Dynamics which are the so-called stylized facts.
10. **Stylized Facts for Simulators**: More recently, Vyetrenko, Byrd, Petosa, Mahfouz, Dervovic, Veloso, and Balch (2020) provide a list of recommended stylized facts for order book simulators. Some of them are briefly described below.
11. **Empirical Distributions**: Densities of prices, returns – particularly of note is the long-tailed distributions of the returns, order volume, arrival rates, joint density of bid and ask queue sizes, time-to-fills.
12. **Auto-correlation of Returns**:  $\text{correlation}(r_{t+\tau, \Delta t}, r_{t, \Delta t})$  is the general formulation where  $\Delta t$  is the time step of the returns calculation – which exhibits behaviors like vanishing auto-correlation of returns at larger timescales; alternatively, the auto-correlation of returns is another interesting aspect where one sees a slower decay.
13. **Auto-correlation of Squared Returns**:  $\text{correlation}(r_{t+\tau, \Delta t}, r_{t, \Delta t})^2$  is the general taken to measure Volatility Clustering.
14. **Correlations**: Volatility and volume have positive correlations whereas volatility and returns have a negative correlation in empirical data.
15. **Intraday Seasonality**: Volumes have a signature intraday U-shaped seasonality with increased trading at Open and Close compared to mid-day.
16. **Signature Plots**: This is defined as the relation between volatility and sampling frequency – generally the empirical observation is that the signature plot decays quite slowly.
17. **Average Shape of the Book**: The mean of volumes at each price level with respect to the mid-price is the quantity of interest here.
18. **'M' vs. 'V' Shaped Books**: Generally, the so-called 'M' shaped average shape is observed for high-spread stocks while an inverted 'V' shape is observed for low-spread stocks.
19. **Price Paths**: The mid-price/ask-price/bid-price from the simulated order book is plotted against time in this stylized fact for a number of independent trials and comparisons are made to the empirical price paths observed.



20. Detailed Focus on the Stylized Facts: Further details are provided in a later section.

## Point Processes Models

1. Aggregate of Temporally Distributed Orders: The order book can be thought of, mathematically, as an aggregate of several individual orders arriving at different points of times.
2. Order Book as a Queue: It is hence quite natural to think of the order book as a queueing system and so there has been a plethora of models using Point processes to model the individual orders' arrivals.
3. Counting Associated with Point Process: A point process has an associated counting process  $N_{t,t+\Delta t}$  which is the number of events occurring in  $(t, t + \Delta t]$ .
4. The Point Process Intensity Function: With the usual conditions defined and satisfied on a complete probability space  $(\Omega, \mathcal{F}, \mathcal{P})$ , the intensity function  $\lambda_t$  is defined as

$$\lambda_t := \lim_{\Delta t \rightarrow 0} \frac{\mathbb{P}[N_{t,t+\Delta t} > 0 | \mathcal{F}_t]}{\Delta t}$$

## Point Processes Models – Point Process and Variants

1. Independence of Order Arrival Events: Poisson Process modeling assumes that order arrivals are independent of each other.
2. Point Processes as a Choice: There are several applications of the Poisson process in the queueing systems' literature so it is a very natural choice to model the LOB as a collection of Poisson processes as well.
3. 'Zero-Intelligence' Nature of Poisson: Usually, the Poisson process is classified as a 'zero-intelligence' model, i.e., a model that uses no prior information about the financial market or expert heuristics.



4. Improving on the ‘Zero-Intelligence’: However, as will be seen later, several ways of adding in the priors have been formulated in the literature to make the Poisson process models richer and more representative of the empirical observations.

## Point Process Models – Zero-intelligence Point Process and Variants

1. Zero-intelligence Models: Bouchaud, Mezard, and Potters (2002) show that, using a zero-intelligence model, they are able to match two stylized facts, which are the shape of the order book and order flow arrival statistics.
2. Extensions provided by Smith, Farmer, Gillemot, and Krishnamurthy (2003): Smith, Farmer, Gillemot, and Krishnamurthy (2003) treat the order flow of Limit Orders and Market Orders as sampling from a uniform probability distribution with Cancels occurring at constant probability unit time. They further develop a stochastic model of accumulated volumes from this order flow.
3. Asymptotic Behavior of the Poisson Model: Luckock (2003) showed the results of the first-order approximations of the Poisson arrival model. They show derivations for the depth of the order-book, time-to-fill, and optimal order type.
4. Poisson Arrivals for Order Processes: In Cont, Stoikov, and Talreja (2020), the authors develop a model for the LOB by assuming Poisson arrival processes for Limit Orders, Market Orders, and Cancellations.
5. Arrivals for Limit/Cancel/Market: The arrival rates for the Limit and the Cancel Orders depend inversely on the distance from the opposite side’s best quote; Cancels further depend on the number of outstanding shares at a level, and Market Order’s arrival rate is considered to be a constant.
6. Analytical Estimate for Order Metrics: The authors show, using Laplace transform, that the probabilities for the quantities of interest like the direction of price move, making the spread and the filling time conditional on the current state of the order book can be calculated analytically.



7. LLN for Order Book Shapes: Building on Cont, Stoikov, and Talreja (2010), Gao and Deng (2018) use fluid approximations of the above model and form a law of large numbers for the order book shapes.
8. Kelly and Yudovina (2018): Similarly, Kelly and Yudovina (2018) also use fluid approximations to study a model with Poisson arrivals with random price drawn from a stationary distribution of orders.
9. IID Price Formation Process: Further, in Cont and de Larrard (2011), the authors show that with a Poisson arrival queueing system for quote dynamics, the price dynamics of a security can be thought of as a sum of independently and identically distributed – IID – random variables, which, under the Central Limit Theorem, form a diffusion process.
10. Brownian Nature of Price Formation: Finally, Abergel and Jedidi (2013) show a detailed analysis of price dynamics converging to a Brownian motion with Poisson queueing system model of the LOB.

## Point Processes Models – Variable Order Intensity Poisson Processes and Variants

1. Non-constant Order Intensities: Moving away from the zero-intelligence approach, one way of addressing priors to the Poisson model is to have non-constant order intensities.
2. Hult and Kiessling (2010): Hult and Kiessling (2010), for example, use a Poisson arrival model with Limit and Cancel Orders' intensities being dependent on the distance from the mid-price while Market Orders' intensities is kept constant.
3. Hult and Kiessling (2010) Order Distribution: Further, to sample the orders' sizes, they use a stationary exponential distribution.
4. Markov Chain Version of the LOB: They use this model to build a Markov Chain Model of the LOB and further create optimal trading strategies using it.



5. Huang, Lehalle, and Rosenbaum (2015): Huang, Lehalle, and Rosenbaum (2015) study a variety of models for the bid and the ask queues around a fixed reference price – the first model they study is a Poisson process.
6. Queue -Reactive Model: They assume independence between bid and ask queues – here the arrival rates depend on the current queue size instead of being constant. This is what they call a *queue-reactive* model, which is quite popular in practice.
7. Lu and Abergel (2018b): Lu and Abergel (2018b) build on the above model and propose a non-Markovian order flow dynamic, albeit still using Poisson arrivals, by considering the order flow intensities to be dependent on not only the current state, but also on the previous history of the order flow which led to this current state.
8. Lu and Abergel (2018b) Unit Order Size: They also address the limitations of having a unit order size by considering the Limit Order sizes to follow a geometric distribution, Cancels a truncated geometric distribution, and Market Orders a mixture of geometric distributions with Dirac delta functions for multiples of 50 to account for traders' preference over round numbers.
9. Handling of Queue-Depletion Events: They further propose that, in the case of a queue depletion, the new limit order not only depends on the side of the cleared queue, but also on the past removal events.

## Point Process Models – Discussion on Point Processes and Variants

1. Inadequacy of Poisson Arrival Models: Despite their simplicity and vast variability, Poisson arrivals do not fit well with some of the observed stylized facts.
2. Auto-correlation between Order Durations: For example, the duration between orders is auto-correlated which leads to a clustering effect which the Poisson processes are unable to explain.
3. Assumption - Independence of Order Durations: The core issue seems to be the assumption that all orders are independent, which is generally contradictory to the practitioner's judgement.



4. Advantages of Queue-reactive Models: Queue-reactive models relax those assumptions to some extent, but are still lacking in some other assumptions which are highlighted in the next section, e.g., endogeneity of order arrivals. For a more detailed analysis, refer to Abergel and Jedidi (2011), and Cont (2011).
5. Popularity of the Poisson Models: However, due to their explainability and their mathematically convenient behavior under the scaling limits, Poisson models remain quite popular.

## Point Process Models – Hawkes Processes

1. Overcoming Shortfalls of Poisson Models: Hawkes process as a way of modeling the LOB queueing system proves to be solving some of these challenges that Poisson processes have.
2. Overview of Hawkes' LOB Usage: In their comprehensive review and tutorial, Bacry, Mastromatteo, and Muzy (2015) outlay the major ideas behind the Hawkes' process, its mathematical theory, some of its crucial properties, and finally applications including a detailed review of the order book models.
3. Empirical Fitting and Testing: Furthermore, they provide insights into the calibration methodologies for fitting and testing.
4. Improvements - Volatility Clustering and Epps Effect: The two major areas of significant improvement seen in the Hawkes Process compared to the Poisson methods is, first, the volatility clustering effect is observed in Hawkes, and second, the Epps Effect, which is the zeroing down the price of the covariance of assets in the limit of timescales to zero.
5. Endogenously Excited Order Flow: Hawkes processes inherently have endogenously excited order flow as well as an implicit form of market impact.
6. Market Impact of the Hawkes Process: A more detailed study of the Market Impact of the Hawkes process is presented in the section on Market Impact. Hawkes (2018) review the financial applications of Hawkes processes.



## Point Processes Models – Hawkes Process Mathematical Overview

1. Relaxation of Independent Poisson Arrivals: Hawkes processes relax the assumption of independent incremental arrivals in Poisson and instead use the fact that the order flow is endogenously excited in its modeling.
2. Multi-dimensional Hawkes Cross-Excitation Terms: A multi-dimensional Hawkes process can also have cross-excitation terms between different dimensions, e.g., a 2D Hawkes Process of  $\{\text{Ask Volume}, \text{Bid Volume}\}$  can have 4 excitation terms:  
 $\{\text{ask} \rightarrow \text{ask}, \text{bid} \rightarrow \text{bid}, \text{ask} \rightarrow \text{bid}, \text{bid} \rightarrow \text{ask}\}$
3. Mathematics of Hawkes Formulation: A brief mathematical description of the Hawkes Process formulation is presented here. The intensity function of the Hawkes process contains the self and the mutual excitation terms mentioned previously.
4. The Intensity and the Counting Processes: For a  $d$ -dimensional Hawkes process, the intensities of the process  $\lambda_{i,t}$  and the associated counting process  $N_{i,t}$  for

$$i = 1, \dots, d$$

are defined as

$$\lambda_{i,t} = \mu_{i,t} + \sum_{j=1}^d \left\{ \sum_{t_j \in T_j} \phi_{j \rightarrow i, t-t_j} \right\}$$

$$T_j := \{t_j : t_j \leq t\}$$

denotes the set of past event times in the  $j^{th}$  dimension of the Hawkes process.

5. Exogenous Intensity and Excitation Terms: Here  $\mu_{i,t}$  is the exogenous intensity of the  $i^{th}$  dimension and  $\phi_{j \rightarrow i, t-t_j}$  is the excitation term from the  $j^{th}$  dimension to the  $i^{th}$  dimension.



6. Time-dependent Excitation Terms: The excitation terms are a function of time from the incidence of the event – generally a decaying function in time like exponential decay or power law decay.
7. Continuous Counting Processes: An alternate but similar formulation is the following:

$$\lambda_{i,t} = \mu_{i,t} + \sum_{j=1}^d \left\{ \int_0^t \phi_{j \rightarrow i, t-s} dN_{j,s} \right\}$$

8. Choice of the Kernel Functions: There have been several discussions in the literature regarding the choice of the kernel functions. Nystrom and Zhang (2022) show that a power law kernel fits much better to the empirical data than exponential kernels.
9. Distributions of Empirical Inter-arrival Times: da Fonseca and Zaatour also show using a Q-Q plot the comparison between empirical inter-arrival times to the exponential distribution that exponentially decaying kernels are probably insufficient in representing the empirical data.

## Point Processes Models – Multi-dimensional Hawkes Process

1. Multi-dimensional Enhancement of Hawkes: In recent years, there has been a significant increase in LOB models using Hawkes Process albeit with vastly varied formulations.
2. Toke (2010) Market/Limit Orders: Toke (2010) creates a two-agent based model where the liquidity takers, i.e., market orders, and liquidity providers, i.e., limit orders, are each modeled as 1D Hawkes process.
3. Toke (2010) Cancel Orders and Price Distribution: Cancels are modeled to be Poisson arrivals and the price for the Limit Orders and the Cancels are sampled from a probability distribution.
4. Toke (2010) Model Performance: They compare performance improvement of using Hawkes process against Poisson process and show that the Hawkes process with the



three following excitations gives better fits to empirical data: Limit and Market Orders' self-excitation, and Market Orders exciting Future Limit Orders.

5. Bacry, Jaisson, and Muzy (2016): Bacry, Jaisson, and Muzy (2016) divide the events in the order book into two categories – those which change the mid-price and those which do not.
6. Bacry, Jaisson, and Muzy (2016) 8-dimensional Hawkes: They use an 8-dimensional Hawkes process with orders changing the mid-price being modeled by one dimension, and for events which don't change the mid-price are modeled by 3 dimensions, i.e., Market Order, Limit Orders, and Cancellations. This is done for both the bid and the ask sides giving us a total of 8 dimensions.
7. Large (2007): Previously, Large (2007) followed a similar technique by using a 10D Hawkes Process:  $\{(Limit\ Order, Market\ Order)\} \times (Change\ Mid, Don't\ Change\ Mid), (Cancel\ Order)\} \times (Bid \times Ask)$
8. Large (2007) Order Book Resiliency: They formalize the *resiliency* of the order book which is the ability of the order book to replenish after being depleted by a large trade.
9. Kirchner (2017): Kirchner (2017) proposed an alternative, non-parametric way of estimating the Hawkes process and showed the applicability of their method in LOB data.
10. Kirchner (2017) Model Selection: Particularly noteworthy is the technique they show in model selection and the usage of the AIC statistic to optimize the hyperparameters.
11. da Fonseca and Zaatour (2014): da Fonseca and Zaatour (2014) propose an alternate strategy to fit the Hawkes process by using the generalized method of moments to fit the first four moments of various quantities of interest.
12. da Fonseca and Zaatour (2014) Estimation Speed: This method is claimed to be much faster than the traditional Maximum Likelihood Estimation MLE methods used in literature.
13. da Fonseca and Zaatour (2014) Parameter Convergence: They show the weak convergence of the fit parameters to the true unknown parameters of the Hawkes process.



14. da Fonseca and Zaatour (2014) Tests: They use several key stylized facts to test the realism of their simulations. They also compare the fit parameters to the MLE baseline.

## Point Processes Models – Constrained Hawkes Process

1. Zheng, Roueff, and Abergel (2014): Zheng, Roueff, and Abergel (2014) create a 4-dimensional Hawkes process for Level 1 Order Book simulation with two for each of bid and ask queues and they construct a spread process to control events where bid becomes greater than ask. Thus, they create and provide analysis for a Hawkes process with constraints.
2. Lee and Seo (2022): Lee and Seo (2022) create a 4D Hawkes process for Level 1 LOB simulator similar to Zheng, Roueff, and Abergel (2014), but instead of constructing a separate spread process, they propose that the exogenous intensity of the spread-narrowing events is a function of the spread relative to the price.
3. Lee and Seo (2022) Single Tick Excitation: This implies that at 1-tick wide spread, the exogenous intensities of the spread-narrowing events is zero.
4. Lee and Seo (2022) Kernel Modeling: Further, they also let the decay kernels' – for self and cross excitation – to be again dependent on spread, but also stochastic. This ensures that the intensity of the spread-narrowing events is exactly zero when the spread is 1-tick wide.
5. Properties of Spread/Price Process: They derive some properties of the price and the spread processes. Further, they provide some techniques for estimating the estimator's biases and also provide a comprehensive empirical study and derive several economic explanations for the observed phenomena.

## Point Processes Models – Other Hawkes Process Variants



1. Kaj and Caglar (2017): Kaj and Caglar (2017) model the order book events in the following manner: Market orders as queue-reactive Poisson arrivals, Limit orders as Hawkes with excitation from Market orders, and Cancels as constant intensity Poisson arrivals. They call this a *Buffer-Hawkes process*.
2. Morariu-Patrichi and Pakkanen (2022): Morariu-Patrichi and Pakkanen (2022) develop a state-dependent Hawkes process where two different types of states are considered; first, on the basis of spread being one-tick, and second, on the basis of order flow imbalance.
3. Morariu-Patrichi and Pakkanen (2022) State-dependent Excitation: They conclude that the excitation effects are observed to be highly dependent on the current state.
4. Morariu-Patrichi and Pakkanen (2022) Kernel Analysis: They perform a number of analyses on their fitted Hawkes process to infer the economic rationale behind the observed effects in the kernels of the Hawkes process.
5. Kirchner and Vetter (2022): Kirchner and Vetter (2022) also formulate a marked state-dependent Hawkes process but they use non-parametric methods to estimate the excitation kernel's shape, although they use power laws to fit the shape later.
6. Kirchner and Vetter (2022) Imbalance as State Indicator: They use the current imbalance as the state indicator, and they also work towards creating a parsimonious model by zeroing out smaller excitations they observe in the data.
7. Muccianti and Sancetta (2023): Another state-dependent model was proposed by Muccianti and Sancetta (2023) where they consider the intensity as a product of a Hawkes process driven intensity and a linear function on some observables in the market environment.
8. Muccianti and Sancetta (2023) Time-of-the-day Incorporation: A key feature of this treatment is the incorporation of the time-of-day into the modeling – it is well-known that the order arrival intensities are non-stationary intra-day and therefore most models clip the data to exclude open and close effects.
9. Wu, Rambaldi, Muzy, and Bacry (2019): Wu, Rambaldi, Muzy, and Bacry (2019) take inspiration from Huang, Lehalle, and Rosenbaum (2015) queue reactive model to build a Hawkes process with exogenous intensity being queue-reactive in one model,



and a queue-reactive multiplier on top of the Hawkes process intensity in the second model.

10. Wu, Rambaldi, Muzy, and Bacry (2019) Queue-reactiveness Impact: They show that adding queue-reactiveness improves the goodness of fit models against Huang, Lehalle, and Rosenbaum (2015) as well as the non-queue-reactive Hawkes model.
11. Rambaldi, Bacry, and Lillo (2017): Rambaldi, Bacry, and Lillo (2017) show that the order size – not just the order count – of each individual order is important in the excitation of future orders.
12. Rambaldi, Bacry, and Lillo (2017) Marked Hawkes: They show that, using a marked Hawkes process, with the marks corresponding to the various bins of the order sizes, they are able to follow the Bacry, Jaisson, and Muzy (2016) non-parametric estimation technique to create a more realistic simulator.

## Point Processes Models – Hawkes Process Scaling Limits

1. Abergel and Jedidi (2015): In their influential paper, Abergel and Jedidi (2015) create a model of the full LOB where each level's Market Order and Limit Order intensities are modeled as Hawkes Processes and Cancels with a queue-reactive Poisson intensity.
2. Abergel and Jedidi (2015) SDE Form: They show through a mathematical analysis that this model can be used to create Stochastic Differential Equations SDE for aggregate features.
3. Horst and Xu (2019): Horst and Xu (2019) further show that under some scaling limits, the Hawkes model for an LOB converges to an SDE for bid and ask prices while intraday volume follows a system of ordinary Differential Equations ODE.
4. Horst and Xu (2019) Stationary Intensities: They also show that the stationary intensities of the different types of events form Volterra-Fredholm Integral Equations.



## Point Processes Models – Non-linear Hawkes Process

1. Lu and Abergel (2018a): Lu and Abergel (2018a) create a 12D Hawkes process:  
 $\{Limit\ Order, Market\ Order, Cancels\} \times \{Change\ Mid, Don't\ Change\ Mid\} \times \{Bid, Ask\}$
2. Lu and Abergel (2018a) Linear/Non-linear Hawkes: They compare the performance of a Linear and a non-linear Hawkes process with the linear one having novel *inhibiting* kernels for negative excitation. They floor the intensities of the non-linear Hawkes model to zero.
3. Lu and Abergel (2018a) Sum of Exponential Functions: Another notable novelty in their research is the use of the sum of exponential functions with varying half-lives as kernels.
4. Mounjid, Rosenbaum, and Saliba (2019): Mounjid, Rosenbaum, and Saliba (2019) also create a non-linear Hawkes process to simulate the order book with the non-linear transformation being dependent on the event type, the current type of the LOB, the current time, and a sum over the past event's excitations.
5. Mounjid, Rosenbaum, and Saliba (2019) Excitation Kernels: These excitation kernels are allowed to depend on the event type and the current state of the order book.
6. Mounjid, Rosenbaum, and Saliba (2019) Framework Comparison: They perform a mathematical comparison of this framework with different kinds of intensity models: Poisson, queue-reactive Poisson, Hawkes, and Quadratic Hawkes.

## Point Processes Models – Neural Hawkes Process

1. Kumar (2021): More recently, Kumar (2021) developed a Deep Neural Hawkes Process for Market Making in a simulated order book. The order book is simulated through a combination of agent-based traders and sampling a Hawkes Process fitted to historical data.



2. Kumar (2021) Improvement using LSTM: They use LSTMs to improve upon the Neural Hawkes Process proposed by Mei and Eisner (2017). Their hypothesis is that LSTMs are able to capture the more complex dynamics of feedback loops between various orders in the market since they inherently have this feature in their structure.
3. Shi and Cartlidge (2022): Shi and Cartlidge (2022) develop a neural Hawkes process with each type's intensity being modeled by continuous time LSTM units. The process' intensity rates evolve in such a way that the current market state influences it. They draw the price and the size of the order from stationary distributions in their simulations.

## Point Processes Model – Hawkes Process Discussion

1. More Comprehensive Point Process Methodology: Hawkes process, with their high adaptability, provide a more comprehensive point process methodology to model the order book arrivals without having to model the individual traders' behaviors in the market.
2. Reproducing the Microstructure: More importantly, their ability to reproduce important microstructure details like volatility clustering and Epps effect make them great candidates for the LOB models.
3. Ease of Deriving Intuition: Since the point process models can be described mathematically, they are fully explainable in their nature and hence are suitable for applications where black-box solutions are not preferred.
4. Bacry, Bompaire, Gaiffas, and Poulsen (2017): Recently, Bacry, Bompaire, Gaiffas, and Poulsen (2017) have published a Python library for calibrating Hawkes process.
5. Difficulty in Model Calibration: The key challenge that the practitioner may face in using the Hawkes process is the difficulty of calibration of these models. This stems from the fact that the likelihood function is quite complex.
6. Choice of the Excitation Kernels: In addition, the choice of the excitation kernels in the Hawkes process can make a large difference in the model's predictive power.



7. Model Parsimony: Further, the question of model parsimony becomes quite relevant since the number of kernels scale as  $\mathcal{O}(n^2)$  for  $n$ -dimensional Hawkes Process.

## Agent-Based Models

1. Large Number of Heterogenous Agents: Considering that the order book is constituted by a large number of heterogenous agents, examples of heterogeneity being differences in their trade frequency, trading objectives, access to financial data, and access to low latency trading hardware, the key idea in this category of LOB models is that each agent needs to be modeled in a separate category.
2. Distinct Clusters of Agent Types: For example, Cont, Cucuringu, Glukhov, and Prenzel (2023) make use of clustering techniques to show from anonymized trade execution data that there exist at least 4 different clusters of agents in the lit LOB.
3. Categories of Participating Agents: The usual set of categories in addition to the informed vs. uninformed traders include high-frequency traders, trend followers, mean reverters, noise traders, and algorithmic traders.
4. LOB Agent Model Reviews: The reader is referred to expansive reviews by Chakraborti, Toke, Patriarca, and Abergel (2011) and Abergel, Anane, Chakraborti, Jedidi, and Toke (2016) on Agent-based models.

## Agent-based Models – Recent Work

1. Paddrik, Hayes, Todd, Yang, Beling, and Scherer (2012): Paddrik, Hayes, Todd, Yang, Beling, and Scherer (2012) use order speed and order characteristics to identify and model various types of agents.
2. Paddrik, Hayes, Todd, Yang, Beling, and Scherer (2012) Flash Crash Simulation: They use the 6 May 2010 *Flash Crash* of E-Mini S&P Futures to support their claim that many agents behave in a correlated manner.



3. Paddrik, Hayes, Todd, Yang, Beling, and Scherer (2012) Traders' Categories: They create 6 categories of traders ranging from fundamental traders trading at very low frequency to market makers of HFTs. They model each category to be zero-intelligence Poisson process.
4. Huang, Lehalle, and Rosenbaum (2015): Huang, Lehalle, and Rosenbaum (2015) assume in their second model that institutional agents post their limit orders at the top of the book while HFTs, market makes, and arbitrageurs post it in deeper levels.
5. Huang, Lehalle, and Rosenbaum (2015) Level's Order Intensity: Hence, they propose that the order level of an intensity depends on whether the level is the best bid/offer or not.
6. Huang, Lehalle, and Rosenbaum (2015) Arrival Rate: Further, they enhance the model by adding the order arrival rate's dependency on the opposite queue size by discretizing the opposite queue size into 4 categorical quantiles.
7. Huang, Lehalle, and Rosenbaum (2015) Reference Price Change: In their queue-reactive model they further relax the assumptions by allowing the mid-price or the reference price to change by one tick at some constant probability, and with a constant re-initialization probability event.
8. Byrd, Hybinette, and Balch (2019): Byrd, Hybinette, and Balch (2019) propose a software framework for simulating tens of thousands of agents with various types of objectives and trading patterns. They also introduce latency and an exchange agent for transactions to make the simulator more realistic.
9. Belcak, Calliess, and Zohren (2020): Belcak, Calliess, and Zohren (2020) create a software package ABIDES and provide a Python API with C++ at the backend of the simulations.
10. Belcak, Calliess, and Zohren (2020) Market Impact and Summary Statistics: They further study a number of simulation statistics and also provide a methodology to measure market impact with temporary and permanent components.

## Combining ABMs with Other Models



1. Lehalle, Gueant, and Razafinimanana (2011): Lehalle, Gueant, and Razafinimanana (2011) describe the drawbacks of both ABM – computational constraints and lack of analytical results – and Point Process Modeling – stationarity assumption and imperfect representation of stylized facts – and propose using a mixed model.
2. Lehalle, Gueant, and Razafinimanana (2011) Zero-intelligence Pegged to an ABM: They create a zero-intelligence model – conditioned on the distance between the investor's view of the order book and the real order book – *pegged* to an ABM with scaling limits taken as a Mean Field Game.
3. Kumar (2021): As detailed in the previous section, Kumar (2021) uses a hybrid approach for modeling the LOB with Hawkes as the background process for different types of agents to interact with.
4. Kumar (2021) Description of the Agents Represented: The propose segregating the market participants into the following classes:
  - a. The fundamental trader who flows a mean-reversion strategy
  - b. The chartist trader who follows a momentum strategy
  - c. The noise trader
  - d. Three different kinds of market makers:
    - i. One which uses the Deep Hawkes Process to quote bid-ask orders
    - ii. Second which uses the Neural Hawkes Process proposed by Mei and Eisner (2017)
    - iii. Third being a probabilistic market make whose order placement is based on their view of the fundamental price of the security.
5. Shi and Cartlidge (2023): Shi and Cartlidge (2023) show that combining a stochastic simulator for the background process of an LOB with a multi-agent simulation built on top of this background simulator has benefits over ABMs or pure stochastic models.
6. Shi and Cartlidge (2023) Neural Hawkes + ABIDES: They create a Neural Hawkes Process for the background simulator and use the ABIDES platform (Byrd,



Hybinette, and Balch (2019)) for the multi-agent simulation. They perform studies on the price impact and observe herding behaviors in their simulations.

## Agent Based Models – Discussion

1. The Econophysics Modeling Approach: The interplay between the plethora of market participants has naturally led the LOB to be modeled in a statistical physics way. Agent-based modeling of the LOB rise from the popularity of econophysics modeling.
2. Capture of Agent Heuristics: The common difficulty with this kind of modeling is the heavy use of heuristics in defining the behavior of an individual agent or a class of agents.
3. Computational Cost of Agent Simulations: In addition to that, the computational cost of simulating individual agents is higher than that of the alternatives. Although, Mean Field Games analysis of the ABM system does help in the analytical tractability of this category of models.
4. Mixed ABMs and Background Models: The usage of ABMs in combination with background models is a promising area of future research since that combines the best out of both these contrasting modeling techniques.

## Deep Learning Based Models

1. Popularity of Large Parameterized Models: Owing to several sources of possible complexities and non-linearities of the order book as well as the distribution of prices/returns and volumes, large parametrized models like deep learning networks have found recent surge in popularity in LOB simulation.



2. Predictive Power of Neural Networks: There has been a significant amount of research done to use the predictive power of neural networks for predicting the mid-price, the volatility, and the direction of price moves.
3. Popular Architectures Considered: Some of the more popular architectures considered are Convolutional Neural Networks CNNs, Long Short-Term Neural Networks LSTMs, Recurrent Neural Networks RNNs, and Generative Adversarial Networks GANs.
4. Goodfellow, Bengio, and Courville (2016): For a detailed description of these architectures, the reader is referred to Goodfellow, Bengio, and Courville (2016).
5. Copponi and Lehalle (2023): For a focused review of machine learning applications encompassing both the traditional machine learning and the deep learning models in finance, the reader is referred to Copponi and Lehalle (2023).

## Deep Learning based Mid-price Prediction from LOB

1. Sirignano and Cont (2018): Sirignano and Cont (2018) use deep learning techniques like LSTMs to model the price formation mechanism with historical price and order flow as inputs.
2. Sirignano and Cont (2018) Path-dependent Price Dynamics: They show that their price dynamics is highly path-dependent since increasing performance was observed with increasing history.
3. Sirignano and Cont (2018) Modeling the Mid-price: Although they do not model the limit order book state explicitly but rather model the next mid-price which is just a property of the order book as a whole, their universality results shown the promise of deep-learning in ingesting tick data.
4. Zhang, Zohrer, and Roberts (2019): Zhang, Zohrer, and Roberts (2019) use deep learning structures like CNNs coupled with LSTMs and Inception Modules to predict future price movements from the current state of the order book.



5. Zhang, Lim, and Zohren (2021): Zhang, Lim, and Zohren (2021) use Deep Learning on Market by order data – Level 3 data – to predict future price movements’ category among up, down, or flat.
6. Briola, Turiel, and Aste (2020): A detailed comparative analysis on price prediction from LOB states is presented Briola, Turiel, and Aste (2020).

## Deep Learning Based Models – Recurrent Neural Networks

1. Order Book Level Volume Prediction: Shi, Chen, and Cartlidge (2021) make use of Recurrent Neural Network RNN structures like the Gated Recurrent Unit GRU and on ODE-RNN – Ordinary Differential Equations RNN – to predict the volume at the different levels of the order book. They use top of the book data – Level 1 – to simulate 5 levels of data – Level 2.
2. Importance of ODE-RNN: The authors claim that the ODE-RNN usage here is of particular importance since the traditional RNN’s are unable to handle non-uniform time intervals in history.
3. Cross-Security Transfer Learning: Further, they use transfer learning to show that the parameters learnt by training the network with one security’s data can be fine tuned to a different security’s data to get reasonably good performance.
4. Shi and Cartlidge (2021): Further, in Shi and Cartlidge (2021), they propose the usage of exponential kernels instead of the ODE kernels to make the model more parsimonious and to reduce the computational cost.
5. Shi and Cartlidge (2021) Bias Removal: They enrich their testing inverse by using a wider set of stocks and they remove look-ahead biases from their previous model.
6. Shi and Cartledge (2021) Volume Prediction Accuracy: They find that the order volume prediction accuracy decreases with increase in volatility.
7. Kumar (2021): As mentioned in the previous sections, Kumar (2021) use LSTMs in their Hawkes process model to capture more complex feedback loops dynamics which exist in various event types in the market.



## Deep Learning Based Generative Models

1. Finance Time Series using GANs: Takahashi, Chen, and Tanaka-Ishii (2019), Wiese, Knobloch, Korn, and Kretschmer (2020), Ni Szpruch, Sabate-Vidales, Xiao, Wiese, and Liao (2022) use GANs and its variants to generate financial time series.
2. Wiese, Knobloch, Korn, and Kretschmer (2020): Particularly noteworthy is Wiese, Knobloch, Korn, and Kretschmer (2020)'s use of the DY metric (Dragulescu and Yakovenko (2002)) to test the performance of the generative network.
3. Li, Wang, Lin, Sinha, and Wellman (2020): Li, Wang, Lin, Sinha, and Wellman (2020) use conditional Wasserstein GANs to create the stock-GAN model which simulates the orders in the market by conditioning on some finite window of historical orders.
4. Li, Wang, Lin, Sinha, and Wellman (2020) LSTM to encode History: They use an LSTM to encode the history and claim that the time dependence of the order flow intensity is captured by this recurrent network.
5. Li, Wang, Lin, Sinha, and Wellman (2020) Continuous Double Auction Neural Network: They further add a continuous double-auction approximation neural network to evolve the order book from the order streams that are simulated.
6. Lim and Gorse (2021): Lim and Girse (2021) use Sequence GANs – SeqGANs – to model the order flow. They argue that SeqGAN is a better choice in handling discrete sequences of events like order flow than conditional GANs.
7. Lim and Gorse (2021) Order Book Simulation: They model the order book data using SeqGAN and further apply this simulation to do an analysis of the macro-level mid-price movements in this model.
8. Prenzel, Cont, Cucuringu, and Jochems (2022): Prenzel, Cont, Cucuringu, and Jochems (2022) created a methodology to calibrate GANs for order flow data dynamically for different market conditions instead of a single calibration over the entire dataset.



9. Prenzel, Cont, Cucuringu, and Jochems (2022) Dynamic Intensity Estimation: They assume that the order flow follows Poisson arrivals but rather than setting the intensity to a constant value, they use GANs to estimate the probability distribution function of the intensities for different conditions such as the time of the day and market volatility.
10. Cont, Cucuringu, and Kochems (2023): Cont, Cucuringu, and Kochems (2023) model the transition between two LOB snapshot using conditional Wasserstein GANs – conditional on the current state of the LOB, therefore making the simulation Markovian.
11. Cont, Cucuringu, and Kochems (2023) Market Impact Capture: As shown in the next section, they particularly focus on creating a model which has implicit market impact in its order book transitions.
12. Cont, Cucuringu, and Kochems (2023) Simulations vs Empiricals: They provide a thorough analysis of their models by comparing the simulations' and the real-world's empirical facts.
13. Coletta, Moulin, Vyetrenko, and Balch (2022): Colette, Moulin, Vyetrenko, and Balch (2022) make use of conditional GANs – CGANs – to create an LOB model – *world model* – which is compared against a baseline ABM.
14. Coletta, Moulin, Vyetrenko, and Balch (2022) Trader Action: They train the CGAN to generate the next trading action given the current features of the state of the order book.
15. Coletta, Moulin, Vyetrenko, and Balch (2022) Adversarial Attacks on CGAN: They further perform *adversarial attacks* on CGANs model in Coletta, Jerome, Savani, and Vyetrenko (2023) to highlight the dependence of the model on its input features.

## Deep Learning Based Large Language Models



1. Nagy, Frey, Sapora, Li, Calinescu, Zohren, and Foerster (2023): Nagy, Frey, Sapora, Li, Calinescu, Zohren, and Foerster (2023) use the recently popular auto-regressive generative models on the order book message data to simulate order flow.
2. Nagy, Frey, Sapora, Li, Calinescu, Zohren, and Foerster (2023) LM Order Book Modeling: They tokenize the LOB messages and treat sequences of these tokens to simulate order flow as a Large Language Model LLM would treat words in a language to create a comprehensible sentence. They perform several out-of-sample tests on their simulator to test its efficiency.

## Deep Learning Based Models Discussion

1. Handling Complexity using Deep Learning: Deep learning models for simulating the order book are natural candidates to solve for the vast complexity of the order book dynamics.
2. Reason for Deep Learning Popularity: The ability of deep neural networks to model the convoluted time evolution of Markov processes coupled with astronomical increase in recent years of the ease of training such models has popularized this category of models.
3. Reproduction of the Stylized Facts: It is to be noted that a number of these models are able to reproduce the stylized facts quite well in their simulations.
4. Concave Market Impact: Some of them also exhibit the concave Market Impact characteristic that practitioners observe in the real world.
5. Black-box: Lack of Explainability: However, these models, like any other deep learning model, present many challenges such as the lack of explainability owing to their black-box nature, high sensitivity to carefully calibrated hyperparameters, and a very high model complexity with millions – or even billions – of parameters.
6. Evolving the Order Book State: A more parsimonious, explainable way of order book modeling is to model the transition of the order book itself by taking the continuous time limit and creating a differential equation evolution of the stochastic process.



## Stochastic Differential Equations Based Models

1. Time Evolution of an LOB State: Since the LOB transitions are probabilistic in nature, the time evolution of the LOB state can be modeled as a set of differential equations.
2. Continuous Approximation of State Transition: Here, usually a continuous approximation is made of the state transition in time which, although is quite different from the reality of discrete time steps, serves in larger time scales as an approximation since the frequency of LOB events is quite high.
3. Long-Time Order Book Dynamics: A major focus of this set of models is studying the long-time dynamics of order book – the so-called steady state or the absence of it – is a major point of interest.
4. Popular Variants of SDE Used: Some of the popular types of components used the differential equations include diffusion and convection.
5. Explaining the LOB Dynamics: These set of models is particularly important if one is concerned about the explainability of the LOB simulation.
6. Integration with Optimal Control Schemes: They also provide good segues into utilizing model dynamics for optimal control problems like portfolio management, wealth management, optimal liquidation, and market making.

## Stochastic Differential Equations Based Models – Continuous Limit of Point Process

1. Korolev, Chertok, Korchagin, and Zeifman (2015): Korolev, Chertok, Korchagin, and Zeifman (2015) model the order flow arrivals as a Cox process – double stochastic Poisson arrivals – and form stochastic differential equations for order flow imbalance.



2. Lakner, Reed, and Stoikov (2016): Lakner, Reed, and Stoikov (2016) study one-sided order books assuming Poisson arrival of limit orders with intensity conditional on the current price.
3. Lakner, Reed, and Stoikov (2016) Weak Price/LOB Process Limits: They provide weak limits on the price and the LOB processes and show that the limit order book process is a solution to a stochastic differential equation.
4. Huang and Rosenbaum (2017): Huang and Rosenbaum (2017) extend their previous Poisson arrival framework Huang, Lehalle, and Rosenbaum (2015) to create a more general stochastic dynamic model.
5. Huang and Rosenbaum (2017) Order Book/Price Jumps: They consider two separate jump processes for the order book state and a reference price which is not the *mid-price* but rather the so-called *fair-value* of the security as perceived by the users.
6. Huang and Rosenbaum (2017) State Transition Matrix: They create a state-transition matrix based on the queue-reactive flow assumption and further also incorporate a re-initialization probability of the reference price attributing it to exogenous jumps.
7. Huang and Rosenbaum (2017) Tests of Ergodicity: They perform tests of ergodicity and conclude by proving that under certain scaling limits, the price dynamics converges to that of a Brownian motion.
8. Cont de Gond, and Xuan (2023): More recently, Cont, de Gond, and Xuan (2023) form a more general LOB model with the order flow modeled as a point process and the trade execution modeled as a deterministic mass transport operator.
9. Cont de Gond, and Xuan (2023) Generalized Framework: Under certain scaling limits they show that their framework generalizes a number of LOB models and show that they are in fact special cases of their framework.

## Stochastic Differential Equations Based Models – Volumes of Orders Processes



1. Cont and de Larrard (2011): Cont and de Larrard (2011) use heavy traffic limits to show that depending on the scaling behavior of the order flow, LOBs can act as deterministic under the fluid limit and stochastic under the diffusive limit.
2. Cont and de Larrard (2011) Diffusive Limit: They find that the diffusive case occurs much more often in the empirical data. They are able to approximate the diffusion limit's differential equation by a 2D Brownian motion.
3. Cont and de Larrard (2011) Analytical Solution to Dynamics: They further derive analytical solution to quantities like price dynamics, duration between price moves, and the probability of a price move.
4. Chavez-Casillas and Figueroa-Lopez (2017): Building on the model in Cont and de Larrard (2011), Chavez-Casillas and Figueroa-Lopez allow for variable spread in the simulated dynamics as well as for in-spread orders.
5. Cont and Muller (2021): Cont and Muller (2021) develop a model for continuous limit of volume at a time  $t$  and price  $p$  by using a volume density  $v(t, p)$ .
6. Cont and Muller (2021) Centered Volume Density: They center the volume density to  $u_t(x) := v(t, S_t + x)$  where  $S_t$  is the mid-price.
7. Cont and Muller (2021) Treatment of Cancels: They follow a data-driven approach to model the order book dynamics – they categorize the cancelations into deletions and modifications.
8. Cont and Muller (2021) Treatment of Modifications: Further, modifications are bifurcated into symmetric, i.e., cancel and place at a near-by level, and anti-symmetric, i.e., cancel and replace at mid-price.
9. Cont and Muller (2021) Decomposition into Diffusion/Convection: They model symmetric modifications as a diffusion and anti-symmetric modifications as a convection.

## Stochastic Differential Equations Based Models – Probabilistic Properties under Scaling Limits



1. Horst and Paulsen (2017): Horst and Paulsen (2017), by assuming generalized time-dependent order arrival intensities, develop limit theorems for price and volume densities at bid and ask.
2. Horst and Paulsen (2017) ODE-PDE System: They conclude that, given some regularity conditions, the two processes converge to a coupled ODE-PDE system of equations until scaling limits.
3. Horst and Kreher (2017): Horst and Kreher (2017) further generalize the previous work to develop a weak LLN by considering the order flow as Markovian dynamics which are state dependent.
4. Horst and Kreher (2017) Order Book Quantities: Specifically, they conjecture that the type of the order, its size, and price are all a function of the mid-price and the standing volume of the order book.
5. Horst and Kreher (2018): In Horst and Kreher (2018) the authors further show that the model in Horst and Kreher (2017) can be used as a first order approximation of liquidity in the order book to construct optimal liquidation trajectories.
6. Horst and Kreher (2018) Optimal Trajectory Ranges: They develop second order approximations to obtain confidence intervals around these trajectories.
7. Horst and Kreher (2018) Dual Scaling Limits: To that extent, they formulate two scaling limits accounting for the fact that price change fluctuations are much slower than order arrival and cancelation fluctuations.
8. Horst, Kreher, and Starovoitovs (2023): Horst, Kreher, and Starovoitovs (2023) create a second order approximation with a single scaling instead of the previous two. They do so by assuming a non-varying Market Order to Limit Order ratio.
9. Horst, Kreher, and Starovoitovs (2023) Price-Volume Process: They show that the price-volume process in the limit converges to a solution of an infinite dimension PDE.
10. Ma and Noh (2023): Ma, Wang, and Zhang (2014) and Ma and Noh (2023) show that a one-sided order book can be modeled as an SDE and equilibrium characteristics can be calculated for them.



11. Ma and Noh (2023) Optimal Trading Trajectories: They further take the limit of trades in the market to infinity and show that the Mean-Field Game for this trader can be constructed using this SPDE, and under some conditions can be solved using the viscosity solutions of the Hamilton-Jacobi-Bellman HJB equations.
12. Rojas, Logachov, and Yambartsev (2020): On the other hand, Rojas, Logachov, and Yambartsev (2020) develop LLN, CLT, and large deviations for a stressed order book – in their case they look at the liquidity fluctuations.

## Stochastic Differential Equations Based Models – Connecting Various Timescales

1. Hambly, Kalsi, and Newbury (2020): Hambly, Kalsi, and Newbury (2020) build a set of models to connect the various timescales – microscopic, then mesoscopic, and finally macroscopic.
2. Hambly, Kalsi, and Newbury (2020) Microscopic View #1: In the microscopic model, they assume a Poisson arrival model for all order types with two intensities based on the frequency of trading; a common intensity for all types of orders is used for high frequency, and an intensity dependent on the queue-price, mid-price, and the number of price changes ahead of the current order is used at lower frequencies.
3. Hambly, Kalsi, and Newbury (2020) Microscopic View #2: Further, they add a diffusion dynamic for order diffusing to nearby price levels.
4. Hambly, Kalsi, and Newbury (2020) Microscopic View #3: They show that as the order arrival rate goes to infinity and the size of each order goes to zero, which is the continuous limit they formulate by looking at very small timescales, they can form a Markovian diffusion process which is described by a system of reflected SPDEs.
5. Hambly, Kalsi, and Newbury (2020) Mesoscopic View: They look at the price changes of these SPDEs to look at the mesoscopic models of these SPDEs.



6. Hambly, Kalsi, and Newbury (2020) Macroscopic View: Finally, they take the limit of tick sizes going to zero to create a macroscopic continuous price process SPDE from the above model.

## Stochastic Differential Equations Based Model – Discussion

1. Mathematically Tractable Formulation: SPDEs provide a mathematically tractable formulation of the time evolution of the order book.
2. Advantages of and Challenges with SPDEs: The feature of this category of models makes them attractive but at the same time also brings out certain issues.
3. Lack of Explicit Solutions: More often than not, SPDEs do not have an explicit solution. There are several approximations made in the literature to circumvent this problem like using viscosity solutions, considering the similarities to heat equation in physics, and even some inspiration from energy models in statistical physics.
4. Ease of State Evolution Formulation: Despite the difficulty in finding exact solution to the system of equations, these models can be used to build a simulator since it is readily possible to evolve a stochastic variable with a set of SPDE for its dynamics.
5. Core Assumptions of the SPDE Model: The practitioner should be careful of the core assumptions of the SPDE model they are using.
6. Model and Computational Complexities: Challenges in using these models include high model complexity and a need for a high amount of computing resources to perform these simulations.
7. Multiple Simulation Timescales: Further, a class of point process models can be scaled to large timescales to produce a set of SPDEs.
8. Analytically Tractable Dynamics: Another desirable property of this category of models is that since the dynamics are analytically tractable, much like Hawkes Process models, one can use optimal control theory to participate in trading.



## Responsiveness to Trades: Market Impact – Introduction

1. Definition of Market Impact: Market Impact or Price Impact is defined as the price movement due to one's own trading.
2. Depletion of Price Levels: Suppose a large market order is submitted by an agent on the buy side and it depletes a few levels of prices in the order-book, the new best ask price will be a few ticks higher than the previous best ask.
3. 'Walking the Order Book': This suggests that the agent's trade 'walked the order book' and moved the price in the opposite direction. If the same agent wishes to buy again, they will have to pay a higher price.
4. Information Leakage from Limit Order: Similarly, one can think of posting of limit orders as showing that market one's intention to trade at that price which gives *information* to the market which can react against the agent.
5. Reducing the Market Impact: Reducing Market Impact has been one of the pillar stones of all agency and electronic trading activities with years of research spent on building algorithms to reduce the impact of large orders.
6. Techniques to reduce Impact Market: Common techniques to reduce market impact include batching of orders following the market, i.e., targeting the Volume Weighted Average Price, and using alpha signals to place orders at a *smarter* price.
7. Estimated Empirical of Market Impact: Market Impact is one of the phenomenon in the markets which cannot be measured immediately – at least not in any meaningful sense – the agent will probably wait for some time for the market to *settle down* or in other words, wait for the price to come back to the previous levels.
8. Modeling the Market Impact: This makes Market Impact one of the harder aspects of the LOB to model. Not only does it depend on the supply and demand of the order book at that point of time but also the volatility of the security.
9. Literature on Market Impact: The impact of Market Impact has been highlighted in the literature since a long time (Biais, Hillion, and Spatt (1999), Foucault, Kadan, and Kendel (2005), Cont, Kukanov, and Stoikov (2014)).



10. Relation to the Price Process: Indeed, Market Impact, at the minutest scales, quite closely related to the process of price formation (Lillo (2023)).
11. Cont, Kukanov, and Stoikov (2014): Cont, Kukanov, and Stoikov (2014) show that the price impact can be explained by order imbalance and with a scaling they show that the ‘square-root law’ heuristic that traders in practice have can be derived quite easily.

## Responsive to Trades: Zero-intelligence Model Market Impact

1. Smith, Farmer, Guillard, and Krishnamurthy (2003): Smith, Farmer, Guillard, and Krishnamurthy (2003) posit that the instantaneous price impact function  $\phi(\omega, t)$  is nothing but the inverse of the cumulative depth profile  $N(p, t)$  of the order book where  $\omega$  is the order size.
2. Smith, Farmer, Guillard, and Krishnamurthy (2003) Analytical Price Impact Function: They show that by using the Taylor’s expansion  $\omega(\delta_p)$ , i.e., order size needed to move the mid-price by  $\delta_p$ , they get an analytical price impact function.
3. Smith, Farmer, Guillard, and Krishnamurthy (2003) Comparison to Empiricals: Their results show good matching with observed price impact when the Taylor series is expanded to two degrees.
4. Smith, Farmer, Guillard, and Krishnamurthy (2003) Order Arrival Intensities: Further, they assume that the order arrival intensities are dependent on the distance from the mid-price level.
5. Smith, Farmer, Guillard, and Krishnamurthy (2003) Noise from Rapid Orders: In addition to that, they add an additive noise from rapid submissions and deletions of high-frequency traders in the form of a Brownian motion dependent on the centered volume density.
6. Smith, Farmer, Guillard, and Krishnamurthy (2003) Convergence of SPDE: They show that if the mid-price is an Arithmetic Brownian Motion, the volume density converges to a SPDE with a moving boundary’s solution.



7. Smith, Farmer, Guillard, and Krishnamurthy (2003) 2-Factor SPDE: They generalize this methodology to formulate an SPDE of the centered-volume density and create a 2-factor model by creating SPDEs for both bid and ask side queues.
8. Smith, Farmer, Guillard, and Krishnamurthy (2003) Long-term Order-Book Shape: They show that the long-term order-book shape can be explicitly solved for and they show a first order approximation of the shape in their results.

## **Responsiveness to Trades: Poisson Process Market Impact**

1. Huang, Lehalle, and Rosenbaum (2015): Huang, Lehalle, and Rosenbaum (2015) study the market impact of VWAP liquidation and exponential scheduling liquidation in their simulated model. They see concavity in their market impact observations against time and volume both.
2. Match with Empirical Data: This shows that their model has intrinsic market impact and it matches some real-world behavior of the markets. The same characteristics are observed in the model by Lu and Abergel (2018a).
3. Market Impact of Poisson Impact: In general, except for certain queue-reactive and state-dependent variants, Poisson models do not have market impact as a feature since the core assumption in the Poisson model is that each order event count is independent in increments.

## **Responsiveness to Trades: Hawkes Process Market Impact**

1. Order Flow and Price Formation: As pointed out by Lillo (2023), given a reference price  $P_t$ , the time evolution of the price is a deterministic function of the order flow point process.



2. Temporary Market Impact: They show that the market impact can be modeled by the transient impact model, i.e., the trading velocity impacts the price in a decaying function of time, if the order flow is considered to be exogenous of the price process.
3. Price Movement vs. Future Order Flow: However, they argue that the empirical data shows correlation of price movement with the future order flow and hence the assumption that order flow is exogenous to price impact is probably not correct.
4. Cross-excitation in the Hawkes Process: Hawkes process – with price and order flow as its dimensions – relaxes this assumption by considering the cross-excitation of order flow from price movement and vice versa.
5. Implicit Hawkes Model Impact: Hawkes models are by definition impact to past events and hence order intensities are influenced by any past orders happening. This can be thought of as an implicit form of the market impact in these models.
6. Bacry, Iuga, Lasnier, and Lehalle (2015): Bacry, Iuga, Lasnier, and Lehalle (2015) study the so-called Hawkes impact model by considering a simple 2D Hawkes process of price.
7. Concave Impact followed by Convex Price Relaxation: They conclude that while liquidating a meta-order, i.e., an order made of multiple child-orders which are individual market/limit orders, they observe a concave market impact followed by a convex relaxation of the price after the agent has stopped trading. This behavior is very much in line with the expectation of traders.
8. Le and Seo (2017): Also noteworthy is the recent work by Lee and Seo (2017) where they study the market impact using the Hawkes process in a more realistic sense by including the tick-size discretization of price levels.
9. Lee and Seo (2017) Volatility: They further develop formulae for realized volatility and compare with the empirical volatility.

## **Responsive to Trades: Agent-based Models Market Impact**



1. Giamouridis, Papaioannou, and Rosenzweig (2023): A recent study on how different categories of agents have different kinds of price impact has been done by Giamouridis, Papaioannou, and Rosenzweig (2023).
2. Paddrik, Hayes, Todd, Yang, Beling, and Scherer (2012): Paddrik, Hayes, Todd, Yang, Beling, and Scherer (2012) provide an interesting case – they do not explicitly study the market impact of an exogenous order, but they perform similar study to replicate the circumstances around the Flash Crash in 2020 by placing a large exogenous trade that *moves* the market.
3. Paddrik, Hayes, Todd, Yang, Beling, and Scherer (2012) Impact from HFT/Market Makers: However, this impact is due to the behavior of the high-frequency and market makers modeled in the ABM and their reaction to the changing liquidity of the order book.
4. Byrd, Hybinette, and Balch (2019): Byrd, Hybinette, and Balch (2019) also show that agent-based modeling can be utilized to study and estimate market impact models.
5. Byrd, Hybinette, and Balch (2019) Market/Limit Orders: They perform a case study with one exogenous trader placing orders, with varying proportions of market and limit orders, and they observe the price evolution during and after the trading is done.
6. Byrd, Hybinette, and Balch (2019) Path-wise Market Impact: They perform a comparison of this price path with and without this exogenous trader and show that this price has meaningfully changed with trading activity.
7. Byrd, Hybinette, and Balch (2019) Concave Market Impact: The market impact they observe in their model is concave and is a decreasing function of the proportion of market orders in the trading strategy.
8. Shi and Cartlidge (2023): Shi and Cartlidge (2023) follow the same framework to test the market impact on their model.
9. Coletta, Moulin, Vyettrenko, and Balch (2022): Colette, Moulin, Vyettrenko, and Balch (2022) also follow ABIDES framework of studying Market Impact.
10. Coletta, Jerome, Savani, and Vyettrenko (2023): Further, in Coletta, Jerome, Savani, and Vyettrenko (2023), they breakdown the impact into market and the limit orders and compare it to historical replay method of order book simulations.



11. Belcak, Calliess, and Zohren (2020): Belcak, Calliess, and Zohren (2020) study their model's market impact by analyzing the average spread, the variance of the spread, and the variance of the best price as a function of time since a large market order happened in the past. They also report a concave shaped market impact curve.

## **Responsiveness to Trades: Stochastic PDEs Based Models Market Impact**

Horst and Kreher (2018) show that, using their second order approximation, two different forms of market impact are found under different scaling limits – they term them to be temporary and permanent forms of impact.

## **Responsiveness to Trades: Deep-Learning Based Models Market Impact**

1. Cont, Cucuringu, and Kochems (2023): In Cont, Cucuringu, and Kochems (2023), market impact study is done by analyzing the price paths observed while executing a varying quantity of orders using three strategies: Market order TWAP – time-weighted average price is the target price; Limit order TWAP; and POV – percentage of volume, i.e., the trader targets maintaining their traded volume to be a constant ratio of the market volume.
2. Cont, Cucuringu, and Kochems (2023) Trend Lines Observed: There are clear trend lines observed in all three and comparisons are made to Poisson and Hawkes in the former two where it is shown that there is no clear trend in Poisson or Hawkes.
3. Cont, Cucuringu, and Kochems (2023) Model Sensitivity to Market Impact: There has been a general rise in order book models being sensitive to the price impact; however, there is still room for improvement.



4. Cont, Cucuringu, and Kochems (2023) Market Impact in Order Book Simulations: It is important to stress the need for being aware of market impact in the order book simulation, especially if the aim of building the simulator is to perform back tests or algo trading strategies.
5. The ABIDES Framework: The ABIDES framework from Byrd, Hybinette, and Balch (2019) proves to be quite useful in the study of market impact since it is available in a open-source code repository.

## Comparative Study

1. Building Model Priors from Observations: Frequently, researchers investigating some of the stylized facts observed in empirical data use their reasonings for the observed distributions as priors in their modeling techniques.
2. Empiricals as Fit-Quality Determinants: It makes sense that, building from those priors, they use these stylized facts as goodness-of-fit metrics as well.
3. Tables of Models/Stylized Facts: The next section enumerates the stylized facts used in the models reviewed in this chapter. It also provides comments on how the authors use the stylized facts in their testing of the model against real data.
4. Popular Empirical Stylized Facts: The various majority of the researchers make use of empirical probability distribution functions for various properties of the order book, the most popular ones being the inter-order arrival times, spread, and volumes as their primary stylized facts.
5. Qualitative Tests: The technique for testing against these stylized facts is usually a qualitative test where the two distributions – empirical data and simulations – are plotted against each other.
6. Q-Q Plots for Shape Comparison: While this method is useful to test whether the general shape of the distribution matches between the two subsets, several researches also use Q-Q plots, which are far more sensitive to the tail events.



7. Region and Quantile Check: This technique not only matches the dense regions of the distribution but also the lower and the upper quantiles.
8. Order Arrival Times Q-Q Plot: Indeed, most of the Hawkes process methodology mentioned in the works below make use of the Q-Q plot of the inter-order arrival times to refute the Poisson model. Some notable observations are listed below.
9. Farmer, Patelli, and Zovko (2005): Farmer, Patelli, and Zovko (2005) test the model proposed by Smith, Farmer, Gillemot, and Krishnamurthy (2003) by using empirical data from the London Stock Exchange. Notably, their zero-impact model is able to reproduce the concave market impact function.
10. Cont, Stoikov, and Talreja (2010): Cont, Stoikov, and Talreja (2020) test their model's quality of fit by comparing the average LOB profile and the realized volatility against real-world data from the Tokyo Stock Exchange.
11. Cont, Stoikov, and Talreja (2010) Transition Probabilities: They further show that the conditional probabilities from their model matches the empirical frequencies observed of the direction of the price moves and one-step transition.
12. Abergel and Jedidi (2013): Abergel and Jedidi (2013) demonstrate a series of tests to compare simulation results to real-world data. These include comparisons of average depth profiles, probability distribution of spreads in ticks and price changes, autocorrelation of price changes, and Q-Q plot of mid-price changes.
13. Paddrik, Hayes, Todd, Yang, Beling, and Scherer (2012): Paddrik, Hayes, Todd, Yang, Beling, and Scherer (2012) show that their agent-based approach reveals volatility clustering phenomenon which is quite remarkable. It seems that a mixed-timescales approach to LOB modeling naturally leads to volatility clustering.
14. Paddrik, Hayes, Todd, Yang, Beling, and Scherer (2012) Flash Crash: Interestingly, they are also able to simulate crashes in the market when a large sell order is traded and HTFS and market makes withdraw.
15. Huang, Lehalle, and Rosenbaum (2015): Huang, Lehalle, and Rosenbaum (2015) model is tested against empirical data for two high spread-in stocks on the French exchange.



16. Huang, Lehalle, and Rosenbaum (2015) Zero Intelligence Model I: The zero-intelligence model I fits better to the asymptotic order distribution of the stocks compared to the constant arrival rate models.
17. Huang, Lehalle, and Rosenbaum (2015) Realized Volatility: Interestingly, in their third model with moving mid-price, without a re-initialization event, they find that the realized volatility in simulations is much lower than the empirical realized volatility. They conclude that their mode probably suffers from the mean-reverting behavior to mid-price which is not necessarily the case in practice.
18. Li, Wang, Lin, Sinha, and Wellman (2020): Li, Wang, Lin, Sinha, and Wellman (2020) perform tests on distributions of price, quantity, inter-arrival times, and spectral bid/ask prices.
19. Li, Wang, Lin, Sinha, and Wellman (2020) Statistical Comparisons: This is of note, since, instead of using the methodology adopted by researches before, they make use of statistical tests to provide a more robust learning methodology.
20. Kirchner (2017): Another notable mention is that of Kirchner (2017) method of model selection. He provides hyperparameters for models to tune in each such use case and shows the use of the AIC metric in model selection in choosing the hyperparameters.
21. Lee and Seo (2022): An impressive study on the calibrated results is performed by Lee and Seo (2022) where they perform stationarity checks on their parameters and also compared the calibrated parameters against a baseline model. They also perform model selection by comparing 5 different proposed models.
22. Muccianti and Sancetta (2023): Muccianti and Sancetta (2023) show case a testing methodology in which they measure the sped of convergence in their fitting method as well as show how sensitive their parameters are to various conditions to check the robustness of their model. They also perform tests on out-of-sample data.
23. Comparisons to Baseline Models: Comparisons to baseline models are also quite beneficial when the practitioner wishes to identify mathematical performance gains against simpler, more explainable methods.



24. Shi and Cartlidge (2021); Shi, Chen, and Cartlidge (2021): A good example are the works of Shi and Cartlidge (2021) and Shi, Chen, and Cartlidge (2021) where the authors compare their method against a number of low complexity baselines.
25. Comparison against Neural Nets: They also perform several studies to validate their use of ODE-RNNs instead of the more traditional LSTM/GRUs and further provide an ablation study on the parameters.
26. Effectiveness of the Poisson Model: As can be seen from the Table in the next section, Poisson models are generally successful in representing a number of *first-order* stylized facts like distribution of spread, volumes, average depth, and average order book profile.
27. Inadequacy of the Poisson Model: However, Abergel and Jedidi (2013) show that several other key stylized facts such as auto-correlation of price changes, signature plots, and long-term volatility are insufficiently represented in a Poisson model.
28. Improvement Offered by Hawkes: They show that a self-exciting process like Hawkes process could be one candidate solution.
29. Stylized Facts Modeled by Hawkes: Hawkes process models are seen to be much better than Poisson in representing the above-mentioned stylized facts. They also fit the tails of the inter-order arrival times distribution quite well.
30. Morariu-Patrichi and Pakkanen (2022): It has been shown in Morariu-Patrichi and Pakkanen (2022) that the residuals are fitting the Hawkes process should follow an exponential distribution; however, it can be seen in their work that this is generally not the case for Hawkes process.
31. Enhancement to Hawkes Process: Therefore, more complex models such as state-dependent Hawkes and Neural Hawkes have been proposed.
32. Improving Excitation Kernels: A frequent property that have been tested across all Hawkes models is the nature of the excitation kernel.
33. Exponential vs. Power Law Kernels: The most common choice, i.e., exponential kernels, have been shown to be insufficient (Nystrom and Zhang (2022)), hence power law kernels have become more popular (Bacry, Jaisson, and Muzy (2016)), although they are harder to calibrate.



34. Non-parametric Hawkes Estimations: A separate class of Hawkes models is the non-parametric model where the excitation kernels are estimated without any prior shape assigned to them. A number of authors report the shape of these kernels in time – or log-time.
35. DL Training and Validation Losses: In the Deep Learning category of models, it can be seen that training and validation losses are reported, which is generally the industry standard.
36. Qualitative Comparison of Simulated Paths: Also notable, particularly in the models using GANs, is the usage of simulated price paths and their qualitative comparisons to empirical data.
37. Stochastic Partial Differential Equations: The SPDE category of models generally use stylized facts to create priors in the differential equation dynamics; a general trend towards testing the first order and long-term asymptotic features can be seen.

## Comparison - Zero-intelligence Poisson Models

1. Bouchaud, Mezard, and Potters (2002) Stylized Facts: Order Flow Statistics and Average LOB Shape.
2. Bouchaud, Mezard, and Potters (2002) Analysis: Qualitative tests performed.
3. Luckock (2003) Stylized Facts:
  - a. PDF's of trade and best bid/ask prices
  - b. Density if unexecuted orders wrt price
4. Luckock (2003) Analysis: Qualitative tests performed.
5. Farmer, Patelli, and Zovko (2005) Stylized Facts:
  - a. Average spread
  - b. Price Diffusion Rate
6. Farmer, Patelli, and Zovko (2005) Analysis: All stylized facts compared between empirical and predicted (Smith, Farmer, Gillemot, and Krishnamurthy (2002))
7. Cont, Stoikov, and Talreja (2010) Stylized Facts:



- a. Average rates wrt distance from opposite quotes
  - b. Average LOB Shape
  - c. Probability of price increase wrt queue size
  - d. Probability of execution – one side or both – before mid-price movement
8. Cont, Stoikov, and Talreja (2010) Stylized Facts:
- a. Tests done to compare empirical vs simulated results for probability of price increase wrt queue size across various levels
  - b. Comparisons are made with empirical data as well as theoretical results by the Laplace transform method
9. Abergel and Jedidi (2013) Stylized Facts:
- a. Average depth profile
  - b. PDF of spread-in-ticks
  - c. Autocorrelation of price movements
  - d. Price paths
  - e. Signature Plots
  - f. Average depth
  - g. Spread
  - h. Long-term Volatility
10. Abergel and Jedidi (2013) Analysis:
- a. Qualitative tests performed on all stylized facts
  - b. Q-Q tests are performed on mid-price movements
11. Cont and de Larrard (2013) Stylized Facts:
- a. Joint distribution of best bid and ask size
  - b. Probability of price increase conditional on current bid/ask sizes
12. Cont and de Larrard (2013) Analysis: Diffusive coefficient of price from simulation and from empirical data is compared
13. Huang, Lehalle, and Rosenbaum (2015) Stylized Facts:
- a. Order intensities distribution by queue size
  - b. PDF of queue size for 3 levels of bid/ask
  - c. Joint distribution of first two-levels queue sizes



- d. Joint distribution of best bid-ask queue sizes
- 14. Huang, Lehalle, and Rosenbaum (2015) Analysis: Qualitative tests performed
- 15. Lu and Abergel (2018b) Stylized Facts:
  - a. Order intensities and order size distribution by queue size
  - b. Conditional distributions and other statistics of various types of orders and their arrival times
  - c. PDF of best bid/ask sizes
- 16. Lu and Abergel (2018b) Analysis: Quantitative tests performed using Monte Carlo simulations

## Comparison - Hawkes Models

- 1. Toke (2010) Stylized Facts: PDF of spread, inter-arrival times, variance of mid-price
- 2. Toke (2010) Analysis: Qualitative tests
- 3. Zheng, Roueff, and Abergel (2010) Stylized Facts: Signature plots (Bid1, Ask1, Mid)
- 4. Zheng, Roueff, and Abergel (2010) Analysis: Qualitative tests
- 5. da Fonseca and Zaatour (2014) Stylized Facts:
  - a. PDF and histograms of inter-arrival times
  - b. Autocorrelation of number of trades in a time window
  - c. Signature Plots
- 6. da Fonseca and Zaatour (2014) Analysis:
  - a. Q-Q plot comparison made of inter-arrival times to the exponential distribution
  - b. Comparisons of the fitted parameters with an MLE baseline with robustness testing of the parameters by investigating the standard deviations
  - c. Qualitative tests
- 7. Bacry, Jaisson, and Muzy (2016) Stylized Facts: Event cross and self-excitation versus time



8. Rambald, Bacry, and Lillo (2017) Stylized Facts: PDF of inter-arrival times, order volumes
9. Rambald, Bacry, and Lillo (2017) Analysis: Plots and calibration results for their estimated Hawkes Process
10. Lu and Abergel (2018a) Stylized Facts:
  - a. Conditional probabilities of events
  - b. Mid-price signature Plots
  - c. PDF of inter-arrival times
11. Lu and Abergel (2018a) Analysis:
  - a. Residuals' Q-Q plot tested to follow  $\exp(1)$  distribution
  - b. Qualitative tests
12. Wu, Rambaldi, and Bacry (2019) Stylized Facts: PDF of queue size
13. Wu, Rambaldi, and Bacry (2019) Analysis:
  - a. Q-Q plot of inter-arrival times are compared with empirical data as well as a Queue-reactive Poisson model baseline
  - b. Qualitative tests
14. Kirchner and Vetter (2022) Stylized Facts:
  - a. Average order intensity by time of decay
  - b. Unconditional transitional probabilities between all orders
  - c. Market orders against imbalance
15. Kirchner and Vetter (2022) Analysis: Properties of calibrated excitation kernels are plotted against time of day
16. Morariu-Patrichi and Pakkanen (2022) Stylized Facts: State transition probabilities
17. Morariu-Patrichi and Pakkanen (2022) Analysis:
  - a. Residuals' Q-Q plot tested to follow  $\exp(1)$  distribution
  - b. Comparison between generic Hawkes and state-dependent Hawkes
18. Lee and Seo (2022) Stylized Facts: Bid/ask price plots
19. Lee and Seo (2022) Analysis:
  - a. Checks for stationary of estimated properties
  - b. Comparison between normal Hawkes and proposed state-conditioned Hawkes



- c. Tests are also performed in further model selection between the 5 proposed models
- d. Residuals' Q-Q plot tested to follow  $\exp(1)$  distribution

20. Nystrom and Zhang (2022) Stylized Facts:

- a. Price changes in tick frequency
- b. Price paths
- c. PDFs of number of jumps

21. Nystrom and Zhang (2022) Analysis:

- a. Price paths compared between exponential and power law kernels
- b. Statistical significance tests as well as qualitative tests on estimated parameters
- c. Computing time for both types of kernel's fitting process
- d. Distributions of the number of jumps are compared qualitatively between empirical data and simulated paths

22. Muccianti and Sancetta (2022) Analysis:

- a. Convergence speed and sensitivity to parameters
- b. Out-of-sample testing using statistical significance tests

## Comparison – Agent-based Models

1. Paddrik, Hayes, Todd, Yang, Beling, and Scherer (2012) Stylized Facts:

- a. Volume Percentage by Agent Type
- b. Cancelation rates by Agent Type
- c. PDF of returns
- d. Auto-correlation of absolute returns – indicative of Volatility Clustering
- e. Auto-correlation of returns
- f. PDF of returns at various frequencies

2. Paddrik, Hayes, Todd, Yang, Beling, and Scherer (2012) Analysis: Q-Q Plots of returns vs. Gaussian distribution



3. Byrd, Hybinette, and Balch (2019) Stylized Facts:
  - a. Price Paths
  - b. Trade Paths
4. Byrd, Hybinette, and Balch (2019) Analysis:
  - a. Price Paths are compared qualitatively
  - b. OOS Errors and Accuracy
5. Belcak, Calliess, and Zohren (2020) Analysis: Comparisons against ABIDES

## Comparison – Deep-Learning Models

1. Li, Wang, Lin, Sinha, and Wellman (2020) Stylized Tests: PDFs of Mid-price
2. Li, Wang, Lin, Sinha, and Wellman (2020) Analysis:
  - a. Qualitative tests performed like KS test, Jarque-Bera test, student t-test
  - b. t-statistic and p-values calculated for volatility measures – realized volatility, realized volatility per trade, and intraday volatility
3. Shi and Cartlidge (2020); Shi, Chen, and Cartlidge (2021) Analysis:
  - a. Several baseline models like Ridge Regression, SVR, Random Forests, 1-layer feed-forward Neural Networks, etc. to test performance of the model against ML techniques
  - b. Further tests are conducted on the effectiveness of using an ODE-RNN by comparing the performance against LSTMs and GRUs
  - c. Ablation study is performed
4. Lim and Gorse (2021) Stylized Tests: PDFs of Mid-price returns
5. Lim and Gorse (2021) Analysis:
  - a. Qualitative tests performed like KS test, Jarque-Bera test, student t-test
  - b. t-statistic and p-values calculated for volatility measures – realized volatility, realized volatility per trade, and intraday volatility
6. Coletta, Moulin, Vyetrenko, and Balch (2022) Stylized Facts:
  - a. PDFs



- i. Log Returns
  - ii. Order Type
  - iii. Time-to-fill
  - iv. Top-of-the-book Volumes
  - v. Spread by time of the day
- b. Price Paths
  - c. Autocorrelation of:
    - i. Returns
    - ii. Square Returns
7. Coletta, Moulin, Vyetrenko, and Balch (2022) Analysis: Simulation's conditional and unconditional PDFs of the stylized facts is compared with empirical PDFs and two baseline models
8. Prenzel, Cont, Cucuringu, and Kochems (2022) Stylized Facts:
- a. Conditional and Unconditioned PDFs by time-of-the day and market volatility of order intensities of each order type
  - b. Price Paths
9. Prenzel, Cont, Cucuringu, and Kochems (2022) Analysis:
- a. Comparisons between GAN model that the authors propose, Poisson process, and Hawkes process
  - b. Qualitative Tests
10. Nagy, Frey, Sapora, Li, Calinescu, Zohren, and Foerster (2023) Stylized Tests: PDFs of returns, order types, arrival times
11. Nagy, Frey, Sapora, Li, Calinescu, Zohren, and Foerster (2023) Analysis:
- a. Perplexity scores are used to test the LLM
  - b. Simulation's conditional and unconditional PDFs of the stylized facts is compared with compared PDFs

## Comparison – Combined Hawkes and Deep-Learning Models



1. Kumar (2021) Stylized Facts: PDF and Autocorrelation of returns
2. Kumar (2021) Analysis: Qualitative tests
3. Shi and Cartlidge (2022) Stylized Facts:
  - a. Price Paths
  - b. Volatility Clustering
  - c. Empirical PDFs of inter-arrival times
  - d. Volume, volatility, log returns, volatility correlations
4. Shi and Cartlidge (2022) Analysis: Comparison against several baseline models

## Comparison – Combined Hawkes and Deep-Learning Models

1. Shi and Cartlidge (2023) Stylized Facts:
  - a. Hurst exponents of absolute returns
  - b. Auto-correlation
  - c. Order Flow Imbalance Impact
  - d. Price Impact Function
  - e. Spread wrt Time
2. Shi and Cartlidge (2023) Analysis: A sensitivity analysis of the parameters with respect to the various stylized facts mentioned is performed to check the robustness of the parameters.

## Comparison – Stochastic Partial Differential Equation Models

1. Cont and de Larrard (2011) Stylized Facts:
  - a. Q-Q Plot of inter-arrival times compared with exponential distribution
  - b. Q-Q Plot number of shares per event to showcase clustering
  - c. Spreads' Timelines – 1 tick vs. >1 ticks
  - d. Joint PDF of bid/ask volumes



- e. Autocorrelation of absolute order sizes
  - f. Autocorrelation of inverse of inter-arrival times
  - g. PDF of inter-arrival times
2. Cont and de Larrard (2011) Analysis: Stylized facts are used to create priors on the model
  3. Korolev, Chertok, Korchagin, and Zeifman (2015) Stylized Facts:
    - a. PDE of Order Arrival Rate
    - b. PDE of ratio of intensities at bid/ask
  4. Korolev, Chertok, Korchagin, and Zeifman (2015) Analysis: Qualitative tests
  5. Chavez-Casillas and Figueroa-Lopez (2017) Stylized Facts:
    - a. Asymptotics of mid-price process
    - b. PDE of price
    - c. PDE of time for price change
    - d. Probability of price change conditional on current state
  6. Gao and Deng (2018) Stylized Facts: Average shape of the order book
  7. Gao and Deng (2018) Analysis: Shapes of order book at various timestamps are compared qualitatively
  8. Rojas, Logachov, and Yambartsev (2020) Stylized Facts:
    - a. Joint distribution of best bid and ask size
    - b. PDE of spread
    - c. PDE of lifetime for spread
    - d. Price Paths
    - e. Autocorrelation
  9. Rojas, Logachov, and Yambartsev (2020) Analysis: Price paths are compared qualitatively
  10. Hambly, Kalsi, and Newbury (2020) Stylized Facts:
    - a. Price Paths
    - b. Average LOB shape
  11. Hambly, Kalsi, and Newbury (2020) Analysis:



- a. Qualitative tests performed and volatility between empirical and simulated results is compared
  - b. Decomposition of volatility into exogenous movements and local imbalance
12. Cont and Muller (2021) Stylized Facts: Average LOB shape
13. Cont and Muller (2021) Analysis: Intraday price volatility and compared with empirical observations qualitatively

## References

- Abergel, F., and A. Jedidi (2013): *A Mathematical Approach to Order Book Modeling* arXiv
- Abergel, F., and A. Jedidi (2015): Long-time Behavior of a Hawkes Process-based Limit Order Book *SIAM Journal on Financial Mathematics* **6** (1) 1026-1043
- Abergel, F., M. Anane, A. Chakraborti, A. Jedidi, and I. M. Toke (2016): *Limit Order Books 1<sup>st</sup> Edition Cambridge University Press* Cambridge UK
- Arjovsky, M., S. Chintala, and L. Bottou (2017): [Wasserstein Generative Adversarial Networks](#)
- Bacry, E., A. Iuga, M. Lasnier, and C. A. Lehalle (2015): Market Impacts and Life-cycle of Investors' Orders *Market Microstructure and Liquidity* **1** (2) 1550009
- Bacry, E., I. Mastromatteo, and J. F. Muzy (2015): *Hawkes Processes in Finance* arXiv
- Bacry, E., T. Jaisson, and J. F. Muzy (2016): Estimation of Slowly Decreasing Hawkes Kernels: Applications to High-Frequency Order Book Dynamics *Quantitative Finance* **16** (8) 1179-1201
- Bacry, E., M. Bompaire, S. Gaiffas, and S. Poulsen (2018): *Tick: A Python Library for Statistical Learning, with a particular Emphasis on Time-dependent Modeling* arXiv



- Belcak, P., J. P. Calliess, and S. Zohren (2020): *Fast Agent-based Simulation Framework of Limit Order books with applications to Pro-rate Markets and the Study of Latency Effects* [arXiV](#)
- Biais, B., P. Hillion, and C. Spatt (1999): Price Discovery and Learning during the Pre-opening Period in the Paris Bourse *Journal of Political Economy* **107** (6) 1218-1248
- Bouchaud, J. P., M. Mezard, and M. Potters (2002): Statistical Properties of Stock Order Books: Empirical Results and Models *Quantitative Finance* **2** (4) 251-256
- Briola, A., J. Turiel, and T. Aste (2020): *Deep Learning Modeling of the Limit Order Book: A Comparative Perspective* [arXiV](#)
- Byrd, D., M. Hybinette, and T. H. Balch (2019): *ABIDES: Towards High-fidelity Market Simulation for AI Research* [arXiV](#)
- Chakraborti, A., I. M. Toke, M. Patriarca, and F. Abergel (2011): Econophysics Review II: Agent-based Models *Quantitative Finance* **11** (7) 1013-1041
- Chavez-Casillas, J. A., and J. E. Figueroa-Lopez (2017): A One-level Limit Order Book with Memory and Variable Speed *Stochastic Processes and Applications* **127** (8) 24472481
- Coletta, A., A. Moulin, S. Vyettrenko, and T. Balch (2022): Learning to simulate Realistic Limit Order Book Markets from Data as a World Agent *Proceedings of the 3<sup>rd</sup> ACM International Conference on AI in Finance* 428-436
- Coletta A., J. Jerome, R. Savani, and S. Vyettrenko (2023): *Conditional Generators for Limit Order Book Environments: Explainability, Challenges, and Robustness* [arXiV](#)
- Cont, R., S. Stoikov, and R. Talreja (2010): A Stochastic Model for Order Book Dynamics *Operations Research* **58** (3) 549-563
- Cont, R. (2011): Statistical Modeling of High-frequency Data *IEEE Signal Processing Magazine* **28** (5) 16-25
- Cont, R., and A. de Larrard (2011): [Order Book Dynamics in Liquid Markets: Limit Theorems and Diffusion Approximations](#)



- Cont, R., Kukanov A., and S. Stoikov (2014): The Price Impact of Order Book Events *Journal of Financial Econometrics* **12** (1) 47-88
- Cont R., and M. S. Muller (2021): A Stochastic Partial Differential Equation Model for Limit Order Book Dynamics *SIAM Journal of Financial Mathematics* **12** (2) 744-787
- Cont, R., M. Cucuringu, V. Glukhov, and F. Prenzel (2023): Analysis and Modeling of Client Order Flow in Limit Order Markets *Quantitative Finance* **23** (2) 187-205
- Cont, R., M. Cucuringu, and J. Kochems (2023): *Limit Order Book Simulation with Generative Adversarial Networks* **arXiV**
- Cont, R., P. de Gond, and L. Xuan (2023): *A Mathematical Framework for Modeling Order Book Dynamics* **arXiV**
- Copponi, A., and C. A. Lehalle (2023): *Machine Learning and Data Sciences for Financial Markets: A Guide to Contemporary Practices* **Cambridge University Press** Cambridge UK
- da Fonseca, J., and R. Zaatour (2014): Hawkes Process: Fast Calibration, Application to Trade Clustering, and Diffusive Limit *Journal of Futures Markets* **34** (6) 548-579
- Dragulescu, A. A., and V. M. Yakovenko (2002): Probability Distribution of Returns in the Heston Model with Stochastic Volatility *Quantitative Finance* **2** (6) 443-453
- Farmer, J. D., P. Patelli, and I. I. Zovko (2005): The Predictive Power of Zero-intelligence in Financial Markets *Proceedings of the National Academy of Sciences* **102** (6) 2254-2259
- Foucault, T., O. Kadan, and E. Kandel (2005): Limit Order Book as a Market for Liquidity *Review of Financial Studies* **18** (4) 1171-1217
- Gao, X., and S. J. Deng (2018): Hydro-dynamic Limit of Order Book Dynamics *Probability in the Engineering and Informational Sciences* **32** (1) 96-125
- Giamouridis, D., G. V. Papaionnaou, and B. Rosenzweig (2023): Deciphering how Investors' Daily Flows are Forming Prices, in: *Part II – How Learned Flows from Prices* (editors: *Copponi, A., and C. A. Lehalle*)
- Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Ward-Farley, S. Ozair, A. Courville, and Y. Bengio (2014): [Generative Adversarial Nets](#)



- Goodfellow, I., Y. Bengio, and A. Courville (2016): *Deep Learning* MIT Press  
Cambridge MA
- Gould, M. D., M. A. Porter, S. Williams, M. McDonald, D. J. Fenn, and S. D. Howison (2013): Limit Order Books *Quantitative Finance* **13 (11)** 1709-1742
- Hambly, B., J. Kalsi, and J. Newbury (2020): Limit Order Books, Diffusion Approximations, and Reflected SPDEs: From Microscopic to Macroscopic Models *Applied Mathematical Finance* **27 (1-2)** 132-170
- Hawkes, A. (2018): Hawkes Processes and their Applications to Finance: A Review *Quantitative Finance* **18 (2)** 193-198
- Horst, U., and M. Paulsen (2017): A Law of Large Numbers for Limit Order Books *Mathematics of Operations Research* **42 (4)** 1280-1312
- Horst, U., and D. Kreher (2017): A Weak Law of Large Numbers for a Limit Order Book Model with Fully State Dependent Order Dynamics *SIAM Journal of Financial Mathematics* **8 (1)** 314-343
- Horst, U., and D. Kreher (2018): Second Order Approximations for Limit Order Books *Finance and Stochastics* **22** 827-877
- Horst, U., D. Kreher, and K. Starovoitovs (2023): *Second-order Approximation of Limit Order Books in a Single-Scale Regime* arXiv
- Horst, U., and W. Xu (2019): A Scaling Limit for Limit Order Books Driven by Hawkes Processes *SIAM Journal on Financial Mathematics* **10 (2)** 350-393
- Huang, W., C. A. Lehalle, and M. Rosenbaum (2015): Simulating and Analyzing the Order Book: The Queue-Reactive Model *Journal of the American Statistical Association* **110 (509)** 107-122
- Huang, W., and M. Rosenbaum (2017): Ergodicity and Diffusivity of the Markovian Order Book Models: A General Framework *SIAM Journal of Financial Mathematics* **8 (1)** 874-900
- Hult, H., and H. Kiessling (2010): [Algorithmic Trading with Markov Chains](#)
- Jain, K., N. Firoozye, J. Kochems, and P. Treleaven (2024): *Limit Order Book Simulations: A Review* arXiv
- Kaj, I., and M. Caglar (2017): *A Buffer Hawkes Process for Limit Order Books* arXiv



- Kelly, F. and E. Yudovina (2018): A Markov Model of a Limit Order Book: Thresholds, Recurrence *Mathematics of Operations Research* **43** (1) 181-203
- Kirchner, M. (2017): AN Estimation Procedure for the Hawkes Process *Quantitative Finance* **17** (4) 571-595
- Kirchner, M., and S. Vetter (2022): Hawkes Model Specification for Limit Order Books *Quantitative Finance* **28** (7) 642-662
- Korolev, V. Y., A. V. Chertok, A. Y. Korchagin, and A. I. Zeifman (2015): Modeling High-frequency Order Flow Imbalance by Functional Limit Theorems by Two-sided Risk Processes *Applied Mathematics and Computation* **253** 224-241
- Kumar, P. (2021): *Deep Hawkes Process for High-frequency Market Making* arXIV
- Lakner, P., J. Reed, and S. Stoikov (2016): High Frequency Asymptotics for the Limit Order Book *Market Microstructure and Liquidity* **2** (1) 1650004
- Large, J. (2007): Measuring the Resiliency of an Electronic Limit Order Book *Journal of Financial Markets* **10** (1) 1-25
- Lee, K., and B. K. Seo (2017): Marked Hawkes Process Modeling of Price Dynamics and Volatility Estimation *Journal of Empirical Finance* **40** 174-200
- Lehalle, C. A., O. Gueant, and J. Razafinimanana (2011): High-frequency Simulation of an Order Book: A Two-scale Approach *Econophysics of Order-driven Markets: Proceedings of Econophysics Kolkata V* 73-92
- Li, J., X. Wang, Y. Lin, A. Sinha, and M. P. Wellman (2020): Generating Realistic Stock Market Order Streams *Proceedings of the AAAI Conference on Artificial Intelligence* **34** (1) 727-734
- Lillo, F. (2023): Order Flow and Price Formation, in: *Part II – How Learned Flows from Prices* (editors: Capponi, A., and C. A. Lehalle)
- Lim, Y. S., D. Gorse (2021): Intra-day Price Simulation with Generative Adversarial Modeling of the Order Flow *2021 20<sup>th</sup> IEEE International Conference on Machine Learning and Applications (ICMLA)* 397-402
- Lu, X., and F. Abergel (2018a): High-dimensional Hawkes Processes for Limit Order Books: Modeling, Empirical Analysis, and Numerical Calibration *Quantitative Finance* **18** (2) 249-264



- Lu, X., and F. Abergel (2018b): *Order Book Modeling and Market Making Strategies* [arXiV](#)
- Luckock, H. (2003): A Steady-state Model of the Continuous Double Auction *Quantitative Finance* **3** (5) 385-404
- Ma, J., X. Wang, and J. Zhang (2014): *Dynamic Equilibrium Limit Order Book Model and Optimal Execution Problem* [arXiV](#)
- Ma, J., and E. Noh (2023): [Equilibrium Model of Limit Order Books – A Mean-field Game View](#)
- Mei, H., and J. Eisner (2017): *The Neural Hawkes Process: A Neurally Self-modulating Multivariate Point Process* [arXiV](#)
- Mirza, M., and S. Osindero (2014): *Conditional Generative Adversarial Nets* [arXiV](#)
- Morariu-Patrichi, M., and M. Pakkanen (2022): State-dependent Hawkes Processes and their Applications to Limit Order Book Modeling *Quantitative Finance* **22** (3) 563-583
- Mounjid, O., M. Rosenbaum, and P. Sabila (2019): *From Asymptotic Properties of General Point Processes to the Ranking of the Financial Agents* [arXiV](#)
- Muccianti, L., and A. Sancetta (2023): *Estimation of an Order Book Dependent Hawkes Process for Large Datasets* [arXiV](#)
- Nagy, P., S. Frey, S. Sapora, K. Li, A. Calinescu, S. Zohren, and J. Foerster (2023): *Generative AI for End-to-end Limit Order Book Modeling* [arXiV](#)
- Ni, H., L. Szpruch, M. Sabate-Vidales, B. Xiao, M. Wiese, and S. Liao (2022): SIG-Wasserstein GANs for Time Series Generation *ICAIC '21: Proceedings of the 2<sup>nd</sup> ACM International Conference on AI in Finance* 1-8
- Nystrom, K., and C. Zhang (2022): Hawkes-based Models for High-frequency Financial Data *Journal of the Operational Research Society* **73** (10) 2168-2185
- Paddrik, M., R. Hayes, A. Todd, S. Yang, P. Beling, and W. Scherer (2012): An Agent-based Model of the E-Mini S&P 500 applied to Flash Crash Analysis *2012 IEEE Conference on Computational Intelligence for Financial Engineering & Economics* 1-8



- Prenzel, F., R. Cont, M. Cucuringu, and J. Kochems (2022): Dynamic Calibration of Order Flow Models with generative Adversarial Networks *ICAIF '21: Proceedings of the 3<sup>rd</sup> ACM International Conference in AI in Finance* 446-453
- Rambaldin, M., E. Bacry, and F. Lillo (2017): The Role of Volume in Order-book Dynamics: A Multivariate Hawkes Process Analysis *Quantitative Finance* **17** (7) 999-1020
- Rojas, H., A. Logachov, A. Yambartsev (2020): *Order Book Dynamics with Liquidity Fluctuations: Limit Theorems and Large Deviations* arXiv
- Shi, Z., Y. Chen, and J. Cartlidge (2021): The LOB Creation Model: Predicting the Limit Order Book from TAQ History using an Ordinary Different Equation Recurrent Neural Network *Proceedings of the AAAI Conference on Artificial Intelligence* **35** (1) 548-556
- Shi, Z., and J. Cartlidge (2021): The Limit Order Book Recreation Model (LOBRM): An Extended Analysis *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13-17 2021, Proceedings Part IV* 21 204-220
- Shi, Z., and J. Cartlidge (2022): State-dependent Parallel Neural Hawkes Process for Limit Order Book Event Stream Prediction and Simulation *Proceedings of the 28<sup>th</sup> ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 1607-1615
- Shi, Z., and J. Cartlidge (2023): *Neural Stochastic Agent-based Limit Order Book Simulation: A Hybrid Methodology* arXiv
- Sirignano, J., and R. Cont (2018): *Universal Features of Price Formation in Financial Markets: Perspectives from Deep Learning* arXiv
- Smith, E., J. D. Farmer, L. Gillemot, and S. Krishnamurthy (2003): Statistical Theory of Continuous Double Action *Quantitative Finance* **3** (6) 481-514
- Sullivan, R., A. Timmermann, and H. White (1999): Data-snooping, Technical Trading Rule Performance, and the Bootstrap *Journal of Finance* **54** (5) 1647-1691
- Takahashi, S., Y. Chen, K. Tanaka-Ishii (2019): Modeling Financia Time-series with Generating Adversarial Networks *Physica A: Statistical Mechanics and its Applications* **527** 121261



- Toke, I. M. (2010): *Market Making Behavior in an Order Book Model and its Impact on the Bid-Ask Spread* [arXiV](#)
- Vyetrenko, S., D. Byrd, N. Petosa, M. Mahfouz, D. Dervovic, M. Veloso, and T. Balch (2020): [Get Real: Realism Metrics for Robust Limit Order Book Market Simulations](#)
- White, H. (2000): A Reality Check for Data Snooping *Econometrica* **68** (5) 1097-1126
- Wiese, M., R. Knobloch, R. Korn, and P. Kretschmer (2020): Quant GANs: Deep Generation of Financial Time-series *Quantitative Finance* **20** (9) 1419-1440
- Wu, P., M. Rambaldi, J. F. Muzy, and E. Bacry (2019): *Queue-reactive Hawkes Models for the Order Flow* [arXiV](#)
- Zhang, Z., S. Zohren, and S. Roberts (2019): DeepLOB: Deep Convolutional Neural Networks for Limit Order Books *IEEE Transactions on Signal Processing* **67** (11) 3001-3012
- Zhang, Z., B. Lim, and S. Zohren (2021): Deep Learning for Market by Order Data *Applied Mathematical Finance* **28** (1) 79-95
- Zheng, B., F. Roueff, and F. Abergel (2014): [Ergodicity and Scaling Limit of a Constrained Multivariate Hawkes Process](#)



## Inverted Price Venues

1. Venues Offering Inverted Price Structure: Venues that offer an inverted price structure are grabbing market share away from those that offer traditional pricing (Editorial Staff (2016)).
2. Traditional Maker-Taker Rebate Scheme: In the traditional maker-taker pricing scheme, the liquidity provider is paid a rebate to post a bid or offer.
3. The Taker Maker Rebate Scheme: The newer scheme is the taker-maker, or inverted rebate system where the liquidity taker is paid for facilitating a trade.
4. Motivation behind Penny-Stock Orders: As stock prices go lower, the one-penny minimum tick size grows in percentage terms. Spreads widen out in percentage terms for lower-priced stocks, and traders use inverted price venues to increase their odds of capturing the spread.
5. Consequences of Larger Order Queues: Further, order queues get longer for certain stocks, bolstering the incentive for brokers to use inverted venues to increase their odds of capturing the spread.
6. More Behavioral than Structural Drivers: Brokers may be changing their attitude toward inverted venues and voluntarily use them more, as opposed to changes to the market structure or market environment that would cause them to route orders to inverted venues.

## References

- Editorial Staff (2016): [Inverted Price Venues Grabbing Market Share in US and Canada Traders' Magazine](#)



## Auction On-Demand

### Efficient Trading with On-demand Auctions

1. Volume Discovery from On-demand Auctions: Several venues (NASDAQ (2023)) provide volume discovery service based on lit auctions triggered on demand.
2. Alternative to OTC/Dark Pools: This is an innovative alternative to OTC and dark pool trading.
3. Execution for Large/Sensitive Orders: Developed with professional investors in demand, Auction On-demand (AOD) addresses a broad range of execution challenges for larger and more sensitive orders helping the industry to navigate the current regulatory landscape.

### Highlights

1. Separation from Lit Order Book: No impact on continuous Lit trading.
2. On-demand Triggering of Effective Auctions: Triggered on-demand by crossing orders.
3. At Primary Best Bid Offer: Enables trading at or within the PBBO.
4. Executes Orders with Low Market Impact.
5. Open yet Discrete Trading: Individual orders are not published.
6. Multiple Safety Features: Features for minimizing information leakage risk.

### Features



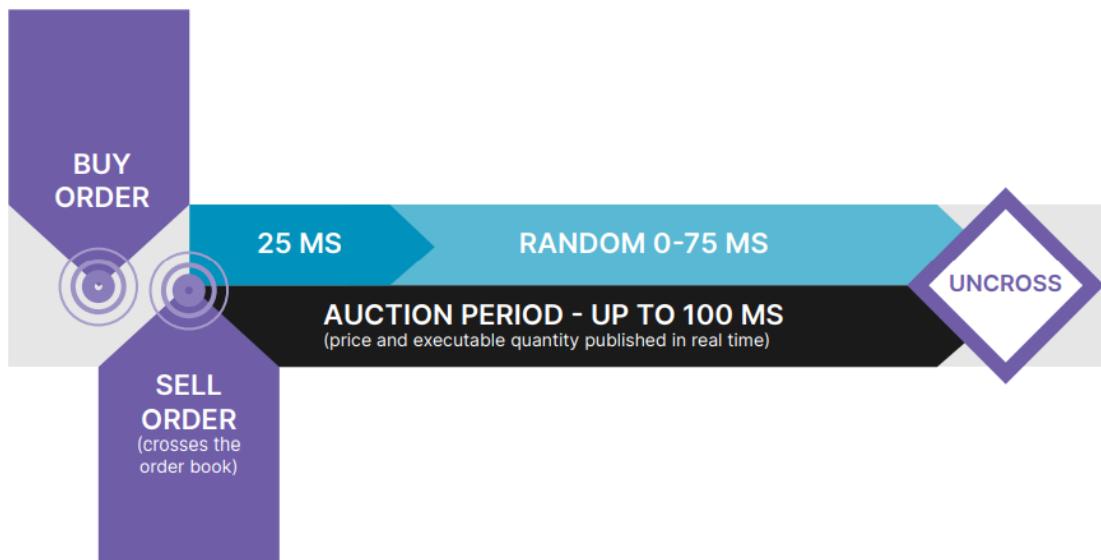
1. Price Discovery Mechanism: Simultaneously maximizes the traded Volume
2. Effective Execution for Larger Orders: This uses unique internal-to-broker/time/size priorities.
3. Accommodation of Pegged Orders: Orders pegged to the Bid, the Mid, and the Offer as well as Limit and Market orders.
4. Pre-trade Transparency: Indicative price and quantity in real-time
5. Post-trade Transparency: Trades in real-time, no counterparty, separate MIC code on execution reports.
6. Executions will contribute to official Member Market Share.

## Safety Features

1. Randomized Uncross after 25 – 100 milliseconds.
2. Speed Bumps on Cancels: This secures the real trading interests in the book.
3. Speed Bumps on Order Modifications: This is used when reducing positions.
4. Minimum Execution Size (MES) Protection.
5. Limit Guard Price Protection.
6. EBBO Collar Protection.

## Auction Overview

1. Duration of Auction Crossing: Auctions are triggered on demand by crossing orders. The duration of the auction is (typically) 25 milliseconds fixed and a random 0-75 milliseconds (again typical), meaning that the auction will last for a maximum of 100 milliseconds.
2. Price of Auction Uncrossing: The auction will uncross at the price where the most volume can be traded.



## Order Types

1. Pegged Orders: Pegged orders are only executed at their pegged price according to the peg instruction, i.e., *at-priced*.
2. Limit and Market Orders: These orders execute at or within the Primary Best Bid Offer PBBO.
3. Optional Limit Price Protection.
4. Tim In Force: DAY, GTT, GFA, and IOC.
5. MES Protection: Opposite order needs to be equal or larger than MES, hence no aggregation or *bulking*.

## References

- NASDAQ (2023): [Auction On-Demand](#)



## Volume-weighted Average Price

### Overview

1. Volume-weighted Average Price (VWAP): Volume-weighted average price (VWAP) is the ratio of a security or financial asset to the volume of transactions during a trading session (Wikipedia (2023)). It is a measure of the average trading price for period (Berkowitz, Logue, and Noser (1988)).
2. Time Period for VWAP Estimation: Typically, the indicator is computed for one day, but it can be measured between any two points in time.
3. VWAP as a Trading Benchmark: VWAP is often used as a trading benchmark by investors who aim to be as passive as possible in their execution. Many pension funds, and some mutual funds, fall into these categories.
4. Aim of VWAP Trading Target: The aim of using a VWAP trading target is to ensure that the trader executing the order does so in line with the volume on the market.
5. Impact on the Transaction Cost: It is sometimes argued that such execution reduces transaction costs by minimizing market impact costs, i.e., the additional cost of the market impact due to the adverse effect of a trader's activities on the price of a security.
6. Best-effort/Guaranteed Algorithmic Execution: VWAP is often used in algorithmic trading.
7. Definition of Guaranteed VWAP Execution: The broker may guarantee the execution of an order at the VWAP price and have a computer program enter the orders into the market to earn the trader's commission and create P&L. This is called a guaranteed VWAP execution.
8. Definition of Best-effort VWAP Execution: The broker can also trade in a best-effort way and answer the client with the realized price. This is called a VWAP target



execution; it incurs more dispersion in the answered price compared to the VWAP price for the client but a lower received/paid commission.

9. Volume Participation Algorithm: Trading algorithms that use VWAP as a target belong to a class of algorithms known as *volume participation algorithms*.

## Formula

VWAP is calculated using the following formula:

$$P_{VWAP} = \frac{\sum_j P_j \cdot Q_j}{\sum_j Q_j}$$

where  $P_{VWAP}$  is Volume Average Price,  $P_j$  is price of trade  $j$ ,  $Q_j$  is quantity of trade  $j$ , and  $j$  is each individual trade that takes place over the defined time period, excluding cross trades and basket cross trades.

## Using VWAP

1. Price Probing as Bullish/Bearish Indicator: The VWAP can be used similar to moving averages, where price above the VWAP reflect a bullish sentiment and prices below the VWAP reflect a bearish sentiment.
2. Position Initiation Signals: Traders may initiate short positions as a stock price moves below VWAP for a given time period or initiate long positions as the price move above VWAP.
3. Other Uses of VWAP: Institutional buyers and algorithms often use VWAP to plan entries and initiate larger positions without disturbing the stock price.



4. VWAP as Broker Performance Metric: VWAP slippage is the performance of a broker, and many Buy-side firms now use a MIFID wheel to direct their flow to the best broker.

## References

- Berkowitz, S. A., D. E. Logue, and E. A. J. Noser (1988): The Total Cost of Transactions on the NYSE *Journal of Finance* **43** (1) 97-112
- Wikipedia (2023): [Volume-weighted Average Price](#)



## Execution Cost and Transaction Trajectories

### Motivation and Practice Overview

1. Definition of Trade Execution Cost: Execution cost is the difference in value between an ideal trade and what was actually done. The execution cost of a single completed trade is typically the difference between the final average trade price, including commissions, fees, and all other costs, and a suitable *benchmark* price representing a hypothetical perfectly executed trade.
2. The Execution Cost Sign Convention: The sign is taken so that the positive cost represents a loss of value; buying for a higher price or selling for a lower price.
3. Value Assigned to Unexecuted Trades: If a trade is not completed either for endogenous reasons (for example, the price moves away from an acceptable level) or for exogenous reasons (the trader gets sick or the system fails), then some value must be assigned to the unexecuted shares.
4. Cumulative Cost of Portfolio Execution: The cost of a portfolio transaction, or a series of transactions, is computed as a suitably weighted average of the cost of individual executions.
5. Direct Costs of Trade Execution: Some of the costs of trading are direct and predictable, such as broker commissions, taxes, and exchange fees. Although these costs can be significant, they are not commonly included in the quantitative analysis of execution costs.
6. Indirect Costs of Trade Execution: “Indirect” costs include all other sources of price discrepancy, such as limited liquidity (market impact) and price motion due to volatility. These are much more difficult to characterize and measure, and are much more amenable to improvement.



7. The Arrival Price Benchmark: The benchmark is commonly taken to be the *arrival price*, that is, the quoted market price in effect at the time that the order was released to the trading desk.
8. Zero Cost Arrival Benchmark: Using that benchmark is equivalent to saying that a perfect trade – one with zero execution cost – would be one that executed instantaneously at the arrival price. The cost measured using the arrival price benchmark is called the *implementation shortfall* – a term introduced by Perold (1988).
9. Arrival Price Execution Cost Example: Suppose that an overnight investment decision assumed that a large quantity of stock could be purchased at the previous day's closing price of \$50 per share; if the trade was fully completed at an average price of \$50.25, then the execution cost would be reported at 25 cents per share.
10. Trade Duration Execution Cost Assignment: But execution costs are only part of the picture; if the stock closed that day at \$51, then the trade would be successful despite its positive cost; a naïve cost model may assign \$1 profit to the portfolio manager and 25 cents cost to the trader.
11. Execution Gain Instead of Cost: Execution costs can be negative, for example, if the price dropped in the course of a purchase program and the asset was acquired at a lower price than was anticipated; or in the above example, if the benchmark price were the day's close.
12. High Uncertainty in Cost Forecast: The forecast execution costs in any particular order have a very high degree of uncertainty, due to market volatility and other random effects.
13. Empirical Validation of the Cost Model: A well-calibrated model for the execution costs is an important part of the quantitative investment process. At a minimum it is a tool for the portfolio manager to evaluate the performance of his or her trading desk and the external brokers; were the results achieved on a particular execution compatible with the costs estimated from the pre-trade model (Almgren (2010))?
14. Cost Model as Decision Tool: Furthermore, the anticipated transaction costs should be a component of the portfolio formulation decisions; turnover should be minimized,



and the expected transaction costs must be incorporated in the portfolio construction model along with the expected alpha. Grinold and Kahn (1999) discuss in depth the use of transaction costs models in investment management.

15. Component of the Execution Process: This chapter is divided into three parts corresponding to the order in which the three aspects should be addressed in designing an investment process, although the order is reverse chronological from the point of view of a single trade. The post-trade cost reporting is looked at first, the optimal trading to minimize execution costs next, and finally the per-trade cost estimation.

## Post-Trade Reporting

1. Benchmark Relative Cost Reporting: The first step in any program to estimate execution costs is to measure them systematically. For each trade executed the cost should be reported relative to a collection of benchmarks.
2. Rolled Up Trade Cost Statistics: In addition, the cost statistics should be computed across all trades over a suitable time-period (daily or weekly) and broken down by any relevant parameters – primary market, size of trade, market capitalization of stock, etc.
3. The VWAP Cost Benchmark: As noted above the most common benchmark is the pre-trade arrival price. Another common choice is the “Volume Weighted Average Price” (VWAP) taken across the time interval over which the trade was executed. Although this most likely does not correspond to an investment goal directly, it is a popular benchmark for assessing the quality of execution, because it largely filters out the effects of volatility.
4. Post-Trade Price Benchmark: One would typically also use a post-trade price – for example the closing price on the day during which the trade was executed. Typical post-trade reporting systems display execution price relative to all of these benchmarks – before, during, and after trading.



5. Aggregating the Trade Cost Numbers: When aggregating the cost numbers across a diverse variety of trades, the individual cost numbers should be weighted so that the result is representative of the overall change in the portfolio value.
6. Pre-trade Prediction Validation: If a pre-trade cost model has been developed, then the realized values should be compared with the forecast values, both on the level of individual execution as well as the overall portfolio level. This will help identify trades that have been badly executed as well as maintaining accurate calibration of the pre-trade model.
7. Standard Deviation of the Trade: In addition to the average, it is useful to report the standard deviation of the costs. This is useful for the reality check of the significance of the mean. For example, if the standard deviation were 25 bp on 100 independent trades, then the expected error in the sample mean would be 2.5 bp and a change 1-2 bp in the mean cost would not be significant.
8. Nomenclature of the Standard Deviation: The standard deviation should also be weighted by the trade size; the most reasonable weights are the same ones used for the average cost. The standard deviation does not have good properties under sub-division.
9. Sub-dividing Aggregate Trade Costs: Ideally aggregate cost numbers would be reported so that the result is indifferent to the sub-division of the trade costs, but this is often not possible. For example, suppose that 100,000 shares were purchased throughout the day; a natural benchmark price would be the day's open price.
10. Aggregation Cost Division Granularity - Caveat: But if this block were considered as 40,000 shares in the morning and 60,000 shares in the afternoon, the arrival price for the second block would be the mid-day price and the overall reported cost would be lower. The choice can only be made with the knowledge of the investor's overall goals. This difficulty does not arise with a VWAP benchmark or with a benchmark price at a fixed time such as at the close.
11. Incomplete Trades - Transaction Cost Reporting: Additional difficulties in cost reporting come from price limits and incomplete trades. Suppose that a buy order is put in for a stock that is currently trading at \$50 but a condition is imposed that no



shares are to be bought at a price higher than \$50.05. It will then be certain that the price impact on this trade will be almost \$0.05 a share, but it may be that only a small fraction of the requested shares is actually executed. If the limit is below the initial price, then the situation may be even more extreme. Similar difficulties arise if the trade is halted for other reasons.

12. Process Oriented Trade Cost Reporting: The solution to this difficulty rests on the valuation of the unexecuted shares, but there is no simple rule. The most straightforward would be to imagine that the unexpected shares were purchased at the day's closing price. In practice this rule is far *too* stringent and does not take into account the variety of possible reasons that the trade may have been halted due to. Trade cost reporting must take into account of the entire investment process.

## Optimal Trading

1. Motivation behind “Best Execution” Practice: Once a system is in place for measuring and reporting trade costs, and after it has been agreed what criteria define a good trade or a collection of trades, one can then design optimal strategies to meet investment goals. “Best execution” is not only advantageous to investors, but is also required by regulation in most markets.
2. Goals of the “Best Execution” Practice: The most important goal is always to reduce the mean expectation cost, which is largely due to market impact and the uncaptured short-term alpha. One also often desires to reduce the standard deviation of the trade costs in order to reduce the overall investment volatility.
3. Reduction in Market Impact Costs: Reducing cost is achieved by searching for liquidity in “space” as well as in time, accessing as many pools of potential liquidity, and as many potential counterparties for each piece of trade. This includes routing to non-exchange trading venues such as “dark pools” and block-crossing services when possible. Rapid changes in the market structure make this increasingly complicated – see Hasbrouck (2007) for a discussion of the US equity markets.



4. Consequence of Altering the Trade Rate: Accessing liquidity in time means willing to slow trading to give potential counterparties the time to appear in the market. If the short-term price drift is not expected to be significant, then the average trading costs can generally be reduced by trading more slowly.
5. Factors Determining the Optimal Execution: Other aspects of trading push for rapid trading – most significantly the anticipated price drift and the underlying market volatility (Alford, Jones, and Lim (2003)). The actual trade schedule is determined by using a quantitative balance of all these factors. This schedule may need to be dynamically adapted depending on the observed liquidity and other market components. For a portfolio the correlation between the assets should be included in addition to the volatility of each one, and the schedule may need to maintain a strict neutrality to a market index or other risk factors.
6. Mean Variance Tradeoff Strategies: In practice, optimal strategies for a mean-variance tradeoff, and optimal strategies that are calculated to optimize expected costs in the presence of short-term drift, are generally similar; at the beginning of each execution rapid trading reduces the exposure to the volatility or alpha, then the trading slows to reduce the overall impact costs (Almgren and Chriss (2000)).
7. Determining the Speed of Execution: The most important question is the overall speed of execution; should it be completed in minutes, hours, or days? Regardless of the model, an approximate quantitative model is an essential element of this decision.

## Pre-trade Cost Estimation

1. Estimating Transaction Cost Mean/Uncertainty: After a measurement system has been running for long enough to generate a useful amount of data, one can then think about developing an analytical model. The goal of the model is to forecast the execution cost to be expected on any anticipated trade, along with an estimate of the uncertainty of the forecast.



2. Market Impact Models in Literature: In its most general form such a model must contain a full explanation of how trading in markets affects prices. This is a rich area of research with a large literature – see Madhavan (2000) for an extensive survey, and Lillo, Farmer, and Mantegna (2003) and Bouchaud, Gefen, Potters, and Wyart (2004) for more subtle models.
3. Prediction of the Transaction Cost: The goal is to give the predicted cost  $\mathcal{C}$  - in cents per share of bp – in terms of input variables, including at a minimum the number of shares traded  $X$ , the daily volume  $V$  of the stock (either historical average of volume on the day of trading), and the effective duration  $T$  over which the trade was executed.
4. Normalization of Input/Output Metrics: In order to normalize across many different stocks, one should also include properties such as volatility  $\sigma$ , perhaps market capitalization, primary exchange/country, and other economic parameters. In addition, one might include trade information across anticipated short-term alpha, and which algorithm or other means was used to trade the stock.
5. Classification using Supervised Learning Algorithms: For any proposed trade the pre-trade model predicts the cost based on the cost that was measured using *similar* trades in the past. “Supervised learning” algorithms (Poggio and Smale (2003)) could in principle be used to carry out this classification.
6. Curse of the Input Dimensionality: In practice because the dimension of the parameter space is so large, there may be no, or very few, trades that are sufficiently similar to the proposed ones in all variables. This argues for a regression approach in which a specific functional form is proposed, calibrated to data, and evaluated using standard statistical tests of significance.
7. Empirical Calibration of Market Impact: The most detailed such study was done by Almgren, Thum, Hauptmann, and Li (2005). The model is based on a separation of costs into “permanent” and “temporary” components. Although both trade components contribute to a realized cost on every trade, it is useful to calibrate each separately.



8. Price Observables used in the Calibration: The calibration relies on three observed prices;  $S_{PRE}$  denotes the arrival price, that is, the market price just before the order begins trading; typically, this would be the bid-ask mid-point before the first execution.  $S_{EXEC}$  is the actual average price for which the order finally executes, which is of course the quantity of most interest to the trader.  $S_{POST}$  denotes the market price just after the order has finished executing, possibly with a short time lag to allow transient effects to dissipate.
9. Temporary/Permanent Market Impact Signals: The model of Almgren, Thum, Hauptmann, and Li (2005) measures two quantities  $I$  and  $J$  for permanent and temporary impacts, respectively. For a buy order these are

$$I = \log \frac{S_{POST}}{S_{PRE}} \approx \frac{S_{POST} - S_{PRE}}{S_{PRE}}$$

$$J = \log \frac{S_{EXEC}}{S_{PRE}} \approx \frac{S_{EXEC} - S_{PRE}}{S_{PRE}}$$

and for a sell order the sign would be reversed.

10. Model Parameter Applicability Range: The approximate equalities are valid for moderate sized order for which the price does not move more than a few percent during execution. These quantities represent the price changes relative to the pre-trade price, as fractions of that benchmark.
11. The Permanent Market Impact Signal: The permanent component of the cost is interpreted as the net displacement of the market due to the buy/sell imbalance introduced by this particular trade. The simplest model sets

$$I = \gamma \sigma \frac{X}{V} + \langle\langle noise \rangle\rangle$$

In this expression the trade size  $X$  is normalized by the trade volume  $V$  so  $\frac{X}{V}$  is the trade size as a function of the typical day's flow.



12. The Universal Permanent Impact Parameters: The impact  $I$  is also normalized by the volatility  $\sigma$ , so that then impact may be represented as a fraction of the typical amount the stock moves without trading. The trading coefficient  $\gamma$  is taken to be constant across all stocks, widely varying daily volumes and volatilities. The noise term arises from intra-day volatility, and the actual cost experienced on any particular trade may be very different from the mean predictions from the model.
13. Linear Permanent Market Impact Coefficient: In this form of the model the permanent impact is linear in the total number of shares traded. This model is particularly convenient for theoretical modeling, but must be justified by empirical analysis of real trade data. Almgren, Thum, Hauptmann, and Li (2005) found a reasonable agreement with their data, but later other empirical studies have suggested that the linear form might not be justified.
14. The Temporary Market Impact Signal: The temporary impact component of the cost is interpreted as the additional premium that must be paid for execution in a finite time, above a slightly prorated fraction of the permanent cost. It is modeled as

$$J = \frac{I}{2} + \eta \sigma \left( \frac{X}{VT} \right)^\beta + \langle \langle \text{noise} \rangle \rangle$$

where  $T$  represents the effective duration of the trade as a fraction of the trading day. Thus  $\frac{X}{VT}$  represents the “participation rate”, of the fraction of the market flow that this trade constitutes during the time it is active.

15. The Universal Temporary Impact Parameters: The factor  $\frac{1}{2}$  on the permanent cost component represents the fraction of the post-trade impact that is paid on the trade itself. As for the permanent cost, the impact cost is expressed as a fraction of typical volatility; the coefficient  $\eta$  is taken to be constant across all stocks. Other factors such as bid-ask spread and market capitalization were not determined to be significant in the US equity markets.
16. Two Stage Permanent/Temporary Calibration: Almgren, Thum, Hauptmann, and Li (2005) calibrate this model to a large sample of US equity trades using a two-step



procedure. First the permanent term  $I$  is calibrated, testing the hypothesis of linear impact and estimating the value of  $\gamma$ . Next the temporary impact term  $J$  is calibrated, determining the values for the exponent  $\beta$  and the coefficient  $\eta$ . A key aspect of the verification is the characterization of the error terms to verify that volatility is an adequate explanation of the residuals.

17. Permanent/Temporary Calibrated Market Parameters: The result has

$$\beta \approx 0.6$$

which is roughly compatible with earlier square root models (Barra (1997)) as well as for the coefficients  $\gamma$  and  $\eta$ . For trades that are a few percent of the daily volume executed across several hours, the predicted impact costs are 10's of basis points.

18. Approximate Nature of the Model #1: A number of factors make such models at best approximate. First is the extremely low values of  $R^2$  in the regression, typically around a few percent at best. That is, the ability of the model to predict the cost of any single trade is very poor, because the market volatility due to other trading activity is usually very large.
19. Approximate Nature of the Model #2: Second is the difficulty of the model in differentiating market impact from alpha; did the price move up because the buy program impacted the price, or was the trade executed because the manager correctly anticipated a price rise?
20. Approximate Nature of the Model #3: A third difficulty is the difference in behavior between small and large trades; although the model claims to be universally valid, in practice a model developed for small trades gives poor results on large trades.
21. Approximate Nature of the Model #4: In any particular application, the model should be critically evaluated by the user and recalibrated and extended as necessary. Despite its intrinsic difficulties and limitations, this model and its extensions are extremely useful in providing approximate anticipated cost values for pre-trade planning, optimal trade scheduling, and post-trade evaluation.



## References

- Alford, A., R. Jones, and T. Lim (2003): Equity Portfolio Management, in: *Modern Investment Management: An Equilibrium Approach – R. Litterman (editor)* Wiley.
- Almgren, R. F., and N. Chriss (2000): Optimal Execution of Portfolio Transactions *Journal of Risk* **3** (2) 5-39.
- Almgren, R.F., C. Thum, E. Hauptmann, and H. Li (2005): Equity Market Impact *Risk* **18** (7) 57-62.
- Almgren, R. F. (2010): Execution Costs, in: *Encyclopedia of Quantitative Finance* Wiley 1177-1181.
- Barra (1997): **Market Impact Handbook**.
- Bouchaud, J. P., Y. Gefen, M. Potters, and M. Wyart (2004): Fluctuations and Responses in Financial Markets: The Subtle Nature of “Random” Price Changes *Quantitative Finance* **4** (2) 176-190.
- Engle, R., and R. Ferstenberg (2007): Execution Risk: It’s the same as Investment Risk *Journal of Portfolio Management* **33** (2) 34-44.
- Grinold, R. C., and R. N. Kahn (1999): *Active Portfolio Management 2<sup>nd</sup> Edition* McGraw-Hill.
- Hasbrouck, J (2007): *The Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading* Oxford University Press.
- Lillo, F., J. D. Farmer, and R. N. Mantegna (2003): Master Curve for Price-Impact Function *Nature* **421** 129-130.
- Madhavan, A. (2000): Market Micro-structure: A Survey *Journal of Financial Markets* **3** 205-258.
- Perold, A. F. (1988): The Implementation Shortfall: Paper vs. Reality *14* (3) 4-9.
- Poggio, T., and S. Smale (2003): The Mathematics of Learning: Dealing with Data *Notices of the American Mathematical Society* **50** (5) 537-544.



## Execution of Portfolio Transactions – Optimal Trajectory

### Overview, Scope, and Key Results

1. Portfolio Transactions under Market Impact: Almgren and Chriss (2000) consider the execution of portfolio transactions with the aim of minimizing a combination of volatility risk and transaction costs arising from temporary and permanent market impact.
2. Efficient Frontier under Linear Cost: For a simple linear cost model, they explicitly construct an *efficient frontier* in the space of time-dependent liquidation strategies, which have the minimum expected cost for a given level of uncertainty.
3. Choice of the Utility Function: This enables one to select optimal strategies either by minimizing a quadratic utility function, or by minimizing the Value-at-Risk.
4. Liquidity Adjusted Value at Risk: The latter choice leads to the concept of liquidity-adjusted VaR, or L-VaR, that explicitly considers the best trade-off between the volatility risk and the liquidity costs.

### Motivation Background, and Synopsis

1. Transactions Changing the Portfolio Composition: Almgren and Chriss (2000) consider the optimal execution of portfolio transactions that move a portfolio from a given starting composition to a specified final composition within a specified period of time.
2. The Bertsimas and Lo Approach: Bertsimas and Lo (1998) define the best execution as the dynamic trading strategy that provides the minimum cost of trading over a fixed period of time, and they also show that under a variety of circumstances one can



find such a strategy by employing a dynamic optimization procedure; but they ignore the volatility of revenues of different trading strategies.

3. Maximization of Expected Trading Revenue: Almgren and Chriss (2000) work in the more general framework of maximizing the *expected revenue* – or equivalently minimizing the costs – with a suitable penalty for the *uncertainty* of revenue (or cost).
4. Market Microstructure Framework: This general framework arises in the market microstructure theory, but with a different purpose in mind. The *uninformed discretionary trader* trades an exogenous endowment over an exogenously specified amount of time to maximize the profits (Admati and Pfleiderer (1988)); the informed strategic trader trades over multiple periods on information not widely available, again to maximize profits (Kyle (1985)). In both cases the literature focuses on the link between the trader and the market maker, and a theory is produced to predict the market clearing price of the security at each period. Thus, a trader's optimal strategy is used as a means to study the price formation in the markets, not as an object of interest in itself.
5. Variance of the Trading Cost: Almgren and Chriss (2000) study the variance of the trading cost in optimal execution because it fits in with the intuition that the trader's utility should figure in the definition of *optimal* in “optimal execution”.
6. Example: Trading Illiquid Volatile Securities: For example, in trading a highly illiquid, volatile security, there are two extreme outcomes; trade everything now at a known, but high, cost, or trade in equal sized packets over a fixed time at a relatively lower cost. The latter strategy has a lower expected cost, but this comes at the expense of greater uncertainty in the final revenue.
7. Estimation of the Trading Uncertainty: How to evaluate the above uncertainty is partly subjective, and is a function of the trader's tolerance for risk. All that can be done is to insist that for a given level of uncertainty that the cost be minimized. This idea extends to a complete theory of optimal execution that includes an efficient frontier of optimal execution strategies.
8. Consistency with Expectations from Intuition: The framework of risk in execution yields several results that are consistent with the intuition. For example, it is evident



that all else equal, a trader will choose to execute a block of illiquid security less rapidly than a liquid security.

9. Models Lacking Consistency with Intuition: While this seems obvious, Almgren and Chriss (2000) demonstrate that a model that ignores risk does not have this property; without enforcing a strictly positive penalty for risk one cannot produce models that trade differently across the spectrum of liquidity.
10. Arithmetic Brownian Motion Price Dynamics: The incorporation of risk into optimal execution does not come without cost. First, in order to be able to produce tractable analytical results, Almgren and Chriss (2000) are forced to work in largely in the framework of price dynamics that are an arithmetic walk with independent increments.
11. Use of Static Optimization Procedures: They obtain results using *static optimization* procedures which they show lead to globally optimal trading trajectories. That is, optimal trading paths may be determined in advance of trading. Only the composition of the portfolio and the trader's utility function figure on the trading path.
12. Why does Static Optimization Work? The fact that the static strategy can be optimal even when the trader has the option to dynamically change his trading mid-course is a direct result of the assumptions of independence of returns and symmetry for the penalty functions for risk.
13. Using Non-Symmetric Penalty Functions: An interesting deviation from the symmetric penalty function was communicated by Ferstenberg, Karchmer, and Malamut at ITG Inc. They argue that the opportunity is a subjective quantity and is measured differently by different traders. Using a trader defined cost function  $g$ , they define opportunity costs as the expected costs of  $g$  applied to the average execution price obtained by the trader relative a benchmark price. They assume that the risk-averse traders will use a convex function  $g$  that is not symmetric in the sense that there is a strictly greater penalty for underperformance than for the same level of outperformance. They show that in this setting, the optimal strategy relative to  $g$  not only depends on the time remaining, but also on the performance of the strategy up to



the present time, and the present price of the security. In particular, this means that in their setting, optimal strategies are dynamic.

14. Serial Correlations among Price Movements: As it is well known that price movements exhibit some serial correlations across various time horizons (Lo and MacKinlay (1988)), that market conditions change, and that some participants possess private information (Bertsimas and Lo (1998)), one may question the usefulness of results that obtain strictly in an independent-increment framework.
15. The Dynamic Nature of Trading: Moreover, as trading is known to be a dynamic process, the conclusion that optimal trading strategies can be statically determined calls for critical examination. Almgren and Chriss (2000) examine what quantitative gains are available that incorporate all the relevant information.
16. Impact of the Serial Correlations: First they consider short term serial correlations in price movements. They demonstrate that the marginal improvements available by explicitly incorporating this information into trading strategies is small, and more importantly, independent of the portfolio sizes; as portfolio sizes increase, the percentage gains possible decrease proportionately.
17. Combining “Correlated” and “Shifting” Strategies: The above is precisely true for linear transaction cost models, and is approximately true for more general models. The results of Bertsimas and Lo (1998) suggest that trading a strategy built to take advantage of serial correlation will essentially be a combination of a “correlation free” strategy and a “shifting strategy” that moves from one trade period to the next based on the information available in the last period’s return. Therefore, Almgren and Chriss (2000) argue that by ignoring serial correlation, they a) preserve the main interesting features of their analysis, and b) introduce virtually no bias away from “truly optimal” solutions.
18. Impact of Scheduled News Events: Second, Almgren and Chriss (2000) examine the impact of scheduled new events on optimal execution strategies. There is ample evidence that anticipated news announcements, depending on their outcome, can have a significant temporary impact on the parameters governing price movements.



19. Scheduled News Events - Literature Review: For a theoretical treatment see Brown, Harlow, and Tinic (1988), Kim and Verrecchia (1991), Easterwood and Nutt (1999), and Ramaswami (1999). For empirical studies concerning earnings announcements, see Patell and Wolfson (1984) for changes in mean and variance of intra-day prices, and Lee, Mucklow, and Ready (1993) and Krinsky and Lee (1996) for changes in the bid-ask spread. For additional studies concerning news announcements, see Charest (1978), Morse (1981), and Kalay and Loewenstein (1985).
20. Model Incorporation of Scheduled Events: Almgren and Chriss (2000) work in a simple extension of their static framework by assuming that the security again follows an arithmetic random walk, but at a time known at the beginning of trading, an uncorrelated event will cause a material shift in price dynamics, e.g., an increase or decrease of volatility.
21. Combining Piece-Wise Static Strategies: In this context they show that optimal strategies are piece-wise static. To be precise, they show that an optimal strategy entails following a static strategy up to the moment of the event, followed by another static strategy that can only be determined once the outcome of the event is known.
22. Variation from the Original Static Strategy: It is interesting to note that the static strategy that one follows in the first leg is in general not the same strategy one would follow in the absence of information concerning the event.
23. Accommodating Unanticipated External “Sudden” Events: Finally, Almgren and Chriss (2000) note that any optimal execution strategy is vulnerable to *unanticipated events*. If such an event occurs during the course of trading and causes a material shift in the parameters of the price dynamics, then indeed a shift in the optimal trading trajectory must also occur.
24. Adaptation at Parameter Shift Edges: However, if one makes a simplifying assumption that all events are either “scheduled” or “unanticipated” one then concludes that optimal execution is always a game of static trading punctuated by shifts in the trading strategies that adapt to material changes in the price dynamics.
25. Pre-determined vs. Active Approaches: If shifts are caused by events that are known ahead of time, then optimal execution benefits from a precise knowledge of the



possible outcomes of the event. If not, the best approach is to be actively “watching” the market for such changes and react swiftly should they occur.

26. Simple Proxy for Unexpected Uncertainty: One approximate way to include such completely unexpected uncertainty into the model is to artificially raise the value of the volatility parameter.
27. Risk Averse Optimal Trading Strategies: As a first step, Almgren and Chriss (2000) obtain closed form solutions for trading strategies for any level of risk aversion.
28. Efficient Frontier of Optimal Strategies: They then show that this leads to an efficient frontier of optimal strategies, where an element of the frontier is represented by a strategy with a minimal level of cost for its level of variance of the cost.
29. Graphical Structure of the Frontier: The structure of the frontier is of some interest. It is a smooth convex function differentiable at its minimal point. The minimal point is what Bertsimas and Lo (1998) call the naïve strategy because it corresponds to trading equally sized packets using all available trading time equally.
30. Differential at the Minimum Point: The differentiability of the frontier at its minimum point indicates that one can obtain a first order reduction in the variance of the trading cost at the expense of only a second order in cost by trading a strategy slightly away from the globally minimal strategy.
31. Curvature at the Minimal Point: The curvature of the frontier at its minimum point is a measure of the liquidity of the security.
32. Half-Life of Optimal Execution: Another ramification of the Almgren and Chriss (2000) study is that for all levels of risk aversion except risk neutrality, optimal execution trades have a “half-life” that fall out of the calculations.
33. Independence from the Time to Complete Execution: A trade’s half-life is independent of the actual specified time to liquidation, and is a function of the security’s liquidity and volatility, and the trader’s level of risk aversion.
34. Half-Life as Execution Time: As such Almgren and Chriss (2000) regard the half-life as an idealized time to execution, and perhaps a guide to the proper amount of time over which to execute a transaction.



35. Time Lesser than Half Life: If the specified time to liquidation is short relative to the trade's half-life, one can expect the cost of trading to be dominated by transaction costs.
36. Time Greater than Half Life: If the time to trade is long relative to the half-life, one can then expect most of the liquidation to take place well in advance of the limiting time.

## The Definition of a Trading Strategy

1. Price Dynamics and Trade Execution: As a starting point, Almgren and Chriss (2000) define a trading strategy, and lay out the dynamics that they study. They start with a formal definition of a strategy for a sell program consisting of liquidating a single security. The definitions and results are analogous for a buy program.
2. Problem Setup - Security Liquidation: Suppose that the seller holds a block of  $X$  units of a security that they want to completely liquidate before time  $T$ . To keep the discussion, Almgren and Chriss (2000) speak of *units* of a security. Specifically, they have in mind shares of stock, futures contract, and units of a foreign currency.
3. Trading Strategy - Price/Unit Strategy: The seller divides  $T$  into  $N$  units of length

$$\tau = \frac{T}{N}$$

and defines the discrete times

$$t_k = k\tau$$

for

$$k = 0, \dots, N$$



The *trading trajectory* is defined to be the list  $x_0, \dots, x_N$  where  $x_k$  is the number of units that the seller plans to hold at time  $t_k$ .

4. Outright/Re-balanced Trajectories: The initial holding is

$$x_0 = X$$

and liquidation at time  $T$  requires

$$x_N = 0$$

A trading trajectory can be thought of as either the ex-post realized trades resulting from some process, or as a plan concerning how to trade a block of securities. In either case one may also consider *re-balancing* trajectories by requiring

$$x_0 = X$$

the initial position, and

$$x_1 = Y$$

the new position, but this is formally equivalent to studying trajectories of the form

$$x_0 = X - Y$$

and

$$x_N = 0$$



5. Outstanding Holdings/Incremental Trade Lists: Equivalently, a strategy may be specified using the “trade list”  $n_1, \dots, n_N$  where

$$n_k = x_{k-1} - x_k$$

is the number of units that the seller will sell between times  $t_{k-1}$  and  $t_k$ . Clearly,  $x_k$  and  $n_k$  are related by

$$x_k = X - \sum_{j=1}^k n_j = \sum_{j=k+1}^N n_j$$

$$k = 0, \dots, N$$

6. Simultaneous Portfolio Buying and Selling: Almgren and Chriss (2000) also consider more general programs of buying and selling simultaneously several securities.
7. Inter-Execution Time Interval Specification: For notational simplicity they consider all the time intervals to be of equal length  $\tau$ , but this restriction is not essential.
8. Behavior at  $N/\tau$  Limits: Although they do not discuss it, in all their results it is easy to take the continuous-time limit of

$$N \rightarrow \infty$$

and

$$\tau \rightarrow 0$$

9. Definition of a Trading Strategy: Almgren and Chriss (2000) define a “trading strategy” to be a rule for determining  $n_k$  in terms of the information available at  $t_{k-1}$ . Broadly speaking they distinguish between two types of trading strategies – static and dynamic.



10. Static vs. Dynamic Trading Strategy: Static strategies are determined in advance of trading, that is the rule for determining each  $n_k$  depends only on information available at  $t_0$ . Dynamic strategies, conversely, depend on all information up to, and including, time  $t_{k-1}$

## Price Dynamics

1. Exogenous/Endogenous Price Move Factors: Suppose that the initial security price is  $S_0$  so that the initial market value of the position is  $XS_0$ . The securities' price evolves according to two exogenous factors – volatility and drift, and one endogenous factor – market impact.
2. Market Forces vs. Trading Impact: Volatility and drift are assumed to be the result of market forces that occur randomly and independent of the trading.
3. Earlier Literature on Market Impact: Almgren and Chriss (2000) discussion largely reflect the work of Kraus and Stoll (1972), and the subsequent works of Holthausen, Leftwich, and Mayers (1987, 1990) and Chan and Lakonishok (1993, 1995). See also Keim and Madhavan (1995, 1997).
4. Origin of the Market Impact: As the market participants begin to detect the volume that the seller (buyer) is selling (buying), they naturally adjust their bids (offers) downward (upward). Almgren and Chriss (2000) distinguish two kinds of market impact.
5. Definition of Temporary Market Impact: *Temporary* impact refers to the temporary imbalances in supply and demand caused by the seller's trading leading to temporary price movements away from equilibrium.
6. Definition of Permanent Market Impact: *Permanent* impact refers to the changes in the “equilibrium” price due to the seller's trading, which remain at least for the life of the liquidation.
7. Price Evolution Stochastic Difference Equation: Almgren and Chriss (2000) assume that the security price evolves according to the discrete random walk



$$S_k = S_{k-1} + \sigma\sqrt{\tau}\xi_k - \tau g\left(\frac{n_k}{\tau}\right)$$

for

$$k = 1, \dots, N$$

8. Glossary of the Equation Terms: Here  $\sigma$  represents the volatility of the asset,  $\xi_k$ 's are draws from independent random variables each with zero mean and unit variance, and the permanent impact function  $g(v)$  is a function of the *average rate* of trading

$$v = \frac{n_k}{\tau}$$

during the interval  $t_{k-1}$  to  $t_k$ .

9. Lack of Explicit Drift Term: In the above equation there is no drift term. Almgren and Chriss (2000) indicate that this is due to the assumption that they have no information about the direction of the future price movements.
10. Trading Term Horizons under Consideration: Over long-term investment time scales, or in extremely volatile markets, it is important to consider *geometric* rather than arithmetic Brownian motion – this corresponds to letting  $\sigma$  in

$$S_k = S_{k-1} + \sigma\sqrt{\tau}\xi_k - \tau g\left(\frac{n_k}{\tau}\right)$$

scale with  $S$ . But over short term “trading” horizons of interest, the total fractional price changes are small, and the differences between arithmetic and geometric Brownian motions are negligible.



## Temporary Market Impact

1. Intuition behind the Temporary Market Impact: The intuition behind the temporary market impact is that a trader plans to sell a certain number of units  $n_k$  between times  $t_k$  and  $t_{k-1}$ , but may work the order in several smaller sizes to locate optimal points of liquidity.
2. Liquidity Reduction Impact on Price: If the total number of units  $n_k$  is sufficiently large, the execution price may steadily decrease between  $t_{k-1}$  and  $t_k$  in part due to the exhaustion of the supply of liquidity at each successive price level. This effect is assumed to be short-lived, and in particular, liquidity is assumed to return back after each period, and a new equilibrium price is established.
3. The Temporary Price Impact Function: This effect is modeled by introducing a temporary price impact function  $h(v)$ , the temporary drop in the average price per share caused by trading at an average rate  $v$  during one time interval.
4. Net Price Received at Execution: Given this, the actual price per share received on sale  $k$  is

$$\tilde{S}_k = S_{k-1} - h\left(\frac{n_k}{\tau}\right)$$

but the effect of  $h(v)$  does not appear in the next “market” price  $S_k$ .

5. Choice of Market Microstructure: The functions  $g(v)$  in

$$S_k = S_{k-1} + \sigma\sqrt{\tau}\xi_k - \tau g\left(\frac{n_k}{\tau}\right)$$

and  $h(v)$  in

$$\tilde{S}_k = S_{k-1} - h\left(\frac{n_k}{\tau}\right)$$



may be chosen to reflect any preferred model of market microstructure, subject only to certain natural convexity conditions.

## Capture and Cost of Trading Trajectories

1. Capture across a Trading Trajectory: Almgren and Chriss (2000) then discuss the profits resulting from trading along a certain trajectory. They define the *capture* of a trajectory to be the full trading revenue upon completion of all trades. Due to the short term horizons that they consider, they do not include any notion of carry or time value of money in their discussions.
2. Full Trading Revenue across Execution: Thus, the capture is the sum of the product of the number of units  $n_k$  sold in each time interval times the effective price per share  $\tilde{S}_k$  received on that sale. It is readily computed as

$$\sum_{k=1}^N n_k \tilde{S}_k = X S_0 + \sum_{k=1}^N \left[ \sigma \sqrt{\tau} \xi_k - \tau g\left(\frac{n_k}{\tau}\right) \right] x_k + \sum_{k=1}^N n_k h\left(\frac{n_k}{\tau}\right)$$

3. Decomposition of the Capture Components: The first term on the RHS above is the initial market value of the position; each additional term represents a gain or a loss due to a specific market factor.
4. The Volatility Price Impact Term: The first term  $\sigma \sqrt{\tau} \xi_k x_k$  represents the total impact from the volatility.
5. The Permanent Market Impact Term: The permanent market impact term  $-\tau x_k g\left(\frac{n_k}{\tau}\right)$  represents the loss in the value of the position caused by a permanent price drop associated with selling a small piece of the position.
6. The Temporary Market Impact Term: And the temporary market impact term  $n_k h\left(\frac{n_k}{\tau}\right)$  is the price drop due to selling, acting only on the units sold during the  $k^{th}$  period.



7. The Total Cost of Trading: The *total cost of trading* is the difference  $XS_0 - \sum_{k=1}^N n_k \tilde{S}_k$  between the initial book value and the capture. This is the standard *ex-post* measure of the performance costs used in performance evaluations, and is essentially what Perold (1988) calls *implementation shortfall*.
8. Estimation of Implementation Short-fall: In this model, prior to trading, the implementation short-fall is a random variable. Write  $\mathbb{E}[X]$  for the expected short-fall and  $\mathbb{V}[X]$  for the variance of the short-fall.
9. Implementation Short-fall Mean/Variance: Given the simple nature of price dynamics, Almgren and Chriss (2000) readily compute

$$\mathbb{E}[X] = \sum_{k=1}^N \tau x_k g\left(\frac{n_k}{\tau}\right) + \sum_{k=1}^N n_k h\left(\frac{n_k}{\tau}\right)$$

$$\mathbb{V}[X] = \sigma^2 \sum_{k=1}^N \tau x_k^2$$

The units of  $\mathbb{E}[X]$  are in dollars, and the units of  $\mathbb{V}[X]$  are dollars squared.

10. Distribution of Implementation Short-fall: The distribution of the short-fall is Gaussian if  $\xi_k$  is Gaussian, in any case if  $N$  is large, it is very nearly Gaussian.
11. Almgren and Chriss Minimizer Utility: Almgren and Chriss (2000) devote much of their paper to finding trajectories that minimize  $\mathbb{E}[X] + \lambda \mathbb{V}[X]$  for various values of  $\lambda$ . They demonstrate that for each value of  $\lambda$  there corresponds a unique trading trajectory  $x$  such that  $\mathbb{E}[X] + \lambda \mathbb{V}[X]$  is minimal.

## Linear Impact Functions

1. Linear Temporary/Permanent Market Impact: Although Almgren and Chriss (2000) formulation does not require it, computing optimal trajectories is significantly easier



if one takes the permanent and temporary impact functions to be *linear* in the rate of trading.

2. Linear Permanent Impact Market Function: For linear permanent impact,  $g(v)$  has the form

$$g(v) = \gamma v$$

in which the constant  $\gamma$  has units of (\$/share)/share.

3. Corresponding Execution Time Security Price: With this form, each  $n$  units sold depresses the price per share by  $\gamma n$  regardless of the time taken to sell  $n$  units.

$$S_k = S_{k-1} + \sigma\sqrt{\tau}\xi_k - \gamma g\left(\frac{n_k}{\tau}\right)$$

readily yields

$$S_k = S_0 + \sigma \sum_{j=1}^k \sqrt{\tau_j} \xi_j - \tau \gamma (X - x_k)$$

4. Permanent Implementation Short-fall Mean: Then summing by parts, the permanent impact term in

$$\mathbb{E}[X] = \sum_{k=1}^N \tau x_k g\left(\frac{n_k}{\tau}\right) + \sum_{k=1}^N n_k h\left(\frac{n_k}{\tau}\right)$$

becomes



$$\begin{aligned}
\sum_{k=1}^N \tau x_k g\left(\frac{n_k}{\tau}\right) &= \gamma \sum_{k=1}^N x_k n_k = \gamma \sum_{k=1}^N x_k (x_k - x_{k-1}) \\
&= \frac{1}{2} \gamma^2 \sum_{k=1}^N [x_{k-1}^2 - x_k^2 - (x_k - x_{k-1})^2] = \frac{1}{2} \gamma X^2 - \frac{1}{2} \gamma \sum_{k=1}^N n_k^2
\end{aligned}$$

5. Linear Temporary Impact Market Function: Similarly, for the temporary impact we take

$$h\left(\frac{n_k}{\tau}\right) = \epsilon sgn(n_k) + \frac{\eta}{\tau} n_k$$

where  $sgn$  is the sign function.

6. Estimating the Fixed Costs of Execution: The units of  $\epsilon$  are \$/share, and those of  $\eta$  are (\$/share)/(share/time). A reasonable estimate for  $\epsilon$  is the fixed cost of selling, such as half of bid-ask spread plus premium.
7. Estimating the Linear Impact Coefficient: It is more difficult to estimate  $\eta$  since it depends on the internal and the transient aspects of the market microstructure. It is in this term that one would expect the on-linear terms to be most important, and the approximation

$$h\left(\frac{n_k}{\tau}\right) = \epsilon sgn(n_k) + \frac{\eta}{\tau} n_k$$

to be most doubtful.

8. Total Temporary Impact Function: The linear model

$$h\left(\frac{n_k}{\tau}\right) = \epsilon sgn(n_k) + \frac{\eta}{\tau} n_k$$

is often called a *quadratic* cost because the total costs incurred by buying or selling  $n$  units in a single unit of time is



$$nh\left(\frac{n}{\tau}\right) = \epsilon|n| + \frac{\eta}{\tau}n^2$$

9. Temporary Implementation Short-fall Mean: With both linear cost models

$$g(v) = \gamma v$$

and

$$h\left(\frac{n_k}{\tau}\right) = \epsilon sgn(n_k) + \frac{\eta}{\tau}n_k$$

the expectation of the impact costs

$$\mathbb{E}[X] = \sum_{k=1}^N \tau x_k g\left(\frac{n_k}{\tau}\right) + \sum_{k=1}^N n_k h\left(\frac{n_k}{\tau}\right)$$

becomes

$$\mathbb{E}[X] = \frac{1}{2}\gamma X^2 + \epsilon \sum_{k=1}^N |n_k| + \frac{\tilde{\eta}}{\tau} \sum_{k=1}^N n_k^2$$

in which

$$\tilde{\eta} = \eta - \frac{1}{2}\gamma\tau$$

10. Strictly Convex Nature of  $\mathbb{E}[X]$ : Clearly  $\mathbb{E}[X]$  is a strictly convex function as long as

$$\tilde{\eta} > 0$$



Note that if  $n_k$  all have the same sign, as would be the case for a pure sell program or a pure buy program, then

$$\sum_{k=1}^N |n_k| = |X|$$

11.  $\mathbb{E}[X]$  and  $\mathbb{V}[X]$  Computation Illustration: To illustrate, Almgren and Chriss (2000) compute  $\mathbb{E}[X]$  and  $\mathbb{V}[X]$  for linear impact functions for two of trajectory schemes at the opposite extremes: sell at a constant rate, and sell to maximize variance without regard to transaction costs.
12. Minimum Impact: Constant Execution Rate: The most obvious trajectory is to sell at a constant rate over the entire liquidation period. Thus, one takes each

$$n_k = \frac{X}{N}$$

and

$$x_k = (N - k) \frac{X}{N}$$

$$k = 1, \dots, N$$

13. Minimum Impact  $\mathbb{E}[X]$  and  $\mathbb{V}[X]$ : From

$$\mathbb{E}[X] = \sum_{k=1}^N \tau x_k g\left(\frac{n_k}{\tau}\right) + \sum_{k=1}^N n_k h\left(\frac{n_k}{\tau}\right)$$

and



$$\mathbb{E}[X] = \frac{1}{2}\gamma X^2 + \epsilon \sum_{k=1}^N |n_k| + \frac{\tilde{\eta}}{\tau} \sum_{k=1}^N n_k^2$$

one has

$$\mathbb{E}[X] = \frac{1}{2}XTg\left(\frac{X}{T}\right)\left(1 - \frac{1}{N}\right) + Xh\left(\frac{X}{T}\right) = \frac{1}{2}\gamma X^2 + \epsilon X + \tilde{\eta} \frac{X^2}{T}$$

and from

$$\mathbb{V}[X] = \sigma^2 \sum_{k=1}^N \tau {x_k}^2$$

$$\mathbb{V}[X] = \frac{1}{3}\sigma^2 X^2 T \left(1 - \frac{1}{N}\right) \left(1 - \frac{1}{2N}\right)$$

14. Minimum Impact N/T Limits: The trajectory minimizes total expected costs, but the variance may be large if the period  $T$  is long. As the number of trading periods

$$N \rightarrow \infty$$

$$v = \frac{X}{T}$$

remains finite, and  $\mathbb{E}[X]$  and  $\mathbb{V}[X]$  have finite limits.

15. Minimum Variance: One Step Execution: The other extreme is to execute the entire position in the first time-step. One then takes

$$n_1 = X$$



$$n_2 = \dots = n_N = 0$$

$$x_1 = \dots = x_N = 0$$

which results in

$$\mathbb{E}[X] = Xh\left(\frac{X}{\tau}\right) = \epsilon X + \eta \frac{X^2}{\tau}$$

and

$$\mathbb{V}[X] = 0$$

16. Minimum Variance N/T Limits: The trajectory has the smallest possible variance – equal to zero – because of the way time has been discretized in the model above. If  $N$  is large and hence  $\tau$  is short, then on the full initial portfolio, one takes a price hit that can be arbitrarily large.
17. Trajectory between the Two Extremes: Almgren and Chriss (2000) show how to effectively compute trajectories that lie between the two extremes.

## The Efficient Frontier of Optimal Execution

1. Computing the Optimal Execution Trajectories: Almgren and Chriss (2000) define and compute optimal execution trajectories and use that to later demonstrate a precise relationship between risk aversion and the definition of optimality.
2. Uniqueness of Optimal Execution Strategy: In particular, they show that each level of risk aversion there is a uniquely determined optimal execution strategy.



## The Definition of the Frontier

1. Minimization of Expected Short-fall: The rational trader will always seek to minimize the expectation of short-fall for a given level of variance of the short-fall. Naturally a trader will prefer a strategy that provides minimum error in its estimate of expected costs.
2. Efficient Optimal Trading Strategy Definition: Thus, a strategy is *efficient* or *optimal* if there is no other strategy that has lower variance for the same or a lower variance of the expected transaction costs, or, equivalently, no strategy which has no lower expected transaction costs for the same or lower level of variance.
3. Static vs. Dynamic Strategy Optimality: This definition of optimality of a strategy is the same whether the strategy is static or dynamic. It will be established later that under this definition and the price dynamics already stated, optimal strategies are in fact static.
4. Efficient Strategies - Constrained Optimization Formulation: One may construct efficient strategies by solving the constrained optimization problem

$$\min_{x: \mathbb{V}[x] \leq V_*} \mathbb{E}[x]$$

That is, for a given maximum level of variance

$$V_* \geq 0$$

one finds a strategy that has the minimum expected levels of transaction costs.

5. Convex Objective Function and Domain: Since  $\mathbb{V}[x]$  is convex, the set

$$\{\mathbb{V}[x] \leq V_*\}$$



is convex – it is a sphere – and since  $\mathbb{E}[x]$  is strictly convex, there is a unique minimizer  $x_*(V_*)$ .

6. Sub-Optimal Trajectory Variance Cost: Regardless of the preferred balance of risk and return, every other solution  $x$  which has

$$\mathbb{V}[x] \leq V_*$$

has higher expected costs than  $x_*(V_*)$  for the same or lower variance, and can never be more efficient.

7. Efficient Frontier of Optimal Strategies: Thus, the family of all possible efficient (optimal) strategies is parametrized by a single variable  $V_*$  representing all possible maximum levels of variance in transaction costs. This family is referred to as *the efficient frontier of optimal trading strategies*.
8. Introducing KKT Type Constraint Multipliers: The constrained optimization problem

$$\min_{x: \mathbb{V}[x] \leq V_*} \mathbb{E}[x]$$

is solved by introducing a constraint multiplier  $\lambda$ , thereby solving the unconstrained problem

$$\min_x (\mathbb{E}[x] + \lambda \mathbb{V}[x])$$

9. Frontier as a Function of  $\lambda$ : If

$$\lambda > 0$$

$\mathbb{E}[x] + \lambda \mathbb{V}[x]$  is strictly convex, and the above minimizer has a unique solution  $x^*(\lambda)$ . As  $\lambda$  varies,  $x^*(\lambda)$  sweeps out the same one parameter family, and thus traces out an efficient frontier.



10.  $\lambda$  as a Risk Aversion Parameter: The Parameter  $\lambda$  has a direct financial interpretation.

It is already apparent from

$$\min_x (\mathbb{E}[x] + \lambda \mathbb{V}[x])$$

that  $\lambda$  is a measure of risk aversion, that is, how much the variance is penalized relative to the cost.

11.  $\lambda$  as an Efficient Frontier Curvature: In fact,  $\lambda$  is the curvature – second derivative – of a smooth utility function, as will be made more precise eventually.

12. Solution given  $h(v)$  and  $g(v)$ : For given values of the parameters, problem

$$\min_x (\mathbb{E}[x] + \lambda \mathbb{V}[x])$$

can be solved by various numerical techniques depending on the functional forms chosen for  $h(v)$  and  $g(v)$ . In the special case that these are *linear* functions, we may write the solution explicitly and gain a great deal of insight into the trading strategies.

## Explicit Construction of Optimal Strategies

1. Optimal Solution in Trajectory Space: With  $\mathbb{E}[x]$  from

$$\mathbb{E}[X] = \frac{1}{2}\gamma X^2 + \epsilon \sum_{k=1}^N |n_k| + \frac{\tilde{\eta}}{\tau} \sum_{k=1}^N n_k^2$$

and  $\mathbb{V}[x]$  from



$$\mathbb{V}[X] = \sigma^2 \sum_{k=1}^N \tau x_k^2$$

and assuming that  $n_j$  does not change sign, the combination

$$\mathbb{U}[x] = \mathbb{E}[x] + \lambda \mathbb{V}[x]$$

is a quadratic function of the control parameters  $x_1, \dots, x_{N-1}$ ; it is strictly convex for

$$\lambda \geq 0$$

2. Finding the Unique Global Minima: Therefore, one determines the unique global minimum by setting its partial derivatives to zero. One readily calculates

$$\frac{\partial \mathbb{U}[x]}{\partial x_j} = 2\tau \left( \lambda \sigma^2 x_j - \tilde{\eta} \frac{x_{j-1} - 2x_j + x_{j+1}}{\tau^2} \right)$$

for

$$j = 1, \dots, N-1$$

3. Combinations of Linear Difference Equations: Then

$$\frac{\partial \mathbb{U}[x]}{\partial x_j} = 0$$

is equivalent to

$$\frac{x_{j-1} - 2x_j + x_{j+1}}{\tau^2} = \tilde{\kappa}^2 x_j$$



with

$$\tilde{\kappa}^2 = \frac{\lambda\sigma^2}{\tilde{\eta}} = \frac{\lambda\sigma^2}{\eta \left(1 - \frac{\gamma\tau}{2\eta}\right)}$$

4.  $\tau$  Abstracted and Re-factored Parameter Set: Note that

$$\frac{x_{j-1} - 2x_j + x_{j+1}}{\tau^2} = \tilde{\kappa}^2 x_j$$

is a linear difference equation whose solution may be written as a combination of the exponentials  $e^{\pm\kappa t_j}$  where  $\kappa$  satisfies

$$\frac{2}{\tau^2} [\cosh(\kappa\tau) - 1] = \tilde{\kappa}^2$$

The tilde's on  $\tilde{\eta}$  and  $\tilde{\kappa}$  denote an  $\mathcal{O}(\tau)$  correction; as

$$\tau \rightarrow 0$$

one has

$$\tilde{\eta} \rightarrow \eta$$

and

$$\tilde{\kappa} \rightarrow \kappa$$

5. Trading Trajectory/Trade List Solutions: The specific solution with



$$x_0 = X$$

and

$$x_N = 0$$

is a trading trajectory of the form

$$x_j = \frac{\sinh(\kappa(T - t_j))}{\sinh(\kappa T)} X$$

$$j = 0, \dots, N$$

and the associated trade list is

$$n_j = \frac{2 \sinh\left(\frac{1}{2} \kappa T\right)}{\sinh(\kappa T)} \cosh\left(\kappa\left(T - t_{j-\frac{1}{2}}\right)\right) X$$

$$j = 1, \dots, N$$

where sinh and cosh are the hyperbolic sine and cosine functions, and

$$t_{j-\frac{1}{2}} = \left(j - \frac{1}{2}\right) \tau$$

These solutions – although not the efficient frontier – have been constructed previously by Grinold and Kahn (1999).

6. Monotonicity of the Trading Trajectory: One has



$$n_j > 0$$

as long as

$$X > 0$$

Thus, for a program of selling a large initial long position, the solution decreases *monotonically* from its initial value to zero at the rate determined by the parameter  $\kappa$ .

7. Consequence of Monotonic Trading Trajectories: For example, the optimal execution of a sell program never involves buying of securities – although this ceases to be true if there is drift or serial correlation in price movements.
8. Approximation under Small Time Step: For a small time-step  $\tau$  one has the approximate expression

$$\kappa \sim \tilde{\kappa} + \mathcal{O}(\tau^2) \sim \sqrt{\frac{\lambda\sigma^2}{\eta \left(1 - \frac{\gamma\tau}{2\eta}\right)}} + \mathcal{O}(\tau)$$

$$\tau \rightarrow 0$$

Thus, if the trading intervals are short  $\kappa^2$  is essentially the ratio of the product of volatility and the risk-intolerance to the temporary transaction cost parameter.

9. Optimal Strategy Expected Cost/Variance: The expectation and the variance of the optimal strategy for a given initial portfolio size  $X$  are then

$$\mathbb{E}[X] = \frac{1}{2} \gamma X^2 + \epsilon X + \tilde{\eta} X^2 \frac{\tanh\left(\frac{1}{2}\kappa\tau\right) [\tau \sinh(2\kappa T) + 2T \sinh(\kappa\tau)]}{2\tau^2 [\sinh(\kappa\tau)]^2}$$

and



$$\mathbb{V}[X] = \frac{1}{2} \sigma^2 X^2 \frac{\tau \sinh(\kappa T) \cosh(\kappa(T-\tau)) - T \sinh(\kappa\tau)}{[\sinh(\kappa T)]^2 \sinh(\kappa\tau)}$$

which reduce to

$$\mathbb{E}[X] = \frac{1}{2} X T g\left(\frac{X}{T}\right) \left(1 - \frac{1}{N}\right) + X h\left(\frac{X}{T}\right) = \frac{1}{2} \gamma X^2 + \epsilon X + \tilde{\eta} \frac{X^2}{T}$$

$$\mathbb{V}[X] = \frac{1}{3} \sigma^2 X^2 T \left(1 - \frac{1}{N}\right) \left(1 - \frac{1}{2N}\right)$$

$$n_1 = X$$

$$n_2 = \dots = n_N = 0$$

$$x_1 = \dots = x_N = 0$$

$$\mathbb{E}[X] = X h\left(\frac{X}{\tau}\right) = \epsilon X + \eta \frac{X^2}{\tau}$$

$$\mathbb{V}[X] = 0$$

in the limits

$$\kappa \rightarrow 0, \infty$$

## The Half-Life of a Trade

1. Definition of the Half-Life: Defining



$$\theta = \frac{1}{\kappa}$$

the trade's "half-life", and using the discussion above, it can be seen that the larger the value of  $\kappa$  and smaller the value of  $\theta$ , the more rapidly the trade list will be depleted. The value  $\theta$  is exactly the amount of time it takes to deplete the holdings by a factor of  $e$ .

2. Half-Life Different from  $T$ : The definition of  $\theta$  is independent of the exogenously specified execution time  $T$ ; it is determined only by the security price dynamics and the market impact factors. If the risk aversion  $\lambda$  is greater than zero, i.e., if the trader is risk-averse, then  $\theta$  is finite and independent of  $T$ .
3. Timeless Initial Portfolio Liquidation Rate: Thus, in the absence of any external time constraint, i.e.

$$T \rightarrow \infty$$

the trader will still liquidate his position on a time scale  $\theta$ . The half-life  $\theta$  is the intrinsic time scale of the trade.

4. Half Life Smaller than  $T$ : For a given  $T$  the ratio

$$\kappa T = \frac{T}{\theta}$$

tells us what factors constrain the trade. If

$$T \gg \theta$$

then the intrinsic half-life  $\theta$  of the trade is small compared to the imposed time  $T$ ; this happens because temporary costs are very small, because volatility is very large, or because of high risk aversion.



5. Impact of Small Half-Life: In this case the bulk of the trading will be done well in advance of the time  $T$ . Viewed on a time scale  $T$  the trajectory will look like a minimum variance solution

$$n_1 = X$$

$$n_2 = \cdots = n_N = 0$$

$$x_1 = \cdots = x_N = 0$$

6. Very High Half Life Limit: Conversely if

$$T \ll \theta$$

then the trade is highly constrained, and is dominated by temporary market impact costs. In the limit

$$\frac{T}{\theta} \rightarrow 0$$

one approaches the straight-line minimum cost strategy

$$n_k = \frac{X}{N}$$

$$x_k = (N - k) \frac{X}{N}$$

$$k = 1, \dots, N$$



7. Trade Size Independent Execution Strategy: A consequence of this analysis is that different sized baskets of the same security will be liquidated in exactly the same fashion, on the same scale, provided the risk aversion parameter  $\lambda$  is held constant.
8. Basket Size Based Liquidity Dependence: This may seem contrary to the expectation that large baskets are effectively less liquid, and should hence be liquidated less rapidly than smaller baskets.
9. Reasons for the Counter-Intuitiveness: This is a consequence of the linear market impact assumption which has the *mathematical* consequence that both variance and market impact scale quadratically with respect to the portfolio size.
10. Higher Order Temporary Impact Function: For large portfolios it may be more reasonable to assume that the temporary impact cost function has higher-order terms, so that such costs increase *super-linearly* with the trade size. With non-linear impact functions, the general framework used here still applies, but one does not obtain explicit exponential solutions as in the linear impact case.
11. Size Dependent Temporary Impact Parameter: A simple practical solution to this problem is to choose different values of  $\eta$  - the temporary impact parameter – depending up on the overall problem size being considered, recognizing that the model is at best only approximate.

## Structure of the Frontier

1. Efficient Frontier and the Corresponding Trajectories: Using a specific choice for the parameters explained below, Almgren and Chriss (2000) produce a sample plot of the efficient frontier – each point on the frontier represents a distinct strategy for optimally liquidating the same basket. Their tangent line represents the optimal solution for a specified risk parameter

$$\lambda = 10^{-6}$$



They also illustrate the trajectories corresponding to a few sample points on the frontier.

2. Trajectory corresponding to Positive  $\lambda$ : Their first trajectory has

$$\lambda = 2 \times 10^{-6}$$

– this would be chosen by a risk-averse trader who wishes to sell quickly to reduce exposure to volatility risk, despite the trading costs incurred in doing so.

3. Trajectory corresponding to Zero  $\lambda$ : Their second trajectory has

$$\lambda = 0$$

They refer to this as the naïve strategy since this represents an optimal strategy corresponding to simply minimizing expected transaction costs without regard to variance.

4. Linear Reduction of the Holdings: For a security with zero drift and linear transaction costs as defined above

$$\lambda = 0$$

corresponds to a simple linear reduction of holdings over the trading period. Since drift is generally not significant over short trading horizons, the naïve strategy is very close to the linear strategy.

5. Sub Optimality of the Strategy: As Almgren and Chriss (2000) demonstrate later, in a certain sense this is *never* an optimal strategy because one can obtain substantial reductions in variance for a relatively small increase in transaction costs.
6. Trajectory corresponding to Negative  $\lambda$ : Finally, their trajectory  $C$  has

$$\lambda = -2 \times 10^{-6}$$



it would only be chosen by a trader who likes risk. He postpones execution, thus incurring higher costs both due to rapid sales at the end, and higher variance during the extended period that he holds the security for.

## The Utility Function

1. The Risk-Reward Trade-off: Almgren and Chriss (2000) offer an interpretation of the efficient frontier of optimal strategies in terms of the utility function of the seller. They do this in two ways – by direct analogy with modern portfolio theory employing a utility function, and by a novel approach: Value-at-risk. This eventually leads to some general observations regarding the importance of utility in forming execution strategies.
2. Utility of Risk-Averse Functions: Suppose one measures utility by a smooth convex function  $u(w)$  where  $w$  is the total wealth. This function may be characterized by its risk-aversion coefficient

$$\lambda_u = -\frac{u''(w)}{u'(w)}$$

3. Approximation in Estimating the  $\lambda$ : If the initial portfolio is fully owned, then as the transfer of assets happens from the risky stock into the alternative riskless investment,  $w$  remains roughly constant, and one may take  $\lambda_u$  to be a constant throughout the trading period. If the initial portfolio is highly leveraged, then the assumption of constant  $\lambda$  is an approximate one.
4. Formulation of the Optimal Execution Strategy: For short time horizons and small changes in  $w$  the higher derivatives of  $u(w)$  may be neglected. Thus, choosing an optimal execution strategy is equivalent to minimizing the scalar function

$$\mathbb{U}_{UTIL}[x] = \lambda_u \mathbb{V}[x] + \mathbb{E}[x]$$



The units of  $\lambda_u$  are  $\$/^{-1}$ ; one is willing to accept an extra square  $\$$  of variance if it reduces the expected cost by  $\$ \lambda_u$ .

5. Constructing Family of Optimal Paths: The combination  $\lambda \mathbb{V}[x] + \mathbb{E}[x]$  is precisely the one used to construct the efficient frontier seen earlier; the parameter  $\lambda$ , introduced as a Lagrange multiplier, has a precise definition as a measure of aversion to risk. Thus, the methodology above used to construct the efficient frontier likewise produces a family of optimal paths, one for each level of risk aversion.
6. Static Nature of Optimal Path: Returning now to an important point raised earlier, the computation of optimal strategies by minimizing  $\lambda \mathbb{V}[x] + \mathbb{E}[x]$  as measured at the initial trading time is equivalent to maximizing the utility at the outset of trading. As one trades, information arrives that could potentially alter the optimal path. The following theorem eliminates that possibility.
7. Time Homogenous Quadratic Utility Theorem: For a fixed quadratic utility function, the static strategies computed above are “time homogenous”. More precisely given a strategy that begins at a time

$$t = 0$$

and ends at a time

$$t = T$$

the optimal strategy computed at

$$t = t_k$$

is simply a continuation from

$$t = t_k$$



to

$$t = T$$

of the optimal strategy computed at time

$$t = 0$$

8. Proof Steps: General/Specific Functions: The proof may be seen in two ways – by the algebraic computations based on the specific solutions above, and by general valid for generic non-linear impact functions.
9. Proof Steps: Function Time Shift: First suppose that at time  $k$  where

$$k = 0, \dots, N - 1$$

If one were to compute a new optimal strategy. The new strategy would precisely be

$$x_j = \frac{\sinh(\kappa(T - t_j))}{\sinh(\kappa T)} X$$

$$j = 0, \dots, N$$

with  $X$  replaced by  $x_k$ ,  $T$  replaced by  $T - t_k$ , and  $t_j$  replaced by  $t_j - t_k$ . Using the subscript  $(k)$  to denote the strategy computed at time  $k$  one would have

$$x_j^{(k)} = \frac{\sinh(\kappa(T - t_j))}{\sinh(\kappa(T - t_k))} x_k$$

$$j = k, \dots, N$$



and the trade lists

$$n_j^{(k)} = \frac{2 \sinh\left(\frac{1}{2}\kappa\tau\right)}{\sinh(\kappa(T - t_k))} \cosh\left(\kappa\left(T - t_{j-\frac{1}{2}}\right)\right) X$$

$$j = k + 1, \dots, N$$

10. Proof Step: Recovering Optimal Solutions: It is then apparent that if  $x_k$  is the optimal solution from

$$x_j = \frac{\sinh\left(\kappa(T - t_j)\right)}{\sinh(\kappa T)} X$$

$$j = 0, \dots, N$$

with

$$j \mapsto k$$

then

$$x_j^{(k)} = x_j^0$$

and

$$n_j^{(k)} = n_j^0$$

where



$$x_j^0 = x_j$$

and

$$n_j^0 = n_j$$

are the strategies from

$$x_j = \frac{\sinh(\kappa(T - t_j))}{\sinh(\kappa T)} X$$

$$j = 0, \dots, N$$

and

$$n_j = \frac{2 \sinh\left(\frac{1}{2} \kappa T\right)}{\sinh(\kappa T)} \cosh\left(\kappa\left(T - t_{j-\frac{1}{2}}\right)\right) X$$

$$j = 1, \dots, N$$

11. Proof Step: Non-linear Impact: For general non-linear impact functions  $g(v)$  and  $h(v)$  the optimality condition

$$\frac{x_{j-1} - 2x_j + x_{j+1}}{\tau^2} = \tilde{\kappa}^2 x_j$$

is replaced by a second-order *non-linear* difference relation. The solution  $x_j^{(k)}$  beginning at a given time is determined by the two boundary values  $x_k$  and



$$x_N = 0$$

It is then apparent that the solution does not change if we re-evaluate it at later times.

12. Origin of Time Stable Solutions: More fundamentally, the solutions are time stable because in the absence of serial correlations in the asset price movements, there is no more information about the price changes at later times than there is at the initial time.
13. Optimality over each Sub-interval: Thus, the solution which was determined to be optimal over the entire time interval is optimal as a solution over each sub-interval. This general phenomenon is well known in the theory of optimal control (Bertsekas (1976)).

## Value at Risk

1. Motivation behind Value at Risk: The concept of value at risk is traditionally used to measure the greatest amount of money – maximum profit or loss - a portfolio will sustain over a given period of time under “normal circumstances”, where “normal” is defined by a confidence level.
2. Trading Value at Risk Definition: Given a trading strategy

$$x = (x_1, \dots, x_N)$$

the value-at-risk of  $x$  defined  $Var_p[x]$  is defined to be the level of transaction costs by the trading strategy  $x$  that will not be exceeded  $p$  percent of the time. Put another way, it is the  $p^{th}$  percentile level of transaction costs for total costs of trading  $x$ .

3. Trading Value at Risk Expression: Under the arithmetic Brownian motion assumption, the total costs – the market value minus capture – are normally distributed with known mean and variance. Thus, the confidence level is determined by the number of standard deviations  $\lambda_v$  from the mean of the inverse of the



cumulative normal distribution function, and the value at risk for the strategy  $x$  is given by

$$Var_p[x] = \lambda_v \sqrt{\mathbb{V}[x]} + \mathbb{E}[x]$$

4. Relation to Implementation Short-fall: That is, with a probability  $p$  the trading strategy will not lose more than  $Var_p[x]$  of its market value in trading. Borrowing from the language of Period (1988), the implementation shortfall of execution will not exceed  $Var_p[x]$  more than a fraction  $p$  of the time. A strategy  $x$  is efficient if it has the minimum possible value at risk for the confidence level  $p$ .
5. Execution Trajectory Optimized for VaR: Note that  $Var_p[x]$  is a complicated non-linear function of  $x_j$  composing  $x$ ; it can be easily evaluated for any given trajectory, but finding the minimizing trajectory directly is difficult.
6. Single Parameter Efficient Frontier Solution: But once the one-parameter family of solutions that form the efficient frontier is obtained, one only needs to solve a one-dimensional problem to find the optimal solutions for the value at risk model, that is, to fund the value of  $\lambda_u$  corresponding to a given value of  $\lambda_v$ . Alternatively, one may characterize the solutions by a simple graphical procedure, or may read off the confidence levels corresponding to any particular point on the curve.
7. Almgren-Chriss Optimal VaR Illustration: Almgren and Chriss (2000) produce an illustration of the above, using the square root of variance in the  $x$ -axis as opposed to the variance in itself. In this co-ordinate system lines of optimal VaR have a constant slope, and for a given value of  $\lambda_v$  they simply find a tangent to the curve where the slope is  $\lambda_v$ .
8. Interim Optimal Execution Re-evaluation: The question of re-evaluation of the strategy is more complicated and subtle. If one re-evaluates the strategy half-way through the execution process, they will choose a new optimal strategy that is not the same as the original optimal one. The reason is that since  $\lambda_v$  is now held constant,  $\lambda_u$  necessarily changes.



9. General Challenges with the VaR Approach: Value at risk has many flaws from a mathematical point of view, as recognized by Artzner, Delbaen, Eber, and Heath (1997). The particular issue encountered here would occur in any problem in which the time of measurement is a fixed date, rather than maintained at a fixed distance in the future. It is an open issue to formulate suitable measures of risk for general time-dependent problems.
10. Liquidity Adjusted Value at Risk: Despite this shortcoming, Almgren and Chriss (2000) propose the smallest possible value of  $Var_p[x]$  as an informative measure of the possible loss associated with the initial position, in the presence of liquidity effects. This value, which they call L-VaR for Liquidity Adjusted Value at Risk, depends on the time to liquidation and the confidence level chosen, in addition to the market parameters such as the impact coefficient (Almgren and Chriss (1999)).
11. Advantages of the L-VaR Approach: The optimal trajectories determined by minimizing the value at risk do *not* have the counter-intuitive scaling behavior seen earlier; even for linear impact functions, large portfolios will be traded closer to the straight-line trajectory.
12. Using L-VaR for Large Portfolios: This is because the cost assigned to uncertainty scales *linearly* with the portfolio size, while the temporary impact cost scales *quadratically* as before. Thus, the latter is more important for large portfolios.

## The Role of Utility in Execution

1. General Observations on Optimal Execution: Almgren and Chriss (2000) use the structure of the efficient frontier in the framework that they have developed to make some general observations concerning optimal executions.
2. The Naïve Strategy Benchmark: They first restrict themselves to the situation where the trader has no directional view on the security being traded. Recall that in this case, the naïve strategy is the simple straight-line strategy in which the trader breaks the blocks being executed into equal sized blocks to be sold over equal time intervals.



They use this strategy as a benchmark for comparison with the other strategies used throughout here.

3. Convex  $\mathbb{E}[x]$  to  $\mathbb{V}[x]$  Mapping: A crucial insight is that the curve defining the efficient frontier is a smooth convex function  $\mathbb{E}[\mathbb{V}]$  mapping the levels of variance  $\mathbb{V}$  to the corresponding minimum mean transaction cost levels.
4. Region around the Naïve Strategy: Write  $(\mathbb{E}_0, \mathbb{V}_0)$  for the mean and variance around the naïve strategy. Regarding  $(\mathbb{E}_0, \mathbb{V}_0)$  as a point on the smooth curve  $\mathbb{E}[\mathbb{V}]$  defined by the frontier,  $\frac{\partial \mathbb{E}}{\partial \mathbb{V}}$  evaluated at  $(\mathbb{E}_0, \mathbb{V}_0)$  is equal to zero. Thus for  $(\mathbb{E}, \mathbb{V})$  near  $(\mathbb{E}_0, \mathbb{V}_0)$  one has

$$\mathbb{E} - \mathbb{E}_0 = \frac{1}{2} (\mathbb{V} - \mathbb{V}_0)^2 \left. \frac{\partial^2 \mathbb{E}}{\partial \mathbb{V}^2} \right|_{\mathbb{V}=\mathbb{V}_0}$$

where

$$\left. \frac{\partial^2 \mathbb{E}}{\partial \mathbb{V}^2} \right|_{\mathbb{V}=\mathbb{V}_0}$$

is positive is positive by the convexity of the frontier at the naïve strategy.

5. Special Feature of the Naïve Strategy: By definition, the naïve strategy has the property that any strategy with lower variance in cost has a greater expected cost. However, a special feature of the naïve strategy is that a first-order decrease in variance can be obtained – in the sense of finding a strategy with a lower variance – while only incurring a second order increase in cost.
6. Disadvantages of Risk Neutral Strategy: From the above it follows that for small increases in variance, one can obtain much larger reductions in cost. Thus, unless the trader is risk-neutral it is always advantageous to execute a strategy that is at least to some degree “to the left” of the naïve strategy. Thus, one concludes that, in this framework, from a theoretical standpoint, it never makes sense to trade a strictly risk-neutral strategy.



7. The Role of a Security's Liquidity: An intuitive proposition is that with all things being equal, a trader will execute a more liquid basket more rapidly than a less liquid one. In the extreme this is particularly clear. A broker given a small order to execute over the course of the day will execute the entire order almost immediately.
8. Executing the Highly Liquid Security: How does one explain this? The answer is that the market impact cost attributable to rapid trading is negligible compared with the opportunity cost incurred in breaking up the order over an entire day. Thus, even if the expected return on a security over the day is zero, the perception is that the risk of waiting is outweighed by any small cost of immediacy.
9. Absence of Risk Reduction Premium: Now if the trader were truly risk neutral, in the absence of any views, he would always use the naïve strategy and employ the allotted time fully. This would make sense because any price to pay for trading immediately is worthless if one places no premium on risk reduction.
10. Limitation of Risk Neutral Approach: It follows that any model that proposes optimal trading behavior should predict that more liquid baskets are traded more rapidly than less liquid ones. A model that only considers the minimization of transaction costs, like that of Bertsimas and Lo (1998), is essentially a model that excludes utility.
11. Optimal Execution Independent of Liquidity: In such a model, and under Almgren and Chriss (2000) basic assumptions, traders will trade all baskets at the same rate irrespective of the liquidity, that is unless they have an explicit directional view on the security, or the security possesses extreme serial correlation in its price movements.
12. Super Linear Market Impact Functions: Almgren and Chriss (2000) do note that their model in the case of linear transaction costs does not predict a more rapid trading for smaller versus larger baskets of the same security. However, this is a consequence of choosing linear temporary impact functions and the problem goes away when one considers more realistic super-linear functions.
13. Risk Neutral Execution Half Life: Another way of looking at this is that the half-life of all black executions, under the assumption of risk-neutral preferences, is infinite.



## Choice of Parameters

1. The Asset Intrinsic Dynamics Parameters: Almgren and Chriss (2000) compute some numerical examples for the purposes of exploring the qualitative properties of the efficient frontier. Throughout the examples they consider a single stock with the current market price of

$$S_0 = 50$$

and that they initially have one million shares, for an initial portfolio size of \$50 million. The stock will have 30% annual volatility, 10% expected annual rate of return, a bid-ask spread of  $\frac{1}{8}$ , and a median daily trading volume of 5 million shares.

2. Stock Asset Daily Return/Volatility: With a trading year of 250 days this gives a daily volatility of

$$\frac{0.3}{250} = 0.019$$

and expected fractional return of

$$\frac{0.1}{250} = 4 \times 10^{-4}$$

To obtain our absolute parameters  $\sigma$  and  $\alpha$  one must scale it by the price, so

$$\sigma = 0.019 \times 50 = 0.95$$

and



$$\alpha = (4 \times 10^{-4}) \times 50 = 0.02$$

The table below summarizes the information.

3. Parameter Values for the Test Case:

Parameter Description	Parameter Symbol	Parameter Value
Initial Stock Price	$S_0$	\$50/share
Initial Holdings	$X$	$10^6$ shares
Liquidation Time	$T$	5 days
Number of Time Periods	$N$	5
30% Annual Volatility	$\sigma$	$0.95 (\$/share)/day^{\frac{1}{2}}$
10% Annual Growth	$\alpha$	$0.02 (\$/share)/day$
Bid Ask Spread $\frac{1}{8}$	$\epsilon$	\$0.0625/share
Daily Volume 5 million shares	$\gamma$	$2.5 \times 10^{-7} \$/share^2$
Impact at 1% of market	$\eta$	$2.5 \times 10^{-6} (\$/share)/(share/day)$
Static Holdings 11,000 shares	$\lambda_u$	$10^{-6}/\$$
VaR Confidence $p = 95\%$	$\lambda_v$	1.645

4. Incremental and Total Execution Times: Suppose that one wants to liquidate this position in one week so that

$$T = 5 \text{ days}$$

This is divided into daily trades such that  $\tau$  is 1 day and

$$N = 5$$



5. Standard Deviation of the Trajectory: Over this period, if one holds the original position with no trading, the fluctuations in the stock value will be Gaussian with a standard deviation of

$$\sigma\sqrt{T} = 2.12 \text{ (\$/share)}$$

and the fluctuations in this value will have an absolute standard deviation of

$$\sqrt{V} = \$2.12M$$

As expected, this is precisely the value of  $\sqrt{V}$  for the lowest point in the efficient frontier, since that point corresponds selling along a linear trajectory rather than holding a constant amount.

6. Temporary Cost Function Parameter -  $\epsilon$ : One then chooses the parameters for the temporary cost function

$$h\left(\frac{n_k}{\tau}\right) = \epsilon sgn(n_k) + \frac{\eta}{\tau} n_k$$

Almgren and Chriss (2000) set

$$\epsilon = \frac{1}{16}$$

that is, the fixed part of the temporary costs will be one-half the bid-ask spread.

7. Temporary Cost Function Parameter -  $\eta$ : For  $\eta$  they suppose that for each 1% of the daily volume traded they incur a price impact equal to one bid-ask spread. For example, trading at a rate of 5% daily volume incurs a one-time cost on each trade of  $\frac{5}{8}$ . Under this assumption



$$\eta = \frac{\frac{1}{8}}{0.01 \times 5 \times 10^6} = 2.5 \times 10^{-6}$$

8. Permanent Cost Function Parameter -  $\gamma$ : For permanent costs, the common rule of thumb is that price effects become significant when 10% of the daily volume is sold. Assuming that “significant” means that the price depression is one bid-ask spread, and that the effect is linear for both smaller and larger trading rates, one has

$$\gamma = \frac{\frac{1}{8}}{0.1 \times 5 \times 10^6} = 2.5 \times 10^{-7}$$

Recall that this parameter gives a fixed cost independent of the path.

9. The Risk Aversion Parameter -  $\lambda$ : Almgren and Chriss (2000) have chosen

$$\lambda = \lambda_u = 10^{-6}$$

For these parameters, from

$$\kappa \sim \tilde{\kappa} + \mathcal{O}(\tau^2) \sim \sqrt{\frac{\lambda \sigma^2}{\eta \left(1 - \frac{\gamma \tau}{2\eta}\right)}} + \mathcal{O}(\tau)$$

$$\tau \rightarrow 0$$

one has for the optimal strategy that

$$\kappa \approx 0.61 \text{ day}$$

so that



$$\kappa T \approx 3$$

Since this value is near 1 in magnitude, the behavior is an interesting intermediate in-between the naïve extremes.

10.  $\lambda_v$  at 95% Confidence Level: For the value at risk representation, as assumed 95% confidence level gives

$$\lambda_v = 1.645$$

## The Value of Information

1. Zero Drift Random Walk Assumption: The discussion carried out so far assumed that the price dynamics followed an arithmetic random walk with zero drift. Since past price paths provide no extra information on future price movements, the conclusion was that the optimal trajectories can be statically determined. There are three ways by which a random walk with zero drift may fail to represent the price process.
2. Non-zero Drift in Dynamics: First the price process may have drift. For example, if the trader has a strong directional view, the trader may want to incorporate this view into the liquidation strategy.
3. Cross Period Serial Correlation Impact: Second, the price process may exhibit serial correlation. The presence of first order serial correlation for example, implies that the price moves in a given period provide non-trivial information concerning the next period movement of the asset.
4. Incorporation of the Investor's Private Information: Bertsimas and Lo (1998) study a general form of this assumption, wherein an investor possesses possibly private information of a serially correlated information vector that acts as a linear factor in the asset returns.



5. Exogenously Induced Material Parameter Shift: Lastly, at the start of trading, it may be known that at some specific point in time, an event will take place whose outcome will cause a material shift in the parameters governing the price process.
6. Literature Survey on Exogenous Events: Such event induced parameter shifts include quarterly and annual earnings announcements, dividend announcements, and share repurchases. Event studies documenting these parameter shifts and providing theoretical grounding for their existence include Beaver (1968), Fama, Fisher, Jensen, and Roll (1969), Dann (1981), Patell, and Wolfson (1984), Kalay and Loewenstein (1985), Kim and Verrecchia (1991), Campbell, Lo, and MacKinlay (1997), Easterwood and Nutt (1999), and Ramaswami (1999).
7. Temporary Shifts on Dynamic Parameters: For example, Brown, Harlow, and Tinic (1988) show that events cause temporary shifts in both the risk and returns of individual securities, and the extent of these shifts depends on the outcome of the event. In general, securities react more strongly to bad news than good news.
8. Probabilistic Event Outcomes/Parameter Shifts: Almgren and Chriss (2000) study a stylized version of the events in which a known event at a known time – e.g., an earnings announcement – has several possible outcomes. The probability of each outcome is known, and the impact that a given outcome will have on the parameters of the price is also known. Clearly, optimal strategies must explicitly use this information, and Almgren and Chriss (2000) develop methods to incorporate event-specific information into their risk-reward framework.
9. Back-to-Back Static Strategies: The upshot is a piece-wise strategy that trades statically up to the event, and then reacts explicitly to the outcome of the event. Thus, the burden is on the trader to determine which of the possible outcomes occurred and then trade accordingly.

## Drift



1. Drift as a Directional View: It is convenient to regard the drift parameter in the price process as a directional view of price movements. For example, the trader charged with liquidating a single security may believe that this security is likely to rise. Intuitively it makes more sense to trade this issue more slowly to take advantage of this view.
2. Incorporating Drift into Price Dynamics: To incorporate drift into the price dynamics Almgren and Chriss (2000) modify

$$S_k = S_{k-1} + \sigma\sqrt{\tau}\xi_k - \gamma g\left(\frac{n_k}{\tau}\right)$$

to

$$S_k = S_{k-1} + \sigma\sqrt{\tau}\xi_k + \alpha\tau - \gamma g\left(\frac{n_k}{\tau}\right)$$

where  $\alpha$  is an expected drift term. If the trading proceeds are invested in an interest-bearing account, then  $\alpha$  should be taken as the *excess* rate of return of the risky asset.

3. Price Expectation over Time Period: One can readily write the modified version of

$$\mathbb{E}[X] = \frac{1}{2}\gamma X^2 + \epsilon \sum_{k=1}^N |n_k| + \frac{\tilde{\eta}}{\tau} \sum_{k=1}^N n_k^2$$

as

$$\mathbb{E}[X] = \frac{1}{2}\gamma X^2 - \alpha \sum_{k=1}^N \tau x_k + \epsilon \sum_{k=1}^N |n_k| + \frac{\tilde{\eta}}{\tau} \sum_{k=1}^N n_k^2$$

4. Updated Objective Function Optimality Condition: The variance is still given by



$$\mathbb{V}[X] = \sigma^2 \sum_{k=1}^N \tau x_k^2$$

The optimality condition

$$\frac{x_{j-1} - 2x_j + x_{j+1}}{\tau^2} = \tilde{\kappa}^2 x_j$$

becomes

$$\frac{x_{j-1} - 2x_j + x_{j+1}}{\tau^2} = \tilde{\kappa}^2 (x_j - \bar{x})$$

in which the new parameter

$$\bar{x} = \frac{\alpha}{2\lambda\sigma^2}$$

is the optimal level of security holding for a time independent portfolio optimization problem.

5. Drift Based Updated Execution Slice: For example, the parameters used in the example above give approximately

$$\bar{x} = 1,100 \text{ shares}$$

or 0.11% of our initial portfolio. One expects this fraction to be very small, since, by hypothesis, the eventual aim is complete liquidation.

6. Drift Based Updated Optimal Solution: The optimal solution

$$x_j = \frac{\sinh(\kappa(T - t_j))}{\sinh(\kappa T)} X$$



$$j = 0, \dots, N$$

becomes

$$x_j = \frac{\sinh(\kappa(T - t_j))}{\sinh(\kappa T)} X + \left\{ 1 - \frac{\sinh(\kappa(T - t_j)) + \sinh(\kappa t_j)}{\sinh(\kappa T)} \right\} \bar{x}$$

for

$$j = 0, \dots, N$$

with the associated trades

$$\begin{aligned} n_j &= \frac{2 \sinh\left(\frac{1}{2}\kappa\tau\right)}{\sinh(\kappa T)} \cosh\left(\kappa\left(T - t_{j-\frac{1}{2}}\right)\right) X \\ &\quad + \frac{2 \sinh\left(\frac{1}{2}\kappa\tau\right)}{\sinh(\kappa T)} \left[ \cosh\left(\kappa t_{j-\frac{1}{2}}\right) - \cosh\left(\kappa\left(T - t_{j-\frac{1}{2}}\right)\right) \right] \bar{x} \end{aligned}$$

7. Initial Position Independent Trajectory Correction: This trading trajectory is a sum of two distinct trajectories – the zero-drift solution as computed before, plus a “correction” which profits by capturing a piece of the predictable drift component. The size of this correction term is proportional to  $\bar{x}$ , and thus to  $\alpha$ : it is independent of the initial portfolio size  $X$ .
8. Practical Incorporation into Program Trading: To place this in an institutional framework, consider a program trading desk that sits in front of customer flow. If this desk were to explicitly generate alphas on all securities that flow through the desk in an attempt to, say, hold securities with high alphas and sell securities with low alphas more rapidly, the profit would not scale in proportion to the average size of the



programs. Rather it would only scale with the number of securities that flow through the desk. An even stronger conclusion is that since the optimal strategy disconnects into a static strategy unrelated to the drift term, and a second strategy related to the drift term, there is no particular advantage to restricting trading in securities which the desk currently holds the positions in.

9. Comparison: Highly Liquid Markets Scenario: The difference between this solution and the no-drift solution in

$$x_j = \frac{\sinh(\kappa(T - t_j))}{\sinh(\kappa T)} X$$

$$j = 0, \dots, N$$

may be understood by considering the case

$$\kappa T \gg 1$$

corresponding to highly liquid markets. Whereas the previous one relaxed from  $X$  to  $\frac{X}{e}$  in a time scale of

$$\theta = \frac{1}{\kappa}$$

this one relaxes instead to the optimal static portfolio size  $\bar{x}$ . Near the end of the trading period the trader sells the remaining holdings to achieve

$$x_N = 0$$

at



$$t = T$$

10. Caveat: Buy-Sell Symmetry Breaking: In this case, one requires

$$0 \leq \bar{x} \leq X$$

in order for all trades to be in the same direction. This breaks the symmetry between a buy program and a sell program, if one wanted to consider buy programs it would be more logical to set

$$\alpha = 0$$

## Gain due to Drift

1. Gain from Drift – Calculation Motivation: Now suppose that the price dynamics is given by

$$S_k = S_{k-1} + \sigma\sqrt{\tau}\xi_k + \alpha\tau - \gamma g\left(\frac{n_k}{\tau}\right)$$

with

$$\alpha > 0$$

but one chooses to determine the solution as though

$$\alpha = 0$$



The situation may arise, for example, in case where the trader is trading a security with non-zero drift, but *unknowingly* assumes that the security has no drift. Almgren and Chriss (2000) explicitly calculate the loss associated with ignoring the drift term.

2. Gain adjusted  $\mathbb{E}[x]$  and  $\mathbb{V}[x]$ : Write  $x_j^*$  for the optimal solution

$$x_j = \frac{\sinh(\kappa(T - t_j))}{\sinh(\kappa T)} X + \left\{ 1 - \frac{\sinh(\kappa(T - t_j)) + \sinh(\kappa t_j)}{\sinh(\kappa T)} \right\} \bar{x}$$

with

$$\alpha > 0$$

$x_j^0$  for the sub-optimal solution

$$x_j = \frac{\sinh(\kappa(T - t_j))}{\sinh(\kappa T)} X$$

$$j = 0, \dots, N$$

or

$$x_j = \frac{\sinh(\kappa(T - t_j))}{\sinh(\kappa T)} X + \left\{ 1 - \frac{\sinh(\kappa(T - t_j)) + \sinh(\kappa t_j)}{\sinh(\kappa T)} \right\} \bar{x}$$

with

$$\alpha = 0$$



Also write  $\mathbb{E}^*[x]$  and  $\mathbb{V}^*[x]$  for the optimal expected cost and its variance measured by

$$\mathbb{E}[X] = \frac{1}{2}\gamma X^2 - \alpha \sum_{k=1}^N \tau x_k + \epsilon \sum_{k=1}^N |n_k| + \frac{\tilde{\eta}}{\tau} \sum_{k=1}^N n_k^2$$

and

$$\mathbb{V}[X] = \sigma^2 \sum_{k=1}^N \tau x_k^2$$

with

$$x_j = x_j^*$$

and write  $\mathbb{E}^0[x]$  and  $\mathbb{V}^0[x]$  for the sub-optimal values of

$$\mathbb{E}[X] = \frac{1}{2}\gamma X^2 - \alpha \sum_{k=1}^N \tau x_k + \epsilon \sum_{k=1}^N |n_k| + \frac{\tilde{\eta}}{\tau} \sum_{k=1}^N n_k^2$$

and

$$\mathbb{V}[X] = \sigma^2 \sum_{k=1}^N \tau x_k^2$$

evaluated with

$$x_j = x_j^0$$



3. Objective Function Gain from Drift: The corresponding objective functions are

$$\mathbb{U}^*[X] = \mathbb{E}^*[X] + \lambda \mathbb{V}^*[X]$$

and

$$\mathbb{U}^0[X] = \mathbb{E}^0[X] + \lambda \mathbb{V}^0[X]$$

One can then define the *gain due to drift* to be the difference  $\mathbb{U}^0[X] - \mathbb{U}^*[X]$ ; this is the reduction in the cost and the variance by being aware of and taking into account of the drift term. Clearly

$$\mathbb{U}^0[X] - \mathbb{U}^*[X] \geq 0$$

since  $x^*$  is the unique optimal strategy for the model with

$$\alpha > 0$$

4. Upper Bound for the Gain: Now the value of the terms in  $\mathbb{U}^0[X]$  that come from

$$\mathbb{E}[X] = \frac{1}{2} \gamma X^2 - \alpha \sum_{k=1}^N \tau x_k + \epsilon \sum_{k=1}^N |n_k| + \frac{\tilde{\eta}}{\tau} \sum_{k=1}^N {n_k}^2$$

and

$$\mathbb{V}[X] = \sigma^2 \sum_{k=1}^N \tau {x_k}^2$$



is only *increased* by going from  $x^0$  to  $x^*$  since  $x^0$  and not  $x^*$  was the optimum strategy with

$$\alpha = 0$$

Therefore, an *upper bound* for the gain is

$$\mathbb{U}^0[X] - \mathbb{U}^*[X] \geq \alpha\tau \sum_{k=1}^N (x_k^* - x_k^0)$$

5. Adjustment Applied to the Holdings: That is, in response to positive drift, one should increase the holdings throughout the trading. This reduces the net cost by the amount of the increase in the asset price one captures, at the expense of slightly increasing the transaction costs and the volatility exposure. An upper bound for the possible benefit is the amount of increase one captures.
6. Explicit Expression for the Bound: But  $x_k^* - x_k^0$  is just the term in the square brackets in

$$x_j = \frac{\sinh(\kappa(T - t_j))}{\sinh(\kappa T)} X + \left\{ 1 - \frac{\sinh(\kappa(T - t_j)) + \sinh(\kappa t_j)}{\sinh(\kappa T)} \right\} \bar{x}$$

times  $\bar{x}$ , which is clearly independent of  $X$ . Indeed, this can be explicitly evaluated to get

$$\alpha\tau \sum_{k=1}^N (x_k^* - x_k^0) = \alpha\bar{x}T \left[ 1 - \frac{\tau}{T} \frac{\tanh\left(\frac{1}{2}\kappa T\right)}{\tanh\left(\frac{1}{2}\kappa\tau\right)} \right]$$



7. Gain Comparison against Execution Cost: Since  $\frac{\tanh x}{x}$  is a positive decreasingly function, this quantity is positive and bounded above by  $\alpha \bar{x}T$ , the amount one would gain by holding  $\bar{x}$  for a time  $T$ . Any reasonable estimates for the parameters show that this quantity is negligible compared to the impact costs incurred in liquidating an institutional sized portfolio over a short period.

## Serial Correlation

1. Prior Period Price Increment Component: Now one supposes that the asset prices exhibit serial correlation, so that at each period one discovers a component of predictability of the asset price in the next period.
2. Methodology behind the Price Increment Estimation: In the model

$$S_k = S_{k-1} + \sigma\sqrt{\tau}\xi_k - \gamma g\left(\frac{n_k}{\tau}\right)$$

with a drift

$$\alpha = 0$$

one now supposes that the  $\xi_k$  are serially correlated with period-to-period correlation  $\rho$

$$|\rho| < 1$$

One can determine  $\xi_k$  at time  $k$  based on the obtained  $S_k - S_{k-1}$  and sale  $n_k$

3. Optimal Strategy no more Static: With serial correlation the optimal trajectory is no longer a static trajectory determined in advance of trading; since each price



movement gives some information about the immediate future price movements, the optimal trade list can be determined only one period at a time.

4. Estimation of the Realized Gain: Thus, a full optimal solution requires the use of dynamic programming methods. However ,since the information is still roughly local in time, one can estimate the optimal gain attainable by an optimal strategy.
5. Almgren and Chriss (2000) Conclusions: Almgren and Chriss (2000) state their conclusion in advance of their estimation. The value of information contained in pure movements due to serial correlations is independent of the size of the portfolios being traded. The calculation demonstrated below lends intuition to this counter-intuitive statement.
6. Per Period Price Change Impact: Consider two consecutive periods during which the base strategy has the trader trading the same number of shares  $n$  in each period. With a linear impact price model, in each period price changes by  $\left[\epsilon + \eta \frac{n}{\tau}\right]$  dollars/share. The trader pays this cost in each of the  $n$  shares, so the total cost due to market impact per period is  $\left[\epsilon + \eta \frac{n}{\tau}\right] n$
7. Price Change from Serial Correlation: Suppose on has some price information due to serial correlations. If one knows  $\xi_k$  at the previous period, then the predictable component of the price change is roughly  $\rho\sigma\sqrt{\tau}\Delta n$ .
8. Incremental Cost of the Adapted Strategy: But this adaptation increases the impact costs. After the shift in the first period the price change is  $\epsilon + \eta \frac{n-\Delta n}{\tau}$  while in the second period  $\epsilon + \eta \frac{n+\Delta n}{\tau}$ . These costs are paid on  $n - \Delta n$  and  $n + \Delta n$  shares respectively, so the market impact per period is now

$$\left[ \frac{1}{2} \left( \epsilon + \eta \frac{n - \Delta n}{\tau} \right) (n - \Delta n) + \frac{1}{2} \left( \epsilon + \eta \frac{n + \Delta n}{\tau} \right) (n + \Delta n) \right] = \left[ \epsilon + \eta \frac{n}{\tau} \right] n + \frac{n}{\tau} \Delta n^2$$

9. Optimal Per-Period Execution Shift: To determine has many shares one should shift, one solves the quadratic optimization problem



$$\max_{\Delta n} \left[ \rho \sigma \sqrt{\tau} \Delta n - \frac{n}{\tau} \Delta n^2 \right]$$

The optimal  $\Delta n$  is readily found as

$$\Delta n^* = \frac{\rho \sigma \tau^{\frac{3}{2}}}{2\eta}$$

and the maximum possible gain per period is  $\frac{\rho^2 \sigma^2 \tau^2}{4\eta}$ . This heuristic can be confirmed by a detailed dynamic programming computation that accounts for optimal shifts across multiple periods.

10. Limitation of Optimal Gain Execution: Almgren and Chriss (2000) also explain briefly the limitation of the above approximation. When  $\rho$  is close to zero, clearly this approximation is extremely close to accurate, because the persistence of the serial correlation effect dies down very quickly after the first period. When  $|\rho|$  is too large to ignore, the approximation is too small for

$$\rho > 0$$

That is,  $\frac{\rho^2 \sigma^2 \tau^2}{4\eta}$  understates the possible gains over ignoring serial correlation.

Conversely when

$$\rho < 0$$

$\frac{\rho^2 \sigma^2 \tau^2}{4\eta}$  overstates the possible gains due to serial correlation. As

$$\rho > 0$$



is more frequently the case Almgren and Chriss (2000) assert that  $\frac{\rho^2 \sigma^2 \tau^2}{4\eta}$  is useful for bounding the possible gains in most situations available from serial correlations.

11. Position Independence of Gain/Cost: Note that both the size of the adaptation, and the resulting gain, are independent of the amount of shares  $n$  that would be sold under an unadapted strategy. That is they are also independent of the size of the initial portfolio.
12. Gain/Cost Liquidity/Correlation Dependence: Instead, the binding constraint is the liquidity of the security being traded, and the magnitude of the correlation coefficient. The more information available due to correlation and the more liquid the security, the more overall gain that is available due to adapting the strategy to the correlations.
13. Higher Order Impact Function Optimality: The results above are especially simple because of the assumption of linear impact functions. Almgren and Chriss (2000) also show briefly what happens in the more general case of nonlinear market impact functions

$$h(v) = h\left(\frac{n}{\tau}\right)$$

The cost per period due to market impact is

$$\begin{aligned} & \left[ \frac{1}{2} h\left(\frac{n - \Delta n}{\tau}\right)(n - \Delta n) + \frac{1}{2} h\left(\frac{n + \Delta n}{\tau}\right)(n + \Delta n) \right] \\ & \approx h\left(\frac{n}{\tau}\right)n + \left[ \frac{1}{2} h''\left(\frac{n}{\tau}\right)\frac{n}{\tau} + h'\left(\frac{n}{\tau}\right) \right] \frac{\Delta n^2}{\tau} \end{aligned}$$

for small  $\Delta n$ . Now the optimal shift and the maximal gain are given by

$$\Delta n^* = \frac{\frac{3}{\rho \sigma \tau^2}}{v h'' + 2 h'}$$



and  $\frac{\rho^2 \sigma^2 \tau^2}{2(\nu h'' + 2h')}$  respectively, where  $h'$  and  $h''$  are evaluated at the base execution rate of

$$\nu = \frac{n}{\tau}$$

The linear case is recovered by setting

$$h(\nu) = \epsilon + \eta\nu$$

This has the special property that  $h'$  is independent of  $\nu$  and

$$h'' = 0$$

14. Optimality Dependence on Impact Exponent: In general, suppose

$$h(\nu) \sim \mathcal{O}(\nu^\omega)$$

as

$$\nu \rightarrow \infty$$

$$\omega > 0$$

is required so that  $h(\nu)$  is increasing; selling the share always pushes the price down more. The marginal cost is

$$h'(\nu) \sim \mathcal{O}(\nu^{\omega-1})$$

$$\omega > 1$$



corresponds to an increasing marginal impact, and

$$\varpi < 1$$

corresponds to a decreasing marginal impact. Then the per-period cost one pays on the base strategy is

$$\sim \mathcal{O}(\nu^{\varpi+1})$$

for large initial portfolios, and hence large rates of execution. The marginal gain from adapting to evolution is

$$\sim \mathcal{O}(\nu^{\varpi-1})$$

in the same limit.

## Parameter Shifts

1. Price Dynamics Parameter Set Shift: Almgren and Chriss (2000) discuss the impact on optimal execution of scheduled news earnings such as earnings and dividend announcements. Such events have two features that make them an important object of study. First the outcome of the event determines the shift in the parameters governing the price dynamics – see Brown, Harlow, and Tinic (1988), Easterwood and Nutt (1999), and Ramaswami (1999).
2. Determining an Event' Full Impact: Second, the fact that they are scheduled increases the likelihood that one can detect what the true outcome of the event is. This situation is formalized below, and explicit formulas are given for price trajectories before and after the event takes place.



3. Scheduled Event Occurrence Time  $T_*$ : Suppose at some time  $T_*$  between now and the specified final time  $T$  an event will occur, the outcome of which may or may not cause a shift in the parameters of price dynamics.
4. New Regime Shifted Parameter Set: The term *regime set* or *parameter set* refers to the collection

$$R = \{\sigma, \eta, \dots\}$$

of the parameters that govern the dynamics at any particular time, and the events of interest are those that have the possibility of causing *parameter shifts*.

5. Initial to Final Regime Shift: Let

$$R_0 = \{\sigma_0, \eta_0, \dots\}$$

be the parameters of price dynamics at the time the execution begins. Suppose the market can shift to one of possible new sets of parameters  $p$  so that  $R_1, \dots, R_p$  is characterized by parameters  $\sigma_j, \eta_j, \dots$  for

$$j = 1, \dots, p$$

6. Probability of a Regime Switch: One also supposes that probabilities can be assigned to these possible new states, so that  $p_j$  is the probability that regime  $R_j$  occurs. The probabilities are *independent* of the short-term market fluctuations represented by  $\xi_k$ . Of course, it is possible that some  $R_j$  has the same values as  $R_0$  in which case  $p_j$  is the probability that no change occurs.
7. Globally Optimal Dynamic Trading Strategy: Almgren and Chriss (2000) consider a dynamic trading strategy the yields globally optimal strategies in the presence of a parameter shift at time  $T_*$ . Taking

$$T_* = t_s = s\tau$$



one precomputes an initial trajectory

$$x^0 = \{x_0^0, \dots, x_s^0\}$$

with

$$x_0^0 = X$$

Denote

$$X_* = x_s^0$$

8. Landscape of Switchable Trajectories: They also compute a family of trajectories

$$x^j = \{x_0^j, \dots, x_s^j\}$$

for

$$j = 1, \dots, p$$

all of which have

$$x_s^j = X_*$$

and

$$x_N^j = 0$$



They follow the trajectory  $x^0$  until the time of the shift. Once the shift occurs they assume they can quickly identify the outcome of the event and the new set of parameters governing the price dynamics. With this settled, the complete trading using the corresponding trajectory  $x^j$  is determined.

9. Key Almgren and Chriss (2000) Results: Almgren and Chriss (2000) show that it is possible to determine each trajectory using static optimization; although one cannot choose which one to use until the event occurs. Also, the starting trajectory  $x^0$  will *not be the same* as the trajectory one would use if they believed the regime  $R_0$  would hold through the entire time  $T$ .
10. Trajectory Conditional on Fixed  $X_*$ : To determine the trajectories  $x^0, x^1, \dots, x^p$  they reason as follows. Suppose that the common value of

$$X_* = x_s^0 = x_s^j$$

is fixed. Then by virtue of the independence of the regime shift in itself from the security motions, the optimal trajectories conditional on the values of  $X_*$  are simply those that have already been computed with a small modification to include the given non-zero final value.

11. Sequential Pair of Static Strategies: One can immediately write

$$x_k^0 = \frac{\sinh(\kappa_0(T_* - t_k))}{\sinh(\kappa_0 T_*)} X + \frac{\sinh(\kappa_0 t_k)}{\sinh(\kappa_0 T_*)} X_*$$

$$k = 0, \dots, s$$

where  $\kappa_0$  is determined from  $\sigma_0, \eta_0, \dots$ . The trajectory is determined the same way as seen before; it is the unique combination of the exponentials  $x^{\pm\kappa_0 t}$  that has

$$x_0^0 = X$$



and

$$x_s^0 = X_*$$

Similarly

$$x_k^j = \frac{\sinh(\kappa_j(T - t_k))}{\sinh(\kappa_j(T - T_*))} X_*$$

$$k = s, \dots, N$$

$$j = 1, \dots, p$$

Thus, one only needs to determine  $X_*$ .

12. Principle behind the Estimation of  $X_*$ : To determine  $X_*$  one needs to determine the expected loss and the variance of the combined strategy. Let  $\mathbb{E}_0$  and  $\mathbb{V}_0$  denote the expectation and the loss incurred by the trajectory  $x^0$  on the first segment

$$k = 0, \dots, s$$

The quantities can be determined readily using

$$\mathbb{E}[X] = \frac{1}{2}\gamma X^2 + \epsilon \sum_{k=1}^N |n_k| + \frac{\tilde{\eta}}{\tau} \sum_{k=1}^N n_k^2$$

and

$$\mathbb{V}[X] = \sigma^2 \sum_{k=1}^N \tau n_k^2$$



13. Mean Variance of the Compound Strategy: Then by virtue of the regime shift and the security motion's independence, the expected loss of the compound strategy is

$$\mathbb{E} = \mathbb{E}_0 + \mathbb{P}_1 \mathbb{E}_1 + \cdots + \mathbb{P}_p \mathbb{E}_p$$

and its variance is

$$\mathbb{V} = \mathbb{V}_0 + \mathbb{P}_1 \mathbb{V}_1 + \cdots + \mathbb{P}_p \mathbb{V}_p + \frac{1}{2} \sum_{i,j=1}^{p_0} \mathbb{P}_i \mathbb{P}_j (\mathbb{E}_i - \mathbb{E}_j)^2$$

One can now do a one-variable optimization in  $X_*$  to maximize  $\mathbb{E} + \lambda \mathbb{V}$ . Almgren and Chriss (2000) provide a pictorial representation of the above.

## Conclusions and Further Extensions

1. Efficient Frontier of Transaction Costs: The central feature of the Almgren and Chriss (2000) analysis has been to construct an *efficient frontier* in a two-dimensional plane whose axes are the expectation of the total cost and its variance.
2. Linear Impact Functions Analytical Solutions: Regardless of an individual's tolerance to risk, the only strategies which are candidates for being optimal solutions are found in this one-parameter set. For linear impact functions, they give complete analytical solutions for the strategies in this set.
3. Efficient Frontier Optimal Operating Characteristic: Then, considering the details of risk aversion, they have shown how to select an optimal point on the frontier either by classical mean-variance optimization, or by the concept of value at risk. These solutions are easily constructed numerically, and interpreted graphically by examining the frontier.



4. First Conclusion: Sub-optimal Strategies: Because the set of attainable strategies, and hence the efficient frontier, are generally *smooth* and *convex*, a trader who is at all risk-averse should never trade according to the naïve strategy of minimizing expected cost. This is because in the neighborhood of that strategy, the first order reduction in the variance is attained at the expense of only a second order increase in the expected cost.
5. Second Conclusion: Custom Risk Optimization: Almgren and Chriss (2000) also observe that this careful analysis of the costs and risks of liquidation can be used to give a more precise characterization of the risk of holding the initial portfolio. As an example, they define a Liquidity-Adjusted VaR (L-VaR) to be, for a given time horizon, the minimum VaR of any static liquidation strategy.
6. Actual Gains of Dynamic Trading: Although it may seem counter-intuitive that the optimal strategies can be determined in advance of trading, Almgren and Chriss (2000) argue that only very small gains can be realized by adapting the strategy to the information as it is needed.
7. First Extension: Continuous Time Trading: The limit

$$\tau \rightarrow 0$$

is immediate in all of their solutions. Their trading strategy is characterized by a holdings function  $x(t)$  and a *trading rate*

$$x(t) = \lim_{\tau \rightarrow 0} \frac{n_k}{\tau}$$

Almgren and Chriss (2000) minimum variance strategy has infinite cost, but the optimal strategies for finite  $\lambda$  have finite cost and variance. However, this limit is at best a mathematical convenience, as the market model is implicitly a “coarse-grained” description of the real dynamics.



8. Second Extension: Nonlinear Cost: The conceptual framework outlined by Almgren and Chriss (2000) is not limited to the linear permanent and temporary impact functions

$$g(v) = \gamma v$$

and

$$h(v) = \epsilon sgn(n_k) + \frac{\eta}{\tau} n_k$$

though the exact exponential/hyperbolic solutions are specific to that case. For nonlinear functions  $g(v)$  and  $h(v)$  that satisfy suitable convexity conditions, optimal risk-averse trajectories are found by solving a non-quadratic optimization problem; the difficulty of the problem depends on the specific functional forms chosen.

9. Third Extension Time Varying Coefficients: Almgren and Chriss (2000) framework also covers the case in which the volatility, the market impact parameters, and perhaps the expected drift are all time-dependent; finding the optimal strategy entails solving a linear system of size equal to the number of time periods (times the number of assets, for a portfolio problem). One example in which this is useful is if the price is expected to jump up or down on a known future date – say, an earnings announcement – as long as one has a good estimate of the expected *size* of this jump.

## Numerical Optimal Trajectory Generation

1. Varying Time Interval Cost Distribution:

$$\mathbb{E}[x] = \sum_{k=1}^N \tau_k x_k g\left(\frac{n_k}{\tau_k}\right) + \sum_{k=1}^N n_k h\left(\frac{n_k}{\tau_k}\right)$$



$$\mathbb{V}[x] = \sum_{k=1}^N \tau_k \sigma_k^2 x_k^2$$

$$\tau_k = t_k - t_{k-1}$$

$$n_k = x_k - x_{k-1}$$

2. Time Varying Interval Linear Impact:

$$\mathbb{E}[x] = \frac{1}{2}\gamma X^2 + \epsilon \sum_{k=1}^N |n_k| + \eta \sum_{k=1}^N \frac{n_k^2}{\tau_k} - \frac{1}{2}\gamma \sum_{k=1}^N n_k^2$$

$$\mathbb{V}[x] = \sum_{k=1}^N \tau_k \sigma_k^2 x_k^2$$

$$\tau_k = t_k - t_{k-1}$$

$$n_k = x_k - x_{k-1}$$

3. Varying Interval Linear Impact Objective:

$$\mathbb{U}[x] = \mathbb{E}[x] + \lambda \mathbb{V}[x]$$

implies

$$\mathbb{U}[x] = \frac{1}{2}\gamma X^2 + \epsilon \sum_{k=1}^N |n_k| + \eta \sum_{k=1}^N \frac{n_k^2}{\tau_k} - \frac{1}{2}\gamma \sum_{k=1}^N n_k^2 + \lambda \sum_{k=1}^N \tau_k \sigma_k^2 x_k^2$$



4. Varying Time Interval Linear Impact Jacobian:

$$\frac{\partial \mathbb{U}[x]}{\partial x_j} = 2 \left\{ \lambda \tau_j \sigma_j^2 x_j - \tilde{\eta} \left[ \frac{x_{j+1}}{\tau_{j+1}} - \frac{x_{j-1}}{\tau_j} - x_j \left( \frac{1}{\tau_j} + \frac{1}{\tau_{j+1}} \right) \right] \right\} + \frac{1}{2} \gamma [x_{j-1} - 2x_j + x_{j+1}]$$

Extremum requires that

$$\frac{\partial \mathbb{U}[x]}{\partial x_j} = 0 \quad \forall j = 1, \dots, N-1$$

5. Estimation Quantities for Numerical Optimization: To carry out the numerical optimization, one needs the Jacobian (i.e., gradient) and the Hessian of the optimizer objective function in terms of

$$x_j \quad \forall j = 1, \dots, N-1$$

This has to be computed for each of the following quantities:

- a. Trajectory Slice Permanent Impact Function Expectation Left Holdings
- b. Trajectory Slice Permanent Impact Function Expectation Right Holdings
- c. Trajectory Slice Permanent Impact Function Expectation Cross Holdings Jacobian
- d. Trajectory Slice Temporary Impact Function Expectation Left Holdings
- e. Trajectory Slice Temporary Impact Function Expectation Right Holdings
- f. Trajectory Slice Temporary Impact Function Expectation Cross Holdings Jacobian
- g. Trajectory Slice Permanent Impact Function Variance Left Holdings
- h. Trajectory Slice Permanent Impact Function Variance Right Holdings
- i. Trajectory Slice Permanent Impact Function Variance Cross Holdings Jacobian
- j. Trajectory Slice Temporary Impact Function Variance Left Holdings



- k. Trajectory Slice Temporary Impact Function Variance Right Holdings
- l. Trajectory Slice Temporary Impact Function Variance Cross Holdings Jacobian
- m. Trajectory Slice Core Market Function Expectation Left Holdings
- n. Trajectory Slice Core Market Function Expectation Right Holdings
- o. Trajectory Slice Core Market Function Expectation Cross Holdings Jacobian
- p. Trajectory Slice Core Market Function Variance Left Holdings
- q. Trajectory Slice Core Market Function Variance Right Holdings
- r. Trajectory Slice Core Market Function Variance Cross Holdings Jacobian
- s. Trajectory Permanent Impact Function Expectation Left Holdings
- t. Trajectory Permanent Impact Function Expectation Right Holdings
- u. Trajectory Permanent Impact Function Expectation Cross Holdings Jacobian
- v. Trajectory Temporary Impact Function Expectation Left Holdings
- w. Trajectory Temporary Impact Function Expectation Right Holdings
- x. Trajectory Temporary Impact Function Expectation Cross Holdings Jacobian
- y. Trajectory Permanent Impact Function Variance Left Holdings
- z. Trajectory Permanent Impact Function Variance Right Holdings
- aa. Trajectory Permanent Impact Function Variance Cross Holdings Jacobian
- bb. Trajectory Temporary Impact Function Variance Left Holdings
- cc. Trajectory Temporary Impact Function Variance Right Holdings
- dd. Trajectory Temporary Impact Function Variance Cross Holdings Jacobian
- ee. Trajectory Core Market Function Expectation Left Holdings
- ff. Trajectory Core Market Function Expectation Right Holdings
- gg. Trajectory Core Market Function Expectation Cross Holdings Jacobian
- hh. Trajectory Core Market Function Variance Left Holdings
- ii. Trajectory Core Market Function Variance Right Holdings
- jj. Trajectory Core Market Function Variance Cross Holdings Jacobian
- kk. Objective Utility Function Permanent Impact Function Expectation Left Holdings



- ll. Objective Utility Function Permanent Impact Function Expectation Right Holdings
- mm. Objective Utility Function Permanent Impact Function Expectation Cross Holdings Jacobian
- nn. Objective Utility Function Temporary Impact Function Expectation Left Holdings
- oo. Objective Utility Function Temporary Impact Function Expectation Right Holdings
- pp. Objective Utility Function Temporary Impact Function Expectation Cross Holdings Jacobian
- qq. Objective Utility Function Permanent Impact Function Variance Left Holdings
- rr. Objective Utility Function Permanent Impact Function Variance Right Holdings
- ss. Objective Utility Function Permanent Impact Function Variance Cross Holdings Jacobian
- tt. Objective Utility Function Temporary Impact Function Variance Left Holdings
- uu. Objective Utility Function Temporary Impact Function Variance Right Holdings
- vv. Objective Utility Function Temporary Impact Function Variance Cross Holdings Jacobian
- ww. Objective Utility Function Core Market Function Expectation Left Holdings
- xx. Objective Utility Function Core Market Function Expectation Right Holdings
- yy. Objective Utility Function Core Market Function Expectation Cross Holdings Jacobian
- zz. Objective Utility Function Core Market Function Variance Left Holdings
- aaa. Objective Utility Function Core Market Function Variance Right Holdings
- bbb. Trajectory Slice Core Market Function Variance Cross Holdings Jacobian



6. Permanent Impact Expectation Left Sensitivity:

$$\mathbb{E}_{P,k} = s\tau_k x_k g\left(\frac{x_k - x_{k-1}}{\tau_k}\right)$$

$$\frac{\partial \mathbb{E}_{P,k}}{\partial x_{k-1}} = s\tau_k x_k \frac{\partial g\left(\frac{x_k - x_{k-1}}{\tau_k}\right)}{\partial x_{k-1}}$$

$$\frac{\partial^2 \mathbb{E}_{P,k}}{\partial x_{k-1}^2} = s\tau_k x_k \frac{\partial^2 g\left(\frac{x_k - x_{k-1}}{\tau_k}\right)}{\partial x_{k-1}^2}$$

$$s = sign\left(\frac{x_k - x_{k-1}}{\tau_k}\right)$$

7. Permanent Impact Expectation Right Sensitivity:

$$\mathbb{E}_{P,k} = s\tau_k x_k g\left(\frac{x_k - x_{k-1}}{\tau_k}\right)$$

$$\frac{\partial \mathbb{E}_{P,k}}{\partial x_k} = s\tau_k g\left(\frac{x_k - x_{k-1}}{\tau_k}\right) + s\tau_k x_k \frac{\partial g\left(\frac{x_k - x_{k-1}}{\tau_k}\right)}{\partial x_k}$$

$$\frac{\partial^2 \mathbb{E}_{P,k}}{\partial x_k^2} = 2s\tau_k \frac{\partial g\left(\frac{x_k - x_{k-1}}{\tau_k}\right)}{\partial x_k} + s\tau_k x_k \frac{\partial^2 g\left(\frac{x_k - x_{k-1}}{\tau_k}\right)}{\partial x_k^2}$$

8. Permanent Impact Expectation Cross Jacobian:

$$\mathbb{E}_{P,k} = s\tau_k x_k g\left(\frac{x_k - x_{k-1}}{\tau_k}\right)$$



$$\frac{\partial^2 \mathbb{E}_{P,k}}{\partial x_{k-1} \partial x_k} = s\tau_k x_k \frac{\partial^2 g\left(\frac{x_k - x_{k-1}}{\tau_k}\right)}{\partial x_{k-1} \partial x_k} + s\tau_k \frac{\partial g\left(\frac{x_k - x_{k-1}}{\tau_k}\right)}{\partial x_{k-1}}$$

9. Temporary Impact Expectation Left Sensitivity:

$$\mathbb{E}_{T,k} = (x_k - x_{k-1})h\left(\frac{x_k - x_{k-1}}{\tau_k}\right)$$

$$\frac{\partial \mathbb{E}_{T,k}}{\partial x_{k-1}} = -h\left(\frac{x_k - x_{k-1}}{\tau_k}\right) + (x_k - x_{k-1}) \frac{\partial h\left(\frac{x_k - x_{k-1}}{\tau_k}\right)}{\partial x_{k-1}}$$

$$\frac{\partial^2 \mathbb{E}_{T,k}}{\partial x_{k-1}^2} = -2 \frac{\partial h\left(\frac{x_k - x_{k-1}}{\tau_k}\right)}{\partial x_{k-1}} + (x_k - x_{k-1}) \frac{\partial^2 h\left(\frac{x_k - x_{k-1}}{\tau_k}\right)}{\partial x_{k-1}^2}$$

10. Temporary Impact Expectation Right Sensitivity:

$$\mathbb{E}_{T,k} = (x_k - x_{k-1})h\left(\frac{x_k - x_{k-1}}{\tau_k}\right)$$

$$\frac{\partial \mathbb{E}_{T,k}}{\partial x_k} = h\left(\frac{x_k - x_{k-1}}{\tau_k}\right) + (x_k - x_{k-1}) \frac{\partial h\left(\frac{x_k - x_{k-1}}{\tau_k}\right)}{\partial x_k}$$

$$\frac{\partial^2 \mathbb{E}_{T,k}}{\partial x_k^2} = 2 \frac{\partial h\left(\frac{x_k - x_{k-1}}{\tau_k}\right)}{\partial x_k} + (x_k - x_{k-1}) \frac{\partial^2 h\left(\frac{x_k - x_{k-1}}{\tau_k}\right)}{\partial x_k^2}$$

11. Temporary Impact Expectation Cross Jacobian:

$$\mathbb{E}_{T,k} = (x_k - x_{k-1})h\left(\frac{x_k - x_{k-1}}{\tau_k}\right)$$



$$\frac{\partial^2 \mathbb{E}_{T,k}}{\partial x_{k-1} \partial x_k} = -\frac{\partial h\left(\frac{x_k - x_{k-1}}{\tau_k}\right)}{\partial x_k} + \frac{\partial h\left(\frac{x_k - x_{k-1}}{\tau_k}\right)}{\partial x_{k-1}} + (x_k - x_{k-1}) \frac{\partial^2 h\left(\frac{x_k - x_{k-1}}{\tau_k}\right)}{\partial x_k \partial x_{k-1}}$$

12. Trajectory Jacobian and Hessian Computation: In general, the trajectory Jacobian's and the Hessian's may be computed as sequential, aggregate accumulations over the corresponding slices, with one very critical caveat. In the automated sensitivity generation schemes, all sensitivities to the left-most and the right-most nodes must be excluded, since these do not constitute the control nodes.

13. Power Objective Function Rationale/Formulation: A generalization of the mean-variance optimization and the Value-at-risk schemes is the power objective function formulation

$$\mathbb{U}[x] = \mathbb{E}[x] + \lambda(\mathbb{V}[x])^p$$

$$p = 1$$

corresponds to the regular mean-variance optimization scheme, and

$$p = 0.5$$

corresponds to the liquidity based VaR formulation.

14. Liquidity VaR Control Jacobian/Hessian:

$$\mathbb{U}[x] = \mathbb{E}[x] + \lambda(\mathbb{V}[x])^p$$

$$\frac{\partial \mathbb{U}[x]}{\partial x_i} = \frac{\partial \mathbb{E}[x]}{\partial x_i} + \lambda p(\mathbb{V}[x])^{p-1} \frac{\partial \mathbb{V}[x]}{\partial x_i}$$

$$\frac{\partial^2 \mathbb{U}[x]}{\partial x_i \partial x_j} = \frac{\partial^2 \mathbb{E}[x]}{\partial x_i \partial x_j} + \lambda p(p-2)(\mathbb{V}[x])^{p-2} \frac{\partial \mathbb{V}[x]}{\partial x_i} \frac{\partial \mathbb{V}[x]}{\partial x_j} + \lambda p(\mathbb{V}[x])^{p-1} \frac{\partial^2 \mathbb{V}[x]}{\partial x_i \partial x_j}$$



## References

- Admati, A., and P. Pfleiderer (1988): A Theory of Intra-day Patterns: Volume and Price Variability *Review of Financial Studies* **1** 3-40.
- Almgren, R., and N. Chriss (1999): Value under Liquidation *Risk* **12 (12)** 61-63.
- Almgren, R., and N. Chriss (2000): Optimal Execution of Portfolio Transactions *Journal of Risk* **3 (2)** 5-39.
- Artzner, P., F. Delbaen, J. M. Eber, and D. Heath (1997): Thinking Coherently *Risk* **10 (11)** 68-71.
- Beaver, W. (1968): The Information Content of Annual Earnings Announcements, in: *Empirical Research in Accounting; Selected Studies*. Supplement to *Journal of Accounting Research* 67-92.
- Bertsekas, D. P. (1976): *Dynamic Programming and Stochastic Control* Academic Press.
- Bertsimas, D., and A. W. Lo (1998): Optimal Control of Execution Costs *Journal of Financial Markets* **1** 1-50.
- Brown, K., W. Harlow, and S. Tinic (1988): Risk Aversion, Uncertain Information, and Market Efficiency *Journal of Financial Economics* **22** 355-385.
- Campbell, J. Y., A. W. Lo, and A. C. MacKinlay (1997): *Econometrics of Financial Markets* Princeton University Press.
- Chan, L. K. C., and J. Lakonishok (1993): Institutional Trades and Intra-day Stock Price Behavior *Journal of Financial Economy* **33** 173-199.
- Chan, L. K. C., and J. Lakonishok (1995): The Behavior of Stock Prices around Institutional Trades *Journal of Finance* **50** 1147-1174.
- Charest, G. (1978): Dividend Information, Stock Returns, and Market Efficiency – II *Journal of Financial Economics* **6** 297-330.
- Dann, L. (1981): Common Stock Re-purchases: An Analysis of Returns to Bond-holders and Stock-holders *Journal of Financial Economics* **9** 113-138.



- Easterwood, J. C., and S. R. Nutt (1999): Inefficiency in Analysts' Earnings Forecasts: Systematic Mis-reaction or Systematic Optimism? *Journal of Finance* **54** (5) 1777-1797.
- Fama, E., L. Fisher, M. Jensen, and R. Roll (1969): The Adjustment of Stock Prices to new Information *International Economic Review* **10** 1-21.
- Grinold, R. C., and R. N. Kahn (1999): *Active Portfolio Management 2<sup>nd</sup> Edition* McGraw-Hill.
- Holthausen, R. W., R. W. Leftwich, and D. Mayers (1987): The Effects of Large Block Transactions on Security Prices: A Cross-Sectional Analysis *Journal of Financial Economy* **19** 237-267.
- Holthausen, R. W., R. W. Leftwich, and D. Mayers (1987): Large Block Transactions, the Speed of Response, and Temporary and Permanent Stock-Price Effects *Journal of Financial Economy* **26** 71-95.
- Kalay, A., and U. Loewenstein (1985): Predictable Events and Excess Returns: The case of Dividend Announcements *Journal of Financial Economics* **14** 423-449.
- Keim, D. B., and A. Madhavan (1995): Anatomy of the Trading Process: Empirical Evidence on the Behavior of Institutional Traders *Journal of Financial Economy* **37** 371-398.
- Keim, D. B., and A. Madhavan (1997): Transaction Costs and Investment Style: An Inter-exchange Analysis of Institutional Equity Trades *Journal of Financial Economy* **46** 265-292.
- Kim, O., and R. Verrecchia (1991): Market Reaction to Anticipated Announcements *Journal of Financial Economics* **30** 273-310.
- Kraus, A., and H. R. Stoll (1972): Price Impacts of Block Trading at the New York Stock Exchange *Journal of Finance* **27** 569-588.
- Krinsky, I., and J. Lee (1996): Earnings Announcements and the Components of the Bid-Ask Spread *Journal of Finance* **51** 1523-1555.
- Kyle, A. S. (1985): Continuous Auctions and Insider Trading *Econometrica* **53** 1315-1336.



- Lee, C. M. C., B. Mucklow, and M. J. Ready (1993): Spreads, Depths, and the Impact of Earnings Information: An Intra-day Analysis *Review of Financial Studies* **6** 345-374.
- Lo, A. W., and A. C. MacKinlay (1988): Stock Market Prices do not follow Random Walks: Evidence from a Simple Specification Test *Review of Financial Studies* **1** 41-66.
- Morse, D. (1981): Price and Trading Volume Reaction surrounding Earnings Announcements: A Close Examination *Journal of Accounting Research* **19** 374-383.
- Patell, J. M., and M. A. Wolfson (1984): The Intra-day Speed of Adjustment of Stock Prices to Earnings and Dividend Announcements *Journal of Financial Economics* **13** 223-252.
- Perold, A. F. (1988): The Implementation Short-fall: Paper versus Reality *Journal of Portfolio Management* **14** 4-9.
- Ramaswami, M. (1999): Stock Volatility declines after Earnings are announced – Honest! *Global Weekly Investment Strategy Lehman Brothers*.



## Non-linear Impact and Trading-Enhanced Risk

### Abstract

1. Price and Market Impact Volatility: Almgren (2003) determines optimal trading strategies for the liquidation of a large single-asset portfolio to minimize a combination of volatility risks and market impact costs.
2. Power Law Market Impact Function: The market impact cost is taken to be a power law of the trading rate with an arbitrary positive exponent. This includes, for example, the square root law that has been proposed based on market microstructure theory.
3. Holdings Size Dependent Characteristic Time: In analogy with the linear model, a *characteristic time* is defined for optimal trading, which now depends on the initial portfolio size and decreases as the execution proceeds.
4. Trade Size Dependent Liquidity Volatility: Also considered is a model in which the uncertainty of the realized price is increased by demanding rapid execution; it is shown that the optimal trajectories are defined by a *critical portfolio size* above which this effect is dominant and below which this effect may be neglected.

### Introduction

1. Active vs. Passive Execution Strategy: In the execution of large portfolio transactions, a trading strategy must be determined that balances the risk of delayed execution against the cost of rapid execution; the choice is roughly between an *active* and *passive* trading strategy (Hasbrouck and Schwartz (1988), Wagner and Banks (1992)).
2. Construction of Optimal Execution Strategies: Several papers have constructed optimal strategies for the problem (Almgren and Chriss (1999), Grinold and Kahn



(1999), Almgren and Chriss (2000), Konishi and Makimoto (2001) under the assumption that the liquidity costs per share traded are a linear function of trading rate or block size, and that the only source of volatility in execution is the price volatility of the underlying asset.

3. Price Effect of Block Trades: There is an extensive literature studying the effects of block trades on prices – see Kraus and Stoll (1972), Holthausen, Leftwich, and Mayers (1987, 1990), Chan and Lakonishok (1993, 1995), Keim and Madhavan (1995, 1997), and Koski and Michaely (2000).
4. Assumption of Linear Trading Costs: In practice, linearity of trading costs is an unrealistic assumption. Perold and Salomon Jr. (1991) have argued that the liquidity premium per share demanded by the market will either be a convex or a concave function of the block size depending on whether the market's perception is that the trader is information driven or liquidity driven, respectively.
5. Barra Market Impact Liquidity Premium: In the Barra Market Impact Model (Loeb (1983), Kahn (1993), Barra (1997), Grinold and Kahn (1999)) it is argued, based upon the detailed analysis of the risk-reward choice faced by the equity market maker that the liquidity premium per share should grow as the square root of the block size traded.
6. Block Size Dependent Liquidity Premium: Electronic trading systems such as Optimark (Rickard and Torre (1999)) have been constructed to allow traders specify precisely what liquidity premium they are willing to pay as a function of the block size, and search for clearing opportunities in the mismatch between profiles of different market participants. These effects can be captured by introducing *nonlinear impact functions* into the cost function which is minimized to determine the optimal trading strategies.
7. Approaches of Nonlinear Models: Although linear models are commonly used in empirical regression analyses for simplicity, nonlinear models can often be emulated by dividing trades into categories by size (Bessembinder and Kaufmann (1997), Huang and Stoll (1997)). In fact, Chakravarthy (2001) argues medium-sized trades have a disproportionately large effect on prices.



8. Handling Non-Deterministic Liquidity Premiums: An additional effect not considered in the theoretical strategies considered in the previous work is that the liquidity premium demanded by the market is not deterministic. In fact, the premium will depend on the presence in the market at that instant of participants who are willing to take the other side of the trade.
9. Motivation for Trading Enhanced Risk: Since the presence of these counterparties cannot be predicted in advance, it represents an additional source of risk incurred by the trading profile. That is, a more complete model should include *trading-enhanced risk* representing the increased uncertainty in the execution price incurred by demanding rapid execution of large blocks.
10. Manifestation of Stochastic Liquidity Premiums: Trading-enhanced risk is an implicit feature in the model described in Rickard and Torre (1999). Chordia, Subrahmanyam, and Anshuman (2001) and Hasbrouck and Seppi (2001) argue that liquidity fluctuates due to intrinsic variations in the market activity independent of the trade size. This effect is included in the model by Almgren (2003) via the constant term  $f(0)$  but additional interest is in the *increase* in the execution price uncertainty due to larger block sizes.
11. Nonlinear Block Size Dependence: Thus Almgren (2003) extends the models of Grinold and Kahn (1999) and Almgren and Chriss (2000) in a few important ways. First, the liquidity premium, expressed as an unfavorable motion of the price per share, may be an increasing nonlinear function of the trading rate and the block size – one is considered to be a proxy for the other.
12. Power Law Premium - Closed Form: This cost is reduced by trading slowly, but it must be balanced against the volatility risk incurred by holding the initial portfolio longer than is necessary. In particular, exact solutions are provided in the case this function is a power law with an arbitrary positive exponent, which covers the range of behavior outlined above.
13. Dependence on Initial Portfolio Size: Whereas in the linear case optimal trajectories are characterized by a single *characteristic time* independent of the initial portfolio



size, in the nonlinear case the characteristic time depends upon the initial portfolio size, and scales appropriately as the remaining portfolio diminishes during trading.

14. Comparison with Price Volatility Risk: The realized price per share itself is a random variable, whose variance increases with the increased rate of trading. This introduces an additional source of risk in addition to the volatility. In contrast to the effect of market volatility, this additional risk is *decreased* by trading slowly, submitting small blocks for execution at each time.
15. Volatile Liquidity - Closed Form Solutions: Nearly explicit optimal solutions including this effect can be constructed, and an asymptotic analysis can be used to show that the effect of trading-enhanced risk is most important for large initial portfolios. Indeed, for any given set of parameters there is a characteristic portfolio size above which the optimal strategy is determined by the need to reduce trading-enhanced risk, and below which this effect may be ignored.

## The Model

1. Holdings Trade and Liquidation Time: The general framework followed is that from Almgren and Chriss (2000). At time

$$t = 0$$

$X$  shares of an asset are held, which are to be completely liquidated by the time

$$t = T$$

The initial size  $X$  is positive for a sell program and negative for a buy program; in the former case, there is a long exposure to the market until all the holdings have been eliminated, while in the latter case there is short exposure to the market until the purchase to which the trader has committed to at



$$t = 0$$

is completed. The focus here is on the case

$$X > 0$$

In the case of a portfolio trading problem  $X$  may be a vector, but the consideration here is only on a single asset.

2. The Problem Trade List Determination:  $x(t)$  denotes the holdings at time  $t$  with

$$x(0) = X$$

and

$$x(T) = 0$$

The problem is to choose an optimal function  $x(\cdot)$  so as to minimize a chosen cost functional. Later the limit

$$T \rightarrow \infty$$

will be taken in which the natural execution time emerges as a result of the analysis, but for now the consideration is on an exogenously imposed time horizon.

3. Origin of the *Static* Strategy: It is a rather surprising fact that in the absence of serial correlation in the asset price movements, the optimal price may be determined *statically* at the start of the trading. Unless the market parameters change, observations of price movements in the course of trading do not convey any information that would lead to a change in the strategy.



4. Evenly Spaced Discrete Time Intervals: The analysis starts with the construction of a discrete time model. Thus, for a given trading interval

$$\tau > 0$$

$$t_k = k\tau$$

for

$$k = 0, \dots, N$$

with

$$N = \frac{T}{\tau}$$

and let  $x_k$  be the holdings at time  $t_k$  with

$$x_0 = X$$

and

$$x_N = 0$$

The sales between times  $t_k$  and  $t_{k-1}$  are

$$n_k = x_k - x_{k-1}$$

corresponding to the velocity



$$v_k = \frac{n_k}{\tau} \text{ shares per unit time}$$

Thus

$$x_k = X - \sum_{j=1}^k n_j = \sum_{j=k+1}^N n_j$$

$$k = 0, \dots, N$$

5. Generating the Optimal Trade List: In the discrete time model there is no assumption that the shares are traded at a uniform rate *within* each interval. Rather the assumption is that the trader achieves the optimal execution possible subject to the constraint that  $n_k$  shares are to be traded in the next time interval  $\tau$ . The functions introduced below are a model to describe the trader's best efforts.
6. Temporary/Permanent Market Impact Components: On a standard manner (Stoll (1985)) the impact is divided into a permanent and a temporary component. Thus  $S_k$  describes the price per share of the asset that is publically available in the market.
7. Discrete Arithmetic Permanent Impact Component: The price satisfies the arithmetic random walk

$$S_k = S_{k-1} + \sigma \sqrt{\tau_k} \xi_k - \tau_k g\left(\frac{n_k}{\tau_k}\right) = S_0 + \sigma \sum_{j=1}^k \sqrt{\tau_j} \xi_j - \sum_{j=1}^k \tau_j g(v_j)$$

where  $\xi_j$  are independent random variables with zero mean and unit variance,  $\sigma$  is an *absolute* (not percentage) volatility,  $g(v)$  is the *permanent impact function* representing the effect of the share price of the information conveyed by the trade. This effect is generally small, and below  $g(v)$  is taken to be a linear function, in which case it will have no effect on determining the optimal strategy.



8. Nonlinear Temporary Impact Component: The price that one actually gets on the  $k^{th}$  trade is

$$\tilde{S}_k = S_{k-1} - h\left(\frac{n_k}{\tau_k}\right) + \frac{1}{\sqrt{\tau_k}} f\left(\frac{n_k}{\tau_k}\right) \xi_k$$

$$k = 1, \dots, N$$

Here  $h(v)$  is a nonlinear *temporary impact function* representing the price concession one must accept in order to trade

$$n_k = v_k \tau_k$$

shares in time  $\tau_k$ . The random variables  $\tilde{\xi}_k$  are independent of each other and of  $\xi_k$  with zero mean and unit variance. The new function  $f(v)$  represents the uncertainty of the trade execution as a function of the block size.

9. Liquidity Volatility Term Time Dependence: The factor  $\frac{1}{\sqrt{\tau_k}}$  in the last term of

$$\tilde{S}_k = S_{k-1} - h\left(\frac{n_k}{\tau_k}\right) + \frac{1}{\sqrt{\tau_k}} f\left(\frac{n_k}{\tau_k}\right) \tilde{\xi}_k$$

simply represents a scaling of the parameters if  $\tau_k$  is fixed and finite. When  $\tau_k$  varies, for example when the continuous time

$$\tau_k \rightarrow 0$$

is taken, this factor is necessary to preserve the effect of the trading-enhanced risk.

10. Liquidity Volatility Incremental Time Dependence: If the above term were not present, then breaking a block into several smaller blocks would diversify away the risk due to the uncertainty of each one, regardless of the form of the risk.



11. *Capture* of the Trade Program: The *capture* of the trade program is the total cash received

$$\begin{aligned} \sum_{k=1}^N n_k \tilde{S}_k &= XS_0 + \sigma \sum_{k=1}^N \sqrt{\tau_k} x_k \xi_k - \sum_{k=1}^N \tau_k x_k g(v_k) + \sum_{k=1}^N \sqrt{\tau_k} v_k f(v_k) \tilde{\xi}_k \\ &\quad - \sum_{k=1}^N \tau_k v_k h(v_k) \end{aligned}$$

12. *The Trade Program Implementation Cost*: Discounting is ignored since the trading horizon is short. The *implementation cost* is  $XS_0 - \sum_{k=1}^N n_k \tilde{S}_k$  - a random variable due to uncertainties in price movements and realized prices.
13. *Components of the Implementation Cost*: Note that the implementation cost includes both the costs of finite liquidity and price uncertainty due to delayed execution. This is the *implementation shortfall* of Perold (1988) – see also Jones and Lipson (1999).
14. *Implementation Cost Mean and Variance*: Its expectation and variance at

$$t = 0$$

depend on the free parameters  $x_1, \dots, x_{N-1}$  of the trade strategy:

$$\mathbb{E}[x_1, \dots, x_{N-1}] = \sum_{k=1}^N \tau_k x_k g(v_k) + \sum_{k=1}^N \tau_k v_k h(v_k)$$

$$\mathbb{V}[x_1, \dots, x_{N-1}] = \sum_{k=1}^N \tau_k \sigma_k^2 x_k^2 + \sum_{k=1}^N \tau_k v_k^2 f^2(v_k)$$

15. *Mean Variance Optimal Static Strategies*: A rational trader will construct his or her own strategies to minimize some combination of  $\mathbb{E}[x]$  and  $\mathbb{V}[x]$ . As  $t$  advances the values of  $\mathbb{E}[x]$  and  $\mathbb{V}[x]$  change, but if  $\mathbb{E}[x]$  and  $\mathbb{V}[x]$  are constructed using a classic



mean-variance approach, the optimal strategy continues to be the one determined initially (Almgren and Chriss (2000), Huberman and Stanzl (2005)).

16. Continuous Time Limit Trading Strategy: Now, for analytical convenience, the continuous time limit

$$\tau \rightarrow 0$$

is taken. The trade strategy becomes a continuous path  $x(t)$  and the block sizes  $n_k$  are assumed to be well-behaved so that

$$v_k \rightarrow v(\tau_k k)$$

with

$$v(t) = -\dot{x}(t)$$

17. Continuous Time Mean and Variance: The above expressions have finite limits

$$\mathbb{E}[x] = \int_0^T [x(t)g(v(t)) + v(t)h(v(t))] dt$$

$$\mathbb{V}[x] = \int_0^T [\sigma^2 x^2(t) + v^2(t)f^2(v(t))] dt$$

where the square brackets indicate that these are *functionals* of the entire continuous-time path  $x(t)$ .

18. Caveat behind Continuous Time Analytics: It needs to be emphasized that the continuous time limit is simply an analytical device for obtaining solutions when  $\tau_k$  is



reasonably small; in reality the discreteness of the trading intervals must be taken into account in order to correctly describe trading-enhanced risk.

19. Mean Variance Optimization Objective Function: Introducing the risk-aversion parameter  $\lambda$ , the combined quantity

$$\mathbb{U}[x] = \mathbb{E}[x] + \lambda \mathbb{V}[x]$$

is minimized. Whether or not mean variance optimization is appropriate for a particular case,  $\lambda$  may be considered to be a Lagrange/KKT multiplier for the constrained problem of minimizing  $\mathbb{E}[x]$  for a given  $\mathbb{V}[x]$  and used to construct an efficient frontier in the space of trading trajectories.

20. VaR Based Optimization Objective Functions: More general weightings of risk, including Value-at-Risk, present thorny conceptual problems for time-dependent strategies (Artzner, Delbaen, Eber, and Heath (1999), Basak and Shapiro (2001)).

21. The Calculus of Variations Approach: Minimizing  $\mathbb{U}[x]$  is a standard problem in the calculus of variations:

$$\min_{x(t)} \mathbb{U}[x(t)] = \min_{x(t)} \int_0^T F(x(t), -\dot{x}(t)) dt$$

with

$$F(x, v) = xg(v) + vh(v) + \lambda\sigma^2x^2 + \lambda v^2f^2(v)$$

22. Perturbation Stationarity: Euler-Lagrange Equation: Stationarity to small perturbations requires that the optimal  $x(t)$  solve the Euler-Lagrange equation



$$\begin{aligned}
0 &= \frac{\partial F(x(t), -\dot{x}(t))}{\partial x(t)} + \frac{d}{dt} \left[ \frac{\partial F(x(t), -\dot{x}(t))}{\partial v(t)} \right] \\
&= \frac{\partial F(x(t), -\dot{x}(t))}{\partial x(t)} + \dot{x}(t) \frac{\partial^2 F(x(t), -\dot{x}(t))}{\partial x(t) \partial v(t)} \\
&\quad - \ddot{x}(t) \frac{\partial^2 F(x(t), -\dot{x}(t))}{\partial v^2(t)}
\end{aligned}$$

– a second order ordinary differential equation to be solved with respect to the given endpoints  $x(0)$  and  $x(T)$

23. Integration into the First Order Form: Since  $F(x(t), -\dot{x}(t))$  does not depend explicitly on  $t$  multiplying throughout by  $\dot{x}(t)$  and integrating results in the first-order equation

$$F(x(t), -\dot{x}(t)) + \dot{x}(t) \frac{\partial F(x(t), -\dot{x}(t))}{\partial v(t)} = \text{constant}$$

24. Application to Optimal Trajectory Determination: In the current case one obtains

$$P(-\dot{x}(t)) - P(v_0) = x[g(-\dot{x}(t)) + \dot{x}(t)g'(-\dot{x}(t))] + \lambda\sigma^2x^2$$

with

$$P(v) = v^2 \frac{\partial h(v)}{\partial v} + \lambda v^2 \left[ f^2(v) + 2vf(v) \frac{\partial f(v)}{\partial v} \right] = v^2 \frac{\partial [h(v) + \lambda v f^2(v)]}{\partial v}$$

25. Properties of the Almgren “P” Function: The constant of integration

$$v_0 = -\dot{x}(t)|_{x=0}$$

is the velocity with which  $x(t)$  hits



$$x = 0$$

For a sell program with

$$X > 0$$

$$v_0 \geq 0$$

and conversely for a buy program. Note that

$$P(0) = 0$$

additional assumption is that  $P(v)$  is always an *increasing* function of  $v$  and hence invertible.

26. Explicit Solutions - Key Simplifying Assumptions: Almgren (2001) makes two simplifying assumptions to obtain explicit solutions.

- a. Permanent impact is linear in the trading rate.
- b. The imposed time horizon is infinite.

27. Linear Permanent Market Impact Function: A linear cost function

$$g(v) = \gamma v$$

gives a total cost  $\gamma X$  independent of the path  $x(t)$ . The first term on the right side of

$$P(-\dot{x}(t)) - P(v_0) = x[g(-\dot{x}(t)) + \dot{x}(t)g'(-\dot{x}(t))] + \lambda\sigma^2 x^2$$

vanishes, and then since  $\dot{x}(t)$  appears only on the left side and  $x$  itself appears only on the right side, the general solution can be written in the quadrature form as



$$\int_{x(t)}^X \frac{dt}{P^{-1}[\lambda\sigma^2 x^2 + P(v_0)]} = t$$

28. Bid Ask Spread Dependence Absent: The constant  $v_0$  is to be chosen so that

$$x = 0$$

corresponds to

$$T = 0$$

Note also that any constant in  $h$  disappears; the bid-ask spread does not affect the optimal strategy.

29. No Extraneously Specified Liquidation Time: Since  $P(\cdot)$  is an increasing function, so is  $P^{-1}(\cdot)$ . It is thus clear that as  $v_0$  decreases towards zero, the liquidation time  $T$  increases.
30. Invoking Longest Possible Liquidation Time: If no time horizon is exogenously imposed, the longest possible liquidation time can be obtained by setting

$$v_0 = 0$$

which leads to the quadrature problem

$$\int_{x(t)}^X \frac{dt}{P^{-1}[\lambda\sigma^2 x^2]} = t$$

31. Tractability of the above Solution: Often analytic solutions to the above problem can be found when



$$\int_{x(t)}^X \frac{dt}{P^{-1}[\lambda\sigma^2x^2 + P(v_0)]} = t$$

with

$$v_0 \neq 0$$

would be too intractable. These solutions will still give nearly complete liquidations in finite time determined by market parameters.

## Nonlinear Cost Functions

1. Power Law Temporary Impact Functions: Restricting the attention to the sell program, with

$$v \geq 0$$

the temporary impact functions are taken to be

$$h(v) = \eta v^k$$

$$f(v) = 0$$

with

$$k > 0$$



- for a buy program the signs will be changed in an obvious way. The linear case corresponds to

$$k = 1$$

2. Temporary Impact Almgren “P” Function: As noted above a possible constant in  $h$  corresponding to the bid-ask spread has been neglected. Then

$$P(v) = \eta k v^{k+1}$$

which, for the case of a general finite time horizon with

$$v_0 \geq 0$$

leads to the quadrature problem

$$\int_{x(t)}^X \left( \frac{\lambda \sigma^2}{\eta k} x^2 + v_0^{k+1} \right)^{-\frac{1}{k+1}} dx = t$$

3. Longest Optimal Trajectory Explicit Solution: Taking

$$v_0 \geq 0$$

explicit solutions for the longest optimal trajectories can be obtained:



$$\frac{x(t)}{X} = \begin{cases} \left(1 + \frac{1-k}{1+k} \frac{t}{T_*}\right)^{-\frac{1+k}{1-k}} & 0 < k < 1 \\ e^{-\frac{t}{T_*}} & k = 1 \\ \left(1 - \frac{k-1}{k+1} \frac{t}{T_*}\right)^{\frac{k+1}{k-1}} & k > 1 \end{cases}$$

4. Characteristic Time for Optimal Execution: Here the *characteristic time* is

$$T_* = \left( \frac{k\eta X^{k-1}}{\lambda\sigma^2} \right)^{\frac{1}{k+1}}$$

This is the analog of the *half-life* in the linear case. Only in the linear case

$$k = 1$$

is  $T_*$  independent of the initial portfolio size  $X$ . For

$$k \neq 1$$

the characteristic time depends on the initial size as

$$T_* \sim X^{\frac{k-1}{k+1}}$$

5. Sub Linear Power Law Exponent: For

$$k < 1$$



rapid trading is *under*-penalized relative to the linear case. As the portfolio size increases, volatility risk dominates the trading costs, and the optimal trading time *decreases* since the exponent is negative.

6. Supra Linear Power Law Exponent: For

$$k > 1$$

rapid trading is *over*-penalized relative to the linear case. As the portfolio size increases, the trading cost dominates the volatility risk, and the optimal trading time *increases*, since the exponent is positive. For example, if

$$k = 3$$

then

$$T_* \sim \sqrt{X}$$

7. Characteristic Time vs Half Life: As the portfolio size decreases to zero, reconciliation of the optimal trajectory would use a different starting value  $X$  and hence a different time  $T_*$ . The meaning of  $T_*$  is thus a little less fundamental than in the linear case. However,  $T_*$  scales in exactly the right way to make  $x(t)$  still a static solution.
8. Intuition behind the Characteristic Time: For more intuition, note that the initial rate of selling is

$$-\dot{x}(0) = \frac{X}{T_*}$$

and  $T_*$  is the solution to the relation



$$\lambda\sigma^2 X^2 T = k\eta \left(\frac{X}{T}\right)^k X$$

9. Characteristic Time as a Cost Balance: The left side is the risk penalty associated with holding  $X$  shares for a time  $T$ , and the right side, up to a factor  $k$ , is  $Xh\left(\frac{X}{T}\right)$ , the impact cost associated with selling  $X$  shares over a time  $T$  (without the constant term representing the bid-ask spread, which does not impact the optimal solution).
10. The Longest Optimal Execution Time: For

$$k > 1$$

the trajectory reaches

$$x = 0$$

with

$$v = 0$$

at a finite time

$$T_{MAX} = \frac{k+1}{k-1} T_*$$

Thus, these trajectories are the solution for finite imposed time  $T$  if

$$T > T_{MAX}$$

the trajectory reaches 0 at  $T_{MAX}$  and stays there till  $T$



11. Self-Similar Scaling Trajectory Form: Almgren (2003) contains a graphical illustration of the optimal trajectories generated from

$$\frac{x(t)}{X} = \begin{cases} \left(1 + \frac{1-k}{1+k} \frac{t}{T_*}\right)^{\frac{1+k}{1-k}} & 0 < k < 1 \\ e^{-\frac{t}{T_*}} & k = 1 \\ \left(1 - \frac{k-1}{k+1} \frac{t}{T_*}\right)^{\frac{k+1}{k-1}} & k > 1 \end{cases}$$

The form of the portfolio is independent of a particular choice of time scale  $T_*$  and initial portfolio size  $X$ ; these solutions may be easily scaled to any case.

12. Time Realization of Trajectory Differences: A sense of the differences between the solutions may be gained by noting that for short times, all optimal trajectories are fairly close to each other; but the *tail* of the trajectories is extended for small value of  $k$  which strongly penalize trading at slow rates.

13. Example: Sell-Order Exponent Dependence: For example, as shown by Almgren (2003), at

$$t = T_*$$

the optimal trajectories reduce holdings to 30%, 37%, and 42% of the initial portfolio for

$$k = 2$$

$$k = 1$$

and



$$k = \frac{1}{2}$$

respectively.

14. Example Sell Time Exponent Dependence: Further as demonstrated by Almgren (2003), at

$$\frac{t}{T_*} = 3$$

the trajectory for

$$k = 2$$

has reached

$$x = 0$$

and remains there, the trajectory for

$$k = 1$$

retains 5% of its initial holdings, and the trajectory for

$$k = 2$$

retains 12.5% of the initial holdings. The relative differences become even more pronounced as time continues.

## Objective Function



1. Determination of  $\mathbb{E}[x]$  and  $\mathbb{V}[x]$ :  $\mathbb{E}[x]$  and  $\mathbb{V}[x]$  can be explicitly computed for these solutions from

$$\mathbb{E}[x] = \int_0^T [x(t)g(v(t)) + v(t)h(v(t))]dt$$

$$\mathbb{V}[x] = \int_0^T [\sigma^2 x^2(t) + v^2(t)f^2(v(t))]dt$$

and hence the frontier can be drawn. In doing this the contributions from  $g(v)$  and the term  $\epsilon X$  in  $\mathbb{E}[x]$  are neglected.

2. Closed Form for  $\mathbb{E}_\lambda[x]$  and  $\mathbb{V}_\lambda[x]$ : Then for a general  $k$

$$\mathbb{E}_\lambda[x] = \frac{k+1}{3k+1} \eta \left( \frac{X}{T_*} \right)^{k+1} T_* = \frac{k+1}{3k+1} \eta \left( \frac{\eta \sigma^{2k} X^{3k+1}}{k^k} \lambda^k \right)^{\frac{1}{k+1}} T_*$$

$$\mathbb{V}_\lambda[x] = \frac{k+1}{3k+1} \sigma^2 T_* X^2 = \frac{k+1}{3k+1} \left( \frac{k \eta \sigma^{2k} X^{3k+1}}{\lambda} \right)^{\frac{1}{k+1}}$$

3. The  $(\mathbb{E}, \mathbb{V})$  Efficient Frontier Curve: As  $\lambda$  varies  $(\mathbb{E}, \mathbb{V})$  moves along the hyperboloid-like curve

$$\mathbb{E}_\lambda[x](\mathbb{V}_\lambda[x])^k = \left( \frac{k+1}{3k+1} \right)^{k+1} \eta \sigma^{2k} X^{3k+1}$$

For any positive  $\lambda$  there is a unique solution.

4. Asymptotics of  $\lambda \rightarrow 0$  and  $\lambda \rightarrow \infty$ : As



$$\lambda \rightarrow 0$$

one gets

$$T_* \rightarrow \infty$$

$$\mathbb{E}_\lambda[x] \rightarrow 0$$

and

$$\mathbb{V}_\lambda[x] \rightarrow \infty$$

i.e., optimizing expected cost without regard to variance leads to use of all available time. As

$$\lambda \rightarrow \infty$$

it can be seen that

$$T_* \rightarrow 0$$

$$\mathbb{E}_\lambda[x] \rightarrow \infty$$

and

$$\mathbb{V}_\lambda[x] \rightarrow 0$$

i.e., uncertainty is minimized regardless of the cost.



## Almgren (2003) Example

1. Reference Trading Rate/Market Depth: Almgren (2003) provides a sample methodology for the estimation of the parameters. The first is to choose a representative level of trading rate  $v_{REF}$ . If a specific time period  $\tau$  is chosen  $v_{REF}$  is equivalent to a certain block size

$$n_{REF} = \tau v_{REF}$$

traded in the time period; it may be interpreted as the market *depth* in the sense of Kyle (1985) or Bondarenko (2001).

2. Choice of Reference Trading Rate: The examples below consider stocks that trade one million shares a day, and  $v_{REF}$  is taken to be 10% of that rate, or

$$v_{REF} = 100,000 \text{ share/day}$$

For time period

$$\tau = 1 \text{ hour}$$

with 6.5 *periods per day* this rate is equivalent to trading a block of approximately 15,300 shares in each hour.

3. The Corresponding Temporary Price Impact: Next the price impact  $h_{REF}$  which would be incurred by the steady trading at the reference rate  $v_{REF}$  would be chosen. The share price is assumed to be \$50/share and the assumption is that trading

$$v_{REF} = 100,000 \text{ shares/day}$$

incurs a price impact of 1% or \$0.50/share



4. Choosing the Exponent - The Rationale: Finally, a choice for the value of the exponent  $k$  is made that best fits the belief of how the price impact would depend upon the trading rate for rates smaller or larger than  $v_{REF}$ .
5. Reminder Trading Rate  $k$  Dependence: The choice

$$k = 1$$

corresponds to the linear dependence of price impact on rate;

$$k > 1$$

means that *large* trading rates or block sizes have a disproportionately *large* effect on price; while

$$k < 1$$

means that large trading rates or block sizes have a relatively *smaller* impact.

6.  $h_{REF}/v_{REF}$  Based Impact Model: This impact model is then written as

$$h(v) = h_{REF} \left( \frac{v}{v_{REF}} \right)^k$$

or

$$\eta = \frac{h_{REF}}{v_{REF}^k}$$

7. Daily Volatility and Initial Portfolio: It is also supposed that the stock has an annual volatility of 32% for the daily price change of



$$\sigma = \$1/share \cdot \sqrt{day}$$

A portfolio of initial size

$$X = 100,000$$

is considered, equal to  $\frac{1}{10}^{th}$  of the daily volume.

8. Construction of the Efficient Frontier: There is now enough information to construct the efficient frontier

$$\mathbb{E}_\lambda[x] = \frac{k+1}{3k+1} \eta \left( \frac{X}{T_*} \right)^{k+1} T_* = \frac{k+1}{3k+1} \eta \left( \frac{\eta \sigma^{2k} X^{3k+1}}{k^k} \lambda^k \right)^{\frac{1}{k+1}} T_*$$

$$\mathbb{V}_\lambda[x] = \frac{k+1}{3k+1} \sigma^2 T_* X^2 = \frac{k+1}{3k+1} \left( \frac{k \eta \sigma^{2k} X^{3k+1}}{\lambda} \right)^{\frac{1}{k+1}}$$

from  $\mathbb{E}_\lambda[x]$  and  $\mathbb{V}_\lambda[x]$  for any chosen  $k$  describing the family of solutions as the risk aversion parameter  $\lambda$  ranges over all the possible values

$$0 < \lambda < \infty$$

To construct particular optimal solutions a specific value for  $\lambda$  needs to be set.

9. Variance/Cost Dependence on  $\lambda$ : The results are shown numerically in the table below. For any value of  $k$  the natural liquidation time  $T_*$  increases with the *risk tolerance* parameter  $\frac{1}{\lambda}$ ; as both increase the expected cost decreases and the variance increases.

10.  $\lambda$  Impact:  $T_*, \mathbb{E}_\lambda[x], \sqrt{\mathbb{V}_\lambda[x]}$ : The table below shows that the optimal time scale  $T_*$ , the expected cost  $\mathbb{E}_\lambda[x]$ , and the standard deviation of the cost  $\sqrt{\mathbb{V}_\lambda[x]}$  as functions of the



risk tolerance parameter  $\frac{1}{\lambda}$  and the temporary impact exponent  $k$ . Market and portfolio parameters are as given in the treatment above (the initial portfolio value is \$5 million). As  $k$  is varied, the reference values  $h_{REF}$  and  $v_{REF}$  are held constant; thus, the coefficient  $\eta$  varies as in

$$h(v) = h_{REF} \left( \frac{v}{v_{REF}} \right)^k$$

or

$$\eta = \frac{h_{REF}}{v_{REF}^k}$$

Time  $T_*$  is measured in days;  $\frac{1}{\lambda}$ ,  $\mathbb{E}_\lambda[x]$ , and  $\sqrt{\mathbb{V}_\lambda[x]}$  are in thousands of dollars.

#### 11. Table of $T_*$ , $\mathbb{E}_\lambda[x]$ , $\sqrt{\mathbb{V}_\lambda[x]}$ :

	$\frac{1}{\lambda}$	$k = \frac{1}{2}$	$k = 1$	$k = 2$
$T_*$	1	0.02	0.07	0.22
	10	0.09	0.22	0.46
	100	0.40	0.71	1.00
	1000	1.84	2.24	2.15
	10000	8.55	7.07	4.64
$\mathbb{E}_\lambda[x]$	1	221	354	462
	10	103	112	99
	100	48	35	21
	1000	22	11	5
	10000	10	4	1
$\sqrt{\mathbb{V}_\lambda[x]}$	1	11	19	30
	10	23	33	45



	100	49	59	65
	1000	105	106	96
	10000	226	188	141

12. Large  $\lambda$  Execution Time Dependence: For large values of  $\lambda$ , optimal trajectories all execute rapidly, to reduce the volatility risk associated with the portfolio. When the trading rate is larger than  $v_{REF}$  costs *increase* with increasing  $k$  so larger  $k$  leads to slower trading.
13.  $\lambda$  and  $k$  Combination Impact: Conversely for small  $\lambda$  generally trading proceeds *more slowly* than  $v_{REF}$  in order to minimize the total expected cost. In this regime *smaller*  $k$  is more expensive and leads to relatively slower trading.
14. Cross-Over  $\lambda$  Execution Time: In the intermediate parameter regime the trajectories cross-over from one behavior to the other; larger  $k$  suggests slower trading at the beginning when the rate is large, then relatively more rapid in the tail.
15. Characteristic Time Reference Rate Dependence: Note that

$$T_* = \left( \frac{k\eta X^{k-1}}{\lambda\sigma^2} \right)^{\frac{1}{k+1}}$$

may be re-written as

$$T_* = \left( \frac{kh_{REF}}{\lambda\sigma^2 X} \right)^{\frac{1}{k+1}} \left( \frac{X}{v_{REF}} \right)^{\frac{k}{k+1}}$$

from which it is clear that

$$T_* \rightarrow \frac{X}{v_{REF}}$$

as



$$k \rightarrow \infty$$

regardless of the values of the other parameters.

16. Trade Speed vs. Cost Balance: In this limit, trading more rapidly than the reference rate is very strongly penalized, while trading more slowly is almost without cost, so the optimal strategy is to always trade exactly at the critical rate.
17.  $\lambda$  Estimation from Reference Parameters: Finally, since  $\lambda$  is a difficult parameter to select in practice, it may be observed that it can be estimated if a time scale  $T_*$  is chosen from

$$T_* = \left( \frac{k\eta X^{k-1}}{\lambda\sigma^2} \right)^{\frac{1}{k+1}}$$

and

$$h(v) = h_{REF} \left( \frac{v}{v_{REF}} \right)^k$$

or

$$\eta = \frac{h_{REF}}{v_{REF}^k}$$

- it can be found that

$$\lambda = k \frac{h_{REF} \left( \frac{X/T_*}{v_{REF}} \right)^k X}{\sigma^2 T_* X^2}$$



18. Trading Cost vs. Variance Balance: The numerator is the price concession per share for trading at a concession rate  $\frac{X}{T_*}$  multiplied by the total number of shares  $X$  to get the expected cost; the denominator is the variance that would be incurred by holding  $X$  shares for time  $T_*$ . This ratio is multiplied by  $k$  to correct for nonlinearities which are ignored by this simple description.

## Trading-Enhanced Risk

1. Liquidity Volatility: Functional Form Considered: Now the following functional form is taken for a sell program with

$$v \geq 0$$

$$h(v) = \eta v$$

$$f(v) = \alpha + \beta v$$

The deterministic part of the temporary impact is the linear case

$$k = 1$$

of the previous section.

2. Trading Rate Independent Volatility Component: The constant term in  $f(v)$ , with coefficient  $\alpha$ , represents a constant uncertainty in the realized sale price independent of the rate of selling and of the underlying process. The total risk associated with this term is minimized by splitting the sale into as many parts as possible; thus, this term pushes towards the linear trajectory.



3. Trading Rate Dependent Volatility Component: The linear term, with coefficient  $\beta$ , represents the increase in variance caused by non-zero amounts of selling. This term can even more strongly push toward the linear trajectory.
4. Liquidity Volatility Almgren “P” Function: Then with

$$\dot{x} = -v \leq 0$$

$$P(-\dot{x}(t)) - P(v_0) = x[g(-\dot{x}(t)) + \dot{x}(t)g'(-\dot{x}(t))] + \lambda\sigma^2x^2$$

becomes

$$P(v) = (\eta + \lambda\alpha^2)v^2 + 4\lambda\alpha\beta v^3 + 3\lambda\beta^2 v^4$$

5. Behavior of the Almgren “P” Function: The polynomial  $P(v)$  has

$$P(0) = 0$$

and is increasing for

$$v \geq 0$$

so the graph of the trajectory is always convex, and the inverse of  $P^{-1}$  is well-defined. For a buy program with

$$x \geq 0$$

the sign of the odd term in  $P(v)$  is reversed.

6. No Hard Maximum Execution Time: Since

$$P(v) \sim \mathcal{O}(v^2)$$



for  $v$  near zero, the integrand appearing in the quadrature formulation

$$\int_{x(t)}^X \frac{dt}{P^{-1}[\lambda\sigma^2x^2 + P(v_0)]} = t$$

behaves as  $\mathcal{O}(x^{-1})$  as

$$x \rightarrow 0$$

for

$$v_0 = 0$$

and there is no “hard” maximum time as was found above for

$$k > 1$$

## Constant Enhanced Risk

1. Analytical Solution for  $\beta = 0$  Case: Two special cases are considered for obtaining analytical solution. The first is

$$\beta = 0$$

With this assumption the price uncertainty on each trade is independent of the size of the trade.

2. Trading Trajectory and Execution Time: A solution can then be found for



$$v_0 = 0$$

$$x(t) = X e^{-\frac{t}{T_*}}$$

$$T_* = \sqrt{\frac{\eta + \lambda \alpha^2}{\lambda \sigma^2}}$$

3. Comparison with  $f(v) = 0, k = 1$  Case: This is a pure exponential solution, except that the time constant has been increased by adding the additional variance per transaction to the impact coefficient

$$\eta \mapsto \eta + \lambda \alpha^2$$

4. Expressions for  $\mathbb{E}_\lambda[x]$  and  $\mathbb{V}_\lambda[x]$ : The value functions are

$$\mathbb{E}_\lambda[x] = \frac{1}{2} \eta \frac{X^2}{T_*} = \frac{1}{2} X^2 \sqrt{\frac{\lambda \eta^2 \sigma^2}{\eta + \lambda \alpha^2}}$$

$$\mathbb{V}_\lambda[x] = \frac{1}{2} X^2 \sigma^2 T_* \left( 1 + \frac{\alpha^2}{\sigma^2 T_*} \right) = \frac{1}{2} X^2 \frac{\sigma}{\sqrt{\lambda}} \sqrt{\frac{\eta + 2\lambda \alpha^2}{\eta + \lambda \alpha^2}}$$

5.  $\lambda \rightarrow 0$   $\mathbb{E}_\lambda[x]$  and  $\mathbb{V}_\lambda[x]$  Behavior: The optimal value functions change in a more complicated way than the trajectory. As

$$\lambda \rightarrow 0$$

the behavior is the same as that found earlier;



$$\mathbb{E}_\lambda[x] \rightarrow 0$$

and

$$\mathbb{V}_\lambda[x] \rightarrow \infty$$

since there is less care about the enhanced risk.

6.  $\lambda \rightarrow \infty$   $\mathbb{E}_\lambda[x]$   $T_*$   $\mathbb{V}_\lambda[x]$  Behavior: In contrast as

$$\lambda \rightarrow \infty$$

all quantities have finite limits;

$$T_* \rightarrow \frac{\sigma}{\alpha}$$

$$\mathbb{E}_\lambda[x] \rightarrow \frac{1}{2} \eta \frac{X^2}{T_*}$$

and

$$\mathbb{V}_\lambda[x] \rightarrow \alpha \sigma X^2$$

Since trading itself introduces risk, risk-aversion and cost reduction both encourage spreading the trade over several periods; the minimum variance solution takes finite time and has finite cost.

## Linear Enhanced Risk

1. Analytical Solutions for the  $\alpha = 0$  Case: The next special case is



$$\alpha = 0$$

and

$$P(v) = 0$$

becomes

$$P(v) = \eta v^2 + 3\lambda\beta^2v^4$$

and thus

$$P^{-1}(\omega) = \sqrt{\frac{\sqrt{\eta^2 + 12\lambda\beta^2\omega} - \eta}{6\lambda\beta^2}}$$

2. Trading Trajectory and Characteristic Fields: This can be integrated to obtain

$$\frac{t}{T_*} = F\left(\frac{X}{X_*}\right) - F\left(\frac{x}{X_*}\right)$$

in which the characteristic time and the characteristic share level are

$$T_* = \sqrt{\frac{\eta}{\lambda\sigma^2}}$$

$$X_* = \frac{1}{\sqrt{3}} \frac{\eta}{\lambda\sigma\beta} = \frac{1}{\sqrt{3}} \frac{\sigma T_*^2}{\beta}$$

and the nonlinear function is



$$F(u) = 2z - \coth^{-1} z$$

where

$$z = \sqrt{\frac{1}{2}(1 + \sqrt{1 + 4u^2})}$$

3. Intuition behind the Characteristic Size: The characteristic time is the same as in the earlier section for

$$k = 1$$

and does not depend on the new coefficient  $\beta$ . To understand the characteristic level  $X_*$  note that

$$\sqrt{3}\beta \frac{X_*}{T_*} \frac{1}{\sqrt{T_*}} = \sigma \sqrt{T_*}$$

4. Trading Enhanced Market Volatility Balance: In this expression the left side is the trading induced variance in the share price given by the model

$$\tilde{S}_k = S_{k-1} - h \left( \frac{n_k}{\tau_k} \right) + \frac{1}{\sqrt{\tau_k}} f \left( \frac{n_k}{\tau_k} \right) \tilde{\xi}_k$$

$$k = 1, \dots, N$$

if an initial portfolio of size  $X_*$  were sold in a single period  $T_*$ . The right side is the variance in the share price due to the volatility in the same time interval; at the characteristic share level these two quantities are of comparable size.



5.  $X_*$  much bigger than  $x, X$ : To compare with the previous results note that

$$F(u) \sim \log u + \text{constant} + \mathcal{O}(u^2)$$

$$u \rightarrow 0$$

If this limit is attained by taking a limit of the *parameters* so that

$$\frac{X_*}{X} \rightarrow \infty$$

so that

$$F(u) \sim \log u + \text{constant} + \mathcal{O}(u^2)$$

is valid uniformly over  $x$ , since

$$0 \leq x \leq X$$

- a pure exponential solution results:

$$\frac{t}{T_*} = \log \frac{X}{x(t)} + \mathcal{O}\left[\left(\frac{\lambda\alpha\beta}{\eta}\right)^2\right]$$

$$\frac{\lambda\alpha\beta}{\eta} \rightarrow 0$$

which, in particular, recovers the previous result with

$$k = 1$$



in the limit

$$\beta \rightarrow 0$$

6. Behavior Towards  $x \rightarrow 0$ ; Trajectory Tail: And for any fixed value of the parameters

$$F(u) \sim \log u + \text{constant} + \mathcal{O}(u^2)$$

$$u \rightarrow 0$$

describes the tail of the solution as

$$x \rightarrow 0$$

the time constant of the decay is not affected by the addition of  $\beta$ .

7.  $X_*$  much smaller than  $x, X$ : For

$$x \gg X_*$$

i.e., the initial behavior when

$$X \gg X_*$$

using the expansion

$$F(u) \sim 2\sqrt{u} - \mathcal{O}\left(\frac{1}{\sqrt{u}}\right)$$

$$u \rightarrow \infty$$



gives

$$x(t) \sim X_* \left( C - \frac{1}{2} \frac{t}{T_*} \right)^2$$

$$x \gg X_*$$

with

$$C = \frac{1}{2} F(1)$$

This is the same solution constructed in the earlier Section with

$$k = 3$$

with

$$\eta = \lambda \beta^2$$

#### 8. Almgren (2003) Asymptotic Solution Illustration:

$$\frac{t}{T_*} = F\left(\frac{X}{X_*}\right) - F\left(\frac{x}{X_*}\right)$$

together with

$$\frac{t}{T_*} = \log \frac{X}{x(t)} + \mathcal{O}\left[\left(\frac{\lambda \alpha \beta}{\eta}\right)^2\right]$$



$$\frac{\lambda\alpha\beta}{\eta} \rightarrow 0$$

and

$$x(t) \sim X_* \left( C - \frac{1}{2} \frac{t}{T_*} \right)^2$$

$$x \gg X_*$$

are illustrated in elaborate figures in Almgren (2003) Figure 5.

9. Strategy Construction Approach: Starting Trajectory: Thus, the optimal strategy for construction would be as follows. Assuming

$$x > X_*$$

the initial trades are done using the trajectories of the temporary impact power law with

$$k = 3$$

with

$$\eta = \lambda\beta^2$$

That is, the volatility due to trading completely dominates the intrinsic volatility  $\sigma$

10. Strategy Construction Approach: Tail Trajectory: As  $x(t)$  reaches the level  $X_*$  switch is done to the optimal solution in the linear case

$$k = 1$$



with the other parameters taking their market values. In the tail trading-enhanced risk is a negligible quantity compared to the volatility.

### Almgren (2003) Nonlinear Example Sample

1. Working out the  $\alpha = 0$  Case: In this case the focus is on the previous section in which

$$\alpha = 0$$

and

$$\beta \neq 0$$

so that trading enhanced risk increases linearly with block size with no constant term.

2. The Corresponding Discrete Price Equation: To estimate the coefficients, one starts with the discrete time model. With

$$h(v) = \eta v$$

and

$$f(v) = \beta v$$

the price model

$$\tilde{S}_k = S_{k-1} - h\left(\frac{n_k}{\tau_k}\right) + \frac{1}{\sqrt{\tau_k}} f\left(\frac{n_k}{\tau_k}\right) \tilde{\xi}_k$$



$$k = 1, \dots, N$$

becomes

$$\tilde{S}_k = S_{k-1} - \eta \frac{n_k}{\tau_k} + \beta n_k \tau_k^{-\frac{3}{2}} \tilde{\xi}_k$$

3. Liquidity Risk as Price Volatility Fraction: Assuming that for a particular choice of the trading interval  $\tau$  the standard deviation of price concession associated with trading-enhanced risk is a fraction  $\varrho$  of the deterministic impact – a plausible assumption since both quantities are linearly proportional to the block size.
4. The Corresponding Characteristic Price: That is

$$\beta n_k \tau_k^{-\frac{3}{2}} = \varrho \eta \frac{n_k}{\tau_k}$$

or

$$\beta = \varrho \eta \sqrt{\tau}$$

which gives

$$X_* = \frac{1}{\sqrt{3}\varrho} \frac{1}{\lambda \sigma \sqrt{\tau}}$$

5.  $X_*$  Dependence on Risk Aversion: At this portfolio size, the volatility risk of holding the portfolio roughly balances the risk of selling along the optimal trajectory. Although both of these quantities are risks,  $X_*$  involves  $\lambda$  through its influence on trading time  $T_*$ .



6. Choice of  $\lambda, \tau, \varrho$ : The market parameters are taken to be the same as in the earlier section, with

$$\frac{1}{\lambda} = \$10,000$$

corresponding to the case where the liquidation is one a day. The trading is divided into one hour time intervals, so

$$\tau = \frac{2}{13} \text{ days}$$

and

$$\varrho = \frac{1}{2}$$

7. Estimate of  $\beta$  and  $X_*$ : One obtains

$$\beta = 10^{-6} \$ \cdot \text{day}^{\frac{3}{2}} \cdot \text{share}^{-2}$$

and

$$X_* = 30,000 \text{ shares}$$

corresponding to a portfolio size of \$1.5m. A liquidation problem with initial value greater than will begin in the large  $x$  regime where trading-enhanced risk is dominant and end in the small  $x$  regime where it is negligible.

## Conclusions: Summary and Extensions



1. Summary: Power Law Temporary Component: The treatment seen above obtains explicit analytical solutions for certain cases of the impact model. First, it neglects the effects of trading-enhanced risk, and takes the impact function to be a simple power law. The solutions in this case are straightforward nonlinear extensions of Almgren and Chriss (2000); the exponential solutions obtained there are a particular dividing case of these power law solutions.
2. Summary: Constant Trading-Enhanced Risk: With trading-enhanced risk, two particular cases with linear impact functions were considered. If the price uncertainty per transaction is independent of the transaction size, then the optimal trajectories are given by the previous results, simply augmenting the impact coefficient by the additional variance. A risk-averse trader lengthens his trade program, diversifying some variance away by spreading the execution over more different transactions at the expense of slightly higher volatility risk.
3. Summary: Linear Trading-Enhanced Risk: If price uncertainty per transaction is linearly proportional to the transaction size, then a characteristic portfolio size emerges, above which reduction of the added variance is the dominant effect. In this regime trade trajectories are equivalent to the previous power law solutions with exponent equal to 3. For portfolios smaller than this size the new effect may be neglected compared to the deterministic impact costs and volatility.
4. Extension #1: Optimal Numerical Trajectories: Throughout this treatment, the focus has been on obtaining explicit solutions for the sake of analytical insight. Numerical solutions would be quite straightforward, and allow lifting the restrictions described above, and consideration of a more general class of models.
5. Extension #2: Linear Impact Portfolios: Portfolios of assets are an interesting extension. Already in the linear case (Almgren and Chriss (2000)), to obtain explicit solutions it is necessary to make simplifying assumptions about cross-impacts, for example, that trading in each asset affects only the price of that asset.
6. Extension #3: Nonlinear Impact Portfolios: Even with that assumption, the nonlinear formulation opens a wide class of possible models; for example, should the exponent



be the same for each asset? Determination and characterization of optimal trajectories in this case is a topic for future work.

## References

- Almgren, R. F., and N. Chriss (1999): Value under Liquidation *Risk* **12 (12)** 61-63.
- Almgren, R. F., and N. Chriss (2000): Optimal Execution of Portfolio Transactions *Journal of Risk* **3 (2)** 5-39.
- Almgren, R. F. (2003): Optimal Executions with Nonlinear Impact Functions and Trading-Enhanced Risk *Applied Mathematical Finance* **10 (1)** 1-18.
- Artzner, P., F. Delbaen, J. M. Eber, and D. Heath (1999): Coherent Measures of Risk *Mathematical Finance* **9** 203-228.
- Barra (1997): *Market Impact Model Handbook*.
- Basak, S., and A. Shapiro (2001): Value-at-risk Based Risk management: Optimal Policies and Asset Prices *Review of Financial Studies* **14** 371-405.
- Bessembinder, H., and H. M. Kaufmann (1997): A Comparison of Trade Execution Costs for NYSE and NASDAQ-Listed Stocks *Journal of Financial and Quantitative Analysis* **32** 287-310.
- Bondarenko, O. (2001): Competing Market Makers, Liquidity Provisions, and Bid-Ask Spreads *Journal of Financial Markets* **4 (3)** 269-308.
- Chakravarthy, S. (2001): Stealth Trading: Which Traders' Trades moves Prices? *Journal of Financial Economics* **61** 289-307.
- Chan, L. K. C., and J. Lakonishok (1993): Institutional Trades and Intra-day Stock Price Behavior *Journal of Financial Economy* **33** 173-199.
- Chan, L. K. C., and J. Lakonishok (1995): The Behavior of Stock Prices around Institutional Trades *Journal of Finance* **50** 1147-1174.
- Chordia, T., A. Subrahmanyam, and V. R. Anshuman (2001): Trading Activity and Expected Stock Returns *Journal of Financial Economics* **59** 3-32.



- Grinold, R. C., and R. N. Kahn (1999): *Active Portfolio Management 2<sup>nd</sup> Edition* McGraw-Hill.
- Hasbrouck, J., and R. A. Schwartz (1988): Liquidity and Execution Costs in Equity Markets *Journal of Portfolio Management* **14** 10-16.
- Hasbrouck, J., and D. J. Seppi (2001): Common Factors in Prices, Order Flows, and Liquidity *Journal of Financial Economics* **59** 383-411.
- Holthausen, R. W., R. W. Leftwich, and D. Mayers (1987): The Effects of Large Block Transactions on Security Prices: A Cross-Sectional Analysis *Journal of Financial Economy* **19** 237-267.
- Holthausen, R. W., R. W. Leftwich, and D. Mayers (1990): Large Block Transactions, the Speed of Response, and Temporary and Permanent Stock-Price Effects *Journal of Financial Economy* **26** 71-95.
- Huang, R. D., and H. R. Stoll (1997): The Components of the Bid-ask Spread: A General Approach *Review of Financial Studies* **10 (4)** 995-1034.
- Huberman, G., and W. Stanzl (2005): Optimal Liquidity Trading *Review of Finance* **9 (5)** 165-200.
- Jones, C. M., and M. L. Lipson (1999): Execution Costs of Institutional Liquidity Orders *Journal of Financial Intermediation* **8** 123-140.
- Kahn, R. N. (1993): How the Execution of Trades is best Operationalized, in: *Execution Techniques, True Trading Costs, and the Microstructure of Markets*, K. F. Sherrerd (Editor) AIMR.
- Keim, D. B., and A. Madhavan (1995): Anatomy of the Trading Process: Empirical Evidence on the Behavior of Institutional Traders *Journal of Financial Economy* **37** 371-398.
- Keim, D. B., and A. Madhavan (1997): Transaction Costs and Investment Style: An Inter-exchange Analysis of Institutional Equity Trades *Journal of Financial Economy* **46** 265-292.
- Konishi, H., and N. Makimoto (2001): Optimal Slice of a Block Trade **3 (4)** 33-51.
- Koski, J. L., and R. Michaely (2000): Price, Liquidity, and the Information Content of Trades *Review of Financial Studies* **13** 659-696.



- Kraus, A., and H. R. Stoll (1972): Price Impacts of Block Trading at the New York Stock Exchange *Journal of Finance* **27** 569-588.
- Kyle, A. S. (1985): Continuous Auctions and Insider Trading *Econometrica* **53** 1315-1336.
- Loeb, T. F. (1983): Trading Costs: The Critical Link between Investment Information and Results *Financial Analysts Journal* **39** 39-44.
- Perold, A. F. (1988): The Implementation Short-fall: Paper versus Reality *Journal of Portfolio Management* **14** 4-9.
- Perold, A. F., and R. S. Salomon Jr. (1991): The Right Amount of Assets under Management *Financial Analysts Journal* **47** 31-39.
- Rickard, J. T., and N. G. Torre (1999): Information Systems for Optimal Transaction Implementation *Journal of Management Information Systems* **16** 47-62.
- Stoll, H. R. (1989): Inferring the Components of the Bid-Ask Spread: Theory and Empirical tests *Journal of Finance* **44** 115-134.
- Wagner, W. H., and M. Banks (1992): Increasing Portfolio Effectiveness via Transaction Cost Management *Journal of Portfolio Management* **19** 6-11.



## Market Impact Function/Parameters Estimation

### Introduction, Overview, and Background

1. Power Law Temporary Impact Function: The impact of large trades on prices is very important and widely discussed, but rarely measured. Using a large data set from a major bank and a simple but theoretical model, Almgren, Thum, Hauptmann, and Li (2005) propose that the impact is a  $\frac{3}{5}$  law of the block size, with specific dependence on trade duration, daily volume, volatility, and shares outstanding.
2. Incorporation into Scheduling/Cost Estimation Algorithms: The results can be directly incorporated into an optimal trade scheduling algorithms and into pre- and post-trade estimation systems.
3. Performance Impact of Transaction Costs: Transaction costs are widely recognized as an important determinant of investment performance (see, for example, Freyre-Sanders, Guobuzaitė, and Byrne (2004)). Not only do they affect the realized results of an active investment strategy, they also control how rapidly assets can be converted into cash should the need arise.
4. Direct Fixed Transaction Cost Component: Such costs generally fall into two categories. First are the direct costs such as commissions and fees that are explicitly stated and easily measured. These are important and should be minimized, but are not the focus here.
5. Indirect Controllable Transaction Cost Component: Indirect costs are not explicitly stated. For large trades the most important component of these is the impact of the traders' own actions on the market. These costs are notoriously difficult to measure but they are most amenable to careful trade management and execution.
6. Calibration of Market Impact Costs: Almgren, Thum, Hauptmann, and Li (2005) present a quantitative analysis of the market impact costs based on a large of



Citigroup US brokerage executions. A simple theoretical model that brings in the very important role of execution is used.

7. Out-of-Sample Cross Validation: The model and its calibration are constructed to satisfy two criteria. First, the predicted costs are quantitatively accurate, as determined by direct fit and out-of-sample back testing, as well as extensive consultations with the traders and the other market participants.
8. Deployability with External Execution Schedulers: The results may be directly used as an input into optimal portfolio scheduling systems, although the scheduling algorithm itself may be non-trivial.
9. Use in Citigroup's BECS System: The results of this study have been incorporated into Citigroup's Best Execution Consulting Services (BECS) software for use internally at all desks as well as the clients of the equity division. While this work has focused on US markets, it has been extended to global equities. BECS is the delivery platform for Citigroup's next generation of trading analytic tools, both pre- and post-execution.
10. Extension to the Standard Trading Model: This pre-trade analyzer is an extension to the market standard existing model that has been delivered through the Stock Facts Pro software for the past 25 years (Sorensen, Price, Miller, Cox, and Birnbaum (1998)).
11. Solid Empiricals and Real-Data Verifications: This model is based on better developed empirical foundations; it is based on real trading data taking time into consideration while verifying the results through post-trade analysis. The table below summarizes some of the advantages/disadvantages of this approach.
12. Distinguishing Features of the Model:
  - a. Advantages
    - i. Calibrated from Real Data
    - ii. Includes Time Component
    - iii. Incorporates Intra-day Profiles
    - iv. Uses non-linear Impact Functions
    - v. Confidence Levels for Coefficients



- b. Disadvantages
  - i. Based only on Citigroup Data
  - ii. Little Data for Small-Cap Stocks
  - iii. Little Data for very large Trades
- 13. Academic Industrial Market Data Quantification: Much work in both the academic and industrial communities has been devoted to understanding and quantifying market impact costs. Many academic studies have only worked with publicly available data in Trade and Quote (TAQ) tick-record from the New York Stock Exchange (NYSE).
- 14. Buy-Sell Market Imbalance Analysis: Breen, Hodrick, and Korajczyk (2002) regress the net markets movements over five-minute and half-hour time periods against a net buy-sell impact during the same period, using a linear impact model. A similar model is developed in Kissell and Glantz (2003).
- 15. Impact Cost Function Dependence Analysis: Rydberg and Shephard (2003) develop a rich econometric foundation for describing price motions; Dufour and Engle (2000) investigate the key role of waiting period between successive trades. Using techniques from statistical physics, Lillo, Farmer, and Mantegna (2003) look for a power law scaling in the impact cost function, and find significant dependence on total market capitalization as well as daily volume, and Bouchaud, Gefen, Potters, and Wyart (2004) discover non-trivial serial correlation ion volume and price data.
- 16. Limitations of Public Data Sets: The publicly available data sets lack the reliable classification of individual trades as buyer- or seller- initiated. Even more significantly, each transaction exists in isolation; there is no information on the sequence of trades that form part of the larger transaction.
- 17. Transaction Trade Sequence Incorporation Studies: Some academic studies have used limited data sets made available by asset managers that do have this information, where the date, but not the time duration of the trade is known (Holthausen, Leftwich, and Mayers (1990), Chan and Lakonishok (1995), and Keim and Madhavan (1996)).
- 18. Comparable Studies on Smaller Samples: No other study is known to have carried out this fit explicitly, although various models in use in the industry are based on



regressions of smaller samples (Weisberger and Kreichman (1999), Alba (2002), and de Ternay (2002)).

19. Permanent and Temporary Impact Costs: The transaction cost model embedded in this analysis is based on the model presented by Almgren and Chriss (2000) with non-linear extensions from Almgren (2003). The essential features of this model, as described below, is that it explicitly divides the market impact costs into a permanent component associated with information and a temporary component arising from the liquidity demands made by the execution in a short time.

## Data Description and Filtering Rules

1. Data Generation Period and Universe: The data on which the analysis was based on contains, before filtering, almost 700,000 US stock orders executed by the Citigroup equity trading desks for a 19-month period from December 2001 to June 2003.
2. Orders, Transactions, and Resulting Executions: Each order is broken down into one or more transactions, each of which may generate one or more executions. The information presented below is available for each order.
3. Symbol, Size, and Order Type: The stock symbol, the requested order size, and the sign (buy/sell) of the entire order – the client information is removed.
4. Order Submission Time and Method: The times and the methods by which the transaction was submitted by the Citigroup trader to the market. The time  $t_0$  of the first transaction is taken to be the start of the order. Some of the transactions are sent as market orders, some as limit order, and some are submitted to the Citigroup's automated VWAP server. Except for the starting time and except to exclude the VWAP orders, no use is made of this transaction information.
5. Execution Times, Sizes, and Prices: The times, the prices, and the sizes corresponding to the execution of each transaction is used. Some transactions are cancelled or only partly executed; only completed sizes and prices are used. The execution times are denoted by  $t_1, \dots, t_n$ , sizes by  $x_1, \dots, x_n$ , and prices by  $S_1, \dots, S_n$ .



6. Order Completion Finish Time Frame: All orders are completed within one day, though not necessarily filled.
7. Additional Types of Information Available: In addition, various additional pieces of information – such as instructions given by the client to the trader for the order – e.g., ‘market on close’, ‘market on open’, ‘over the day’, VWAP or blank – are available.
8. Sample Subset Filtering Criteria: The total sample contains 682,582 orders, but only a subset is used for the data analytics.
9. S & P 500 Constituent Stocks Only: To exclude small and thinly traded stocks, only orders on the Standard and Poor’s Index are considered, which represent about half of the total number of orders, but a large majority of the dollar value. Even within this universe there is enough diversity to explore the dependence on the market capitalization, as there are both NYSE and OTC stocks.
10. Exclusion of Highly Volatile Stocks: Also excluded are approximately 400 orders for which the stocks exhibit more than 12.5% daily volatility – 200% annual volatility.
11. Orders that Match the Analysis Objectives: Furthermore, only those orders that are reasonably representative of the actual scheduling strategies that are the ultimate goal are considered.
12. Excluding Market-on-Open/Close: The orders for which the client requested market-on-open or market-on-close executions are excluded. These orders are likely to be executed with strongly non-linear profiles that do not satisfy the modeling assumption – there are only a few hundred of these.
13. Excluding Client Requested VWAP Orders: The orders for which the client requested VWAP execution are excluded. These orders have consistently long execution times and represent very small rates of trading relative to the market volume. These are about 16% of the total number of orders.
14. Excluding Later-in-the-Day Transactions: Also excluded are orders for which any executions are recorded after 16:10 EST, approximately 10% of the total. In many cases these use Citigroup’s block desk for some or all of the transactions, and the fills are reported sometime after the order is completed. Therefore, the time information is not reliable.



15. Caveat - Impact of Orders Exclusion: This exclusion, together with the use of fill size in place of originally requested size, could be a source of significant bias. For example, if clients and traders consistently used limit orders, orders may be filled only if the prices moved in a favorable direction. Analysis of the data set suggests that this effect is not significant – for example the same coefficients are obtained with or without partially filled orders – and informal discussions with the traders confirm the belief that partial fills are not the result of a limit order strategy.
16. Other Minimum Cut off Criteria: Most significantly, small orders are excluded since the goal is to estimate transaction costs in the range where they are significant. Specifically, only the following orders are included:
- a. The order has at least two completed transactions.
  - b. Orders are at least 1,000 shares.
  - c. Orders are at least 0.25% of average daily volume in that stock.
17. Range of Execution Sizes/Times: The results of the model are reasonably stable to changes in the above criteria. After this filtering there are 29,509 orders in the data set; the largest number of executions for any order is 548, and the median is around 5. The median time is around 5 minutes.
18. Order/Volume Ratio Range: The table below shows some descriptive statistics for the sample. Most of the orders constitute only a few percent of the typical market volume, and the model is designed to work within this range of values. Orders greater than a few percent of daily volume have substantial sources of uncertainty that are not modeled here, and the model does not represent them.
19. Summary Statistics of the Sample Orders:

	<b>Mean</b>	<b>Minimum</b>	<b>Q1</b>	<b>Median</b>	<b>Q3</b>	<b>Maximum</b>
Total Cost %	0.04	-3.74	-0.11	0.03	0.19	3.55
Permanent Cost % $I$	0.01	-3.95	-0.17	0.01	0.19	2.66
Temporary Cost % $J$	0.03	-3.57	-0.11	0.02	0.17	2.33
Shares/ADV % $ X $	1.51	0.25	0.38	0.62	1.36	88.62
Time Days	0.39	0.00	0.10	0.32	0.65	1.01



Daily Volatility %	2.68	0.70	1.70	2.20	3.00	12.50
Mean Spread %	0.14	0.03	0.03	0.11	0.16	2.37

## Data Model - Variables

1. Market Impact Input Dependence Estimation: The goal of the study is to determine the market impact in terms of a small number of input variables. Below is a list of precisely which market impacts are measured, and what primary and auxiliary variables will be used to model them.
2. Pre- and Post- Market Prices: Let  $S(t)$  be the price of the asset being traded. For each order the following price points of interest are defined:  $S_0$  is the market price before this order begins executing;  $S_{POST}$  is the market price after this order is completed; and  $\bar{S}$  is the average realized price on the order.
3. The Transaction Weighted Average Price: The realized price

$$\bar{S} = \frac{\sum_{j=1}^N x_j S_j}{\sum_{j=1}^N x_j}$$

is calculated from the transaction data set. The market price  $S_0$  and  $S_{POST}$  are the bid-ask mid points from  $TAQ$ .

4. First Transaction Pre- Trade Price: The pre-trade price  $S_0$  is the price before the impact on the trade begins to be felt (this is an approximation, since some information may leak before any record enters the system).  $S_0$  is computed from the latest quote just preceding the first transaction.
5. Post-trade Price Capture - Caveat: The post-trade price  $S_{POST}$  should capture the permanent effects of the trade program. That is, it should be long enough after the last execution that any effects of temporary liquidity have dissipated.
6. Accounting for the Permanent Impact: In reportedly performing the fits, Almgren, Thum, Hauptmann, and Li (2005) have found that 30 minutes after the last execution



is enough to achieve this. For shorter time intervals, the regressed values depend on the time lag, and about this level the variation stops. That is, they define

$$t_{POST} = t_n + 30 \text{ minutes}$$

7.  $t_{POST}$  Delay Date Roll Over: The price  $S_{POST}$  is taken from the first quote following  $t_{POST}$ . If  $t_{POST}$  is after the market close, it carries over to the next morning. This risks distorting the results by including excessive overnight volatility, but Almgren, Thum, Hauptmann, and Li (2005) have found this to give more consistent results than truncating at the market close.
8. Permanent Realized Impact Variables Definition: Based on these prices the following dimensionless impact variables are defined. The dimensionless permanent impact is

$$I = \frac{S_{POST} - S_0}{S_0}$$

and the dimensionless realized impact is

$$J = \frac{\bar{S} - S_0}{S_0}$$

9. Conversion into Observed Market Impacts: The “effective dimensionless impact  $J$  is the quantity of most interest, since it specifies the actual cash spent or received on the trade. In the model below the temporary impact will be defined to be  $J$  minus a suitable fraction of  $I$  and this temporary impact will be the quantity described by the theory.
10. Signs of the Impact Variables: On any individual order, the signs of  $I, J$  can be positive or negative. In fact, since volatility is a very large contributor to either value, they are almost likely to have either sign. They are defined so that positive cost is experienced if  $I, J$  have the same sign as the total order  $X$ ; for a buy order with



$$X > 0$$

positive cost means that the price  $S(t)$  moves upwards. The average values of  $I, J$  taken across many orders is expected to have the same sign as  $X$ .

11. Intra-day Volume Weighted Time: The level of market activity is known to vary substantially and consistently over different periods of the trading day; this intra-day variation affects both the volume profile and the variance of prices. To capture this effect, all computations are performed in volume time  $\tau$  which represents the fraction of the average day's volume that has executed up to the time  $t$ .
12. Intra-day Volume Weighted Trajectory: Thus, a constant rate trajectory in the  $\tau$  variable corresponds to a VWAP execution in real time. The relationship between  $t$  and  $\tau$  is independent of the daily trading volume; it is scaled so that

$$\tau = 0$$

at market open and

$$\tau = 1$$

at market close.

13. Intra-day Volume Weighted Times: Each of the clock times  $\tau_0, \dots, \tau_n$  in the data set is mapped to the corresponding volume time  $t_0, \dots, t_n$ . Since the stocks in the sample are heavily traded a non-parametric estimator that directly measures the differences in  $\tau$  is used; the shares traded during the period correspond to the execution of each order.
14. Time Volume vs. Price Volatility: Almgren, Thum, Hauptmann, and Li (2005) display an illustration of the empirical profiles. The fluctuations in each time-period in these illustrations correspond to the approximate size of statistical error in the



volume calculation for a 15-minute trade; these errors are typically less than 5%, and are smaller for longer periods.

15. The “Dimensional” Parametric Explanatory Variables: The impacts  $I$  and  $J$  are to be described in terms of the following quantities.

- Total executed size in shares

$$X = \sum_{j=1}^N x_j$$

- Volume Duration of Active Trading:

$$T = \tau_n - \tau_0$$

- Volume Duration of the Impact:

$$T_{POST} = \tau_{POST} - \tau_0$$

16. Caveats around the Explanatory Variables: As noted above,  $X$  is positive for a buy order, and negative for a sell order. Explored defining  $T$  using a size weighted average of execution times, but the results are not substantially different. The intermediate execution times  $\tau_1, \dots, \tau_{n-1}$  are not used, and the execution sizes are not used either except in calculating the order size and the mean realized prices.

17. Fixed Trade - Optimal Time Nodes: In the eventual application for trajectory optimization, the size  $X$  will be assumed given, and the execution schedule here represented by  $T$  will be optimized.

18. Execution Time as Optimizing Parameter: In general, the solution will be a complicated time dependent trajectory parametrized by a time scale  $T$ . For the purposes of data modeling the trajectory optimization is ignored and the schedules are taken to be determined only by a single number  $T$ .



19. Market Core Empirical Parametric Inputs: Although the goal is to explain the dependence of the impact costs  $I, J$  on order size  $X$  and trade time  $T$ , other market variables will influence the solution. The most important of these are:  $V$  – which is the average daily volume in shares, and  $\sigma$  – the daily volatility.
20. Daily Volume/Volatility “Wander” Scale:  $V$  is a 10 day moving average. For volatility, an intra-day estimator that makes use of every transaction in the day is used. It is important to track changes in these variables not only between different stocks but also across time for the same stock.
21. Order Size/Daily Volume Normalization: These values serve primarily to “normalize” the active variables across the stocks with widely varying properties. It seems natural that order size  $X$  should be measured as a fraction of the average daily volume  $V$ :  $\frac{X}{V}$  is a more natural variable than  $V$  itself.
22. Intrinsic Notion of Volume Time: In the model presented below, the order size as a fraction of the average volume traded during the time of execution will also be seen to be important.  $VT$  is estimated directly by taking the average volume that executed between the times  $t_0$  and  $t_n$  over the previous 10 days. In fact, since in the model the trade duration  $T$  appears only in the combination  $VT$  this avoids the need to measure  $T$  directly.
23. “Wander” Scale of Market Impact: The volatility is used to scale the impacts – a certain level of participation in the daily volume should cause a certain level of participation in the ‘normal’ motion of the stock. Empirical investigation by Almgren, Thum, Hauptmann, and Li (2005) shows that volatility is the most important scale factor for cost impact.

## Trajectory Cost Model

1. Constant Volume Time Trading Rate: The model used is based on the framework developed by Almgren and Chriss (2000), and Almgren (2003), with simplifications made to facilitate data fitting. The main simplification is that the rate of trading is



constant (in volume time). In addition, cross impact is neglected, since the data has no information about the effect of trading one stock on the price of the other.

2. The Permanent Impact Market Component: The price impact is decomposed into two components. First is a permanent component that reflects the information transmitted to the market by the buy/sell imbalance. This component is believed to be roughly independent of trade scheduling; ‘stealth’ trading is not admitted by this construction. In the data fit this component will be independent of the execution time  $T$ .
3. The Temporary Market Impact Component: A temporary component reflects the price concession needed to attract counterparties within a specified short time interval. This component is highly sensitive to trade scheduling; here it will strongly depend on  $T$ .
4. Other Elaborate Market Impact Frameworks: More detailed conceptual frameworks have been developed (Bouchaud, Geffen, Potters, and Wyart (2004)), but this easily understood model has become standard in industry and academic literature (Madhavan (2000)).
5. Decomposition of the Realized Market Impact: The realized price impact is a combination of the above two effects. In terms of the realized and the permanent impact defined above and observed from the data, the model may be summarized as

$$\text{Realized} = \text{Permanent} + \text{Temporary} + \text{Noise}$$

with suitable coefficients and scaling depending upon  $T$ . Thus, the temporary impact is obtained as a difference between the permanent impact and the realized impact; it is not directly observed, although there is a direct model for it.

6. Uniform Rate of Order Liquidation: The starting point is the initial order demand of  $X$  shares. This is assumed to be completed by a uniform rate of trading over a volume interval  $T$ . That is, the trade rate in volume units is

$$v = \frac{X}{T}$$



and is held constant until the program is completed.

7. Sign of the Trade Rate: Constant rate in these units is equivalent to VWAP execution during the time of execution. Note that  $v$  has the same sign as  $X$ ; thus

$$v > 0$$

for a buy order and

$$v < 0$$

for a sell order. Market impact will move the price in the same direction as  $v$ .

## Permanent Impact

1. Volatility/Permanent Impact Price Change: The model postulates that the asset price  $S(\tau)$  follows an arithmetic Brownian motion with a drift term that depends on the trade rate term  $v$ . That is

$$\Delta S = S_0 g(v) \Delta \tau + S_0 \sigma \Delta B(\tau)$$

where  $B(\tau)$  is a standard Brownian motion (or a Bachelier process); and  $g(v)$  is the permanent impact function; the only assumptions made are that  $g(v)$  is increasing and has

$$g(0) = 0$$

2. Integrated Form of Price Change: As noted above,  $\tau$  is volume time, representing the fraction of an average day's volume that has executed so far. This expression can be integrated in time taking  $v$  to equal  $\frac{X}{T}$  for



$$0 \leq \tau \leq T$$

to obtain the permanent impact

$$I = Tg\left(\frac{X}{T}\right) + \sigma\sqrt{T_{POST}}\xi$$

where

$$\xi \sim \mathcal{N}(0, 1)$$

is a standard Gaussian variable.

3. Linearity of the Permanent Impact Function: Note that if  $g(v)$  is a linear function, then the accumulated drift at time  $\tau$  is equal to  $\frac{X\tau}{T}$ , the number of shares executed to time  $\tau$ , and the permanent impact  $I$  is proportional to the total order size  $X$  independently of the time scale  $T$ .

## Temporary Impact

1. Temporary Impact Price Change Realization: The actual price received from the trade is

$$\tilde{S}(\tau) = S(\tau) + S_0 h\left(\frac{X}{T}\right)$$

where  $h(v)$  is the temporary impact function. For convenience, it has been scaled by the market price at the start of trading, since the time intervals involved are all less than one day.



2. Discretization of the Price Impact: This expression is a continuous time approximation to a discrete process. A more accurate description would be to imagine that the time intervals would be broken down into intervals such as, say, one hour or 30 minute intervals. Within each interval the average price realized on the trade during that interval would be less favorable than the average price that an unbiased observer would measure during that time interval.
3. Unbiased Price Plus Liquidity Concession: The unbiased price is affected by the previous trades that have been executed before this interval (as well as the volatility) but not on their timing. The additional concession during this time interval is strongly dependent on the number of shares executed in this interval.
4. Closed Form Temporary Impact Expression: At a constant liquidation rate, calculating the time average of the execution price gives the temporary impact expression

$$J - \frac{I}{2} = h\left(\frac{X}{T}\right) + \sigma \left[ \sqrt{\frac{T}{12} \left( 4 - 3 \frac{T}{T_{POST}} \right)} \chi - \frac{T_{POST} - T}{2\sqrt{T_{POST}}} \xi \right]$$

where

$$\chi \sim \mathcal{N}(0, 1)$$

is independent of  $\xi$ . The term  $\frac{I}{2}$  reflects the effect on the later execution prices of permanent impact caused by the earlier parts of the program.

5. Estimate of the Heteroscedastic Corrections: The rather complicated error expression reflects the fluctuations on the middle part of the Brownian motion on  $[0, T]$  relative to the end point at  $T_{POST}$ . It is only used for the heteroscedastic corrections for the regression fits below.
6. Fluctuations and Error Residuals Estimation: The equations



$$I = Tg\left(\frac{X}{T}\right) + \sigma\sqrt{T_{POST}}\xi$$

and

$$J - \frac{I}{2} = h\left(\frac{X}{T}\right) + \sigma \left[ \sqrt{\frac{T}{12} \left( 4 - 3 \frac{T}{T_{POST}} \right)} \chi - \frac{T_{POST} - T}{2\sqrt{T_{POST}}} \xi \right]$$

provide explicit expressions for the permanent and the temporary impact components  $I, J$  in terms of the values of the functions  $g$  and  $h$  at known trade rates, together with the estimates of the magnitude of the error coming from the volatility.

7. Regression Based Impact Form Estimation: The data fitting procedure above is in principle straightforward, the impacts  $I$  and  $J$  are computed from the transaction data, and those values are regressed against order sizes and times as indicated to directly extract the functions  $g(v)$  and  $h(v)$ .

## Choice of the Functional Form

1. Permanent/Temporary Impact Function Structure: The next question that needs to be addressed is what should the structure of the permanent impact function  $g(v)$  and the temporary impact function  $h(v)$  be. Even with a large sample it is not possible to extract these functions purely from data, so a hypothesis must be made about their structure.
2. Power Law Impact Functional Forms: The postulate is that these functions are power laws, that is, that:

$$g(v) = \pm \gamma |v|^\alpha$$

and



$$h(v) = \pm \eta |v|^\beta$$

where the numerical values of the dimensionless coefficients  $\gamma$  and  $\eta$  and the exponents  $\alpha$  and  $\beta$  are to be determined by linear and non-linear regressions on the data. The sign is to be chosen so that  $g(v)$  and  $h(v)$  have the same sign as  $v$ .

3. Range of Power Law Representation: This class of power law is extremely broad. It includes concave functions (exponent  $< 1$ ), convex functions (exponent  $> 1$ ), and linear functions (exponent  $= 1$ ). It is the functional form that is implicitly assumed by fitting straight lines on a log-log plot as is very common in physics, and has been used in this context, for example, by Lillo, Farmer, and Mantegna (2003).
4. Order Type/Exchange Parameter Independence: The same coefficients are taken for buy orders

$$v > 0$$

and sell orders

$$v < 0$$

It would be a trivial modification to introduce different coefficients  $\gamma_{\pm}$  and  $\eta_{\pm}$  for the two sides, but the exploratory analysis by Almgren, Thum, Hauptmann, and Li (2005) has not indicated a strong need for this. Similarly, it would be possible to use different coefficients for stocks traded on different exchanges, but this does not appear to be necessary either.

5. Quasi Arbitrage Permanent Impact Elimination: There is reason to be specific in the choice of the exponents. For the permanent impact function there is a strong reason to prefer the linear model with

$$\alpha = 1$$



This is the only value for which the model is free from quasi-arbitrage (Huberman and Stanzl (2004)).

6. Linearity of the Permanent Impact: Furthermore, the linear function is the only one for which the permanent price impact is independent of the trading time. Of course, this substantial conceptual simplification must be supported by the data.
7. Concave Nature of the Temporary Exponents: For temporary impacts, there is ample evidence indicating that the function should be concave, that is

$$0 < \beta < 1$$

This evidence dates back to Loeb (1983) and is strongly demonstrated by the fits in Lillo, Farmer, and Mantegna (2003). In particular theoretical arguments (Barra (1997)) suggest that the particular value of

$$\beta = \frac{1}{2}$$

is especially plausible, resulting in a square root impact function.

8. Verification of Power Exponent Values: The approach is then as follows. Unprejudiced fits to the power law functions shall be made to the entire data set to determine the best estimates for the exponents  $\alpha$  and  $\beta$ . The validity of the values

$$\alpha = 1$$

and

$$\beta = \frac{1}{2}$$



will then be tested to validate the linear and the square root candidate functional forms.

9. Determination of the Impact Coefficients: Once the exponents have been selected, simple linear regression is adequate to determine the coefficients. In this regression heteroscedastic weightings are used, with the error magnitudes from

$$I = Tg\left(\frac{X}{T}\right) + \sigma\sqrt{T_{POST}}\xi$$

and

$$J - \frac{I}{2} = h\left(\frac{X}{T}\right) + \sigma \left[ \sqrt{\frac{T}{12} \left( 4 - 3 \frac{T}{T_{POST}} \right)} \chi - \frac{T_{POST} - T}{2\sqrt{T_{POST}}} \xi \right]$$

The result of this regression is not only the values for the coefficients, but also a collection of the error residuals  $\xi$  and  $\chi$  which must be tested for normality as the theory supposes.

## Cross-Sectional Description

1. Motivations for the Properties Normalization: The above analysis has assumed an ‘ideal’ asset, all of whose properties remain constant in time. For any real asset, the parameters that determine the market impact will vary with time. For example, one would expect that the execution of a given number of shares would incur higher impact costs on a day with unusually low volume or unusually high volatility.
2. Basis for the Normalizer Choice: That is, the impact of the cost functions should be expressed in terms of the dimensionless quantity  $\frac{X}{VT}$  rather than  $X$  itself, where  $V$  is the average number of shares per day defined above.



3. Normalization of the Price Moves: Furthermore, the motion of the price should not be given as a raw percentage figure, but it should be expressed as a fraction of ‘normal’ daily motion of the price, as expressed by the volatility  $\sigma$ .
4. Normalized Expressions for  $I, J$ : With these assumptions, the equations

$$I = Tg\left(\frac{X}{T}\right) + \sigma\sqrt{T_{POST}}\xi$$

and

$$J - \frac{I}{2} = h\left(\frac{X}{T}\right) + \sigma \left[ \sqrt{\frac{T}{12} \left( 4 - 3 \frac{T}{T_{POST}} \right)} \chi - \frac{T_{POST} - T}{2\sqrt{T_{POST}}} \xi \right]$$

can be modified to

$$I = \sigma T g\left(\frac{X}{VT}\right) + [\text{noise}]$$

and

$$J - \frac{I}{2} = \sigma h\left(\frac{X}{VT}\right) + [\text{noise}]$$

respectively, where  $[\text{noise}]$  is the error expression depending on the volatility.

5. Dimensionless Permanent Temporary Function Inputs: Now  $g$  and  $h$  are dimensionless functions of a dimensionless variable. They are assumed to be constant in time for a single stock across days when  $\sigma$  and  $V$  vary. The next step is to investigate these functions for their dependence on cross-stock variables.

## Model Determination



1. Specification of Extraneous Model Regressors: To bring the full size of the data into play, one must address the more complex and the less precise question of how the impact functions vary across the stocks, that is, how much they depend variables such as market capitalization, shares outstanding, bid-ask spread, or other quantities. Temporary and permanent impact must be considered separately.
2. Permanent Impact Function Liquidity Regressor: A ‘liquidity factor’  $\mathcal{L}$  is inserted into the permanent cost function  $g(v)$ , where  $\mathcal{L}$  depends on the market parameters characterizing each stock (in addition to daily volume and liquidity). There are several candidates for inputs into  $\mathcal{L}$ .
3. Liquidity Regressor Candidate - Inverse Turnover: The form of  $\mathcal{L}$  is constrained to be

$$\mathcal{L} = \left[ \frac{\Theta}{V} \right]^\delta$$

where  $\Theta$  is the total number of shares outstanding, and  $\delta$  is the exponent to be determined. The dimensionless quantity  $\frac{\Theta}{V}$  is the inverse of the ‘turnover’ – the fraction of the company’s value traded each day. This is a natural explanatory variable, and has been used in empirical studies such as Breen, Hodrick, and Korajczyk (2002).

4. Liquidity Regressor Candidate – Bid-Ask: Almgren, Thum, Hauptmann, and Li (2005) did not find any consistent dependence on the bid-ask spread across the sample, so it is not included in  $\mathcal{L}$ .
5. Liquidity Regressor Candidate - Market Capitalization: This differs from the shares outstanding by the price per share, so including this factor is equivalent to including a ‘price effect’. The study by Almgren, Thum, Hauptmann, and Li (2005) found that there is a persistent price effect, as also found by Lillo, Farmer, and Mantegna (2003), but that the dependence is weak enough that it may be neglected in favor of the conceptually simpler quantity  $\frac{\Theta}{V}$ .



6. Temporary Impact Function - Regressor Candidate: In further extensive preliminary exploration, it was found that the temporary cost function  $h(v)$  does not require any stock-specific modification; liquidity costs as a fraction of volatility only depends upon the fraction of the shares traded as a fraction of the average daily volume.
7. Revised I, J Functional Forms: After assuming the functional form defined above, the model is validated and the exponent  $\delta$  is determined by performing a non-linear regression of the form

$$\frac{I}{\sigma} = \gamma T sgn(X) \left| \frac{X}{VT} \right|^{\alpha} \left[ \frac{\Theta}{V} \right]^{\delta} + [\text{noise}]$$

and

$$\frac{1}{\sigma} \left[ J - \frac{I}{2} \right] = \eta sgn(X) \left| \frac{X}{VT} \right|^{\beta} + [\text{noise}]$$

where  $[\text{noise}]$  is the again the heteroscedastic error term from

$$I = T g \left( \frac{X}{T} \right) + \sigma \sqrt{T_{POST}} \xi$$

and  $sgn$  is the sign function.

8. Estimates and Residuals of Exponents: A modified Gauss-Newton optimization algorithm was used to determine the values of  $\alpha$ ,  $\beta$ , and  $\delta$  that minimized the normalized residuals. The results are:

$$\alpha = 0.891 \pm 0.10$$

$$\delta = 0.267 \pm 0.22$$

$$\beta = 0.600 \pm 0.038$$



9. Errors Represented as One Sigma Amounts: Here, as throughout this chapter, the error bars are expressed with  $\pm$  are one standard deviation, assuming Gaussian error model. Thus the ‘true’ value can be expected to be within this range with 67% probability, and within a range twice as large with 95% probability.
10. Choice of Linear Permanent Impact: From these values the following conclusions can be drawn. First the value

$$\alpha = 1$$

for linear impact cannot be reliably rejected. In view of enormous practical simplification of the linear permanent impact

$$\alpha = 1$$

is chosen.

11. Permanent Impact Liquidity Exponent Estimation: The liquidity factor is very approximately

$$\delta = \frac{1}{4}$$

12. Temporary Impact Power Law Exponent: For temporary impact, the analysis confirms the concavity of the function with  $\beta$  strictly less than one. This confirms the fact that the bigger the trades made by the fund managers on the market, the less additional cost they experience per share traded. At 95% confidence level, the square root model

$$\beta = \frac{1}{2}$$



is rejected. The temporary cost exponent is therefore fixed at

$$\beta = \frac{3}{5}$$

In comparison with the square root model, this gives slightly smaller costs for smaller trades, and slightly larger costs for large trades.

13. Permanent Dependence on Shares Outstanding: Note that because

$$\delta > 0$$

for fixed values of number  $X$  of shares in the order, and the average daily volume  $V$ , the cost increases with  $\Theta$ , the total number of shares outstanding. In effect, a large number of outstanding shares means that a smaller fraction of the company is traded each day, so a given fraction of that flow has a greater impact.

14. Linear Permanent Concave Temporary Impact: Therefore, the results confirm empirically the theoretical arguments of Huberman and Stanzl (2004) for permanent impact that is linear in the block size, and the concavity of the temporary impact has been widely described in the literature for both theoretical and empirical reasons.

## Determination of the Coefficients

1. Estimation of the Impact Coefficients: After fixing the exponent values, the values of  $\gamma$  and  $\eta$  are determined by linear regression of the models.

$$\frac{I}{\sigma} = \gamma T sgn(X) \left| \frac{X}{VT} \right|^{\alpha} \left[ \frac{\Theta}{V} \right]^{\delta} + [\text{noise}]$$

and



$$\frac{1}{\sigma} \left[ J - \frac{I}{2} \right] = \eta sgn(X) \left| \frac{X}{VT} \right|^{\beta} + [\![noise]\!]$$

using the heteroscedastic error estimates given in

$$I = T g \left( \frac{X}{T} \right) + \sigma \sqrt{T_{POST}} \xi$$

and

$$J - \frac{I}{2} = h \left( \frac{X}{T} \right) + \sigma \left[ \sqrt{\frac{T}{12} \left( 4 - 3 \frac{T}{T_{POST}} \right)} \chi - \frac{T_{POST} - T}{2\sqrt{T_{POST}}} \xi \right]$$

It is found that

$$\gamma = 0.314 \pm 0.041$$

with

$$t = 7.7$$

and

$$\eta = 0.142 \pm 0.0062$$

with

$$t = 23$$



2. Interpretation of the  $t$  statistic: The  $t$  statistic is calculated assuming that the Gaussian model expressed in

$$I = Tg\left(\frac{X}{T}\right) + \sigma\sqrt{T_{POST}}\xi$$

and

$$J - \frac{I}{2} = h\left(\frac{X}{T}\right) + \sigma \left[ \sqrt{\frac{T}{12} \left( 4 - 3 \frac{T}{T_{POST}} \right)} \chi - \frac{T_{POST} - T}{2\sqrt{T_{POST}}} \xi \right]$$

is valid; the error estimates are the values divided by the  $t$  statistic. Although the actual residuals are fat-tailed as discussed below, these estimates indicate that the coefficient values are highly significant.

3. Permanent Impact Signal Contribution: The  $\mathbb{R}^2$  values are typically less than 1% indicating that only a small part of the dependent variables  $I$  and  $J$  is explained by the model in terms of the independent variables. This is precisely what is expected given the small size of the random impact term relative to the random motion of the price due to the volatility arising from the trade execution.
4. Importance of the Permanent Cost: The permanent persistent cost, though small, is of major importance since it is on the average the cost incurred by the fund managers while trading. Furthermore, since most orders are part of large portfolio trades, the volatility costs experienced on the portfolio level is considerably lower than exhibited in the stock-level analysis, increasing the significance of the fraction of the impact cost estimated. As previously mentioned, the non-linear optimization of the volatility versus impact cost trade-off would reveal additional profitable strategies.
5. Universal Coefficients of Market Impact: The dimensionless numbers  $\gamma$  and  $\eta$  are the universal coefficients of market impact. According to the model, they apply to every order and every asset in the entire data set.
6. Interpretation Caveat  $I$  vs.  $J$ : To summarize, they are to be inserted into the equations



$$I = \gamma\sigma \frac{X}{V} \left[ \frac{\Theta}{V} \right]^{\frac{1}{4}} + [\text{noise}]$$

and

$$J = \frac{I}{2} + \eta\sigma sgn(X) \left| \frac{X}{VT} \right|^{\frac{3}{5}} + [\text{noise}]$$

giving the expectation of the impact costs; in any particular order the realized values will vary greatly due to the volatility. To reiterate,  $I$  does not signify the total cost, but is simply the net price motion from pre-trade to post-trade. The actual cost experienced by the trade is signified by  $J$ .

7. Sub Group Impact Parameters Determination: Almgren, Thum, Hauptmann, and Li (2005) have chosen the simple form above to have a single model that applies reasonably well across the entire data set which consists entirely of large cap stocks in the US market. More detailed models could be constructed to capture more limited data or assets, or to account for variations across global markets. In practice, it is expected that the coefficients, perhaps even the exponents, or maybe even the functional forms, will be continually updated to reflect the most recent data.
8. Example of Impact Cost: The table below shows the impact cost functions and the numerical examples for two large cap stocks when the customer buys 10% of the average daily volume. The permanent cost is independent of the time of execution, the temporary cost depends on the time of execution, but across different asset it is the same fraction of the daily volatility.  $K$  is written as

$$K = J - \frac{I}{2}$$

9. Example of Impact Costs Table:



			IBM			DRI		
Average Daily Volume	Million	$V$	6.561			1.929		
Shares Outstanding	Million	$\Theta$	1728			168		
Inverse Turnover		$\frac{\Theta}{V}$	263			87		
Daily Volatility	%	$\sigma$	1.57			2.26		
Normalized Trade Rate		$\frac{X}{V}$	0.1			0.1		
Normalized Permanent Impact		$\frac{I}{\sigma}$	0.126			0.096		
Permanent Price Impact	bp	$I$	20			22		
Trade Duration	Days	$T$	0.1	0.2	0.5	0.1	0.2	0.5
Normalized Temporary Impact		$\frac{K}{\sigma}$	0.142	0.094	0.054	0.142	0.094	0.054
Temporary Impact Cost	bp	$K$	22	15	8	32	21	12
Realized Cost	bp	$J$	32	25	18	43	32	23

10. Example of Impact Cost – Analysis: In the above example, because DRI turns over  $\frac{1}{87}$  of its float each day, whereas IBM turns over only  $\frac{1}{263}$ , trading 10% of the day's volume causes a permanent price move of only 0.1 times volatility for DRI, but 0.13 times for IBM; half of this is experienced as cost. Because the permanent cost function is linear, the permanent cost numbers are independent of the times of execution.

## Residual Analysis

1. Results of Impact and Market: The results of the above analysis are not simply the values of the coefficients presented. In addition, the error formulation provides



specific predictions for the nature of the residuals  $\xi$  and  $\chi$  for the permanent and the temporary impact respectively from

$$I = Tg\left(\frac{X}{T}\right) + \sigma\sqrt{T_{POST}}\xi$$

and

$$J - \frac{I}{2} = h\left(\frac{X}{T}\right) + \sigma \left[ \sqrt{\frac{T}{12} \left( 4 - 3 \frac{T}{T_{POST}} \right)} \chi - \frac{T_{POST} - T}{2\sqrt{T_{POST}}} \xi \right]$$

2. Validating the Independence of the “Wanderers”: Under the assumption that the asset price is a Brownian motion with the drift caused by the impact, these two variables should be independent standard Gaussians. This assumption has already been used in heteroscedastic regression, now it needs to be verified.
3. Residual Mean and Covariance: Almgren, Thum, Hauptmann, and Li (2005) demonstrate histograms and  $Q - Q$  plots of  $\xi$  and  $\chi$ . The means are quiet close to zero, the variances are reasonably close to 1, and the correlation is reasonably small.
4. Fat-Tailed Nature of the Distribution: But the distribution is extremely fat-tailed, as is normal for returns distributions on short-time intervals (see Rydberg (2000)), and hence does not indicate that the model is poorly specified. Nonetheless, the structure of the residuals confirms that the model is close to the best that can be obtained within the Brownian framework.

## References

- Alba, J. N. (2002): Transaction Cost Analysis: How to achieve the best Execution, in: *Best Execution: Executing Transactions in Securities Markets on Behalf of Investors* European Asset Management Association 6-17.



- Almgren, R. F., and N. Chriss (2000): Optimal Execution of Portfolio Transactions *Journal of Risk* **3 (2)** 5-39.
- Almgren, R. F. (2003): Optimal Executions with Nonlinear Impact Functions and Trading-Enhanced Risk *Applied Mathematical Finance* **10 (1)** 1-18.
- Almgren, R. F., C. Thum, E. Hauptmann, and H. Li (2005): Equity Market Impact *Risk* **18 (7)** 57-62.
- Barra (1997): *The Market Impact Handbook*.
- Bouchaud, J. P., Y. Gefen, M. Potters, and M. Wyart (2004): Fluctuations and Responses in Financial Markets: The “Subtle” Nature of Random Price Changes *Quantitative Finance* **4 (2)** 176-190.
- Breen, W., L. Hodrick, and R. Korajczyk (2002): Predicting Equity Liquidity *Management Science* **48 (4)** 470-483.
- Chan, L. K. C., and J. Lakonishok (1995): The Behavior of Stock Prices around Institutional Trades *Journal of Finance* **50** 1147-1174.
- de Ternay, A. (2002): Orders Execution – Emergence of a New Added Value: Concerns, from Regulator to Operator, in: *Best Execution: Executing Transactions in Securities Markets on Behalf of Investors European Asset Management Association* 31-43.
- Dufour, A. and R. Engle (2000): Time and the Price Impact of a Trade *Journal of Finance* **55 (6)** 2467-2498.
- Freyre-Sanders, A., R. Guobuzaitė, and K. Byrne (2004): A Review of Trading Cost Models *Journal of Investing* **13** 93-115.
- Holthausen, R. W., R. W. Leftwich, and D. Mayers (1990): Large Block Transactions, the Speed of Response, and Temporary and Permanent Stock-Price Effects *Journal of Financial Economy* **26** 71-95.
- Huberman, G., and W. Stanzl (2004): Price Manipulation and Quasi-arbitrage *Econometrica* **72 (4)** 1247-1275.
- Keim, D. B., and A. Madhavan (1995): Anatomy of the Trading Process: The Upstairs Market for Large Block Transactions; Analysis and Measurement of Price Effects *Review of Financial Studies* **9** 1-36.



- Kissell, R., and M. Glantz (2003): *Optimal Trading Strategies* **Amacom**.
- Lillo, F., J. Farmer, and R. Mantegna (2003): Master Curve for Price-Impact Function *Nature* **421** 129-130.
- Loeb, T. (1983): Trading Cost: The Critical Link between Investment Information and Results *Financial Analysts Journal* **39 (3)** 39-44.
- Rydberg, T. (2000): Realistic Statistical Modeling of Financial Data *International Statistical Review* **68 (3)** 233-258.
- Rydberg, T., and N. Shephard (2003): Dynamics of Trade-by-trade Price Movements; Decomposition and Models *Journal of Financial Economics* **1 (1)** 2-25.
- Sorensen, E., L. Price, K. Miller, D. Cox, and S. Birnbaum (1998): *The Solomon Smith Barney Global Equity Impact Cost Model* Technical Report **Solomon Smith Barney**.
- Weisberger, D., and S. B. Kreichman (1999): *The Accurate Measurement of Equity Trading Costs* Technical Report **Solomon Smith Barney Portfolio Trading Strategies**.



## Optimal Execution of Program Trades

### Introduction

1. The Definition of the Program Trade: The program trade involves the sale or the purchase of a basket of stocks that is too large to be traded immediately in the market. When such trades are brokered, they take on of two forms.
2. The Agency Program Trade Type: In the *agency trade* type the broker executes the trade on behalf of the client on a commission basis, and all the risk of the trade is borne by the client.
3. The Principal Program Trade Type: In a *principal trade*, also called a principal basket, principal bid, or risk bid, the broker directly purchases the entire basket for a fixed price, usually expressed as a discount to the fair market value. By design, principal trades transfer all of the risk from the client to the broker in exchange for a single price, which therefore proxies for the risk of the market portfolio.
4. Program Trades Share of NYSE: Program trading represents an increasing percentage of the overall stock market volume. In 2002 program trades averaged over 30% of the New York Stock Exchange trading volume, up from approximately 20% in 1999 and 2000.
5. Share of the Principal Trades: Overall 50% of all trading volume took place in the NYSE, and of this 30-40% was done on a principal basis (NYSE (2002)).
6. Attributes of a Principal Trade: Almgren and Chriss (2003) begin by making two key observations about the program trading business. First the principal trade consists of two attributes – a basket of stocks, and a price.
7. The Basket Components and Price: The basket is determined by the client, but the price – usually expressed as a per share discount to the current market value – is agreed upon by the client and the broker, and the potential profitability of the trade depends upon the precise price that can be secured for trading the basket. Throughout



this chapter, the trade will be treated as the logical unit which consists of both the basket and its price.

8. Corporate Finance View of the Principal Trade: The second observation is that since program trading represents an investment of the firm's capital, the correct way to view performance is on an annualized basis.
9. First Result - The “Optimal” Execution: The treatment here constructs a mathematical framework for pricing and trading of principal baskets, yielding two main results. First it shows that for a broad class of measures of annualized risk-adjusted return there is a unique optimal way to trade the basket.
10. Second Result - The Information Ratio: Second, the measure called the *information ratio* of a trade is introduced, which is the ratio of the annualized expected profit to the annualized standard deviation of the profit.
11. Information Ratio as a Relative Value Metric: Given a proposed trade one can calculate the information ratio from the known information about the basket. Because the information ratio is annualized by the expected time to completion of the liquidation, it provides a way to compute the profitability of trades of different sizes and levels of liquidity. This yields a powerful tool for analyzing principal businesses.
12. Accommodating Various Market Impact Models: One does not need to know all of the constituents of a basket to compute its information ratio – just the volatility, the liquidity, and a proposed price for the basket. In particular, one must know the effect of trading a basket on the prices of its constituents, that is, the market impact. The methods in this treatment fit in with a wide variety of existing models.
13. The “Hurdle Rate” Minimum Price: The information ratio is primarily a pricing tool. By specifying a hurdle rate – a minimum information ratio that every principal basket must exceed – one can determine for a given principal basket what the minimum price will be that exceeds the hurdle rate.
14. Viability of the “Going Price”: Alternatively, it may be used as an evaluation tool. If one knows the “going price” of a trade, one can compute the information ratio of the trade based on that price to determine whether it is worthwhile to submit a winning bid for the business.



## Efficient Frontier Pricing of Program Trades

1. Ex-Ante Risk Adjusted Return: The program trade is essentially the use of capital by a trading desk. In this treatment, the use of this capital is evaluated using an *ex-ante* risk adjusted return ratio – the ratio of the predicted profit to the standard deviation of the predicted profit. This turns out to be the familiar information ratio, analogous to the familiar Sharpe ratio of the Information Theory.
2. The Investment Horizon Cost/Variance: Almgren and Chriss (2003) argue that for a principal desk engaged in an ongoing business, the correct approach is to *annualize* the cost and the variance, placing them in the context of other investment opportunities.
3. Discount Based Optimal Execution Point: Remarkably for each value of the discount received for trading the basket, there is then a *single* optimal point, independent of the risk preferences. This point corresponds to the single best value for the overall information ratio.
4. “Put Out To Bid” Call: The value of this ratio is therefore a potential tool to be used in evaluating whether to accept a certain piece of business at a certain price. In many situations the program trading business is “put out to bid”. That is, the portfolio manager contacts multiple desks about a particular portfolio.
5. Information Ratio as a Decision Tool: Each business responds with a certain bid – the discount to the fair market value required to do the business. Program trading desks often know the level of bid required to win the business, and therefore the information ratio can be used as a hurdle or an evaluation tool to decide whether or not to bid at a level to win the business.

## The Efficient Frontier Including Discount



1. Incorporating the Execution Discount Premium: This section connects the problem of liquidating a basket to the price of a basket. The aim is to eventually compute the information ration of liquidating a basket incorporating the value of the discount to the fair value received in the transaction.
2. Units of the Execution Discount: Thus, the assumption is that the trader will receive a discount of  $D$  dollars per share for a basket in the principal trade, and explicitly calculate the cash received in trading out the basket and its variance. For example, if the program trading desk were to be able to dump the entire portfolio onto the market without any market impact, it would earn a profit of  $DX$  dollars.
3. Execution Profit Under Market Impact: In general, because of the market impact, the total expected profit of a trade would be

$$E = DX - C_{PERM} - C_{TEMP}$$

and the variance is the same as stated in other publications (Almgren and Chriss (2000), Almgren (2003)).

4. Execution Profit Expectation and Variance: The expected profit and its variance can be explicitly calculated as functions of the execution time  $T$ :

$$\mathbb{E}_P[T] = \left( D - \frac{1}{2} \gamma X \right) X - \frac{k+1}{3k+1} \eta \left( \frac{X}{T} \right)^k X$$

$$\mathbb{V}_P[T] = \frac{k+1}{3k+1} \sigma^2 T X^2$$

5. Market Impact Reduction of Profit: Shortly  $\mathbb{E}_P[T]$  and  $\mathbb{V}_P[T]$  will be used to construct the information ratio of a trade, but for now some of its properties are examined. Clearly it is seen that the expected profit of a trade is its total discount  $DX$  reduced by a temporary impact and a permanent impact amount.
6. Difference between the Permanent and the Temporary Impact Costs: The effect of the permanent impact is to reduce the expected profit per share, as reflected in the size of



the discount  $D$ , by an amount equal to the portfolio size, while temporary impact is proportional to a *per share* reduction in the expected profit of  $\left(\frac{X}{T}\right)^k$ .

7. Time Horizon Dependence of Profit: It is worth noting the dependence of  $\mathbb{E}_P[T]$  and  $\mathbb{V}_P[T]$  on  $T$ . Short liquidation times

$$T \rightarrow 0$$

correspond to

$$\mathbb{E}_P[T] \rightarrow -\infty$$

and

$$\mathbb{V}_P[T] \rightarrow 0$$

All profit is dissipated in impact costs, but no variance is incurred.

8. Intuitive Interpretation of Long Times: Long times

$$T \rightarrow \infty$$

correspond to

$$\mathbb{E}_P[T] = \left(D - \frac{1}{2}\gamma X\right)X$$

and

$$\mathbb{V}_P[T] \rightarrow \infty$$



Temporary impact costs are avoided completely by essentially holding the portfolio forever, but at the expense of any certainty of profit.

9. Elimination of Permanent Impact Costs: In principle, if the portfolio were held forever without trading, then the permanent impact costs will also be avoided:

$$T = \infty$$

is not the same as

$$T \rightarrow \infty$$

10. The Zero Net Profit Trade: Assuming that the discount is at least enough to compensate for the permanent impact, there is an intermediate point at which

$$\mathbb{E}_P[T] = 0$$

- the zero profit trade; impact costs are exactly compensated by the discount on average, but risk is taken to achieve this.

11. Applying Customized Mean Variance Objective: The previous work by Almgren and Chriss (2000) focused on various ways to draw the expectation cost expectation/variance combination frontier, in order to maximize either a mean-variance criterion  $\mathbb{E}_P[T] + \lambda_u \mathbb{V}_P[T]$  or a value-at-risk measure  $\mathbb{E}_P[T] + \lambda_v \sqrt{\mathbb{V}_P[T]}$  for a single trade in isolation. The next step is to consider this trade as part of an ongoing business.

## Performance Measures

1. Basket Specific Optimal Liquidation Time: In this section the information ratio of a single trade is determined assuming a given discount of  $D$  dollars a share. First



observe that the above analysis did not take into account the fact that different baskets will have different optimal liquidation times.

2. Comparison across Different Principal Bids: If the principal bids are to be considered in the context of ongoing business in relation to multiple investment opportunities, then the expected profit per trade must be viewed in units that are comparable across different optimal times.
3. Per Basket Annualized Expected Return: This is done by directly annualizing the expected return by the expected amount of time it takes to liquidate substantially all of the basket, as determined by the characteristic time of the trade  $T$ . If a positive profit can be made, i.e., if

$$\mathbb{E}_P[T] > 0$$

then the trader prefers a shorter liquidation time to a longer liquidation time, other things being equal.

## Annualization

1. Characteristic Time as the Investment Horizon: In order to annualize the expected profit and its variance, it is assumed that the entire invested capital becomes available for re-investment after one characteristic time  $T$ .
2. Intermittent Release of Invested Capital: In fact, liquidation is a continuous process; some capital is available immediately, and recovery of the full capital formally requires an infinite amount of time. Nonetheless  $T$  is a reasonable average value, and it is the simplest way to compare different trajectories.
3. Annualized Expected Return and Variance: Assuming that  $T$  is measured in years, annualizing is simply a matter of dividing by  $T$ . The expectation and the variance per year of trading is



$$\frac{\mathbb{E}_P[T]}{T} = \frac{(D - \frac{1}{2}\gamma X)X}{T} - \frac{k+1}{3k+1}\eta\left(\frac{X}{T}\right)^{k+1}$$

$$\frac{\mathbb{V}_P[T]}{T} = \frac{k+1}{3k+1}\sigma^2 X^2$$

4. Annualization of the Discount and the Permanent Costs: The annualized expectation is composed of two terms. The first term is the average rate at which the discount payment  $D$  is accepted, reduced by the cost of the permanent impact; since this is a fixed amount per portfolio, it is increased by rapidly trading.
5. Annualization of the Temporary Costs: The second term is the impact cost incurred by trading at a constant rate  $\frac{X}{T}$  adjusted by a numerical coefficient to account for the non-linear shape of the trajectory.
6. Annualized Variance Independent of the Liquidation Time: Note that the annualized variance is independent of the liquidation time; this can be interpreted as saying that in the course of repeated execution, one is always invested in the market by the same amount on average.
7. Impact on the Efficient Frontier: This has an important implication. That is that if the efficient frontier is re-cast in terms of annualized expectation and annualized variance, it collapses to a single point.
8. Almgren Chriss (2003) Efficient Frontier Illustration: As illustrated in Almgren and Chriss (2003) the feasible region collapses into a half-infinite vertical line and the frontier itself has collapsed into a single point. This is a direct consequence of the annualized variance being independent of the trading time.
9. The Mandatory Capital Market Line Point: The striking consequence of this is that any measure of risk-adjusted profitability as constructed from the tangent line on the curve, regardless of the functional form or the parameter values, will pass through the highest point on this line.
10. The corresponding Optimal Trading Time: This means that, in particular, there is a unique best way to trade for any reasonable risk-adjusted return measure *regardless*



of any particular risk-reward preferences, and is found simply by finding the value of  $T$  that maximizes  $\frac{\mathbb{E}_P[T]}{T}$ . This gives

$$T_{OPT} = \left[ \frac{(k+1)^2}{3k+1} \right]^{\frac{1}{k}} \frac{\eta^{\frac{1}{k}} X}{\left( D - \frac{1}{2} \gamma X \right)^{\frac{1}{k}}}$$

11. Quasi Universal Optimal Trading Time: To emphasize the point made once more,  $T_{OPT}$  is the parameter representing *the* optimal trading strategy across a broad spectrum of possible risk-adjusted return measures. It is independent of the risk/reward preferences, but depends on the discount  $D$ . The aim next is to define and evaluate a particular risk-adjusted return measure.

## Definition of the Information Ratio

1. Mathematical Definition of Information Ratio: Almgren and Chriss (2003) define and compute the *information ratio* of a trade, incorporating the effect of the discount  $D$  received in the transaction. For a given characteristic time  $T$  the information ratio with respect to  $T$  represents the annualized risk-adjusted expected profit that may be achieved by trading along a trajectory with parameter  $T$

$$I(T) = \frac{\frac{\mathbb{E}_P[T]}{T}}{\sqrt{\frac{\mathbb{V}_P[T]}{T}}}$$

2. Risk Adjusted Basket Return: Note that  $I(T)$  is the risk-adjusted return for a basket implicitly assuming a discount  $D$  received for the trade and a trading time parameter  $T$ .



3. Estimating the Maximal Information Ratio: The question is, for which  $T$  is  $I(T)$  maximal? The answer to this is  $T_{OPT}$ , which can be substituted into the expression for  $I(T)$  to get

$$I_{MAX} = \frac{(3k+1)^{\frac{k+2}{2k}} \left(D - \frac{1}{2}\gamma X\right)^{\frac{k+1}{k}}}{(k+1)^{\frac{3k+4}{2k}} \eta^{\frac{1}{k}} X \sigma}$$

4. Alternative Interpretation of  $\mathbb{E}_P[T]$  and  $\mathbb{V}_P[T]$ : Since the numerator and the denominator in the definition of  $I$  are both proportional to the portfolio size, it would be equivalent to considering  $\mathbb{E}_P[T]$  and  $\sqrt{\mathbb{V}_P[T]}$  above as *percentage* return and risk.
5. Units of the Information Ratio: Thus, this quantity allows comparison of baskets and other investment opportunities of arbitrary size. It has units of  $year^{-\frac{1}{2}}$  and thus should be compared only with other annualized measures.

## Applications of the Information Ratio

1. Discount Level Implied Information Ratio: The two main applications of information ratio can now be stated as answers to two questions. First, for a given level of discount than can be demanded for the trade, what is the information ratio of the basket?
2. Information Ratio Implied Discount Hurdle: Second, for a given information ratio *hurdle*, what minimum discount must be demanded in order to clear it?
3. Discount Implied Maximum Information Ratio: The answer to the first question is simply  $I_{MAX}$ , the information ratio of the basket assuming a discount of  $D$ . Put a different way



$$I_{MAX} = \frac{(3k+1)^{\frac{k+2}{2k}} (D - \frac{1}{2}\gamma X)^{\frac{k+1}{k}}}{(k+1)^{\frac{3k+4}{2k}} \eta^{\frac{1}{k}} X \sigma}$$

shows that given a discount  $D$  that a desk can demand for a trade, it will yield a peak information ratio  $I_{MAX}$  which can then be used to determine if the trade clears a particular hurdle.

4. Information Ratio Implied Minimum Discount: The second question may be answered by simply inverting

$$I_{MAX} = \frac{(3k+1)^{\frac{k+2}{2k}} (D - \frac{1}{2}\gamma X)^{\frac{k+1}{k}}}{(k+1)^{\frac{3k+4}{2k}} \eta^{\frac{1}{k}} X \sigma}$$

to yield

$$D_{MIN} = \frac{1}{2}\gamma X \left[ \frac{(k+1)^{\frac{3k+4}{2k}}}{(3k+1)^{\frac{k+2}{2k}}} X \eta^{\frac{1}{k}} \sigma I_{HURDLE} \right]^{\frac{k}{k+1}}$$

The expression for  $D_{MIN}$  above gives the maximum that a desk can be bid for a given basket while still clearing the minimum information ratio threshold of  $I_{HURDLE}$

5. Power Law Process Impact Table:

<b><math>k</math></b>	<b><math>T</math></b>	<b><math>I_{MAX}</math></b>
$\frac{1}{2}$	$0.810 \frac{\eta^2 X}{\tilde{D}^2}$	$1.063 \frac{\tilde{D}^3}{\eta^2 X \sigma}$
1	$\frac{\eta X}{\tilde{D}}$	$0.707 \frac{\tilde{D}^2}{\eta X \sigma}$



2	$1.134 \frac{\eta^{\frac{1}{2}} X}{\tilde{D}^{\frac{1}{2}}}$	$0.449 \frac{\tilde{D}^{\frac{3}{2}}}{\eta^{\frac{1}{2}} X \sigma}$
---	--	---

6. Power Law Process Impact – Legend: The table above shows the optimal trading time and maximum information ratio for three different values of the market exponent  $k$ .  $\tilde{D}$  is computed from

$$\tilde{D} = D - \frac{1}{2} \gamma X$$

the discount reduced by the anticipated permanent impact costs.

7. Execution Time/IR Exponent Dependence: The table above presents the specific forms of these expressions for a few particular and important values of  $k$ . Although the analytical expressions above are complex, for a specific choice of  $k$  they reduce to simple numerical coefficients.
8. Execution Time/IR Discount Dependence: What is particularly noteworthy is the relationship between the price of the principal basket – as embodied in the discount to fair – and both the information ratio and the optimal time for liquidation. It depends on the market impact functions assumed, and can be quite sensitive to small movements.
9. Example - BARRA Market Impact Exponent: For example, for the BARRA model

$$k = \frac{1}{2}$$

the maximum information ratio increases as the *cube* of the discount, after allowing for permanent impact, and the optimal time decreases as the square of the discount.

10. Risk-Adjusted Profit Discount Dependence: One interpretation of these results is that small changes in the price of the principal bid, expressed as cents per share discount to fair value, can have significant impact on both the risk-adjusted profitability of the



trade and the time it takes to liquidate the trade. For instance, a basket that commands  $2.5c$  per share discount to fair is twice as profitable on a risk-adjusted basis versus one that commands a  $2c$  per share discount.

## References

- Almgren, R. F., and N. Chriss (2000): Optimal Execution of Portfolio Transactions *Journal of Risk* **3** (2) 5-39.
- Almgren, R. F. (2003): Optimal Executions with Nonlinear Impact Functions and Trading-Enhanced Risk *Applied Mathematical Finance* **10** (1) 1-18.
- Almgren, R. F., and N. Chriss (2003): Bidding Principles *Risk* **16** (6) 97-102.
- NYSE (2002): Press Release.



## Bayesian Trading with a Daily Trend

### Overview, Motivation, and Synopsis

1. Parametric Estimation using Updated Information: Standard models of algorithmic trading neglect the presence of a daily trend. Almgren and Lorenz (2006) construct a model in which the trader uses information from the observations of price evolution during the day to continuously update his estimate of other traders' target sizes and directions.
2. Constraint Based Optimal Trajectory Generation: The trader uses this information to determine an optimal trade schedule to minimize total expected costs of trading, subject to sign constraints, e.g., never buy as part of a sell program.
3. Dynamic Strategies using Projected Cost: It is argued that these strategies are determined using very simple dynamic reasoning – at each moment they assume that the current condition will last until the end of trading – they are in fact globally optimal strategies as would be determined by dynamic programming.

### Introduction and the Associated Literature

1. Market Information Based Learning Updates: The work of Almgren and Lorenz (2006) presents a model for price dynamics and optimal trading that explicitly includes the daily trend and the trader's attempt to learn the targets of other market participants.
2. Drawbacks of Current Approaches: This is in contrast to most current models of optimal trading strategies that view time as an undifferentiated continuum, and other traders as a collection of random noise sources. This approach has two primary motivations.



3. Incorporating the Explicit Daily Trend: The first set of motivations is the academic articles by Brunnermeier and Pedersen (2005) and Carlin, Lobo, and Viswanathan (2007). In these articles, institutional trading has an explicit daily cycle, based on the assumption that at the beginning of each day, each informed market participant, or institutional investor, is given an exogenously specified trade target.
4. Targets of the Informed Traders: These participants know the targets of the other informed traders, and they must decide whether to cooperate with their peers so as to not lose value to the uninformed traders, or whether to compete and take value from their peers.
5. Dynamic Estimation of Peer Targets: The novel feature of this market is that the participants do not know each other's targets, but must guess them by observing the prices throughout the day. It is taken for granted that informed participants will use all available information to compete with each other.
6. Dynamic Determination of Execution Trajectory: The second set of motivations is the popularity of execution algorithms that adapt to the changes in the prices of the asset being traded, either by accelerating execution when the prices move in the traders' favor, or conversely.
7. Momentum and/or Mean Reversion: Although these optimal trade models may be handled by introducing various forms of risk aversion (Kissell and Malamut (2006), Almgren and Chriss (2007)), the most common justification for them is a belief in mean reversion or momentum of the asset price.
8. Underlying Institutional Investor Drift Factor: The model introduced in this chapter may be interpreted as one plausible way to model price momentum. There is an underlying drift factor, caused by the net positions being executed by the other institutional investors.
9. Daily Institutional Trader Price Momentum: This factor is approximately constant throughout the day because the other traders execute across the entire day. Thus, price increases in the early part of the day suggest that this factor is positive, which suggests that the prices will continue to increase throughout the day.



10. Short Term Price Change Correlation: This is different from the short-term momentum model in which the price change across one short period of time is correlated with the price change across the preceding period; most empirical evidence shows that such correlation is weak if it exists at all.
11. Incorporating the Daily Price Momentum: The strategies presented in this chapter exploit this momentum to minimize the expected value of the trading costs, somewhat in the spirit of Bertsimas and Lo (1998), except that because the focus is on long term momentum, higher gains can be obtained.
12. Origin of the Daily Trend: The daily trend is an essential feature of this model. Large institutional participants make investment decisions overnight and implement them through the following day.
13. Trend Period vs. Implementation Horizon: Within each day morning is different from the afternoon, since an intelligent trader will spend the early hours collecting information about the targets of the other traders, and will use this information to trade in the rest of the day.
14. Random Nature of Trade Decisions: By contrast, in the market that is implicitly assumed by most models, trade decisions are made at random times, and trade programs have random durations, with no regard to the daily trends. Thus, if one observes a buy pressure from the market as a whole, one has no reason to believe that this pressure will last more than a short term. From the point of view of optimal trading, price motions are purely random.
15. Constraint on the Trade Direction: In addition, the very important feature of constraints on the trade direction is incorporated; the trader must never sell as part of a buy program even if this yields lower expected costs – or even expected profit – because of the anticipated negative drift in the price.
16. Reasons for the Constraint Imposition: This is for two reasons. First the point of view of a broker/dealer executing an agency trade for a client is taken. Second, the bid/offer spread and the other fixed costs are neglected, which greatly reduce the profitability of such reversing strategies. These adaptive strategies simply sift the buying or selling from one period to another.



17. Binding Nature of the Constraint: This constraint is often binding, and globally affects the structure of optimal strategies. In many cases it leads to the determination of an optimal end time for trading, and sometimes directs the strategy to stop completely for a finite period in the middle of the execution.
18. Bayesian Learning of Institutional Direction: In the section below a model is presented of Brownian motion with a drift whose distribution is updated continuously using Bayesian inference.
19. Best Estimate Based Optimal Trajectory: Subsequently, optimal strategies are presented which, surprisingly, can be computed by determining a “static” optimal Trajectory at each moment, assuming that the best parameter estimates as of that moment will persist through the end of the trading period.

## Price Model Using Bayesian Update

1. Arithmetic Brownian Motion Price Dynamics: Trading in a single asset is considered, whose price  $S(t)$  obeys an arithmetic random walk

$$S(t) = S_0 + \alpha t + \sigma B(t)$$

for

$$t \geq 0$$

where  $B(t)$  is a standard Brownian motion,  $\sigma$  is the absolute volatility, and  $\alpha$  is the drift. In the presence of intra-day seasonality,  $t$  is interpreted as a volume time relative to a historical profile.

2. Origin of Volatility - Uninformed Traders: The interpretation of volatility is that it comes from the activity of the “uninformed” traders, whose average behavior can be predicted reasonably well. Mathematically, the value of  $\sigma$  is assumed to be known



precisely – for a Brownian process  $\sigma$  can be estimated arbitrarily precisely from an arbitrarily short observation of the process.

3. Origin of Drift - Institutional Traders: The drift is interpreted as coming from the activities of the other institutional traders, who have made trade decisions before the market opens, and who expect to execute these trades throughout the day. If these decisions are in the aggregate weighted to buys, then this will cause a positive price pressure and an upwards drift – conversely for overall selling.
4. Trade Direction Based Drift Estimate: No knowledge of the net direction of the trade estimates is presumed, but inferred by observing the prices. It is implicitly assumed that the traders are using VWAP-like strategies rather than the pure arrival price, so that their trading is not “front-loaded”. This assumption is questionable; if their strategies are front-loaded, then the drift coefficient would vary throughout the day.
5. Drift Belief - Mean and Confidence: Thus, the drift  $\alpha$  is assumed constant throughout the day, but its value is unknown. At the beginning of the day the prior belief

$$\alpha \sim \mathcal{N}(\bar{\alpha}, v^2)$$

will be updated using price observations throughout the day.

6. “Frequentist” Volatility vs. “Bayesian” Drift: There are two sources of randomness in the problem – the continuous Brownian motion representing the uninformed traders, and the single drift coefficient representing the constant trading of the large traders.

## Bayesian Inference

1. Drift Estimate from Realized Price: Intuitively as the trader observes the prices from the beginning of the day onwards, he/she starts to get a feel for the day’s overall flow. Mathematically the stock price trajectory  $S(\tau)$  is known for

$$0 \leq \tau \leq t$$



In fact, all the information about the drift comes from the final value  $S(t)$

2. Bayesian Formulation of the Price Evolution: Conditional on the value of  $\alpha$  the distribution of  $S(t)$  is

$$S(t) - S_0 \sim \mathcal{N}(\alpha t, \nu^2 t)$$

The unconditional distribution can be found after some calculation as

$$S(t) - S_0 \sim \mathcal{N}(\bar{\alpha}t, [\sigma^2 + \nu^2 t]t)$$

3. The Posterior Conditional Drift Distribution: The Bayes' rule is then used:

$$\text{Prob} (\alpha|S(t)) = \frac{\text{Prob} (S(t)|\alpha) \cdot \text{Prob} (\alpha)}{\text{Prob} (S(t))}$$

to obtain the posterior conditional distribution

$$\alpha \sim \mathcal{N}\left(\frac{\bar{\alpha}\sigma^2 + \nu^2[S(t) - S_0]}{\sigma^2 + \nu^2 t}, \frac{\sigma^2}{\sigma^2 + \nu^2 t}\nu^2\right)$$

conditional on  $S(t)$ .

4. Best Estimate of Mean/Variance: This represents the best estimate of the true drift  $\alpha$  as well as the uncertainty in this estimate based on the combination of the prior belief with the price information observed to time  $t$ .
5. Fully Certain Estimate of Drift: This formulation accommodates a wide variety of belief structures. If the belief in the initial information is perfect then one sets

$$\nu = 0$$



and the updated belief is always

$$\alpha = \bar{\alpha}$$

with no incremental updating.

6. Fully Uncertain Estimate of Drift: If one believes that there is no reliable prior information then

$$\nu^2 \rightarrow \infty$$

and the estimate is

$$\alpha \sim \mathcal{N} \left( \frac{S(t) - S_0}{t}, \frac{\sigma^2}{t} \right)$$

coming entirely from intra-day observations.

7. The  $t = 0$  and  $t \rightarrow \infty$  Asymptotes: For

$$t = 0$$

one has

$$S(0) = S_0$$

and the belief is just the prior. As

$$t \rightarrow \infty$$

the estimate becomes



$$\alpha \sim \mathcal{N}\left(\frac{S(t) - S_0}{t}, 0\right)$$

so much information has been accumulated that the prior belief becomes irrelevant.

## Trading and Price Impact

1. The Order Size and Horizon: The trader has an order of  $X$  shares which begins at time

$$t = 0$$

and must be completed by the time

$$t = T < \infty$$

For concreteness it is supposed that

$$X > 0$$

which is interpreted as a buy order.

2. Trade Rate and Trading Trajectory: A *trading trajectory* is a function  $x(t)$  with

$$x(0) = X$$

and

$$x(T) = 0$$

representing the number of shares to buy at time  $t$ . The corresponding *trading rate* is



$$v(t) = -\frac{dx(t)}{dt}$$

It shall be required that

$$v(t) \geq 0$$

for all  $t$  so that a program never sells as part of the buy order. Together with the endpoint constraints this requires

$$0 \leq x(t) \leq X$$

but it may also be binding in the interior of the region.

3. Linear Temporary Market Impact Function: A linear temporary market impact function is used for simplicity, although the empirical work of Almgren, Thum, Hauptmann, and Li (2005) suggests a concave function. The actual execution price is

$$\tilde{S}(t) = S(t) + \eta v(t)$$

where

$$\eta > 0$$

is the coefficient of temporary market impact.

4. Execution Trajectory Implementation Shortfall:  $\mathcal{C}$  is the total cost of executing the buy program relative to the initial value

$$\mathcal{C} = \int_0^T \tilde{S}(t)v(t)dt - XS_0 = \sigma \int_0^T x(t)dB(t) + \eta \int_0^T v^2(t)dt + \alpha \int_0^T x(t)dt$$



5. Deterministic and Random Cost Components: Here  $\alpha$  is the true drift, and this determines cost, whether or not its true value is known.  $C$  is a random variable, both because  $S(t)$  is random, and because the optimal trading trajectory  $v(t)$  may be adapted to  $S$ .

## Optimal Trading Strategies

1. Classic Mean-Variance Risk Aversion: This section addresses the question of what trading strategies are optimal given the above model for price evolution and market impact. In the classic arrival price framework of Almgren and Chriss (2000) trajectories are determined by a trade-off between market impact and aversion to risk caused by volatility.
2. Balance between Slow/Fast Trading: The trader wants to complete the trade quickly to reduce exposure to price volatility; He or she wants to trade slowly to reduce the cost of market impact. The optimal trajectory is determined as a balance between these two effects, parametrized by a coefficient of risk aversion.
3. Optimal Cost Adaptive Trading Strategies: Risk-averse trading strategies can behave strangely in time even in the classic mean variance framework (Almgren and Lorenz (2007)) depending on the precise formulation of the mean-variance trade-off.
4. Complication Introduced by the Drift Variance: In this case the problem is complicated by the need to account for the variance in the estimate of  $\alpha$ . Almgren and Lorenz (2006) claim to have obtained partial solutions for the risk-averse problem, but the resulting complexity obscures the underlying structure.
5. Neglecting the Mean Variance Risk Aversion: To focus on the drift, which is the most important new aspect of this problem, risk aversion is neglected here; only the expectation of the trading cost is sought to be minimized.
6. Cost associated with the Drift: That is, it is assumed that the pressure to complete the trade rapidly comes primarily by a desire to capture the price motion expressed by the



drift  $\alpha$ , and it is this effect that must be balanced against the desire to reduce the impact costs by trading slowly.

7. Positive Baseline Drift Assumption: To support this description it is generally supposed that the original buy decision was made because of the traders' belief that

$$\bar{\alpha} > 0$$

Thus, it is expected

$$\alpha > 0$$

in

$$\mathcal{C} = \int_0^T \tilde{S}(t)v(t)dt - XS_0 = \sigma \int_0^T x(t)dB(t) + \eta \int_0^T v^2(t)dt + \alpha \int_0^T x(t)dt$$

and the term  $\alpha \int_0^T x(t)dt$  is a positive cost. It may be that the true value has

$$\alpha < 0$$

or that the intermediate price movements result in the formation of a negative estimate.

8. Hard Trade Completion Time  $t = T$ : Because the point of view is that of a broker/dealer executing an agency trade, it shall always be required that the trade be completed by

$$t = T$$

unless the instructions are altered.



9. Conditional Expectation of Unrealized Cost: For any deterministic trajectory  $x(t)$  specified at

$$t = 0$$

$\mathcal{C}$  is a Gaussian variable. Conditional on the true value of  $\alpha$  it has the expected value

$$\mathbb{E}[\mathcal{C}] = \eta \int_0^T v^2(t) dt + \alpha \int_0^T x(t) dt$$

10. The Bayesian Estimate for  $\alpha$ : From

$$\alpha \sim \mathcal{N}\left(\frac{\bar{\alpha}\sigma^2 + v^2[S(t) - S_0]}{\sigma^2 + v^2 t}, \frac{\sigma^2}{\sigma^2 + v^2 t}v^2\right)$$

conditional on  $S(t)$  the best estimate at time  $t$  for the value of  $\alpha$  is

$$\alpha_*(t, S) = \frac{\bar{\alpha}\sigma^2 + v^2[S(t) - S_0]}{\sigma^2 + v^2 t}$$

where

$$S \equiv S(t)$$

11. Bayesian Estimate of Unrealized Cost: Because the expectation

$$\mathbb{E}[\mathcal{C}] = \eta \int_0^T v^2(t) dt + \alpha \int_0^T x(t) dt$$



conditional on  $\alpha$  is linear  $\alpha$  one may substitute the expected value of  $\alpha_*$  to see that, conditional on the information available at time  $t$  the expected cost of the remaining program is

$$\mathbb{E}[t, x(t), S, \{x(\tau)\}] = \eta \int_t^T v^2(\tau) d\tau + \alpha_*(t, S) \int_t^T x(\tau) d\tau$$

12. Nomenclature - Description of the Terms: On the left  $t$  is the current time,  $x(t)$  is the number of shares currently remaining to buy,  $S$  is the current price, and  $\{x(\tau)\}$  denotes the liquidation strategy that will be used on the remaining time

$$t \leq \tau \leq T$$

13. Strategy Objective - Minimizing Transaction Cost: The trading goal is to choose the remaining strategy to minimize this expected cost, i.e., determine  $x(\tau)$  for

$$t \leq \tau \leq T$$

so that

$$\min_{\{x(\tau)\}} \mathbb{E}[t, x(t), S, \{x(\tau)\}]$$

14. Invariance of the Drift Estimate: In computing this solution, it is assumed that the drift estimate  $\alpha_*(t, S)$  does not change during the interval

$$t \leq \tau \leq T$$

In fact, it will change as new price is obtained.



15. Dynamically Recomputed Unrealized Trajectory Cost: The actual strategy will only use the instantaneous trade rate of this trajectory, continuously responding to price information. This is equivalent to following the strategy only for a very small-time interval  $\Delta t$  then re-computing. Thus, the strategy is highly dynamic.
16. Equivalence with Full Dynamic Optimization: It shall be argued that the trajectory thus determined is the true optimum strategy that would be computed by a full dynamic optimization. Loosely speaking this will be because the expected value of future updates is zero, and thus they do not change the strategy of a risk-neutral trader.

## Trajectory by Calculus of Variations

1. Trajectory Perturbation Fixed at the End-points: A small perturbation of the path

$$x(\tau) \mapsto x(\tau) + \Delta x(\tau)$$

is considered for

$$t \leq \tau \leq T$$

Since  $x(\tau)$  is fixed at

$$\tau = t$$

and

$$\tau = T$$

this perturbation must have



$$\Delta x(t) = \Delta x(T) = 0$$

2. The Corresponding Bayesian Cost Impact: The associated trade rate perturbation is

$$\Delta v(\tau) = -\Delta x'(\tau)$$

and the perturbation in cost – assuming that  $x(\tau)$  and  $\Delta x(\tau)$  are twice differentiable – is

$$\begin{aligned}\Delta \mathbb{E}[t, x(t), S, \{x(\tau)\}] &= \eta \int_t^T 2v(\tau)[\Delta v(\tau)]d\tau + \alpha_*(t, S) \int_t^T [\Delta x(\tau)]d\tau \\ &= \int_t^T \{-2\eta x''(\tau) + \alpha_*(t, S)\}[\Delta x(\tau)]d\tau\end{aligned}$$

3. Cost Optimized Trajectory - Admissibility Condition: Here

$$\alpha_* \equiv \alpha_*(t, S)$$

is the best available drift estimate using information available at time  $t$  which we assume is constant for

$$t \leq \tau \leq T$$

If  $x(\tau)$  is an optimal solution then there must not exist any admissible  $\Delta x(\tau)$  that gives

$$\Delta \mathbb{E}[t, x(t), S, \{x(\tau)\}] > 0$$



4. Unconstrained Trajectories - The Holdings ODE: For now, the sign constraints on  $x'(\tau)$  is neglected. The  $\Delta x(\tau)$  may have either positive or negative values independently for each  $\tau$  and optimizing  $x(\tau)$  must satisfy the ordinary differential equation (ODE)

$$x''(\tau) = \frac{\alpha_*}{2\eta}$$

$$t \leq \tau \leq T$$

5. Unconstrained Trajectory - The Holding Solution: The solution to this equation that satisfies the boundary conditions is

$$x(\tau) = \frac{T-\tau}{T-t} x(t) - \frac{\alpha_*}{4\eta} (\tau-t)(T-\tau)$$

$$t \leq \tau \leq T$$

and the corresponding instantaneous trade rate is

$$v(t, x) = -x'(\tau)|_{\tau=T} = \frac{x(t)}{T-t} + \frac{\alpha_*}{4\eta} (T-t)$$

as a function of time and shares remaining.

6. Unconstrained Trajectory Holdings Constraint Violation: This solution may violate the constraints; if  $\alpha_*$  is large then the quadratic term in

$$x(\tau) = \frac{T-\tau}{T-t} x(t) - \frac{\alpha_*}{4\eta} (\tau-t)(T-\tau)$$

$$t \leq \tau \leq T$$



may cause  $x(\tau)$  to dip below zero, which would cause  $v(t)$  in

$$v(t, x) = -x'(\tau)|_{\tau=T} = \frac{x(t)}{T-t} + \frac{\alpha_*}{4\eta}(T-t)$$

to become negative.

7. Unconstrained Trajectory - The Holdings Component: The unconstrained solution is the sum of two parts. The first piece is proportional to  $x(t)$  and represents the linear (VWAP) liquidation of the current position; it is the optimal strategy to reduce the expected impact costs with no risk aversion.
8. Unconstrained Trajectory - Holdings Drift Component: The second piece is independent of  $x(t)$  and would exist even if the trader has no initial position. Just as in the solutions of Bertsimas and Lo (1998), this second piece is effectively a proprietary trading strategy superimposed on liquidation.
9. Unconstrained Trajectory - Relative Component Contribution: The magnitude of this strategy, and hence the possible gains, are determined by the ratio between the expected drift and the liquidity coefficient. Imposition of this constraint will couple these pieces together.
10. Constrained Trajectories - Consequences of Violation: If the constraint becomes binding then it is no longer clear that the integration by parts procedure used to derive

$$\begin{aligned}\Delta \mathbb{E}[t, x(t), S, \{x(t)\}] &= \eta \int_t^T 2v(\tau)[\Delta v(\tau)]d\tau + \alpha_*(t, S) \int_t^T [\Delta x(\tau)]d\tau \\ &= \int_t^T \{-2\eta x''(\tau) + \alpha_*(t, S)\}[\Delta x(\tau)]d\tau\end{aligned}$$

is valid.

11. Constrained Trajectories – Non-Smooth Edges: For example, if a trajectory that crosses the axis



$$x = 0$$

is simply clipped to satisfy

$$x \geq 0$$

then the derivative will be discontinuous. A more refined use of the calculus of variations gives the additional condition that  $v(\tau)$  must be continuous though not differentiable.

12. Constrained Trajectories - Smoothing the Edges: Thus, when solutions meet the constraint, they must do so smoothly. Solutions are obtained by combining the ODE's

$$x''(\tau) = \frac{\alpha_*}{2\eta}$$

$$t \leq \tau \leq T$$

in the regions of smoothness, with “smooth pasting” conditions at the boundary points.

13. Constrained Trajectories - The “Critical” Drift: The results may be summarized as follows. There is a critical drift value  $\alpha_c$  such that if

$$|\alpha_*| \leq \alpha_c$$

then the constraint is binding. The solution is the one given in

$$x(\tau) = \frac{T-\tau}{T-t} x(t) - \frac{\alpha_*}{4\eta} (\tau-t)(T-\tau)$$

$$t \leq \tau \leq T$$



and

$$v(t, x) = -x'(\tau)|_{\tau=T} = \frac{x(t)}{T-t} + \frac{\alpha_*}{4\eta}(T-t)$$

14. Constrained Trajectories – Super-critical Drift Horizon: If

$$\alpha_* > \alpha_c$$

the solution is still

$$x(\tau) = \frac{T_* - \tau}{T_* - t} x(t) - \frac{\alpha_*}{4\eta} (\tau - t)(T_* - \tau)$$

$$t \leq \tau \leq T_*$$

and

$$v(t, x) = -x'(\tau)|_{\tau=t} = \frac{x(t)}{T_* - t} + \frac{\alpha_*}{4\eta}(T_* - t)$$

but with a shortened end-time

$$T_* < T$$

determined by

$$T_* - t = \sqrt{\frac{4\eta x(t)}{\alpha_*}}$$



15. Constrained Trajectories Supercritical Drift Value: The values for  $T_*$  above is determined so that

$$x'(T_*) = x(T_*) = 0$$

The threshold value  $\alpha_C$  is the value of  $\alpha_*$  for which

$$T_* = T$$

$$\alpha_C(x(t), T - t) = \frac{4\eta x(t)}{(T - t)^2}$$

16. Constrained Trajectories Subcritical Drift: If

$$\alpha_* < -\alpha_C$$

then the solution is one of

$$x(\tau) = \frac{T - \tau}{T - t_*} x(t) - \frac{\alpha_*}{4\eta} (\tau - t_*)(T - \tau)$$

$$t_* \leq \tau \leq T$$

and

$$v(t, x) = -x'(\tau)|_{\tau=t_*} = \frac{x(t_*)}{T - t} + \frac{\alpha_*}{4\eta} (T - t)$$

except that trading does not begin until a starting time  $t_*$  determined by



$$T - t_* = \sqrt{\frac{4\eta x(t)}{-\alpha_*}}$$

This value is determined so that

$$x'(t_*) = 0$$

and

$$x(t_*) = x(t)$$

The threshold value  $\alpha_c$  is the value  $-\alpha_*$  for which

$$t_* = t$$

17. Constrained Trajectories - Illustration: Almgren and Lorenz (2006) illustrate the constrained solutions  $x(\tau)$  starting at time  $t$  with shares  $x(t)$  and drift estimate  $\alpha$ . For

$$\alpha > 0$$

the trajectories go below the linear profile to reduce the expected purchase cost. As shown in the shaded region in the illustration for

$$|\alpha| \leq \alpha_c$$

the constraint is not binding. At

$$\alpha = \alpha_c$$

the solutions become tangent to the line



$$x = 0$$

at

$$\tau = T$$

and for larger values they hit

$$x = 0$$

with zero slope at

$$\tau = T_* < T$$

For

$$\alpha < -\alpha_c$$

trading does not begin until

$$\tau = t_* > t$$

18. Bayesian Drift Trading Rates - Summary: Thus, the overall trade rate may be summarized as



$$v(t, x, S) = \begin{cases} 0 & \alpha_* < \alpha_c \\ \frac{x(t_*)}{T - t} + \frac{\alpha_*}{4\eta}(T - t) & |\alpha_*| < \alpha_c \\ \frac{x(t)}{T_* - t} + \frac{\alpha_*}{4\eta}(T_* - t) = \sqrt{\frac{\alpha_* x(t)}{\eta}} & \alpha_* > \alpha_c \end{cases}$$

where

$$\alpha_* = \alpha_*(t, S(t))$$

and is given as

$$\alpha_*(t, S) = \frac{\bar{\alpha}\sigma^2 + \nu^2[S(t) - S_0]}{\sigma^2 + \nu^2 t}$$

This is the Bayesian adaptive strategy; it is a specific formula for the instantaneous trade as a function of price, time, and shares remaining.

19. ODE Nature of Optimal Trajectories: Since

$$\Delta x = -v\Delta t$$

this gives an ordinary differential equation for the trajectory  $x(t)$  with a stochastic element due to the presence of  $S(t)$ . It is not a stochastic differential equation since  $\Delta B$  only appears in  $\Delta S$  and not in  $\Delta x$ . Thus  $x(t)$  will have a first time derivative but not a second derivative.

## Optimality of the Bayesian Adaptive Strategy



1. Joint Local and Global Optimality: The Bayes adaptive strategy for  $v(t, x, S)$  is locally optimal in the sense that at any intermediate time all the new available information is used to re-compute the trajectory for the remainder as though the same estimate for the drift will be used until the end of the trading. Since updates to the estimate are expected, it is not obvious if this is the true optimal strategy.
2. Stochastic Optimal Control using HJB: In the next section, using the methods of stochastic optimal control, a Hamilton-Jacobi-Bellman (HJB) partial differential equation (PDE) is formulated for the value function of the corresponding dynamic program.
3. Simplification under the Unconstrained Case: For the unconstrained case the PDE can be solved analytically. The corresponding optimal strategy is calculated by differentiating the value function and agrees precisely with

$$v(t, x) = -x'(\tau)|_{\tau=t_*} = \frac{x(t_*)}{T-t} + \frac{\alpha_*}{4\eta}(T-t)$$

This computation verifies that the local solution is the dynamic optimal solution for the unconstrained case.

4. Gains from the Dynamic Trajectory: Furthermore, the gains due to adaptivity may be explicitly determined. At

$$t = 0$$

with initial shares  $X$  the value function for the dynamic strategy may be written as

$$E_{DYN} = E_{STAT} - \mathcal{G}$$

where

$$E_{STAT} = \frac{\eta X^2}{T} + \frac{\bar{\alpha}XT}{2} - \frac{\bar{\alpha}^2 T^3}{48\eta}$$



is the expected cost of the non-adaptive strategy determined at

$$t = 0$$

using the prior expected drift  $\bar{\alpha}$ .

5. Estimate of the Bayesian Adaptive Gains: The additional term

$$\mathcal{G} = \frac{\sigma^2 T^2}{48\eta} \int_0^1 \frac{(1 - \hat{t})^3}{(\hat{t} + \rho)^2} d\hat{t}$$

$$\rho = \frac{\sigma^2}{\nu^2 T}$$

$$\hat{t} = \frac{t}{T}$$

is the reduction in the expected cost obtained by using the adaptive Bayesian strategy.

Note that

$$\mathcal{G} > 0$$

6. Gains Independent of the Portfolio Size: The gain  $\mathcal{G}$  is independent of the portfolio size  $X$ , and thus as discussed above, represents the gains from a proprietary trading strategy super-imposed on a risk neutral liquidation profile.
7. Time Asymptote of the Gain: It can be seen that

$$\mathcal{G} \sim \mathcal{O}(T^4)$$

when  $T$  is small and



$$\mathcal{G} \sim \mathcal{O}(T^2)$$

when  $T$  is large, so the adaptivity adds very little value when applied to short term correlation. This accounts for the small gains obtained by Bertsimas and Lo (1998), as discussed by Almgren and Chriss (2000).

8. Analytic Tractability of the Constrained Case: For the constrained case the HJB equation has complicated boundary conditions, and Almgren and Lorenz (2006) are not able to determine explicit analytical solutions. However, they believe that the imposition of the constraint should not change the relationship between the static and the dynamic solutions. Thus, they believe that the dynamic solution is the dynamic optimal solution in the constrained case as well.

## Stochastic Optimal Control Treatment

1. Full Dynamic Programming Framework: This section supports the claim that the trade velocity

$$v(t, x) = -x'(\tau)|_{\tau=t_*} = \frac{x(t_*)}{T-t} + \frac{\alpha_*}{4\eta}(T-t)$$

for the adaptive strategy is in fact the optimal strategy for

$$\min_{\{x(\tau)\}} \mathbb{E}[t, x(t), S, \{x(\tau)\}]$$

For that the problem will be formulated in a full dynamic programming framework.

2. Control Variables, State Variables, and the SDE: The control variables, state variable, and the stochastic differential equations of



$$\min_{\{x(\tau)\}} \mathbb{E}[t, x(t), S, \{x(\tau)\}]$$

are given as follows:

$v \Rightarrow$  Rate of Buying

$x \Rightarrow$  Shares Remaining to Buy

$$\Delta x = -v\Delta t$$

$y \Rightarrow$  Dollars Spent so far

$$\Delta y = (s + \eta v)v\Delta t$$

$S \Rightarrow$  Stock Price

$$\Delta s = \alpha\Delta t + \sigma\Delta B$$

where

$$\alpha \sim \mathcal{N}(\bar{\alpha}, \nu^2)$$

is chosen randomly at

$$t = 0$$

3. The Initial and Final Conditions: At

$$t = 0$$



the shares are

$$x(0) = X$$

the cash is

$$y(0) = 0$$

and the initial stock price is

$$S(0) = S$$

The strategy  $v(t)$  must be adapted to the full filtration of  $B$  and must satisfy

$$x(T) = 0$$

The focus is on the unconstrained case, and therefore

$$0 \leq x(0) \leq X$$

is not required.

4. The Value Function to be Optimized: The goal is to find the control function  $v(t)$  that minimizes the amount of final dollars spent:

$$\min_{\{v(\tau)\} \text{ such that } x(T) = 0} \mathbb{E}[y(T)]$$

This is a common problem in stochastic dynamic control, and is solved by dynamic programming.



5. The Stochastic Control HJB PDE: Standard techniques lead to the Hamilton-Jacobi-Bellman (HJB) partial differential equation

$$0 = \frac{\partial u}{\partial t} + \frac{1}{2} \sigma^2 \frac{\partial^2 u}{\partial s^2} + \alpha_* \frac{\partial u}{\partial s} + \min_v \left[ \left( s \frac{\partial u}{\partial y} - \frac{\partial u}{\partial x} \right) + \eta v^2 \frac{\partial u}{\partial y} \right]$$

so that the value function

$$u(t, x, y, s) = \min_{v(t), t \leq \tau \leq T \text{ such that } x(T) = 0} \mathbb{E}[y(T)]$$

$$\alpha_* \equiv \alpha_*(t, S)$$

denotes the estimate of  $\alpha$  at time  $t$  as computed in

$$\alpha \sim \mathcal{N} \left( \frac{\bar{\alpha} \sigma^2 + v^2 [S(t) - S_0]}{\sigma^2 + v^2 t}, \frac{\sigma^2}{\sigma^2 + v^2 t} v^2 \right)$$

conditional on  $S(t)$

6. Incorporating the Optimal Trade Velocity: The optimal trade velocity is found as

$$v_*(t, x, y, s) = \frac{\frac{\partial u}{\partial x} - s \frac{\partial u}{\partial y}}{2\eta \frac{\partial u}{\partial y}}$$

and the final HJB partial differential equation for  $u(t, x, y, s)$  is

$$0 = \frac{\partial u}{\partial t} + \frac{1}{2} \sigma^2 \frac{\partial^2 u}{\partial s^2} + \alpha_* \frac{\partial u}{\partial s} - \frac{\left( \frac{\partial u}{\partial x} - s \frac{\partial u}{\partial y} \right)^2}{4\eta \frac{\partial u}{\partial y}}$$



together with the boundary condition

$$u(T, 0, y, s) = y$$

for all  $y, s$ .

7. Solution to the HJB PDE: It is straightforward to check that

$$\begin{aligned} u(t, x, y, s) &= y + xs + \eta \frac{x^2}{T-t} + \frac{1}{2} \alpha_*(y, s)x(T-t) - \frac{\alpha_*^2(t, S)}{48\eta}(T-t)^3 \\ &\quad - \int_t^T \frac{\sigma^2 v^2 (T-\tau)^3}{48\eta(\sigma^2 + \tau v^2)^2} d\tau \end{aligned}$$

satisfies the PDE

$$0 = \frac{\partial u}{\partial t} + \frac{1}{2} \sigma^2 \frac{\partial^2 u}{\partial s^2} + \alpha_* \frac{\partial u}{\partial s} - \frac{\left( \frac{\partial u}{\partial x} - s \frac{\partial u}{\partial y} \right)^2}{4\eta \frac{\partial u}{\partial y}}$$

and the boundary condition

$$u(T, 0, y, s) = y$$

for all  $y, s$ .

8. Expression for Optimal Trade Velocity: Moreover, the corresponding optimal trade velocity

$$v_*(t, x, y, s) = \frac{\frac{\partial u}{\partial x} - s \frac{\partial u}{\partial y}}{2\eta \frac{\partial u}{\partial y}}$$



which is exactly the trade velocity

$$v(t, x) = -x'(\tau)|_{\tau=t_*} = \frac{x(t_*)}{T-t} + \frac{\alpha_*}{4\eta}(T-t)$$

9. Correspondence with the Earlier Formulation: That is the Bayesian adaptive strategy is in fact the optimal strategy for the optimization problem

$$\min_{\{x(\tau)\}} \mathbb{E}[t, x(t), S, \{x(\tau)\}]$$

10. Extension to the Constrained Case: For the constrained case the optimal velocity

$$v_*(t, x, y, s) = \frac{\frac{\partial u}{\partial x} - s \frac{\partial u}{\partial y}}{2\eta \frac{\partial u}{\partial y}}$$

becomes

$$v_*(t, x, y, s) = \max\left(\frac{\frac{\partial u}{\partial x} - s \frac{\partial u}{\partial y}}{2\eta \frac{\partial u}{\partial y}}, 0\right)$$

This makes the corresponding PDE even more nonlinear, and derivation of explicit solutions is elusive.

## References



- Almgren, R. F., and N. Chriss (2000): Optimal Execution of Portfolio Transactions *Journal of Risk* **3 (2)** 5-39.
- Almgren, R. F., C. Thum, E. Hauptmann, and H. Li (2005): Equity Market Impact *Risk* **18 (7)** 57-62.
- Almgren, R. F., and J. Lorenz (2006): Bayesian Adaptive Trading with a Daily Cycle *Journal of Trading* **1 (4)** 38-46.
- Almgren, R. F., and J. Lorenz (2007): Adaptive Arrival Price, in: *Algorithmic Trading III (B. R. Bruce, editor)* **Institutional Investor** 59-66.
- Bertsimas, D., and A. W. Lo (1998): Optimal Control of Execution Costs *Journal of Financial Markets* **1** 1-50.
- Brunnermeier, M., K., and L. H. Pedersen (2005): Predatory Trading *Journal of Finance* **60 (4)** 1825-1863.
- Carlin, B. I., M. S. Lobo, and S. Viswanathan (2007): Episodic Liquidity Crises: Cooperative and Predatory Trading *62 (5)* 2253-2274.
- Kissell, R., and R. Malamut (2006): Algorithmic Decision-Making Framework *Journal of Trading* **1 (1)** 12-21.



## Smart Order Routing

### Overview

1. Definition of Smart Order Routing: *Smart Order Routing* SOR is an automated process of handling orders, aimed at taking the best available opportunity throughout a range of different trading venues (Wikipedia (2023)).
2. Liquidity Consequence of Venue Availability: The increasing number of various trading venue and the MTFs leads to a surge in liquidity fragmentation, when the stock is traded on several different venues, so the price and the amount of stock can vary among them. SOR serves to tackle liquidity fragmentation, or even benefit from it.
3. Purpose of Smart Order Routing: Smart Order Routing is performed by smart order routers – systems designed to analyze the state of various and the place orders the best available way, relying on the defined rules, configuration, and algorithms.

### Benefits and Disadvantages of Smart Order Routing

1. Advantages of Smart Order Routing: SOR provides the following benefits:
  - a. Simultaneous access to several venues
  - b. Automatic search for the best price
  - c. A good framework for usage of custom algorithms
  - d. Opportunity to get additional validation, control, and statistics
2. Drawbacks of Smart Order Routing: There are, however, some disadvantages:
  - a. Additional Latency
  - b. Additional complexity, and therefore, additional risk of loss/outage
  - c. Transparency of information, concerning transactions, for the third party

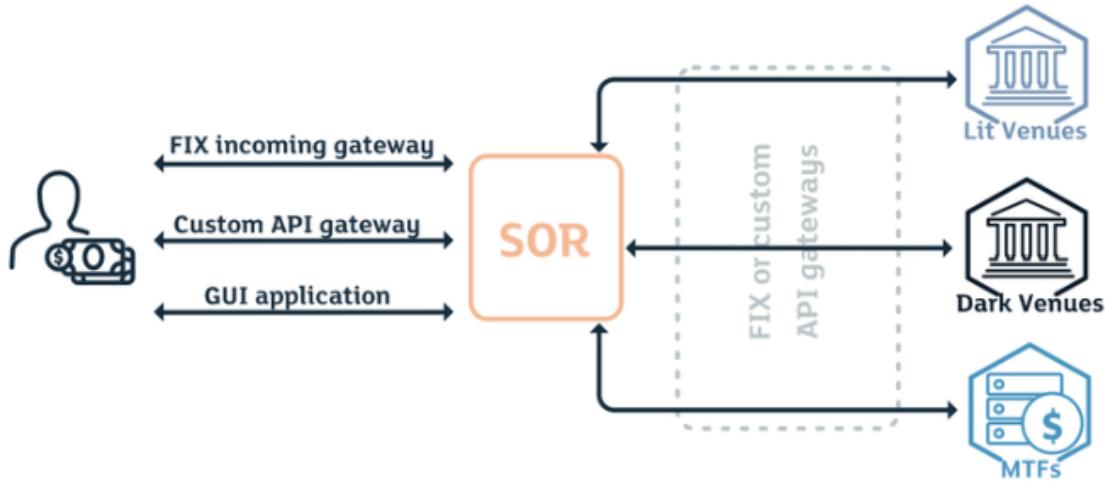


## Brief Concept

1. SOR – Base Idea and Stages: The idea of smart order routing is to scan the markets and find the best place to execute a customer's order, based on price and liquidity. Thus, smart order routing involves the following stages:
  2. Receiving Incoming Orders: This is done through different channels:
    - a. An incoming FIX gateway
    - b. An incoming gateway based on any custom protocol
    - c. A front-end
  3. Processing Orders inside SOR System: This takes into account:
    - a. Characteristics of available venues
    - b. Custom algorithms
    - c. Settings/preferences of a certain client
    - d. State of available markets/market data
  4. Handling the Venue Parameters: Venue parameters, such as average latency, commissions, and rank can be used to prioritize certain venues.
  5. Utilization of Custom Algorithms: Custom algorithms, like synthetic orders – peg, iceberg, spraying, TWAP – can be used to manage order automatically, for instance, if a specific client has certain routing preferences among several brokers, or certain rules for handling of incoming, or creation of outgoing orders.
  6. Incorporation of Venue State: It is also crucial to track the actual venue situation, like the trading phase, as well as the available opportunities.
  7. Integration of Venue Market Data: Thus, any smart order router requires real-time market data from different venues. The market data can be obtained directly by connecting to the venue's feed handlers, or by using market data handlers.
  8. Interfacing with the Target Gateway: Routing the orders to one or several venues according to the decision made at the second step above using either a FIX gateway, or a custom API gateway.



9. Dynamic Routing via Order Creation/Swap: Routing here does not imply just static routing to a certain venue, but dynamic behavior with updates of existing orders, creation of new ones, and sweeping to capture a newly appeared opportunity.
10. Client Gateways:



At a closer look, the structure of the SOR system usually contains *client gateways*, to receive incoming orders of the SOR customers.

11. Market Gateways: This sends order to specific exchanges.
12. SOR Implementation: The SOR implementation keeps the SOR logic and custom algos and tackles the clients' orders.
13. Feed Handlers: These provide market data from exchanges for decision making.
14. Client Front-ends: These provide GUI for SOR.

## Algorithmic Trading and SOR

1. Algorithmic Trading vs. SOR: In some cases, algorithmic trading is rather dedicated to automatic usage of synthetic behavior. Algorithmic trading manages the “parent” order while a smart order router directs the “child” orders to the desired destinations.



2. Slicing Parent Orders to Children: In effect, slicing a big order into a multiplicity of smaller orders and timing these orders to minimize market impact via electronic means.
3. Slice Size and Dispatch Time: Based on mathematical models, and considering historical and real-time market data, algorithms determine ex-ante, or continuously, the optimum size of the next slice and its time of submission to the market.
4. Inputs that Determine the Routing: A variety of principles is used for these algorithms, it is aimed at reaching or beating an implicit or explicit benchmark, e.g., a volume weighted average price VWAP algorithm targets at slicing and timing orders in a way that the resulting VWAP of its own transactions is close to or better than VWAP of all transactions of the respective security throughout the trading day or during a specified period of time.
5. Algorithmic Trading as an SOR Extension: However, smart order routing and algorithmic trading are connected more closely than it seems. Since even Smart Order Routing can be considered the simplest example of an algorithm, it is reasonable to say that algorithmic trading is a logical continuation and an extension of Smart Order Routing.
6. Simple Smart Order Routing Strategy: Shown below is a common example of a simple smart order routing strategy.
7. Child Orders to fulfill Parent: Having the initial Order Book, the SOR strategy will create the child orders, that is orders that aim at completing the initial SOR parent order.
8. Orders can be Passive or Aggressive: These orders can be either passive or aggressive depending on the current context and the SOR algorithm.
9. Illustrative Example of Using IOC Orders:

Preferred venue		Venue 1		Venue 2	
Buy	Sell	Buy	Sell	Buy	Sell
	100@21.5		200@21.5		300@21.6



In the example below, IOC – immediate or cancel – orders are used.

10. SOR Example Step #1: AN SOR Buy Day order for 1000@21.5 arrives.
11. SOR Example Step #2: Aggressive child order to grab opportunity in preferable venue is created: But IOC 100@21.5
12. SOR Example Step #3: Aggressive child order to grab opportunity in venue 1 created: Buy IOC 200@21.5
13. SOR Example Step #4: The remaining part placed passive to the preferred venue:

Preferred venue		Venue 1		Venue 2	
Buy	Sell	Buy	Sell	Buy	Sell
700@21.5					300@21.6

14. SOR Example Step #5: New liquidity on venue 2 appears: Sell 150@21.4

Preferred venue		Venue 1		Venue 2	
Buy	Sell	Buy	Sell	Buy	Sell
700@21.5					150@21.4
					300@21.6

15. SOR Example Step #6: The algo “sweeps” from preferred venue to grab the opportunity on venue 2. Buy 150@21.4 IOC

Preferred venue		Venue 1		Venue 2	
Buy	Sell	Buy	Sell	Buy	Sell
550@21.5					300@21.6



16. SOR Example Step #7: New liquidity on venue 1 appears: Sell [600@21.5](#)

Preferred venue		Venue 1		Venue 2	
Buy	Sell	Buy	Sell	Buy	Sell
550@21.5			600@21.5		300@21.6

17. SOR Example Step #8: The algo “sweeps” from the preferred venue to grab the opportunity on venue 1: Buy [550@21.5](#) IOC

18. SOR Example Step #9: The trade happens, the algo terminates, because all the intended shares were executed:

Preferred venue		Venue 1		Venue 2	
Buy	Sell	Buy	Sell	Buy	Sell
			50@21.5		300@21.6

19. Possibility of Child Order Rejects: As there are latencies involved in constructing and reading from the consolidated order book, child orders may be rejected before the target order was filled before it got there.

20. Handling Rejections and Partial Fills: Therefore, modern smart order routers have callback mechanisms that re-route orders if they are rejected or partially executed.

21. Need for Additional Liquidity: If more liquidity is needed to execute an order, smart order routers will post day limit orders, relying on probabilistic and/or machine learning models to find the best venues.

22. Dark Venues for Child Orders: If the targeting logic supports it, child orders may also be sent to dark venues, although the client will typically have an option to disable this.



23. Tradeoff between Cost and Time: More generally, smart order routing algorithms focus on a tradeoff between execution cost and execution time (Cont and Kukanov (2017)).

## Cross-Border Routing

1. Cross-Border Routing for Stocks: Some institutions offer cross-border routing for inter-listed stocks. In this scenario, the SOR targeting logic will use real-time FX rates to determine whether to route to various venues in different countries that trade in different currencies.
2. Most common Cross-border Routing: The most common cross-border routers typically route to both Canadian and American venues; however, there are some routers that also factor in European venues while they are open during trading hours.

## References

- Cont, R., and A. Kukanov (2017): Optimal Order Placement in Limit Order Markets  
*Quantitative Finance* **17 (1)** 21-39
- Wikipedia (2023): [Smart Order Routing](#)



## Nagar – Algos and Smart Order Router

### Introduction

1. Comprehensive Algos and Model Platform: DROP hosts its algos in a comprehensive platform that centralizes signals, models, and decision making designed to deliver a more efficient and effective trading experience. VWAP, TWAP, POV, VIPANI, and Target-Close algorithms are available currently within the platform.
2. Custom Smart-Order Router Implementation: DROP has developed a smart-order router – PATH – for US equities that leverages its electronic trading platform and NAGAR electronic trading strategies.

### Summary of the Algos Logic

1. Components of the Execution Platform: There are three key logical components within the Algos – Scheduler, Order Placement, and Router (PATH).
2. Role of the Order Scheduler: The scheduler aims to determine the best way to execute a given order based on client instructions, short-term liquidity demands, and a volume prediction model.
3. Role of the Order Placement: The Scheduler sends the Order Placement specific instructions regarding the target quantity to execute over a defined period as well as a preference to be either ahead or behind the schedule.
4. The Order Placement Execution Mode: The Order Placement logic is comprised of several execution modes. Within each execution mode, the Order Placement takes into account the liquidity demands of the Scheduler as well as a short-term price predictor to aggregate quantities and price levels to trade at any given time.



5. Order Router Venue Selection Scheme: Based on the combined instructions of the Scheduler and the Order Placement, specific venue selections are made by the embedded router component PATH as described later.

## Algorithms and Routing Logic

1. Execution Model Invocation of Router (PATH): DROP Algos utilize the PATH Router to route orders for execution. PATH – also available as a stand-alone strategy – is utilized by the Algos in 4 execution modes: Ping, Sweep, Exchange Post, and Dark Post.
2. Marketability Based on Order Execution Invocation: An order will go through Ping, Sweep, and Exchange Post – in that order – depending on the marketability.
3. Execution via the Dark Mode: Dark Post must be explicitly elected on a router order. A Dark Post will not trigger any other execution mode. Dark Post is currently turned on by the Order Placement on DROP algorithms by default.

## Ping Execution Mode

1. Objective of the Ping Mode: This mode is used to locate hidden liquidity at or within the spread.
2. Near and Far Touch Pings: Ping may be performed for marketable orders that could attempt to remove liquidity at the far-touch, i.e., National Best Offer NBO for buys, National Best Bid NBB for sells, or for or for orders that would post at the near-touch or better – at or NBB for a buy, at or below NBO for a sell.
3. Use of Dynamic Ping Module: NAGAR utilizes the Dynamic Ping Module, which utilizes a customized set of venues to find hidden liquidity using mid-point and limit orders.



4. Venues Used in Dynamic Ping: Dynamic Ping leverages feeds provided by VIPANI and other Broker-Dealer IOI Off-Exchange Venues.

## Sweep Execution Mode

1. Objective of the Sweep Mode: This mode takes the available liquidity at the best displayed price or better.
2. Aggregation of the Displayed Liquidity: NAGAR aggregates the displayed liquidity from the exchange market data feeds as well as from proprietary feeds provided by VIPANI, ATSs, and other Broker-Dealer IOI Off-Exchange Venues.
3. Aggregation of the Hidden Liquidity: Sweep may attempt to capture hidden liquidity in addition what may be displayed in the market data and proprietary feeds.
4. Using IOC and ISO Orders: NAGAR routes IOC orders to sweep, using ISOs whenever possible to attempt to improve fill rates and to sweep multiple price levels simultaneously, where relevant.
5. Limit Price Setting in Sweep: Sweep will set the limit price based on the depth to fill the order, unless constrained by a more conservative client-specified limit price.

## Exchange Post Execution Mode

1. Objective of the Exchange Post: The Exchange Post aims to provide liquidity on exchanges when the order is unmarketable and cannot take liquidity via Ping or Sweep modes.
2. Sub-modes within Exchange Post: Exchange Post has two options for Posting – Smart Post and Dynamic Post.
3. Mechanics Underlying Smart Post Mode: When Smart Post is making the allocation decision, NAGAR may determine that the expected increased probability of fill may not outweigh the increased likelihood of increasing the security's price by displaying



an amount above a certain size. This may result in NAGAR allocating both displayed and non-displayed liquidity into the market.

4. Venues with Displayed/Hidden Quantity: When a single venue is allocated using both a displayed and a non-displayed quantity, a single reserve order may be routed to that venue.
5. Use of Dynamic Post Mode: Dynamic Post utilizes a customized set pf profiles to post liquidity using a venue specific market share or allocation weights in the posting profiles.
6. Posting using DAY/ISO Order: NAGAR uses non-routable, DAY limit, displayed orders when doing Exchange Post.
7. Handling Locked and Crossed Orders: When a non-routable order is viewed as locking or crossing the NBBO, the exchange may slide the displayed price to a valid increment at that time.
8. Re-allocating the Quantities among the Venues: NAGAR may re-allocate quantities among venues when a child order routed to a destination receives a full fill, a minimum time has elapsed, or market conditions have changed.
9. Triggering the Ping/Sweep Modes: If the market becomes locked, NAGAR may trigger Ping or Sweep modes to take liquidity.
10. Handling the Opening/Closing Auctions: Shortly before the Opening and the Closing auctions, NAGAR may re-allocate orders to the primary exchange to ensure resting orders participate in the primary exchange's auctions.

## **Dark Post Execution Mode**

1. Objective of the Dark Post: The Dark Post mode aims to provide hidden liquidity across available venues.
2. Modes Used by Dark Post: Similar to the Exchange Post, the Dark Post has two modes for Posting – Smart Post or Dynamic Post.



3. Order Types used in Dark Post: NAGAR uses mid-point pegged DAY orders by default in this mode, but may use passive or market-pegged orders depending on client instruction.
4. Re-allocation of Quantities among Venues: NAGAR may re-allocate quantities among venues when a child order routed to a destination receives a full fill, a minimum time has elapsed, or the market conditions have changed.
5. NAGAR Internal Posts to VIPANI: NAGAR supports the ability to post an entire order initially to VIPANI for a short interval to allow for potential crossing prior to placing orders across multiple venues.

## Venue Selection

1. Real-time Venue Selection Parameters: Each execution mode used by the Order Placement may use any of the following as inputs for real-time venue selection.
2. Available Price: When attempting to take displayed liquidity within the sweep mode, venues displaying the best price will be allocated quantity ahead of venues with worse prices.
3. Fair Price: A proprietary, symbol-specific price prediction based on historical data, market analytics, and order feedback.
4. Historical Data: Information obtained from orders and executions from configured venues and from market activity on previous trading days. Historical data may be used in conjunction with, or in place of, real-time analytics and models.
5. Market Analytics: Intra-day, symbol specific information captured from market data, such as bid/ask imbalances, turnover frequency, execution price momentum, and execution price location within the spread.
6. Order Feedback: Real-time, parent or child-order specific information captured from results of previously placed open and/or executed child orders, such as presence of non-displayed/hidden liquidity and average observed execution price.



7. Order Details: Order-specific information such as size, limit price, symbol being traded, order type, and execution instructions.
8. Probability of Fill: A proprietary model that predicts the likelihood of an execution at a given price based on the relevant inputs mentioned above.
9. Venue Characteristics: Venue-specific information such as venue type, e.g., exchange vs. ATS, pricing structure, e.g., inverted vs. maker/taker, and support for specific order types/features.
10. Venue Cost: Fees charged and/or rebates offered by the venue.

## Accessible Venues

1. Trading Venues used in NAGAR: For Reg NMS securities, the following trading venues are accessible by NAGAR.
2. Exchanges:

Venue	Ping	Sweep	Exchange Post	Dark Post	Conditional
CBOE BZX	Y	Y	Y	Y	N
CBOE BYX	Y	Y	Y	Y	N
CBOE EDG A	Y	Y	Y	Y	N
CBOE EDG X	Y	Y	Y	Y	N
IEX	Y	Y	Y	Y	N
LTSE	Y	Y	N	N	N
MEMX	Y	Y	Y	Y	N
MIAX	Y	Y	N	N	N
NASDAQ	Y	Y	Y	Y	N
NASDAQ BX	Y	Y	Y	N	N
NASDAQ PSX	Y	Y	N	N	N
NYSE	Y	Y	Y	Y	X



NYSE American	Y	Y	Y	X	X
NYSE ARCA	Y	Y	Y	Y	X
NYSE Chicago	Y	Y	X	X	X
NYSE National	Y	Y	Y	Y	X

3. ATS:

Venue	Ping	Sweep	Exchange Post	Dark Post	Conditional
AlphaX US	Y	Y	N	Y	Y
ASPEN	Y	Y	N	Y	N
BOFA Instinct-X	Y	Y	N	Y	Y
BIDS	Y	Y	N	Y	Y
Block Cross	Y	Y	N	N	Y
BNP Cortex	Y	Y	N	Y	N
CODA Markets	Y	Y	N	N	N
Fidelity Cross Stream	Y	Y	N	Y	N
Instinet CBX	N	N	N	Y	N
Intelligent Cross	Y	Y	N	Y	N
GS Sigma-X	Y	Y	N	Y	N
Virtu POSIT	Y	Y	N	Y	Y
JPM-X	Y	Y	N	Y	Y
LeveL	Y	Y	N	Y	Y
Liquid Net H2O	Y	Y	N	N	Y
LX	Y	Y	N	Y	Y
MS POOL	N	N	N	Y	N
One Chronos	Y	Y	N	Y	Y
Pure Stream	N	N	N	Y	Y
UBS	Y	Y	N	Y	N
Virtu MatchIt	Y	Y	N	Y	N



#### 4. Broker-Dealer IOI Off-Exchange Venues:

Venue	Ping	Sweep	Exchange Post	Dark Post	Conditional
BARX Book	Y	Y	N	Y	Y
Citadel Connect	Y	Y	N	N	N
Jane Street LX	Y	Y	N	N	N
Jump Liquidity	Y	Y	N	N	N
Hudson River Trading	Y	Y	N	N	N
Tower Research Capital	Y	Y	N	N	N
Virtu VEQ Link	Y	Y	N	N	N
XTX	Y	Y	N	N	N

## Order Types Used

1. Order Types vs. Execution Modes: This section lists the order type category and their usage in the execution modes.
2. DAY: Execution Post | Dark Post
3. Hidden: Ping | Sweep
4. IOC: Ping | Sweep
5. ISO: Sweep
6. Limit: Ping | Sweep | Exchange Post | Dark Post
7. Non-routable/Do-not-ship: Ping | Sweep | Exchange Post
8. Pegged: Ping | Dark Post
9. Price Sliding: Exchange Post
10. Discretionary: Exchange Post | Dark Post
11. Reserve: Exchange Post
12. Market-on-Open/Limit-on-Open (MOO/LOO): Auction
13. Market-on-Close/Limit-on-Close (MOC/LOC): Auction



#### 14. D-QUOTE: Auction

### Order Handling Scenarios to Consider

1. Re-routing Short Sells on Price-test Securities: Short sell orders on a security that is in price test are routed to the primary exchange, unless a different exchange is specified.
2. Market Order on Price-test Securities: Market orders will be repriced as limit orders, one tick less aggressive than the market impact price, which is defined as more restrictive of either Limit Up/Limit Down Price and the clearly erroneous price.
3. Handling Price Unavailability at Venues: NAGAR may rely on exchange to reroute marketable orders when unable to directly access the venue at the best price due to a system issue; currently, PATH uses BATS for routing in such scenarios.
4. Pre-open and Continuous Trading: Orders entered prior to open will trade in pre-open and continuous sessions, unless the client has opted out of pre-open; continuous trading ends when the security goes into closed state based on the primary exchange.
5. Unopened Security at the Primary: If a security has not opened at the primary venue by a configured time, PATH may systematically treat the security as open for continuous trading.
6. Opening or Closing TIF: Orders with opening or closing times-in-force are automatically routed to the primary exchange.
7. Orders for Halted Securities: Orders for securities that are halted are routed to the primary exchange, unless a different exchange is specified.
8. Randomized Venue Selection: In cases where a decision cannot be made solely based on the analytics and factors described in this chapter, PATH may also use a randomized ranking table to select venues.

### Router Customization Options



1. Customization across all Execution Modes #1: Disable interaction with specific or all non-exchange venues.
2. Customization across all Execution Modes #2: Route exclusively based on exchange cost. Ping mode is disabled as part of this option. When sweeping, venues would be based on price first and then exchange cost.
3. Ping Customization: Disable Ping.
4. Sweep Customization #1: Route one price at a time when sweeping multiple prices.
5. Sweep Customization #2: Minimize odd lot interaction by not routing to displayed odd lot quotes and using a minimum execution size of 100 shares.
6. Exchange Post Customization #1: Specify the aggregate number of shares to display across all posted child venues.
7. Exchange Post Customization #2: Specify the venue to which PATH will post as part of the normal order placement logic.
8. Dark Post Customization #1: Limit routing to selected displayed and non- displayed venues, e.g., “Exclude posting to ATS XYZ”.
9. Dark Post Customization #2: Specify a minimum execution quantity when dark posting, e.g., “Do not execute for less 500 shares at any dark pool”.
10. Dark Post Customization #3: Disable NAGAR’s initial Order Placement to VIPANI via PATH.

## Directed Orders via PATH

1. Client Orders sent to NAGAR: Client orders sent to NAGAR – not including client orders sent to VIPANI via direct electronic order entry connection – are handled through PATH.
2. Client Orders to Individual Exchanges: Client directed orders are routed straight to the indicated exchange and are not subject to any of the PATH’s order placement logic.



3. PATH Control of Market Orders: PATH applies a risk control to directed market orders that results in these orders being sent to the market with a marketable limit – buying at or above the NBO or selling at or below the NBB – based on internally set thresholds.

## Exchange Market Data

1. EXEGY for Protected Venue Feeds: NAGAR uses direct market data feeds from all protected venues that makes such feeds available under Regulation NMS to obtain displayed quotes when making routing decisions. NAGAR employs a third-party vendor, EXEGY, to obtain these feeds.
2. SIP as a Backup Feed: In the event of a disruption in a direct data feed from a protected venue, e.g., if the venue is experiencing technology issues, NAGAR has a backup mechanism to use the SIP feed instead. SIP is also used to obtain last sale information.

## Non-exchange Market Data

1. IOIs from Non-displayed Venues: NAGAR receives IOI from several non-displayed liquidity venues, include venues that send IOIs from registered broker-dealers as well as from their unregistered affiliates.
2. IOI Quantities for Venue Rankings: When sweeping, IOI quantity is used as an input for the non-displayed venue in the venue ranking logic, analogous to how a market data feed is used for a displayed venue.
3. IOC Orders to Non-displayed Venues: In response to these IOIs, NAGAR may send IOC orders to the non-displayed liquidity venues, which may or may not execute the IOC order.



4. Quotes/Sizes from ATS Venues: NAGAR also receives quotes from displayed ATS venues. When sweeping, the displayed quantity is utilized as an input for the venue in the venue ranking logic.
5. IOC Orders to ATS Venues: In response to these quotes, NAGAR may send IOC orders to the displayed ATS venues, which may or may not execute the IOC order.

## Performance Evaluation of Algos

1. Performance Evaluation of Algos is a continuous Process
2. Metrics used for Performance Evaluation: The performance of the Algos is evaluated against several measures, which include but are not limited to, average slippage from the relevant benchmarks, e.g., VWAP and arrival price; effective participation rate; price movements before, during, and after the trade.
3. Cross Market and Outlier Analysis: Performance is also evaluated across various market capitalizations, spreads, durations, and order sizes. Outlier trades may be analyzed to identify potential improvements.

## Information Leakage Prevention and Anti-gaming Protection

1. Information Leakage and Gaming Activity: NAGAR employs various techniques to help minimize information leakage and avoid gaming activity. These techniques, but are not limited to, are the following:
2. Symbol-specific Analysis: Volume profiles, average spreads and quote depth, volatility estimates
3. Dynamic Scheduling and Volume Forecasting: Intraday, real-time adjustments to react to deviations from typical volume profile
4. Chase Protection: Symbol-specific block exclusion for all participation tracking



5. Non-deterministic Child Order Placement: Release time of passive and aggressive child orders does not follow a pre-defined progression.
6. Dynamic Limit Pricing for Non displayed Orders: When routing hidden orders, a local limit is applied to prevent unfavorable executions due to short-term market movements.

## **Capital Commitment Features**

1. Automated Capital Commitment CAPCOMM: CAPCOMM features are available through NAGAR strategies. Once pre-determined threshold or criteria is reached, and certain conditions are met, the remainder of the order is facilitated by NAGAR.
2. Child CAPCOMM for Executions: Child CAPCOMM is turned on by default for TWAP, VWAP, and POB flow, with an option to opt-out.



## Retail SOR Strategy Builder

### Retail SOR Wave Instructions

1. Wave #1: Send quantity to PATH with Sweep venue inclusion list (VIPANI). Order type will be mid-pegging, and time-in-force will be IOC.
2. Wave #1 Attributes/Strategy Parameters:

Time In Force	IOC
Pegging Instruction	Mid-point
Ping Type	SEQUENTIAL
Sweep Include Venue List	VIPANI
Post Type	NO POST

3. Wave #2: Send any remaining quantity to PATH with Sweep venue inclusion list that correspond to inverted venues. Order type will be mid-pegging, and time-in-force will be IOC.
4. Wave #2 Venue Choice Customization: The inverted venue set is designed to be flexible, as exchanges can change fee schedules and can change from inverted to make/take.
5. Wave #2 Attributes/Strategy Parameters:

Time In Force	IOC
Pegging Instruction	Mid-point
Ping Type	SEQUENTIAL
Sweep Include Venue List	XBOS, BATY, EDGA, XCIS
Post Type	NO POST



6. Wave #3: Send any remaining quantity to PATH with Sweep venue inclusion list (VIPANI) with time-in-force of IOC.
7. Wave #3 Attributes/Strategy Parameters:

Time In Force	IOC
Pegging Instruction	NONE
Ping Type	NONE
Sweep Include Venue List	VIPANI
Post Type	NO POST

8. Wave #4: Send any remaining quantity to PATH with IOI Sweep venue inclusion list with time-in-force of IOC.
9. Wave #4 Attributes/Strategy Parameters:

Time In Force	IOC
Pegging Instruction	NONE
Ping Type	NONE
Sweep Include Venue List	IOI
Post Type	NO POST

10. Cancel any remaining Quantity to the User.

## Phase-based Conditional Action

1. Pre-open Phase: Here

$$AsOfTime \leq ContinuousStartTime$$



This corresponds to pre-market opening auction, and takes full fill or partial fill, the rest is canceled. The order is sent pre-market, the auction is skipped, and continuous logic is applied.

2. Continuous Phase: Here

$$\text{ContinuousStartTime} \leq \text{AsOfTime} < \text{ContinuousEndTime}$$

This corresponds to pre-market opening auction, and takes full fill or partial fill, the rest is canceled. The order is sent pre-market, the auction is skipped, and continuous logic is applied.

3. Post-close Phase: Here

$$\text{AsOfTime} \geq \text{ContinuousEndTime}$$

There will not be any trade in the Post market phase.

## Additional Considerations

1. DAY Parent Orders from Client: The client may send TIF as either IOC or DAY, but the underlying wave will always be IOC orders, hence the parent will be treated the same.
2. Processing of Price/Quantity Amendments: Amendments will be processed by canceling the current wave and restarting from Wave #1 in line with other Strategy Builder algos.
3. Processing the Cancels: For cancels, the wave in progress will be allowed to complete, and any leaves will be canceled.
4. Case where Symbol is an IPO: If the symbol is an IPO, the Strategy Builder will reject the order.



5. Rejection from the Strategy Builder: Rejections from the Strategy Builder will not pause at the desk as it is a NAGAR workflow.

## Continuous Trading Scenario Overview

1. Send a Marketable Limit Order with TIF = IOC: All wave sequences above apply.
2. Send a Market Order with TIF = DAY: All wave sequences above apply.
3. Pre-open Order with TIF = DAY: Order received before pre-open should be parked and then trade – waves will be sent – in continuous.
4. Amend Price Up: The current wave will be canceled and restarted from Wave #1 in line with other Strategy Builder algos. The waves will go at the amended price after the amend.
5. Amend Price Down: The current wave will be canceled and restarted from Wave #1 in line with other Strategy Builder algos.
6. Amend TIF to DAY: From IOC to DAY. Wave should not be canceled after the amend, and the order should execute normally.
7. Amend TIF to IOC: From DAY to IOC. Wave should not be canceled after the amend, and the order should execute normally.
8. Amend Quantity Up: The parent quantity is first amended. The current wave will be canceled and restarted from Wave #1 in line with other Strategy Builder algos.
9. Amend Quantity Down but above Leaves: The parent quantity is first amended. The current wave will be canceled and restarted from Wave #1 in line with other Strategy Builder algos.
10. Amend Quantity Down but below Leaves: The parent quantity is first amended. The current wave will be canceled and restarted from Wave #1 in line with other Strategy Builder algos.
11. Closed IPO Order: Order should be rejected.
12. Marketable Limit Order sent to a Halted Symbol: The order will be accepted but will not trade.



13. Marketable Limit Order sent to a Halted-to-Open Symbol: The order will be accepted but will not trade. Once the symbol is open, the full wave scenario will apply.
14. Order Followed by Cancel: The sliced waves are processed, but the leaves are canceled.



## Indifference Price

### Overview

1. Pricing Using a Utility Function: *Indifference Pricing* is a method pricing financial securities with regard to a utility function (Wikipedia (2023)). The *indifference price* is also known as the *reservation price* or *private valuation*.
2. Definition of the Indifference Price: In particular, the indifference price is the price at which the agent would have the same utility level by exercising a financial transaction as by not doing so – with optimal trading otherwise.
3. Indifference Price as a Range: Typically, the indifference price is a pricing range – a bid-ask spread – for a specific agent; this price is an example of good-deal bounds (Birge (2008)).

### Mathematics

1. Utility Scheme Based Pricing Functional: Given a utility function  $u$  and a claim  $C_T$  with a known payoff at some terminal time  $T$ , let the function

$$V : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$$

be defined by

$$V(x, k) = \sup_{X_T \in \mathcal{A}(x)} \mathbb{E}[u(X_T + kC_T)]$$



where  $x$  is the initial endowment,  $\mathcal{A}(x)$  is the set of all self-financing portfolios at time  $T$  starting with endowment  $x$ , and  $k$  is the number of claims to be purchased or sold.

2. Indifference Bid and Ask Prices: The indifference bid price  $v_b(k)$  for  $k$  units of  $C_T$  is the solution to

$$V(x - v_b(k), k) = V(x, 0)$$

and the indifference ask price  $v_a(k)$  is the solution to

$$V(x + v_a(k), -k) = V(x, 0)$$

3. The Indifference Price Bound Range: The indifference price bound is the range  $[v_b(k), v_a(k)]$  (Carmona (2009)).

## Example

1. Portfolio of Riskless/Risky Assets: Consider a market with a risk-free asset  $B$  with

$$B_0 = 100$$

and

$$B_T = 110$$

and a risky asset  $S$  with

$$S_0 = 100$$



and

$$S_T \in \{90, 110, 130\}$$

each with probability  $\frac{1}{3}$ .

2. European Bid/Ask Indifference: Let the utility function be given by

$$u(x) = 1 - e^{-\frac{x}{10}}$$

To find either the bid or the ask indifference price, for a single European call option with strike 110,  $V(x, 0)$  is first calculated.

3. Utility Function Based  $V(x, 0)$  Evaluation:

$$\begin{aligned} V(x, 0) &= \max_{\alpha B_0 + \beta S_0} \mathbb{E} \left[ 1 - e^{-\frac{\alpha B_T + \beta S_T}{10}} \right] \\ &= \max_{\beta} \left[ 1 - \frac{1}{3} \left( e^{-\frac{1.10x - 20\beta}{10}} + e^{-\frac{1.10x}{10}} + e^{-\frac{1.10x + 20\beta}{10}} \right) \right] \end{aligned}$$

which is maximized at

$$\beta = 0$$

therefore

$$V(x, 0) = 1 - e^{-\frac{1.10x}{10}}$$

4. Indifference Based Bid Price Expression: To find the indifference bid price, one solves for  $V(x - v_b(1), 1)$ .



$$\begin{aligned}
 V(x - v_b(1), 1) &= \max_{\alpha B_0 + \beta S_0 = x - v_b(1)} \mathbb{E} \left[ 1 - e^{-\frac{\alpha B_T + \beta S_T + C_T}{10}} \right] \\
 &= \max_{\beta} \left( 1 - \frac{1}{3} \left\{ e^{-\frac{1.10[x-v_b(1)]-20\beta}{10}} + e^{-\frac{1.10[x-v_b(1)]}{10}} + e^{-\frac{1.10[x-v_b(1)]+20\beta}{10}} \right\} \right)
 \end{aligned}$$

which is minimized at

$$\beta = \frac{1}{2}$$

therefore

$$V(x - v_b(1), 1) = 1 - \frac{1}{3} e^{-\frac{1.10x}{10}} e^{-\frac{1.10v_b(1)}{10}} \left[ 1 + \frac{2}{e} \right]$$

Thus

$$V(x, 0) = V(x - v_b(1), 1)$$

occurs when

$$v_b(1) = \frac{10}{1.1} \log \frac{3}{1 + \frac{2}{e}} \approx 4.97$$

5. Indifference Based Evaluation of Ask: The ask price is similarly solved to estimate  $v_a(1)$ .

## Notes



1. Bid-Ask Indifference Claim Symmetry: If  $[v_b(k), v_a(k)]$  are the indifference price bounds for a claim, then, by definition

$$v_b(k) = -v_a(-k)$$

(Carmona (2009)).

2. Comparison against Corresponding Hedge Prices: If  $v(k)$  is the indifference bid price for a claim and  $v_{sup}(k)$  and  $v_{sub}(k)$  are the super-hedging and the sub-hedging prices respectively, then

$$v_{sub}(k) \leq v(k) \leq v_{sup}(k)$$

Therefore, in a complete market, the indifference price is always equal to price to hedge the claim.

## References

- Birge, J. R. (2008): *Financial Engineering* Elsevier Amsterdam, Netherlands
- Carmona, R. (2009): *Indifference Pricing: Theory and Applications* Princeton University Press Princeton, NJ
- Wikipedia (2023): [Indifference Price](#)



## Cost Adaptive Arrival Price Trading

### Synopsis and Key Results

1. Cost Uncertainty Balance in Trading: Electronic trading of Equities and other securities makes heavy use of *arrival price* algorithms that determine optimal trade execution strategies by balancing the market impact cost of rapid execution against the volatility risk of slow execution.
2. Optimal Static Mean Variance Strategies: In the standard formulation mean-variance optimal strategies are static; they do not modify the execution speed in response to price motions observed during trading.
3. Dynamically Adjusting Risk Tolerant Profiles: Almgren and Lorenz (2007) show that with a more realistic formulation of the mean-variance tradeoff, and even with no momentum or mean-reversion in the price process, substantial improvements are possible for adaptive strategies that spend trading gains to reduce risk by accelerating execution when the price moves in the traders' favor. The improvement is larger for large initial portfolios.

### Introduction, Background, and Motivation

1. Breaking-Down a Given Order: Algorithmic trading represents a large and growing fraction of the total order flow, especially in the equity markets. When the size of a requested buy or sell order is larger than what the market can immediately supply or absorb, then the order must be worked across some period of time, exposing the trader to price volatility.



2. Tailoring Execution to Risk Preferences: The algorithm attempts to achieve an average execution price whose profitability is suited to the client's preferences. Almgren and Lorenz (2007) propose a way to dramatically improve this distribution.
3. Benchmark Implementation Shortfall: Arrival price algorithms, which are currently the most widely used framework, take as their benchmark the *pre-trade* or *decision* price. The difference between the execution price and the benchmark is the *implementation shortfall* which is an uncertain quantity since the order execution takes a finite amount of time.
4. Expected Value of Shortfall: In the most straightforward implementation of this model, the expected value of the implementation shortfall is entirely due to the market impact incurred by trading at a non-zero rate (neglecting any anticipated price drift); this expected cost is minimized by trading as slowly as possible, for example, using a VWAP strategy across the maximum allowed time horizon.
5. Variance of Implementation Shortfall: Since market impact is assumed to be deterministic, the variance of the implementation shortfall is entirely due to price volatility; this variance is minimized by trading rapidly.
6. Efficient Frontier of Optimal Trading: This risk-reward trade-off is very common in finance, and a variety of criteria can be used to determine risk-averse optimal solutions. Arrival price algorithms compute the set of *efficient* strategies that compute the risk for a specified level maximum of expected cost or the converse; the set of such strategies is summarized in the *efficient frontier of optimal trading* introduced by Almgren and Chriss (1999, 2000).
7. Independence from the Portfolio Size: This simple mean-variance approach has the advantage that the risk-reward tradeoff is independent of the initial wealth – a useful property in an institutional setting.
8. Static vs. Dynamic Execution Schemes: A central question is whether the trade schedule should be *static* or *dynamic*; should the list of shares to be executed at each time interval be computed and fixed before the trading begins, or should the list be updated in “real time” using information revealed during the execution?



9. Static vs. Dynamic Schemes Equivalence: The observations of Almgren and Chriss (2000) is that, under very realistic assumptions about the price process (arithmetic random walk with no serial correlation), static strategies are *equivalent* to dynamic strategies. No value is added by considering “scaling” strategies in which the execution speed changes in response to price motions.
10. Static Strategy - Initial Time Determination: To be more specific, two different specifications of the trade scheduling problem are considered. For a static strategy the entire trade schedule is required to be fixed in advance – Huberman and Stanzl (2005) suggest that a reasonable example of this is insider trading – where trades must be announced in advance.
11. Pre-determination of the Trajectory Cost Distribution: For any candidate schedule the mean and the variance are calculated at the initial time, and the optimal schedule is determined for a specific risk aversion level.
12. Dynamic Determination of the Trading Schedule: For a dynamic strategy – as is usually understood in dynamic programming – arbitrary modification of the strategy is allowed at any time. To re-calculate the trade list all available information is used at that time, and the strategies are valued using a mean-variance trade-off of the remaining cost, using a constant parameter of risk aversion.
13. Reduction of Dynamic to Static: In the model of Almgren and Chriss (2000), the first and the second strategies have the same solution. Liquidity and volatility are assumed known in advance, so the only information revealed is the asset price motion.
14. Price Distribution Independence from the Realization: Price distribution revealed in the first part of the execution does not change the probability distribution of future price changes. Because the mean-variance trade-off is independent of the initial wealth, trading gains or losses incurred in the first part of the program are “sunk costs” and therefore do not influence the strategy for the remainder.
15. Trade Rate Determination Rule: Almgren and Lorenz (2007) present an alternate formulation. In this, they pre-compute the *rule* determining the trade rate as a function of price, using a mean-variance tradeoff measures at the initial time. Once trading begins the rule may not be modified, even if the trader’s preferences re-



evaluated at an intermediate time would lead him or her to choose a different strategy, as in the second strategy above.

16. Differences among the Computed Trajectories: The optimal solution to the third strategy is generally not the same as the solutions to the first and the second strategies.
17. Example: Comparison of the Strategies: As an illuminating contrast, in the well-known problem of option hedging, the optimal hedge position, once the trade list, depend on the price, and hence are not known until the price is observed, although the rule giving this hedge position is computed in advance using dynamic programming. Thus the first strategy is dramatically sub-optimal, and the second gives the same results as the third.
18. When is the Third Strategy Optimal? For algorithmic trading, the improvement of the third strategy over the first and the second come from introducing the negative correlation between the trading gains or losses in the first part of the execution and the market impact costs incurred in the second part.
19. Extraneously Imposed Serial Correlation Rule: Trading gains and losses due to price movements are serially uncorrelated, but can be correlated with the market impact costs via a simple rule; if the price moves in the traders' favor in the early part of the trading, then those gains are spent on the market impact costs by accelerating the remainder of the program.
20. Adaptive Strategy - Contra Price Move: If the price moves against, then the future costs are reduced by trading more slowly, despite the increased exposure to risk of future fluctuations. The result is an overall decrease in the variance measured at the initial time, which can be traded for a decrease in the expected cost.
21. Ex ante vs ex post Optimization: In practice there are no artificial constraints in the adaptivity of the trading strategies. The key observation contained in this chapter is that the *ex ante* mean-variance optimization expressed by the third formulation corresponds better to the way the trading results are measured in practice, via *ex post* sample mean and variance over a collection of similar programs.



## Adaptive Strategies – A Simple Illustration

1. Universe of Available Sample Bets: Suppose that two bets are available. Bet  $A$  pays 0 or 6 with equal probability; its expected value is 3 and its variance is 9. Bet  $B$  pays 1 with certainty; its expected value is 1 and its variance is 0.
2. Per-Strategy Objective Utility Value: Consider the case of a risk-averse investor whose coefficient of risk aversion is  $\frac{1}{9}$ ; he assigns an *ex ante* value of  $E - \frac{1}{9}V$  to a random payout with an expected value  $E$  and variance  $V$ . For this investor a single pay of  $A$  has a value 2 and a single pay of  $B$  has a value 1 so he prefers  $A$ .
3. Two Plays with Outcome Independence: Now suppose that the investor plays this game twice, with independence between the outcomes. Three ways in which he chooses his bets are considered.
4. First Strategy - Optimal Outcome: In a static strategy, the sequence  $AA$ ,  $AB$ ,  $BA$ , or  $BB$  must be fixed before the game begins. By independence choice  $AA$  has twice the value of  $A$  and is preferred. Its value is 4.
5. Second Strategy - Constant Wealth Effect: In a dynamic strategy, the second bet is chosen after the result of the first play is learnt. By that time the first result will be a constant wealth effect, so  $A$  will always be chosen on the second play.
6. Second Strategy - Optimal Outcome: Knowing that that will be the future choice  $A$  is chosen on the first bet as well to maximize the total value measured at the initial time. Thus the strategy and the payoff are the same as in the static case.
7. Third Strategy - Sequential Play Rule: In the new formulation the investor specifies *three* choices; his bet on the first play, his bet on the second play if he wins the first one, and his bet on the second play if he loses the first.
8. Third Strategy - Optimal Outcome: The optimal rule is to bet  $A$  on the first play, then if he wins to choose  $B$ , if he loses to play  $A$  again, giving payouts of 0, 6, 6, and 7 with equal probability. Its value is 4.06, better than the first two strategies.
9. Optimally using Slow/Fast Trading: In this model bet  $A$  corresponds to slow trading, with high expected value (low cost) and high variance, and  $B$  is fast trading. If the



random outcome (trading gain) in the first period is positive, then the trader spends some of this gain on reducing the variance in the second period.

10. Extension to Multi-Play: Now suppose the investor plays this game many times in sequence, and wishes to optimize the sample mean and variance, combined the coefficient of risk aversion.
11. Ex Post vs Ex Ante Single Play: If the results are reported over individual plays, then the *ex post* sample mean and variance will be close to the *ex ante* expectation and variance of a single play, and the optimal strategy would be to bet A each time, as in the first and the second strategies above.
12. Aggregation Over Play Pairs: However, suppose the results are aggregated over *pairs* of plays. That is, the gains of play 1 and play 2 are added together, play 3 and play 4 are added, *etc.*
13. Pairs Connected via Rule: Then the third strategy above, which is adaptive, will give the best results; within each pair choose the second bet based on the result of the first one. If the results are grouped into larger sets, then a more complicated strategy will be even more optimal.

## Trading in Practice

1. Reporting Driven Aggregation Granularity: As in the simple example, the question of which formulation is more realistic depends on how the trading results are reported. At Banc of America securities, and probably at other firms, clients of the agency trading desk are provided with a post-trade report daily, weekly, or monthly depending on their trading activity.
2. Aggregation along the Reporting Dimensions: Typically these reports show sample average and standard deviation of execution price relative to the implementation shortfall benchmark across all the trades executed for that client during the reporting period. The results are further broken down into subsets across a dozen dimensions



such as strategy type, buy or sell, primary exchange, trade sector, industry sector, market capitalization, *etc.*

3. Supra-Order Reporting Challenges: Because of these kinds of subsets, it is difficult to identify a larger unit than an individual order. Therefore it can be argued that the broker-dealer's goal is to design algorithms that optimize sample mean and variance at the per-order level so that the post-trade report will be as favorable as possible.
4. Order-Level Aggregation corresponds to Third Strategy: As in the simple example this criterion translates to the third strategy above which is not optimized by the typical arrival price algorithms.
5. Reporting Consistency with Client Goals: Of course, the broker also has a responsibility to design the post-trade report so that it will be maximally useful to the client; that is, it corresponds as closely as possible to the client's investment goals.
6. Execution Metrics under Finer Resolution: One interpretation of these results is that the reports should show details with a finer resolution. For instance, it can show the mean and the variance of shortfall for each one thousand dollars of client money spent. The best choice of the reporting intervals is an open question.

## Other Adaptive Strategies

1. “Aggressive-in-the-money” AIM: The new optimal strategies of Almgren and Lorenz (2007) are “aggressive-in-the-money” in the sense of Kissell and Malamut (2006); execution accelerates when the price moves in the traders’ favor, and slows when the price moves adversely.
2. “Passive-in-the-money” PIM: A “passive-in-the-money” (PIM) strategy would react oppositely. Adaptive strategies of this form are called “scaling” strategies, and they can arise for a number of different reasons beyond those considered here.
3. Traders’ Preference and Prospect Theory: A decrease in risk tolerance following a gain, and an increase following a loss, is consistent with traders’ risk preferences



(Shefrin and Statman (1985)) and is well-known in “prospect theory” (Kahneman and Tversky (1979)).

4. Mathematical Foundations of Scaling Strategies: Perhaps for these reasons scaling strategies often seem intuitively reasonable, though such qualitative preferences properly have no place in quantitative institutional trading. The formulation here is straightforward mean-variance optimization.
5. PIM as Optimal Momentum Strategy: One important reason for using an AIM or a PIM strategy would be the expectation of serial correlation in the price process. If the price is believed to have momentum, i.e., positive serial correlation, then a PIM strategy is optimal; if the price moves favorably, one should slow down to capture even more favorable prices in the future.
6. AIM for Mean Reversion Optimality: Conversely if the price is believed to be mean-reverting, then favorable prices should be captured quickly before they mean-revert (Kissell and Malamut (2006)). The strategies presented in this chapter arise from pure random walk with no serial price correlation, using pure classic mean-variance optimization.
7. Caveats behind AIM/PIM Deployment: These models do provide an important caveat for their formulation. The AIM strategy suggest to “cut the gains and let the losses run”.
8. Adverse Impact of “Moneyness” Guesses: If the price process does have any significant momentum, even on a small fraction of the real orders, then this strategy can cause much more serious losses than the gains that it provides. Thus implementing them in practice should be done only after doing extensive empirical tests.
9. The “Market Power” Parameter: The next section presents the market and the trading model, and shows the general importance of the “market power” parameter.
10. Single Update vs. Continuous Time: Two simple “proofs of concept” are considered; first a single update time, then a continuous response function that depends linearly on the asset price. The final section describes some approaches towards a full continuous time model.



## The Market Model

1. Asset Price Arithmetic Random Walk: Trading in a single asset whose price is  $S(t)$  is considered.  $S(t)$  obeys the arithmetic random walk

$$S(t) = S_0 + \sigma B(t)$$

where  $B(t)$  is a standard Brownian motion and  $\sigma$  is an absolute volatility. This process has neither momentum nor mean reversion; future price changes are completely independent of past changes.

2. Intra-day Profile Adapted Randomness: The Brownian motion  $B(t)$  is the only source of randomness in the formulation. In the presence of intra-day seasonality  $t$  is interpreted as a time relative to a historical profile, and volume is assumed to constant under this transformation.
3. The Trade Order Execution Settings: The trader has an order of  $X$  shares, which begins at time

$$t = 0$$

and must be completed by time

$$t = T < \infty$$

$X$  is taken to be

$$X > 0$$



and this is interpreted as a buy order. The benchmark value of this position at the start of the strategy is  $XS_0$

4. The Estimation Output Trading Strategy: A *trading trajectory* is a function  $x(t)$  with

$$x(0) = X$$

and

$$x(T) = 0$$

representing the number of shares remaining to buy at time  $t$ . For a static trajectory  $x(t)$  is determined at

$$t = 0$$

but in general  $x(t)$  may be any non-anticipating random functional of  $B(t)$ .

5. Observability of the Equilibrium Price: Permanent market impact is also important but has no effect on the optimal trade trajectory if it is linear – Almgren and Chriss (2000) carry out a detailed discussion of this model. The model parameters are assumed to be known with certainty and thus the underlying price  $S(t)$  is observable based on the execution price  $\tilde{S}(t)$  and the trade rate  $v(t)$ .
6. The Arrival Price Shortfall: The *implementation shortfall*  $\mathcal{C}$  is the total cost of executing the buy program relative to its initial value.

$$\mathcal{C} = \int_0^T \tilde{S}(t)v(t)dt - XS_0 = \sigma \int_0^T x(t)dB(t) + \eta \int_0^T v^2(t)dt$$

7. Randomness of the Cost Components: The first term above represents the trading gains or losses. Since the trader is buying a positive price motion gives a positive



cost. The second term represents the market impact cost. For an adaptive strategy both terms are random since  $x(t)$  and  $v(t)$  are both random.

8. Adaptive Strategies not necessarily Gaussian: Mean-variance optimization solves the problem

$$\min_{x(t)} \{ \mathbb{E}[\mathcal{C}] + \lambda \mathbb{V}[\mathcal{C}] \}$$

for each

$$\lambda \geq 0$$

where  $\mathbb{E}[\mathcal{C}]$  and  $\mathbb{V}[\mathcal{C}]$  are the expected values of the mean and the variance of  $\mathcal{C}$ . As  $\lambda$  varies the resulting set of points  $\{\mathbb{V}_\lambda[\mathcal{C}], \mathbb{E}_\lambda[\mathcal{C}]\}$  trace out an efficient frontier. For adaptive strategies  $\mathcal{C}$  is not Gaussian, but mean-variance optimization is still used.

## Static Trajectories

1. Non-random Optimal Execution Trajectory: If  $x(t)$  is fixed independently of  $B(t)$  then  $\mathcal{C}$  is a Gaussian random variable with mean and variance

$$\mathbb{E}_\lambda[\mathcal{C}] = \eta \int_0^T v^2(t) dt$$

and

$$\mathbb{V}_\lambda[\mathcal{C}] = \sigma^2 \int_0^T x^2(t) dt$$



2. Optimal Execution Trajectory Closed Form: The solution to

$$\min_{x(t)} \{ \mathbb{E}[\mathcal{C}] + \lambda \mathbb{V}[\mathcal{C}] \}$$

is then obtained as

$$x(t) = Xh(t, T, \kappa)$$

where the static trajectory function is

$$h(t, T, \kappa) = \frac{\sinh[\kappa(T - t)]}{\sinh[\kappa T]}$$

for

$$0 \leq t \leq T$$

and the static *urgency* parameter is

$$\kappa = \sqrt{\frac{\lambda \sigma^2}{\eta}}$$

3. Portfolio Size Dependence of Urgency: The units of  $\kappa$  are inverse time, and  $\frac{1}{\kappa}$  is the desired time scale for liquidation – the “half-life” as described in Almgren and Chriss (2000). The static trajectory is effectively an exponential with adjustments made to reach

$$x = 0$$



at

$$t = T$$

For a fixed  $\lambda$  the optimal time scale is independent of the portfolio size  $X$  since both the expected costs and the variance scale as  $X^2$ .

4. Static vs Dynamic Trajectory Equivalence: Equivalence of the static and the dynamic trajectories is demonstrated by observing that

$$h(t, T, \kappa) = h(s, T, \kappa)h(t - s, T - s, \kappa)$$

for

$$0 \leq s \leq t \leq T$$

That is, the trajectory recomputed at time  $s$ , using the same urgency parameter, is the same as the tail of the original trajectory.

5. Low Urgency Limit VWAP Trading: By taking

$$\kappa \rightarrow 0$$

the linear profile

$$x(t) = X \frac{T - t}{T}$$

is recovered, which is equivalent to a VWAP profile under volume time transformation. The profile has expected cost

$$E_{LIN} = \frac{\eta X^2}{T}$$



and variance

$$V_{LIN} = \frac{\sigma^2 X^2 T}{3}$$

## Non-dimensionalization

1. Dimensional Constants Determining the Solution: The optimal trajectory and the cost depend on 5 dimensional constants; the initial shares  $X$ , the time horizon  $T$ , the volatility  $\sigma$ , the impact coefficient  $\eta$ , and the risk aversion  $\lambda$ . To simplify the structure of the solution, it is convenient to define scaled variables.
2. Non-dimensionalization of Horizon/Time: Time is measured relative to  $T$  and shares relative to  $X$ . That is, the non-dimensional time is defined as

$$\hat{t} = \frac{t}{T}$$

and the non-dimensional holdings as

$$\hat{x}(\hat{t}) = \frac{x(\hat{t}T)}{X}$$

so that

$$0 \leq \hat{t} \leq 1$$

and

$$\hat{x}(0) = 1$$



The non-dimensional velocity is

$$\hat{v}(\hat{t}) = \frac{v(\hat{t}T)}{X/T} = -\frac{d\hat{x}}{d\hat{t}}$$

3. Non-dimensionalization of the Cost: The cost is scaled by a dollar cost of a typical move due to volatility. That is, defining

$$\hat{\mathcal{C}} = \frac{\mathcal{C}}{\sigma X \sqrt{T}}$$

one then has

$$\hat{\mathcal{C}} = \int_0^1 \hat{x}(\hat{t}) d\hat{B}(\hat{t}) + \mu \int_0^1 \hat{v}^2(\hat{t}) d\hat{t}$$

where

$$\hat{B}(\hat{t}) = \frac{B(\hat{t}T)}{\sqrt{T}}$$

and the “market power” parameter is

$$\mu = \frac{\eta X}{\sigma \sqrt{T}} / T$$

4. Definition of the “Market Power”: Here the numerator is the price concession for trading at a constant rate, and the denominator is the typical size of the price motion



due to volatility over the same period. The ratio  $\mu$  is the non-dimensional preference free measure of the portfolio size, in terms of its ability to move the market.

5. “Market Power” Estimation – ATHL Model: To estimate realistic sizes for this parameter one recalls that Almgren, Thum, Hauptmann, and Li (2005) introduced the non-linear model

$$\frac{K}{\sigma} = \eta \left( \frac{X}{VT} \right)^\alpha$$

where  $K$  is the temporary impact (the only kind relevant here),  $\sigma$  is the daily volatility,  $X$  is the trade size,  $V$  is the average daily volume (ADV), and  $T$  is the fraction of the day over which the trade is executed.

6. ATHL Model “Market Power” Estimates: The coefficient was estimated as

$$\eta = 0.142$$

as was the exponent

$$\alpha = \frac{3}{5}$$

Therefore a 100% ADV executed across one full day gives

\

$$\mu = 0.142$$

7. “Market Power” Estimate Typical  $\mu$ : Although the estimate above is only an approximate parallel to the linear model used here, it does suggest that for realistic trade sizes  $\mu$  will be substantially smaller than one.
8. Non-dimensionalization of the Urgency: The problem



$$\min_{x(t)} \{ \mathbb{E}[\mathcal{C}] + \lambda \mathbb{V}[\mathcal{C}] \}$$

has the scaled form

$$\min_{\hat{x}(\hat{t})} \{ \mathbb{E}[\hat{\mathcal{C}}] + \mu \bar{\kappa}^2 \mathbb{V}[\hat{\mathcal{C}}] \}$$

and the static urgency is

$$\bar{\kappa} = \kappa T$$

with  $\kappa$  from

$$\kappa = \sqrt{\frac{\lambda \sigma^2}{\eta}}$$

or

$$\bar{\kappa}^2 = \frac{\lambda \sigma^2 T^2}{\eta}$$

9. Non-dimensionalization of the Risk Aversion: The scaled risk aversion parameter  $\mu \bar{\kappa}^2$  depends on  $X$  via the factor  $\mu$  though the scaled time  $\bar{\kappa}$  is independent of  $X$ .
10. Non dimensionalization of the Trajectory:  $\bar{\kappa}$  will be used as the parameter to trace the frontier in place of  $\lambda$ . The result will be a trajectory  $\hat{x}(\hat{t}; \bar{\kappa}, \mu)$  with the scaled cost values  $\mathbb{E}[\hat{\mathcal{C}}(\bar{\kappa}, \mu)]$  and  $\mathbb{V}[\hat{\mathcal{C}}(\bar{\kappa}, \mu)]$ .
11. Non dimensionalization of Cost Distribution: For each of

$$\mu \geq 0$$



there will be an efficient frontier obtained by tracing  $\mathbb{E}[\hat{\mathcal{C}}(\bar{\kappa}, \mu)]$  and  $\mathbb{V}[\hat{\mathcal{C}}(\bar{\kappa}, \mu)]$  as functions of  $\bar{\kappa}$  over

$$0 \leq \bar{\kappa} < \infty$$

The profile has expected cost

$$\hat{E}_{LIN} = \mu$$

and variance

$$\hat{V}_{LIN} = \frac{1}{3}$$

## Small Portfolio Limit

1. Limit of Small “Market Power”: Next the limit

$$\mu \rightarrow 0$$

is considered, keeping  $\bar{\kappa}$  constant. Since  $X$  appears in  $\mu$  but not in  $\bar{\kappa}$  and all other dimensional variables do appear in  $\bar{\kappa}$  this is equivalent to taking

$$X \rightarrow 0$$

with  $T, \sigma, \eta$ , and  $\lambda$  fixed. Almgren and Lorenz (2006) show that for small portfolios status strategies are optimal.

2. Variance of the Non-dimensional Cost: When  $\mu$  is small, assuming that  $x(t)$  and  $v(t)$  have reasonable limits, the second term in



$$\hat{\mathcal{C}} = \int_0^1 \hat{x}(\hat{t}) d\hat{B}(\hat{t}) + \mu \int_0^1 \hat{v}^2(\hat{t}) d\hat{t}$$

is small compared to the first, and the variance of the non-dimensional cost is approximately

$$\mathbb{V}[\hat{\mathcal{C}}] \sim \mathbb{V} \left[ \int_0^1 \hat{x}(\hat{t}) d\hat{B}(\hat{t}) \right] = \int_0^1 \mathbb{E}[\hat{x}^2(\hat{t})] d\hat{t}$$

$$\mu \rightarrow 0$$

3. Market Impact Contribution to Volatility: That is, the uncertainty in realized price comes primarily from the price volatility. Even if the strategy is adapted to the price process so that  $\hat{x}(\hat{t})$  is random the market impact cost itself is a small number and the uncertainty in that number can be neglected next to the price volatility.
4. Expectation of Non-dimensional Cost: The first term in

$$\hat{\mathcal{C}} = \int_0^1 \hat{x}(\hat{t}) d\hat{B}(\hat{t}) + \mu \int_0^1 \hat{v}^2(\hat{t}) d\hat{t}$$

has strictly zero expected value for any non-anticipating strategy – it is an Ito integral – and hence the expectation comes entirely from the second term.

5. Corresponding Non-dimensional Objective Utility: Thus

$$\mathbb{E}[\hat{\mathcal{C}}] = \mu \mathbb{E} \left[ \int_0^1 \hat{v}^2(\hat{t}) d\hat{t} \right]$$



and the complete risk-aversion cost function is approximately

$$\mathbb{E}[\hat{\mathcal{C}}] + \mu\bar{\kappa}^2\mathbb{V}[\hat{\mathcal{C}}] \sim \mu \int_0^1 \mathbb{E}[\hat{v}^2(\hat{t}) + \bar{\kappa}^2\hat{x}^2(\hat{t})] d\hat{t}$$

$$\mu \rightarrow 0$$

6. Quadratic Nature of Objective Utility: Consider a candidate adaptive strategy  $\hat{x}(\hat{t})$ . Since the quadratic is convex the static strategy

$$\bar{x}(\hat{t}) = \mathbb{E}[\hat{x}(\hat{t})]$$

will give a lower value of the objective function (thus  $x(t)$  and  $v(t)$  have limits, thereby justifying the original assumption).

7. Consequence of the “Market Power”: When  $\mu$  is not small adaptive strategies can create negative correlation between the two terms in

$$\hat{\mathcal{C}} = \int_0^1 \hat{x}(\hat{t}) d\hat{B}(\hat{t}) + \mu \int_0^1 \hat{v}^2(\hat{t}) d\hat{t}$$

thereby reducing the overall variance below its value for purely static strategies.

## Portfolio Comparison

1. Optimal Strategy Portfolio Size Dependence: In the simplest form, the goal is to determine the optimal strategy  $\hat{x}(\hat{t})$  for any specific set of parameters. But to understand results it is useful to compare strategies and costs for portfolios of different sizes.



2. Quadratic Scaling of Static Trajectories: Consider two portfolios  $X_1$  and  $X_2$  with

$$X_2 = 2X_1$$

and all other parameters the same including risk aversion; thus

$$\mu_2 = 2\mu_1$$

and  $\bar{\kappa}$  is the same. Portfolio  $X_2$  will in general cost four times to trade as much as portfolio  $X_1$ . For example, static trajectories for the two portfolios will have identical shapes, and the cost will satisfy

$$\mathbb{E}[\hat{\mathcal{C}}_2] = 4\mathbb{E}[\hat{\mathcal{C}}_1]$$

and

$$\mathbb{V}[\hat{\mathcal{C}}_2] = 4\mathbb{V}[\hat{\mathcal{C}}_1]$$

3. Adaptive Strategies: Sub-Quadratic Scaling: For adaptive strategies, the large portfolio is still more expensive to trade than the small portfolio, but it can take advantage of the negative correlation. Thus one will have

$$\mathbb{E}[\hat{\mathcal{C}}_2] + \lambda\mathbb{V}[\hat{\mathcal{C}}_2] \leq 4(\mathbb{E}[\hat{\mathcal{C}}_1] + \lambda\mathbb{V}[\hat{\mathcal{C}}_1])$$

for each  $\lambda$  although it is generally not true that separately

$$\mathbb{E}[\hat{\mathcal{C}}_2] < 4\mathbb{E}[\hat{\mathcal{C}}_1]$$

AND



$$\mathbb{V}[\hat{\mathcal{C}}_2] < 4\mathbb{V}[\hat{\mathcal{C}}_1]$$

4. Cost Ratio - Adaptive to Static: The ratio of an adaptive cost to a static cost will be less for a large portfolio than for a small portfolio, though all costs are higher for the large portfolio. Therefore these solutions will be of most interest to the large investors.
5. Representative Illustration of the Relative Costs: To highlight the differences in relative costs, Almgren and Lorenz (2007) draw efficient frontiers which show the expectation of the cost and its variance *relative* to their values for the linear trajectory.
6. Static Strategies correspond to  $\mu = 0$ : Thus the static efficient frontiers for all the values of

$$\mu > 0$$

super-impose, since the cost of all static trajectories scale precisely as  $X^2$ . This common static frontier appears as the limit of adaptive frontiers as

$$\mu \rightarrow 0$$

As  $\mu$  increases the adaptive frontiers move down and to the left, away from the static frontier.

## Single Update

1. Non-dimensional Decision Time Instant: Here the assumption is that the urgency update occurs at a single update time  $\hat{T}_*$  where

$$0 < \hat{T}_* < 1$$



2. Starting with the Initial Urgency: On the first trading period

$$0 < \hat{t} < \hat{T}_*$$

an initial urgency  $\bar{\kappa}_0$  is used, that is, the trajectory is

$$\hat{x}(\hat{t}) = h(\hat{t}, 1, \bar{\kappa}_0)$$

with  $h$  from

$$h(\hat{t}, \hat{T}, \bar{\kappa}) = \frac{\sinh[\bar{\kappa}(\hat{T} - \hat{t})]}{\sinh[\bar{\kappa}\hat{T}]}$$

for

$$0 < \hat{t} < \hat{T}_*$$

3. The Set of Decision Urgencies: Let

$$\hat{x}_*(\bar{\kappa}_0, \hat{T}_*) = h(\hat{T}_*, 1, \bar{\kappa}_0)$$

be the shares remaining at the decision time. At time  $\hat{T}_*$  one switches to one of the  $n$  new urgencies  $\bar{\kappa}_1, \dots, \bar{\kappa}_n$ ; with urgency  $\bar{\kappa}_i$  one sets

$$\hat{x}(\hat{t}) = \hat{x}_*(\bar{\kappa}_0, \hat{T}_*)h(\hat{t} - \hat{T}_*, 1 - \hat{T}_*, \bar{\kappa}_i)$$

for

$$\hat{T}_* < \hat{t} < 1$$



4. **Urgency Based on Realized Cost:** The new urgency is chosen based on the non-dimensional realized cost up to time  $\hat{T}_*$ :

$$\hat{\mathcal{C}}_0 = \int_0^{\hat{T}_*} \hat{x}(\hat{t}) d\hat{B}(\hat{t}) + \mu \int_0^{\hat{T}_*} \hat{v}^2(\hat{t}) d\hat{t} + \int_{\hat{T}_*}^1 [\hat{B}(\hat{t}) + \mu \hat{v}(\hat{t})] \hat{v}(\hat{t}) d\hat{t} + \hat{x}_* \hat{B}(\hat{T}_*)$$

5. **Trajectory Cost Decomposition - Terms Explain:** To measure  $\hat{\mathcal{C}}_0$  at time  $\hat{T}_*$ , as can be seen in the second expression above, the first term is the total dollar cost paid to acquire the shares so far, minus the value of those shares at the pre-trade price.
6. **Trading Cost Decomposition - Position Remaining:** The second term is the estimation of the additional cost that will need to be paid on the remaining shares relative to the pre-trade price, due to price movements observed so far.
7. **Observability of the Realized Price Brownian:** As noted before  $\hat{B}(\hat{t})$  is observable if the execution price, the trade rate, and the coefficient of the market impact are all known.
8. **Outcome Partitioning at the Decision Instant:** The real-line is partitioned into  $n$  intervals  $I_1, \dots, I_n$  and  $\bar{\kappa}_j$  is used if

$$\hat{\mathcal{C}}_0 \in I_j$$

For large  $n$  this approaches a continuous dependence

$$\bar{\kappa} = f(\hat{\mathcal{C}}_0)$$

9. **Price vs Cost Bound  $\bar{\kappa}$ :** The intuition seen in the Introduction section indicates that using the accumulated cost should be more effective than using the instantaneous price at time  $\hat{T}_*$ .



10. Pre-fixing Decision Time Urgencies: Before trading begins, the decision time  $\hat{T}_*$ , the interval break-points, and the  $n + 1$  urgencies  $\bar{\kappa}_1, \dots, \bar{\kappa}_n$  are all fixed. However it is not known which trajectory shall actually be executed until  $\hat{\mathcal{C}}_0$  is observed at time  $\hat{T}_*$ .
11. Cost of the Decision Trajectory:  $\hat{\mathcal{C}}_j$  is the cost incurred in the second part of the trajectory if urgency  $\bar{\kappa}_j$  is used:

$$\hat{\mathcal{C}}_j = \int_{\hat{T}_*}^1 \hat{x}(\hat{t}) d\hat{B}(\hat{t}) + \mu \int_{\hat{T}_*}^1 \hat{v}^2(\hat{t}) d\hat{t} + \int_{\hat{T}_*}^1 [\hat{B}(\hat{t}) + \mu \hat{v}(\hat{t})] \hat{v}(\hat{t}) d\hat{t} - \hat{x}_* \hat{B}(\hat{T}_*)$$

12. Total Cost across the Trajectory: The total cost is then

$$\hat{\mathcal{C}} = \hat{\mathcal{C}}_j + \hat{\mathcal{C}}_{\mathcal{J}(\hat{\mathcal{C}}_0)}$$

where

$$\mathcal{J}(\hat{\mathcal{C}}_0) = i$$

if

$$\hat{\mathcal{C}}_0 \in I_i$$

Although the total cost is not Gaussian the optimal frontier is still computed using mean-variance optimization.

## Single Update Mean and Variance

1. Initial Time Decision Mean/Variance: As described before the mean and the variance are calculated at the initial time. Each variable  $\hat{\mathcal{C}}_i$  is Gaussian with mean



$$E_i = \mu F_i$$

and variance  $V_i$  where  $F_i$  and  $V_i$  are integrals of the form

$$F_i = \int_{\hat{T}_*}^1 \hat{v}^2(\hat{t}) d\hat{t}$$

and

$$V_i = \int_{\hat{T}_*}^1 \hat{x}^2(\hat{t}) d\hat{t}$$

which do not depend on  $\mu$ .

2. Initial and Decision Marginal Distributions: Next the mean and the variance of each possible decision trajectory cost needs to be evaluated.
3. Decision Cost Trajectory Mean and Variance: The following integrals are readily determined:

$$F_0 = \bar{\kappa}_0 \frac{\sinh(2\bar{\kappa}_0) - \sinh[2\bar{\kappa}_0(1 - \hat{T}_*)] + 2\bar{\kappa}_0\hat{T}_*}{4 \sinh^2 \bar{\kappa}_0}$$

$$V_0 = \frac{\sinh(2\bar{\kappa}_0) - \sinh[2\bar{\kappa}_0(1 - \hat{T}_*)] - 2\bar{\kappa}_0\hat{T}_*}{4 \sinh^2 \bar{\kappa}_0}$$

and

$$F_i = \frac{\sinh^2[\bar{\kappa}_0(1 - \hat{T}_*)]}{\sinh^2[\bar{\kappa}_i(1 - \hat{T}_*)]} \bar{\kappa}_i \frac{\sinh[2\bar{\kappa}_i(1 - \hat{T}_*)] + 2\bar{\kappa}_i(1 - \hat{T}_*)}{4 \sinh^2 \bar{\kappa}_0}$$



and

$$V_i = \frac{\sinh^2[\bar{\kappa}_0(1 - \hat{T}_*)] \sinh[2\bar{\kappa}_i(1 - \hat{T}_*)] - 2\bar{\kappa}_i(1 - \hat{T}_*)}{\sinh^2[\bar{\kappa}_i(1 - \hat{T}_*)]} \frac{4 \sinh^2 \bar{\kappa}_0}{4 \sinh^2 \bar{\kappa}_0}$$

for

$$i = 1, \dots, n$$

4. Trajectory Cost Distribution Density Expression: Each  $\hat{\mathcal{C}}_i$  is a Gaussian with a mean  $F_i$  and a variance  $V_i$  so its density is

$$f_i(\hat{\mathcal{C}}_i) = \frac{1}{\sqrt{2\pi V_i}} e^{-\frac{(F_i - \hat{\mathcal{C}}_i)^2}{2V_i}}$$

$$i = 1, \dots, n$$

5. Partitioning the Cost Decision Space: The intervals are defined as

$$I_i = \{b_{j-1} < \hat{\mathcal{C}}_0 < b_j\}$$

with

$$b_j = E_0 + a_j \sqrt{V_0}$$

where  $a_0, \dots, a_n$  are fixed constants with

$$a_0 = -\infty$$



and

$$a_n = +\infty$$

#### 6. Cost Convolution over Decision Segments:

$$\begin{aligned} f(c)\Delta c &= \text{Prob}\{\hat{C} \in [c, c + \Delta c]\} \\ &= \sum_{i=1}^n \text{Prob}\{\hat{C}_0 \in I_i \text{ AND } \hat{C}_i \in [c - \hat{C}_0, c - \hat{C}_0 + \Delta c]\} \end{aligned}$$

so

$$\begin{aligned} f(\hat{C}) &= \sum_{i=1}^n \int_{b_{i-1}}^{b_i} f(\hat{C}_0) f(\hat{C} - \hat{C}_0) d\hat{C}_0 \\ &= \sum_{i=1}^n \frac{1}{\sqrt{2\pi V_0 V_i}} \int_{b_{i-1}}^{b_i} e^{-\left[\frac{(\hat{C}_0 - E_0)^2}{2V_0} + \frac{(\hat{C} - \hat{C}_0 - E_i)^2}{2V_i}\right]} d\hat{C}_0 \\ &= \sum_{i=1}^n \frac{1}{\sqrt{2\pi V_0 V_i}} e^{-\frac{1}{2}\left[\frac{E_0^2}{V_0} + \frac{(\hat{C} - E_i)^2}{V_i} - \frac{\{E_0 V_i + (\hat{C} - E_i) V_0\}^2}{V_0 V_i (V_0 + V_i)}\right]} \int_{b_{i-1}}^{b_i} e^{-\frac{1}{2} \frac{V_0 + V_i}{V_0 V_i} \left[\hat{C}_0 - \frac{\{E_0 V_i + (\hat{C} - E_i) V_0\}^2}{V_0 + V_i}\right]} d\hat{C}_0 \\ &= \sum_{i=1}^n \frac{1}{\sqrt{2\pi(V_0 + V_i)}} e^{-\frac{(\hat{C} - \hat{C}_0 - E_i)^2}{2V_i}} \\ &\quad \times \left[ \Phi\left(\frac{\{\hat{C} - E_i - b_{i-1}\}V_0 + \{E_0 - b_{i-1}\}V_i}{\sqrt{V_0 V_i (V_0 + V_i)}}\right) \right. \\ &\quad \left. - \Phi\left(\frac{\{\hat{C} - E_i - b_i\}V_0 + \{E_0 - b_i\}V_i}{\sqrt{V_0 V_i (V_0 + V_i)}}\right) \right] \end{aligned}$$

#### 7. Incremental Cost Distribution and Density: To calculate the mean and the variance of the composite cost $\hat{C}$ the following non-dimensional fixed costs are defined.



$$p_j = \Phi(a_j) - \Phi(a_{j-1})$$

and

$$q_j = \phi(a_{j-1}) - \phi(a_j)$$

for

$$j = 1, \dots, n$$

$\phi$  is the standard normal density, and  $\Phi$  is its cumulative. Thus

$$\text{Prob}\{\hat{C}_0 \in I_i\} = p_j$$

and

$$\mathbb{E}[\hat{C}_0 | \hat{C}_0 \in I_i] = E_0 + \frac{q_j}{p_j} \sqrt{V_0}$$

8. Total Cost Mean and Variance: By linearity of expectation one readily gets

$$E = \mu(F_0 + \bar{F})$$

with

$$\bar{F} = \sum p_i F_i$$



The variance is more complicated because of the dependence between the two terms in

$$\hat{C} = \hat{C}_0 + \hat{C}_{J(\hat{C}_0)}$$

9. The Full Trajectory Cost Variance: One uses the conditional variance expression

$$\mathbb{V}[X] = \mathbb{E}[\mathbb{V}[X|Y]] + \mathbb{V}[\mathbb{E}[X|Y]]$$

to write, using

$$\bar{V} = \sum p_i V_i$$

$$\begin{aligned}\mathbb{V}[\hat{C}] &= \mathbb{E}\left[\mathbb{V}\left[\hat{C}_0 + \hat{C}_{J(\hat{C}_0)}|\hat{C}_0\right]\right] + \mathbb{V}\left[\mathbb{E}\left[\hat{C}_0 + \hat{C}_{J(\hat{C}_0)}|\hat{C}_0\right]\right] \\ &= \mathbb{E}\left[\mathbb{V}\left[\hat{C}_{J(\hat{C}_0)}\right]\right] + \mathbb{V}\left[\hat{C}_0 + \mathbb{E}\left[\hat{C}_{J(\hat{C}_0)}\right]\right] \\ &= \bar{V} + \mathbb{V}[\hat{C}_0] + 2 \operatorname{Covar}\left[\hat{C}_0, \hat{C}_{J(\hat{C}_0)}\right] + \mathbb{V}\left[\mathbb{E}\left[\hat{C}_{J(\hat{C}_0)}\right]\right]\end{aligned}$$

10. Full Trajectory Cost Variance Components: By definition

$$\mathbb{V}[\hat{C}_0] = V_0$$

and

$$\mathbb{V}\left[\mathbb{E}\left[\hat{C}_{J(\hat{C}_0)}\right]\right] = \mu^2 \sum p_i (F_i - \bar{F})^2$$

Further



$$\begin{aligned}
Covar[\hat{\mathcal{C}}_0, \hat{\mathcal{C}}_{\mathcal{I}(\hat{\mathcal{C}}_0)}] &= \mathbb{E}\left[\hat{\mathcal{C}}_0 \mathbb{E}\left[\hat{\mathcal{C}}_{\mathcal{I}(\hat{\mathcal{C}}_0)}\right]\right] - \mathbb{E}[\hat{\mathcal{C}}_0] \mathbb{E}\left[\hat{\mathcal{C}}_{\mathcal{I}(\hat{\mathcal{C}}_0)}\right] \\
&= \sum Prob\{\hat{\mathcal{C}}_0 \in I_i\} \mathbb{E}\left[\hat{\mathcal{C}}_0 \mathbb{E}\left[\hat{\mathcal{C}}_{\mathcal{I}(\hat{\mathcal{C}}_0)}\right] | \hat{\mathcal{C}}_0 \in I_i\right] - \mathbb{E}[\hat{\mathcal{C}}_0] \mathbb{E}\left[\hat{\mathcal{C}}_{\mathcal{I}(\hat{\mathcal{C}}_0)}\right] \\
&= \mu \sqrt{V_0} \sum q_i F_i
\end{aligned}$$

11. Bringing all the Parts Together: Putting all this together one gets

$$V = \mathbb{V}[\hat{\mathcal{C}}] = V_0 + \bar{V} + \mu \sqrt{V_0} \sum q_i F_i + \mu^2 \sum p_i (F_i - \bar{F})^2$$

12. The Non-dimensional Objective Function: The overall objective function is

$$U = \frac{E + \mu \bar{\kappa}^2 V}{\mu}$$

or

$$\begin{aligned}
U(\bar{\kappa}_1, \dots, \bar{\kappa}_n, \hat{T}_*; \bar{\kappa}, \mu) \\
= F_0 + \bar{F} + \bar{\kappa}^2 (V_0 + \bar{V}) + \mu \bar{\kappa}^2 \sqrt{V_0} \sum q_i F_i + \mu^2 \bar{\kappa}^2 \sum p_i (F_i - \bar{F})^2
\end{aligned}$$

13. Negative Two Period Cross Correlation: The  $\mathcal{O}(\mu)$  term is approximately

$2 \sum (\hat{\mathcal{C}}_0 - E_0) p_i E_i$  and can be made negative by making  $E_i$  negatively related to  $\hat{\mathcal{C}}_0$  corresponding to anti-correlation between second period impact costs and the first period trading losses.

14. The Optimizer Input/Search Space: For a given market power  $\mu$  and static urgency  $\bar{\kappa}$ ,  $U$  is minimized numerically over  $\bar{\kappa}_0, \dots, \bar{\kappa}_n$  and the decision time  $\hat{T}_*$

15. Efficient Frontier Curve over  $\bar{\kappa}$ : As  $\bar{\kappa}$  varies the resulting set of points  $(V, E)$  traces the efficient frontier. This results in a one-parameter family of efficient frontiers, depending on  $\mu$ . The static trajectories appear at the limit of



$$\mu = 0$$

## Almgren and Lorenz (2007) Results

1. Decision Urgency Based Efficient Frontier: Almgren and Lorenz (2000) illustrate the complete set of efficient frontiers for the single update problem. Each curve is computed by varying a static urgency parameter  $\bar{\kappa}$  from 0 to  $\infty$  for a fixed value of  $\mu$ .
2. Discretization of the Decision Urgency: The solution for each pair of  $(\bar{\kappa}, \mu)$  is computed using a fixed set of 32 equal-probability breakpoints. As described earlier  $\hat{E}$  and  $\hat{V}$  are plotted relative to their values for the linear trajectories to clearly see the improvement due to the adaptivity.
3. Improved Execution Strategy Cost Distribution: The frontiers are used to obtain adaptive strategies that are better than the cost distribution for any static strategies.
4. Static Urgency Trajectory and Cost: First Almgren and Lorenz (2007) compute a static trajectory using

$$\bar{\kappa} = 8$$

and generate the resulting cost distribution that is Gaussian. For a portfolio with

$$\mu = 0.1$$

this distribution has an expectation

$$\hat{E} \approx 4 \times \hat{E}_{LIN} \approx 4 \times \mu = 0.4$$

and variance



$$\hat{V} \approx 0.2 \times \hat{V}_{LIN} = \frac{0.2}{3} = 0.067$$

5. Market Power Based Efficient Frontier: Likewise they generate adaptive efficient frontiers for different values of the market power  $\mu$ . They identify the set of values accessible to a static strategy as well as the static frontier – which is also the limit

$$\mu \rightarrow 0$$

with a static strategy  $\bar{\kappa}$ . The improved values accessible to the adaptive strategies are also identified; the improvement is greater for larger portfolios. The actual cost distributions corresponding to different  $\bar{\kappa}$  are also estimated.

6. Improvement available over the Static Trajectory: The region in the  $(\hat{V}, \hat{E})$  space accessible to an adaptive strategy with

$$\mu = 0.1$$

that are strictly preferable to a static strategy since they have lower expected cost and/or variance can be readily observed.

7. Cost Profile Adaptive Urgency Range: On the efficient frontier for

$$\mu = 0.1$$

these solutions are obtained by computing adaptive solutions with parameters approximately in the range

$$4.9 \leq \bar{\kappa} \leq 7.1$$

There is no need to use the same value for  $\bar{\kappa}$  for the adaptive strategy as for the static strategy to which it is compared.



8. Adaptive Trajectory Urgency - Cost/Variance: Detailed cost distributions associated with these adaptive strategies can also be generated. For

$$\bar{\kappa} = 4.9$$

the adaptive distribution has a lower expected cost than the static distribution with the same variance. For

$$\bar{\kappa} = 7.1$$

the adaptive distribution has a lower variance than the static distribution with the same mean.

9. Strictly Optimal Adaptive Strategy Urgency: These distributions are the extreme points of a one-parameter family of distributions, each of which is strictly preferable to the given static strategy, regardless of the traders' risk preferences. For example the adaptive solution for

$$\bar{\kappa} = 6$$

has both lower expected cost and lower variance than the static distribution.

10. Strongly Positive Optimal Distribution Skew: These cost distributions are strongly skewed toward positive distribution costs suggesting that the mean-variance optimization may not give the best possible solutions.
11. Realized Static vs. Adaptive Trajectories: Almgren and Lorenz (2007) compare the adaptive trading trajectories for

$$\mu = 0.1$$

and



$$\bar{\kappa} = 6$$

against the static optimal trajectory with urgency

$$\bar{\kappa} = 8$$

The adaptive strategy clearly delivers both lower expectation of cost and lower variance.

12. Urgency Dependence on Trading Cost: They also demonstrate the dependence of the decision urgency on the initial trading cost  $\hat{C}_0$  - in their plots they normalize  $\hat{C}_0$  by its initial *ex ante* expectation and trading cost.
13. Adaptation under Favorable Price Move: The adaptive strategy initially trades more slowly than the optimal static trajectory. At  $\hat{T}_*$ , if the prices have moved in the traders' favor, the adaptive strategy accelerates, spending the investment gains on the impact costs.
14. Adaptation under Unfavorable Price Move: If the prices have moved against the trader, corresponding to positive values of  $\hat{C}_0$ , then the strategy decelerates to save impact costs in the remaining period. The values of  $\bar{\kappa}$  become very large when  $\hat{C}_0$  is large negative, corresponding to the instruction: "if you have gains in the first part of the trading, then finish the program immediately".

## Continuous Response

1.  $\mathbb{R}^1 \rightarrow \mathbb{R}^1$  Dependence on Brownian: Next Almgren and Lorenz (2007) illustrate a simple form of *continuous response* to trading gains or losses. In general one can specify any rule  $\hat{v}(\hat{t})$  as a function of the price history  $\hat{B}(s)$  for

$$0 \leq s \leq \hat{t}$$



Rather than adjusting the rate  $\hat{v}(\hat{t})$  directly it is more convenient to adjust  $\bar{\kappa}$ .

2. Trade Rate Explicit Functional Form: From

$$h(t, T, \kappa) = \frac{\sinh[\kappa(T - t)]}{\sinh[\kappa T]}$$

for

$$0 \leq t \leq T$$

on differentiating

$$x(s) = x(t)h(s - t, 1 - t, \kappa)$$

with respect to  $s$  and evaluating at

$$s = t$$

the following relationship between  $v$  and  $\kappa$  is obtained.

$$v(t) = x(t)\kappa(t) \coth[\kappa(t)(1 - t)]$$

For all choices of  $\kappa(t)$  the trajectories hit

$$x = 0$$

at

$$t = 1$$



3. Exponential Price Brownian Functional Form: Determining the full optimal dependence of  $\kappa(t)$  on  $B(s)$  for

$$0 \leq s \leq t$$

is difficult. Thus the following relationship is considered:

$$\kappa(t) = ae^{bB(t)}$$

Thus the instantaneous urgency depends on the instantaneous price level. Other functional relationships for  $\kappa(t)$  in terms of  $B(t)$  are possible as well. Here  $\kappa(t)$  is always positive, and is monotone in  $B(t)$ .

4. Corresponding Trade Rate Shortfall: From

$$v(t) = x(t)\kappa(t) \coth[\kappa(t)(1-t)]$$

one readily obtains  $x(t)$  and finally the shortfall  $\mathcal{C}$  by integration as in

$$\mathcal{C} = \sigma \int_0^T x(t) dB(t) + \eta \int_0^T v^2(t) dt$$

5. Need for Numerical Framework: However, because of the highly nonlinear dependence of  $\kappa(t)$ , and thus  $v(t)$  and  $x(t)$ , on the Brownian motion  $B(t)$ , analytic solution of this stochastic integral is beyond reach.

## Continuous Response Numerical Results



1. Price Move Brownian Bridge Construction: For numerical solutions one generates a fixed collection of sample paths using a Browning bridge construction with quasi-random variables.
2. Objective Value Function Numerical Evaluation: For any candidate values of  $a$  and  $b$  the stochastic integrals are evaluated numerically, and the sample mean  $E$  and the variance  $V$  are calculated. The objective function  $E + \bar{\kappa}^2 \mu^2 V$  is then numerically minimized over  $a$  and  $b$ .
3. Generation of the Efficient Frontier: By solving for a series of values of

$$0 < \bar{\kappa} < \infty$$

the efficient frontier can again be traced for different values of  $\mu$ , yielding similar results as in the single update framework.

4. Execution Cost Gain/Loss Adaptation: Again the optimal strategies are “aggressive in the money”, having

$$b < 0$$

When the stock price goes down, an unexpected smaller shortfall is incurred, and a reaction occurs with increasing urgency  $\bar{\kappa}(t)$  whereas for rising stock prices the trading is slowed down.

5. Exponential Urgency Response Trajectory Sample: As an illustration, Almgren and Lorenz (2007) generate optimal trading trajectories using the adaptation rule

$$\bar{\kappa}(t) = ae^{bB(t)}$$

with

$$a = 5.9$$



and

$$b = -1.7$$

for a static urgency

$$\bar{\kappa} = 6$$

As the stock price goes down the trading is accelerated compared to the optimal static trajectory, whereas for rising stock price it is slowed down.

## Discussion and Conclusions

1. Rule Based Adaptive Scaling Strategies: The simple update rules presented in the previous sections demonstrate that price adaptive scaling strategies can lead to significant improvements over static trade schedules, and illustrate the importance of the market power parameter  $\mu$ .
2. Dynamic Programming Based Optimal Trajectory: However neither of these rules is the fully adaptive optimal trading strategy. A fully optimal adaptive trading strategy would use stochastic dynamic programming to determine the trading rate as a general function of the continuous state variables such as the number of shares remaining, time remaining, current stock price, and trading gains or losses experienced to date.
3. Infeasibility of Mean Variance Optimization: One subtlety is that the mean-variance optimization cannot be used directly in this context; it involves the square of an expectation, which is not amenable to dynamics programming techniques.
4. Quadratic Utility Family of Optimization: However Li and Ng (2000) have shown how to embed mean-variance optimization into a family of optimizations that use the quadratic utility function.



5. MVO as a Family Member: The mean-variance solution is recovered as one element of this family. The need to solve this family of problems is an additional degree of complication.
6. HJB PDE Based Stochastic Control: The calculation uses tools of optimal stochastic control and requires the numerical solution of a highly nonlinear Hamilton-Jacobi-Bellman partial differential equation.
7. Adaptive Strategies as a Simplifier: Partial formulation of this problem, and the solution of the resulting equations, is an involved undertaking and the focus of a later chapter. The examples shown here show that even with very simple adaptive strategies substantial improvement is possible over static strategies.

## References

- Almgren, R. F., and N. Chriss (1999): Value under Liquidation *Risk* **12 (12)** 61-63.
- Almgren, R. F., and N. Chriss (2000): Optimal Execution of Portfolio Transactions *Journal of Risk* **3 (2)** 5-39.
- Almgren, R. F., and J. Lorenz (2007): Adaptive Arrival Price, in: *Algorithmic Trading III (B. R. Bruce, editor)* **Institutional Investor** 59-66.
- Huberman, G., and W. Stanzl (2005): Optimal Liquidity Trading *Review of Finance* **9 (2)** 165-200.
- Kahneman, D., and A. Tversky (1979): Prospect Theory: An Analysis of Decision under Risk *Econometrica* **47 (2)** 263-291.
- Kissell, R., and R. Malamut (2006): Algorithmic Decision-Making Framework *Journal of Trading* **1 (1)** 12-21.
- Li, D., and W. L. Ng (2000): Optimal Dynamic Portfolio Selection: Multi-period Mean-Variance Formulation *Mathematical Formulation* **10 (3)** 387-406.
- Perold, A. F. (1988): The Implementation Shortfall: Paper versus Reality *Journal of Portfolio Management* **14 (3)** 4-9.



- Shefrin, H., and M. Statman (1985): The Disposition to sell Winners too early and ride Losers too Long: Theory and Evidence *Journal of Finance* **40** (3) 777-790.



## Mean Variance Optimal Adaptive Execution

### Background, Synopsis, and Key Results

1. Trade Cost Expectation vs. Volatility: Electronic trading of equities and other securities makes heavy use of “arrival price” algorithms that balance market impact costs of rapid execution against the volatility risk of slow execution.
2. Static Optimal MVO Execution Strategies: In the standard formulation, mean variance optimal trading strategies are static, they do not modify the execution speeds in response to the price motions observed during trading.
3. Dynamic Optimal MVO Execution Strategies: Lorenz and Almgren (2011) show that substantial improvement is possible by using dynamic trading strategies, and that the improvement is larger for large initial positions.
4. Tree Discretization of Asset Price: They develop a technique for computing optimal dynamic strategies to any desired degree of precision. The asset price process is observed on a discrete tree with an arbitrary number of levels.
5. Control Variables for Dynamic Optimization: A novel dynamic programming techniques is introduced in which the control variables are not only the shares traded at each step, but also the maximum expected cost for the remainder of the program; the value function is the variance for the remainder of the program.
6. Aggressive-in-the-Money Strategies: The resulting adaptive strategies are “aggressive-in-the-money”; they accelerate the execution when the price moves in the traders’ favor, spending parts of the trading gains to reduce risk.

### References



- Lorenz, J., and R. F. Almgren (2011): Mean-Variance Optimal Adaptive Execution  
*Applied Mathematical Finance* **18 (5)** 395-422.



## Optimal Trading in a Dynamic Market

### Introduction, Overview, and Motivation

1. Stock Market Liquidity and Volatility: This chapter considers the problem of mean-variance agency execution strategies when the market volatility and the liquidity vary randomly in time.
2. Cost/Strategy Dynamic Optimal Trajectory: Under specific assumptions for the stochastic processes satisfied by these parameters, a Hamilton-Jacobi-Bellman equation is presented for the optimal cost and the strategy.
3. Trajectory Adoption to Market Conditions: This equation is solved numerically, and optimal strategies are illustrated for varying risk aversion. These strategies adapt optimally to the instantaneous variations of market quality.

### Limitations of Arrival Price Frameworks

1. Optimal Order Execution Trade Scheduling: A fundamental part of agency algorithmic trading in equities and other asset classes is trade scheduling. Given a trade target, that is a number of shares of a trade that must be bought or sold before a fixed time horizon, trade scheduling refers to how many shares that will be bought or sold by each time instant between the beginning of trading and the horizon.
2. Optimal Measure of Execution Quality: The optimal execution is done so as to optimize some measure of execution quality, usually measured as the final average execution price relative to some benchmark price.
3. Definition of the Arrival Price: One of the most popular benchmarks is the “arrival price”, i.e., the price prevailing in the market the time the order was received into the



trading system. The difference between the execution price and this pre-trade price is the *implementation shortfall* (Perold (1988)) or *slippage*.

4. Balancing Transaction Costs and Variance: Grinold and Kahn (1995) and Almgren and Chriss (2000) suggested that the optimal trajectory could be determined by balancing the market impact cost, which leads toward slow trading, versus volatility risk, which leads toward rapid completion of the order.
5. Risk Aversion Based Efficient Frontier: This framework leads to an efficient frontier in which the trade schedule is selected from a one-parameter family based on a risk-aversion parameter that must be specified by the trading client.
6. Risk Aversion Based Front Loading: Optimal trading strategies are typically *front-loaded*. They execute as much as possible early in the program to reduce risk relative to the benchmark price. The degree of front-loading depends on the risk aversion parameter that must be specified by the trading client. The exact shape of the schedule depends on the form of the market impact model.
7. Market Price Based Benchmarks: The largest alternative category of benchmarks is composed of some form of average market price during the trading interval; usually either time-weighted average price (TWAP) or volume weighted average price (VWAP).
8. Close Tailing of the Benchmarks: For these benchmarks optimal strategy follows the benchmarks quite closely, since deviation from the profile both increases the risk relative to the benchmark and the impact costs. Determining optimal response to the short-term price is an interesting topic for optimization, but is not the focus here.
9. Use of Arrival Price Frameworks: While other factors such as anticipated price drift, serial correlation or other short-term signals, and daily patterns are certainly important, this fundamental “arrival price” framework has proven remarkably robust and useful in designing practical trading systems.
10. Time Profiles of Liquidity/Volatility: A fundamental assumption of most of this work has been that the market parameters are constant, or at least have known predictable profiles. This assumption is reasonably accurate for large-cap US stocks.



11. Static Nature of the Framework: Under that assumptions optimal strategies are *static*; that is the trade schedule can be determined before the trading starts and is not modified by the new information revealed by price moves during trading. Almgren and Chriss (2000) did consider a model in which the market parameters updated at a single time to one of a known set of possible new values.
12. Algorithmic Trading of Less Liquid Assets: over the last few years, a major push of providers of algorithmic trading services has been to extend their functionality to smaller stocks and less liquid assets.
13. Random Intra-day Volatility/Liquidity: A distinguishing feature of these assets is that their liquidity and volatility vary randomly in time. That is, there will be times during the trading day when trading becomes very expensive, and times when trading is cheap; similarly there will be times when delaying trading introduces large amounts of volatility risk and other times when the delay is relatively costless.
14. Optimal Mean-Variance Tradeoff: The modeling challenge is to determine optimal strategies that adapt to the instantaneous market state, while retaining the mean-variance trade-off inherent in the arrival price framework.
15. Continuous Time and State Treatment: Walia (2006) solved this problem in a discrete time discrete state model. Almgren (2009) provides a systematic mathematical solution to the problem in continuous time and continuous state.
16. Coordinated Liquidity/Volatility Joint Moves: The first section of this chapter presents the basic price and the impact models used, and presents the optimal trading problem. Also presented is the “coordinated variation” approximation in which the liquidity and the volatility vary together, which is very realistic and greatly simplifies the mathematical problem.
17. Dynamic HJB Optimal Cost Function: The second section uses a Hamilton-Jacobi-Bellman PDE describing the optimal cost function and the trade rate. The third section examines some aspects of the numerical solution to this PDE, and presents example solutions.



## The Liquidation Problem

1. The Continuous Holdings Rate Trajectory: The trader begins trading at a time

$$t = 0$$

with a purchase order of  $X$  shares which must be completed by

$$t = T$$

The number of shares remaining to purchase at the time  $t$  is the remaining trajectory  $x(t)$  with

$$x(0) = X$$

and

$$x(T) = 0$$

The rate of buying is

$$v(t) = -\frac{dx(t)}{dt}$$

2. The Trajectory as a Random Variable: Thus, for a buy program

$$X > 0$$



$$x(t) \geq 0$$

and decreasing, and

$$v(t) \geq 0$$

– sell program may be modeled similarly. In general, the trajectory conditions  $x(t)$  may be determined depending on price motions and market conditions discovered during trading, so it is a random variable.

3. The Arithmetic Brownian Price Dynamics: The price  $S(t)$  follows the arithmetic Brownian motion

$$\Delta S(t) = \sigma(t) \Delta B(t)$$

$$S(0) = S_0$$

where  $B(t)$  is a standard Brownian motion, and the instantaneous volatility  $\sigma(t)$  depends on time either deterministically or stochastically.

4. Volatility and Permanent Impact Parameters: Note that  $\sigma(t)$  is an absolute volatility rather than fractional; it contains an implicit factor of the reference price  $S_0$ . It is possible to include the permanent impact into the price equation, but it is not central to the problem.
5. Execution Price - Incorporating the Temporary Impact: The price actually received on each trade is

$$\tilde{S}(t) = S(t) + \eta(t)v(t)$$

where  $\eta(t)$  is the coefficient of the temporary market impact, also time varying. Again  $\eta(t)$  is an absolute coefficient rather than fractional.



6. More Elaborate Market Impact Models: Much richer market impact models have been considered in the literature (Gatheral (2010)), but this simple one is adequate to highlight the response to stochastic liquidity.
7. Estimation of  $\sigma(t)$  and  $\eta(t)$ : Both  $\sigma(t)$  and  $\eta(t)$  are assumed to be observable in real-time with some degree of confidence. There is a variety of techniques available for doing this estimation.
8. Techniques for Estimation of  $\sigma(t)$ : For volatility  $\sigma(t)$  there is an extensive literature on estimation using high-frequency market data (for example, Gatheral and Oomen (2010)). The primary focus there is to find effective means to filter out noise associated with market details such as bid and offer prices so as to obtain reliable estimates on time intervals that are as short as possible. Thus, for example, one could estimate  $\sigma(t)$  by using market data from the preceding five minutes, which would typically contain hundreds of trades and potentially thousands of quote updates.
9. Techniques for Estimation of  $\eta(t)$ : Instantaneous liquidity, the inverse of  $\eta(t)$  is more difficult to estimate, since it is an estimation of what *would* happen if one were to submit trades to the market rather than being an observable in itself. One proxy for the instantaneous trade history would be the realized trade volume over the last few minutes; if more people are trading actively in the market then one would be able to move a given number of shares with less slippage.
10. Trade Volume as Liquidity Proxy: A refined version of the above would be to measure the trade volume at or near the bid price if one is a buyer (or at the ask if one is a seller); large volume there would indicate the presence of a motivated seller and a good opportunity to go in as a buyer with low impact. Although these measures are not quantitatively very precise, they are often adequate to distinguish *good* opportunity from *bad*.
11. Persistence of the Market Properties: Both of these estimators rely on the presence of market properties (volatility and liquidity), so that information about the past provides reasonable forecasts for the future. Such persistence, at least across short horizons, is well documented (Bouchard, Farmer, and Lillo (2009) contain a review).



12. Time Dependent Liquidity/Volatility Patterns: Two broad classes of problems may be addressed. First is the case in which  $\sigma(t)$  and  $\eta(t)$  are both known non-random functions in time. This would accommodate the well-known intra-day profiles of volatility and liquidity; generally markets are more active in the mornings and in the close than in the middle of the day. This case is not the primary focus.
13. Stochastic Liquidity and Volatility Processes: The second case is when the volatility and the liquidity vary randomly through the day, so that  $\sigma(t)$  and  $\eta(t)$  follow some stochastic processes. This effect is very important in small and medium capitalization stocks' algorithmic trading, and other assets that are less heavily traded than the large-cap US stocks.

## Cost of Trading

1. Expression for the Transaction Cost: The *cost of trading* is the total cost paid to purchase  $X$  shares relative to the initial market value of  $XS_0$

$$\mathbb{C} = \int_0^T \tilde{S}(t)v(t)dt - XS_0 = \int_0^T \sigma(t)x(t)dB(t) - \int_0^T \eta(t)v^2(t)dt$$

where

$$x(t) = \int_t^T v(t)dt$$



2. Dynamic Optimal Trading Cost Control: The cost  $\mathbb{C}$  is a random variable, both because of the price uncertainty in  $B(t)$  in the first and because of the liquidity uncertainty. The strategy  $x(t)$  is to tailor the properties of this random variable to meet some optimal criterion.
3. Forward Time Cost of Trading: More generally, starting at time

$$t \geq 0$$

with  $x(t)$  shares remaining to purchase, the cost of a strategy  $x(s)$  on

$$t \leq s \leq T$$

is

$$\mathbb{C} = \int_t^T \sigma(s)x(s)dB(s) + \int_t^T \eta(s)v^2(s)ds$$

4. Trading Cost Expectation and Variance: The optimal trajectory is defined by the mean-variance criterion

$$\min_{x(s): t \leq s \leq T} \{\mathbb{E}[\mathbb{C}] + \lambda \mathbb{V}[\mathbb{C}]\}$$

where

$$\lambda \geq 0$$

is a risk-aversion coefficient. Note that



$$\mathbb{E}[\mathbb{C}] = \int_t^T \eta(s)v^2(s)ds$$

since the first term is an Ito's integral, and

$$\mathbb{V}[\mathbb{C}] = \int_t^T \sigma^2(s)x^2(s)ds + \{Terms from the Uncertainty of \eta(s) and \sigma(s)\}$$

5. Components of the Transaction Cost Variance: The first term in the variance contains the largest source of uncertainty, which corresponds to the price changes during execution. The other terms arise from the uncertainty in the market impact  $\eta(s)$  that will be paid on the transaction in the future, in the volatility  $\sigma(s)$  that will be experienced at a later time, and in the trade strategy  $v(s)$  itself if it is determined in response to uncertain market conditions.
6. Domination of the Market Volatility Term: Almgren (2009) argues that the first term in  $\mathbb{V}[\mathbb{C}]$  above dominates the other terms.
7. Practical Use of Risk Aversion: The risk aversion coefficient  $\lambda$  is rarely defined in terms of fundamental investment preferences (Engle and Ferstenberg (2007)). Rather it is a parameter used to adjust the trajectories to a form that seems reasonable by other criteria such as representing a desired fraction of the market volume.

## Constant Coefficients

1. Constant Volatility and Market Impact: The classic problem of Almgren and Chriss (2000) takes  $\sigma$  and  $\eta$  constant. Then for a strategy  $x(t)$  that is fixed in advance and does not adapt to price motions



$$\mathbb{E}[\mathbb{C}] + \lambda \mathbb{V}[\mathbb{C}] = \int_t^T [\eta(s)v^2(s) + \sigma^2(s)x^2(s)]ds$$

2. Application of the Calculus of the Variations: Using the calculus of variations to minimize this over trajectories  $x(s)$  gives the second order ODE

$$\frac{d^2x}{ds^2} = \kappa^2 x(s)$$

with

$$\kappa^2 = \frac{\lambda\sigma^2}{\eta}$$

3. Optimal Trading Rate and Trajectory: The solution is a combination of the exponentials  $e^{\pm\kappa s}$

$$x(s) = x(t) \frac{\sinh[\kappa(T-s)]}{\sinh[\kappa(T-t)]}$$

$$v(s) = \kappa x(t) \frac{\cosh[\kappa(T-s)]}{\sinh[\kappa(T-t)]}$$

Thus  $\frac{1}{\kappa}$  is the characteristic time scale of liquidation.

4. The Corresponding Cost of Trading: The strategy may also be expressed as a rule for

$$v(t) = \kappa x(t) \coth[\kappa(T-t)]$$

The corresponding cost function is



$$\mathbb{C}(x, t, \eta, \sigma) = \eta \kappa x^2 \coth[\kappa(T - t)] = \eta v x$$

5. Components of the Trading Cost: The total cost is equal to the impact cost component – neglecting the volatility term – incurred by trading  $x$  shares at a price concession given by the instantaneous velocity  $v$ . The actual trajectory slows down as the position size decreases, thus reducing market impact costs, but the total cost includes volatility risk as well as impact costs, giving the above value.
6. Non-dimensionalization of the Time Scales: The shape of the solution is governed by the non-dimensional quantity  $\kappa(T - t)$  – the ratio of time remaining to the intrinsic time scale determined by the market's parameters and the trader's risk aversion.
7. Limit of Long Execution Time: In the infinite horizon limit

$$\kappa(T - t) \gg 1$$

the strategy has the limit

$$x(s) = x(t) e^{-\kappa(s-t)}$$

with

$$v(t) = \kappa x(t)$$

and the cost function

$$\mathbb{C} \rightarrow \eta \kappa x^2$$

Trading is substantially completed well before the expiration, and the precise value of  $T$  is not controlling.



8. Risk Neutral Optimal Execution Trajectory: Note that discounting has not been included, so the only motivation for rapid execution is risk aversion. In the limit of complete risk neutrality

$$\lambda \rightarrow 0$$

minimization of market impact costs would lead the trader to use all available time, and no infinite-horizon limit would exist; since

$$\kappa \rightarrow 0$$

the regime

$$\kappa T \gg 1$$

would never be achieved.

9. Limit of Short Execution Time: In the short horizon limit

$$\kappa(T - t) \ll 1$$

the strategy has the linear form

$$x(s) = x(t) \frac{T - s}{T - t}$$

$$v(s) = \frac{x(t)}{T - t}$$

and the cost function is essentially non-random with the value



$$\mathbb{C} \sim \frac{\eta x^2}{T - t}$$

$$\kappa(T - t) \rightarrow 0$$

If a time change is applied to match the market average profile then this is equivalent to the volume weighted average price (VWAP) execution.

10. Short-Term Limit – Higher Orders: In the same limit, the cost function has the higher order local behavior

$$\mathbb{C} \sim \frac{\eta x^2}{T - t} + \frac{\lambda \sigma^2 x^2}{3} (T - t) + \mathcal{O}((T - t)^3)$$

where  $\sim$  denotes asymptotic equivalence, that is, equal up to the terms that are asymptotically smaller than the displayed expressions in the given limit.

11. Short Term Limit Cost Components: The first term in this expression is the transaction cost associated with selling  $x$  shares at a price concession of

$$\eta v = \frac{\eta x}{T - t}$$

The second term is the risk penalty for holding an average of  $\frac{x^2}{3}$  shares across time  $T - t$ .

12. Rolling Forward Dynamically Optimal Strategies: Whether adaptive strategies are better than a fixed one is a subtle question. Almgren and Chriss (2000) showed that if the strategy is re-evaluated at an intermediate time using the mean and variance measured at that time, then the optimal strategy is the remaining part of the initial strategy, and hence the optimal strategy is fixed. This is the context of Almgren (2009, 2012), since it is appropriate for dynamic programming.

13. Adaptive Strategies for Large Portfolios: In contrast, Almgren and Lorenz (2007), Lorenz (2008), and Lorenz and Almgren (2011) showed that adaptive strategies are



optimal if the mean and the variance are measured at an initial time for portfolios that are large enough so that their impact is a substantial fraction of volatility. This framework is appropriate for *ex post* measurement of historical mean and variance across a large collection of trades. Tse, Forsyth, Kennedy, and Windcliff (2013) have given a fuller description of optimal solutions in the latter framework.

14. Cost Reductions from Adaptive Strategies: Schied and Schoneborn (2009) and Schied, Schoneborn, and Tehranchi (2010) showed that improvement from adaptivity depends on the risk profile; for example, it vanishes for a CARA utility function.

## Coordinated Variation

1. Inverse Relation between Liquidity/Variance: Suppose  $\sigma(t)$  and  $\eta(t)$  vary inversely so that

$$\sigma^2(t)\eta(t) = \text{constant} = \bar{\sigma}^2\bar{\eta}$$

where  $\bar{\sigma}$  and  $\bar{\eta}$  are constant reference values.

2. Joint Arrival trading Time Model: For example, this relationship would be a natural consequence of a *trading time* model (Jones, Kaul, and Lipson (1994), Geman, Madan, and Yor (2001)) in which the single source of uncertainty is the arrival rate of trade events. If each trade event brings both a fixed amount of price variance, and the opportunity to trade a fixed number of shares for a particular cost, then one obtains the above relation.
3. Change of Drift/Wander Variables: Time may then be changed to an artificial variable defined by  $\hat{t}(t)$  defined by

$$\Delta\hat{t} = \sigma^2(t)\Delta t$$

In this time frame a modified Brownian motion  $\hat{B}(\hat{t})$  results, with



$$\Delta \hat{B}(\hat{t}) = \sigma(t) \Delta B(t)$$

4. Change of Variables - The Trading Rate: The holdings are the same trajectory at different times, so

$$\hat{x}(\hat{t}) = x(t)$$

The trade rate is modified to

$$\hat{v}(\hat{t}) = -\frac{d\hat{x}}{d\hat{t}} = \frac{v(t)}{\sigma^2(t)}$$

5. Change of Variables: Transaction Cost: In terms of these new variables the trading cost is

$$\mathbb{C} = \int_0^{\hat{T}} \hat{x}(\hat{t}) d\hat{B}(\hat{t}) + \bar{\sigma}^2 \bar{\eta} \int_0^{\hat{T}} \hat{v}^2(\hat{t}) d\hat{t}$$

where

$$\hat{T} = \hat{t}(T)$$

is the time horizon in the changed variable.

6. Deterministic Liquidity and Volatility Profiles: The above is easy to solve in two cases. First if the time varying volatility and liquidity have known non-random profiles, then the upper bound  $\hat{T}$  may be computed exactly. The problem reduces exactly to the constant coefficient problem



$$\mathbb{E}[\mathbb{C}] + \lambda \mathbb{V}[\mathbb{C}] = \int_t^T [\eta(s)v^2(s) + \sigma^2(s)x^2(s)]ds$$

and the solution is the exponential functions computed there.

7. The Corresponding Trade Rate/Time Scale: The rule

$$v(t) = \kappa x(t) \coth[\kappa(T-t)]$$

becomes

$$v(t) = \kappa(t)x(t) \coth \left[ \frac{\kappa(t)}{\sigma^2(t)} \int_t^T \sigma^2(s)ds \right]$$

where

$$\kappa(t) = \sqrt{\frac{\lambda\sigma^2(t)}{\eta(t)}}$$

is the time scale formed with the instantaneous values of the parameters.

8. Special Random Case - Infinite Horizon: If the coefficients vary randomly, then the problem is not the same as the constant-coefficient problem, because of the uncertainty in the end. But in the infinite horizon case

$$T = \infty$$

one also has

$$\hat{T} = \infty$$



under very mild assumptions on  $\sigma(t)$ . The trade rate is

$$v(t) = \kappa(t)x(t)$$

and the cost is

$$\mathbb{C} = \eta(t)\kappa(t)x^2 = x^2\sqrt{\lambda\bar{\sigma}^2\bar{\eta}}$$

9. Infinite Horizon Trade Rate/Cost: Somewhat surprisingly the optimal cost in the coordinated variation random market infinite horizon case does not depend on the instantaneous market state  $\eta(t)$  and  $\sigma(t)$  though the instantaneous trade rate does depend on the market state.
10. Trading Only Under Favorable Conditions: In effect, since volatility is low whenever the market impact is high, the strategy is always able to wait for favorable market conditions without incurring very much risk from the delay.
11. Absence of Variance/Liquidity Constraint: Thus, the interesting problems come from two sources. First is the case of variation of the profiles of  $\sigma(t)$  and  $\eta(t)$  away from the “base case”

$$\sigma^2(t)\eta(t) = \text{constant}$$

with non-random coefficients. Kim and Boyd (2008) contain a fuller discussion of optimal trading with general market profiles.

12. Departure from “Trading Time” Approximation: For example, even within the *trading time* framework intra-day profiles may vary from the base case because different market participants are active at different times of the day. This leads to problems in the ODE’s that are not the focus of this chapter.
13. Finite Horizon Adaptation to Randomness: With random coefficients, the proper handling of the uncertainty is needed as the end time is approached. For example, if



liquidity is temporarily poor, is it worthwhile to wait for a better opportunity to trade, or is there a large risk that the opportunity will not come before expiration?

## Rolling Time Horizon Approximate Strategy

1. Piecewise Constant Time Realizations: One way to determine a plausible strategy is to use

$$v(t) = \kappa x(t) \coth[\kappa(T - t)]$$

to compute  $v(t)$  using instantaneous values of  $\eta(t)$ ,  $\sigma(t)$ , and hence  $\kappa(t)$ .

2. Explicitly Adapted On-Change Re-evaluation: That is, the assumption is that the values observed at each instant will remain constant throughout the liquidation period, and determine the optimal strategy using those values. When the values change, a stationary solution is re-computed.
3. Piecewise Adaptation Approach - Caveats: This strategy is strictly optimal only in the infinite horizon case, and only when the market parameters co-vary in the appropriate way. In general it is not optimal, but provides a reasonable solution that is easy to implement.

## Small Impact Approximation

1. Approximating the Trading Cost Variance: In order to do dynamic programming when  $\eta(t)$  and  $\sigma(t)$  vary randomly, the variance term needs to be approximated. Almgren (2009) approximates it as



$$\mathbb{V}[\mathbb{C}] = \mathbb{E} \left[ \int_t^T \sigma^2(s) x^2(s) ds \right]$$

That is, the variance comes primarily by the price volatility represented by  $\sigma$  with lesser contributions from the uncertainty in  $\eta(t)$  and  $\sigma(t)$ .

2. The Small Impact Approximation Assumption: When the market impact is small in absolute terms, it is nonetheless significant because it is always positive, but the uncertainty in the market impact is negligible compared to price volatility. This is called the *small-impact approximation*.
3. “Market Power” Non-dimensional Quantity: In the language of Almgren and Lorenz (2007) and Lorenz (2008) this is a small value of the “market power”. This is also equivalent to the “small portfolio” approximation used in Lorenz and Almgren (2011) to neglect uncertainty in impact cost when the portfolio is small enough. Mathematically, the approximation relies on small values of “market power” parameter

$$\mu = \frac{\eta X / T}{\sigma \sqrt{T}}$$

the amount by which trading moves the market compared to its intrinsic motion due to volatility across time  $T$ .

4. The Corresponding Optimal Value Function: The value function is then taken as

$$\mathbb{C}(t, x, \eta, \sigma) = \min_{x(s): t \leq s \leq T} \mathbb{E} \left[ \int_t^T \lambda \sigma^2(s) x^2(s) ds + \int_t^T \eta(s) v^2(s) ds \right]$$

From this point on this approximation shall be made.



## Dynamic Programming – Fully Co-ordinated Version

1. PDE Derivation and Numerical Solution: It is simplest to derive the PDE for the coordinated variation case directly, since that the only one that is solved here numerically. The extensions that are necessary to handle the general case are treated soon after.
2. Dynamics of Liquidity and Volatility: Since  $\eta(t)$  and  $\sigma(t)$  are positive it is convenient to write

$$\eta(t) = \bar{\eta} e^{\xi(t)}$$

and

$$\sigma(t) = \bar{\sigma} e^{-\frac{1}{2}\xi(t)}$$

where  $\bar{\eta}$  and  $\bar{\sigma}$  are typically values for  $\eta$  and  $\sigma$ , and  $\xi(t)$  is a single non-dimensional variable indicating the *market state*. When  $\xi(t)$  is large positive, the market is non-volatile and illiquid, and trading should be done more slowly; when  $\xi(t)$  is large negative, the market is volatile and liquid, and trading should be done fast.

3. Dynamics of the Intrinsic Time Scale: The intrinsic time scale in the mean market is written as

$$\bar{\kappa} = \sqrt{\frac{\lambda \bar{\sigma}^2}{\bar{\eta}}}$$

and

$$\kappa(t) = \sqrt{\frac{\lambda \sigma^2(t)}{\eta(t)}} = \bar{\kappa} e^{-\xi(t)}$$



for the instantaneous value. The value function  $c(t, x, \xi)$  then depends only on three variables.

4. Coordinated Market State Evolution:  $\xi(t)$  is assumed to solve and SDE of the form

$$\Delta\xi(t) = a(\xi(t))\Delta t + b(\xi(t))\Delta B_L(t)$$

where  $B_L(t)$  is a Brownian motion independent of the one driving the price motion.

5. The Corresponding HJB Variational Increment: Then by standard dynamic programming applied to

$$c(t, x, \eta, \sigma) = \min_{v(s), t \leq s \leq T} \mathbb{E} \left[ \int_t^T \{\lambda\sigma^2(s)x^2(s) + \eta(s)v^2(s)\} ds \right]$$

with instantaneous trade rate

$$v = -\frac{dx}{dt}$$

as the control parameter, one writes

$$c(t, x, \xi) = \min_v \{\lambda\sigma^2x^2\Delta t + \eta v^2\Delta t + \mathbb{E}[c(t + \Delta t, x + \Delta x, \xi + \Delta\xi)]\}$$

giving the HJB PDE

$$0 = \frac{\partial c(t, x, \xi)}{\partial t} + \lambda\sigma^2x^2 + \min_v \left[ \eta v^2 - v \frac{\partial c}{\partial x} \right] + a \frac{\partial c}{\partial \xi} + \frac{1}{2} b^2 \frac{\partial^2 c}{\partial \xi^2}$$

6. Value Function Optimal HJB PDE: The minimum is clearly



$$v = \frac{1}{2\eta} \frac{\partial c}{\partial x}$$

and the PDE for  $c(t, x, \xi)$  becomes

$$-\frac{\partial c(t, x, \xi)}{\partial t} = \lambda\sigma^2 x^2 - \frac{1}{4\eta} \left( \frac{\partial c}{\partial x} \right)^2 + a \frac{\partial c}{\partial \xi} + \frac{1}{2} b^2 \frac{\partial^2 c}{\partial \xi^2}$$

7. Near Expiration Asymptotic Cost Behavior: The initial data for the HJB PDE above is in fact a local asymptotic condition and must be treated with some care. Near expiration, liquidation must happen on a linear trajectory

$$v = \frac{x}{T-t}$$

As with constant coefficients

$$\mathbb{C} \sim \frac{\eta x^2}{T-t} + \frac{\lambda\sigma^2 x^2}{3} (T-t) + \mathcal{O}((T-t)^3)$$

the cost comes primarily from market impact.

8. Approximating the Terminal Cost Estimate: To accurately approximate the cost one must account for the expected changes in the impact coefficient during the time

$$t \leq s \leq T$$

A simple application of the Ito's lemma shows that

$$\mathbb{E}[\eta(s)] = \eta(t) \left[ 1 + \left( a + \frac{1}{2} b^2 \right) (s-t) \right]$$

$$s-t \rightarrow 0$$



and hence the average value of  $\eta(s)$  for  $s$  between  $t$  and  $T$  is

$$\eta \sim \eta(t) \left[ 1 + \frac{1}{2} \left( a + \frac{1}{2} b^2 \right) (T - t) \right]$$

$$T - t \rightarrow 0$$

9. The Terminal Cost Estimate Approximation: Using this value in

$$\mathbb{C} \sim \frac{\eta x^2}{T - t} + \frac{\lambda \sigma^2 x^2}{3} (T - t) + \mathcal{O}((T - t)^3)$$

$$\kappa(T - t) \rightarrow 0$$

with

$$\eta(t) = \bar{\eta} e^{\xi(t)}$$

one gets the cost expansion

$$\mathbb{C} \sim \frac{\bar{\eta} e^{\xi(t)} x^2}{T - t} + \frac{1}{2} \left( a + \frac{1}{2} b^2 \right) \bar{\eta} e^{\xi(t)} x^2 + \mathcal{O}(T - t)$$

$$\kappa(T - t) \rightarrow 0$$

In the  $\mathcal{O}(T - t)$  term there would appear both the risk contribution and a further expansion of the impact cost.

## Log-Normal Model and Non-dimensionalization



1. Market State Explicit Drift/Wander: The next step is to assume that  $\xi(t)$  evolves according to an Ornstein-Uhlenbeck mean-reverting process of zero mean. Thus

$$a(\xi(t)) = -\frac{\xi(t)}{\delta}$$

and

$$b(\xi(t)) = \frac{\beta}{\sqrt{\delta}}$$

2. Coordinated Market State Behavior: Here  $\delta$  is a market relaxation time, and  $\beta$  is a *burstiness* parameter describing the dispersion of liquidity and volatility around their average levels. In the steady state  $\xi(t)$  is normal with unconditional moments

$$\mathbb{E}[\xi(t)] = 0$$

and

$$\mathbb{V}[\xi(t)] = \frac{1}{2}\beta^2$$

3. Cost Function Dependence on  $x$ : Clearly the value function is strictly proportional to  $x^2$  – the square of the number of shares remaining to execute. This is a consequence of the linear market impact model, since both the variance and the expected cost are quadratic in quantity.
4. Non-dimensionalization of the Cost Function: The non-dimensionalization may now be done using  $\delta$  as the time scale and incorporating the factor  $x^2$ . One this defines



$$\tau = \frac{T-t}{\delta}$$

the time remaining to expiration as a multiple of the market relaxation time, and sets

$$c(t, x, \xi) = \frac{\bar{\eta}x^2}{\delta} u\left(\frac{T-t}{\delta}, \xi\right)$$

where  $u(\tau, \xi)$  is a non-dimensional function of non-dimensional variables.

##### 5. Non-dimensionalization of the HJB PDE: Then

$$-\frac{\partial c(t, x, \xi)}{\partial t} = \lambda \sigma^2 x^2 - \frac{1}{4\eta} \left( \frac{\partial c}{\partial x} \right)^2 + a \frac{\partial c}{\partial \xi} + \frac{1}{2} b^2 \frac{\partial^2 c}{\partial \xi^2}$$

becomes

$$\frac{\partial u(\tau, \xi)}{\partial \tau} + \xi \frac{\partial u(\tau, \xi)}{\partial \xi} = e^{-\xi} [K^2 - u^2(\tau, \xi)] + \frac{1}{2} \beta^2 \frac{\partial^2 u(\tau, \xi)}{\partial \xi^2}$$

in which the non-dimensional risk aversion parameter  $K$  is given as

$$K = \bar{\kappa}\delta = \frac{\text{Market Relaxation Time}}{\text{Trade Time Scale in Mean Market State}}$$

##### 6. The Corresponding Dimensional Trade Velocity: From

$$v = \frac{1}{2\eta} \frac{\partial c}{\partial x}$$

the dimensional trade velocity becomes



$$v = \frac{x}{\delta} e^{-\xi} u(\tau, \xi)$$

7. Terminal Asymptote Behavior of  $u(\tau, \xi)$ : Substituting the above expression into

$$\mathbb{C} \sim \frac{\bar{\eta} e^{\xi(t)} x^2}{T-t} + \frac{1}{2} \left( a + \frac{1}{2} b^2 \right) \bar{\eta} e^{\xi(t)} x^2 + \mathcal{O}(T-t)$$

$$\kappa(T-t) \rightarrow 0$$

the initial condition is determined as

$$u(\tau, \xi) = \frac{e^\xi}{\tau} - \frac{1}{2} \left( \xi - \frac{1}{2} \beta^2 \right) e^\xi + \mathcal{O}(\tau)$$

as

$$\tau \rightarrow 0$$

for each fixed  $\xi$ .

8. Terminal  $u(\tau, \xi)$  Asymptote at  $\xi < 0$ : For

$$\xi < 0$$

when trading is fast, the region of approximate validity of this trading is limited by the rate of trading itself, and this expression should be replaced by

$$u(\tau, \xi) \sim K \coth(K \tau e^{-\xi})$$

$$\xi \rightarrow -\infty$$



for

$$\tau > \mathcal{O}(e^{-\xi})$$

9. Terminal  $u(\tau, \xi)$  Asymptote at  $\xi > 0$ : For

$$\xi > 0$$

when trading is slow, the region of validity is

$$\tau \ll \mathcal{O}(1)$$

since the market itself changes on times of scale  $\mathcal{O}(1)$ . Almgren (2012) illustrates using pictorial summary the various asymptotic behavior of solutions to

$$\frac{\partial u(\tau, \xi)}{\partial \tau} + \xi \frac{\partial u(\tau, \xi)}{\partial \xi} = e^{-\xi} [K^2 - u^2(\tau, \xi)] + \frac{1}{2} \beta^2 \frac{\partial^2 u(\tau, \xi)}{\partial \xi^2}$$

## Constant Market

1. Non-volatile Steady-State Market: The steady-state market takes

$$\beta = 0$$

Along the line

$$\xi = 0$$

the PDE



$$\frac{\partial u(\tau, \xi)}{\partial \tau} + \xi \frac{\partial u(\tau, \xi)}{\partial \xi} = e^{-\xi} [K^2 - u^2(\tau, \xi)] + \frac{1}{2} \beta^2 \frac{\partial^2 u(\tau, \xi)}{\partial \xi^2}$$

reduces to the ODE

$$\frac{\partial u(\tau, \xi)}{\partial \tau} = K^2 - u^2(\tau, \xi)$$

with

$$u(\tau) \sim \frac{1}{\tau} + \mathcal{O}(\tau)$$

as

$$\tau \rightarrow 0$$

whose solution is

$$u(\tau) \sim K \coth(K\tau)$$

On undoing the change of variables, this reduces exactly to

$$c(t, x, \eta, \sigma) = \eta \kappa x^2 \coth[\kappa(T-t)] = \eta v(t)x$$

2. The  $\xi \rightarrow -\infty$  Case: Fast Trading: To generalize the above relation one considers the limit

$$\xi \rightarrow -\infty$$



That is, the market impact is very small, and the volatility is very large, thus the optimal strategy is trade very quickly.

3. Fast Relative to Market Relaxation: Since the market relaxation time scales are fixed, fast trading means that the program is completed before the market parameters have had time to change. Thus, the cost is the static cost

$$c(t, x, \eta, \sigma) = \eta \kappa x^2 \coth[\kappa(T - t)] = \eta v(t)x$$

using the instantaneous market parameters, which, in the transformed functions becomes

$$u(\tau, \xi) \sim K \coth(K\tau e^{-\xi})$$

$$\xi \rightarrow -\infty$$

4. Relaxation to Rolling Horizon Strategy: The corresponding trade rate is the *rolling horizon* strategy, which is always an admissible, though sub-optimal, strategy. The expression for  $u(\tau, \xi)$  accurately describes the optimal cost only in the indicated limit when indeed the market coefficients do not change substantially before trading is completed.

## Long Time

1. Asymptotic  $u(\tau, \xi)$  far from Expiration: As noted before, with coordinated variation, when time is far from expiry, the value of the function is

$$C = x^2 \sqrt{\lambda \bar{\sigma}^2 \bar{\eta}}$$

or



$$u(\tau, \xi) \rightarrow \infty$$

as

$$\tau \rightarrow \infty$$

in non-dimensional terms. Certainly

$$u = K$$

is a steady state solution of the PDE.

2. Validity of the Terminal Asymptote: And since the value function must be decreasing in  $\tau$  it is clear that

$$u(\tau, \xi) \geq K$$

for all

$$\tau \geq 0$$

As a consequence, the initial expression

$$u(\tau, \xi) = \frac{e^\xi}{\tau} - \frac{1}{2} \left( \xi - \frac{1}{2} \beta^2 \right) e^\xi + \mathcal{O}(\tau)$$

as

$$\tau \rightarrow 0$$



for each fixed  $\xi$  can only be valid when

$$\frac{e^\xi}{\tau} \geq K$$

or

$$\tau \leq \mathcal{O}(e^\xi)$$

- a very thin region when  $\xi$  is negative.

3. Uniqueness of the Solution to the PDE: Provided that a unique solution  $u(\tau, \xi)$  to

$$\frac{\partial u(\tau, \xi)}{\partial \tau} + \xi \frac{\partial u(\tau, \xi)}{\partial \xi} = e^{-\xi} [K^2 - u^2(\tau, \xi)] + \frac{1}{2} \beta^2 \frac{\partial^2 u(\tau, \xi)}{\partial \xi^2}$$

exists, a standard verification argument establishes that this function does indeed give the optimal control to the original control problem.

4. Uniqueness of the Solution to the PDE: Since

$$\frac{\partial u(\tau, \xi)}{\partial \tau} + \xi \frac{\partial u(\tau, \xi)}{\partial \xi} = e^{-\xi} [K^2 - u^2(\tau, \xi)] + \frac{1}{2} \beta^2 \frac{\partial^2 u(\tau, \xi)}{\partial \xi^2}$$

is a non-degenerate diffusion equation with lower-order terms, it certainly has smooth unique solutions locally in time if the solution at some positive time satisfies

$$u(\tau, \xi) < C e^{\alpha \xi^2}$$

5. Decomposition of the Initial Term: To understand the initial behavior near the initial term, writing



$$u(\tau, \xi) = \frac{e^\xi}{\tau} [1 + w(\tau, \xi)]$$

$w(\tau, \xi)$  for

$$\tau > 0$$

satisfies the PDE

$$\begin{aligned} \frac{\partial w(\tau, \xi)}{\partial \tau} + \frac{w(\tau, \xi)}{\tau} [1 + w(\tau, \xi)] \\ = K^2 \tau e^{-2\xi} - \left( \xi - \frac{1}{2} \beta^2 \right) - \left( \xi - \frac{1}{2} \beta^2 \right) w(\tau, \xi) - (\xi - \beta^2) \frac{\partial w(\tau, \xi)}{\partial \tau} \\ + \frac{1}{2} \beta^2 \frac{\partial^2 w(\tau, \xi)}{\partial \xi^2} \end{aligned}$$

as

$$\underset{\tau \rightarrow 0}{\text{Limit}} w(\tau, \xi) = 0$$

for each  $\xi$ .

6. Uniqueness of the Decomposed Solution: It is in this sense that  $u(\tau, \xi)$  satisfies its PDE

$$\frac{\partial u(\tau, \xi)}{\partial \tau} + \xi \frac{\partial u(\tau, \xi)}{\partial \xi} = e^{-\xi} [K^2 - u^2(\tau, \xi)] + \frac{1}{2} \beta^2 \frac{\partial^2 u(\tau, \xi)}{\partial \xi^2}$$

and the singular boundary condition. Although Almgren (2012) does not formally present a proof for the existence and the uniqueness of the function  $w(\tau, \xi)$  and hence of  $u(\tau, \xi)$  there do not appear to be any obstacles.



7. Perturbation of the Decomposed Solution: A search for the perturbation expansion of the form

$$w(\tau, \xi) \sim \tau w_1(\xi) + \tau^2 w_2(\xi) + \dots$$

$$\tau \rightarrow 0$$

readily determines

$$w_1(\xi) = -\frac{1}{2} \left( \xi - \frac{1}{2} \beta^2 \right)$$

$$w_2(\xi) = \frac{1}{3} K^2 e^{-2\xi} + \frac{1}{12} \left[ \xi^2 + (2 - \beta^2)\xi + \frac{1}{4} \beta^4 - 2\beta^2 \right]$$

8. Local Behavior Description for  $u(\tau, \xi)$ : The construction of this asymptotic behavior is strong evidence that the solution exists and has the associated local behavior. Thus a description of the local behavior of  $u(\tau, \xi)$  slightly fuller than

$$u(\tau, \xi) = \frac{e^\xi}{\tau} - \frac{1}{2} \left( \xi - \frac{1}{2} \beta^2 \right) e^\xi + \mathcal{O}(\tau)$$

as

$$\tau \rightarrow 0$$

for each fixed  $\xi$  is

$$\begin{aligned} u(\tau, \xi) &\sim \frac{e^\xi}{\tau} - \frac{1}{2} \left( \xi - \frac{1}{2} \beta^2 \right) e^\xi \\ &+ \tau \left\{ \frac{1}{3} K^2 e^{-2\xi} + \frac{1}{12} \left[ \xi^2 + (2 - \beta^2)\xi + \frac{1}{4} \beta^4 - 2\beta^2 \right] \right\} + \mathcal{O}(\tau^2) \end{aligned}$$



## Dynamic Programming – Custom $\eta(t)$ and $\sigma(t)$

1. Non-dimensional Liquidity and Volatility: Since  $\eta(t)$  and  $\sigma(t)$  are positive it is convenient to use

$$\xi(t) = \log \frac{\eta(t)}{\bar{\eta}}$$

and

$$\zeta(t) = \log \frac{\sigma(t)}{\bar{\sigma}}$$

as state variables.

2. Mean Market State Time Scale: Here  $\bar{\eta}$  and  $\bar{\sigma}$  are typical values of  $\eta(t)$  and  $\sigma(t)$ , and  $\xi(t)$  and  $\zeta(t)$  are non-dimensional values that fluctuate around zero.  $\bar{\kappa}$  is written as

$$\bar{\kappa} = \sqrt{\frac{\lambda \bar{\sigma}^2}{\bar{\eta}}}$$

for the intrinsic time scale in the mean market state.

3. Evolution of  $\xi(t)$  and  $\zeta(t)$ :  $\xi(t)$  and  $\zeta(t)$  are taken to evolve according to the stochastic differential equations (SDE) of the forms

$$\Delta \xi = a_\xi \Delta t + b_\xi \Delta \beta_L$$

and



$$\Delta\zeta = a_\zeta \Delta t + b_\zeta \Delta \beta_V$$

where  $a_\xi$ ,  $b_\xi$ ,  $a_\zeta$ , and  $b_\zeta$  are coefficients whose values may depend on  $\xi(t)$  and  $\zeta(t)$ .

4. Correlated Liquidity/Volatility Brownian Processes:  $\beta_L(t)$  and  $\beta_V(t)$  are Brownian motions that are independent of the process  $B(t)$  driving the price process, but possibly correlated with each other, with

$$\mathbb{E}[\Delta\beta_L \Delta\beta_V] = \rho \Delta t$$

5. Applying the Dynamic Programming Criterion: Then, by dynamic programming, it follows that

$$\mathbb{C}(t, x, \xi, \zeta) = \min_v \{ \lambda \sigma^2 x^2 \Delta t + \eta v^2 \Delta t + \mathbb{C}(t + \Delta t, x + \Delta x, \xi + \Delta \xi, \zeta + \Delta \zeta) \}$$

giving the PDE

$$\begin{aligned} 0 &= \frac{\partial \mathbb{C}(t, x, \xi, \zeta)}{\partial t} + \lambda \sigma^2 x^2 + \min_v \left[ \eta v^2 - v \frac{\partial \mathbb{C}(t, x, \xi, \zeta)}{\partial x} \right] + a_\xi \frac{\partial \mathbb{C}(t, x, \xi, \zeta)}{\partial \xi} \\ &\quad + a_\zeta \frac{\partial \mathbb{C}(t, x, \xi, \zeta)}{\partial \zeta} + \frac{1}{2} b_\xi^2 \frac{\partial^2 \mathbb{C}(t, x, \xi, \zeta)}{\partial \xi^2} + \frac{1}{2} b_\zeta^2 \frac{\partial^2 \mathbb{C}(t, x, \xi, \zeta)}{\partial \zeta^2} \\ &\quad + \frac{1}{2} b_\xi b_\zeta \frac{\partial^2 \mathbb{C}(t, x, \xi, \zeta)}{\partial \xi \partial \zeta} \end{aligned}$$

6. Optimality in the Trade Rate Space: The minimum is clearly

$$v = \frac{1}{2\eta} \frac{\partial \mathbb{C}(t, x, \xi, \zeta)}{\partial x}$$

and the PDE for  $\mathbb{C}(t, x, \xi, \zeta)$  is



$$\begin{aligned}
& - \frac{\partial \mathbb{C}(t, x, \xi, \zeta)}{\partial t} \\
& = \lambda \sigma^2 x^2 - \frac{1}{4\eta} \left[ \frac{\partial \mathbb{C}(t, x, \xi, \zeta)}{\partial x} \right]^2 + a_\xi \frac{\partial \mathbb{C}(t, x, \xi, \zeta)}{\partial \xi} + a_\zeta \frac{\partial \mathbb{C}(t, x, \xi, \zeta)}{\partial \zeta} \\
& + \frac{1}{2} b_\xi^2 \frac{\partial^2 \mathbb{C}(t, x, \xi, \zeta)}{\partial \xi^2} + \frac{1}{2} b_\zeta^2 \frac{\partial^2 \mathbb{C}(t, x, \xi, \zeta)}{\partial \zeta^2} + \rho b_\xi b_\zeta \frac{\partial^2 \mathbb{C}(t, x, \xi, \zeta)}{\partial \xi \partial \zeta}
\end{aligned}$$

7. Exogenous Expiration Trade Rate Asymptote: Near expiration the liquidation must happen on a linear trajectory

$$v = \frac{x}{T - t}$$

The cost comes entirely from market impact in the market conditions at that time, since volatility risk is negligible across a short time. Thus

$$\mathbb{C}(t, x, \xi, \zeta) \sim \frac{\eta x^2}{T - t}$$

$$\kappa(T - t) \rightarrow 0$$

applies, and

$$\mathbb{C}(t, x, \xi, \zeta) \sim \frac{\eta x^2}{T - t} = \frac{\bar{\eta} e^\xi x^2}{T - t}$$

$$(T - t) \rightarrow 0$$

8. De-dimensionalization of the PDE State Variables: To non-dimensionalize the cost function and the differential equation, a time scale needs to be defined, which also helps define a cost scale; the market parameters are already non-dimensionalized by their mean values.



9. Time Scale Choice - Liquidity Reversion: So far, the only two-time scales are the intrinsic liquidation time in the mean market state  $\frac{1}{\kappa}$  and the imposed horizon  $T$ . Since both of these depend on a trader's preferences for a particular trade order, it will be more natural to use a time scale based on market dynamics.

## Log-Normal Model

1. Ornstein-Uhlenbeck Mean-Reverting Dynamics: Here the assumption is that  $\xi(t)$  and  $\zeta(t)$  evolve according to the mean-reverting process of zero mean.
2.  $\xi(t)$  and  $\zeta(t)$  Drift/Wander: Accordingly

$$a_\xi = -\frac{\xi}{\delta_L}$$

$$a_\zeta = -\frac{\zeta}{\delta_V}$$

$$b_\xi = \frac{\beta_L}{\sqrt{\delta_L}}$$

$$b_\zeta = \frac{\beta_V}{\sqrt{\delta_V}}$$

3. Liquidity/Volatility Relaxation Time Scales: Here  $\delta_L$  and  $\delta_V$  are relaxation time scales for liquidity and volatility, and  $\beta_L$  and  $\beta_V$  are non-dimensional “burstiness” parameters.
4.  $\xi(t)$  and  $\zeta(t)$  Steady State: In the steady state,  $\xi(t)$  and  $\zeta(t)$  are normal with

$$\mathbb{E}[\xi(t)] \rightarrow 0$$



$$\mathbb{V}[\xi(t)] \rightarrow \frac{1}{2} \beta_L^2$$

$$\mathbb{E}[\zeta(t)] \rightarrow 0$$

$$\mathbb{V}[\zeta(t)] \rightarrow \frac{1}{2} \beta_V^2$$

as

$$t \rightarrow \infty$$

Thus  $\beta_L$  and  $\beta_V$  describe the liquidity and the volatility around their mean levels.

5.  $\delta_L$  as the Reference Time Scale: One may non-dimensionalize using  $\delta_L$  as the time scale. That is, defining

$$\tau = \frac{T - t}{\delta_L}$$

and setting

$$\mathbb{C}(t, x, \xi, \zeta) = \frac{\bar{\eta}x^2}{T - t} = u\left(\frac{T - t}{\delta_L}, \xi, \zeta\right)$$

where  $u\left(\frac{T-t}{\delta_L}, \xi, \zeta\right)$  is a non-dimensional function of non-dimensional variables, one gets



$$\begin{aligned}
& \frac{\partial u(\tau, \xi, \zeta)}{\partial \tau} + \xi \frac{\partial u(\tau, \xi, \zeta)}{\partial \xi} + \mu \zeta \frac{\partial u(\tau, \xi, \zeta)}{\partial \zeta} \\
&= K^2 e^{2\zeta} - e^{-\xi} u^2(\tau, \xi, \zeta) + \frac{1}{2} \beta_L^2 \frac{\partial^2 u(\tau, \xi, \zeta)}{\partial \xi^2} + \frac{1}{2} \beta_V^2 \frac{\partial^2 u(\tau, \xi, \zeta)}{\partial \zeta^2} \\
&\quad + \rho \sqrt{\mu} \beta_L \beta_V \frac{\partial^2 u(\tau, \xi, \zeta)}{\partial \xi \partial \zeta}
\end{aligned}$$

6. Volatility Market State Time Scales: On writing

$$\mu = \frac{\delta_L}{\delta_V}$$

the multi-dimensional risk aversion parameter becomes

$$K = \bar{\kappa} \delta_L = \frac{\text{Relaxation Time for Market Liquidity}}{\text{Trade Time Scale in Mean Market State}}$$

Of course  $\tau$  is the time remaining to expiration measured in units of the market relaxation time.

7. Initial Condition Re-cast: The initial condition becomes

$$u(\tau, \xi, \zeta) = \frac{e^\xi}{\tau}$$

$$\tau \rightarrow 0$$

8. The Corresponding Dimensional Trade Rate: From

$$v = \frac{1}{2\eta} \frac{\partial \mathbb{C}(t, x, \xi, \zeta)}{\partial x}$$

the dimensional trade velocity is



$$v = \frac{xe^{-\xi}}{\delta_L} u(\tau, \xi, \zeta)$$

9. Deterministic Liquidity and Volatility Processes: The constant volatility market takes

$$\beta_L = \beta_V = 0$$

Along the line

$$\xi = \zeta = 0$$

the PDE

$$\begin{aligned} \frac{\partial u(\tau, \xi, \zeta)}{\partial \tau} + \xi \frac{\partial u(\tau, \xi, \zeta)}{\partial \xi} + \mu \zeta \frac{\partial u(\tau, \xi, \zeta)}{\partial \zeta} \\ = K^2 e^{2\zeta} - e^{-\xi} u^2(\tau, \xi, \zeta) + \frac{1}{2} \beta_L^2 \frac{\partial^2 u(\tau, \xi, \zeta)}{\partial \xi^2} + \frac{1}{2} \beta_V^2 \frac{\partial^2 u(\tau, \xi, \zeta)}{\partial \zeta^2} \\ + \rho \sqrt{\mu \beta_L \beta_V} \frac{\partial^2 u(\tau, \xi, \zeta)}{\partial \xi \partial \zeta} \end{aligned}$$

reduces to the ODE

$$\frac{\partial u(\tau, \xi, \zeta)}{\partial \tau} = K^2 - u^2(\tau, \xi, \zeta)$$

with

$$u(\tau) \rightarrow \frac{1}{\tau}$$

as



$$\tau \rightarrow 0$$

whose solution is

$$u(\tau) = K \coth K\tau$$

10. Constant Coefficient Cost Function Reduction: On undoing the change of variables, this reduces exactly to

$$\mathbb{C}(t, x, \eta, \sigma) = \eta \kappa x^2 \coth \kappa(T - t) = \eta v x$$

11. Low Market Impact/High Volatility Scenario: To generalize the above solution the limits

$$\xi \rightarrow -\infty$$

and

$$\zeta \rightarrow +\infty$$

are considered. That is, the market impact is temporarily very small and the volatility is very large; the optimal strategy would be to trade very quickly.

12. Consequence of the Fast Trading: Since the market relaxation times are fixed, fast trading means that the program is completed before the market parameters have had time to change.
13. The Corresponding Non-dimensional Cost: Thus, the cost is the static cost

$$\mathbb{C}(t, x, \eta, \sigma) = \eta \kappa x^2 \coth \kappa(T - t) = \eta v x$$



using instantaneous market parameters, which in the transformed functions becomes

$$u(\tau, \xi, \zeta) \sim K e^{\zeta + \frac{1}{2}\xi} \coth K\tau e^{\zeta - \frac{1}{2}\xi}$$

$$\xi \rightarrow -\infty$$

$$\zeta \rightarrow +\infty$$

14. Comparison with the “Rolling Horizon” Approximation: The corresponding trade rate is the “rolling horizon” strategy seen earlier, which is always an admissible, although sub-optimal, strategy. The expression

$$u(\tau, \xi, \zeta) \sim K e^{\zeta + \frac{1}{2}\xi} \coth K\tau e^{\zeta - \frac{1}{2}\xi}$$

$$\xi \rightarrow -\infty$$

$$\zeta \rightarrow +\infty$$

accurately describes the optimal cost only in the indicated limit, when indeed the market coefficients do not change substantially before the trading is completed.

## Coordinated Variation

1. Coordinated Variation Case Reduction: Rather than solve the full PDE



$$\begin{aligned}
& \frac{\partial u(\tau, \xi, \zeta)}{\partial \tau} + \xi \frac{\partial u(\tau, \xi, \zeta)}{\partial \xi} + \mu \zeta \frac{\partial u(\tau, \xi, \zeta)}{\partial \zeta} \\
&= K^2 e^{2\zeta} - e^{-\xi} u^2(\tau, \xi, \zeta) + \frac{1}{2} \beta_L^2 \frac{\partial^2 u(\tau, \xi, \zeta)}{\partial \xi^2} + \frac{1}{2} \beta_V^2 \frac{\partial^2 u(\tau, \xi, \zeta)}{\partial \zeta^2} \\
&\quad + \rho \sqrt{\mu} \beta_L \beta_V \frac{\partial^2 u(\tau, \xi, \zeta)}{\partial \xi \partial \zeta}
\end{aligned}$$

in two space dimensions and one time dimension, more insight can be attained by considering the coordinated variation model described above.

2. Equal Liquidity/Volatility Time Scales: Thus, the following assumptions are made on the stochastic processes. First the time scales of liquidity and volatility are assumed to be equal.

$$\delta_L = \delta_V = \delta$$

so

$$\mu = 1$$

3. Fully Correlated Liquidity/Volatility Brownian: The Brownian motions driving the liquidity and the volatility have perfect positive correlation

$$\rho = 1$$

4. Liquidity/Volatility Wander Intensity Ratio: For now the fluctuation magnitudes  $\beta_L$  and  $\beta_V$  are arbitrary. The setting is

$$y = -\frac{\beta_L}{\beta_V} = \frac{\text{Signed Fractional Variation of } \sigma^2}{\text{Signed Fractional Variation of } \eta}$$

so that the coordinated variation case takes



$$\gamma = 1$$

The assumption here is that

$$\gamma > 0$$

5. Wander Intensity Scaled Volatility/Liquidity: Then, from

$$\Delta \xi = a_\xi \Delta t + b_\xi \Delta \beta_L$$

and

$$\Delta \zeta = a_\zeta \Delta t + b_\zeta \Delta \beta_V$$

$$a_\xi = -\frac{\xi}{\delta_L}$$

$$a_\zeta = -\frac{\zeta}{\delta_V}$$

$$b_\xi = \frac{\beta_L}{\sqrt{\delta_L}}$$

$$b_\zeta = \frac{\beta_V}{\sqrt{\delta_V}}$$

one gets

$$\Delta(\beta_V \xi - \beta_L \zeta) = -(\beta_V \xi - \beta_L \zeta) \frac{\Delta t}{\delta}$$



and hence, after at most an initial transient, the solutions satisfy

$$(\beta_V \xi - \beta_L \zeta) = 0$$

or

$$\zeta = -\frac{\gamma}{2} \xi$$

## 6. The Coordinated Variation Cost PDE: The PDE

$$\begin{aligned} \frac{\partial u(\tau, \xi, \zeta)}{\partial \tau} + \xi \frac{\partial u(\tau, \xi, \zeta)}{\partial \xi} + \mu \zeta \frac{\partial u(\tau, \xi, \zeta)}{\partial \zeta} \\ = K^2 e^{2\zeta} - e^{-\xi} u^2(\tau, \xi, \zeta) + \frac{1}{2} \beta_L^2 \frac{\partial^2 u(\tau, \xi, \zeta)}{\partial \xi^2} + \frac{1}{2} \beta_V^2 \frac{\partial^2 u(\tau, \xi, \zeta)}{\partial \zeta^2} \\ + \rho \sqrt{\mu} \beta_L \beta_V \frac{\partial^2 u(\tau, \xi, \zeta)}{\partial \xi \partial \zeta} \end{aligned}$$

maybe re-cast as

$$\begin{aligned} \frac{\partial u(\tau, \xi, \zeta)}{\partial \tau} + \left( \xi \frac{\partial}{\partial \xi} + \zeta \frac{\partial}{\partial \zeta} \right) u(\tau, \xi, \zeta) \\ = K^2 e^{2\zeta} - e^{-\xi} u^2(\tau, \xi, \zeta) + \frac{1}{2} \left( \beta_L \frac{\partial}{\partial \xi} + \beta_V \frac{\partial}{\partial \zeta} \right)^2 u(\tau, \xi, \zeta) \end{aligned}$$

Ignoring cross-variation (see the reasons below) the PDE for  $u(\tau, \xi, \zeta)$  is, with

$$\beta = \beta_L$$

$$\frac{\partial u(\tau, \xi, \zeta)}{\partial \tau} + \xi \frac{\partial u(\tau, \xi, \zeta)}{\partial \xi} = K^2 e^{-\gamma \xi} - e^{-\xi} u^2(\tau, \xi, \zeta) + \frac{1}{2} \beta^2 \frac{\partial^2 u(\tau, \xi, \zeta)}{\partial \xi^2}$$



7. Change of Variable - Initial Condition: The initial condition is still

$$u(\tau, \xi) = \frac{e^\xi}{\tau}$$

as

$$\tau \rightarrow 0$$

8. Change of Variable - Wander Differential: To trace the change of variables in detail, introduce  $w(\tau, \xi, \zeta)$  with

$$u(\tau, \xi, \zeta) = w(\tau, \xi, \beta_V \xi - \beta_L \zeta)$$

so that

$$\left( \xi \frac{\partial}{\partial \xi} + \zeta \frac{\partial}{\partial \zeta} \right) u(\tau, \xi, \zeta) = \left( \xi \frac{\partial}{\partial \xi} + \chi \frac{\partial}{\partial \chi} \right) w(\tau, \xi, \chi)$$

and

$$\left( \beta_L \frac{\partial}{\partial \xi} + \beta_V \frac{\partial}{\partial \zeta} \right)^2 u(\tau, \xi, \zeta) = \beta_L^2 \frac{\partial^2 w(\tau, \xi, \chi)}{\partial \xi^2}$$

9. Change of Variables Zero Wander: Thus

$$\frac{\partial w(\tau, \xi, \chi)}{\partial \tau} + \left( \xi \frac{\partial}{\partial \xi} + \chi \frac{\partial}{\partial \chi} \right) w(\tau, \xi, \chi) = K^2 e^{2\xi} - e^{-\xi} w^2(\tau, \xi, \chi) + \frac{1}{2} \beta_L^2 w(\tau, \xi, \chi)$$

and on the plane



$$\chi = 0$$

this reduces to

$$\frac{\partial u(\tau, \xi, \zeta)}{\partial \tau} + \xi \frac{\partial u(\tau, \xi, \zeta)}{\partial \xi} = K^2 e^{-y\xi} - e^{-\xi} u^2(\tau, \xi, \zeta) + \frac{1}{2} \beta^2 \frac{\partial^2 u(\tau, \xi, \zeta)}{\partial \xi^2}$$

## Asymptotic Behavior

1. Short Time Frame Asymptotic Behavior: To study the behavior as

$$\tau \rightarrow 0$$

for a fixed  $\xi$  one writes

$$u(\tau, \xi) = \frac{e^\xi}{\tau} + u_0(\xi) + \tau u_1(\xi) + \dots$$

$$\tau \rightarrow 0$$

2. Short Time Higher Order Dependence: One finds that at  $\mathcal{O}\left(\frac{1}{\tau}\right)$

$$u_0(\xi) = -\frac{1}{2} \left( \xi - \frac{1}{2} \beta^2 \right) e^\xi$$

and at  $\mathcal{O}(1)$

$$3u_1(\xi) = K^2 e^{-y\xi} - e^{-\xi} u_0^2(\xi) - \xi u_0'(\xi) + \frac{1}{2} \beta^2 u_0''(\xi)$$



or

$$u_1(\xi) = \frac{1}{3}K^2 e^{-y\xi} + \frac{1}{12}e^\xi \left[ \left( \xi + 1 - \frac{1}{2}\beta^2 \right)^2 - (1 + \beta^2) \right]$$

3. Implications for the Solution Robustness: Thus  $u(\tau, \xi)$  has a regular expansion on the powers of  $\tau$ . This reassures that the singular initial data is indeed enough to define the solution.
4. Long Time Asymptotic PDE Dependence: As

$$\tau \rightarrow \infty$$

presumably there is a steady state cost and a strategy in which the horizon is not controlling. The steady state solution  $u(\xi)$  will be determined by the second order nonlinear ODE

$$\xi u' = K^2 e^{-y\xi} u^2 + \frac{1}{2}\beta^2 u''$$

5. Explicit Long Time Asymptote PDE: In the coordinated variation case

$$y = 1$$

this has a constant solution

$$u(\xi) = K$$

For

$$y \neq 1$$



an explicit solution cannot be given, but for

$$\gamma > -1$$

the asymptotic behavior

$$u(\xi) \sim K e^{-\frac{1}{2}(\gamma-1)\xi}$$

as

$$\xi \rightarrow -\infty$$

can be identified, based on the balance

$$0 = K^2 e^{-\gamma \xi} - e^{-\xi} u^2$$

## 6. Low Impact High Volatility Simplification: As

$$\xi \rightarrow -\infty$$

with

$$\gamma > 0$$

one also has

$$\zeta \rightarrow +\infty$$

and thus the asymptotic solution



$$u(\tau, \xi, \zeta) \sim K e^{\zeta + \frac{1}{2}\xi} \coth\left(K \tau e^{\zeta - \frac{1}{2}\xi}\right)$$

$$\xi \rightarrow -\infty$$

$$\zeta \rightarrow +\infty$$

is valid, and simplifies to

$$u(\tau, \xi) \sim K e^{-\frac{1}{2}(y-1)\xi} \coth\left(K \tau e^{-\frac{1}{2}(y+1)\xi}\right)$$

$$\xi \rightarrow -\infty$$

7. Consistency with Long/Short Time Asymptote: At leading order this is consistent with both the short-term and the long-term behavior above.

## Numerical Solution

1. Numerical Solution to the HJB: Since explicit analytical solutions cannot be given, Almgren (2009) resorts to numerical solutions to generate solution solutions to

$$\frac{\partial u(\tau, \xi, \zeta)}{\partial \tau} + \zeta \frac{\partial u(\tau, \xi, \zeta)}{\partial \xi} = K^2 e^{-y\xi} - e^{-\xi} u^2(\tau, \xi, \zeta) + \frac{1}{2} \beta^2 \frac{\partial^2 u(\tau, \xi, \zeta)}{\partial \xi^2}$$

for a range of the given parameters.

2. Choice for the Fluctuation Ratio: The coordination parameter  $y$  should be chosen as part of the market structure. Since there is no particular reason to choose other values



$$\gamma = 1$$

shall be considered.

3. Solution to the Generalized HJB: Similarly, Almgren (2009) does not illustrate numerical solutions to the two-variable problem

$$\begin{aligned} \frac{\partial u(\tau, \xi, \zeta)}{\partial \tau} + \xi \frac{\partial u(\tau, \xi, \zeta)}{\partial \xi} + \mu \zeta \frac{\partial u(\tau, \xi, \zeta)}{\partial \zeta} \\ = K^2 e^{2\zeta} - e^{-\xi} u^2(\tau, \xi, \zeta) + \frac{1}{2} \beta_L^2 \frac{\partial^2 u(\tau, \xi, \zeta)}{\partial \xi^2} + \frac{1}{2} \beta_V^2 \frac{\partial^2 u(\tau, \xi, \zeta)}{\partial \zeta^2} \\ + \rho \sqrt{\mu} \beta_L \beta_V \frac{\partial^2 u(\tau, \xi, \zeta)}{\partial \xi \partial \zeta} \end{aligned}$$

since a truly complete study would also consider a broader range of market dynamics models.

4. Choice of the Burstiness Parameters: The burstiness parameter  $\beta$  is stock-specific. A large cap stock will have a  $\beta$  near zero, for a near-uniform profile. A small cap stock will have

$$\beta = 1$$

or larger. For the sample calculations below  $\beta$  is fixed at 2 – a relatively large value to better illustrate the effects of market variation.

5. Choice of the Risk Aversion: The risk aversion parameter  $K$  must range across non-negative values, since the actual choice of the trajectory will be determined by the trader's risk preference. Values of  $K$  smaller than 1 are the most realistic, so that the algorithm has time to adapt to at least one market reversion time.
6. Technical Issues behind the Solution: Almgren (2009) briefly discusses a few issues with space and time discretization, and presents example solutions, which is covered below.



## Time Discretization

1. Rationale for the Modified Euler Scheme: The first obstacle is that the initial condition is given as singular behavior. A simple modification to the Euler's forward scheme handles this problem. This is illustrated below using an ordinary differential equation (ODE).
2. The ODE and its Solution: Consider

$$\frac{du(t)}{dt} = -(u - a)(u - b)$$

$$u(t) \sim \frac{1}{t}$$

as

$$t \rightarrow 0$$

whose exact solution is

$$u(t) = \frac{ae^{-bt} - be^{-at}}{e^{-bt} - e^{-at}}$$

3. Local Expansion at the Origin: Either by expanding the solution, or directly from the ODE, the local expansion is determined as

$$u(t) \sim \frac{1}{t} + \frac{1}{2}(a + b) + \frac{t}{12}(a - b)^2 + \dots$$

as



$$t \rightarrow 0$$

4. Euler Scheme on Modified ODE: For the numerics a forward Euler scheme is applied to

$$w(t) = tu(t)$$

which is regular near

$$t = 0$$

With

$$w'(t) = tu'(t) + u(t)$$

and denoting

$$u_n \approx u(t_n)$$

this gives

$$u_{n+1} = u_n + \frac{t_n}{t_{n+1}}(t_{n+1} - t_n)u_n'$$

5. Correction to the Euler Update: Thus, a correction is applied to the Euler update formula, which becomes small as one moves away from the initial singular time and

$$\frac{t_n}{t_{n+1}} \rightarrow 1$$



6. Evolution on a Time Grid: The test is done on a regular grid with

$$t_n = n\Delta t$$

starting at

$$n = k \geq 1$$

$k$  is chosen to satisfy the stability criterion for the forward Euler scheme.

7. The First Time Node: For an ODE

$$\frac{du}{dt} = f(u)$$

the stability criterion requires that

$$\Delta t < \frac{1}{\left| \frac{df(u)}{du} \right|}$$

In this case

$$\frac{df(u)}{du} \sim 2u \sim \frac{2}{t}$$

so the stability criterion translates to

$$\Delta t < \frac{t}{2}$$

or



$$t > 2\Delta t$$

Thus one can expect the scheme to be stable for

$$k \geq 2$$

8. Discretization Modification to the Scheme: Four cases, resulting from all combinations of the two parameters, are explored. First:
  - a. The forward Euler discretization scheme can be applied directly to  $u$
  - b. The forward Euler discretization scheme can be applied directly to  $tu$  as seen above.
9. First Time Step Data Update: Second, for the data at the first-time step:
  - a.

$$u_k = \frac{1}{t_k}$$

using the given initial conditions

b.

$$u_k = \frac{1}{t_k} + \frac{1}{2}(a + b)$$

using the local expansion

$$u(t) \sim \frac{1}{t} + \frac{1}{2}(a + b) + \frac{t}{12}(a - b)^2 + \dots$$

as

$$t \rightarrow 0$$



10. Improvements from Discretization/Data Modifications: Almgren (2009) carries a demonstration of the example solutions. The combination of improved initial data, with a time discretization that takes into account the initial singularity, yields far more accurate results than naïve discretization.

## Space Discretization

1. Diffusion and Convection Terms Discretization: Almgren (2009) uses a 3-point standard discretization scheme for the diffusion term, and upwind differencing for the convection term (see, for example, Le Veque (1992)). A forward Euler time discretization scheme with the correction seen earlier is used; thus, a small-time step is used for stability.
2. Initial Time Node Discretization Scheme: For initial data, the asymptotic expressions

$$u(\tau, \xi) = \frac{e^\xi}{\tau} + u_0(\xi) + \tau u_1(\xi) + \dots$$

$$\tau \rightarrow 0$$

$$u_0(\xi) = -\frac{1}{2} \left( \xi - \frac{1}{2} \beta^2 \right) e^\xi$$

and

$$3u_1(\xi) = K^2 e^{-y\xi} - e^{-\xi} u_0^2(\xi) - \xi u_0'(\xi) + \frac{1}{2} \beta^2 u_0''(\xi)$$

or



$$u_1(\xi) = \frac{1}{3}K^2 e^{-y\xi} + \frac{1}{12}e^\xi \left[ \left( \xi + 1 - \frac{1}{2}\beta^2 \right)^2 - (1 + \beta^2) \right]$$

are used at an initial time

$$t = k\Delta t$$

3. Discretizing the Initial Trade Date: It is more convenient to discretize

$$v(\tau, \xi) = e^{-\xi} u(\tau, \xi)$$

rather than  $u$  directly; from

$$v = \frac{x}{\delta_L} e^{-\xi} u(\tau, \xi, \zeta)$$

this is the instantaneous trade rate, except for the dimensional factor. The PDE for  $v$  is easily derived from

$$\frac{\partial u(\tau, \xi, \zeta)}{\partial \tau} + \xi \frac{\partial u(\tau, \xi, \zeta)}{\partial \xi} = K^2 e^{-y\xi} - e^{-\xi} u^2(\tau, \xi, \zeta) + \frac{1}{2} \beta^2 \frac{\partial^2 u(\tau, \xi, \zeta)}{\partial \xi^2}$$

4. Left Far Field Boundary Condition: A finite spatial domain

$$-\Xi \leq \xi \leq \Xi$$

is used. At the left boundary

$$\xi = -\Xi$$

the far field solution



$$u(\tau, \xi) \sim K e^{-\frac{1}{2}(\gamma-1)\xi} \coth \left( K \tau e^{-\frac{1}{2}(\gamma+1)\xi} \right)$$

$$\xi \rightarrow -\infty$$

is used.

5. Right Far Field Boundary Condition: At the right boundary

$$\xi = +\Xi$$

the “natural” boundary conditions

$$\frac{\partial^2 v(\tau, \xi, \zeta)}{\partial \xi^2} = 0$$

are used. Since the convective term is flowing outwards, the effect of the boundary conditions is confined to a narrow boundary layer.

## Almgren (2009, 2012) Sample Solutions

1. Runs for  $\gamma = 1$ ,  $\beta = 1$ , and  $K = 0.1$ : Almgren (2009, 2012) illustrate the computed solution of the PDE for

$$\gamma = 1$$

$$\beta = 1$$

and



$$K = 0.1$$

As noted above, the natural log of  $e^{-\xi} u$  - the dimensionless trade rate as a fraction of the shares remaining – is examined.

2. High Impact/Low Volatility Behavior: As expected, when  $\tau$  is small, the trade rate becomes large like  $\frac{1}{\tau}$ . When  $\xi$  is large positive the market impact is high and the volatility is low, so the optimal strategy trades very slowly except near expiration.
3. Low Impact High Volatility Behavior: When  $\xi$  is large negative, the market impact is low and the volatility is high, so the optimal strategy trades rapidly. As

$$\tau \rightarrow \infty$$

the solution approaches the steady state

$$\log(e^{-\xi} u) = \log K - \xi$$

4. Realization of the Market State: Almgren (2009, 2012) also show a realization of the market state process  $\xi(t)$  used for the trajectory simulations. With the “coordinated variation” approximation, the market moves back and forth between a high-activity regime with low impact and high volatility (small  $\xi$ ) and a low-activity regime with high impact and low volatility (large  $\xi$ ).
5. Multi-Market Cycle Simulation Span:

$$\beta = 1$$

has been assumed so that the root mean-square fluctuation of  $\xi(t)$  is  $\frac{1}{2}$ . The mean reversion time is

$$\delta = 1$$



so that with a time

$$T = 10$$

several market cycles are experienced.

6. Response Dynamics across Risk Aversion: The optimal trading trajectories are examined for several values of the risk aversion parameter  $K$  and are compared with the non-adaptive trajectories computed in the mean market state. From these examinations the dynamic nature of the response is very clear.
7. Trading Slow Down/Speed Up: For instance, in the simulation, Almgren (2009, 2012) shows that around

$$t = 1$$

the market state is poor, so all trajectories trade slowly and fall behind the static ones.

Around

$$t = 1.5$$

there is a brief burst of liquidity, and all the trajectories accelerate in response.

8. Impact of Urgency of the Trajectory: The trajectories with lower urgency have more shares remaining to trade, so they are able to react more than the high-urgency trajectories, which have completed a substantial fraction of the goal by that time. Thus the lowest urgency trajectory is able to adapt and is able to benefit from an eventual large and prolonged burst of liquidity.
9. “Rolling Horizon” vs. Fast Trading: For large risk aversion (fast trading), this approximate strategy is almost identical to the optimum.



10. “Rolling Horizon” vs. Slow Trading: For smaller risk-aversion (slow trading), the rolling horizon strategy almost rigidly follows a straight-line trajectory, while the true optimum is able to adapt to the varying market state even when its profile is linear.
11. Validity of the “Rolling Horizon” Approximation: In general, the rolling horizon strategy seems to be an adequate approximation when the risk aversion is relatively high.

## References

- Almgren, R. F., and N. Chriss (2000): Optimal Execution of Portfolio Transactions *Journal of Risk* **3** (2) 5-39.
- Almgren, R. F., and J. Lorenz (2007): Adaptive Arrival Price, in: *Algorithmic Trading III* (B. R. Bruce, editor) **Institutional Investor** 59-66.
- Almgren, R. F. (2009): [Optimal Trading in a Dynamic Market](#).
- Almgren, R. F. (2012): Optimal Trading with Stochastic Liquidity and Volatility *SIAM Journal of Financial Mathematics* **3** (1) 163-181.
- Bouchaud, J. P., J. D. Farmer, and F. Lillo (2009): How Market slowly Digest Changes in Supply and Demand, in: *Handbook of Financial Markets: Dynamics and Evolution* 57-160 **North Holland** San Diego.
- Engle, R., and R. Ferstenberg (2007): Execution Risk: It is the same as Investment Risk *Journal of Portfolio Management* **33** 34-44.
- Gatheral, J (2010): No Dynamic Arbitrage and Market Impact *Quantitative Finance* **10** 749-759.
- Gatheral, J., and R. C. A. Oomen (2010): Zero Intelligence Realized Variance Estimation *Finance and Stochastics* **14** 249-283.
- Geman, H., D. B. Madan, and M. Yor (2001): Time Changes for Levy Processes *Mathematical Finance* **11** (1) 79-96.
- Grinold, R. C., and R. N. Kahn (1995): *Active Portfolio Management* **Probus Publishing**.



- Jones, C. M., G. Kaul, and M. L. Lipson (1994): Transactions, Volume, and Volatility *Review of Financial Studies* **7** 631-651.
- Le Veque, R. J. (1992): *Numerical Methods for Conservation Laws 2<sup>nd</sup> Edition* Birkhauser Basel.
- Lorenz, J. (2008): *Optimal Trading Algorithms* Ph. D. ETH Zurich.
- Lorenz, J., and R. F. Almgren (2011): Mean-Variance Optimal Adaptive Execution *Applied Mathematical Finance* **18** 395-422.
- Perold, A. F. (1988): The Implementation Shortfall: Paper versus Reality *Journal of Portfolio Management* **14 (3)** 4-9.
- Schied, A., and T. Schoneborn (2009): Risk Aversion and Dynamics of Optimal Liquidation Strategies in Illiquid Markets *Finance and Stochastics* **13 (2)** 181-204.
- Schied, A., T. Schoneborn, and M. Tehranchi (2010): Optimal Basket Liquidation for CARA Investors is Deterministic *Applied Mathematical Finance* **17** 471-489.
- Tse, S. T., P. A. Forsyth, J. S. Kennedy, and H. Windcliff (2013): Comparison between Mean-Variance Optimal and Mean Quadratic Variation Optimal Trading Strategies *20 (5)* 415-449.
- Walia, N. (2006): *Optimal Trading: Dynamic Stock Liquidation Strategies* Senior Thesis Princeton University.



## Order Placement in Limit Order Markets

### Overview

1. Limit/Market Orders across Exchanges: To execute a trade, participants in electronic equity markets may choose to submit limit order or market orders across various exchanges where a stock is traded.
2. Parameters Influencing the Submission Decision: The decision is influenced by the characteristics of the order flow and queue sizes in each limit order book, as well as the structure of transaction fees and rebates across exchanges.
3. Framework for Order Placement Problem: Cont and Kukanov (2017) propose a quantitative framework for studying this *order placement* problem by formulating it as a convex optimization problem.
4. Interplay between the Placement Parameters: This formulation allows studying how the interplay between the state of the order books, the fee structure, the order flow properties, and the preferences of a trader determine the optimal placement decision.
5. Case of a Single Exchange: In the case of a single exchange, an explicit solution for the optimal order split between limit and market orders is derived.
6. Case of Multiple Exchanges: For the general problem of order placement across multiple exchanges, they propose a stochastic algorithm for computing the optimal policy and study the sensitivity of the solution to various parameters using a numerical implementation of the algorithms.

### Introduction

1. Decomposition of the Trading Process: In automated electronic financial markets, the trading process is divided into several stages, each taking place on a different time



horizon; portfolio allocation decisions are usually made on a monthly or a daily basis and translate into trades that are executed over minutes to several days.

2. Modeling the Optimal Trade Execution: Studies on optimal trade execution (Bertsimas and Lo (1998), Almgren and Chriss (2000)) have investigated how the execution cost of a large trade may be reduced by splitting it into multiple *orders* spread in time.
3. Order Scheduling Followed by Placement: Once this *order scheduling* decision is taken, one still needs to specify how each individual order should be *placed*, this order placement decision involves the choice of an *order type* – limit order, market order – order size and destination, when multiple trading venues are available.
4. Timeframe for Order Filling: Orders are filled over short time intervals of few seconds to several minutes and the mechanics of how the orders are filled in the limit order book are relevant for such order placement decisions.
5. Impact of the Order Decisions: Market participants need to make such decisions thousands of time each day, and their outcomes have a large impact on each participant's transaction cost as well as on aggregate market dynamics.
6. Process for Filling the Order: Early work on optimal trade execution (Bertsimas and Lo (1998), Almgren and Chriss (2000)) did not explicitly model the process by which each order is filled, but more recent formulations have tried to incorporate some elements in this direction.
7. First Approach - Use of Market Orders: In one stream of literature – see Obizhaeva and Wang (2006), Alfonsi, Fruth, and Schied (2010), Predoiu, Shaikhet, and Shreve (2011) – a trader is restricted to using market orders whose execution costs are given by an idealized order book shape function.
8. Second Approach - Random Order Filling: Another approach is to model the process through which a order is filled as a dynamic random process (Cont (2011), Cont and de Larrard (2013)) and thus formulate the optimal execution problem as a stochastic control problem; this formulation has been studied in various settings in limit orders (Bayraktar and Ludkovski (2012), Gueant and Lehalle (2013)) or limit and market



orders (Guilbaud and Pham (2012), Huitema (2014)) but its complexity makes it intractable unless restrictive assumptions are made on price and order book dynamics.

9. Focus of this Chapter: This chapter adopts a simpler, more tractable approach; assuming that the trade execution schedule has been specified, the focus is on the task of filling each order.
10. Decoupling Order Scheduling from Placement: Decoupling the scheduling problem from the order placement problem leads to a more tractable approach which is closer to the market practice and allows us to incorporate some realistic features which matter for the order placement decisions while preserving analytical tractability.
11. Order Placement and Routing Decisions: Individual order placement and routing decisions play an important role in modern financial markets.
12. Legally Mandated Best Execution Quality: Brokers are commonly obliged by law to deliver the best execution quality to their clients and empirical evidence confirms that a large percentage of market orders in the US and Europe is sent to trading venues providing lower execution costs or smaller delays (Boehmer and Jennings (2007), Foucault and Menkveld (2008)).
13. Use of Market vs. Limit Orders: Market orders gravitate towards exchanges with larger posted quote sizes and low fees, while limit orders are submitted to exchanges with high rebates and lower execution waiting times – see Maglaras, Moallemi, and Zheng (2011).
14. Investors Aggregate Order Routing Decisions: The studies demonstrate how investors' aggregate order routing decisions have a significant influence on the market dynamics, but a systematic study of the order routing problem from the investor's perspective is lacking.
15. Limit Order Reduced Form Model: A reduced-form model for routing an infinitesimal order to a single destination is used by Maglaras, Moallemi, and Zheng (2011), while Laruelle, Lehalle, and Pages (2010) and Ganchev, Nevmyvaka, Kearns, and Vaughan (2010) propose numerical algorithms to optimize order executions across multiple dark pools, where supply/demand is unobserved.



16. Order Type Placement across Exchanges: Cont and Kukanov (2017) are the first to provide a detailed treatment of investor's order placement decision in a multi-exchange market, unified with market/limit order choice.
17. Quantitative Formulation of Order Placement: The key contribution of this chapter is the quantitative formulation of the order placement problem which takes into account multiple important factors – the size of an order to be executed, lengths of order queues across exchanges, statistical properties of order flows in these exchanges, trader's execution preferences, and the structure of liquidity rebates across trading venues.
18. Tractable Optimal Allocation Solution: The problem formulation is tractable and intuitive, and blends the aforementioned factors into an optimal allocation of limit and market orders across available trading venues.
19. Routing Heuristics from Past Behavior: Order routing heuristics employed in practice depend on past order fill rates at each exchange and are inherently backward-looking.
20. Order Routing Forward-looking Treatment: In contrast, the approach here is forward-looking – the optimal order allocation depends on the current queue sizes and distributions of future trading volumes across exchanges.
21. Case of a Single Exchange: When only a single exchange is available for execution, the order placement problem is reduced to the problem of choosing an optimal split between market and limit orders.
22. Explicit Solution for Single Exchange: An explicit solution to this problem is derived and its sensitivity to the order size, the trader's urgency for filling the order, and other factors is analyzed.
23. Case of Two Trading Venues: Similar results are also established in the case of two trading venues under some approximations on order flow distributions.
24. Case of Multiple Exchanges: Finally, a stochastic approximation method is proposed for solving the order placement problem in the general case and its efficiency is demonstrated through examples.
25. Numerical Solution using Stochastic Approximation: The numerical examples demonstrate that the use of optimal order placement method allows for substantial



trading cost decreases in comparison with the various ‘naïve’ order placement strategies.

26. Modeling the Execution Risk: An important aspect of the framework is to account for *execution risk*, through the incorporation of a penalty for under- or over-falling an order.
27. Penalizing Time-sensitive Executions: This penalty is high for time-sensitive executions or when it is costly up on the unfilled portion of the order.
28. When Market Orders are Preferred: Although market orders are executed at a less favorable price, it becomes optimal to use them when execution risk is a primary concern.
29. Determining Optimal Limit-Order Sizes: Optimal limit order sizes are strongly influenced by total quantities of orders queuing for execution at each exchange and by the distribution of order outflows from these queues.
30. When Limit Orders are Preferred: For example, if at one of the exchanges the queue size is much smaller than the expected future order outflow, it is optimal to place a large limit order there.
31. Impact of Total Order Size: Finally, the total order size plays an important role – limit orders are used predominantly to execute small order sizes and market orders are used for medium and large sizes.
32. Limitations in Filling Limit Orders: The amount that can be realistically filled with a limit order at each exchange is naturally constrained by the corresponding queue size and order outflow distribution, so the share of market orders in the optimal allocation increases as the total order size increases.
33. Limit Order Split across Exchanges: It is found that the optimal order allocation always splits the total quantity among all available exchanges, suggesting that there is a benefit in having multiple markets.
34. Formulation of Order Placement Problem: The next section describes the formulation of the order placement and shows that it has a global optimum.
35. Optimal Market/Limit Order Split: The section following that derives an optimal split between market and limit orders for a single exchange.



36. Order Placement on Multiple Venues: The penultimate section analyzes the general case of order placement on multiple trading venues.
37. Numerical Algorithm for Order Placement: The final section presents a numerical algorithm for solving the order placement problem in a general case and the simulation results, and follows with the conclusion.
38. Proofs: All proofs are presented right where the proposition is stated.

## The Order Placement Problem

1. Mandate to Buy  $S$  Shares: Consider a trader who has a mandate to buy  $S$  shares of a stock within a short time interval  $[0, T]$ . The deadline  $T$  may be a fixed horizon, e.g., 1 minute, or a stopping time triggered by a market activity.
2. Submission of  $K$  Limit Orders: To gain queue priority the trader may immediately submit  $K$  limit orders of sizes  $L_k$  to various exchanges

$$k = 1, \dots, K$$

or submit one market order of size  $M$ .

3. Optimal Order Placement State Vector: The trader's *order placement* decision is thus summarized by a vector

$$X \triangleq (M, L_1, \dots, L_k) \in \mathbb{R}_+^{K+1}$$

whose components are non-negative, i.e., only buy orders are allowed.

4. Goal - Formulation of Comprehensive Framework: The objective is to define a meaningful framework in which the trader may choose the various settings for this decision.



5. Filling Market Order with Certainty: The focus is on limit order placement and execution and the assumption is that a market order of size  $M$  can be filled immediately and with certainty.
6.  $S$  Small relative to Depth: This assumption is reasonable if  $S$  is small relative to the prevailing market depth. Under the assumption of immediate and certain market order execution, it is easy to show that sending market orders to exchanges with high fees is always sub-optimal.
7. Single Exchange for Market Order: One therefore considers a single exchange – with the smallest liquidity fee – for the purpose of sending a single market order.
8. Limit Orders on Pre-existing Queues: Limit orders with quantities  $(L_1, \dots, L_k)$  join queues of pre-existing limit orders of sizes  $(Q_1, \dots, Q_k)$  at the best bids of  $K$  exchanges, where

$$Q_k \geq 0$$

9. Bid Queues at Best Prices: To simplify the notation, one makes an assumption that all available  $K$  bid queues are lined up at the best price, but this is easily relaxed.
10. Modeling the Queue Outflow Process: Denote by

$$x_+ \triangleq \max(x, 0)$$

If  $L_k$  is constant within  $[0, T]$ , the amount purchased by a limit order on exchange  $k$  by time  $T$  is equal to  $(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k)_+$  where

$$\xi_k \triangleq C_k + D_k$$

is an order outflow from the front of the  $k^{th}$  bid queue consisting of

$$C_k \in [0, Q_k]$$



cancelations of pre-existing orders from that queue and  $D_k$  trades with contra-side marketable orders reaching the queue.

11. Random Limit Order Fill Process: It is specifically noted that the limit order fill amounts are random, and partial fills are allowed.
12. Total Amount Purchased by the Placement: The *total amount*  $A(X, \xi)$  bought by the trader by time  $T$  with all of his orders is a function of his order allocation  $X$  and an overall bid queue outflow

$$\xi = (\xi_1, \dots, \xi_K)$$

$$A(X, \xi) = M + \sum_{k=1}^K [(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k)_+]$$

13. Expression for the Transaction Execution Cost: The total price of this purchase is divided into a benchmark cost paid regardless of the trader's decisions, computed using a mid-quote price level, and an execution cost given by

$$(s + f)M - \sum_{k=1}^K (s + r_k)[(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k)_+]$$

where  $s$  is half of the bid-ask spread at time 0,  $f$  is the lowest available liquidity fee, and  $r_k$

$$k = 1, \dots, K$$

are liquidity rebates for all exchanges.

14. Limit Orders Lower the Cost: The trader can reduce the execution cost by sending more limit orders, but this leads to a risk of falling behind the target quantity  $S$  because the fills are random.



15. Penalization of the Execution Cost: To capture this *execution risk* one includes, in the objective function, a penalty for violations of target quantity in both directions

$$\lambda_U [S - A(X, \xi)]_+ + \lambda_O [A(X, \xi) - S]_+$$

where  $\lambda_U, \lambda_O$  are marginal penalties for falling behind or exceeding the execution target  $S$ , respectively.

16. Incorporating Adverse Selection Price Movements: These penalties are motivated by a correlation that exists between limit order executions and price movements, the so-called adverse selection.

17. Consequence of Under-filling the Order: If

$$A(X, \xi) < S$$

the trader has to purchase the remaining  $S - A(X, \xi)$  shares at time  $T$  with market orders.

18. Transaction Cost Impact of Under-filling: Adverse selection implies that conditionally on the event

$$\{A(X, \xi) < S\}$$

prices likely have moved up and the transaction cost of market orders at time  $T$  is higher than at time 0, i.e.,

$$\lambda_U > s + f$$

19. Incorporation of the Buyer's Remorse: Alternatively, if

$$A(X, \xi) < S$$



the trader experiences buyer's remorse, i.e., conditional on this event, prices have moved down and the trader could have achieved better execution by being more patient.

20. Tuning Parameters to Capture “Alpha”: Besides adverse selection, the parameters  $\lambda_U$ ,  $\lambda_O$  may reflect the trader's execution preferences. For example, a trader with a positive forecast of short-term returns may prefer to trade early with a market order and set a larger value for  $\lambda_U$ .
21. Optimal Order Placement Problem - Statement: An optimal order placement is a vector

$$X^* \in \mathbb{R}_+^{K+1}$$

solution to

$$\min_{X \in \mathbb{R}_+^{K+1}}$$

where

$$\begin{aligned} v(X, \xi) = & (s + f)M - \sum_{k=1}^K (s + r_k)[(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k)_+] \\ & + \lambda_U [S - A(X, \xi)]_+ + \lambda_O [A(X, \xi) - S]_+ \end{aligned}$$

is the sum of the execution cost and penalty for the execution risk.

22. Range Restriction on the Parameters: Denoting

$$V(X) = \mathbb{E}[v(X, \xi)]$$

this section begins by placing certain economically reasonable restrictions on the parameter values.



## The Optimal Order Problem – Assumptions

1. Penalty for Under/Over Filling:

$$\lambda_U > 0$$

$$\lambda_O > 0$$

The trader is penalized for falling behind or exceeding the target quantity.

2. Suboptimal to exceed the Target:

$$\lambda_O > s + \max_k \{r_k\}$$

and

$$\lambda_O > -(s + f)$$

It is suboptimal to exceed the target quantity  $S$  regardless of fees and rebates.

3. Enforcing Limit Order Cost Reduction:

$$s + \min_k \{r_k\} > 0$$

Even if some  $r_k$  are negative, limit orders still reduce the execution cost.

4. Convex Nature of the Problem: Proposition 1 below shows that it is not optimal to submit market or limit orders that are *a priori* too large or too small, i.e., larger than the target size  $S$  or whose sum is less than  $S$ . Proposition 2 guarantees the existence of an optimal solution.



## The Optimal Order Problem – Proposition 1

1. Compact Convex Subset of State Vector: Consider  $\mathcal{C}$  – a compact convex subset of  $\mathbb{R}_+^{K+1}$  defined by

$$\mathcal{C} \triangleq \left\{ X \in \mathbb{R}_+^{K+1} \mid 0 \leq M \leq S, 0 \leq L_k \leq S - M, k = 1, \dots, K, M + \sum_{k=1}^K L_k \geq S \right\}$$

2. Infimum achieved under Compact Set: Under the assumptions

$$\lambda_U > 0$$

$$\lambda_O > 0$$

$$\lambda_O > s + \max_k \{r_k\}$$

$$\lambda_O > -(s + f)$$

for any

$$\tilde{X} \notin \mathcal{C}$$

$$\exists \tilde{X}' \in \mathcal{C}$$

has

$$V(\tilde{X}') \leq V(\tilde{X})$$



3. Strictness of the Inequality: Moreover, if

$$\min_k \mathbb{P}[\xi_k > Q_k + S] > 0$$

the inequality is strict:

$$V(\tilde{X}') < V(\tilde{X})$$

4. Proof Start - Case where  $\tilde{M} > S$ : First, for any allocation  $\tilde{X}$  that has

$$\tilde{M} > S$$

one automatically has

$$A(\tilde{X}) > S$$

and it can be shown that the random cost and penalty  $\tilde{X}$  is larger than those of

$$X_{naive} \triangleq (S, 0, \dots, 0) \in \mathcal{C}$$



$$v(\tilde{X}, \xi) - v(X_{naive}, \xi)$$

$$\begin{aligned}
&= (s + f)(\tilde{M} - S) - \sum_{k=1}^K (s + r_k)[(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k)_+] \\
&\quad + \lambda_o \left\{ \tilde{M} - S + \sum_{k=1}^K (s + r_k)[(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k)_+] \right\} \\
&= (\lambda_o + s + f)(\tilde{M} - S) \\
&\quad + \sum_{k=1}^K (\lambda_o - s - r_k)[(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k)_+] > 0
\end{aligned}$$

which holds for all random  $\xi$ . Therefore

$$V(\tilde{X}) > V(X_{naive})$$

5. Case where  $L_k > S - \tilde{M}$ : Similarly, for any allocation  $\tilde{X}$  with

$$\tilde{L}_k > S - \tilde{M}$$

define a different allocation  $\tilde{X}'$  by

$$\tilde{M}' = \tilde{M}$$

$$\tilde{L}'_j = \tilde{L}_j \quad \forall j \neq k$$

and

$$\tilde{L}'_k = S - \tilde{M}$$

6.  $V(X, \xi)$  Difference when  $\xi_k > Q_k + S - M$ : Then



$$v(\tilde{X}, \xi) - v(\tilde{X}', \xi) = 0$$

on the event

$$B = \{\omega \mid \xi_k(\omega) < Q_k + S - M\}$$

7.  $V(X, \xi)$  Difference in the Complementary Case: On its complementary event  $B_c$

$$\begin{aligned} v(\tilde{X}, \xi) - v(\tilde{X}', \xi) \\ = -(s + r_k)[(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k)_+] \\ + \lambda_O[(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k)_+] \end{aligned}$$

8. Directionality of the above Difference: Therefore

$$\begin{aligned} V(\tilde{X}) - V(\tilde{X}') &= \mathbb{E}[v(\tilde{X}, \xi) - v(\tilde{X}', \xi) \mid B] \mathbb{P}[B] + \mathbb{E}[v(\tilde{X}, \xi) - v(\tilde{X}', \xi) \mid B_c] \mathbb{P}[B_c] \\ &= 0 + \mathbb{E}[\{\lambda_O - (s + r_k)\} \{(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k)_+\}] \mathbb{P}[B_c] \geq 0 \end{aligned}$$

with a strict inequality when

$$\mathbb{P}[B_c] > 0$$

9. Truncation across Exchanges with  $L'_j > S - \tilde{M}$ : If

$$\tilde{X}' \in \mathcal{C}$$

one can continue truncating limit order sizes

$$L'_j > S - \tilde{M}$$



using the same argument. Each time the truncation does not increase the objective function and finally one obtains

$$\tilde{X}'' \in \mathcal{C}$$

such that

$$V(\tilde{X}) \geq V(\tilde{X}'')$$

10. Assigning the Under-filled Order Amounts: Next, if  $\tilde{X}$  is such that

$$\tilde{M} + \sum_{k=1}^K \tilde{L}_k < S$$

define

$$s = S - \tilde{M} - \sum_{k=1}^K \tilde{L}_k$$

take

$$\tilde{M}' = \tilde{M}$$

$$\tilde{L}'_k = \tilde{L}_k$$

$$k = 1, \dots, K-1$$

and



$$\tilde{L}'_K = \tilde{L}_K + s$$

11.  $v(\tilde{X}, \xi)$  when  $\xi_K < Q_K + \tilde{L}_K$ : Then, on the event

$$B = \{\omega \mid \xi_K(\omega) < Q_K + \tilde{L}_K\}$$

where one has

$$v(\tilde{X}', \xi) = v(\tilde{X}, \xi)$$

12. Proof End -  $v(\tilde{X}, \xi)$  when  $\xi_K > Q_K + \tilde{L}_K$ : However, on the complementary event  $B_c$

$$\begin{aligned} v(\tilde{X}, \xi) - v(\tilde{X}', \xi) \\ = (s + r_K) \left[ (\xi_K - Q_K - \tilde{L}_K)_+ - (\xi_K - Q_K - \tilde{L}_K - s)_+ \right] \\ + \lambda_U \left[ (\xi_K - Q_K - \tilde{L}_K)_+ - (\xi_K - Q_K - \tilde{L}_K - s)_+ \right] \end{aligned}$$

therefore

$$\begin{aligned} V(\tilde{X}) - V(\tilde{X}') &= \mathbb{E}[v(\tilde{X}, \xi) - v(\tilde{X}', \xi) \mid B] \mathbb{P}[B] + \mathbb{E}[v(\tilde{X}, \xi) - v(\tilde{X}', \xi) \mid B_c] \mathbb{P}[B_c] \\ &= 0 \\ &+ \mathbb{E} \left[ \{\lambda_U + (s + r_K)\} \left\{ (\xi_K - Q_K - \tilde{L}_K)_+ \right. \right. \\ &\quad \left. \left. - (\xi_K - Q_K - \tilde{L}_K - s)_+ \right\} \mid B_c \right] \mathbb{P}[B_c] \geq 0 \end{aligned}$$

with a strict inequality of

$$\mathbb{P}[B_c] > 0$$



13. Penalizer Terms as Soft Penalties: The penalty function  $\lambda_U[S - A(X, \xi)]_+ + \lambda_O[A(X, \xi) - S]_+$  implements a soft constraint for order sizes and effectively focuses the search for an optimal order allocation to the set  $\mathcal{C}$ .
14. Incorporation of Additional Meaningful Constraints: Specific economic or operational considerations could also motivate hard constraints, e.g.

$$M = 0$$

or

$$\sum_{k=1}^K L_k = S$$

Such constraints can be easily included in this framework, but absent the aforementioned considerations, they are not imposed here.

## The Optimal Order Problem – Proposition 2

1. Existence of a Global Minimum: Under assumptions

$$\lambda_U > 0$$

$$\lambda_O > 0$$

$$\lambda_O > s + \max_k \{r_k\}$$

$$\lambda_O > -(s + f)$$



$$s + \max_k \{r_k\} > 0$$

$V(X)$  is a convex function on  $\mathbb{R}_+^{K+1}$  bounded below, and has a global minimizer

$$X^* \in \mathcal{C}$$

2. Proof Start - Concave Limit Order Purchase Functions: First, note that  $(\xi_k - Q_k)_+$  –  $(\xi_k - Q_k - L_k)_+$  are concave functions of  $L_k$ . Therefore,  $A(X, \xi)$  is concave as a sum of concave functions.
3. Criteria Determining Cost Term Convexity: Similarly, the cost term in  $v(X, \xi)$  is a sum of convex functions, as long as

$$r_k \geq -s$$

$$k = 1, \dots, K$$

and is itself a convex function.

4. Convex Cost Function Terms: Further, since  $S - A(X, \xi)$  is a convex function of  $X$ , and the function

$$h(x) \triangleq [\lambda_U(x)]_+ - [\lambda_O(-x)]_+$$

is convex in  $x$  for positive  $\lambda_U$  and  $\lambda_O$ , the penalty term  $h(S - A(X, \xi))$  is also convex.

5. Case where  $V(X)$  has Lower Bound: If

$$\lambda_O > s + \max_k \{r_k\}$$

the function  $V(X)$  is also bounded from below since



$$v(X, \xi) \geq - \left( s + \max_k \{r_k\} \right) S$$

6. Local/Global Minimum on  $\mathcal{C}$ : Finally, since  $V(X)$  is convex, it is also continuous and reaches on local minimum  $V_{min}$  on the compact set  $\mathcal{C}$  at some point

$$X^* \in \mathcal{C}$$

By convexity,  $V_{min}$  is a global minimum of  $V(X)$  on  $\mathcal{C}$ .

7. Proof End - Global Minimum across  $\mathbb{R}_+^{K+1}$ : Moreover, since

$$\lambda_o > s + \max_k \{r_k\}$$

Proposition 1 guarantees that

$$V_{min} < V(\tilde{X})$$

for any

$$\tilde{X} \in \mathcal{C}$$

so  $V_{min}$  is also a global minimum of  $V(X)$  on  $\mathbb{R}_+^{K+1}$ .

## Choice of Order Type: Limit Orders vs Market Orders

1. Trade-off – Limit against Market: To highlight the tradeoff between limit and market order executions in the optimization setup, one first considers the case where the asset is traded on a single exchange, and the trader has to choose an optimal split between limit and market orders.



2. Elimination of the Exchange Subscript: Since

$$K = 1$$

the subscript 1 is suppressed throughout this section.

### **Proposition 3 – Single Exchange: Optimal Split between Limit and Market Orders**

1. Lower and Upper  $\lambda_U$  Bounds: Assume that  $\xi$  has a continuous distribution and that

$$\lambda_U > 0$$

$$\lambda_O > 0$$

$$\lambda_O > s + \max_k \{r_k\}$$

$$\lambda_O > -(s + f)$$

$$s + \max_k \{r_k\} > 0$$

Denote

$$\underline{\lambda}_U \triangleq \frac{2s + f + r}{F(Q + S)} - (s + r)$$

and



$$\overline{\lambda_U} \triangleq \frac{2s + f + r}{F(Q)} - (s + r)$$

2. Case where  $\lambda_U \leq \underline{\lambda_U}$ : If

$$\lambda_U \leq \underline{\lambda_U}$$

the optimal allocation is

$$(M^*, L^*) = (0, S)$$

3. Case where  $\lambda_U \geq \overline{\lambda_U}$ : If

$$\lambda_U \geq \overline{\lambda_U}$$

the optimal allocation is

$$(M^*, L^*) = (S, 0)$$

4.  $\lambda_U$  Values between the Bounds: If

$$\lambda_U \in (\underline{\lambda_U}, \overline{\lambda_U})$$

the optimal allocation is:

$$M^* = S - F^{-1} \left( \frac{2s + f + r}{\lambda_U + s + r} \right) + Q$$

$$L^* = F^{-1} \left( \frac{2s + f + r}{\lambda_U + s + r} \right) - Q$$



where  $F(\cdot)$  is a cumulative distribution function of the bid queue outflow  $\xi$ .

5. Proof Start – Existence of an Optimal Split: By Proposition 1, there exists an optimal split

$$(M^*, L^*) \in \mathcal{C}$$

between limit and market orders.

6. The Market/Limit Order Split: Moreover, for

$$K = 1$$

the set  $\mathcal{C}$  reduces to the line

$$M^* + L^* = S$$

so it is sufficient to find  $M^*$ .

7. Objective Function along the Optimum: Restricting

$$L = S - M$$

implies that

$$\{A(X, \xi) < S\} = \emptyset$$

$$\{A(X, \xi) < S, \xi > Q + L\} = \emptyset$$

and the objective function can be re-written as



$$V(M) = \mathbb{E} \left[ (s + f)M - (s + r)\{(\xi - Q)_+ - (\xi - Q - S + M)_+\} + \lambda_U \{S - M - (\overline{\xi - Q}_+ - \overline{\xi - Q - S + M}_+)\}_+ \right]$$

8. Differentiability of the Objective Function: For

$$M \in (0, S)$$

the expectation is bounded for all  $\xi$  and differentiable with respect to  $M$  for almost all  $\xi$ , so one can compute

$$V'(M) = \frac{dV(M)}{dM}$$

by interchanging the order of differentiation and integration – see e.g. Aliprantis and Burkinshaw (1998):

$$\begin{aligned} V'(M) &= \mathbb{E}[(s + f) + (s + r)\mathbb{I}_{\xi > Q + S - M} - \lambda_U \mathbb{I}_{\xi < Q + S - M}] \\ &= 2s + f + r - (s + r + \lambda_U)F(Q + S - M) \end{aligned}$$

9. Case where  $\lambda_U$  undershoots Lower Bound: Note that if

$$\lambda_U \leq \frac{2s + f + r}{F(Q + S)} - (s + r)$$

then

$$V'(M) \geq 0$$

for



$$M \in (0, S)$$

and therefore  $V$  is non-decreasing at these points.

10.  $M^*$  Resulting from Lower Breach: Checking that

$$V(S) - V(0) \geq (s + f - \lambda_U)S + (\lambda_U + s + r) = S[1 - F(Q + S)] \geq 0$$

one concludes that

$$M^* = 0$$

11. Case where  $\lambda_U$  Overshoots Upper Bound: Similarly, if

$$\lambda_U \geq \frac{2s + f + r}{F(Q)} - (s + r)$$

then

$$v(M) \leq 0$$

for all

$$M \in (0, S)$$

and  $V(M)$  is non-increasing at these points.

12.  $M^*$  Resulting from Upper Breach: Checking that

$$V(S) - V(0) \leq (s + f - \lambda_U)S + (\lambda_U + s + r) = S[1 - F(Q)] \leq 0$$

one concludes that



$$M^* = S$$

13. Case of  $\lambda_U$  between Bounds: Finally, if  $\lambda_U$  is between these two values

$$\exists c > 0$$

such that

$$V'(\epsilon) < 0$$

$$V'(S - \epsilon) > 0$$

and by continuity of  $V'$  there is a point where

$$V'(M^*) = 0$$

14. Proof End -  $M^*$  Resulting from No Breach: Thus  $M^*$  is optimal by convexity of  $V(M)$  and

$$M^* = S - F^{-1} \left( \frac{2s + f + r}{\lambda_U + s + r} \right) + Q$$

$$L^* = F^{-1} \left( \frac{2s + f + r}{\lambda_U + s + r} \right) - Q$$

solves equations where

$$v(M^*, \xi) = 0$$

$$L^* = S - M^*$$

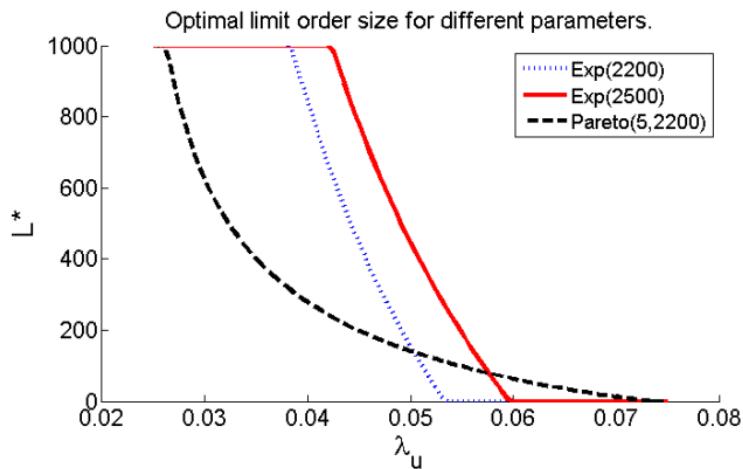


15. No Overfill with Single Exchange: In the case of a single exchange, Proposition 1 implies that

$$M^* + L^* = S$$

therefore, there is no risk of exceeding the target size and  $\lambda_O$  does not affect the optimal solution.

16. Single Exchange Risk of Underfill: The trader is only concerned with the risk of falling behind the target quantity, and balances this risk with fee, rebate, and other market information.
17.  $\lambda_U$  as Fill Urgency Tuner: The parameter  $\lambda_U$  can be interpreted as the trader's urgency to fill the orders, and higher values of  $\lambda_U$  lead to smaller order sizes, as illustrated in the figure below.



**Figure 2** Optimal limit order size  $L^*$  for one exchange. The parameters for this figure are:  $Q = 2000$ ,  $S = 1000$ ,  $h = 0.02$ ,  $r = 0.002$ ,  $f = 0.003$ ,  $\theta = 0.0005$ . Colors correspond to different order outflow distributions - exponential with means 2200 and 2500 and Pareto with mean 2200 and a tail index 5.

18. Optimal Market Size  $\lambda_U$  Dependency: In contrast, the optimal market order size increases with  $\lambda_U$ .



19. Ratio Determining Market/Limit Split: The optimal split between market and limit orders depends on the ratio  $\frac{2s+f+r}{\lambda_U+s+r}$  which balances marginal costs and savings from a market order.
20. Ratio Dependence on  $F/Q$ : It also depends on the distribution  $F$  and the queue length  $Q$  – keeping all else constant, a trader would submit a larger limit order if its execution is more likely and vice versa.
21. Optimal Limit Size  $\lambda_U$  Dependence: The optimal limit order size decreases with  $\lambda_U$  and it becomes more expensive to underfill the order and increases with  $f$  as market orders become more expensive.
22.  $L^*$  Independent of  $S$ : Another interesting feature is that  $L^*$  is full determined by  $Q$ ,  $F$ , and pricing parameters  $s$ ,  $r$ ,  $f$ , and  $\lambda_U$  while  $M^*$  increases with  $S$ .
23.  $M^*$  increases with  $S$ : As a consequence of this solution feature, as the order size  $S$  increases, a larger fraction  $\frac{M^*}{S}$  of that order is executed with a market order.
24. Dependence on Distribution of  $\xi$ : The solution  $(M^*, L^*)$  depends on the entire distribution of  $\xi$  and not just its mean, as illustrated on the above figure for a pair of exponential and Pareto distributions with equal means.

## Optimal Routing of Limit Orders across Multiple Exchanges

1. More Venues Greater Fill Likelihood: When multiple trading venues are available, dividing the target quantity among them provides better execution quality by reducing the risk of not filling the order.
2. More Venues, Size Breach Likelihood: However, sending too many orders leads to an undesirable possibility of exceeding target size.
3. Optimality Criterion for Multiple Exchanges: Proposition 4 gives a criterion for optimality of an order allocation

$$X^* = (M^*, L_1^*, \dots, L_K^*)$$



that balances these risks.

### **Proposition 4**

1. General Criteria for Solution: Assuming

$$\lambda_U > 0$$

$$\lambda_O > 0$$

$$\lambda_O > s + \max_k \{x_k\}$$

$$\lambda_O > -(s + f)$$

and

$$s + \max_k \{x_k\} > 0$$

also assume that the distribution of  $\xi$  is continuous:

$$\max_k \{F_k(Q_k + s)\} < 1$$

and

$$\lambda_U < \max_k \left\{ \frac{2s + f + r_k}{F_k(Q_k)} - (s + r_k) \right\}$$



Then:

2. Optimal Positive Market Order Quality: Any optimal allocation  $X^*$  has a positive market order quantity

$$M^* > 0$$

if

$$\lambda_U \geq \frac{2s + f + \max_k \{x_k\}}{\mathbb{P}[\bigcap_k \{\xi_k \leq Q_k\}]} - \left( s + \max_k \{x_k\} \right)$$

3. Optimal Positive Limit Order Quality: Any optimal allocation  $X^*$  has a positive limit order quantity

$$L_j^* > 0$$

if

$$\mathbb{P} \left[ \bigcap_k \{\xi_k \leq Q_k\} \mid \xi_j > Q_j \right] > \frac{\lambda_O - (s + r_j)}{\lambda_U + \lambda_O}$$

4. Necessary/Sufficient Conditions for Optimality: If the above criteria for

$$M^* > 0$$

and

$$L_j^* > 0$$

holds for all exchanges



$$j = 1, \dots, K$$

a necessary and sufficient condition for optimality of an order allocation

$$X^* \in \mathcal{C}$$

is that it solves the following equations:

$$\mathbb{P} \left[ M^* + \sum_{k=1}^K \{(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k^*)_+\} < S \right] = \frac{\lambda_o + f + s}{\lambda_u + \lambda_o}$$

$$\mathbb{P} \left[ M^* + \sum_{k=1}^K \{(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k^*)_+\} \mid \xi_j > Q_j + L_j \right] = \frac{\lambda_o - (s + r_j)}{\lambda_u + \lambda_o}$$

$$j = 1, \dots, K$$

5. Proof Start - Existence of Optimal Order Allocation: Proposition 2 implies the existence of an optimal order allocation

$$X^* \in \mathcal{C}$$

6. Non-optimality of Naïve Allocation: First, one defines

$$X_M \triangleq (S, 0, \dots, 0)$$

and proves that

$$X \neq X_M$$



by contradiction.

7. Implication of the Naïve Allocator: If  $X_M$  were optimal in the problem

$$\min_{X \in \mathbb{R}_+^{K+1}} \mathbb{E}[v(X, \xi)]$$

it would also be optimal in the same problem with a consistent

$$L_k = 0$$

$$k \neq j$$

for any one  $j$ .

8. Optimal Solution for Single Exchange: In other words, the solution  $(S, 0)$  would be optimal for any one-exchange problem, defined by using only exchange  $j$ .
9. Proof by Contradiction of Naïve Allocation Optimality: But by earlier assumption, there exists a  $J$  such that

$$\lambda_U < \frac{2s + f + r_J}{F_J(Q_J)} - (s + r_J)$$

and Proposition 3 implies that  $(S, 0)$  is not optimal for the  $J^{th}$  single-exchange subproblem, leading to a contradiction.

10. Bounded, Differentiable Nature of  $v(X, \xi)$ : The function  $v(X, \xi)$  is bounded for

$$X \in \mathcal{C}$$

and for all  $\xi$ , differentiable with respect to  $M$  and  $L_k$

$$k = 1, \dots, K$$



for

$$X \in \mathcal{C} \setminus \{X_M\}$$

for almost all  $\xi$ .

11. Differentiability of the Expected Cost: Applying the same theorem as in the proof of Proposition 3, one concludes that  $V(X)$  is differentiable for

$$X \in \mathcal{C} \setminus \{X_M\}$$

one can compute all of its partial derivatives by interchanging the order of differentiation and integration.

12. KKT Criteria for Expected Cost: The KKT conditions for problem

$$\min_{X \in \mathbb{R}_+^{K+1}} \mathbb{E}[v(X, \xi)]$$

and

$$X \in \mathcal{C} \setminus \{X_M\}$$

are

$$s + f - \lambda_U \mathbb{P}[A(X^*, \xi) < S] + \lambda_O \mathbb{P}[A(X^*, \xi) > S] - \mu_0 = 0$$

$$\begin{aligned} -(s + r_k) \mathbb{P}[\xi_k > Q_k + L_k^*] - \lambda_U \mathbb{P}[A(X^*, \xi) < S, \xi_k > Q_k + L_k^*] \\ + \lambda_O \mathbb{P}[A(X^*, \xi) > S, \xi_k > Q_k + L_k^*] - \mu_k = 0 \end{aligned}$$

$$k = 1, \dots, K$$



$$M \geq 0$$

$$L_k \geq 0$$

$$\mu_0 \geq 0$$

$$\mu_k \geq 0$$

$$\mu_0 M = 0$$

$$\mu_k L_k = 0$$

$$k = 1, \dots, K$$

13. Necessary and Sufficient Conditions for Optimality: Since the objective function  $V(\cdot)$  is convex, the KKT criteria above are both necessary and sufficient for optimality.

14. Optimal Positive Market Order Proof: The final result of this proposition follows from considering any  $\tilde{X}$  with

$$\tilde{M} = 0$$

$$\begin{aligned} V(\tilde{X}) &\geq \lambda_U S^{\mathbb{P}} \left[ \overline{\bigcap_k \{\xi_k \leq Q_k\}} \right] - \left( s + \max_k \{x_k\} \right) S^{\mathbb{P}} \left[ \overline{\bigcap_k \{\xi_k \leq Q_k\}} \right] \geq (s + f) S \\ &= V(X_M) \end{aligned}$$

and it has already been argued that  $\exists X^*$  with

$$V(X^*) < V(X_M)$$



so

$$X^* \neq \tilde{X}$$

and therefore

$$M^* > 0$$

15. Re casting the Exchange KKT Criterion: Re-arranging terms in a  $j^{th}$  equality in

$$\begin{aligned} -(s + r_k) \mathbb{P}[\xi_k > Q_k + L_k^*] - \lambda_U \mathbb{P}[A(X^*, \xi) < S, \xi_k > Q_k + L_k^*] \\ + \lambda_O \mathbb{P}[A(X^*, \xi) > S, \xi_k > Q_k + L_k^*] - \mu_k = 0 \end{aligned}$$

$$k = 1, \dots, K$$

one obtains

$$\begin{aligned} \mathbb{P}[\xi_k > Q_k + L_k^*] \{ \lambda_O - (s + r_j) + (\lambda_U + \lambda_O) \mathbb{P}[A(X^*, \xi) < S \mid \xi_k > Q_k + L_k^*] \} - \mu_j \\ = 0 \end{aligned}$$

16. Optimal Positive Limit Order Proof: The term in the curly brackets above is negative for any

$$X \in \mathcal{C} \setminus \{X_M\}$$

with

$$L_j = 0$$



because

$$\mathbb{P}[A(X^*, \xi) < S \mid \xi_j > Q_j + L_j^*] > \mathbb{P}\left[\bigcap_k \{\xi_k \leq Q_k \mid \xi_j > Q_j\}\right] > \frac{\lambda_o - (s + r_j)}{\lambda_u + \lambda_o}$$

by assumption and since

$$\mu_j \geq 0$$

the condition

$$-(s + r_k)\mathbb{P}[\xi_k > Q_k + L_k^*] - \lambda_u \mathbb{P}[A(X^*, \xi) < S \mid \xi_k > Q_k + L_k^*] \\ + \lambda_o \mathbb{P}[A(X^*, \xi) > S \mid \xi_k > Q_k + L_k^*] - \mu_k = 0$$

$$k = 1, \dots, K$$

cannot be satisfied with

$$L_j^* = 0$$

17. Proof End - Complementary Slackness implies KKT Multiplier  $\mu_j = 0$ : It was shown earlier that

$$M^* > 0$$

$$L_j^* > 0$$

for all

$$j = 1, \dots, K$$



and therefore

$$\mu_0 = \mu_1 = \cdots = \mu_k = 0$$

by complementary slackness. Then the KKT conditions

$$s + f - \lambda_U \mathbb{P}[A(X^*, \xi) < S] + \lambda_O \mathbb{P}[A(X^*, \xi) > S] - \mu_0 = 0$$

$$\begin{aligned} -(s + r_k) \mathbb{P}[\xi_k > Q_k + L_k^*] - \lambda_U \mathbb{P}[A(X^*, \xi) < S, \xi_k > Q_k + L_k^*] \\ + \lambda_O \mathbb{P}[A(X^*, \xi) > S, \xi_k > Q_k + L_k^*] - \mu_k = 0 \end{aligned}$$

$$k = 1, \dots, K$$

$$M \geq 0$$

$$L_k \geq 0$$

$$\mu_0 \geq 0$$

$$\mu_k \geq 0$$

$$\mu_0 M = 0$$

$$\mu_k L_k = 0$$

$$k = 1, \dots, K$$

reduce to



$$\mathbb{P} \left[ M^* + \sum_{k=1}^K \{(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k^*)_+\} < S \right] = \frac{\lambda_o + f + s}{\lambda_u + \lambda_o}$$

$$\mathbb{P} \left[ M^* + \sum_{k=1}^K \{(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k^*)_+\} \mid \xi_j > Q_j + L_j \right] = \frac{\lambda_o - (s + r_j)}{\lambda_u + \lambda_o}$$

$$j = 1, \dots, K$$

18. Target Breach Probability as Bounds: The equations

$$\mathbb{P} \left[ M^* + \sum_{k=1}^K \{(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k^*)_+\} < S \right] = \frac{\lambda_o + f + s}{\lambda_u + \lambda_o}$$

and

$$\mathbb{P} \left[ M^* + \sum_{k=1}^K \{(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k^*)_+\} \mid \xi_j > Q_j + L_j \right] = \frac{\lambda_o - (s + r_j)}{\lambda_u + \lambda_o}$$

$$j = 1, \dots, K$$

show that an order allocation is optimal as long as it sets the probabilities of falling behind the target quantity equal to specific thresholds computed with pricing parameters.

19. Challenges generating Closed-Form Solutions: When the number of exchanges  $K$  is large, the probabilities

$$\mathbb{P} \left[ M^* + \sum_{k=1}^K \{(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k^*)_+\} < S \right] = \frac{\lambda_o + f + s}{\lambda_u + \lambda_o}$$



and

$$\mathbb{P} \left[ M^* + \sum_{k=1}^K \{(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k^*)_+\} \mid \xi_j > Q_j + L_j \right] = \frac{\lambda_o - (s + r_j)}{\lambda_u + \lambda_o}$$

$$j = 1, \dots, K$$

are difficult to compute in closed-form.

20. Analyzing the Two-Exchange Case: However, before turning to numerical procedures, the next section investigates how these equations can be solved in a tractable case of two exchanges.

## Optimal Routing of Limit Orders Across Multiple Exchanges - Proposition 4 Corollary

1. Treating the Two-exchange Case: Consider the case of two exchanges with  $\xi_1, \xi_2$  that are independent and have continuous distributions. Assume the following conditions apply.
2. Condition #1 - Bounding  $F_k(Q_k + s)$ :

$$\max_{k=1,2} F_k(Q_k + s) < 1$$

3. Condition #2 - Lower/Upper Bounds on  $\lambda_u$ :

$$\lambda_u < \max_{k=1,2} \left\{ \frac{2s + f + r_k}{F_k(Q_k)} - (s + r_k) \right\}$$



$$\lambda_U \geq \frac{2s + f + \max_{k=1,2} \{r_k\}}{F_1(Q_1)F_2(Q_2)} - \left( s + \max_{k=1,2} \{r_k\} \right)$$

4. Condition #3 - Bounds on  $F_1$  and  $F_2$ :

$$F_1(Q_1) < 1 - \frac{s + r_2}{\lambda_o}$$

$$F_2(Q_2) < 1 - \frac{s + r_1}{\lambda_o}$$

5. Existence of Optimal Order Allocation: Then there exists an optimal order allocation

$$X^* = (M^*, L_1^*, L_2^*) \in \text{int}\{\mathcal{C}\}$$

and it verifies the following.

6. Explicit Expression for Optimal  $L_1$ :

$$L_1^* = Q_2 + S - M^* - F_2^{-1} \left( \frac{\lambda_o - [s + r_1]}{\lambda_U + \lambda_o} \right)$$

7. Explicit Expression for Optimal  $L_2$ :

$$L_2^* = Q_1 + S - M^* - F_1^{-1} \left( \frac{\lambda_o - [s + r_2]}{\lambda_U + \lambda_o} \right)$$

8. Mixed Integral for Solving  $M^*$ :

$$\begin{aligned} & \bar{F}_1(Q_1 + L_1^*) \bar{F}_2(Q_2 + S - M^* - L_1^*) \\ & + \int_{Q_1 + S - M^* - L_2^*}^{Q_1 + L_1^*} \bar{F}_2(Q_1 + Q_2 + S - M^* - x_1) dF_1(x_1) = \frac{\lambda_o - (s + f)}{\lambda_U + \lambda_o} \end{aligned}$$



where  $F_1(\cdot)$  and  $F_2(\cdot)$  are the CDF of  $\xi_1$  and  $\xi_2$  respectively.

9.  $L^*$  as Linear Function of  $M^*$ : In this solution the optimal limit order quantities  $L_1^*, L_2^*$  are linear functions of an optimal order quantity  $M^*$ .

10.  $M^*$  Non-linear on  $\xi$  Distribution: When

$$L_1^* = Q_2 + S - M^* - F_2^{-1} \left( \frac{\lambda_o - [s + r_1]}{\lambda_u + \lambda_o} \right)$$

and

$$L_2^* = Q_1 + S - M^* - F_1^{-1} \left( \frac{\lambda_o - [s + r_2]}{\lambda_u + \lambda_o} \right)$$

are substituted into

$$\begin{aligned} & \bar{F}_1(Q_1 + L_1^*) \bar{F}_2(Q_2 + S - M^* - L_1^*) \\ & + \int_{Q_1 + S - M^* - L_2^*}^{Q_1 + L_1^*} \bar{F}_2(Q_1 + Q_2 + S - M^* - x_1) dF_1(x_1) = \frac{\lambda_o - (s + f)}{\lambda_u + \lambda_o} \end{aligned}$$

one obtains a non-linear equation for  $M^*$ , which can be solved for a given distribution of  $(\xi_1, \xi_2)$ .

11. Proof Start – Non-optimal  $M^*/L^*$  at Boundaries: Solutions on the boundary of  $\mathcal{C}$  are sub-optimal:

$$M^* = 0$$

and

$$M^* = S$$



are ruled out by the assumptions

$$\lambda_U < \max_{k=1,2} \left\{ \frac{2s + f + r_k}{F_k(Q_k)} - (s + r_k) \right\}$$

and

$$\lambda_U \geq \frac{2s + f + \max_{k=1,2} \{r_k\}}{F_1(Q_1)F_2(Q_2)} - (s + \max_{k=1,2} \{r_k\})$$

$$L_1^* = S - M$$

and

$$L_2^* = S - M$$

are ruled out by the assumptions

$$F_1(Q_1) < 1 - \frac{s + r_2}{\lambda_o}$$

and

$$F_2(Q_2) < 1 - \frac{s + r_1}{\lambda_o}$$

and

$$\begin{aligned} -(s + r_k)\mathbb{P}[\xi_k > Q_k + L_k^*] - \lambda_U \mathbb{P}[A(X^*, \xi) < S, \xi_k > Q_k + L_k^*] \\ + \lambda_O \mathbb{P}[A(X^*, \xi) > S, \xi_k > Q_k + L_k^*] - \mu_k = 0 \end{aligned}$$



$$k = 1, \dots, K$$

12. Non optimal  $M/L$  at Boundary: Solutions with

$$M^* + \sum_{k=1}^K L_k^* = S$$

are ruled out by directly checking

$$\begin{aligned} & -(s + r_k) \mathbb{P}[\xi_k > Q_k + L_k^*] - \lambda_U \mathbb{P}[A(X^*, \xi) < S, \xi_k > Q_k + L_k^*] \\ & + \lambda_O \mathbb{P}[A(X^*, \xi) > S, \xi_k > Q_k + L_k^*] - \mu_k = 0 \end{aligned}$$

$$k = 1, \dots, K$$

13. Non optimality of  $L_1^* = 0$  and  $L_2^* = 0$ : Finally

$$L_1^* = 0$$

and

$$L_2^* = 0$$

are ruled out by

$$\begin{aligned} & -(s + r_k) \mathbb{P}[\xi_k > Q_k + L_k^*] - \lambda_U \mathbb{P}[A(X^*, \xi) < S, \xi_k > Q_k + L_k^*] \\ & + \lambda_O \mathbb{P}[A(X^*, \xi) > S, \xi_k > Q_k + L_k^*] - \mu_k = 0 \end{aligned}$$

$$k = 1, \dots, K$$



14. Proof of  $L^* = 0$  non-optimality: For example, if

$$L_1^* = 0$$

then by Proposition 1

$$M^* + L_2^* = S$$

and in

$$\begin{aligned} & -(s + r_k) \mathbb{P}[\xi_k > Q_k + L_k^*] - \lambda_U \mathbb{P}[A(X^*, \xi) < S, \xi_k > Q_k + L_k^*] \\ & + \lambda_O \mathbb{P}[A(X^*, \xi) > S, \xi_k > Q_k + L_k^*] - \mu_k = 0 \end{aligned}$$

$$k = 1, \dots, K$$

$$\mu_2 = 0$$

by complementary slackness

$$\mathbb{P}[A(X^*, \xi) < S, \xi_2 > Q_2 + L_2^*] = \mathbb{P}[A(X^*, \xi) > S, \xi_2 > Q_2 + L_2^*] = 0$$

But then

$$\begin{aligned} & -(s + r_k) \mathbb{P}[\xi_k > Q_k + L_k^*] - \lambda_U \mathbb{P}[A(X^*, \xi) < S, \xi_k > Q_k + L_k^*] \\ & + \lambda_O \mathbb{P}[A(X^*, \xi) > S, \xi_k > Q_k + L_k^*] - \mu_k = 0 \end{aligned}$$

$$k = 1, \dots, K$$

cannot hold, because



$$\mathbb{P}[\xi_2 > Q_2 + L_2^*] > 0$$

15. Necessary Criteria for Excess Allocation: For any

$$X \in \text{int } \{\mathcal{C}\}$$

$$A(X^*, \xi) > S$$

the following inequalities need to be satisfied:

$$\xi_1 > Q_1 + S - M - L_2$$

$$\xi_2 > Q_2 + S - M - L_1$$

$$\xi_1 + \xi_2 > Q_1 + Q_2 + S - M$$

16. Verifying Scenarios Resulting in Excess Allocation: These inequalities give a simple characterization of the event

$$\{A(X, \xi) > S\}$$

and their equivalence is directly verified by considering subsets  $\{\xi_1, \xi_2\}$  forming a complete partition of  $\mathbb{R}_+^2$ .

17. Case I - Both Limit Orders Completely Filled:

$$\xi_1 > Q_1 + L_1$$

$$\xi_2 > Q_2 + L_2$$

18. Trivial Fulfillment of Excess Criteria: Since



$$L_1 + L_2 + M > S$$

one has

$$A(X, \xi) = L_1 + L_2 + M > S$$

and at the same time, all of the inequalities

$$\xi_1 > Q_1 + S - M - L_2$$

$$\xi_2 > Q_2 + S - M - L_1$$

$$\xi_1 + \xi_2 > Q_1 + Q_2 + S - M$$

are satisfied, so they are trivially equivalent in this case.

19. Case II - Second Limit Order Partially Filled:

$$\xi_1 > Q_1 + L_1$$

$$Q_2 \leq \xi_2 \leq Q_2 + L_2$$

20. First Limit Order Completely Filled: Because of the condition

$$\xi_1 > Q_1 + L_1$$

$$\xi_1 > Q_1 + S - M - L_2$$

is satisfied.

21. Condition where Allocation exceeds Target: In this case, one has that



$$A(X, \xi) = L_1 - \xi_2 - Q_2 + M$$

and thus

$$A(X, \xi) > S$$

if and only if

$$\xi_2 > Q_2 + S - M - L_1$$

is satisfied.

22. Fulfillment of Excess Allocation Criteria: Finally

$$\xi_1 > Q_1 + L_1$$

together with

$$\xi_2 > Q_2 + S - M - L_1$$

imply

$$\xi_1 + \xi_2 > Q_1 + Q_2 + S - M$$

so

$$A(X, \xi) > S$$

and



$$\xi_1 > Q_1 + S - M - L_2$$

is also satisfied.

23. Case III - First Limit Order Partially Filled:

$$\xi_2 > Q_2 + L_2$$

$$Q_1 \leq \xi_1 \leq Q_1 + L_1$$

24. Treatment Mirrors the Previous Case: Similar to Case II, it can be shown that the inequalities

$$\xi_1 > Q_1 + S - M - L_2$$

$$\xi_2 > Q_2 + S - M - L_1$$

$$\xi_1 + \xi_2 > Q_1 + Q_2 + S - M$$

are satisfied if and only if

$$A(X, \xi) > S$$

25. Case IV - Both Limit Orders are Partially Filled:

$$Q_1 + S - M - L_2 < \xi_1 \leq Q_1$$

$$Q_2 + S - M - L_1 < \xi_2 \leq Q_2 + L_2$$

26. Valid Scenario in this Case: This set is non-empty because



$$0 < S - M - L_1 \leq L_2$$

and, similarly, for  $L_1, L_2$  reversed.

27. Criterion that needs Verification: The inequalities

$$\xi_1 > Q_1 + S - M - L_2$$

$$\xi_2 > Q_2 + S - M - L_1$$

hold trivially, only

$$\xi_1 + \xi_2 > Q_1 + Q_2 + S - M$$

needs to be checked.

28. Verification of the Third Criterion: One can write

$$A(X, \xi) = \xi_1 - Q_1 + \xi_2 - Q_2 + M > S$$

if and only if

$$\xi_1 + \xi_2 > Q_1 + Q_2 + S - M$$

holds, so

$$A(X, \xi) > S$$

is equivalent to all the necessary criteria.

29. Case V - The Residual Scenario: This captures the case outside of 1-4, and neither

$$\xi_1 > Q_1 + S - M - L_2$$



nor

$$\xi_2 > Q_2 + S - M - L_1$$

is satisfied.

30. First Criterion Violation + Partial Second Fill: If

$$\xi_1 \leq Q_1 + S - M - L_2$$

and

$$\xi_2 \leq Q_2 + L_2$$

then

$$A(X, \xi) \leq S - M - L_2 + L_2 + M = S$$

31. Converse to the Case above: The case

$$\xi_2 \leq Q_2 + S - M - L_1$$

$$\xi_1 \leq Q_1 + L_1$$

is completely symmetric, and it shows that neither

$$A(X, \xi) > S$$

nor the necessary excess allocation criteria hold in this case.



32. Relating Excess Allocation Criteria to Optimality: Next, the necessary excess allocation criteria are used to characterize the set

$$\{A(X, \xi) > S\}$$

in the first order conditions

$$\mathbb{P} \left[ M^* + \sum_{k=1}^K \{(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k^*)_+\} < S \right] = \frac{\lambda_o + f + s}{\lambda_u + \lambda_o}$$

and

$$\mathbb{P} \left[ M^* + \sum_{k=1}^K \{(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k^*)_+\} \mid \xi_j > Q_j + L_j \right] = \frac{\lambda_o - (s + r_j)}{\lambda_u + \lambda_o}$$

$$j = 1, \dots, K$$

33. Excess Allocation to Outflow Probabilities: One observes that in the two-exchange case

$$\{A(X, \xi) > S, \xi_1 > Q_1 + L_1\} = \{\xi_1 > Q_1 + L_1, \xi_2 > Q_2 + S - M - L_1\}$$

$$\{A(X, \xi) > S, \xi_2 > Q_2 + L_2\} = \{\xi_2 > Q_2 + L_2, \xi_1 > Q_1 + S - M - L_2\}$$

and then use the independence of  $\xi_1$  and  $\xi_2$  to compute

$$\mathbb{P}[A(X, \xi) > S \mid \xi_1 > Q_1 + L_1] = \bar{F}_2(Q_2 + S - M - L_1)$$

$$\mathbb{P}[A(X, \xi) > S \mid \xi_2 > Q_2 + L_2] = \bar{F}_1(Q_1 + S - M - L_2)$$



34. Solution to the Limit Order Allocations: Together with

$$\mathbb{P} \left[ M^* + \sum_{k=1}^K \{(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k^*)_+\} < S \right] = \frac{\lambda_o + f + s}{\lambda_u + \lambda_o}$$

and

$$\mathbb{P} \left[ M^* + \sum_{k=1}^K \{(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k^*)_+\} \mid \xi_j > Q_j + L_j \right] = \frac{\lambda_o - (s + r_j)}{\lambda_u + \lambda_o}$$

$$j = 1, \dots, K$$

this leads to a pair of equations for limit order sizes:

$$\bar{F}_2(Q_2 + S - M - L_1) = \frac{\lambda_u + s + r_1}{\lambda_u + \lambda_o}$$

and

$$\bar{F}_1(Q_1 + S - M - L_2) = \frac{\lambda_u + s + r_2}{\lambda_u + \lambda_o}$$

whose solution is given by  $L_1^*$ ,  $L_2^*$  using the inequalities of the necessary excess allocation criteria.

35. Non-linear Equation for  $L^*$ : To obtain the equation

$$\begin{aligned} & \bar{F}_1(Q_1 + L_1^*) \bar{F}_2(Q_2 + S - M^* - L_1^*) \\ & + \int_{Q_1 + S - M^* - L_2^*}^{Q_1 + L_1^*} \bar{F}_2(Q_1 + Q_2 + S - M^* - x_1) dF_1(x_1) = \frac{\lambda_o - (s + f)}{\lambda_u + \lambda_o} \end{aligned}$$



one re-writes

$$\mathbb{P} \left[ M^* + \sum_{k=1}^K \{(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k^*)_+\} < S \right] = \frac{\lambda_o + f + s}{\lambda_u + \lambda_o}$$

using the necessary excess allocation criteria.

36. Excess Allocation to Quadrature Space: Then

$$\mathbb{P}[A(X, \xi) > S]$$

may be computed as the integral of the product measure  $F_1 \otimes F_2$  over the region defined by

$$U(Q, S, M, L_1, L_2) = \left\{ \begin{array}{l} (x_1, x_2) \in \mathbb{R}^2 \\ x_1 > Q_1 + S - M - L_2 \\ x_2 > Q_2 + S - M - L_1 \\ x_1 + x_2 > Q_1 + Q_2 + S - M \end{array} \right\}$$

37. Explicit Form for the Quadrature: This integral is given by

$$\begin{aligned} \mathbb{P}[A(X, \xi) > S] &= F_1 \otimes F_2(U(Q, S, M, L_1, L_2)) \\ &= \bar{F}_1(Q_1 + L_1) \bar{F}_2(Q_2 + S - M - L_1) \\ &\quad + \int_{Q_1 + S - M - L_2}^{Q_1 + L_1} \bar{F}_2(Q_1 + Q_2 + S - M - x_1) dF_1(x_1) = \frac{\lambda_u - (s + f)}{\lambda_u + \lambda_o} \end{aligned}$$

## Optimal Routing of Limit Orders Across Multiple Exchanges – Example



1. Explicit Form for Exponential Distribution: If  $\xi_1, \xi_2$  are exponentially distributed with means  $\mu_1, \mu_2$  respectively, then an optimal order allocation is given by

$$M^* = Q_1 + Q_2 + S - z$$

$$L_1^* = z - Q_1 + \mu_2 \log \frac{\lambda_U + s + r_1}{\lambda_U + \lambda_O}$$

$$L_2^* = z - Q_2 + \mu_1 \log \frac{\lambda_U + s + r_2}{\lambda_U + \lambda_O}$$

where  $z$  is a solution to the transcendental equation

$$1 + \log \left( \frac{\lambda_U + s + r_2}{\lambda_U + \lambda_O} \cdot \frac{\lambda_U + s + r_1}{\lambda_U + \lambda_O} \right) + \frac{z}{\mu} = \frac{\lambda_U - (s + f)}{\lambda_U + \lambda_O} e^{\frac{z}{\mu}}$$

if

$$\mu_1 = \mu_2 = \mu$$

or

$$\begin{aligned} & \frac{\mu_1}{\mu_1 - \mu_2} e^{-\frac{z}{\mu_1}} \left( \frac{\lambda_U + s + r_1}{\lambda_U + \lambda_O} \right)^{\frac{\mu_1 - \mu_2}{\mu_1}} + \frac{\mu_2}{\mu_1 - \mu_2} e^{-\frac{z}{\mu_2}} \left( \frac{\lambda_U + s + r_2}{\lambda_U + \lambda_O} \right)^{\frac{\mu_2 - \mu_1}{\mu_2}} \\ &= \frac{\lambda_U - (s + f)}{\lambda_U + \lambda_O} \end{aligned}$$

if

$$\mu_1 \neq \mu_2$$



2.  $X^*$  Dependence on Queue Size: Similar to the case of single exchange, in this example an optimal market order size  $M^*$  is an increasing linear function of queue sizes  $Q_1, Q_2$  and the target quantity  $S$ , while optimal limit order sizes  $L_i^*$  are decreasing functions of the corresponding queue sizes  $Q_i$ .
3. Optimal Limit/Market Dependence on  $S$ : As in the case of a single exchange, the optimal order sizes do not depend on the target quantity, but the optimal market order size increases with it.
4. Dependence of  $L^*$  on  $\mu$ : In addition, it is to be noted that each  $L_i^*$  depends on the order flow distributions on both exchanges through  $\mu_{1,2}$  and  $z$ .

## Numerical Solution to the Optimization Problem

1. Objective Function and its Gradient: Computing the objective function in the order placement problem

$$\min_{X \in \mathbb{R}_+^{K+1}} \mathbb{E}[\nu(X, \xi)]$$

or its gradient at any point requires calculating an expectation – a multi-dimensional integral – which, aside from specific examples, is generally not analytically tractable.

2. Approach using the Stochastic Approximation: Stochastic approximation methods, developed specifically for problems where the objective function is an expectation, turn out to be very useful for this problem.
3. Order Placement without Outflow Distribution: This section proposes a procedure for computing the order placement problem which does not require specifying an order outflow distribution.
4. The Stochastic Approximation Method: The stochastic approximation approach and the specific method used are briefly described here.
5.  $V(X)$  and the Gradient  $g(X, \xi)$ : Consider an objective function



$$V(X) \triangleq \mathbb{E}[v(X, \xi)]$$

to be minimized and denote by

$$g(X, \xi) \triangleq \nabla v(X, \xi)$$

where the gradient is taken with respect to  $X$ .

6. Robbins and Monro Stochastic Approximation Algorithm: The Robbins and Monro (1951) stochastic approximation algorithm tackles the problem in the following way.
7. Initial Allocation and Step Size: Choose  $X_0$  and a sequence of ‘step sizes’  $\{\gamma_n\}$ .
8. For  $n = 1, \dots, N$ : Do the next 2 steps.
9. Simulating Random Variables from its Distribution: Simulate an independent random variable  $\xi^n$  with a distribution  $F$ .
10. Stepping to the Next Objective Value: Set

$$X_n = X_{n-1} - \gamma_n g(X_{n-1}, \xi^n)$$

11. Convergence to the Optimal Allocation: This algorithm produces an estimate

$$\hat{X}^* \triangleq X_N$$

which converges to the optimal point  $X^*$  as

$$N \rightarrow \infty$$

under some technical assumptions.

12. Technical Conditions needed for Convergence: Specifically,  $V(X)$  need to be strongly convex, twice continuously differentiable and



$$X^* \in \text{int}\{\mathcal{X}\}$$

where  $\mathcal{X}$  is a non-empty bounded closed convex set, then

$$\mathbb{E}[V(X_n) - V(X^*)] \rightarrow 0$$

Kushner and Yin (2003) contain more details.

13. Constants Impacting Rate of Convergence: The sequence of constants  $\{\gamma_n\}$  affects the rate of convergence.
14. Reducing the Step Size Sensitivity: This sensitivity can be overcome by using, for example, the robust stochastic approximation of Nemirovski, Juditsky, Lan, Shapiro (2009) which follows the same sequence as above with a constant step size

$$\gamma_n = \gamma$$

and uses an average of iterates

$$\hat{X}^* = \frac{1}{N} \sum_{n=1}^N X_n$$

instead of  $X_N$  as an estimate of the optimal point.

15. Performance Bound for the Convergence: Under some weak assumptions, this method achieves a performance bound

$$V(\hat{X}^*) - V(X^*) \leq \frac{D\mathcal{M}}{\sqrt{N}}$$

for a finite  $N$ , where

$$D = \max_{X, X' \in \mathcal{C}} \|X - X'\|_2$$



and

$$\mathcal{M} = \sqrt{\max_{X \in \mathcal{C}} \mathbb{E}[\|g(X, \xi)\|^2]}$$

with  $\|\cdot\|$  defining the  $L_2$  norm.

16.  $V(X)$  as a Well-behaved Function: This method assumes that

$$\min_{X \in \mathcal{X}} \{V(X)\}$$

is sought, where  $V(X)$  is a well-defined and finite-valued expectation for every  $X \in \mathcal{X}$  and  $\mathcal{X}$  is a non-empty bounded closed convex set.

17. The Optimal Iteration Step-size: Moreover  $V(X)$  needs to be continuous and convex in  $\mathcal{X}$ . The optimal step size is

$$\gamma = \frac{D}{\sqrt{N} \cdot \mathcal{M}}$$

18. Robustness to Step Size Mis-specification: Multiplying the optimal step size  $\gamma$  by a constant

$$\theta > 0$$

leads to a performance bound of the same order of magnitude

$$\max(\theta, \theta^{-1}) \cdot \frac{D\mathcal{M}}{\sqrt{N}}$$

i.e., the method is *robust* to step-size mis-specifications.



19. Step Size from “Secular” Bounds: For the current problem one can further bound

$$D \leq \sqrt{K} \cdot S$$

and

$$\mathcal{M} = \sqrt{(s + f + \lambda_O + \lambda_U)^2 + \sum_{k=1}^K (s + r_k + \lambda_O + \lambda_U)^2}$$

20. Explicit Formula for the Stochastic Gradient: It is assumed that on each iteration

$$X_n \in \text{int}\{\mathcal{C}\}$$

– this is enforced by rescaling  $X_n$  when needed and does not affect the convergence – then the stochastic gradient in the current problem is given by:

$$g(X_n, \xi)$$

$$= \begin{pmatrix} s + f - \lambda_U \mathbb{I}_{\{A(X_n, \xi) < S\}} + \lambda_O \mathbb{I}_{\{A(X_n, \xi) > S\}} \\ -(s + r_1) \mathbb{I}_{\{\xi_1 > Q_1 + L_{1n}\}} - \lambda_U \mathbb{I}_{\{A(X_n, \xi) < S, \xi_1 > Q_1 + L_{1n}\}} + \lambda_O \mathbb{I}_{\{A(X_n, \xi) > S, \xi_1 > Q_1 + L_{1n}\}} \\ -(s + r_K) \mathbb{I}_{\{\xi_K > Q_K + L_{Kn}\}} - \lambda_U \mathbb{I}_{\{A(X_n, \xi) < S, \xi_K > Q_K + L_{Kn}\}} + \lambda_O \mathbb{I}_{\{A(X_n, \xi) > S, \xi_K > Q_K + L_{Kn}\}} \end{pmatrix}$$

21. Gradient Dependence on Outflow Variate: Note that  $g(X_n, \xi)$  depends on random variables  $\xi$  only through the indicator functions, which have a simple economic meaning.

22. Economic Meanings of Indicator Functions: For example

$$\mathbb{I}_{\{A(X_n, \xi) < S\}} = 1$$



if the target quantity was not fully executed on the  $n^{th}$  step and  $\mathbb{I}_{\{\xi_k > Q_k + L_{kn}\}}$  if there was an opportunity to execute a larger limit order at exchange  $k$  on that step.

23. Corresponding Interpretation of Numerical Iterations: This leads to a simple interpretation of numerical iterations – at each step order sizes are increased or decreased depending on whether or not the executed quantity was smaller or larger than the target size and whether or not a larger limit order can be filled at the given exchange.
24. Explicit Model for Outflow Distribution: If a model for  $\xi$  is available, one can use it to simulate  $\xi$  and compute a numerical solution that takes into account specific order flow assumptions, e.g., forecasts of future trading volumes.
25. Historical Model for Outflow Distribution: Alternatively, one can use past order fill data to compute indicator functions in  $g(X_n, \xi)$  and obtain a non-parametric numerical solution to the order placement problem, using the empirical distribution of past order fills instead of assuming a functional form for  $F$ .
26. Comparing Numerical and Analytical Solutions: The numerical stability and convergence estimates for  $\hat{X}^*$  are analyzed by comparing them with an analytical solution in the case of a single exchange.
27. Quantifying the Process Parameters: This computation uses

$$Q = 2000 \text{ shares}$$

$$\xi \sim \text{Poisson}(\mu T)$$

$$\mu = 2200 \text{ shares per minute}$$

$$T = 1 \text{ minute}$$

and

$$S = 1000 \text{ shares}$$



28. Pricing Parameters Used in Comparison: The pricing parameters in dollar per share are

$$s = 0.02$$

$$r = 0.002$$

$$f = 0.003$$

and fall in a typical value range for US equities.

29. Overfill and Underfill Penalty Costs: Finally, the penalty costs, also in dollars per share, are set to

$$\lambda_o = 0.024$$

$$\lambda_u = 0.026$$

30. Optimal Allocation from Analytical Expression: According to

$$M^* = S - F^{-1} \left( \frac{2s + f + r}{\lambda_u + s + r} \right) + Q$$

$$L^* = F^{-1} \left( \frac{2s + f + r}{\lambda_u + s + r} \right) - Q$$

the optimal allocation is

$$(M^*, L^*) = (728, 272) \text{ shares}$$



31. Starting Allocation and Step Size: Numerical estimates  $\hat{X}^*$  were computed for five initial points  $X_0$  and different number of iterates  $N$  in the stochastic approximation, using a step size

$$\gamma \triangleq \frac{\sqrt{K} \cdot S}{\sqrt{N(s + f + \lambda_O + \lambda_U)^2 + N \sum_{k=1}^K (s + r_k + \lambda_O + \lambda_U)^2}}$$

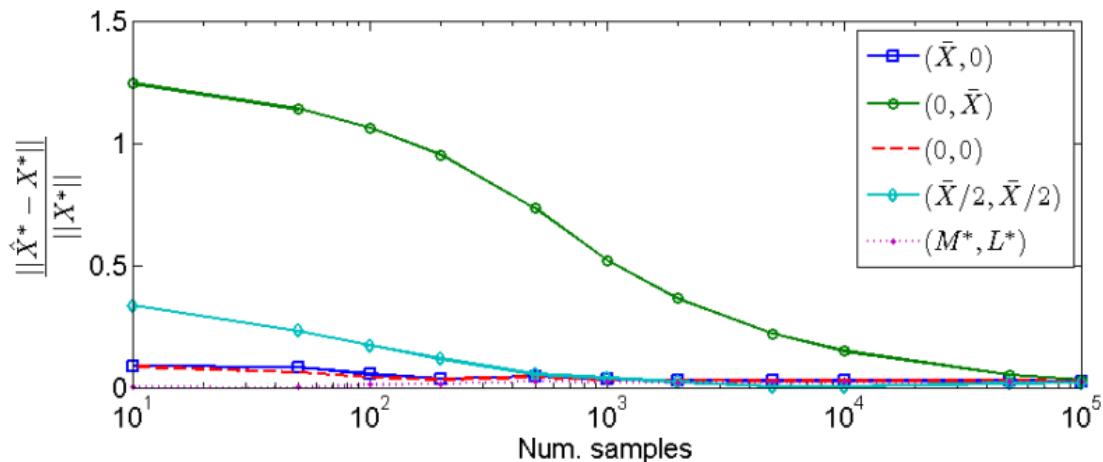
32. Averaging the Cost over Observations: For each choice of  $X_0$  and  $N$ , an additional

$$L = 1000 \text{ observations}$$

of  $\xi$  are simulated to estimate the objective values  $V(X)$  with sample averages

$$V(X) = \frac{1}{L} \sum_{i=1}^L v(X, \xi_i)$$

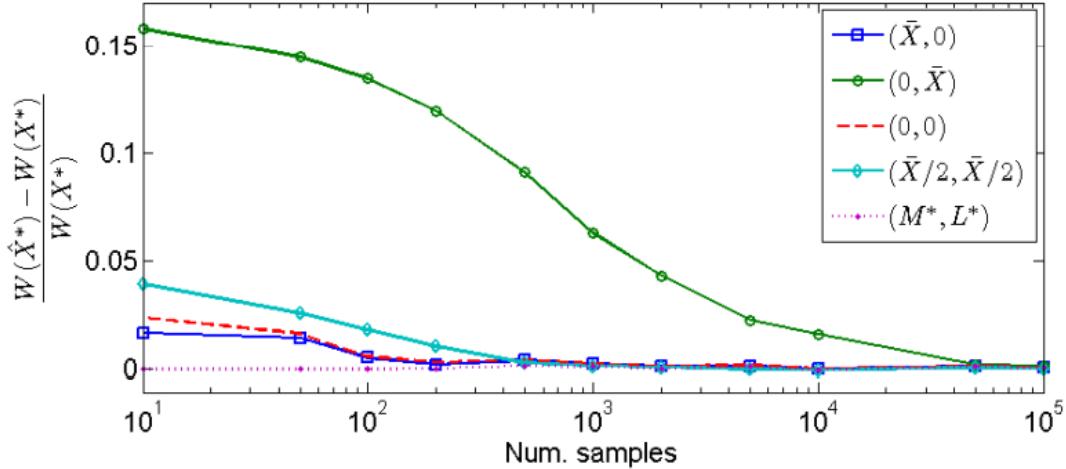
33. Convergence of Order Allocation Vectors:



Convergence of order allocation vectors to an optimal point for different initial points.



34. Convergence of Objective Values:



Convergence of order allocation vectors to an optimal point for different initial points.

35. Illustration of the Allocation Convergence: The above figures show that estimates converge to  $X^*$  regardless of  $X_0$ . When

$$X_0 = X^*$$

estimates remain close to that point.

- 36. Speed of the Allocation Convergence: Convergence is also quiet fast – after as few as 50 samples the algorithm can be within 2% of the optimal objective value.
- 37. Starting Allocation at the Boundary: In the worst case of initial points at the boundary it can take a few thousand samples to converge.
- 38. Convergence of the Execution Cost: It is also worth noting that the convergence in terms of the objective value occurs significantly faster than convergence in terms of order allocation vector.
- 39. Limit Order across Multiple Exchanges: The savings from optimal limit order and from dividing a limit order across multiple exchanges was also estimated.



40. Comparison across Naïve Allocation Schemes: A pure market order allocation is denoted by

$$X_M = (S, 0, \dots, 0)$$

single limit order allocation by

$$X_L = (0, S, \dots, 0)$$

and an equal split allocation by

$$X_E = \left( \frac{S}{K+1}, \frac{S}{K+1}, \dots, \frac{S}{K+1} \right)$$

41. Table - Savings from Order Splitting:

K	Order allocation in % of S						Average cost, in cents per share				
	$M^*$	$L_1^*$	$L_2^*$	$L_3^*$	$L_4^*$	$L_5^*$	Total	$W(X_M)$	$W(X_L)$	$W(X_E)$	$W(X^*)$
$S = 500$											
1	82%	18%					100%	2.35	2.19	0.83	1.54
2	15%	44%	44%				103%	2.35	2.22	-0.57	-0.85
3	1%	34%	34%	34%			102%	2.35	2.21	-1.02	-1.99
4	0%	26%	25%	26%	26%		102%	2.35	2.20	-1.25	-2.06
5	0%	22%	21%	20%	20%	20%	103%	2.35	2.22	-1.40	-2.05
$S = 1000$											
1	94%	6%					100%	2.35	3.65	2.27	2.07
2	56%	27%	27%				111%	2.35	3.64	1.28	0.77
3	35%	23%	23%	23%			104%	2.35	3.65	0.09	-0.07
4	15%	23%	23%	22%	23%		106%	2.35	3.64	-0.88	-0.90
5	1%	21%	21%	21%	21%	21%	106%	2.35	3.65	-1.29	-1.64
$S = 5000$											
1	97%	3%					100%	2.35	4.81	3.44	2.22
2	88%	8%	8%				104%	2.35	4.81	3.60	2.10
3	83%	9%	9%	9%			110%	2.35	4.81	3.55	1.95
4	79%	11%	11%	11%	11%		124%	2.35	4.81	3.39	1.79
5	75%	11%	11%	11%	11%	11%	129%	2.35	4.82	3.22	1.62



42. Simulation using Equal Allocation Start: The table above presents outputs from the numerical algorithm with

$$X_O = X_E$$

$$N = 1000$$

$$L = 1000$$

for different order sizes  $S$  and number of exchanges

$$K = 1, \dots, 5$$

43. Process and Exchange Parameters: The parameters  $s, f, r, \lambda_U$ , and  $\lambda_O$  are same as in the previous simulation and exchanges are identical replicas of each other:

$$r_k = r$$

$$Q_k = Q$$

and

$$\xi_{nk} \sim Poisson(\mu T)$$

are i.i.d. copies of each other, where

$$k = 1, \dots, K$$

$$n = 1, \dots, N$$



44. Cost Comparison with Naïve Allocation: Order allocations produced by stochastic approximation clearly outperform the naïve benchmarks, especially when the target quantity  $S$  is relatively small and cost savings of limit orders can be fully explained.
45. Comparing  $X_L$  and  $X_E$  Allocations: Comparing  $W(X_L)$  and  $W(X_E)$  one can also see that splitting a limit order across multiple exchanges can be very advantageous when limit order fills are independent.
46. Equal Limit Split across Exchanges: Since multiple exchanges in this example are copies of each other, the algorithm splits the total limit order amount equally among them.
47. Distinction between  $X_E$  and  $X^*$ : This is not the same as the allocation  $X_E$  because the latter sets the market order size to  $\frac{S}{K+1}$ , which may be too large or too small depending on  $S$  and the number of exchanges available.
48. Optimal Order is Over-allocated: Another interesting feature of the numerical solution  $\hat{X}^*$  is the tendency to oversize the total quantity of limit orders, which is clearly observed for

$$S = 1000, 5000$$

and

$$K = 4, 5$$

49. Consequence of Outflow Distribution Independence: This may be a consequence of assumed independence between  $\xi_k$  – by submitting large orders to multiple exchanges the algorithm reduces the probability of falling behind the target quantity with a relatively low probability of exceeding it.
50. Sensitivity Analysis on Two-Exchanges: To illustrate the structure of a numerical structure, Cont and Kukanov (2017) performed a sensitivity analysis with

$$K = 2$$



exchanges and parameters

$$Q_1 = Q_2 = 2000$$

$$S = 1000$$

$$\xi_{1,2} \sim Poisson(\mu_{1,2}T)$$

$$\mu_1 = 2600$$

$$\mu_2 = 2200$$

$$T = 1$$

$$s = 0.02$$

$$r_1 = r_2 = 0.002$$

$$f = 0.003$$

$$\lambda_U = 0.26$$

and

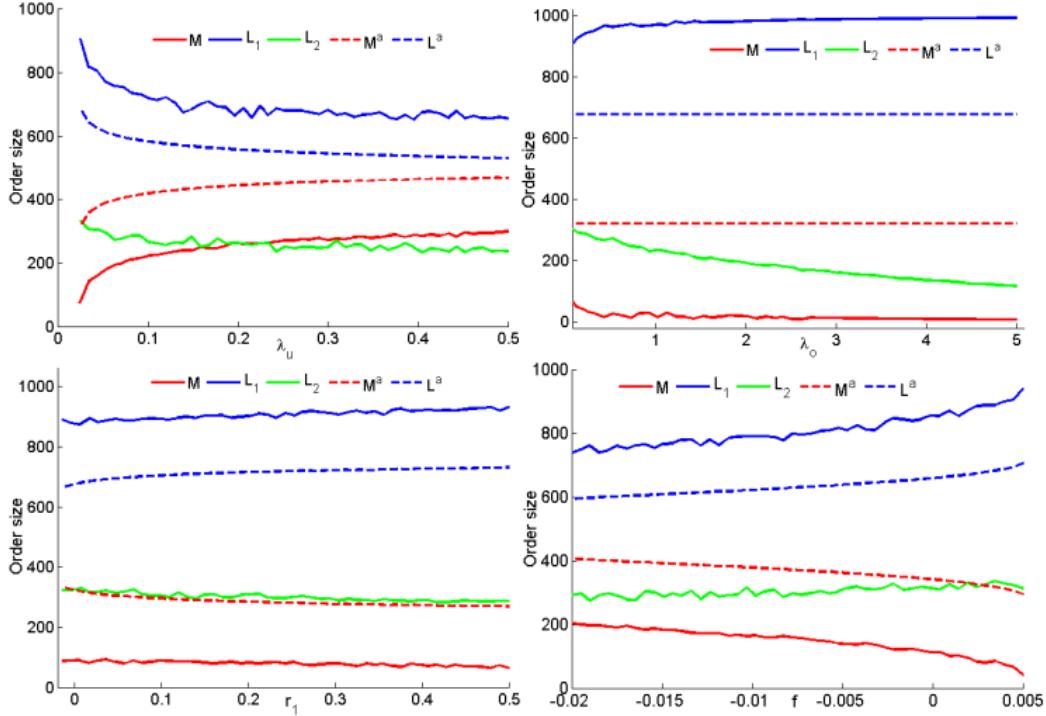
$$\lambda_O = 0.24$$

51. Parameter Sensitivity Analysis #1: Sensitivity analysis for a numerical solution

$$\hat{X}^* = (M, L_1, L_2)$$



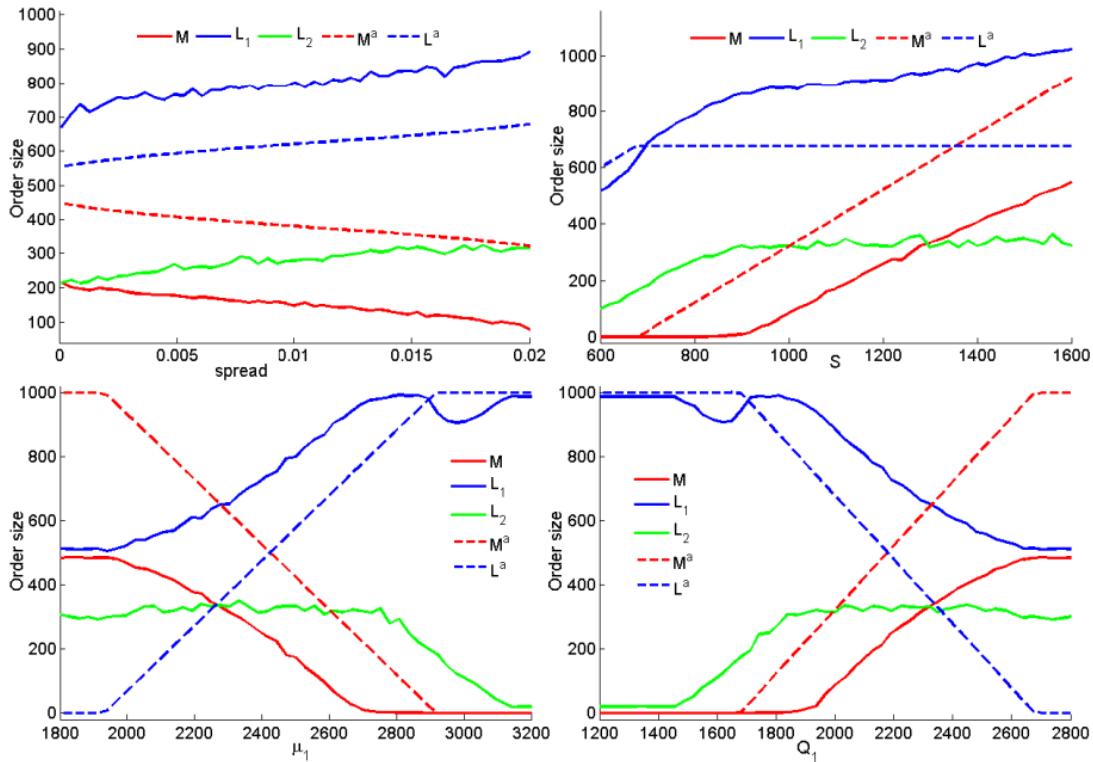
with two exchanges and an optimal solution  $(M_a, L_a)$  with the first exchange only.



## 52. Parameter Sensitivity Analysis #2: Sensitivity analysis for a numerical solution

$$\hat{X}^* = (M, L_1, L_2)$$

with two exchanges and an optimal solution  $(M_a, L_a)$  with the first exchange only.



53. Sensitivity Comparison with Single Exchange: Varying some of these parameters one at a time, the above figures show a plot of the numerical solution  $\hat{X}^*$  after

$$N = 1000$$

iterations, together with an analytical solution for a single exchange.

54. Allocation Vector Dependence on  $\lambda_U$ : Similar to the single exchange case, limit order sizes on two exchanges  $L_1, L_2$  decrease and market order size  $M$  increases as the penalty  $\lambda_U$  increases.

55.  $s, r_1$ , and  $f$  Dependence: Increasing the half-spread  $s$ , the rebate  $r_1$ , of the fee  $f$  makes a limit order on exchange number one more attractive, so  $L_1$  increases and  $M$  decreases.



56. Execution Risk Dominates with  $\lambda_U$ : Because the penalty  $\lambda_U$  is large in this example, execution risk is more important fees and rebates, therefore the queue size  $Q_1$  and the order outflow men  $\mu_1$  have a much stronger effect on the optimal solution than  $r_1$ .
57. Dependence on  $Q_1$  and  $\mu_1$ : Both decreasing the  $Q_1$  and increasing  $\mu_1$  make a limit order fill more likely at exchange manner one and  $L_1$  increases. The observed drop in  $L_1$  for large  $\mu_1$  and small  $Q_1$  is a feature of this example, Cont and Kukanov (2017) were not able to replicate it for other distributions of  $\xi$ .
58. Allocation Vector Dependence on  $S$ : Finally, as in the case of a single exchange, the target size  $S$  has a strong effect on the optimal order allocation.
59. Market Order increases with  $S$ : Only limit orders are used while  $S$  is small, but as soon as it becomes larger it is difficult to fill that amount solely with limit orders and the optimal market order size begins to grow to limit the execution risk.

## Conclusion

1. Limit Order across Multiple Names: This chapter formulated and solved the problem of placing a small batch of orders on multiple trading venues.
2. Analytical Allocation for Single Exchange: In the case when there is a single exchange, an optimal split between a limit and a market-order sizes we derived.
3. Fast Converging Stochastic Approximation Scheme: For more general cases, a stochastic approximation algorithm that is shown to quickly converge to an optimal point was proposed and tested.
4. Optimal Allocation across Multiple Venues: The chapter explored the properties of an optimal order allocation policy and showed that splitting an order across multiple exchanges can lead to a substantial reduction in transaction costs.

## References



- Alfonsi, A., A. Fruth, and A. Schied (2010): Optimal Execution Strategies in Limit Order Books with General Shape Functions *Quantitative Finance* **10** (2) 143-157
- Aliprantis, C., and O. Burkinshaw (1998): *Principles of Real Analysis 3<sup>rd</sup> Edition* Academic Press Cambridge, MA
- Almgren, R. F., and N. Chriss (2000): Optimal Execution of Portfolio Transactions *Journal of Risk* **3** (2) 5-39.
- Bayraktar, E., and M. Ludkovski (2012): Liquidation in Limit Order Books with Controlled Intensity **arXiv**
- Bertsimas, D., and A. W. Lo (1998): Optimal Control of Execution Costs *Journal of Financial Markets* **1** 1-50
- Boehmer, E., and R. Jennings (2007): Public Disclosure and Private Decisions: Equity Market Execution Quality and Order Routing *Review of Financial Studies* **20** (2) 315-358
- Cont, R. (2011): Statistical Modeling of High-frequency Data *IEEE Signal Processing* **28** (5) 16-25
- Cont, R., and A. de Larrard (2013): Price Dynamics in a Markovian Limit Order Market *SIAM Journal on Financial Mathematics* **4** (1) 1-25
- Cont, R., and A. Kukanov (2017): Optimal Order Placement in Limit Order Markets *Quantitative Finance* **17** (1) 21-39
- Foucault, T., and A. J. Menkveld (2008): Competition for Order Flow and Smart Order Routing Systems *Journal of Finance* **63** (1) 119-158
- Ganchev, K., Y. Nevmyvaka, M. Kearns, and J. W. Vaughan (2010): Censored Exploration and the Dark Pool Problem *Communications of the ACM* **53** (5) 99-107
- Gueant, O., and C. A. Lehalle (2013): General Intensity Shapes in Optimal Liquidation **arXiv**
- Guilbaud, F., and H. Pham (2012): Optimal High Frequency Trading in a Pro-rata Microstructure with Predictive Information **arXiv**
- Huitoma, R. (2014): Optimal Portfolio Execution using Market and Limit Orders **eSSRN**



- Kushner, H., and G. Yin (2003): *Stochastic Approximation and Recursive Algorithms and Applications* Springer New York, NY
- Laruelle, S., A. A. Lehalle, and G. Pages (2010): Optimal Split of Orders across Liquidity Pools: A Stochastic Algorithm Approach
- Maglaras, C., C. Moallemi, and H. Zhang (2011): [Optimal Order Flow Routing, Exchange Competition, and the Effect of Make/Take Fees](#)
- Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro (2009): Robust Stochastic Approximation Approach to Stochastic Programming *SIAM Journal on Optimization* **19 (4)** 1574-1609
- Obizhaeva, A., and J. Wang (2006): [Optimal Trading Strategy and Supply/Demand Dynamics](#)
- Predoiu, S., G. Shaikhet, and S. Shreve (2011): Optimal Execution in a General One-sided Limit-Order Book *SIAM Journal on Financial Mathematics* **2 (1)** 183-212
- Robbins, H., and S. Monro (1951): A Stochastic Approximation Method *Annals of Mathematical Statistics* **22 (3)** 400-407



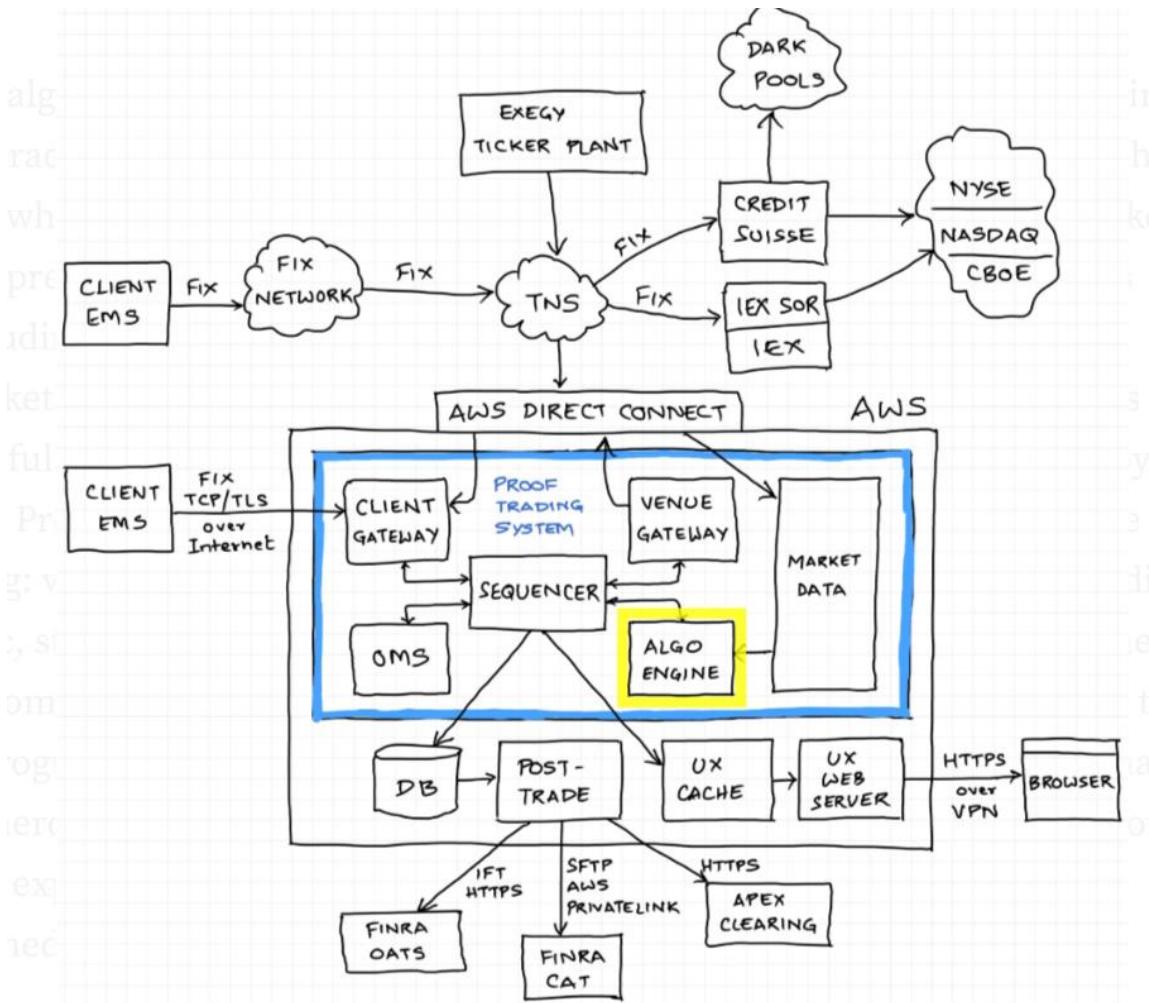
## Trading Strategy Algo Framework

### Introduction

1. The Institutional Trading Platform: From the outside, the institutional trading platform may seem like a giant black box of code (Aisen (2021)).
2. Role of the Trading Platform: An asset manager sends an order to buy or sell a particular stock, the black box enriches some numbers and in turn sends out smaller orders to the street. If the algo does a good job, those child orders trade at relatively good prices.
3. Specialized Applications working in Tandem: Beneath the hood, there are dozens of specialized software applications performing various essential tasks such as FIX translation, order validation, risk management, market data consumption, and post-trade processing.
4. Focus on the Algo Engine: This chapter focusses on the key suite of components responsible for making trading decisions – the algo engine.

### The Algo Engine

1. Role of the Algo Container: The algo engine – or container – is the core piece of the system that contains the trading logic.

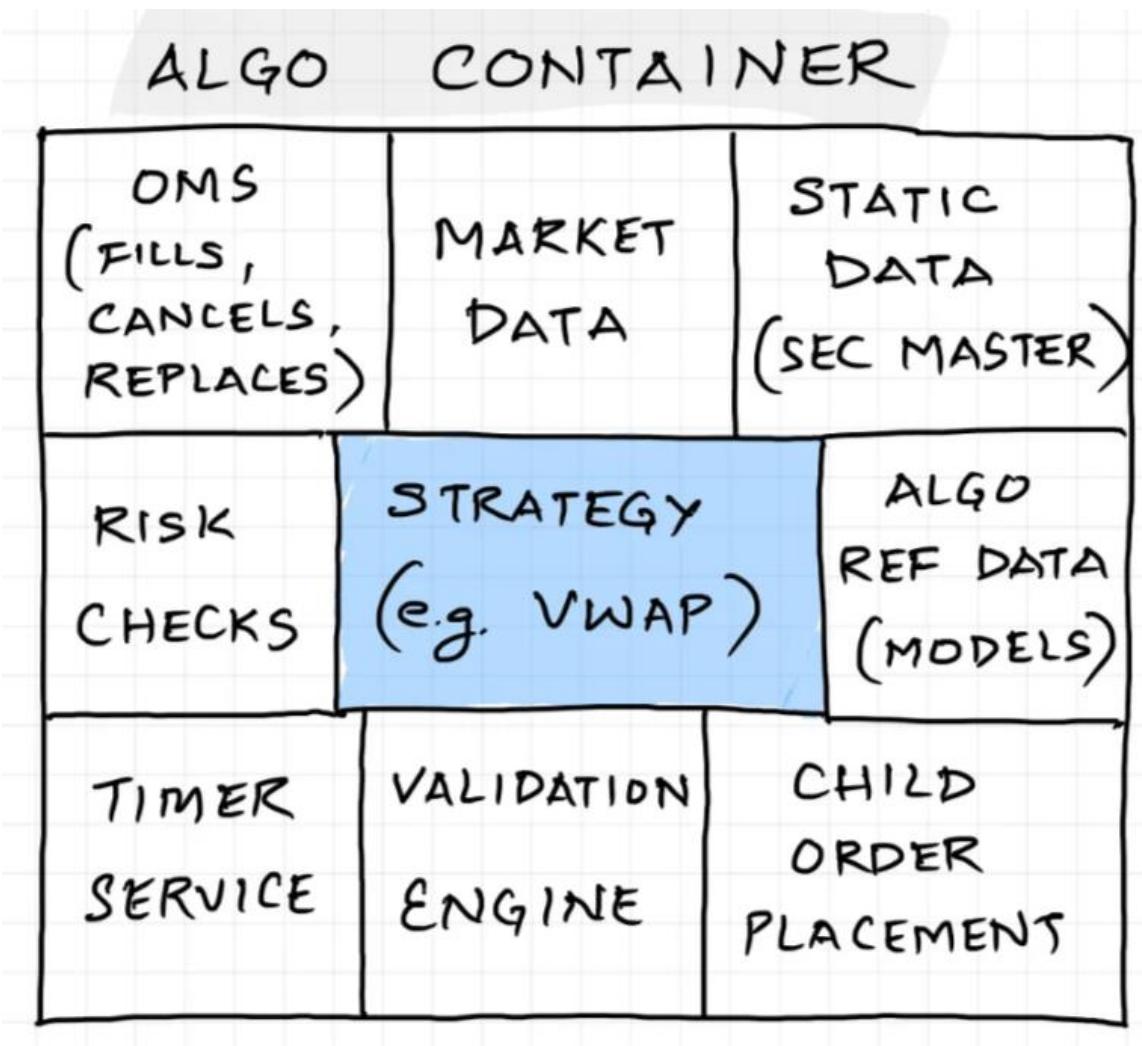


2. **Role of the Algo Engine:** The Algo Engine takes the client's high-level instruction, and decides how and when to slice out child orders based on the current state of the market and the pre-loaded quantitative models.
3. **Tasks Performed by the Engine:** The engine also performs many related tasks including order validation, local order management, consumption of market and reference data, and risk checks.
4. **Algorithms Hosted by the Engine:** The Algo Engine provides helpful callbacks and contains numerous safety mechanisms, so that the algo can be safely and reliably hosted.
5. **Safety Features Offered by the Engine:** The engine keeps a tight leash on the trading logic, strictly enforcing that it behaves appropriately and stays within the client's



instructions. Even if the algo has an unfortunate bug or tries to “go rogue”, the engine blocks it from causing damage.

6. Checks within the Algo itself: The algo typically has numerous checks within itself to ensure that it acts properly, but this is augmented by additional robustness provided by the engine.
7. Component of the Engine:



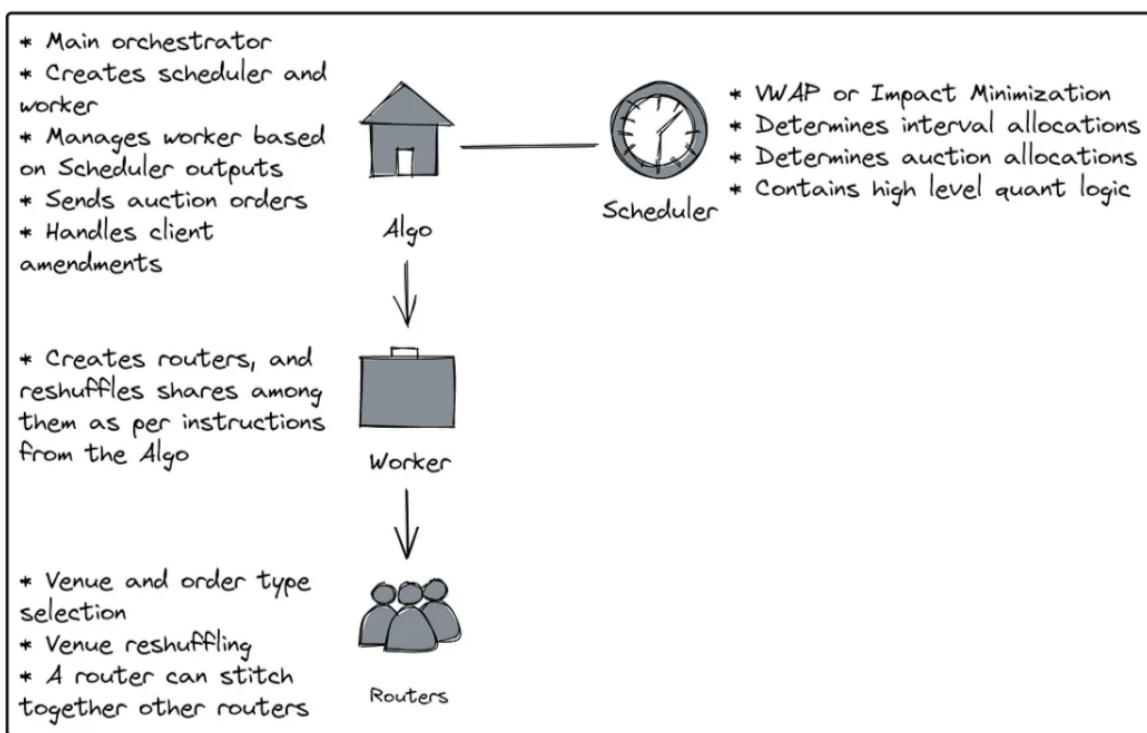
## The Trading Strategies (Algos)



1. Trading Strategies Available in the Engine: At a high level, there are two strategies/trading instructions that the user can specify.
2. VWAP Strategy: This strategy trades in line with the market volume based on a dynamic volume prediction model.
3. DROP LIQUIDITY SEEKER Strategy: This is a hybrid liquidity seeker/impact minimization algo that balances the speed of execution with avoiding undue market impact as per its equivalent model (Bishop (2021)), with an opportunistic component searching for block liquidity in parallel.
4. Safety Check inside these Strategies: Additionally, both algos have two available override check boxes:
  - a. They must complete.
  - b. They must exclude auctions.The ticket is maintained as simple as is possible, while still covering the most common use cases.
5. Modules inside these Algo Strategies: Both of these top-level strategies are implemented via a common software architecture. The trading strategy is in itself a collection of modular components, each with an important role to play.

## Trading Logic Components

1. Algo, Scheduler, Worker, and Routers: The Strategy component inside the Algo Engine contains four main pieces – the *Algo*, the *Scheduler*, the *Worker*, and various *Routers*.
2. Components of the Trading Strategy:



## Trading Logic Components – Algo

1. Main Orchestrator of the Trading Logic: The Top Algo layer manages all upstream interactions with the user, e.g., handling new orders and amendments, and it creates and manages the Scheduler and the Worker.
2. Opening/Closing Auction Child Orders: The Algo layer is also responsible for slicing shares to the opening and the closing auctions.

## Trading Logic Components – Scheduler

1. Role of the Scheduler: The scheduler is the bridge between the various quantitative models and the trading logic.



2. Strategy-specific Customization of the Scheduler: The primary difference between the trading strategies is that they invoke different versions of the scheduler. The following 3 models are encapsulated by the scheduler.
3. Pre-trade Model: This model suggests the total number of shares to trade over the duration of the order. This is used for need-not-compete orders.
4. Impact Minimization Model: This model suggests the pace at which the Algo should trade at any given time. This is used for DROP LIQUIDITY SEEKER orders.
5. Dynamic Volume Prediction Model: This model predicts the amount of volume throughout the day and at the auctions based on what has happened historically and intra-day so far. This is used for VWAP orders.

## Trading Logic Component – Worker

1. Management of Intra-day Trading: This middle layer manages and deploys three layers of intra-day trading tactics, i.e., routers, and shuffles between them as dictated by the Algo.
2. POST, TAKE, and OPPO Routers: The three intra-day tactics/routers are for:
  - a. POST – Passively adding liquidity
  - b. TAKE – immediately removing liquidity
  - c. OPPO – opportunistically seeking block liquidity

## Trading Logic Component – Routers

1. Allocation at the Price Level: The lower layer tactics allocate and reshuffle orders to one or multiple external destinations at a single price level.
2. Custom Use of Multiple Sub-routers: A router can also stitch together multiple sub-routers, for example a serial mid-point all-or-none router + a spread crossing order to the IEX router.



3. POST #1: POST generally starts with a 2:1 split-post on the near-side between IEX D-Limit and the primary exchange.
4. POST #2: POST proportionately reshuffles between two legs based on where it gets filled, and re-pegs to the inside roughly every 30-60 seconds if the stock drifts away.
5. TAKE: TAKE first serially pings a handful of dark-pools using all-or-none IOCs, and if that fails, it crosses the spread to remove liquidity using the IEX router.
6. OPPO: OPPO searches for block liquidity by split-posting mid-point orders across several dark pools and exchanges using a high minimum quantity, generally in the thousands of shares. This may be supplemented with other order types such as conditionals in the future.

## Life-cycle of an Order

To really understand the interaction between these components, it is helpful to walk through the specific logic at various phases of an order, from creation through completion or cancellation.

### Life-cycle of an Order – Arrival

1. Market Data Subscription and Scheduler Creation: Upon receipt of a valid order, the Algo subscribes to market data and then creates the Scheduler.
2. Algo-specific Scheduler Customizer: For a VWAP order, the Algo creates VWAP scheduler; for a DROP LIQUIDITY SEEKER order, it creates an Impact Minimization Scheduler.
3. Algo Engine Wakeup Call-back Invocations: Then, the Algo registers with the Algo Engine for various wakeup callbacks; at the start time, shortly before auction cutoff times, at the cleanup time, i.e., shortly before the end, and at the end time.



## Life-cycle of an Order – Wakeup Logic

1. Times the Callback is Invoked: The Algo Engine wakes the Algo at the following times as requested.
2. Upon a Pre-auction Wakeup: The Algo slices shares to the primary auction.
3. At the Start Time: The Algo creates the worker and begins the first interval – the Interval logic is shown below.
4. Upon the Cleanup wakeup: The Algo tells the Worker to complete any remaining scheduled shares.
5. Upon the End Time Wakeup: The Algo cancels any outstanding orders and then provides an out to the user.

## Life-cycle of an Order – Interval Logic

1. Order Life as Discrete Intervals: The Algo strategies treat the life of an order as a series of distinct “intervals”, generally about 5-10 minutes windows of time that share common trading behavior.
2. Information Needed by the Scheduler: At the start of a new interval, the Algo requests the following two pieces of information from the Scheduler.
3. Length of the Interval #1: The *length of the interval* is used to request a wakeup at the end.
4. Length of the Interval #2: Intervals are generally about 5-10 minutes in length, but depend on other factors too like the time of the day – intervals are stretched longer in the morning when spreads are wider, and compressed as the day progresses.
5. Length of the Interval #3: Additionally, smaller orders have longer intervals and vice versa. All interval durations are randomized.
6. Number of Shares Scheduled #1: This represents the number of shares scheduled to trade in the interval. This number is the first key difference among the algo strategies.



7. Number of Shares Scheduled #2: The VWAP Scheduler uses the dynamic volume prediction model to predict what percentage of the day's volume will have traded by the end of the current interval.
8. Number of Shares Scheduled #3: The Impact Minimization Scheduler gets this value by running a dynamic programming cycle of the impact minimization model.
9. Number of Shares Scheduled #4: The Algo then uses this information to tell the Worker to complete any shares outstanding from previous interval and start working a new interval quantity.
10. Liquidity Seeking Scheduler Augmentation: The second key difference between the strategies is the liquidity seeking piece.
11. DROP LIQUIDITY SEEKER Block Liquidity: Throughout the life of a DROP LIQUIDITY SEEKER order, the Algo tells the Worker to search for block liquidity at the mid-point with all remaining unscheduled shares using an Opportunistic router.
12. Algo “catch up” Callbacks: Additionally, the Algo requests “catch up” wakeups throughout the interval where it checks if the Worker is falling behind, and if so, tells it to cross the spread to keep up pace.

## Life-cycle of an Order – Order Amendment

1. Order Replacement Request: Upon receipt of a client replace request, the Algo generally cancels or amends the Worker to comply with the new instructions.
2. New Scheduler Instance to Handle the Amend: In most cases, the Algo also creates a new Scheduler to use from this point forward.

## Trading Objectives/Design Process

1. Buy-side Perspective on the Execution: This algo was designed by thinking how one would approach the execution if they were on the buy-side.



2. Driving Principle of Best Execution: The driving principle, of course, is best execution with less emphasis on a “consistent” user experience, e.g., the Algo doesn’t immediately trade 100 shares so that the user can see it’s working.

## Trading Objectives/Design Process – Low Level

1. Objectives at the Micro-level: At the micro-level, the objectives are two-fold; avoiding adverse selection when adding liquidity, and capturing as much volume at the best possible price(s) while removing liquidity.
2. Avoiding Adverse Selection - Prop Firm Perspective: When the market transitions to a new price level, there is a flood of trading activity where the fastest proprietary firms race to pick off the resting orders at the old price level and establish queue position at the new price level – these are two similar but different strategies.
3. Avoiding Adverse Selection - Sell-side Perspective: As a sell-side firm, it is not plausible to effectively compete in these races, so the only viable way to prevent adverse selection is to utilize built-in protections available in trading venues, such as IEX D-Peg and D-Limit, and NASDAQ MELO.
4. Removing Liquidity using IEX Router: Rather than re-invent the wheel, the Algo starts off by using the IEX Router when crossing the spread, which costs only 1 mil and gets ~99% fill rates.
5. Removing Liquidity - Dark Pools Check: In most cases, various dark pools are checked using the all-or-none mid-point order prior to crossing the spread.
6. Impact of All-or-None Mid-point Peg: Because these mid-point pegs are all-or-none, either the full amount gets done at the mid, or nothing happens and an immaterial amount of time has been wasted.

## Trading Objectives/Design Process – High Level



1. Algo Execution through the Day: Even more important than the low-level strategy is the high-level strategy, i.e., how should the Algo spread out trading throughout the day.
2. Intra-day Factors to be Researched: The research stoked by intra-day spread factors can be summarized using the following four determinants.
3. Determinant #1 - Distilled Impact: Addresses the question: “How does one measure success at a high level?”
4. Determinant #2 - Volume Prediction Model: Addresses the question: “How much volume does one expect to trade during the life of the order”
5. Determinant #3 – Pre-trade Model: “How much should the Algo be willing to trade during the life of the order”
6. Determinant #4 - Impact Minimization Model: “How should one pace the trading activity throughout the life of the order”
7. Challenges evaluating High-Level Performance: Demonstrable performance is the ultimate goal, but the above questions are all extremely difficult to answer as higher-level trading data is sparse and noisy.
8. Street-wide Point of Performance Comparison: That, combined with an opaque industry at large, makes it tough to build confidence in the performance numbers or find a reliable point of reference.
9. Optimization at the Micro level: In the meantime, the approach has been to bite off the significant low-hanging fruit at the micro-level by avoiding harmful/conflicted practices and properly utilizing exchange and dark pool order types.

## References

- Aisen, D. (2021): [The Trading Strategy](#)
- Bishop, A. (2021): [Rejecting the Black Box: An Inside Look at the Design](#)



## A Volume-Weighted Average Approach

### Introduction

1. VWAP as a Metric for Algorithm Performance: Comparing the Strategy VWAP to the general market VWAP has pros and cons.
2. Pros of the VWAP Metric #1: For one thing, it does considerably remove noise from measurements.
3. Cons of the VWAP Metric: However, it also introduces circularity; the measurement is against a benchmark that is self-influenced. If a participant trades heavily, the participant's activity impacts the prices across the market.
4. Search for Alternate Performance Benchmark: This creates a blind spot for the price impact to hide; whatever impact makes it into the benchmark will not show up in the difference between the participant's performance and the benchmark. For this reason, alternate metrics such as distilled impact are explored elsewhere.
5. Pros of the VWAP Metric #2: Another pro of VWAP is the it suggests a natural secondary goal; distributing the trading volume throughout the day in a way that is similar to how the overall trading volume is distributed.
6. Formulation of the Volume Prediction Algorithm: The rest of the sections formulate the prediction problem, introduce the preliminary approach, and summarize the results and the first version of the VWAP predictor.

### Defining the Prediction Problem

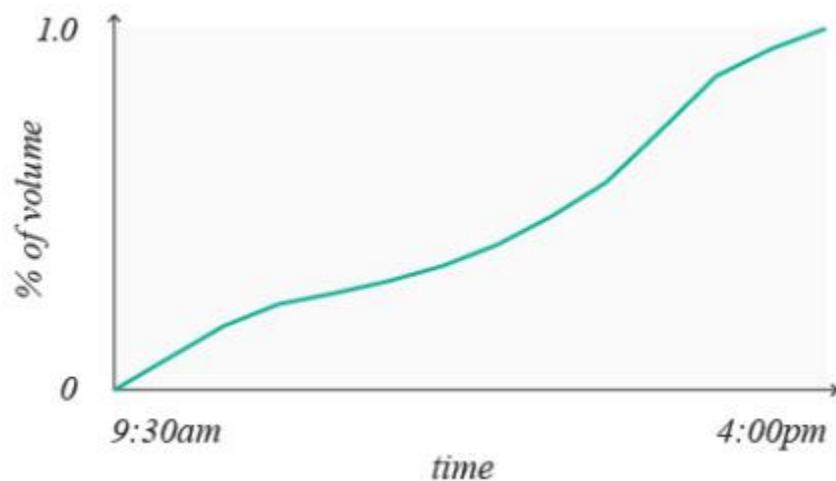
1. Problem Setup – Stock Purchase Schedule: Suppose a participant wants to buy  $X$  shares of a particular stock over the course of a day, and wants to match the volume-weighted average price as closely as possible.



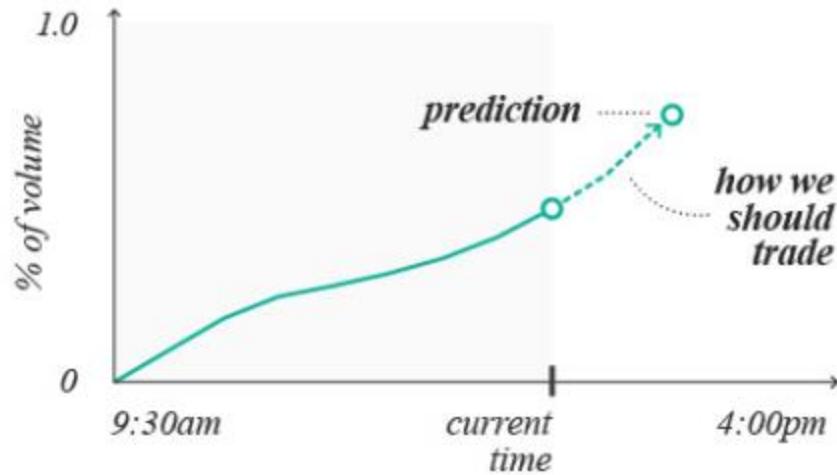
2. Start/Close of the Trading Day: For now, it is imagined that the participant is going to start trading at the beginning of the trading day – including the opening auction – and stop trading at the close – putting any remaining volume into the close auction.
3. Trading over a Smaller Window: Later discussion considers where one is seeking to trade over a smaller time interval that may start later or end earlier.
4. Volume Curve Based Trading Approach: A typical approach to this is to try to trade roughly in line with the *volume curve*, meaning that if 10% of the day's volume trades within a certain window, then buy  $\frac{X}{10}$  shares within that time window.
5. Explicit Trading may be Unneeded: It must be noted upfront that this may not be necessary; someone could offer a crossing mechanism, for example, that matches shares to be traded and waits for the true VWAP price of the day to be determined before pricing the trades.
6. Challenge #1 with this Approach: Sticking to the volume curve, the challenge is visible on zooming into the phrase “10% of the day’s volume ...” in the above example. This is something that will not be known until the end of the day.
7. Trading Task/Daily Volume Units: There is a mismatch here between the units of the trading task – buy  $X$  shares, and the units of the volume to match, i.e., trade  $Y\%$  of volume in this interval.
8. Trading Task in ADV Percentage Terms: If one wanted to buy  $X\%$  of the volume over the course of the day but didn’t care about how many absolute shares this represented, progress can be tracked against this metric in time and try to slow down or catch up as desired. Likewise, progress can be tracked in real-time if one wanted to buy  $X$  shares out of the next  $Y$  traded shares.
9. Proportional Participation across the Day: Wanting to buy an absolute number of shares and distributing it proportionately makes it less obvious how one might use real-time data to adjust the actions.
10. VWAP’s Use of Volume Curve: This is why many VWAP algorithms largely ignore real-time data and rely upon *volume curve* predictions informed by recent completed trading days.



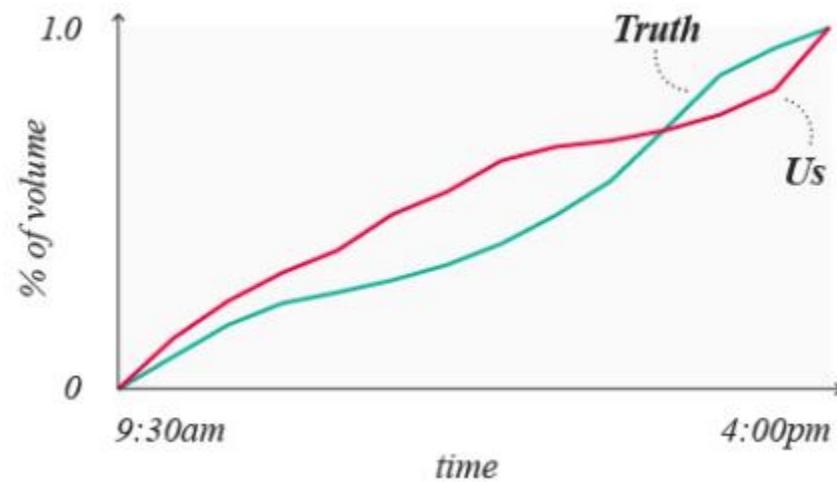
11. Bucket-wise Realized Volume Percentage: A typical default is to look at, say, the last 20 completed trading days and compute the average percentage of the volume falling into each time bucket for each symbol.
12. Initial Estimate of Interval Volume: This may be equivalently viewed as giving a prediction for the cumulative percentage of the volume one expects to have traded by the end of each time interval.
13. Trading Action for the Interval: If the above prediction is treated as the truth, it is then clear what should be done at the beginning of the interval. If one has traded  $Y\%$  of the  $X$  shares so far, and expects  $Z\%$  of the day's volume to be traded by the end of the interval, then one should try to trade an additional  $(Z - Y)\%$  of the  $X$  shares in this interval.
14. Piece wise Linear Interval Trade Representation: For initialization purposes, one makes the assumption that over each interval, the volume trades at a steady rate. This means that the cumulative volume percentage is approximated by a piece-wise linear function curve.



15. Interval Trading Volume Prediction: If the prediction has been on target so far, and has correctly predicted where one needs to be by the end of the next interval, the volume curve is as below.

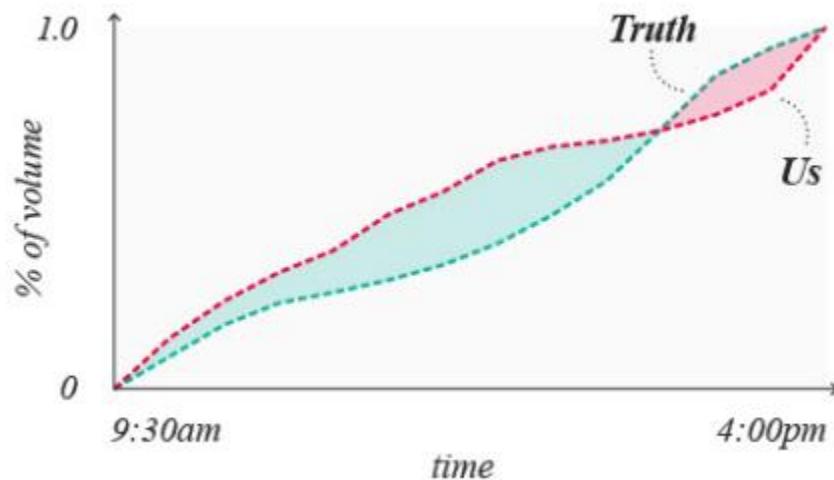


16. Incorporating Volume Prediction Mismatches: Using solely pre-computed averages leaves one no mechanism for adjusting when off. After the trading day is completed, one knows the true picture, and the performance can be evaluated.
17. Cumulative Relative Volume vs. Time: One can plot the true curve of the cumulative relative volume versus time, vs. the curve of one's own trading, as below.





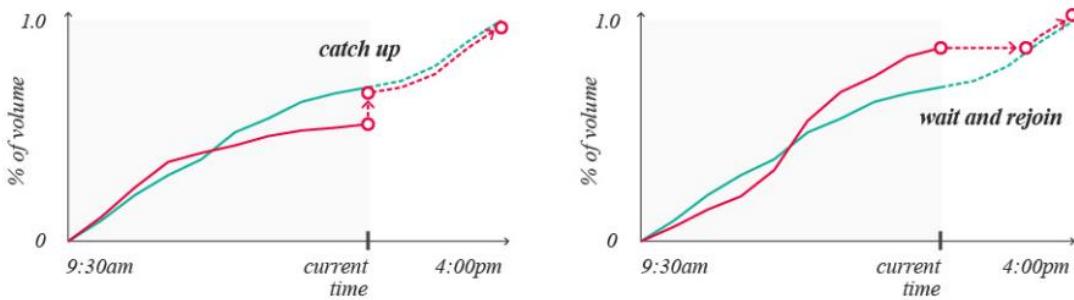
18. Quantification of the Trading Mismatch Error: There are many reasonable choices for how to quantify the error, but this section focusses on the total area between the curves.



19. Integrand Based Error Metric: This is the integral of  $|Truth(t) - Us(t)|$ . This metric is easy to visualize, and intuitively punishes equally for being ahead or behind, and it also reflects the desire to correct errors as quickly as possible.
20. Maximization of the Metric Integrand: If one is currently behind and seeking to minimize  $|Truth(t) - Us(t)|$ , one should want to catch up as quickly as possible, rather than, say, gradually making up for the deficit over the remaining time.
21. Reducing the Trade Volume Drift: This is desirable behavior because the price of later trading is likely to deviate further and further away from the current price, so making up the deficit is likely to result in greater deviation from the TWAP price than making it up now. Setting a goal of minimizing the integral  $|Truth(t) - Us(t)|$  over time is one way of incorporating this.
22. Exception - Significant Market Impact: There may be an exception that if one is so behind that quickly catching up would be unusual and likely to cause inordinate price impact. Generally, however, it is believed that catching up small amounts sooner than later is a sound strategy under this metric.



23. Practical Use of the Strategy: Suppose one could use real-time data to make higher quality predictions of what percentage of the day's volume has been traded so far as well as what percentage of the day's volume will have traded by the end of the next time interval. The predictions might suggest that one is ahead or behind the target.
24.  $Us(t)$  is behind  $Truth(t)$ : If the goal is to minimize the area between the  $Truth(t)$  and the  $Us(t)$  functions over time, and the participant is behind, one should attempt to immediately catch up and then follow a linear path to the next predicted interval.
25.  $Us(t)$  is ahead  $Truth(t)$ : If the participant is ahead, he should wait for the prediction of  $Truth(t)$  to catch up and then follow a linear path to the next predicted value.



26. Steps involved in the VWAP Strategy: Following steps are needed to implement this strategy:
- A prediction of the current cumulative volume percentage
  - A prediction of the next cumulative volume percentage
27. Use of 20-day Rolling Average: Clearly, the typical 20-day rolling averages could be used as a default for these, but the next section shows how to use the real-time data to improve the quality of these predictions.

## Predicting Cumulative Relative Volume with Real-time Data



1. Cumulative Volume up to time  $t$ : For a given symbol over s given trading day, let  $R(t)$  denote the cumulative relative volume traded up to time  $t$ . In other words

$$R(t) := \frac{\text{Volume traded from 9:30 AM up until time } t}{\text{Volume traded from 9:30 AM up through 4:10 PM}}$$

2. Incorporation of Post-trading Hours: To accommodate the fact that the closing auction sometimes occurs slightly after 4 PM, a cutoff time of 4:10 PM instead of 4 PM sharp is used. Letting  $V(t)$  denote the volume traded up to time  $t$ , so  $R(t)$  can be re-written as

$$R(t) := \frac{V(t)}{V(4:10 \text{ PM})}$$

Note that at time  $t$   $V(t)$  is known, but  $R(t)$  is not yet known.

3. Breakdown in 10-minute Intervals: The trading day is broken down in 10 minute intervals, so the focus is now on the discrete sequence of values:

$$R(9:30 \text{ AM}), R(9:40 \text{ AM}), R(9:50 \text{ AM}), \dots, R(3:50 \text{ PM}), R(4:00 \text{ PM}), R(4:10 \text{ AM})$$

4. Challenges Handling Open and Close: How to best handle open and close is a delicate issue. These should be treated individually and mixed in with the rest of the trading day.
5. Handling Non-Open/Close Intervals: For now, with a slight abuse of notation,  $R(9:30 \text{ AM})$  is thought of as indicating the relative volume of the opening auction,  $R(4:00 \text{ PM})$  as indicating the relative volume of the opening auction and regular trading day exclusive of the closing auction, and  $R(4:10 \text{ PM})$  as including the closing auction.
6. Discrete Time Estimation Sequence: Let  $t_1, t_2, \dots$  etc. denote the discrete sequence of times. At time  $t_i$ , the prediction is required for both  $R(t_i)$  and  $R(t_{i+1})$ . It is



emphasized that  $R(t_i)$  is not known at time  $t_i$ , as only the numerator of  $R(t_i)$  is known at time  $t_i$ , not the denominator.

7. Suite of Historical Information Series: The first step is to assemble some prices of historical information that may be highly relevant to predicting  $R(t_i)$  and  $R(t_{i+1})$  on a given day.
8. Relevant Rolling Averages:

$$AVG_R(t_i) := \text{Rolling 20 day average of } R(t_i)$$

$$AVG_R(t_{i+1}) := \text{Rolling 20 day average of } R(t_{i+1})$$

$$ADV = AVG_V(4:10 PM) := \text{Rolling 20 day average of daily volume}$$

9. ADV Normalized Trading Volume: In order to start with a minimalist set of features, the values of  $ADV$  and  $V(t_i)$  are used to compute a simple feature with an intuitive interpretation:

$$A(t_i) := \frac{V(t_i)}{ADV}$$

10. Proxy for  $R(t_i)$  Series: The numerator  $V(t_i)$  here is the actual volume that has occurred. Dividing by the average daily volume makes this a rough indicator of how far the participant might be from where they are expected to be at this point.
11. Observations Grouped by  $AVG_R(t_i)$  and  $A(t_i)$ : The next step is to group the observations of the training data by their values of  $AVG_R(t_i)$  and  $A(t_i)$ .
12. Averaging  $R(t_i)$  across the Groups: The average value of  $R(t_i)$  is computed across all groups – weighting all observations equally with each symbol, but averaging across symbols weighted by notional value.
13. Generating  $R(t_{i+1})$  across the Groups: Analogously, the average value of  $R(t_{i+1})$  for groups is determined by their rounded values of  $AVG_R(t_{i+1})$  and  $A(t_i)$ .



14. The “Learned” Look-up Tables: This gives two look-up tables conditioned on the values of  $AVG_R(t_i)$  and  $A(t_i)$ , one may look up a prediction for  $R(t_i)$ , and conditioned on the values of  $AVG_R(t_{i+1})$  and  $A(t_i)$ , one can look up a prediction for  $R(t_{i+1})$ .
15. Eliminating Small Changes: Any variable combinations that have a sample size less than 100 are removed. In these cases, the 20-day averages are used as defaults.
16. Impact of using  $A(t_i)$  as a Conditional: By simple inspection of these tables, it can be seen that the predictions do vary considerably from the  $AVG_R(t_i)$  and  $AVG_R(t_{i+1})$  values as the  $A(t_i)$  variable is indeed meaningful for predicting  $R(t_i)$  and  $R(t_{i+1})$ .
17. Ease of Implementing the Algorithm: It may be noted that  $V(t_i)$ ’s,  $R(t_i)$ ’s, and the ADV are the only variables that need to be referenced for each symbol, which makes the resulting algorithm relatively easy to implement.

## Extension to Partial Trading Days

1. Trades after 9:30 AM / Before 4:00 PM: This section considers extending the reasoning above to cases where an intra-day order with a start time that is past 9:30 AM, and/or an end time after 4:00 PM.
2. X shares between 1 PM/2 PM: To make things concrete, say that the order arrives at 1 PM and should finish trading by 2 PM, and it is an order to buy  $X$  shares.
3. Volume Estimates for 1 PM/1:10 PM: At 1 PM when the order starts, based on historical as well as real-time data so far, the estimate for the daily traded volume exists for both 1 PM and 1:10 PM.
4. Estimate for 1 PM and 1:10 PM: To decide how much of  $X$  shares should be purchased between 1 PM and 1:10 PM, one more piece of information is needed: what percentage of the day’s volume does the participant have traded by 2 PM?
5. 0/1 Full Trading Day Volume Fraction: This reveals two new variables that were held constant when talking about full trading days: the percentage of volume traded



before the start was fixed at 0, and the percentage of volume traded by the end was fixed at 1.

6. Volume Fraction at other Times: While trading volume percentages are these particular times were held constant, estimated or percentages at other times could change and become as new data arrives.
7. Example of Inconsistency in Estimate: For example, it may be estimated that by 10 AM that 10% of the day's volume has already been traded by 10 AM. But by 11 AM, based on updated real-time data, the participant might estimate that 8% of the day's volume has traded by 11 AM.
8. Order Start/End Times Incorporation #1: This consideration was not a problem earlier, since the participant bases new decisions based on updated estimates and tries to catch up or slow down as indicated by the newer data.
9. Order Start/End Times Incorporation #2: However, here the order start and the end times need to be referenced throughout.
10. Example - Traded Volumes at different Times: If the participant thinks, for instance, that 20% of the day's volume traded before the start time, 60% of the day's volume will be traded before the end time, 30% of the day's volume has traded by the current moment, and that 40% will have traded by the next 10 minutes, and  $Y$  of the  $X$  shares has already been traded, what does the participant do?
11. Strategy based on Realized  $Y/X$ : First, the participant examines  $\frac{Y}{X}$  and compares it to

$$\frac{30 - 20}{60 - 20} = \frac{1}{4}$$

If  $\frac{Y}{X}$  is greater than this, he should slow down. Otherwise, he should catch up.

12. Calculating the corresponding Shares to Trade at Start: Assuming it is less, the participant immediately buys  $Z$  shares so that  $\frac{Y+Z}{X}$  is approximately  $\frac{1}{4}$ .
13. Number of Shares to Trade at End: Next, let  $Z'$  denote the number of shares such that  $\frac{Y+Z+Z'}{X}$  is approximately



$$\frac{40 - 20}{60 - 20} = \frac{1}{2}$$

Then  $Z'$  is how many shares the participant wants to buy spread out over the next 10 minutes.

14. Problem using the Earlier Approach: Employing the logic with the earlier approach runs into a problem. That approach described a way to make new estimates of  $R(t_i)$  and  $R(t_{i+1})$  at time  $t$  based on up-to-date data, but did not describe a way to also estimate  $R(start)$  and  $R(end)$ .
15. Estimate for a general  $R(t)$ : In fact, since the order start times and the order end times are arbitrary, one would need to have a general method for re-estimating the entire function  $R$  over all values of  $t$  at any point in time, based on the real-time data so far.
16. Estimation of  $R(start)$  and  $R(end)$ : Performing the general  $R(t)$  is a task for the future. For now, the estimates of  $R(start)$  and  $R(end)$  are fixed at the start of the order. These estimates come from the default 20 day average.
17. Scenarios where Regularity is Violated: Once these are fixed, later estimates for  $R(t)$  for times  $t$  between the start and the end may end up violating basic regularity conditions. For instance, there could be a situation where the estimate for  $R(t)$  is less than the fixed estimate for  $R(start)$ , or where the current estimate for  $R(t)$  is greater than the fixed estimate for  $R(end)$ .
18. Enforcing the Regularity Condition: To enforce regularity, any estimated value  $r$  for  $R(t)$  is replaced by the maximum of  $r$  and  $R(start)$ , or by the minimum of  $r$  and  $R(end)$  as necessary.
19. 20 day Averages of  $R(start)$  and  $R(end)$ : It is to be noted that

$$R(start) \leq R(end)$$

is guaranteed to hold when both values are estimated from the 20 day averages on the same data set.



20. Linear Interpolation across 10 minute Granularity: So far, this section has described how to derive estimates for  $R(t)$  at times  $t$  that fall on 10-minute boundaries, but an order might arrive at a time like 9:42 AM. For this, one can linearly interpolate the estimates between 9:40 AM and 9:50 AM to arrive at an estimate for 9:42 AM.
21. Reliance on Discrete Aggregation Levels: An improvement would be to make the method less reliant on discretized intervals of time and react to real-time data on a more streaming fashion.
22. Continuous Function over the Predictors: One natural way to do this is to fit a continuous function to the predictor/response data, rather than stopping at empirical aggregation.



1. Tight Skew:  $\alpha_r$
2. Loose Skew:  $\alpha_L$
3. Tight Width:  $\omega_r$
4. Loose Width:  $\omega_L$
5. Algorithmically generated Ideal Mid Cash Price:  $\Pi_{ideal}$
6. Position:  $P$  (expressed in cumulative net position per unit under consideration – firm/desk/trader)
7. Position Pivot:  $P_{pivot}$ . Dimensionless ontological view of the scaling position metric – roughly equivalent to the Reynolds' number of market making position units. Expressed in currency units.
8. Risk:  $R$  (expressed in cumulative net risk per unit under consideration – firm/desk/trader)
9. Risk Pivot:  $R_{pivot}$ . Dimensionless ontological view of the scaling risk metric – roughly equivalent to the Reynolds' number of market making risk units. Expressed in PV01 currency units.

## Framework Glossary

1. Equilibrium quantity: Quantity that only changes with the macro drivers/factors, and not the technical factors. Typically stable, but jumpy and undergoes changes when drivers shift – and introduces perturbations on the disequilibrium quantities.
2. Disequilibrium quantity: Quantity that changes with the technical, transient factors.

## Width/Skew/Size Estimation Models



1. Tight Models:

- Tight models estimate the market making quantities on a trader/firm/desk independent manner.
- They estimate the “secular” market making parameters – width, skew, and size for either the Market Making Outputs or the Axe Outputs – estimate them based on classes of input parameters.
- For each input parameter class, the following are needed:
  - a. A proxy that serves as a quantitative estimate of the desired parameter class.
  - b. Segmentation of the proxy over the sub-classified parameter set.

2. Input Class => Risk Profile:

- Captures all the cumulative risk components => the credit/solvency, market, and liquidity risk behind the issue.
- Proxy => CDS Spread, rating, bond basis
- Sub-classification => Issue, issuer, and sector.

3. Input Class => Liquidity:

- Captures the frequency and volume of the trade flow of a given issue, and the ease of getting in and getting out at the given side.
- Proxy:
  - i. Aggregated periodic (e.g., daily) volume for each side (buy/sell).
  - ii. Aggregated periodic (e.g., daily) notional for each side (buy/sell).
- Sub-classification => Issue, issuer, sector, and the instrument universe.

4. Firm/Desk/Trader level parameters: These provide aggregated controls for trading.

- Net Position => vital metric for inventory control.
- Risk limits => to control/manage exposure to specific granules – issue, issuer, tenor, sector, unit etc.

5. Monitor Mobility: Certain measures such as PV01 based risk, inventory, etc. are more easily human-monitored, so they are done daily. Others (such as tenor 01s) are less easily monitored, so they are done infrequently.



## Market Making System SKU

1. Intra day Curve Generation Scheme
2. Mid Price Estimation Models
  - i. Accommodate different mid price estimation models, and their respective parameters
3. Algorithmic Quote Construction => used for generating venue/ECN independent width/skew/size [composed of tight/loose components]. Broadly speaking achieves the following:
  - i. Specific parameters to control skew for targeted alpha generation strategies
  - ii. Accommodate different width and size estimation models, and their respective parameters
  - iii. Venue-independent base quote synthesis/construction
  - iv. Circuit breaker heuristics
  - v. Policy driven/policy enforcement/policy control applied at this level
4. Quote Management: Publishing/tailoring the constructed quote towards specific venues (possibly with order routing applied at this stage).
  - i. Venue specific rules (and thereby external vendor incorporations, like Broadway etc. at this stage).

## Market Making Parameter Types

1. Model Parameters: Parameters for generation of algorithmic generation of width, skew, and size.
2. Quote Generation Control Parameters
3. Quote Heuristics Control



#### 4. Quote Management Control

### Intra-day Pricing Curve Generation Schemes

1. Issue Benchmark Bonds: The following set of threshold criteria are used to determine the issuer specific benchmark bonds:

- ii. Threshold of daily TRACE volume/number of trades
- iii. Threshold of outstanding notional
- iv. Only senior obligations
- v. Some combination of the following threshold of the ratios:
  - $\frac{\text{CUMULATIVE DAILY ISSUE TRACE VOLUME}}{\text{OUTSTANDING NOTIONAL}}$
  - $\frac{\text{CUMULATIVE DAILY ISSUE TRACE VOLUME}}{\text{CUMULATIVE DAILY ISSUE TRACER VOLUME}}$

- 2. Benchmark bonds basis tracking: Track the bid side and ask side credit basis of the benchmark bonds from each TRACE print, using EMA VWAP/TWAP from the intra-day rates/credit curves. This will be the attempt to estimate the mid credit basis for the, and it is generally well behaved.
- Need to find a way to accommodate the institutional closing CDS mid marks and the benchmark bonds into the credit curve construction – these are highly valid points.
- 3. Liquid vs. illiquid: Typical liquid securities' quote may be proxied out of print (or at least EMA'd). Intra-day quote generation, however, is materially important for illiquid securities.
- 4. Intra-day credit curve generation inputs: Need a way to generate the credit curve from
  - i. The CDS marks
  - ii. The basis-adjusted benchmark bonds
  - iii. It always needs to be used in conjunction with tension splines.



- iv. Also need intra-day TRACE series to update the basis (direct or EMA) – will use this to establish the intra-day relationship between the CDS nodes and the TRACE cut-off threshold).
- 5. Intra-day credit curve updating:
  - a. Use the relationship grid between CDS 5Y, the off-tenors, and the benchmark bonds
  - b. Any change in any of them automatically re-adjusts using the set relationships.
  - c. CDS Curves are trader set; bond basis are EMA'd from the TRACE series using the prior credit curve
  - d. Relationships are either reviewed daily EOD
- 6. Live updating of bond prices: Use the live curve (either pure CDS, or a mixture of CDS/bond instruments) to extract the basis of each print, and then EMA that to generate the bond live prices.

## Mid-Price Models

- i. Definition: Computed theoretical mid-price, as to where the next print should be – assuming zero transaction costs, zero position/risk constraints, and infinite liquidity. Mid -Price is an ***Equilibrium Quantity***.
- ii. Estimation parameters: Typical mid price estimation parameters are: the IR curve, the survival curve, and the recovery curve. The other possible drivers are: funding curve – typically for long position, and repo curve – typically for shorts.

## Width Models



1. Tight Width: Computed theoretical width, after accounting for the issue liquidity and the issue riskiness. Tight width is the first in the set of disequilibrium quantities. Tight width is:
  - a. Proportional to issue risk (combination of credit and market risk – not counter party risk).
  - b. Inversely proportional to liquidity

## Skew Models

1. Tight Skew: This measure how far the last print has been OFF from the theoretical mid-price. Thus, Tight Skew is representative of the alpha potential – for a theoretical mid-price that chases the print in a sequence, the tight skew is zero.
2. Tight Bid Skew and Tight Ask Skew: This is an alternative SKU – instead of tight width and tight skew cognitive view, tight bid/ask skew parameters are determined only from their corresponding liquidity and flow metrics (i.e., bid/ask liquidity metrics).
3. Loose Skew: Simply put, loose skew is:

$$\alpha_L = \max\left(\frac{P}{P_{Pivot}}, \frac{R}{R_{Pivot}}\right)$$

4. Heuristic Checks on Loose Skew: Following checks applied to round out quoting:
  1. Ceiling/floor applied
  2. Maximum cutoff for width
  3. Best right skew – bid becomes ask.
  4. Best left skew – ask becomes bid.

## Size Models



1. Tight bid size/tight ask size: Basically, tight size is inversely proportional to tight width, to within normalized bounds.

## Heuristics Control

1. Can Buy/Can Short: Can Buy/Can Short => whether the bid/ask stays within the LONGABLE/SHORTABLE cutoff.
2. ECN Threshold Cross: Check to see if there is a cross between the published bid/ask and a given ECN's bid/ask.

## Published Market Quote Picture

1. Bid/Ask Sizes: Truncated to their appropriate rounding.
2. Bid Price:  $\Pi_{Ideal} - \frac{1}{2}\omega_L\alpha_L$
3. Ask Price:  $\Pi_{Ideal} + \frac{1}{2}\omega_L\alpha_L$
4. Bid/Ask Prices rounded downward/upward to their appropriate increments.

## Flow Analysis

1. Dimensionless flow classifier: If the metric (ADV etc.) is greater than a specific threshold, then the flow becomes "turbulent", else it is "laminar".
2. Flow Potential: Skew of all kinds is related to the flow driver/equilibration strength.



## Corporate Bond Auto-Responder (CBAR)

### Summary

1. Focus of the CBAR Algorithm: The aim of this chapter is to describe the auto-response mechanism for incoming client RFQ's for corporate bond trades less than or equal to a threshold size. A number of checks are made, e.g., risk, price, macro, before the trade is accepted by the algorithm.
2. Corporate Bond Desk Trade Flow: Typically, a Corporate Bond Desk receives on average over 10,000 client RFQ's daily, many of which are less than 100K in size. Automating the flow for smaller sizes allows traders to focus on larger, higher-value tickets.

### Reference Data

1. Instruments Traded: Secured Debentures.
2. Markets/Trade Venues: Market Axess (MKA), Bloomberg (BOLT.BBG), Trade Web (TWB), Merrill Lynch Bond Markets (MLBM), ICE Bond Point (KBP), Trade Web Direct (TWD), The Muni Center (TMC), MTS Bond Pro, TruMid.

### Flow Diagram - Client RFQ Flow

1. Primary Workflow of the Algorithm: The primary workflow that the algorithm handles is responding to client inquiries – RFQ's. The path of the ticket through the credit e-trading systems is described below.



2. Customer RFQ through the ECN: Customer sends in RFQ through external ECN, e.g., Market Axess.
3. TEX Routing of the RFQ's: TEX – the Trade execution – listens for the RFQ's and routes them to the auto-responder and the traders. A waterfall logic determines whether and how the algorithm responds to RFQ's with line traders having the first refusal on whether to respond. The waterfall logic is described later.
4. Autoresponder determines Eligibility of RFQ: The Auto-responder listens to all the RFQ's and performs numerous control checks when an RFQ is received to determine auto-response eligibility. It repeats the checks with 20 seconds – configurable – left in the quoting window. Details of the checks can be found in the later sections. If and only if all checks pass, the auto-responder sends back a level that was computed by the algo pricer.
5. Trader's Ability to Overwrite RFQ: The trader can see all RFQ's and has the ability to overwrite any level before the quoting window closes.
6. Time for Client to Accept: The client has a set amount of time to accept.
7. STP Based Standard Trade Booking: If DONE, the trade is booked through standard STP.
8. High Level Architecture Diagram of CBAR:
9. Waterfall Logic Determining Algo Response:

## Streaming Flow

1. Streaming Algorithmic Prices to Venues: The algorithmic prices are streamed to various ECN's that support the *click-to-trade* functionality, e.g., BOLT, KBP, etc. The workflow is described below and is illustrated as follows.
2. High-Level Algo Streaming Architecture: The streamer has waterfall logic that decides which of the internal levels – algo, portfolio, or institutional – to publish to ECN's.



3. Streamer Control Checks on Algo: The streamer listens to all algo price updates and performs numerous control checks before it sends the price to TEX.
4. TEX Waterfall for Outgoing Levels: TEX utilizes a waterfall to determine whether algo, institutional, or portfolio trader levels go out to the external trading platforms. Sometimes, dealers show just a single level – bid and offer – for bonds on all ECN's.
5. Auto-price Checks on Trader Acceptance: If a trader clicks the price on the external platform within the on-the-wire time, TEX sends the request to trade to the auto-execution module where numerous checks are performed. Details of the check can be found in a later section.
6. STP on Passing Control Checks: If all control checks – including checks in common with the RFQ workflow – pass, the trade is accepted and booked through standard STP. If any control fails, the acceptance decision is passed to a manual trader.

## Algorithm Operational Logic Description

1. Workflows of the Auto-responder Module: The corporate bond auto-responder has two main workflows.
2. Algo Levels for Client Inquiry: An RFQ auto-responding component that responds to the client inquiry with algorithmically generated levels.
3. Broadcast Algo Levels by Streamer: A streaming component that broadcasts algorithmically generated or manually derived levels to external venues. These levels can be automatically executed.
4. High Level Algo Components Architecture: The figure below illustrates the high-level architecture of the algo components and their connectivity to the wider credit e-trading platform.
5. Schematics of Typical Credit Algo Platform: All real-time components of the algorithm that contain the business logic are part of a typical Credit Algo Platform and are illustrated above. The only exception would be the streamer waterfall logic that is part of the TEX platform.



6. Interfacing with External Venues: Connectivity to external markets, normalization of messaging, and delivery of pricing and control information is part of the TEX platform. The configuration and control of the algorithm is provided through a number of user interfaces, listed in the Table below.
7. Interfaces that Control the Algorithm:

Interface	Type	Description
Parameter Controller	Config and Control	<ul style="list-style-type: none"><li>• Desk-level Control and Enablement</li><li>• Algo Global Parameter Control</li></ul>
Trader Work Queue	Control	Trader Portfolio-level On/Off Switch
Bond Switch	Control	Bond Level On/Off Switch
Bond Controller	Config	Algo Bond-level Parameter Control

## Algorithm Processing Logic Operating Detail – RFQ Quoting

1. Algo Response to Trading Strategies: The high-level workflow for client RFQ quoting is described earlier. The algo responds on behalf of the following three separate *trading strategies*.
2. Line/Flow Traders: For bonds enabled, the algo will respond with the line trader algo price if all checks are passed – and the trade is executed – the risk will flow into the bond owner’s book.
3. Portfolio Trader: For bonds that are enabled and are not responded by the line-trader algo, the client RFQ’s will be answered with the portfolio algo price. Executed trades will be routed to the portfolio trading book.
4. Algo Trader: All remaining trades will be routed to the automated market making algorithm. RFQ’s that pass the checks are returned to the client with the *pure* algo price, incorporating risk-management logic and trader configuration. All executed trades are routed to the algo book.



## Algorithm Processing Logic Operating Detail - Streaming

1. Role of the Streaming Component: The streaming component of the algorithm allows for the bid and the ask levels to be sent to the various ECN's, e.g., KBP, MLBM, Bloomberg BOLT, and TruMid. Internally, there are a number of possible levels that could be sent to the markets from both algorithmically and trader marked levels.
2. Waterfall Logic for Choosing Levels: The streamer contains waterfall logic to choose a single level for each side based on business preference, for example:
  - a. Institutional Trader Axe
  - b. Algo Risk Reducing Levels
  - c. Gold Levels
3. RFT's that are Auto-executable: Client RFT's based on either axed or algo levels can be auto-executed, whereas those initiated when *gold levels* – which are just the standard marked levels for gold clients – are subject to manual execution.
4. Checks to Ensure Cross-Consistency: Further rules are applied to ensure consistency between the internal levels:
  - a. If two sides are streamed, they must not cross each other;
  - b. If the algo crosses the axe or the gold level, do not send out the algo;
  - c. If gold crosses with the axe level, do not send out gold.
5. Streaming On/Off Market Checks: The streamed algo levels go through the following workflow to ensure that they are on-market and not stale.
6. Standard Checks on Algo Levels: These are performed in accordance with the criteria described for level validation for bond/algo prices/spreads below.
7. Outgoing Price Streamed Checks: Outgoing prices must pass streamed price checks, as shown in the additional level validation sequence presented later.
8. Perform E-trading and Macro Checks.
9. Pulling Back Bond Streaming Prices: Pull back the streaming prices for a bond when a price is pulled back or if a check will fail, even if there is no new price, e.g., just surpassed the DONE only limit, including 10 minute bond notional/DV01 limit or macro checks.



10. Limiting/Controlling Streamed Sizes: Limit streamed size so that DONE + stream size does not exceed DONE sizes for Notional/DV01 check at the CUSIP level. Stream sizes of inventory on algo topic – initially make sure it is the same direction and its absolute value is less than or equal to the auto-responder's estimate of position plus DONE but not booked.
11. Check Levels of Treasury Yields: Check Treasury Desk Yields against second source as is done for RFQ's, i.e., the level and the bid/offer spread.
12. Maximum Stream Size Per Bond: 500K initially.
13. Streaming under Risk Accepting RFQ: Streaming is turned on using the risk-accepting RFQ setting.
14. Check Timestamps on Prices Published: 10 minutes.
15. Latency Check for Stream Validation: A latency check of 2 seconds for validating streamed prices.
16. Checks for Streaming Heartbeat Signal: Check streaming heartbeat signal is within last 6 seconds.

## Auto-Execution of Streamed Levels

1. Checks for Auto-executed Streamed Levels: When a client clicks on a streamed level on the ECN, a *Request-for-Trade* is sent to the dealer – this is technically equivalent to an RFQ, but sent to the dealer only and where the level is already set. Assuming this is eligible for auto-execution, the usual RFQ workflow and checks are performed along with the following additions.
2. Check Trade Size versus Streamed Size.
3. Compare Price Sent out versus Clicked: Compare sent price against clicked on price and make sure they match. If there is a more recent price, make sure

$$\text{Absolute Value of the PnL} = \text{abs}(\text{Price Difference} \times \text{Notional}) < \$500$$



4. Price Currentness Check for Axes: For axes, check that the price is within last 30 minutes and ensure price does not go past algo mid. Perform algo level/price and treasury yield tests to validate mid.
5. Auto-execution 5-minute Limit: Auto-execution 5-minute limit of size sent out per CUSIP.
6. Reversion to a Manual Flow: If any of these checks fail, the ticket then reverts to a manual flow.

## Inputs

1. Message Input into Auto-responder: All the ticket information relevant to an RFQ – which includes the RFT type received when answering inquiry on streamed levels -is communicated via an RPC call to the auto-responder service. All other necessary inputs are provided to the algorithm via messaging. These are listed below.
2. List of Messaging Based Inputs into the Algorithm:
  - a. DV01
  - b. Third Party Levels (CBBT, BVAL, etc.)
  - c. ECN Levels
  - d. TRACE Prints
  - e. Treasury Levels
  - f. Trader Levels/Runs
  - g. Algo Levels
  - h. Bond Reference Data
  - i. Bond Position
  - j. STW Order
  - k. TWQ Config
  - l. Credit Macro Signals
  - m. Equity Macro Signals
  - n. TWQ Heartbeat



- o. E-trading User Control
- p. ECN Heartbeat
- q. Bond Auto-quote State
- r. Bond Control Market
- s. Accumulation Instruction

## Outputs

1. Quote-Response from Auto-Responder: The quote response is returned to TEX via an RPC call. The response includes the high-level decision on whether to quote or not as well as detailed information on the decision-making process.
2. Book Picked from Waterfall Rules: The algo can accept risk or reduce risk for a number of different trading books – as defined in the waterfall rules – so the final decision on which book to use for the RFQ is passed back to TEX to allow them to STP any done trades accordingly.
3. Outputs from Auto-responder and Streamer: In addition to the quote response, the messaging framework is used to broadcast outputs of both the auto-response and the streaming components.
4. List of Messaged Outputs from the Algorithm:
  - a. Algo Levels
  - b. Streaming Level Output
  - c. Streaming BAC Level Streaming Heartbeats

## Benchmarks

IG bonds are priced in spread-to-benchmark terms which means that accurate and timely US Treasury Levels are needed. The standard on-the-run treasuries are used in the valuation.



## Market Phases

Corporate bonds are traded in an OTC market; therefore, there are no official market phases. The algorithm is operational between 8:30 and 16:15 EST on US trading days for both the IG and the HY trading desks.

## Algorithm Operating Constraints

1. Enablement Needed for CUSIP Trading: The algorithm is used for bonds owned by the trader. The trader has to be enabled using the Parameter Controller interface. In addition, the trader must have the offerings enabled on the ECN's. The following checks are required to pass before a bond can be considered eligible for auto-response.
2. Market and Owner Group On: Checks to see if the Market and the Owner group is switched on using the Parameter controller interfaces, which also acts as a kill switch. An owner group can be turned on risk-reducing, axe-matching, and/or auto-trading.
3. Control - Trader Enabled: Checks to see if the current owner of the bond is switched on.
4. Control - Trader Logged In: Checks to see if the current owner of the bond is logged in.
5. Control - Bond Enabled: Checks to see if the bond is enabled for auto-response and/or for streaming.
6. Control - Ticket Type: Checks to see if the trade is a CLIENT trade.
7. Control - Restricted: Checks to make sure the bond is not restricted.
8. Control - TEX Heartbeat: Checks that the TEX engine for that market is heart-beating.



9. Control - Protocol: Checks that the PFQ used the BIN protocol – this control is now relaxed so that non-BIN protocol RFQ's will be in scope.

## Model Use Constraints

1. Models the Algorithm Depends on: The algorithm is not a model, but does rely on a number of models for live price and risk calculation. These are:
  2. Yield to Maturity Model: Used for DV01 calculation. DV01 calculation in itself does not typically happen within the algorithm. Pre-calculated DV01 values can be fed into the algorithm from an upstream system fast pricer, within which the model is implemented. There is no ETF NAV related calculation that needs to happen within the algorithm either. This applies to both the Auto-responder and the Auto-pricer modules.
  3. Signal Mixer for Corporate Bonds: This is used to generate algorithmic levels for bonds as well as for hit-rate calculations. The RFM model used for generating the algorithmic levels is implemented in the *auto pricer*.
  4. State Space Model for Market Prints: The SMAP model is used to produce a clean estimate of a bond's mid-price and bid/ask spread from TRACE market prints. SMAP levels are used as part of the *auto pricer*.

## Operational Risk

1. Trades at Off-Market Levels: The key risk associated with the algorithm is that it trades at off-market levels resulting in financial loss. Off-market levels could occur due to a breakdown in the algorithmic pricing engine or due to stale pricing.
2. On-Market Level Validation Checks: The risks are mitigated by comparing algo levels with reference levels from a number of other sources. The validation checks



performed on RFQ responses are listed in the sections below for spread/yield and price marked bonds.

3. Verifying Levels sent to ECN: The outgoing streamed prices are also checked to make sure that the levels sent to the ECN are not off-market. These checks are listed in the following section.

## **Level Validation Checks for Algo Levels on Bonds Marked in Spread or Yield**

1. Notes on Level Offset Criterion: Positive means more aggressive.

$$\text{Bid Spread} \geq \text{Reference Spread} - \text{Level Offset}$$

$$\text{Ask Spread} \leq \text{Reference Spread} + \text{Level Offset}$$

2. Notes on TRADE Level Check: Last same day TRACE print with size  $\geq 0.5\text{ MM}$ , 10 bp from same side, 5 bp from opposite side, 7.5 bp from dealer-to-dealer.
3. Notes on Level-Type Check: All benchmarks must match; spreads should be from the same source when possible, e.g., IBX, CBBT, TRACE.
4. Notes on Field Requirement Check: Only 2 of the mandatory Credit Level Checks need to be satisfied.
5. Reference Source - RUN:

Reference Side	Mid
Level Type (see Notes above)	Spread
Level Offset (see Notes above)	2.0
Offset Units	Basis Points
Valid Time	4 Business Days
Required?	Mandatory (see Notes above)



6. Reference Source – Trader Mark:

Reference Side	N/A
Level Type (see Notes above)	Spread
Level Offset (see Notes above)	2.0
Offset Units	Basis Points
Max Bid-Offer	15.0
Valid Time	4 Business Days
Required?	When available

7. Reference Source – IBX:

Reference Side	Mid
Level Type (see Notes above)	Spread
Level Offset (see Notes above)	2.0
Offset Units	Basis Points
Max Bid-Offer	15.0
Valid Time	2.50 hours
Required?	Mandatory (see Notes above)

8. Reference Source – IDC:

Reference Side	Mid
Level Type (see Notes above)	Spread
Level Offset (see Notes above)	2.0
Offset Units	Basis Points
Max Bid-Offer	15.0
Valid Time	2.50 hours
Required?	Mandatory (see Notes above)

9. Reference Source – CBBT:

Reference Side	Mid
----------------	-----



Level Type (see Notes above)	Spread
Level Offset (see Notes above)	4.0
Offset Units	Basis Points
Max Bid-Offer	15.0
Valid Time	5 minutes
Max Latency	30 seconds
Required?	When available

10. Reference Source – ECN (Top of Book 100K):

Reference Side	Mid
Level Type (see Notes above)	Spread
Level Offset (see Notes above)	2.0
Offset Units	Basis Points
Max Bid-Offer	15.0
Valid Time	15 minutes
Max Latency	30 seconds
Required?	When available

11. Reference Source – TRACE:

Reference Side	See Notes above
Level Type (see Notes above)	Spread
Level Offset (see Notes above)	See Notes above
Offset Units	Basis Points
Max Bid-Offer	See Notes above
Valid Time	Today
Required?	When available

12. Reference Source – Treasury Benchmark:

Reference Side	N/A
----------------	-----



Level Type (see Notes above)	N/A
Level Offset (see Notes above)	N/A
Offset Units	N/A
Max Bid-Offer	N/A
Valid Time	10 minutes
Max Latency	30 seconds
Required?	Mandatory

## Level Validation Checks on Algo Levels for Bonds Marked in Price

1. Notes on Level Offset Criteria: Positive means aggressive:

$$\text{Bid Price} \leq \text{Reference Price} + \text{Level Offset}$$

$$\text{Ask Price} \geq \text{Reference Price} - \text{Level Offset}$$

2. Notes on TRACE Print Criteria: Last Same Day TRACE Print – Size  $\leq 0.5MM$ ; USD 2 from same side, USD 1 from opposite side; USD 1.5 dealer-to-dealer.
3. Reference Source – RUN:

Reference Side	Mid
Level Type	Price
Level Offset (see Notes above)	0.15
Offset Units	Dollars
Max Bid-Offer	2.5
Valid Time	Today
Required?	Mandatory

4. Reference Source – Trader Mark:



Reference Side	N/A
Level Type	Price
Level Offset (see Notes above)	2.0
Offset Units	Dollars
Max Bid-Offer	N/A
Valid Time	4 Business Days
Required?	Mandatory

5. Reference Source – IBX:

Reference Side	Mid
Level Type	Price
Level Offset (see Notes above)	0.30
Offset Units	Dollars
Max Bid-Offer	2.50
Valid Time	2.50 hours
Required?	Mandatory

6. Reference Source – CBBT:

Reference Side	Mid
Level Type	Price
Level Offset (see Notes above)	0.60
Offset Units	Dollars
Max Bid-Offer	2.50
Valid Time	Today
Max Latency	30 seconds
Required?	When available

7. Reference Source – ECN (Top of Book 100K):

Reference Side	Mid
----------------	-----



Level Type	Price
Level Offset (see Notes above)	0.15
Offset Units	Dollars
Max Bid-Offer	2.50
Valid Time	Today
Max Latency	30 seconds
Required?	When available

8. Reference Source – TRACE:

Reference Side	See Notes above
Level Type	Price
Level Offset	See Notes above
Offset Units	Dollars
Max Bid-Offer	See Notes above
Valid Time	Today
Required?	When available

## Level Validation Checks for Outgoing Streamed Prices

1. Note #1 on Level Offset:

$$PV\ Offset = \min(DV01 \times LevelOffset / 100.0, 0.75)$$

2. Note #2 on Level Offset:

$$Price\ To\ Check = Price + DV01 \times (TSY\ Yield\ Quote\ \% - TSY\ Yield\ Now\ \%)$$

3. Note on the Required Check: Two of Three must pass.



4. Note #3 on Level Offset: In addition to the checks of Treasury Desk Yield vs. CBBT Yield, the Streamer Treasury Yield is checked against the Treasury Desk Yield.
5. Reference Source – ECN (Top of the Book 100K):

Reference Side	Mid
Level Type	Price
Level Offset	4 (See Note #1 on Level Offset above)
Offset Units	Dollars
Max Bid-Offer	2.0
Max Latency	30 seconds
Required?	When available

6. Reference Source – CBBT:

Reference Side	Mid
Level Type	Price
Level Offset	6 (See Note #1 on Level Offset above)
Offset Units	Dollars
Max Bid-Offer	2.0
Max Latency	30 seconds
Required?	When available

7. Reference Source – BMRK:

Reference Side	Mid
Level Type	Price
Level Offset	6 (See Note #1 on Level Offset above)
Offset Units	Dollars
Max Bid-Offer	2.0
Max Latency	30 seconds
Required?	Mandatory (see Notes above)

8. Reference Source – IDC:



Reference Side	Mid
Level Type	Price
Level Offset	4 (See Note #2 on Level Offset above)
Offset Units	Dollars
Max Bid-Offer	2.0
Required?	Mandatory (see Notes above)

9. Reference Source – IBOXX:

Reference Side	Mid
Level Type	Price
Level Offset	4 (See Note #2 on Level Offset above)
Offset Units	Dollars
Max Bid-Offer	2.0
Required?	Mandatory (see Notes above)

10. Reference Source – Treasury:

Reference Side	Mid
Level Type	Yield
Level Offset	1 (See Note #3 on Level Offset above)
Offset Units	Basis Points
Max Bid-Offer	2.0
Required?	Mandatory

## Reputational Risk

1. Main Source of Reputational Risk: The main source of reputational risk arises from the dealer not honoring prices streamed to ECNs. This can occur for RFT's when the last-look PnL check is triggered.



2. Streamed RFQ Level Response Monitoring: The following metrics are monitored to ensure that excessive rejects are not seen and that the checks are applied symmetrically. These are reported at regular governance forums.
3. Same Customer/Current Algo Level: RFT's that were accepted and where the level sent by the customer and the current algo level were the same.
4. Different Levels - Dealer Loss < USD 500: RFT's that were accepted and where the level sent by the customer and the current algo level were different and where the PnL represented a loss of less than USD 500 for the dealer.
5. Different Levels - Dealer Gain < USD 500: RFT's that were accepted and where the level sent by the customer and the current algo level were different and where the PnL represented a win of less than USD 500 for the dealer.
6. Customer Rejects - Dealer Gain > USD 500: RFT's that were rejected by the algo, i.e., went manual, because the level sent by the customer and the current algo level were different and the resulting PnL would have represented a win of > 500 USD for the dealer.
7. Customer Rejects - Dealer Loss > USD 500: RFT's that were rejected by the algo, i.e., went manual, because the level sent by the customer and the current algo level were different and the resulting PnL would have represented a loss of > 500 USD for the dealer.
8. Other RFT Rejects/Manual Switchover: RFT's that were rejected by the algo, i.e., went manual, for other reasons, e.g., a limit was breached.

## Market Risk

1. Sources of Algorithm Market Risk: Sources of market risk considered for the algorithm are the following.
2. Excessive Exposure to a Particular Bond or Ticker: These risks are mitigated by notional and position controls, as detailed in the next section.



3. Trading During Periods of Excessive Volatility: These risks are mitigated by checking that Equity, CDX, and Treasury moves have not exceeded a threshold.

## Sample Risk/Notional Controls for an IG Desk

1. Notes on the Listed Numbers: Percentages are relative to the desk/day headline in that row – for clarity, the calculated limit is in brackets.
2. Level – Desk; Time Period - Day:

Gross Notional (MM)	200
Net Notional (MM)	50
Gross DV01	USD 120K/bp
Net DV01	USD 30K/bp

3. Level – Desk; Time Period – 10 minutes:

Ticket Count	300
Gross Notional (MM)	15% (30)
Net Notional (MM)	20% (10)
Gross DV01 (USD 1K/bp)	10% (12)
Net DV01 (USD 1K/bp)	20% (6)

4. Level – Ticker:

Time Period	Day
Gross Notional (MM)	10% (20)
Net Notional (MM)	20% (10)
Gross DV01 (USD 1K/bp)	10% (12)
Net DV01 (USD 1K/bp)	20% (6)

5. Level – Bond:



Time Period	Day
Gross Notional (MM)	5% (10)
Net Notional (MM)	20% (10)
Gross DV01 (USD 1K/bp)	5% (6)
Net DV01 (USD 1K/bp)	20% (6)

## Typical Risk/Notional Controls for HY Desk

### 1. Level – Desk; Time Period - Day:

Gross Notional (MM)	112.5
Net Notional (MM)	58.5
Gross DV01	USD 58.5K/bp
Net DV01	USD 31.5K/bp

### 2. Level – Desk; Time Period – 10 minutes:

Ticket Count	113
Gross Notional (MM)	16% (18)
Net Notional (MM)	23% (13.5)
Gross DV01 (USD 1K/bp)	15% (9)
Net DV01 (USD 1K/bp)	29% (9)

### 3. Level – Ticker:

Time Period	Day
Gross Notional (MM)	12% (13.5)
Net Notional (MM)	23% (13.5)
Gross DV01 (USD 1K/bp)	15% (9)
Net DV01 (USD 1K/bp)	29% (9)

### 4. Level – Bond:



Time Period	Day
Gross Notional (MM)	12% (13.5)
Net Notional (MM)	23% (13.5)
Gross DV01 (USD 1K/bp)	15% (9)
Net DV01 (USD 1K/bp)	29% (9)

## Risk/Notional Controls

1. Algo Risk and Notional Controls: There are a number of risk and notional controls that prevent the algorithm from accumulating too much risk within a particular bond, ticker, or in total. These are measured over a single 10-minute period to mitigate against excessive client requests as well as over the whole trading day.
2. IG DV01 Normalized to LQD: For IG, the bond DV01's are normalized often to a major IG Bond ETF – LQD – by applying a *beta-adjustment* to the raw bond DV01. These betas are calculated by regression to historical spreads between the bonds and the ETF. The previous two sections show some typical control numbers.
3. Single Ticket Size Constraints: A typical maximum ticket notional handled by the algorithm is USD 3MM for the standard client RFQ flow. Streamed prices are subject to a USD 500K maximum, for instance.

## Volatility Controls

1. Checks for Market State Swings: To guard against wider market volatility, a number of checks are performed on credit, rates, and equity market benchmarks which act as indicators for wider market activity. The algorithm fails to respond if any of these conditions are violated.
2. CDX High-Low in the Past 15 Minutes: Less than 1 bp.
3. CDXIG Timestamp within Last 15 Minutes:



*Latency  $\leq$  30 seconds*

4. ESA High-Low in Past 15 Minutes: Less than 1%
5. ESA Timestamp within Last 5 Minutes:

*Latency  $\leq$  30 seconds*

6. 10-Year Treasury High-Low in Past 5 Minutes: Less than 5 bp
7. 10-Year Treasury Timestamp within Last 10 Minutes:

*Latency  $\leq$  30 seconds*



## Corporate Bond Skewer

### Major Components

1. Algo Pricer: An application that produces real-time algo levels on the data backbone.
2. Auto-responder: An application that listens for RFQ's from the Tech framework with the goal of auto-responding. It pulls levels from the data backbone, performs all control checks, and if all checks pass, responds automatically with a level when there is 20 seconds left in the BIN protocol.
3. Trader Work Queue: Application that displays RFQ's to the manual trader and enables the trader to submit levels.
4. TEX: An application that is an intermediary between the outside world and the trader. It maintains the state of the RFQ's, and interacts with manual traders via Trader Work Queue and the automated trader via the Auto-responder.
5. Streamer/Auto-execution Engine: Streamer is an application that pulls algo clean prices from the data backbone, performs all control checks, and sends it downstream to be streamed out the door for click to trade on ECNs, e.g., ALLQ, TW/BD, KBP, MLBM. The auto-execution engine performs a few checks after a client clicks on our level, e.g., ensures that the level has not changed significantly, before it accepts an automated trade.

### Maximum Auto-Responder Sizes

1. Maturity/Level/Beta DV01 Discriminator: RFQ Maximum Size = 3M for maturity <= 5Y, level <= 100 bp, BetaDV01 <= 2200
2. Beta High Water Mark Discriminant: RFQ Maximum Size = 2MM, if BetaDV01 <= 2200, 0 otherwise.



3. Maximum Streaming Size Discriminant: Maximum streaming size = 500K for algo generated streams.

## Waterfalls

1. Ax Request from Current Day: If a trader has an ax from today, the auto-responder responds with the trader's ax level if the ax size is greater than or equal to the RFQ size. A risk-accepting ax from today is first in the waterfall. A risk-accepting ax is ignored if it is not from today.
2. RFQ that is Risk-Reducing: If an RFQ is risk-reducing to an institutional trader, the institutional trader gets the first preference unless the RFQ is small, e.g.,  $\leq 100K$  for long end. Otherwise, the algo book gets preference if it is also risk-reducing to the algo book.
3. Level for Risk Reducing RFQ: If the auto-responder is quoting on behalf of the institutional trader for a risk-reducing RFQ, it will use the ax if it is from today. Otherwise, it will use the algo level.
4. Risk-reducing followed by Risk-accepting: The remaining waterfall in order of preference is risk-reducing for algo book, risk-reducing for portfolio book, risk-accepting for portfolio book if a set of bonds with their corresponding skews was uploaded by the portfolio trader that day, and then risk-accepting for algo book.
5. RFQs in Trader Work Queue:
  - a. Ax/Risk-reducing for institutional trader
  - b. Risk-reducing for algo or portfolio book
  - c. Risk-accepting for portfolio book
  - d. Risk-accepting for algo book
6. Overrides using Manual Trader Input: If a trader puts in a manual level, it will override all other levels and the trade will be booked into the trader's book.
7. Limits on the Algo Levels: Algo generated streams are only generated for sized up to 500K and for algo and portfolio book inventory.



8. Waterfall Sequence after the Ax: If a trader has an ax, the ax has preference over the algo on ALLQ. An ax can be auto-executed up to 500K if it passes a test against the algo mid. After the ax, the algo streams are next in the waterfall, followed by the trader's gold levels.

## Skewing Flow

1. Bond-level Pricing Set List: There are 5 different pricing sets for each bond. Each set has multiple client tiers. The five sets are as follows:
  - a. Institutional Trader/Portfolio Trader Risk-reducing
  - b. Portfolio Trader Risk-accepting
  - c. Algo Book Risk-reducing
  - d. Algo Book Risk-accepting
  - e. Algo Book Risk-accepting for small size
2. Small Size Risk-accepting Algo Book: The algo book risk-accepting levels for small sizes are the regular algo-book risk-accepting levels backed off a little. The reason for this initially was that when the platform first went live, RFQ's were the only mechanism to get rid of risk. Now that it streams to multiple ECNs and can exit small positions, this condition can be relaxed. This will improve the ticket count hit rate.
3. Signal Waterfall Order - ECN, CBBT, BMK: All of the signals are based upon an initial waterfall that consists of ECN, CBBT, and BMK. Since BMK is based purely on TRACE prints and not executable levels – CBBT is indirectly based on executable levels – it is not allowed to be better than the second most aggressive reference level. BMK is also further constrained by the Runs for the non-algo book levels.
4. Bid/Offer Spread Computation/Reduction: For ECN levels, the average bid and offer up to a threshold size is based upon the stack and the resulting bid/offer spread is reduced by a percentage, e.g., 30%.



5. Estimating the BO % Reduction: This percentage was set based upon looking back at historical RFQ's, backing out where they traded, and then looking at the resulting PnL profiles based upon reducing the ECN bid/offer spread by different percentages.
6. Separate BO for Different Pricing Sets: Separate percentages are used for the following three scenarios for different pricing sets.
7. Algo Book Risk-reducing: Risk-reducing for the algo books is the most aggressive as we are often trading near the mid.
8. Institutional/Portfolio Trader Risk-reducing: Risk-reducing for the institutional/portfolio trader is more about optimizing the PnL.
9. Algo Book Risk-accepting: Finally, risk-accepting for the algo book is the least aggressive and is about maximizing the PnL and constraining the balance sheet.
10. Adjusting Bid and Offer Levels: The bid and the offer levels are then adjusted in a number of different ways.
11. Improved Levels for Risk-reducing: First, for risk-reducing only, if there are more aggressive Run levels, the bid or the offer is improved up to a basis point if the Runs are recent.
12. Accuracy Impact of Incorporating Runs: The degree to which Runs are updated on a timely basis varies widely across the desk. Of course, Runs would be most accurate if there was a position in the CUSIP and hence limited this to risk-reducing RFQs.
13. Adjustments Based on Recent TRACE: The levels are then adjusted based upon TRACE. The levels are backed up in the direction of the last print if it is a 5 MM print or if the aggregate net TRACE size of the last 5 prints is greater than a threshold.
14. Constraints Placed on TRACE Prints: The level is also constrained if it is out of the range of recent TRACE prints or if it goes past the reference mids – Runs, IBOXX, IDC, BVAL – by more than a margin.
15. Non Algo-Book Risk-reducing: For all price sets other than algo-book risk reducing, if there is a Run from the last few days, the level will not go more than a threshold past the Runs mid. This may require a revision for the portfolio book as stale tuns can cause uncompetitive bids for certain bonds.



16. Idea Behind the RFM Signal: A second signal – RFM – is incorporated into pricing where the bid and the offer for each pricing source – e.g., TRACE, ECN, CBBT, IBOXX, IDC, BMRK, is reduced to minimize on average the difference between historical TRACE prints and the historical levels of a pricing source at the same point in time.
17. Proximity of Adjusted Levels to Median: The median value of these bids or offers is determined and a score is computed based upon how many of these adjusted pricing source levels are near the median.
18. Determining Hit-Rate Based Level: Based upon the historical variation of TRACE prints around this median signal, a level can be determined based on the current median level, the historical variability of the signal, and a desired hit rate. This signal is used to improve the first algo signal if there is a high score.
19. Algo Risk-accepting Liquidity Premium: For algo risk-accepting levels, a liquidity premium is added ranging from 0 to 20 bp.
20. Steering Close to Neutral Risk: The second component of pricing/skewing allows the risk to close to market neutral across the risk categories.
21. Skewing applied for Own Books: There is skewing logic only when the algo responds to its own books as it only considers risk in its own books.
22. Algo Book Risk Limit Table: A representative limit table for the algo book across different risk features is as follows:

	<b>Net (K/bp)</b>	<b>Gross (K/bp)</b>
ISIN	2.5	2.5
Ticker	3.0	6.0
Maturity Bucket per Sector	6.0	25.0
Sector	6.0	35.0
Maturity Bucket	15.0	NONE
All	25.0	NONE



23. Estimating Category-Based Risk-Fraction: The algo computes risk fractions for these risk factors. For example, if the sector risk is 3K/bp and the sector limit is 6 K/bp, then the category risk is 0.5, i.e., it is at 50% of the limit.
24. Category Risk Fraction Scaling Table: Each risk fraction is then scaled by the following factors for each bond:

<b>CUSIP</b>	1.0
<b>Ticker</b>	0.35
<b>Maturity Bucket per Sector</b>	0.075
<b>Sector</b>	0.075

25. Category Based Risk Fraction Impact: Clearly, the bond's own risk will be the largest risk factor in skewing the levels followed by the ticker, and then benchmark and sector.
26. Sum of Scaled Risk Fractions: The algo then sums up these scaled risk fractions. The sum is translated to a skew percentage using a piecewise linear skew table.
27. Weighted Sum vs Skew Table:

<b>Weighted Sum</b>	<b>Skew</b>
<= -0.67	-0.50
-0.40	-0.20
-0.20	-0.20
0.00	0.00
+0.20	+0.20
+0.40	+0.20
>= +0.67	+0.50

28. Adjustment for Risk-accepting RFQ: For a risk-accepting RFQ, the algo applies a small percentage improvement to the raw level, e.g., the ECN levels, to the base



signal methodology discussed above, improves upon the RFQ signal, and then applies this skew percentage to the resulting bid/offer spread to determine the skew.

29. Scenario-specific Logic Walk-through: The levels then go through a crawling logic which has a few scenarios. One such scenario sequence is laid down below.
30. Maximum Risk Fraction across Categories: The maximum risk fraction across all of the categories is above a threshold:
31. Minimum Risk Fraction across Categories: The minimum risk fraction exceeds another threshold:
32. Recent Hit-rate Threshold Breach: The hit-rate for the past hour and day is less than a threshold – say 10%:
33. CUSIP-Level Liquidity Fraction Check: The bond is liquid, i.e.

$$LiquidityFraction \geq 1$$

34. Cross Category Maximum Gross Risk: The maximum gross risk across categories is less than 0.75.
35. Actions on Scenario Criteria Match: Then: the offer side spread for that bond will be ramped up, i.e., cheaper, over the next hour by up to a percentage of the bid-offer spread.
36. Actions on Net Risk Reduction: If the algo reduces this net risk by more than a threshold, the crawler backs up and begins the cycle again, assuming all of the above conditions are still met.
37. Risk-reducing RFQ - Initial Stance: For risk-reducing RFQs, the algo starts with a more aggressive stance, improving the raw levels in the base methodology signal further.
38. Risk-reducing RFQ - Adjustment #1: It then improves the level on one side for positions that are aged more than 60 days by 15% of the bid/offer spread.
39. Risk-reducing RFQ - Adjustment #2: It then further improves the level if the maximum net risk fraction is greater than 0.33 and goes through different crawling strategies that are a function of risk across all of the categories and the hit rate.



40. Checks across the ECN Levels: The levels then go through reference checks – IBOXX, Runs, IDC, TRACE, and BVAL – as the level is not allowed to go past the second most aggressive mid by more than a threshold; however, risk-reducing is allowed to go up to 2 bp further through the mid if risk is built up on one side.
41. Final Checks for Risk-Control: The following are some final rules to control risk.
42. 80% Gross Risk Limit Threshold: If 80% of the gross limit is hit, the algo backs up all balance sheet increasing levels by 10 bp.
43. Threshold for Net Risk Limit: If 33% of the net limit is hit, it backs up 2 bp, and if 50% of the net limit is hit, it backs up 10 bp. These are clearly very constraining.
44. Impact of Imposing above Limits: The practical consequence of the above is that net limits are really half of the limits specified. So, once 10 K/bp is hit across all books, quoting is done purely for the sake of quoting.
45. Consequence of Beta-weighted Risk: These limits are further constrained by the fact that risk is beta-weighted, which prevents quoting larger sizes for long-dated, high-spread CUSIPs.
46. Migration to LQD-Based Betas: The betas used in the algo are really volatility adjustments that are dependent on the spread of the bond. A better idea would be to migrate to the LQD betas that also take into account correlation.



## High-Frequency Trading in a Limit Order Book

### Introduction

1. Role of a Securities Dealer: The role of a dealer in securities market is to provide liquidity on the exchange by quoting bid and ask prices at which he is willing to buy or sell a specific quantity of assets. Traditionally, this role has been filled by market maker or specialist firms (Avellaneda and Stoikov (2008)).
2. Expansion of the Dealer Role: In the recent years, with the growth of electronic exchanges such as NASDAQ's INET, anyone willing to submit limit orders in the system can effectively play the role of a dealer.
3. Transparency among Limit Order Books: Indeed, the availability of high frequency data in limit order book – see [www.inetats.com](http://www.inetats.com) – ensures a fair playing field where various agents can post limit orders at prices they choose.
4. Optimal Bid/Ask Submission Strategies: This chapter the optimal submission strategies of bid and ask orders in such a limit-order book.
5. Main Risks Faced by Dealers: The pricing strategies of dealers have been studied extensively in the microstructure literature. The two most often addressed sources of risk facing the dealers are: a) the inventory risk coming from uncertainty of the asset's values, and b) the asymmetric information risk arising from informed trades.
6. Focus of the Chapter - Inventory Risk: Useful surveys of their results can be found in O'Hara (1997), Stoll (2003), and Biais, Glosten, and Spatt (2005). This chapter will focus on the inventory risk.
7. Optimal Prices for a Monopolistic Dealer: In fact, the model presented here is closely related to that of Ho and Stoll (1981), which analyzes the optimal prices for a monopolistic dealer in a single stock.
8. Accounting for the Inventory Effect: Ho and Stoll (1981) specify a *true* price for the asset, and derive optimal bid and ask quotes around this price to account for the effect



of the inventory. This inventory effect was found to be significant in an empirical study of the AMEX options (Ho and Macris (1984)).

9. Problem of Dealers under Competition: In another work by Ho and Stoll (1980), the problem of dealers under competition is analyzed and the bid and the ask prices are shown to be related to the reservation – or indifference – prices of the agents.
10. True Price Given by Market: This framework assumes that the agent is but one player in the market, and that the *true*\_price is given by the mid-market price.
11. Arrival Rates of Buy/Sell Orders: Of crucial importance will be the arrival rates of the buy and the sell orders that will reach the agent. In order to model these arrival rates, results will be drawn from econophysics.
12. Statistical Properties of the LOB: One of the important achievements of the current literature has been to explain the statistical properties of the limit order book (Bouchaud, Mezard, and Potters (2003), Luckock (2003), Potters and Bouchaud (2003), and Smith, Farmer, Gillemot, and Krishnamurthy (2003)).
13. Modeling the Observed Statistics: The focus of these studies has been to reproduce the observed patterns in the markets by introducing *zero intelligence* agents, rather than modeling optimal strategies of rational agents.
14. Optimal Strategy without Utility: One possible exception is the work of Luckock (2003), who defines the notion of optimal strategies without resorting to utility functions.
15. Inferring Buy/Sell Arrival Rates: Though the objective of this chapter is different to that of the econophysics literature, results will be drawn on them to infer reasonable arrival rates of buy and sell orders.
16. Size Distribution and Price Impact: In particular, the results that will be most useful are the size distribution of market order (Maslow and Mills (2001), Weber and Rosenow (2005), Gabaix, Gopikrishnan, Plerou, and Stanley (2006)), and the temporary price impact of market orders (Bouchaud, Mezard, and Potters (2002), Weber and Rosenow (2005)).



17. Combined Utility and Microstructure Approach: The approach, therefore, is to combine the utility framework of Ho and Stoll approach with the microstructure of actual limit order books as described in the econophysics literature.
18. Two-step Procedure: The main result is that the optimal bid and ask quotes are derived in an intuitive two-step procedure.
19. Personal Inventory-Based Indifference Value: First, the dealer computes a personal indifference value for the stock given his current inventory.
20. Calibrating Bid/Ask Side Quotes: Second, the dealer calibrates his bid and ask quotes to the limit order book by considering the probability with which his quotes will be executed as a function of their distance from the mid-price.
21. Unifying Inventory and Market Effects: In the balancing act between the dealer's personal risk and the market environment lies the essence of the solution.
22. Main Building Blocks of the Model: The next section describes the main building blocks of the model: the dynamics of the mid-market price, the agent's utility objective, and the arrival rate of orders as a function of the distance to the mid-price.
23. Optimal Bid and Ask Quotes: The next section solves for the optimal bid and ask quotes, and relates them to the reservation price of the agent given his current inventory.
24. Numerical Solution and PnL Comparison: An approximate solution is then presented, and a numerical simulation of the agent's strategy is then compared with the PnL profile of that of a benchmark strategy.

## The Mid-Price Model of the Stock

1. Money Market with no Interest: For simplicity, it is assumed that the money market pays no interest.
2. Modeling the Mid-market Price: The mid-market price, or the mid-price, of the stock evolves according to



$$\Delta S_u = \sigma \Delta W_u$$

with initial value

$$S_t \equiv s$$

Here  $W_u$  is a standard one-dimensional Brownian motion and  $\sigma$  is constant.

3. Choice of Arithmetic Price Model: This model is chosen over the standard geometric Brownian motion to ensure that the utility functionals introduced remain bounded.
4. Evolution using Geometric Price Model: In practical applications, one could also use a dimensionless model such as

$$\frac{\Delta S_u}{S_u} = \sigma \Delta W_u$$

with the initial value

$$S_t = s$$

To avoid mathematical infinities, exponential utility functions could be modified to a standard mean/variance objective with the same Taylor-series expansion. The essence of the results would remain. More details regarding the model with mean/variance utility will be given in a subsequent section.

5. Assumptions Underlying the Evolution Scheme: Underlying this continuous-time model is the implicit assumption that our agent has no opinion on the drift or any auto-correlation structure for the stock.
6. Agent's Terminal Investment Horizon: This mid-price will be used solely to value the agent's assets at the end of the investment period.



7. Measuring the Impact of the Inventory: The dealer may not trade costlessly at this price, but this source of uncertainty allows measuring the risk of his inventory in stock. A later section will introduce the possibility of trading through limit orders.

## The Optimizing Agent with Finite Horizon

1. Focus of the Agent's Utility: The agent's objective is to maximize the expected exponential utility of his PnL profile at a terminal time  $T$ .
2. Choice of the Exponential Utility Function: This choice of convex risk measure is particularly convenient, since it helps define the reservation – or indifference – prices which are independent of the agent's wealth.

## References

- Avellaneda, M., and S. Stoikov (2008): High-frequency Trading in a Limit-order Book *Quantitative Finance* **8** (3) 217-224
- Biais, B., L. Glosten, and C. Spatt (2005): Market Micro-structure: A Survey of Micro-foundations, Empirical Results, and Policy Implications *Journal of Financial Markets* **8** (2) 217-264
- Bouchaud, J. P., M. Mezard, and M. Potters (2002): Statistical Properties of Stock Order Books: Empirical Results and Models *Quantitative Finance* **2** (4) 251-256
- Gabaix, X., P. Gopikrishnan, P. Plerou, and H. E. Stanley (2006): Institutional Investors and Stock Market Volatility *Quarterly Journal of Economics* **121** (2) 461-504
- Ho, T., and R. Macris (1984): Dealer Bid-ask Quotes and Transaction Prices: An Empirical Study of some AMEX Options *Journal of Finance* **39** (1) 23-45
- Ho, T., and H. R. Stoll (1980): On Dealer Markets under Competition *Journal of Finance* **35** (2) 259-267



- Ho, T., and H. R. Stoll (1981): Optimal Dealer Pricing under Transactions and Returns Uncertainty *Journal of Financial Economics* **9 (1)** 47-73
- Luckock, H. (2003): A Steady-state Model of the Continuous Double Auction *Quantitative Finance* **3 (5)** 385-404
- Maslow, S., and M. Mills (2001): Price Fluctuations from the Order-book Perspective: Empirical Facts and a Simple Model *Physics A* **299 (1-2)** 234-246
- O'Hara (1998): *Market Microstructure Theory* Wiley New York, NY
- Potters, M., and J. P. Bouchaud (2003): More Statistical Properties of Order Books and Price Impact *Physica A: Statistical Mechanics and Applications* **324 (1-2)** 133-140
- Smith, E., J. D. Farmer, L. Gillemot, and S. Krishnamurthy (2003): Statistical Theory of Continuous Double Auction **3 (6)** 481-514
- Stoll, H. R. (2003): Market Microstructure, in: *Handbook of the Economics of Finance* (editors: G. M. Constantinides, M. Harris, and R. M. Stulz) **North Holland** Amsterdam, Netherlands
- Weber, P., and B. Rosenow (2005): Order-book Approach to Price Impact *Quantitative Finance* **5 (4)** 357-364