Imperial College London

Department of Mathematics

Optimal Execution with Intraday Liquidity Changes and Transient Nonlinear Impact

Author: George Coxon (CID: 022741149)

A thesis submitted for the degree of

MSc in Mathematics and Finance, 2022-2023

Declaration

The work contained in this thesis is my own work unless otherwise stated.

Acknowledgements

First and foremost, I would like express my deepest gratitude to my academic supervisor Dr Ofelia Bonesini. Her generous investment of time and invaluable feedback were instrumental in shaping the quality and direction of this thesis.

I am also grateful to Fei Wang, Mark Richardson, and the whole Multi Strategy team at Janus Henderson Investors, for their support and for providing the opportunity to work on such an interesting problem.

I would like to thank my friend Louis for opening my eyes to the world of quantitative finance and encouraging me to apply to the MSc, and my girlfriend Agota for her unwavering support this last year. Finally, I would like to thank my Mum and Dad for their unconditional love and support. I would not have made it this far without them.

Abstract

This thesis delves into price impact modelling and optimal execution strategies within the framework of electronic financial markets, particularly focusing on the futures market where the evolution from physical trading to electronic platforms has transformed market dynamics. The Discrete Propagator Model (DPM) offers an alternative approach to traditional complex models, capturing essential market impact characteristics with the ability to calibrate through regression on trade data. This research extends the DPM to futures markets across various asset classes, revealing insights into intraday liquidity changes and impact concavity.

The study's core findings lie in optimising market impact parameters and exploring execution strategies for different financial instruments. The DPM's calibration unveils intraday liquidity shifts, cross-sectional impact differences, and an evolving Limit Order Book microstructure - advocating a static recalibration approach. Optimal execution strategies, considering intraday liquidity dynamics, highlight significant cost savings compared to industry benchmarks. Such execution strategies are particularly advantageous for mid-frequency institutions with a high portfolio turnover.

Contents

List of Figures

List of Tables

Introduction

Centralised exchanges have always played an important role in the trading of financial products, bringing together buyers and sellers in one location to trade. Historically, traders and brokers would physically gather and trade verbally on behalf of clients who would communicate orders to the brokers via telephone. This practice has all but died out in modern day trading, with much of the activity being moved to *electronic markets* where orders can be executed in a faster, more accurate and transparent way [2, 3]. One common form of electronic market is a Limit Order Book (LOB) [1] where market makers publicly post quotes of volume and price they would be willing to trade at. Market participants seeking to trade instantly match these orders, taking liquidity from the LOB and potentially affecting the best bid/offer price. With market orders impacting both price and liquidity, careful thought needs to be put into how to optimally execute large volumes, and this requires a comprehensive understanding of the dynamics between traded volume and market impact.

With the evolution of market microstructure came the development of price impact models. Beginning with simple permanent impact models [4], more features have been added over time including temporary impact [5] and transient impact [6]. Over the years, accessibility to market data and the capability to systematically analyse it has improved, leading to observed stylised facts of market impact [1, 7, 8, 9] such as intraday liquidity changes, impact propagation and highly concave instantaneous impact. Advances have been made using Order Flow Imbalance (OFI) to both explain and predict price changes to a high degree of accuracy [10, 11]. Whilst these models can be used by market makers and High Frequency Trading (HFT) firms to create sophisticated dynamic execution algorithms [12], they require multiple levels of order book data to calibrate and may not be useful for institutions without the data accessibility, technical ability and exchange connectivity. However, the use of regression to calibrate OFI models has been applied to the Discrete Propagator Model (DPM) which can capture the aforementioned stylised facts whilst requiring trade data only [13, 14]. We intend to further this research, with a particular focus on the futures market - versatile instruments which provide large amounts of liquidity over a range of asset classes, including Equity Index, Commodity, Currency and Interest Rate futures.

This thesis is structured as follows. In Chapter 1, we introduce the dynamics of the LOB and briefly cover the important history of price impact models and optimal execution plans. We recount some empirical observations found and introduce the DPM. In previous investigations into price impact modelling, many papers have studied single stocks, and there is little evidence to support the validity of such results in the futures market. In Chapter 2, we discuss the origin of futures contracts and explain the rationale behind investors trading them, providing the motivation for studying these instruments.

We introduce the data used for calibration in Chapter 3, and using tick level trade data from Tick Data we reconstruct order signs and aggregate the data. We outline the data transformations performed for model calibration and backtesting, and discuss the processes of creating synthetic normalised futures.

In Chapter 4, we explain in detail the process of calibrating the model, including the method to find an optimal scale (λ) and concavity (δ) of market impact. We find supporting evidence for the existence of intraday liquidity changes in the futures market for all asset classes, and provide novel investigation into the changes in liquidity and concavity throughout time. We evaluate the cross-sectional differences of impact parameters, both between asset classes and futures within the same asset class and show the model's ability to explain the effects of periods of market dysfunction such as the pandemic crash, and central bank rate hikes. We show however that the optimal calibration is rarely stable throughout time, and find the optimal frequency and training length to recalibrate

the model.

Using calibrated parameters, we turn to the problem of execution in Chapter 5 using the SLSQP solver to find optimal trading schedules. We study the effectiveness and drawbacks of this method, and discuss the optimal trading plan assuming constant liquidity under a wide range of settings. Finally, we calculate the optimal execution schedule when accounting for intraday liquidity changes, and find cost savings of over 80% compared to common industry benchmarks. We also quantify the savings obtained by including intraday liquidity compared to assuming a constant impact parameter, and find savings of up to 32% for some futures.

All code written for this thesis is made public on GitHub, including additional plots and calibration results.

Chapter 1

Literature Review

1.1 Electronic Markets and Limit Order Books

In modern day electronic markets, participants are generally categorised into two groups: *liquidity takers* and *liquidity providers*. Liquidity takers decide if and when they want to trade, and look for a counterparty who can provide the best price. On the other hand, liquidity providers provide quotes of prices that they would be willing to both buy and sell at. Most liquidity providers come in the form of *market makers*, market participants who stand ready to both buy and sell an instrument at any and all times, but at a price they set themselves [1, Chapter 1]. Liquidity takers are not guaranteed to find a market maker that is willing to trade with them at a price they are happy with, and similarly market makers may not find a counterparty that is willing to trade with them at their quoted prices.

In an electronic *Limit Order Book* (LOB), quotes are visible to all market participants at all times (as opposed to RFQ markets) allowing for higher transparency. Quotes are provided by submitting a *limit order* which contains the crucial information of order price, direction (buy/sell) and quantity. Figure 1.1 shows a visual representation of limit orders in a LOB, with bid (buy) quotes on the left and ask (sell) quotes on the right. The largest bid and smallest ask price are called the best bid and best ask price respectively, with the difference between the two being known as the *bid-ask spread* and the average known as the *mid-price*. The volume of a given order can only be in integer multiples of the *lot size*, and order prices must be in integer multiples of the *tick size*. One key way in which trades occur in a LOB exchange is when liquidity takers place *market orders*, orders of a specific volume and direction that are instructed to be matched with one or more limit orders immediately. Whilst there are a number of ways in which different types of trades can be matched, we will only focus on market orders being matched with limit orders in this thesis, but direct the reader to [1, Chapter 3] for a thorough explanation of LOB mechanics.

When a buy market order is submitted, a number of outcomes can occur depending on the order volume compared to that of the current limit orders:

- If the volume is smaller than that of the volume at the best ask, the order will be matched at the best ask price, reducing the volume at that level but leaving prices unchanged.
- If the volume is exactly equal to that of the best ask, the two orders will be fully matched at the best ask price, but leaving the best ask price higher after taking liquidity.
- If the volume of the market order is higher than the volume at the best ask price, the limit order at the best ask will be fully matched, and the remaining volume of the market order will be traded with the new best ask price. In this instance the order will be matched at an average price higher than the best ask.

Clearly, trading in the LOB can have a direct impact on price and for large orders this price difference could be material. The increase in price for a buy order will adversely affect the liquidity taker who is trying to buy for the lowest price possible, and so thought needs to be put into how and when the order is executed. Not only can a market order push the price up for future trades, but large enough orders can affect the average price of the current trade.

Studies have found that only a small fraction of the daily traded volume is actually quoted on a LOB, with Bouchaud & Bonart [1, Table 4.1] finding that within 1% of the mid-price, total volume

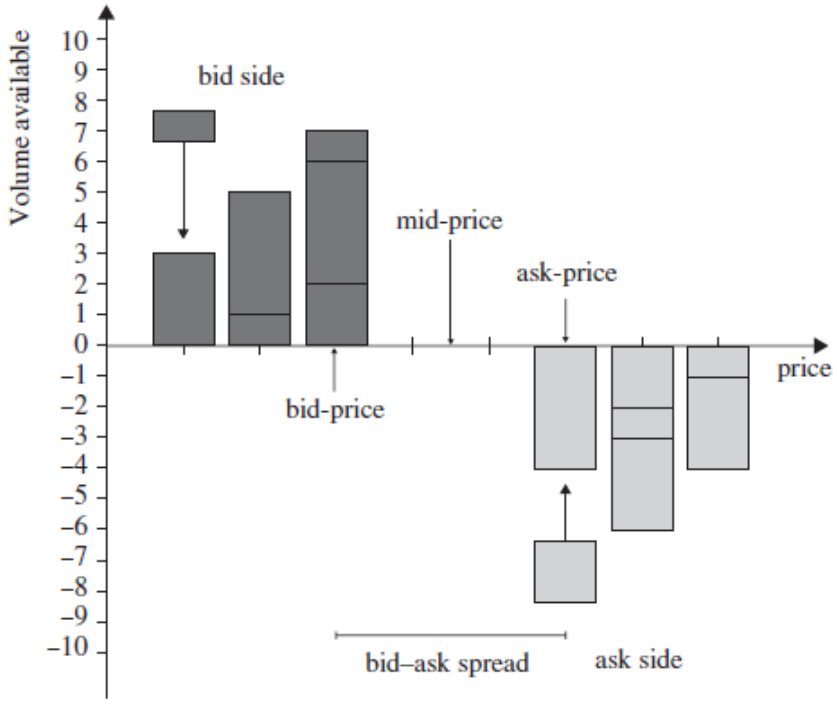


Figure 1.1: Schematic of a Limit Order Book [1, Figure 3.1].

quoted was between 1% and 3% of daily traded volume. Clearly, trading a large amount instantly would drain liquidity from the order book, and so large single orders called are often split up into many child orders, with whole parent order (or *Meta-Order*) being executed over a long period of time. To attempt to construct an optimal trading strategy, one needs to first quantitatively understand exactly how prices react to traded volume.

1.2 Definitions

Definition 1.2.1 (Trading Strategy)**.** [15, Section 2.3] A continuous trading strategy Π is a sequence of positions x_t where at time t a trader holds a position of size x_t in the underlying security, namely

$$\Pi := \{x_t\}_{t=0}^{T}. \tag{1.2.1}$$

The speed of trading at time t is $v_t = \frac{dx}{dt}$.

Definition 1.2.2 (Impact of trading)**.** At time t, the expected impact of executing Π is

$$I_t := \mathbb{E}\left[r_t \mid \Pi\right], \tag{1.2.2}$$

where r_t is some measure of price returns.

The definition of impact is somewhat loosely defined in the literature, with some studies measuring the quantity with absolute returns $S_t - S_0$, percentage returns $\frac{S_t - S_0}{S_0}$, or log returns $\ln(S_t) - \ln(S_0)$. Viewing impact as percentage or log returns can be helpful when comparing impact across financial instruments with different prices. There is a dependence of Π on S_t as the price of the asset will be impacted by the trades executed. The cost of future trades may be impacted by the aggressiveness of trading up to that point, and to formalise the total cost incurred by the trader's own activity we define the following.

Definition 1.2.3 (Cost of Trading)**.** [15, Section 2.3] The expected cost of executing Π is $C[\Pi]$.

$$C[\Pi] := \mathbb{E}\left[\int_0^T v_t I_t \mathrm{d}t\right] = \mathbb{E}\left[\int_0^T v_t (S_t - S_0)\mathrm{d}t\right] \tag{1.2.3}$$

The definitions in discrete time are analogous with individual trades of size v_i.

1.3 Models

There is a rich history of price impact and optimal execution models in financial literature with both markets and models evolving over time, and in this chapter we summarise some of the key progress made from both a modelling and empirical view. For readability we attempt to keep the notation consistent between different models, slightly deviating from the original papers when required.

1.3.1 Kyle Model

Whilst the problem of optimal execution and price impact modelling is a thoroughly studied area of modern day quantitative finance, its roots began with much simpler models in the 1900s. One of the first papers to discuss a framework surrounding the problem emerged in 1985 where Kyle proposed the optimal liquidation problem involving one noise trader and one insider. In equilibrium, the price of an asset S would react to the aggregate trade of size v from the noise trader and insider by the linear relationship [4, Theorem 2]

$$S_1 = S_0 + \lambda v. \tag{1.3.1}$$

The model implies that price is permanently affected by trades, and the scale of this impact λ could be dependent on the variance of the noise trader. In the context of a LOB however, λ has the interpretation of a liquidity parameter, with a smaller λ indicating smaller market impact and therefore more liquidity. The notion of a liquidity parameter λ is a common theme in a number of subsequent models, and is often referred to as *Kyle's lambda*.

1.3.2 Almgren & Chriss

In [5], Almegren and Chriss develop a discrete time model including both permanent and temporary market impact. In the original paper the trader's objective is to liquidate an initial position X_0 before time T; however, for consistency we will present the problem from the point of view of a trader who needs to *buy* a position of size X_0 before time T. The time period $[0, T]$ is divided into N intervals of length T/N with $t_j = \tau j$. At time t_j the trader's position is x_j, and between times t_{j-1} and t_j the trader buys v_j stock. The price of the security evolves in response to a trader's trading schedule and through random fluctuations. [5, Eqn. 1]

$$S_j = S_{j-1} + \sigma \tau^{\frac{1}{2}} \xi_j + \tau g(\frac{v_j}{\tau}), \tag{1.3.2}$$

where $g(\frac{v_j}{\tau})$ represents the permanent market impact caused by average rate of trading between t_{j-1} and t_j, σ is the volatility of the security, and ξ_j are random variables drawn from a distribution with 0 mean and unit variance. Whilst there is an effect of permanent price impact in the model, part of the depletion of liquidity is assumed to be temporary and so the price that the trader can purchase the asset for at time t_j is [5, Eqn. 2]

$$\tilde{S_j} = S_{j-1} + h(\frac{v_j}{\tau}). \tag{1.3.3}$$

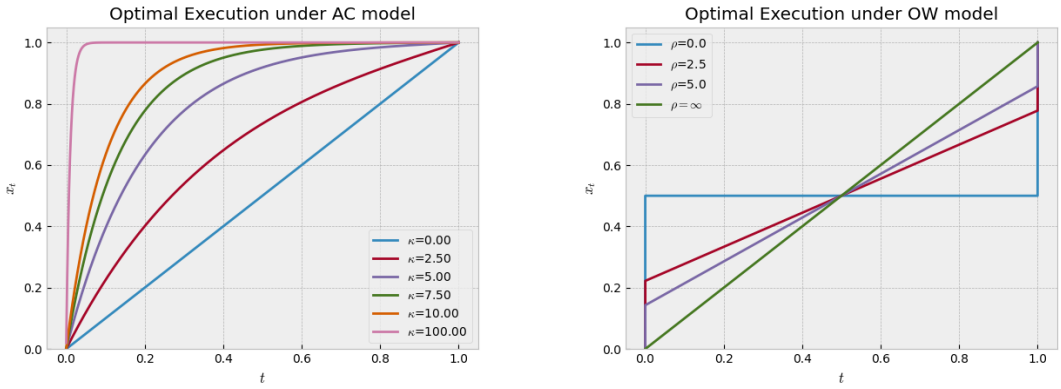
After choosing $g(x) = \gamma x$ and $h(x) = \epsilon \text{sgn}(x) + \eta x$, and optimising the mean variance problem with coefficient λ, the optimal trading schedule under the AC model in the limit of $\tau \to 0$ is shown to be

$$x_j = \left(1 - \frac{\sinh(\kappa(T - t_j))}{\sinh(\kappa T)}\right) X_0 \qquad j = 0, ..., N \tag{1.3.4}$$

where

$$\kappa = \sqrt{\frac{\lambda \sigma^2}{\eta}}. \tag{1.3.5}$$

In Figure 1.2a the effect of the parameter κ can be seen on the optimal trading schedule. As κ becomes large, the trader becomes more risk averse and prefers to accumulate the position quickly and eventually purchases the whole position in the first trading round. Conversely, as $\kappa \to 0$, the trader prefers to trade slowly in an attempt to minimise temporary transaction costs. In the limit the trader will accumulate his position linearly throughout the time interval.



(a) Almgren & Chriss model, $\kappa \in (0, 100]$.

(b) Obizhaeva & Wang model, $\rho \in [0, \infty)$.

Figure 1.2: Optimal buy schedules using AC and OW model.

1.3.3 Obizhaeva & Wang

Obizhaeva & Wang [6] view the problem of market impact through the lens of the limit order book. They propose that a stock price has a fundamental value S_t, and the ask price A_t evolves in response to a trader's buy orders. A buy order of size v_0 would "eat into" the ask side of the LOB moving the price of the security higher, but over time more limit orders would come from market participants and decrease the ask price. The speed of this replenishment of orders is described as the *resilience* of the order book. They propose that not all of the limit orders will return to the order book, but the ask price would tend to a new steady state higher than the original price, indicating the initial trade of v_0 had caused permanent market impact. Formally [6, Eqn. 6]

$$A_t = S_0 + \lambda v_0 + \frac{s}{2} + v_0 \kappa e^{-\rho t}, \quad \kappa = \frac{1}{q} - \lambda, \tag{1.3.6}$$

where s denotes the bid-ask spread of the security and ρ describes the resilience of the order book, with a higher ρ indicating limit orders return to the order book quicker after a buy market order. The density of the limit order book is represented by q; for a higher value of q there would be less immediate market impact. Finally, λ represents the size of permanent market impact caused by the trade.

To solve for the optimal buy schedule of size X_0 the setup is similar to that of the Almgren & Chriss model, however the price of the current trade is now affected by both the size of previous buy orders and the time passed since the orders. We look to minimise the cost function, over the set of buy orders v_i [6, Eqn. 14].

$$\begin{aligned} \min_{\{v_0, v_1, \ldots, v_N\}} \quad & \mathbb{E}\left[\sum_{j=0}^{N}[S_j + \frac{s}{2} + \sum_{i=0}^{j-1}\left(\lambda v_i + v_i \kappa e^{-\rho\tau(j-i)}\right) + \frac{v_j}{2q}]v_j\right] \\ \text{s.t.} \quad & \sum_{j=0}^{N} v_j = X_0, \end{aligned} \tag{1.3.7}$$

where S_j is the unaffected mid-price at time t_j. We see in Figure 1.2b the optimal buy schedule of the trader, and we will discuss this figure in more detail in the following section.

1.3.4 Transient Impact Model

Gatheral proposes a generic continuous time model for modelling the price of a stock given a trader's execution. [15, Eqn. 1]

$$S_t = S_0 + \int_0^t G(t-s) f(v_s) ds + \sigma \int_0^t dW_s, \tag{1.3.8}$$

where v_s is the rate of trading, $f(v_s)$ is the *instantaneous market impact function*, and $G(t-s)$ is the *decay kernel* allowing for the effect of trading at time s to be felt at some time $t \geq s$.

There are a number of special cases of the TIM. Permanent impact can be represented with $G(t-s) = 1$, and temporary impact can be represented with $G(t-s) = \delta(t-s)$ (where $\delta(\cdot)$ denotes the Dirac delta function). Choosing $G(t-s) = e^{-\rho(t-s)}$ - an exponential decay kernel, and $f(v_s) = \lambda v_s$ - linear market impact, yields the continuous time limit of the Obizhaeva & Wang model [15, Section 2.4]. The continuous OW model benefits from including realistic characteristics of transient market impact, and is one of the only continuous models that has explicit solutions for the optimal execution plan. The cost of trading under the OW model is

$$C[\Pi] = \int_0^T v_t \int_0^t \lambda v_s e^{-\rho(t-s)} \mathrm{d}s\mathrm{d}t, \tag{1.3.9}$$

and the optimal trading speed can be shown to be [6, Proposition 2]

$$v_s^* = \frac{X_0}{\rho T + 2}[\delta(s) + \rho + \delta(s-T)]. \tag{1.3.10}$$

The execution strategy is made up of two block trades: one at the beginning and one at end of the trading window, with continuous linear trading in between. The impact due to v_s under the OW model is

$$I_t = \mathbb{E}[S_t - S_0] = \int_0^t \lambda v_s e^{-\rho(t-s)} \mathrm{d}s, \tag{1.3.11}$$

and when following the optimal execution strategy v_s^* (Equation 1.3.10) we have $I_t^* = \frac{X_0}{\rho T+2}$ for $t \in [0, T)$, and $I_T^* = \frac{2X_0}{\rho T+2}$. The initial bulk trade is used to get to a desired impact state, and the trades for $t \in (0, T)$ correspond to maintaining the desired impact state. The final bulk trade leaves the trader at a higher impact state, but given there are no more trades after this, there will be no future cost to the trader for this large bulk trade.

The parameter ρ represents the speed of impact decay, and we can see the effect of ρ on the execution strategy in Figure 1.2b. $\rho = 0$ represents no impact decay, and the optimal strategy is to buy $\frac{X_0}{2}$ shares at the beginning of the trading period and $\frac{X_0}{2}$ at the end. In this setting, any strategy is an optimal strategy as the impact is permanent and the trader has no ability to wait until prices have returned to their equilibrium. Conversely, as $\rho \to \infty$ the decay of past trades is immediate and the optimal execution plan is to trade linearly throughout time with no bulk trades, and $v^*(s) = \frac{X_0}{T}$ for all $t \in [0, T]$.

1.4 Empirical Observations

Whilst it should be expected that buy market orders should cause prices to go up, it is not clear to what extent they should rise. Both the Almgren & Chriss and Obizhaeva & Wang models assume that market impact is linear as a function of trading speed, and the former assumes that impact decays exponentially throughout time. The TIM (Equation 1.3.8) provides great flexibility in the choice of both decay kernel and instantaneous impact function, and there have been a number of studies to measure certain stylised facts.

1.4.1 Impact Concavity

Lillo et al. [7] study 1000 stocks traded on the NYSE during 1995-1998 using both trade and quote data. They analyse the change of the log mid-price $\Delta S_i = \log\left(\frac{S_{i+1}}{S_i}\right)$ due to a trade of size V_i in trade time. They find on average that $\Delta S_i \propto V_i^\delta$ with $\delta \in [0.1, 0.5]$ for buyer initiated trades, with a higher exponent for smaller trade sizes in stocks with large market cap, and lower exponent for large trades in small cap stocks. Similar results were found for seller initiated trades.

Bouchaud et al. similarly use both trade and quote data, but for a single French stock between 2001 and 2002. They also considered price changes in trade time, but looked to quantify the average price change as a function of signed volume [8, Section 2.3].

$$\mathcal{R}(l, V) = \langle (S_{i+l} - S_i) \cdot \epsilon_i \rangle \mid_{V_i = V} ,$$

where ϵ_i is the sign of the i-th trade. The response function $\mathcal{R}(l, V)$ is factorised into individual components, and the following functional form is fit for the volume dependence.

$$\mathcal{R}(l, V) \approx \mathcal{R}(l) f(V) \propto \mathcal{R}(l) \ln V$$

Both studies provide evidence for highly concave instantaneous market impact at the transaction level, contradicting the linear impact used in previous theoretical models [5, 6] with analytic solutions.

1.4.2 Impact Decay

Impact decay measures how the effect of a given trade 'propagates' throughout time. There have been a number of investigations into the long-memory of order flow, which show a correlation between the signs of order imbalance. Taranto et al. study the slowly decaying correlation of the sign of order flow imbalance in trade time for two stocks (MSFT & AAPL) between February and April 2013. For MSFT, the autocorrelation increased for small lags before reaching a peak [9, Section 4.2]. The increasing autocorrelation could be attributed to market participants splitting meta-orders up into smaller trades, or herding effects/copy-cat traders. The autocorrelation slowly decayed for long time periods for both stocks.

Busseti & Lillo model price impact considering the log mid-price p_n [16, Eqn. 17].

$$p_n = p_0 + \sum_{j=0}^{n-1} f(v_j) G_0(n-j) + \eta_j$$

$$G_0(l) = \frac{\Gamma_0}{(l_0^2 + l^2)^{\beta/2}}$$

The model parameters are calibrated to data in both trade time and aggregated real time, and β is found to be between 0.07 and 0.23 [16, Table 2] for the calibration using 5 minute aggregated real time. The fit varies when considering the different aggregations of time, with the highest R^2 values with 64 aggregated trades (0.229-0.440), and the worst with 5 minute aggregation (0.153-0.292). The study provides reasonable evidence for the decay of market impact throughout time.

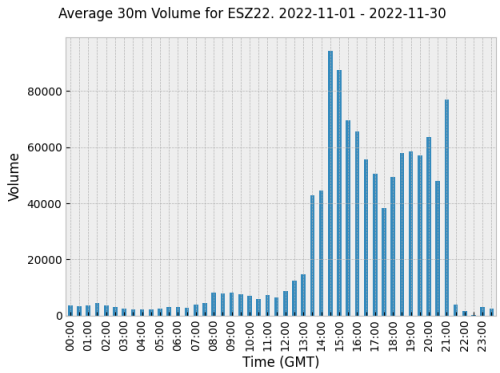
1.4.3 Coefficient of Impact

Whilst the shape of the decay kernel and impact function are important aspects of an impact model, the scale of impact λ is also a key component to understand. A number of questions arise: is this a universal quantity shared between all assets in an asset class, and does this quantity change throughout time or throughout the day?

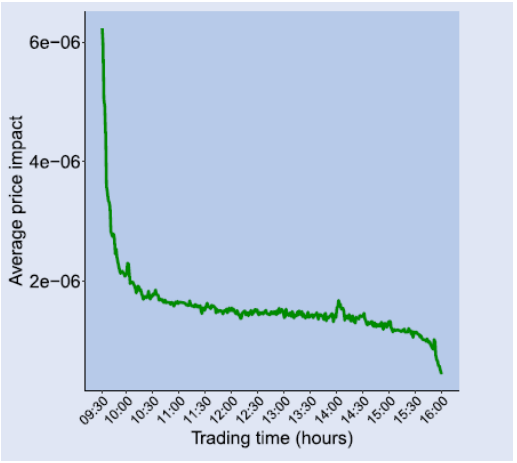
It is well known that the traded volume is extremely dependent on the time of day, with more volume traded at the beginning and end of the trading day [1, Section 4.2.1]. In Figure 1.3a we plot the average volume for each 30 minute period for an S&P futures contract over the period of a month. We see that the volume is highest during US market hours, when the underlying index constituents actively trade on the spot market. There is a U shaped pattern in this period, with heightened volume at the start and end of the US trading session. Whilst there is less volume traded in the Europe and Asia sessions, we still observe a U shape in these intervals.

Using Order Flow Imbalance (OFI) as a predictor for price change, Mertens et al. [11] found extreme time of day effects for single stocks traded on the NASDAQ. We see in Figure 1.3b the price impact being the highest at the start of the day, and decreasing to a minimum at the end of the trading session. OFI is made up of market orders, limit orders and order cancellations, thus providing more information about the state of the limit order book and trades than other empirical studies. Using a linear model of the OFI in each 10 second window to explain the price change in the same time period, the model was found to have an R^2 of 83% when refitting for every hour of the trading day for a single stock (MSFT) over a 125 day period in 2016 [11, Figure 1].

As λ is used to measure the scale of impact (ie the shallowness of the LOB), it seems reasonable to assume this could change if underlying market liquidity were to change. A report from Goldman Sachs [17] investigates the cost to buy a fixed amount of contracts for a variety of different futures. This quantity is measured in basis points and studied across time, and shows a decreasing trend between September 2022 and July 2023 for both equity indices and government bonds. For certain futures, the maximum cost can be over four times bigger than the minimum cost measured during the sample period, providing strong evidence for liquidity changes throughout time.



(a) Average intraday volume for ESZ22 in Nov 2022.



(b) Average price impact using OFI for MSFT [11, Figure 2].

Figure 1.3: Intraday volume and intraday price impact.

1.5 Limitations of Transient Impact Model

The Transient Impact Model (TIM) for price impact (Equation 1.3.8) provides a generic and flexible framework in which we can incorporate the stylised facts of price impact that are discussed in Section 1.4. The observed decay can be expressed in the TIM decay kernel, the nonlinear impact in the impact function, and we can incorporate changes in order book depth with λ. Figure 1.2b shows that the optimal execution schedule under the OW model consists of only buy orders and this result is due to our choice of $f(v_s)$ and $G(t-s)$, however in this section we will summarise the results of [15] and discuss scenarios that result in unrealistic optimal schedules.

Definition 1.5.1 (Round trip trade)**.** [15, Section 3]
A round trip trade is a trading strategy Π such that $x_0 = x_T = 0$. Equivalently,

$$\int_0^T v_t \mathrm{d}t = 0.$$

Definition 1.5.2 (Price Manipulation)**.** [15, Section 3]
A round trip trade that has negative expected cost, ie. $C[\Pi] < 0$.

In the search for a price impact model that accurately describes the reaction of price to a trading strategy in the real world, it is reasonable to seek a model that does not allow price manipulation. Such a model implies a trader can repeatedly execute a strategy Π with positive expected returns each time. We say a price impact model is *consistent* if it does not not induce price manipulation, and in the setting of the TIM (Equation 1.3.8) this suggests that the relationship between $f(\cdot)$ and $G(\cdot)$ defines whether a specific model is consistent.

Lemma 1.5.3. *[15, Lemma 4.1] Under the transient impact model, if $G(t-s) = e^{-\rho(t-s)}$ and $f(v)$ is not linear in v, then the model is not consistent.*

Example: [15, Section 4]
Let $f(v) = \sqrt{v}$, $\rho = 1$ and $T = 1$. Defining the trading strategy Π by $v_t = 0.2$ on $[0, \frac{5}{6})$ and $v_t = -1$ on $[\frac{5}{6}, 1]$, then $C[\Pi] = -0.001705$.

Lemma 1.5.3 rules out exponential decay for the continuous TIM, removing one class of functions from the possibility of choices. Gatheral proposes a more general result in [18].

Proposition 1.5.4. *[18, Proposition 6] Under the transient impact model, if $G(t-s)$ is finite and continuous at 0, and $f(v)$ is not linear in v, then the model is not consistent.*

With this result, and the empirical observations found in Section 1.4.2, it would seem reasonable to suggest a power law decay. Combing this with evidence for a power law function for instantaneous market impact, a potential model could be

$$I_t = \int_0^t \lambda(t-s)^\gamma \mathrm{sgn}(v_s) \left|v_s\right|^\delta \mathrm{d}s. \tag{1.5.1}$$

However, there are necessary conditions for this model to be consistent.

Proposition 1.5.5. *[15, Section 5.2, 5.3] If either of the following conditions are true, the model (1.5.1) is not consistent:*

- $\gamma + \delta \leq 1$
- $\gamma \leq \gamma^* = 2 - \frac{\log(3)}{\log(2)} \approx 0.415$

Given the empirical evidence of $\gamma \in [0.07, 0.23]$ [16, Table 2] and $\delta \in [0.1, 0.5]$, any reasonable combination of δ and γ in Equation 1.5.1 would result in condition 1 of Proposition 1.5.5 being satisfied, leaving the model inconsistent. Whilst certain characteristics of market impact may be described well, optimal execution will not be possible without forcefully imposing unidirectional conditions on the trading velocity.

1.6 Discrete Propagator Model

As discussed in Section 1.4, a number of stylised facts have been observed regarding both the concavity of volume traded on instantaneous market impact and the decay of this impact throughout time. Whilst the continuous TIM (Equation 1.3.8) allows great flexibility in choosing both the impact function and decay kernel, it is limited in it's ability to calibrate to real world data where individual trades are executed at discrete points in time.

Muhle-Karbe, Wang, and Webster discuss a framework for estimating a number of the parameters in the market impact model using the following discrete model: [13, Definition 2.1]

$$\Delta I_n = -\beta \Delta t I_{n-1} + f(\Delta Q_n), \tag{1.6.1}$$

where $\Delta I_n = I_n - I_{n-1}$, ΔQ_n is the signed volume traded in the interval $[t_{n-1}, t_n)$, $f \in C^1$ is an odd function, concave on $[0, \infty)$, and $\Delta t = t_n - t_{n-1}$. The term $-\beta \Delta t I_{n-1}$ allows the impact to *propagate* throughout time, and in the absence of further trades the impact will tend to 0. Due to these characteristics, we will refer to the model as the Discrete Propagator Model (DPM). We will slightly abuse the notation by letting $I_n = I_{t_n}$ depending on the context in question. Unlike other discrete transient models, the DPM has the advantage of a recursive relationship between subsequent impacts

$$I_n = (1 - \beta \Delta t) I_{n-1} + f(\Delta Q_n), \tag{1.6.2}$$

allowing for the impact state at each time to be calculated in $O(n)$ time. Other discrete models such as the one discussed by Busseti & Lillo use a lower triangular toeplitz matrix to represent the decay kernel [16, Eqn. 36], resulting in $O(n^2)$ to calculate the impact state. Whilst the DPM shares similar characteristics of impact decay and concave instantaneous market impact as the continuous TIM, we propose that these are in fact a separate class of models.

Proposition 1.6.1. *For nonlinear market impact, the limit of the discrete propagator model (1.6.1) is not equivalent to the continuous TIM with exponential decay (1.3.8).*

Proof. (Outline)
Beginning with the continuous TIM (Equation 1.3.8):

$$\begin{aligned} I_t &= \mathbb{E}[S_t - S_0] = \int_0^t G(t-s) f(v_s) \mathrm{d}s \\ \frac{dI_t}{dt} &= G(t-t) f(v_t) + \int_0^t \frac{d}{dt}(G(t-s) f(v_s)) \mathrm{d}s \\ &= G(0) f(v_t) + \int_0^t \frac{dG}{dt}(t-s) f(v_s) \mathrm{d}s \end{aligned}$$

Choosing $G(t-s) = e^{-\beta(t-s)}$

$$\begin{aligned} \frac{dI_t}{dt} &= f(v_t) + \int_0^t -\beta e^{-\beta(t-s)} f(v_s) \mathrm{d}s \\ &= f(v_t) - \beta I_t \\ dI_t &= -\beta I_t dt + f(v_t) \mathrm{d}t \end{aligned}$$

We compare this to the Discrete Propagator Model (Equation 1.6.1)

$$\Delta I_{t_n} = -\beta I_{t_{n-1}} \Delta t + f(\Delta Q_{t_n})$$

and conclude that, in general, the TIM is not the continuous time limit of the DPM, as for $\Delta t \to 0$,

$$f(\Delta Q_{t_n}) \to f(dQ_t) \neq f(v_t)dt \tag{1.6.3}$$

since $v_t = \frac{dQ_t}{dt}$.
It is worth noticing that for a linear f the identification in the limit holds true.

□

1.6.1 Parameter Estimation

Muhle-Karbe, Wang, and Webster calibrated the DPM on the constituent stocks of the S&P 500 in 2019. This calibration was performed by regressing onto observed market price changes, a technique that has been used in other literature [10, 11]. Specifically, $\Delta^h P_n$ is regressed onto $\Delta^h I_n$, where

$$\Delta^h P_n = \frac{P_{n+h} - P_n}{P_n}, \tag{1.6.4}$$

$$\Delta^h I_n = I_{n+h} - I_n. \tag{1.6.5}$$

The signed order flow ΔQ_n over 10 second bins was used to predict the price returns $\Delta^h P_n$ over a series of horizons, h =1, 15, and 60 minutes [13, Section 5]. A number of impact functions were used, including linear impact $f(\Delta Q_n) = \lambda \Delta Q_n$, and strictly concave impact $f(\Delta Q_n) = \lambda \text{sgn}(\Delta Q_n)\sqrt{|\Delta Q_n|}$. The predictive power of traded volume was found to be better at smaller time horizons; the in-sample R^2 decreased from 19% (1 minute) to 10% (60 minutes) for the best performing model. The results also showed that the square root impact function outperformed linear, with in-sample R^2 of 19% and 11% respectively. Refitting the model to every hour in the trading day resulted in an in-sample R^2 of 25%, but Muhle-Karbe et al. do not go into further detail on the magnitude of intraday impact λ_t.

Whilst an aim of [13] was to investigate using time dependent λ_t as a proxy for nonlinear impact, the highest R^2 value was obtained using both time dependent λ_t *and* square root impact, indicating the benefit of both features in the model.

Chapter 2

Futures

In the previous section, we have provided motivation for the problem of optimal execution and discussed a number of models that can help to describe the relationship between traded volume and price impact. A number of empirical studies have been performed to both observe impact characteristics and calibrate models, but have focused largely on single stocks from the UK and France, or constituents of the US-based S&P 500 Index. Fewer studies have been performed on the derivatives market, least so the futures market. A thorough search of the relevant literature yielded only one related article [19] in which Webb et al. study the intraday effect of price impact for one Korean Index future. We provide a brief introduction to the futures market and it's history, with the intention to further the study of price impact when trading this type of instrument.

2.1 Origin of Futures

Futures are exchange traded contracts that give the buyer the obligation to buy and receive the underlying asset for a pre-specified price at an agreed upon date. The trading futures contracts has a rich global history, with the Dojima Rice Exchange opening in 1730 in Japan [20], which gave local producers the ability to lock in a price for their rice months before the harvest and delivery of their produce. Both the rice farmer and the wholesaler buying the rice are incentivised to engage in a futures contract as it removes the volatility of spot rice prices in the future, giving them financial certainty. However, each party could have received a better deal if they had traded spot on the expiration date. If the spot price of rice had increased, the farmer would have been able to sell the rice at a higher price, and conversely if the spot price had gone down, the wholesaler would have been able to buy at a lower price.

The popularity of trading futures grew over the centuries with the opening of the New York Cotton Exchange (NYCE) in 1870, London Metals and Market Exchange in 1877, and the Chicago Mercantile Exchange (CME) in 1970. Modern day exchanges list contracts for not only commodities, but other asset classes, including FX, interest rates and equity indices [21]. These futures are extremely standardised, with contracts coming in an exact quantity (e.g. 5000 bushels of wheat, 1000 barrels of WTI Crude Oil), and expiring on specific dates. Futures contracts often expire on the third Friday of each month and, depending on the underlying spot, there will be a futures contract expiring on specific months throughout the year. For example, WTI Crude Oil has futures expiring every month of the year, whereas futures contracts on the S&P 500 Index expire quarterly. Futures tickers are made up of three parts; for example, ESU22 is a contract on the S&P Index (ES) expiring in September (U) 2022 (22).

Historically, on the expiration date of a future, one party was expected to deliver the agreed upon cash amount and the other expected to deliver the agreed upon amount of the physical good. Current contracts can also be cash settled, where the two parties exchange the difference between the agreed future price and the current spot price allowing for simpler settlement.

2.2 Pricing of Futures

We define the price of a future at time t expiring at time T as $F(t,T)$. Clearly the price of a future should be closely linked to the spot price of the underlying, however there are a number

of additional factors to consider when pricing a future. Following the methodology of pricing by replication, a replicator would have to borrow cash S_t to buy the underlying security at time t, and at time T return the borrowed money plus any interest incurred.

The replicator may also incur extra fees - "the cost of carry" - whilst holding the asset between times t and T, which can vary depending on the underlying asset. For example, purchasing barrels of oil will require the replicator to pay for both storage costs and insurance costs for the duration of the futures contract, and contracts with high carry fees will therefore often trade above the spot price - in *contango*. There can also be financial benefits to holding the underlying asset. Replicating equity index futures would require the replicator to buy a basket of stocks and hold them for the duration of the contract, during which time the stocks are likely to issue dividends. In this case the future may trade below the spot price - in *backwardation*. As the time to expiry decreases, the price of the future will tend to the spot price of the underlying S_t, and $F(T,T) = S_T$.

2.3 Futures Traders

Futures traders are categorised into two types: hedgers and speculators. Hedgers, like the rice farmer example in Section 2.1, look to hedge a risk they are exposed to, trading potential financial upside for certainty. Speculators take on such risk, looking to make a profit from the price changes. As mentioned in Section 2.2, futures prices are exposed to much more than the price of the underlying asset, making them a versatile security to speculate on a number of different exposures.

If the speculator is only looking to get exposure to the price of the underlying instrument, there are still a number of reasons why they would trade futures contracts instead of trading the underlying spot [22]:

- **Leverage** - Given futures are agreements to trade cash for an underlying product on a future date, there is no need to exchange cash when entering the contract, only on the expiry date. To mitigate counterparty risk, futures exchanges require each participant to provide a small amount of cash, called the *margin*, which is usually a small percentage of the futures contract value. This allows futures traders to gain more exposure than the initial margin, making them a leveraged security.

- **Difficulty in spot trading** - For some futures contracts it may be practically difficult to trade the underlying. To replicate an equity index future, one would have to buy every single stock in the index, which is much more expensive and less efficient. There may also be practical difficulties in shorting stocks, including short selling bans or broker requirements. Selling a future can give the seller positive exposure to the downside moves of the underlying, without any of the difficulties of trading spot.

- **Liquidity** - Futures can be much more liquid than the spot exchange for certain instruments [23], meaning institutional investors can enter and exit large positions with less fear of moving the market in an adverse direction. Futures contracts will often trade even when the underlying spot market is closed, e.g. equity index futures.

Chapter 3

Data Preparation

3.1 Motivation

Using the framework discussed in Section 1.6, we seek to calibrate a propagator model to different securities. We chose to focus on futures contracts given the range of benefits they offer investors, summarised in Section 2. Whilst [13] focused on impact exponents of 0.5 and 1, there is evidence to show that the instantaneous market impact function is strongly concave, with exponents ranging from 0.1 to 0.5. To calibrate the impact models we will need normalised binned data for both price returns and volume, and the impact state to be calculated with $\lambda = 1$ for a range of impact exponents. With a calibrated model for each security, we then intend to answer the underlying question of 'Given a security and an amount to trade, what is the optimal buy schedule that minimises transaction costs due to market impact?'.

3.2 Tick Data

Tick Data provides both tick level trade data (last executed trade price & volume) and Level 1 Quotes (bid & ask prices with size) for over 280 futures contracts [24]. The company works directly with exchanges (e.g. CBOE, ICE, CME) to consolidate and standardise consumption of separated data sources. The time period for the data we use ranges from 2012 - present, and data is collected for every trading day during this period. It is worth noting that Tick Data uses different codes for some futures compared to Bloomberg, for example FTSE 100 Index futures are 'FT' in Tick Data but 'Z' in Bloomberg. We will reference the Tick Data tickers for the purpose of this thesis, but we provide a lookup between the two in Table A.3.

Table 3.1 describes the format of the trade data provided. We make use of the 'Exclude Record Flag' by removing any trades that are flagged for exclusion. Tick Data provides a thorough report on the reasons for needing exclusions and the methodology and algorithms behind the flag [25]. We also use the 'Price' column as the true price, as opposed to the 'Unfiltered Price' as Tick Data recommends. To provide some context on the scale and frequency of the data being used, we consider the trade data for the contract ESH23 on 02/17/2023. There are 520,057 individual trades and when stored as a *.csv*, the file size is 27MB. Out of the 280 possible futures offered by Tick Data, we consider only 61 of these from a variety of asset classes: Interest Rates, Equity Indices, FX and Commodities.

3.2.1 Order Sign Reconstruction

An important quantity when modelling price impact is the sign of the trade volume, ie if the trade was a buyer initiated market or seller initiated. Exchanges rarely provide data explicitly stating who initiated the trade, and it is often up to additional data providers or practitioners to reconstruct this information themselves. The trade file from Tick Data only contains unsigned volume data, whereas previous studies [13, Section 5][11, Section 3] have used data from LOBSTER, which provides signed volume [26]. There are a number of algorithms to reconstruct the signed order flow such as the Lee and Ready algorithm used in [16], but this, and other algorithms, use Level 1 LOB (quote) data to infer trade volume sign. For example, if a trade of size 20 was recorded

Column name	Data type	Description
Date	MM/DD/YYYY	Date of the trade.
Time	HH:MM:SS.000	Trade time to the nearest millisecond.
Price	float	Trade price per contract, up to 7 decimal places.
Volume	integer	Number of contracts traded.
Market Flag	char(1)	P (Pit Trade) or E (Electronic Trade).
Sales Condition	char(4)	Exchange specific Sales Condition code.
Exclude Record Flag	char(4)	Identifies trades executed away from the exchange, including block trades etc.
Unfiltered Price	float	Raw price before any Tick Data filtering (see [24])

Table 3.1: Tick Data trade file format.

in the trade data, and at the same time the volume at the first level of the bid side of the LOB reduced by 20, this trade would be classified as seller initiated.

Whilst Tick Data provide such LOB data, we do not have access; therefore for the purpose of this thesis will have to look for other algorithms to infer order signs. Holthausen et al. [27, Section 3.3] describe what is now known as the *tick classification rule*, or *tick test*, an algorithm for classifying trade direction based on trade data alone.
The sign $\epsilon_i = \pm 1$ of trade i is defined as

$$\epsilon_i = \begin{cases} \text{-1}, & p_i < p_{i-1} \\ 1, & p_i > p_{i-1} \end{cases}.$$

Clearly there are a number of drawbacks to this algorithm, most notably the inability to classify trades that did not move the mid-price. Holthausen et al. [27, Table 4] were unable to classify 33.7% of their transactions, with most being trades of small volume. The tick test also creates the risk of misclassifying buys as sells (or vice versa), as the mid-price can change due to new limit orders being submitted and existing orders being cancelled. Extensions of this algorithm exist, such as the using the previous tick change if there was no change to the current tick, or the *reverse tick rule* which uses the next tick to classify the current trade [28, Section 1]. We chose not to use these extensions as we prefer the risk of being unable to classify the order sign, compared to the risk of misclassification. As such, we remove any data that we are unable to sign using the tick test.

Whilst previous studies [9, 16] calibrate impact models in trade time, we opt to view the problem in aggregated real time by aggregating the data into 10 second intervals. This follows the approach taken in [13], and aims to strike a balance between detail retained and computational efficiency. One disadvantage with this method is that aggregating signed volume further reduces the absolute volume, ie if there were three trades in one time bin $V_1 = 1, V_2 = -2, V_3 = 2$, the aggregate volume would be $V_{bin} = 1$, giving the appearance that less volume had been traded in the given time interval. We compare the traded volume for a number of futures at three different stages:

- Original - Unsigned volume data from Tick Data.
- Tick test (gross) - Unsigned volume data with trade direction inferred from the tick test.
- Tick test (net) - Unsigned volume data with trade direction inferred from the tick test, aggregated into 10 second bins.

For each of the above, we then look at the absolute volume traded at the end of each transformation. We find that amount of unclassified trades is higher for extremely liquid instruments where there is a lot of volume quoted at the best bid and offer, and the netting effect is most notable in frequently traded instruments such as S&P E-mini futures, where there are a large number of signed trades that net in a 10 second interval.

Although there can be a sharp decrease in observed traded volume, we would ideally like to see the intraday volume patterns remain. To analyse this decrease in volume, we plot the intraday profiles of the above metrics, normalised by total volume traded. We see in Figure 3.1 that for

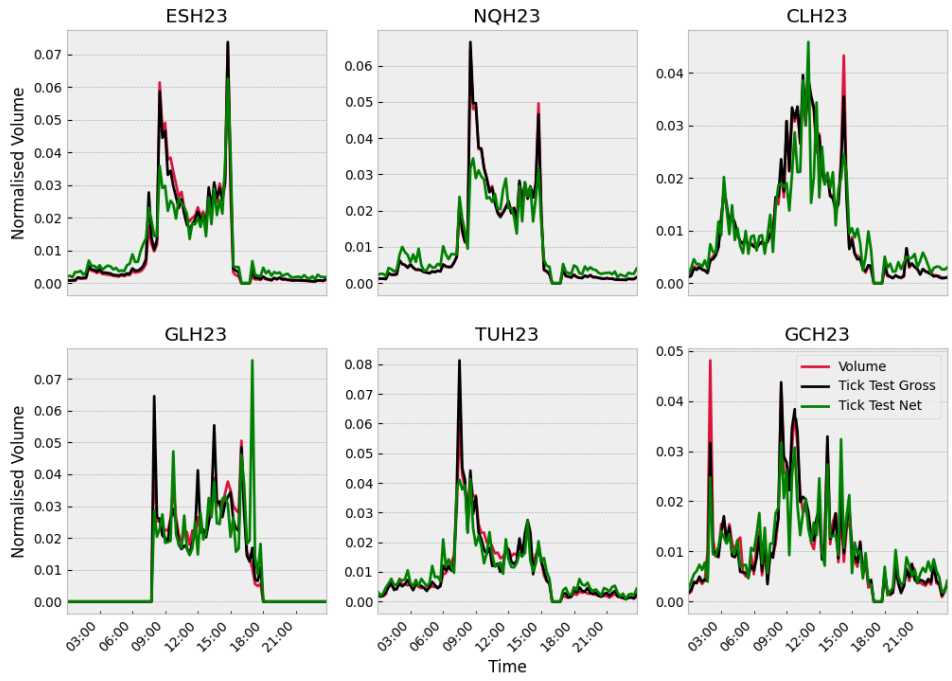


Figure 3.1: Average 15 minute normalised volume in February 2023 for different volume methods.

a range of different futures, binned to 15m and averaged over all days in Feb 2023, the intraday patterns from the original volume largely still remain after reduction from both the tick test and netting.

3.3 Future Combining

For any given underlying, there are always a number of futures available to trade at a given time. For example, in August for S&P futures, the soonest expiring contract is the September expiry, but there is also open interest in the December contract, and the March, June and September expiry contracts for the following year. Whilst there are reasons to trade multiple futures at once, such as gaining exposure to interest rate and dividend risk, traders looking to gain exposure to the underlying may choose to trade the most liquid contract, which could allow them to more easily enter and exit large positions. The most liquid contract can either be defined as the contract with the most open interest or the most actively traded volume, and whilst the most liquid contract is often the one which expires the soonest, this is not always the case. Traders will often roll their position before the current contract expires to avoid holding a contract to expiry, with common rolling windows varying from contract to contract. With this in mind, we look to create a 'synthetic' future by slicing the volume and return data of many contracts for the same underlying, which will represent the instrument traded by traders looking to trade the *most active* contract.

We will construct the synthetic future using rule based methods using observable patterns in the exchange traded volume. We use volume instead of open interest as this is a key quantity used in price impact modelling, however the process is slightly nuanced and very asset class dependant. We explain some of the challenges and differences between futures and provide a complete list of the rules in Table A.3.

- **Expiration months and dates** - Most equity indices will have contracts expiring quarterly. Figure 3.2 shows the extremely predictable three month roll of S&P futures, whereas commodities such as Crude Oil and Natural Gas will have contracts expiring monthly. Food commodity futures can have regular expiration contracts with extra dependant on the harvest, such as soybean futures that expire every other month as well as an extra contract for August. For a given expiration month, futures usually expire on the third Friday of the

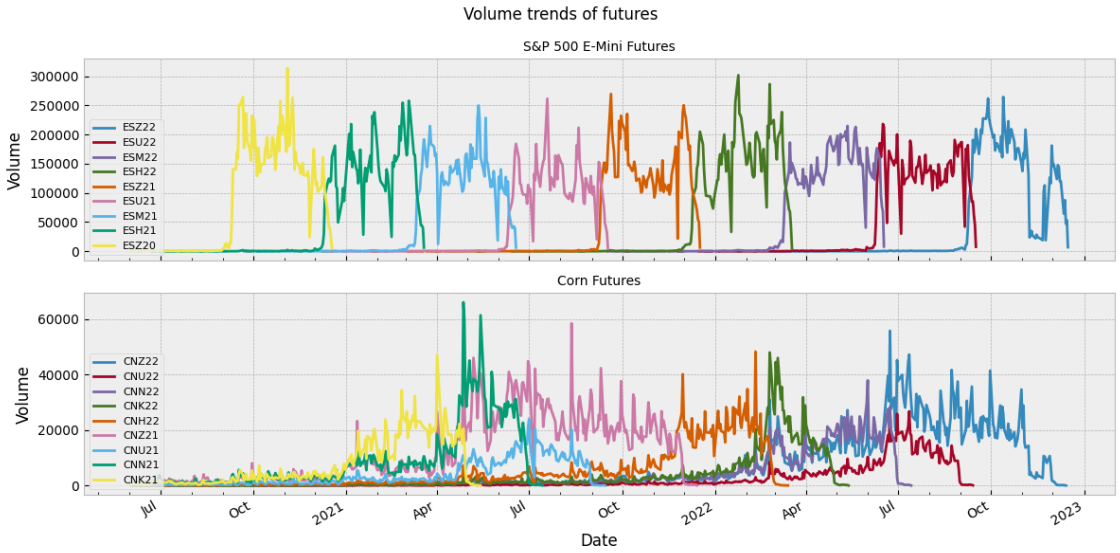


Figure 3.2: Volume trends for futures contracts.

month, except Hang Seng Index futures, which expire on the last trading day of each month.

- **Active Months** - Even if a future exists, it may never be the most actively traded one. This is observed in grain futures, such as corn. In Figure 3.2 we see how the September contract is dominated by the December contract for both 2022 and 2021.

- **Roll Time** - Equity Index and other cash settled futures will usually roll 1-2 weeks before expiration. Commodities such as Low Sulphur Gasoil and Cotton futures can roll 4+ weeks prior to expiry due to the physical settlement of the underlying.

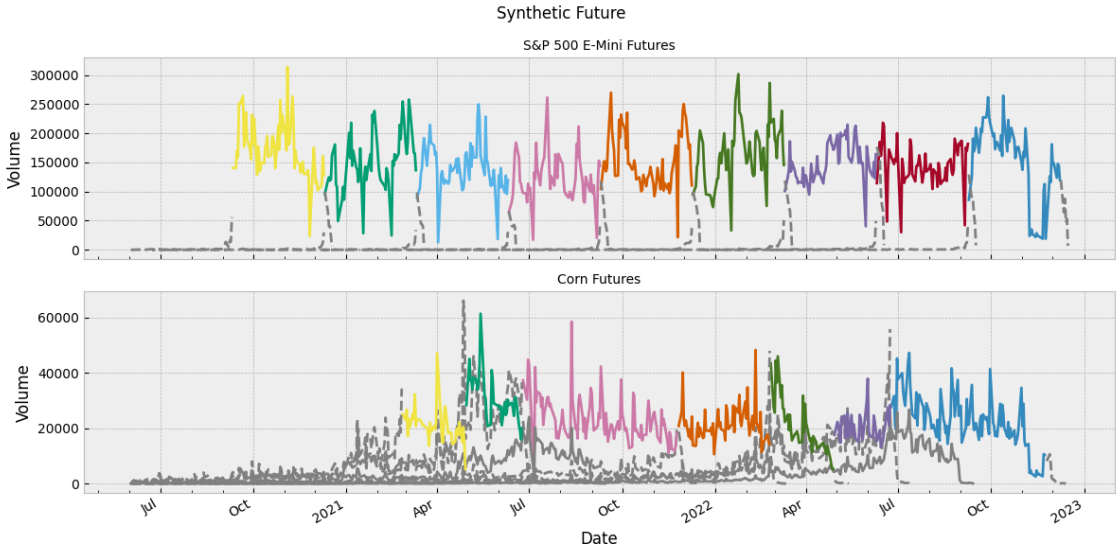


Figure 3.3: Synthetic Future based on rule-based roll.

Figure 3.3 displays the effect of applying our rules based method on the futures displayed in Figure 3.2. We see for CN that the September contract is never used in the synthetic future and the future chosen is usually the one with the largest daily traded volume. The same rules are applied to the price time series.

3.4 Data Normalisation

Market impact measures the change in price due to trades, but a number of different units of measurements can be used. In older models such as Almgren & Chriss (Equation 1.3.2) the stock price S_t was modelled, whereas in [7, 16] the log stock price is used. In [13], the change in impact $\Delta^h I_n$ is regressed on to $\Delta^h P_n$ (Equation 1.6.4), making impact a measure of return. Viewing impact in this way is a step in the right direction towards being able to compare different calibrated models, but more needs to be done. In both [13, Section 5.2] and [14, Section 4.1] a normalisation of impact and volume is done, where traded volume ΔQ_n is normalised by the average daily traded volume for a security, and impact is normalised by the average standard deviation of 10s returns. This transforms (1.6.1) into

$$\frac{\Delta I_n}{\sigma} = -\beta \Delta t \frac{I_{n-1}}{\sigma} + f\left(\frac{\Delta Q_n}{ADV}\right), \tag{3.4.1}$$

$$\Delta I_n = -\beta \Delta t I_{n-1} + \sigma f\left(\frac{\Delta Q_n}{ADV}\right). \tag{3.4.2}$$

With this normalisation Muhle-Karbe et al. were able to fit a universal model to all stocks in the S&P, with comparable in and out of sample performance to the stock specific calibration [13, Table 3]. This implies that, once we have accounted for the standard deviation of returns and average daily trading volume, different stocks roughly react the same to incoming market orders, regardless of tick size, market cap, financial sector or other characteristics.

To use this method in practice we will need a volume and volatility normalisation that can be known prior to our period of trading, and so we must enforce this for our historical backtest. We define the quantities

$$\mathrm{ADV}(m,k) = \frac{1}{k}\sum_{j=1}^{k} V_{m-j},$$

$$\sigma(m,k) = \frac{1}{k}\sum_{j=1}^{k} \sigma_{m-j},$$

ie. the k day moving average prediction for day m, where V_i is the gross volume on day i, and σ_i is the standard deviation of binned returns on day i. This normalisation restricts our data usage by requiring k days of data to calculate moving averages, meaning impact state can not be computed for the first k days. In Figure 3.4 we see the effect of a 20 day simple moving average for the synthetic future ES which will be used for normalising the binned data.

Using these normalising quantities, we explicitly redefine the DPM which we will use when calibrating to market data in future sections.

$$\Delta I_n^m = -\beta \Delta t I_{n-1}^m + \sigma(m,20) f\left(\frac{\Delta Q_n^m}{\mathrm{ADV}(m,20)}\right), \tag{3.4.3}$$

where the superscript m is used to denote trading and impact activity on day m.

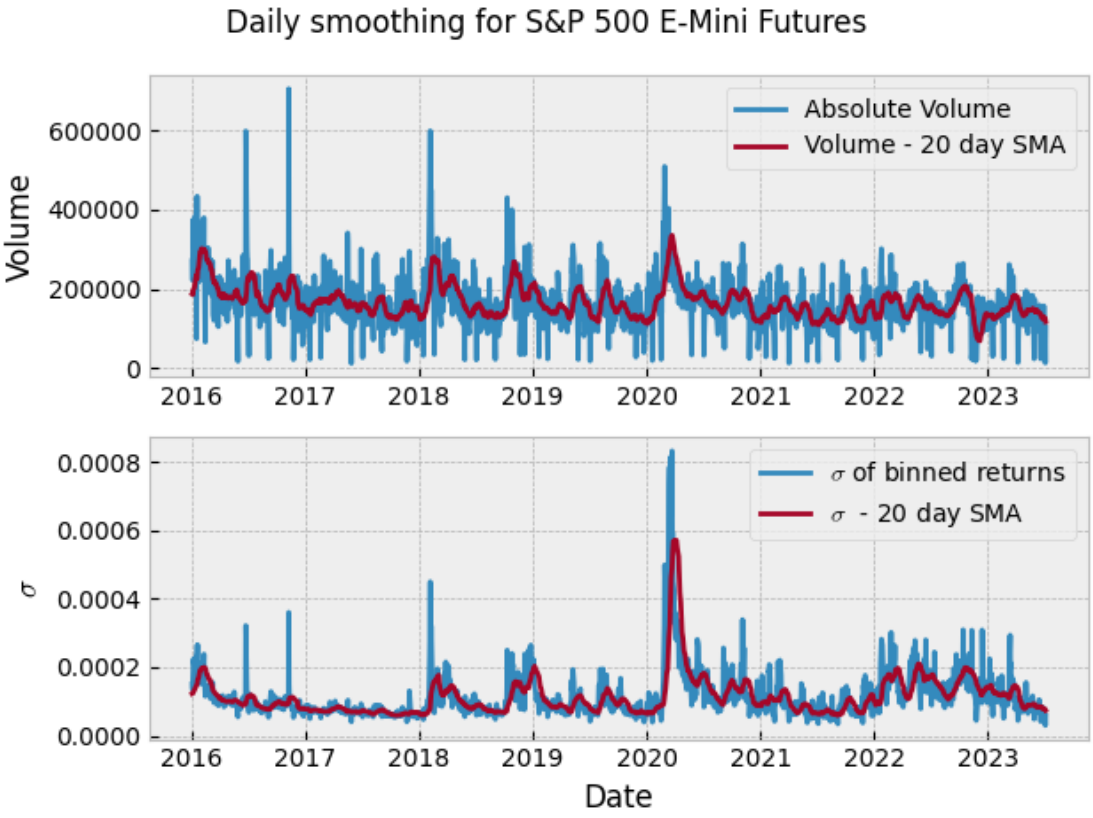


Figure 3.4: 20 Day smoothing of ES synthetic future.

Chapter 4

Model Calibration

4.1 Methodology

We aim to calibrate the model

$$\Delta I_n^m = -\beta \Delta t I_n^m + \sigma(m, 20) f\left(\frac{\Delta Q_n^m}{\text{ADV}(m, 20)}\right). \tag{4.1.1}$$

We drop the day superscript and moving average notation and assume a functional form of $f(x) = \lambda \text{sgn}(x) |x|^\delta$. We also assume that the impact state at the start of every trading day is zero, ie $I_0^m = 0$, $\forall m$. To emphasise the dependency of impact on it's parameterisation, we have

$$\Delta I_n = \Delta I_n(\lambda, \delta, \beta) = -\beta \Delta t I_{n-1} + \sigma \lambda \text{sign}\left(\frac{\Delta Q_n}{\text{ADV}}\right) \left|\frac{\Delta Q_n}{\text{ADV}}\right|^\delta. \tag{4.1.2}$$

After fixing a value for δ and β, we assume that the observed price change (for any time horizon) due to volume order flow is exactly equal to the expected impact, plus some random variable ϵ with 0 mean and constant variance.

$$\Delta^h P_n = \Delta^h I_n(\lambda, \delta, \beta) + \epsilon, \tag{4.1.3}$$

where both quantities are defined in Equations 1.6.4 and 1.6.5. We note from Equation 4.1.2 that the quantity $\Delta I_n(\lambda, \delta, \beta)$ is linear in λ, ie $\Delta I_n(\lambda, \delta, \beta) = \lambda \Delta I_n(1, \delta, \beta)$, which allows us to pre-compute $\Delta I_n(1, \delta, \beta)$, and then estimate λ using linear regression.

4.1.1 Impact Decay

One of the key parameters in Equation 4.1.2 is β, which allows us to control the speed of decay of the impact caused. This is one of the notable stylised facts observed in previous studies, as discussed in Section 1.4.2. In [13, Section 5.1], different decay speeds were tried with half lives between 1 and 120 minutes with "negligible impact on the results", and Hey et al. [14] went further to quantify the effect of decay speed on the quality of fit to data. [14, Figure 3] shows that a change in decay speed has little effect on the R^2 compared to a similar scaled change in the impact concavity. Whilst the model used is slightly different to Equation 4.1.2 (specifically, the concavity is applied to the final impact state as opposed to the instantaneous impact), the evidence suggested by both papers imply that the quality of fit is affected most by the coefficient of impact λ and the concavity of impact δ. Due to the added computational complexity of jointly calibrating all three parameters, we chose to fix β to have a half life of one hour ($\beta = \frac{\ln(2)}{3600}$) and optimise for (λ, δ) only.

4.1.2 Bin Size

Whilst there have been a number of studies prior calibrating price impact models, some even using a regression-based framework, there is little consensus on the time intervals to consider. One method is to compare price in trade time [16, 8], where price change is measured between every individual trade. Alternatively, trades and volume can be aggregated into time bins of fixed size,

such as 10 seconds [10, 13]. Mertens et al. found no conclusive difference in R^2 between 60 and 300 second time bins [11, Table 3], and Busseti & Lillo found little difference in R^2 between bins of size 10 to 60 seconds [16, Figure 1]. Given these results, we chose to use $\Delta t = 10$ seconds for the calibration of 4.1.2, which should help to maintain as much of the volume information as possible, and not lose information due to volume netting.

4.1.3 Prediction Horizon

For $h = 1$, the model is *explanatory*, it attempts to explain why the price changed in an interval due to the volume traded in the same interval. For $h > 1$, the model becomes *predictive*, it attempts to predict future price changes due to the volume in a previous interval. [13, Table 3] found the model fit quality to decrease as h increased, which makes practical sense because the larger h is the more the model is attempting to predict future order flow caused by current order flow. For our initial analysis we fix $h = 1$ and try to quantify the explanatory capability of the model.

4.1.4 Optimising for (λ, δ)

As described, after fixing a δ we can find λ by linear regression. Optimising jointly for (λ, δ) is more difficult because for each δ we try, we need to pre-compute the impact states $\Delta^h I_n(1, \delta, \beta)$, then fit for λ. Due to this computational constraint we choose to calculate the impact states for $\delta \in \{0.1, 0.15, 0.2, ..., 0.65, 0.7\}$ and then fit a λ for each hour of the trading period. This range for δ covers the range found in previous studies mentioned in Section 1.4. As for frequency of calibration, previous studies have fit λ to a range of intervals including 90 minutes [13, Section 5.4], 30 minutes [10, Section 4.1] and 60 seconds [11, Section 4.1].

4.1.5 Test - Train Split

Whilst it is important to have a model which fits well to in-sample data, we plan to use the calibrated model to decide volume schedules in the future. It is therefore important that we understand how well our calibrated model performs on unseen data. It is natural due to the time series nature of the data to train the model using n sequential months [29, Section 2] and then test on the next m months, allowing the backtest strategy to best mimic the production use of such a model. We demonstrate this in Figure 4.1 with a series of sliding windows.
A longer training window allows the regression to be trained on more data points, potentially leading to a better understanding of the data. If the underlying relationship between the dependant and independent variable changes over time, then a smaller training window may be preferred so the model can train on more recent and relevant data. The length of the testing window largely depends on both how the relationship changes over time, and the practicality of retraining the model frequently. We arbitrarily choose the backtesting strategy to be a series of sliding windows with 6 months training and 3 months testing, with each period rolling forward by 3 months.

Figure 4.1: Example sliding window.

4.2 Calibration Results

4.2.1 Initial Observations

Table 4.1 displays the metrics we record for each model fit per future, δ and test start date. The data shown is for ES, $\delta = 0.1$ and test start 2022-01-01. We record the coefficient of impact λ and the train and test R^2 for each hour a regression was performed. In most cases, hour 0 represents the first hour in the trading session, but for futures such as GL which begin and end trading in one single day (BST), we leave the hour unchanged. We also plot some fit results for a selection of futures in Figure 4.2, using 6 months training data and 3 months testing data, with the test period starting on 2022-01-01. The in and out-of-sample R^2 is plotted, along with the impact coefficient λ. We chose to show the δ for each future that had the best average in-sample R^2. For each future we look to observe any intraday patterns in λ, the value and any intraday patterns in the R^2 values, and the relationship between train and test R^2.

Hour	λ	Test R^2	Train R^2
0	1.90	0.30	0.45
1	1.89	0.35	0.49
2	1.94	0.34	0.46
3	1.90	0.34	0.45
4	1.82	0.34	0.51
5	1.86	0.37	0.57
6	1.89	0.33	0.55
7	2.03	0.32	0.48
8	1.98	0.35	0.44
9	2.04	0.30	0.35
10	1.88	0.31	0.40
11	1.87	0.33	0.42
12	1.89	0.31	0.42
13	1.72	0.30	0.40
14	1.69	0.26	0.30
15	1.59	0.17	0.13
16	1.56	0.17	0.13
17	1.39	0.18	0.14
18	1.30	0.19	0.17
19	1.31	0.19	0.17
20	1.43	0.18	0.17
21	1.54	0.18	0.16
22	1.27	0.16	0.22

Table 4.1: Example Calibration output for ES futures using 6 months training and 3 months testing data. $\delta = 0.1$.

For ES futures, the impact coefficient is highest for the first 14 hours of the trading session, which represents the Asian and European market hours. As the US market opens, λ is still higher at the start of the trading day and decreases throughout the day, coming to a minimum in the final hour of trading. The calibration for NQ futures follows a very similar intraday pattern, possibly due to the fact they are both equity index futures with US stocks as index constituents. The minimum value for λ is 56% and 38% less than the maximum value for ES and NQ respectively, potentially implying a trader could benefit from waiting for periods of low impact before executing trades. As for the goodness of fit for both tickers, both have R^2 values between 0.4 and 0.6 for pre-market hours, this is consistent throughout the whole day for NQ, but the quality of fit drops substantially for the US market hours for ES. As for test train stability, both futures exhibit similar R^2 for both in and out-of-sample data, implying the model is able to generalise well and the relationship between traded volume and price impact is similar between the train and test periods.

For GL (10-Year United Kingdom Government Bonds) and TU (2-Year United States Government Bonds) there appears to be much less intraday changes in impact, with the difference between maximum and minimum impact being 20% for both futures despite the fact that TU

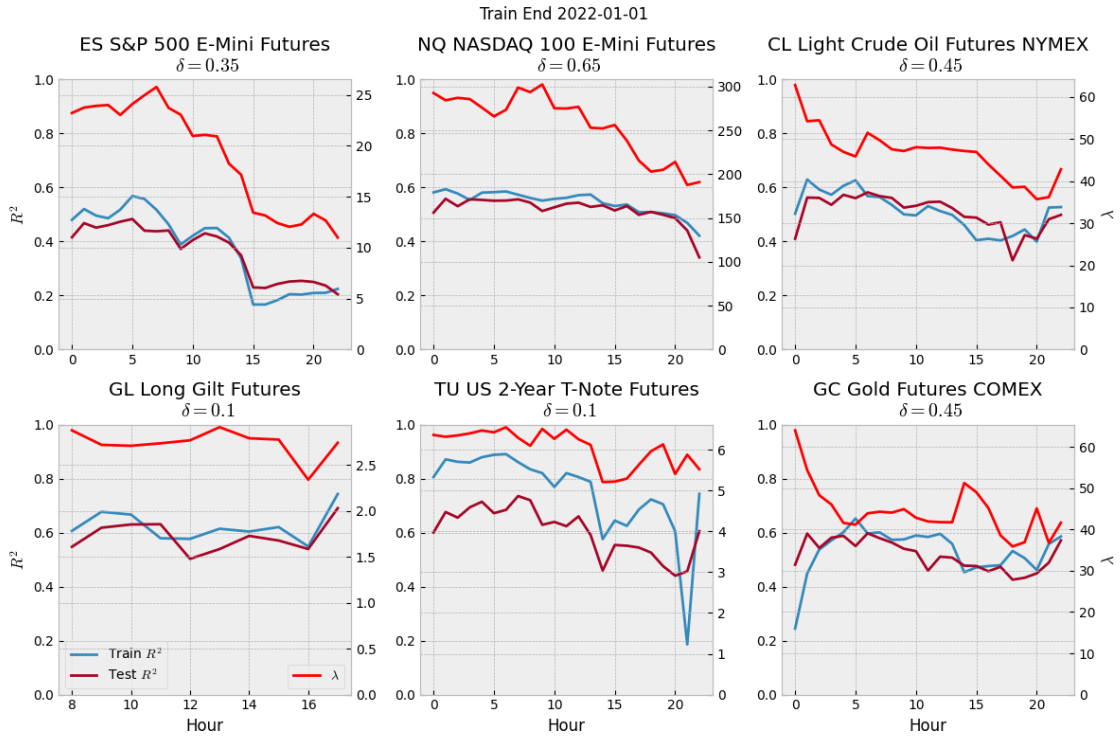


Figure 4.2: Model fit and R^2 for 6 futures using 6 months training and 3 months testing data.

trades outside US market hours, whereas GL generally does not trade outside UK market hours. GL futures do see a small dip towards the end of the trading day, finding a minimum at hour 16 followed by a sharp increase in hour 17. It is worth noting that most UK trading ends at 16:30 when the London Stock Exchange closes, but GL futures trade until 18:00 making the last 90m of trading similar to 'out of market' hours. Train and test R^2 are very similar for GL, but for TU there is a clear decrease in out-of-sample performance. This could be due to overfitting of the model or changes in market regime between the test period and the train period. Given the simple linear regression model used to optimise for λ, we suggest that there could be structural market changes between these two time periods, and one potential remedy for this could be to reduce the number of months used in the training and test period.

CL (Light Crude Oil) and GC (Gold) futures both exhibit stability between in and out-of-sample performance and have large coefficients of price impact soon after the trading session begins. Impact continues to decrease for CL once the US markets open, whereas GC's impact initially increases for US market open and decreases throughout the US trading day.

Whilst each future exhibits unique characteristics, there are some general themes we can see throughout all examples above.

- Price impact is often higher during 'out of market' hours, and decreases during US trading hours. This makes some sense as there is usually much less volume and liquidity during these hours, meaning trades can eat deeper into the LOB causing a large price change.

- Price impact is often higher at the beginning of US trading hours, and decreases throughout the day. At US market open (hour 14), λ often begins high and decreases throughout the trading day which could be slightly counter-intuitive to traders who believe price impact is only dependant on volume traded. US market open is one of the most traded times in the day, but is one of the 'worst' times to trade. This analysis aligns with other studies such as [1, Figure 4.2] where the volume at the best bid and ask quotes is higher nearer to the end of the trading day, and [11, Figure 1] where impact is found to be highest at the start of the trading session and lowest at the end. Possible reasons for low impact at the end of the day include: market makers looking to offload inventory to reduce overnight risk, and ETF providers trading nearer to the end of the day to reduce tracking error.

- R^2 is often higher during illiquid hours. There is a general trend of higher R^2 values in

hours 0-16, and smaller values in US market open hours. Due to the smaller trading volume and liquidity in overnight these hours, price impact seems to conform better to the concave propagator model. In Section 3.2.1 we discussed some issues that arose when reconstructing order signs, and the reduction in volume during overnight hours could have lead to a higher rate of tick classification and lower effect of netting, allowing for less information loss.

4.2.2 Changes in λ throughout time

The difference between train and test R^2 for TU in Figure 4.2 raises the possibility of λ changing throughout time. This is an important feature of price impact to understand; if we are deciding volume schedules for tomorrow from a model that has been trained on the last 6 months of data, we would like to be confident that the shape and scale of impact will be similar between the two time periods.

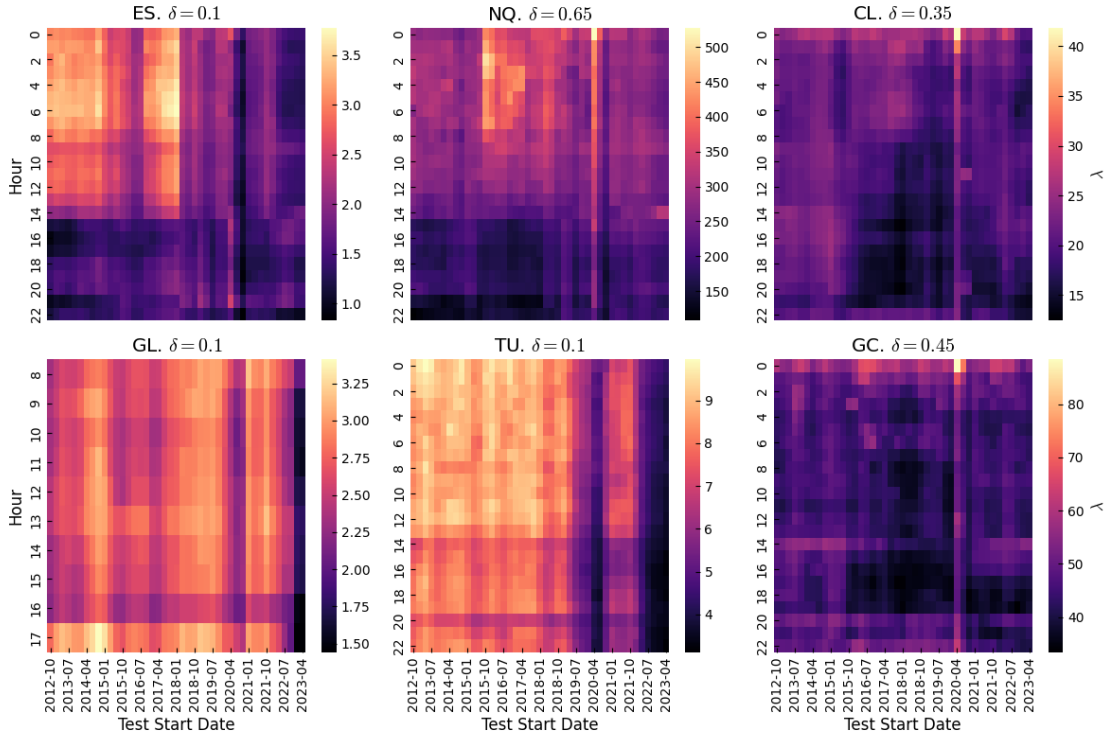


Figure 4.3: Coefficient of impact over time for 6 futures using 6 months training and 3 months testing data.

Figure 4.3 shows how the fit for λ changes throughout the different backtest windows. One thing that can immediately be seen in all plots is a stark change in λ for test periods starting in mid-late 2020. This represents a model that has been trained on data in early-mid 2020 where most equity markets crashed during the beginning of the COVID pandemic, as well as most interest rate and commodities futures experiencing higher levels of both volume and volatility. However, not all markets seem to have changed in the same way during this time. Looking at both ES and NQ, there appears to be a period of slightly higher λ followed by a large decrease. The period of high λ, where prices reacted very strongly to volume, could indicate lower liquidity during that time, potentially due to more uncertainty from investors on the future direction of the global economy. Government Bond futures GL and TU only show a decrease in λ during this period however, implying price was less sensitive to traded volume.

Aside from the time period covering the beginning of the COVID outbreak, in some futures we can see general long term shifts in λ. For TU in earlier years, the coefficient of impact is much higher than more recent years. For GL we can see λ has been historically quite low, with distinct increases in 2014 and 2017-2020, which could be attributed to uncertainty in the UK economy regarding the effects of Brexit and the decoupling of the UK from the EU. The only periods of low impact were in 2020 and late 2022, the latter of which could be caused by changes in interest rates

by the Bank of England (BoE) in response to rising inflation during this period. We also see this decrease in impact for TU but slightly earlier, which reinforces our suspicion that λ decreased in this time due to central bank interest rate rises, as the Federal Reserve also increased interest rates sharply and slightly before the BoE. We investigate this effect in further detail in Section 4.4.1.

Whilst the magnitude of λ changes over time for GL and TU, the intraday pattern persists throughout time. This cannot be said for ES and TU however, with ES showing a steep decrease in λ during US market close hours in 2018. This indicates a large structural change within the overnight market whereas the US 'market open' impact remained relatively stable over a similar period. NQ also experienced a change in overnight impact only in the years 2015-2018, where λ increased making trades more impactful on price during this time. GC and CL experienced decreased λ in 2015-2019, but during US market hours only.

Figure 4.3 provides evidence that frequent retraining of the price impact model is necessary to capture frequent changes in λ. If retraining does not happen, then traders are at risk of the following:

- Trading at the wrong time - if changes to the intraday pattern have not been captured by re-calibrating, traders may trade too much at periods of high impact and too little during periods of low impact, adding to their transaction costs.

- Trading the wrong size - some trading strategies are capacity constrained, meaning the strategy only works up to a certain amount of volume, due to the fact that a higher volume can erode the alpha the trader is intending to capture. If the model is not calibrated to recent data and the future impact is less than the model assumes, then the trader would overestimate their price impact and leave money on the table. Conversely, if a trader underestimates their price impact they could trade too aggressively and erode their alpha, losing money.

4.2.3 Optimising for δ

So far we have thoroughly investigated the characteristics of λ, both in terms of its intraday profile (Section 4.2.1) and how the scale and shape changes throughout time (Section 4.2.2). To do this we chose a value of δ somewhat arbitrarily for each future, but many questions still remain about the optimal choice for this parameter over time. In Figure 4.4 we plot the train R^2 values for every backtest window for a range of different deltas, using the same futures contracts we have studied thus far, and using 6 month training window and a 3 month testing window. For each future we look to see if there is any δ that is consistently optimal, if there are distinct regime changes in optimal δ, or if there are frequent/sporadic changes.

The scale of δ implies certain properties about the order book for a given future. Extremely small values mean that not much trading is required to have an impact, but there is little difference in impact between a small trade and a big trade. This indicates that the volume quoted at the best bid and offer could be quite thin, but with very large volume immediately after. Conversely for higher values of δ, the difference in volume between each level of the order book is much less extreme, with gradual increases in volume at each level.

For TU and GL (both government bond futures), there is a clear sequential nature to the performance of each δ, with lower values performing better. For TU this is more pronounced with each increase of δ providing a large increase in performance and $\delta = 0.1$ giving the best fit in all backtest windows. Given the optimal value is on the boundary of our delta grid, this implies we could find a better fit for TU with $\delta \in [0, 0.1)$. For GL, however, the improvements are slightly smaller, with a saturation level occurring at 0.6 R^2 for $\delta \in \{0.1, 0.15, 0.2\}$.

Similar behaviour of 'lower δ is better' can be observed for ES in the years before 2021 where $\delta = 0.1$ consistently outperformed other values of δ. However, in 2021 we can observe a distinct regime change, where lower values of δ became worse and the optimal value of δ changed between backtest windows taking values in $\{0.2, 0.25, ..., 0.5, 0.55\}$. Interestingly, $\delta = 0.1$ is the worst performer in these later years, despite being the best fit in previous years. In this period we can see that the R^2 increases as δ increases up to a critical value δ^*, after which the R^2 begins to decrease. For NQ, there is not a clear optimal value for δ for any large period of time. Pre 2019, δ^* ranges between 0.3 and 0.5, and in later years larger values for δ appear to fit the data best, with $\delta = 0.65$ being most often the best fit. There is one thing to note for both these futures: in periods where δ^* is changing frequently in consecutive backtest windows, there is little difference in R^2 between the top 5 performing values of δ.

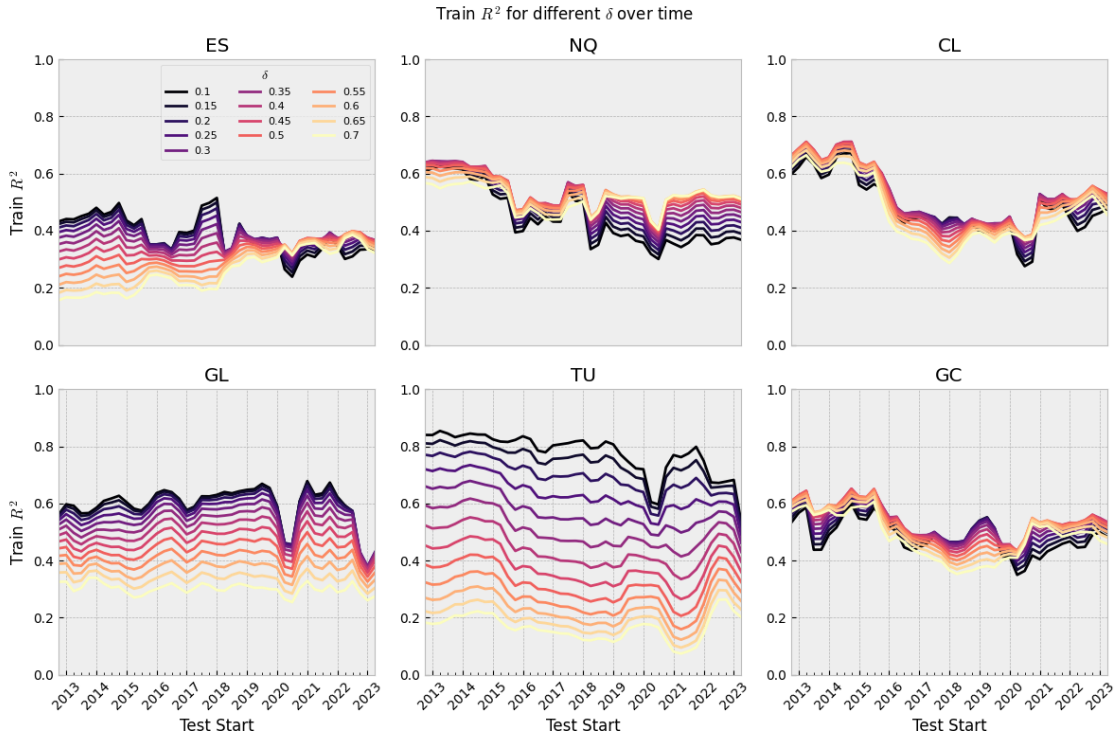


Figure 4.4: Train R^2 over time for different δ using 6 months training and 3 months testing data.

For CL and GC the optimal δ often changes, but usually stays within the range $[0.2, 0.5]$. They both exhibit similar traits with both the optimal delta and value of R^2 being the same at a given time. Whilst each future has unique characteristics, all seem to experience worse performance during early-mid 2020 during the beginning of the COVID pandemic. We saw in Figure 4.3 that this period caused clear changes in the value of λ, but here we see widespread deterioration in model fit, further adding evidence that this was a clear period of market irregularity.

4.3 Model Selection

In the previous section, we saw there can be a lot of variability in the optimal value of both δ and λ for a given future between months. As we will be creating volume schedules for the future based on data from the past, we will need to choose one combination of (λ, δ) in each backtest window to use in the next time period. We would like to be able to quantify for each backtest window if we could (with perfect knowledge) have picked a different parameterisation, and if so, how good that fit could have been relative to our chosen fit. We use a simple methodology for picking the 'best' parameters in a given backtest window - the combination of (λ, δ) that gave the highest train R^2 score. We define the following:

- $R^2_{train}(\lambda, \delta)$: Average Train R^2 over all hours for a given backtest window and λ, δ
- $R^2_{test}(\lambda, \delta)$: Average Test R^2 over all hours for a given backtest window and λ, δ
- $\lambda^*, \delta^* = \arg\max_{\lambda,\delta} R^2_{train}(\lambda, \delta)$: Optimal parameters according to train R^2
- $\lambda', \delta' = \arg\max_{\lambda,\delta} R^2_{test}(\lambda, \delta)$: Optimal parameters according to test R^2

To help us understand the stability between test and train period, we consider the 'Test Train Ratio' $R^2_{test}(\lambda^*, \delta^*)/R^2_{train}(\lambda^*, \delta^*)$ for each backtest window. Values close to and above 1 indicate that the optimal parameters chosen from in-sample data accurately describe out-of-sample data. Whilst this metric helps to explain the stability of the model between test and train periods, it does not provide any indication on the quality of our model selection; if the value is low, it could be because there was an alternative parameterisation providing higher R^2 in the test period, or

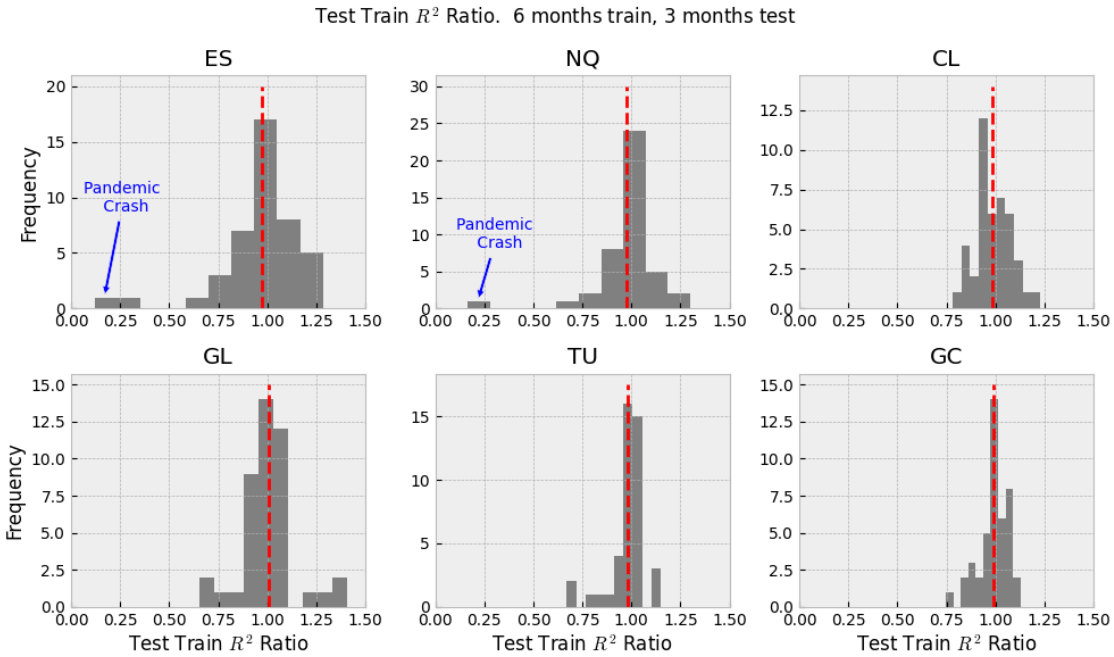


Figure 4.5: Test Train Ratio using 6 months training and 3 months testing data.

it could be that all models performed poorly in the test period and we could not have selected better. We also consider the metric 'Max Test Regret' $R^2_{test}(\lambda^{'}, \delta^{'}) - R^2_{test}(\lambda^*, \delta^*)$, for each backtest window and future. This metric is bound below by 0 given both values are positive, and higher values indicate a higher level of regret, suggesting we could have theoretically chosen a much better model.

We see in Figure 4.5 that all futures have a test train ratio centred around 1, an ideal property implying that on average the training dataset and fit accurately describes the testing set. There are, however, some instances where the test R^2 is much worse than that seen in the training data, with ES and NQ both having test R^2 75%+ worse than the train R^2. Notably, these two instances of poor out-of-sample performance relate to a test start date of 2020-04-01, where the model was trained on the equity market crash caused by the COVID 19 pandemic. The other low performing backtest window for ES was for the test start date of 2018-01-01. In Figure 4.3 we can see this was the period where overnight hours transitioned from a period of extremely high impact to a period of lower impact.

In Figure 4.6 we plot histograms for the 'Max Test Regret' metric. For futures GL and TU we see that the model selection algorithm selects the best parameters for test R^2 the vast majority of the time, and when it does not the difference between chosen and best R^2 is very small. This is also largely true for the remaining four futures, but some have long right tails with ES having one instance of a test regret of 0.1.

Summarising the results from both plots, we can make two conclusions.

- The model selection algorithm of picking (λ,δ) based on the best train R^2 does a reasonably good job, with rare instances of the model being able to pick a much better set of parameters with perfect knowledge.

- There are a number of instances where the test R^2 is much worse than the train R^2. This implies the training dataset does not characterise the testing data well and, given the sequential time series nature of the underlying data, supports the idea that it may be beneficial to use a smaller training and testing window. This would mean a larger proportion of the training dataset would be 'relevant' to the testing set, at the risk of reducing the amount of data points seen by the model for training, allowing for potential overfitting.

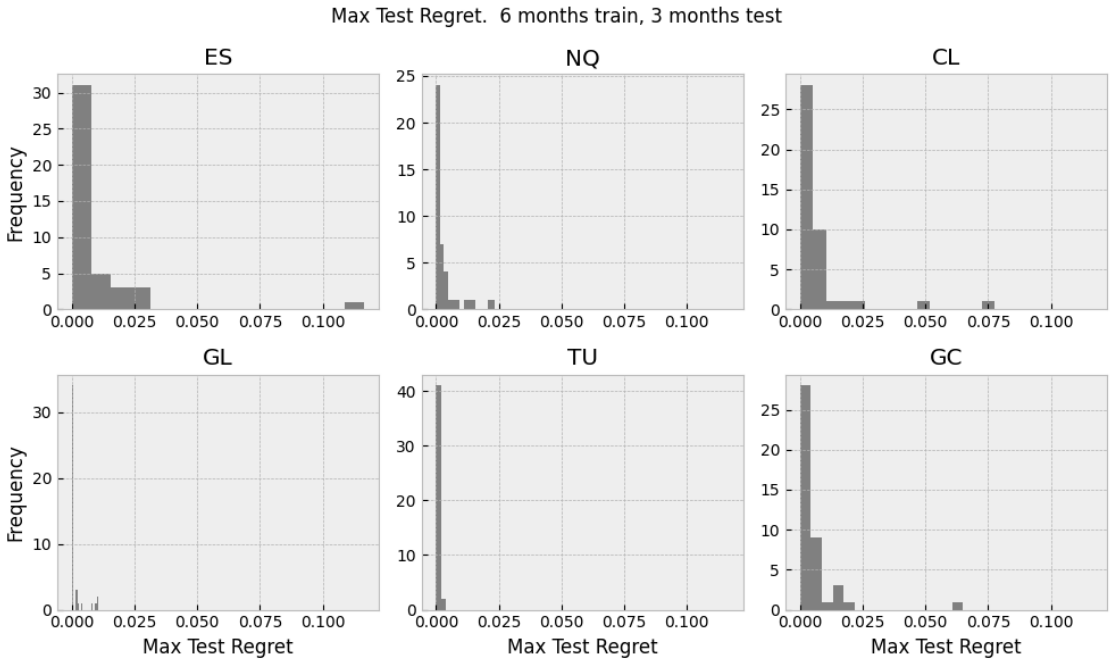


Figure 4.6: Max Test Regret using 6 months training and 3 months testing data.

4.4 Cross-Sectional λ Analysis

In [13, Table 3], it is suggested that all stocks in the S&P 500 could share a 'universal' price impact coefficient, and out-of-sample R^2 were found to be the same if not better than fitting a λ for each specific stock. Based on evidence that we have seen in Figure 4.3, it is unclear whether this could be true for the futures market. Each future plotted is for a different value of δ, so the λ are not always comparable, but TU, GL and ES all have $\delta = 0.1$ and the values of λ vary greatly between the futures. In Figure 4.7, we show a similar plot but with all futures having the same $\delta = 0.3$ allowing for easier comparison. One could make the case that the two commodities CL and GC share the same coefficient of impact, possibly due to the fact they are both commodities. However, just because two futures are in the same asset class, it does not mean they must have a similar coefficient of impact. GL and TU are both government bond futures, but their values of λ are greatly different. Whilst GL futures mostly trade during UK market open and TU futures trade for 23 hours a day, it is only fair to compare the λ during the US market open hours of TU. Even taking this into account, the US market open values of λ range between 15 and 25 for TU, but for GL the values are all between 8 and 15. For the equity index futures ES and NQ, there is also evidence that the two futures do not share a common λ, as during the US market open hours λ_{ES} ranges between 5 and 10, whereas λ_{NQ} frequently sees values of 10 to 16.

Whilst we do not find evidence against the results found in Muhle-Karbe et al. [13], we do find empirical evidence that this result *does not* extend to all futures - even those in the same asset class.

4.4.1 Interest Rate Futures

In Section 4.2.2 we observed the changes in market impact for some interest rate futures and speculated that this could be due to changes in central bank policy. Here we investigate this hypothesis further by expanding the analysis to a broader range of tenures and futures.

In Figure 4.8 we plot the average intraday λ for different interest rate futures from the UK, Eurozone, and US using $\delta = 0.1$. This is plotted for each backtest window using a one month training period, and in general there is little change in the value of λ over time, which can be seen especially in UK, EU 2/5 year and US 5/10/30 year futures. There also appears to be an inverse relationship between tenure and λ, which can be seen in both US and Euro futures. This implies that the price of shorter dated interest rate futures are more sensitive to trades compared to their longer dated counterparts. If an investor is looking to get a large exposure to the fixed income

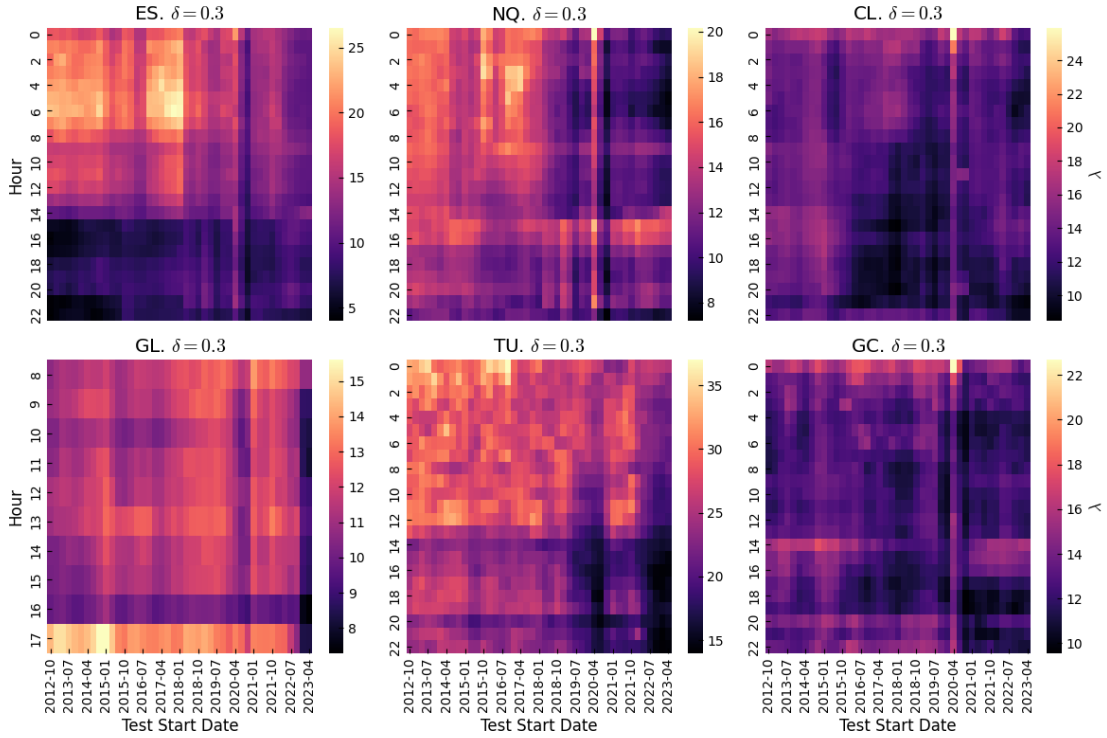


Figure 4.7: Coefficient of impact over time with fixed δ for 6 futures using 6 months training and 3 months testing data.

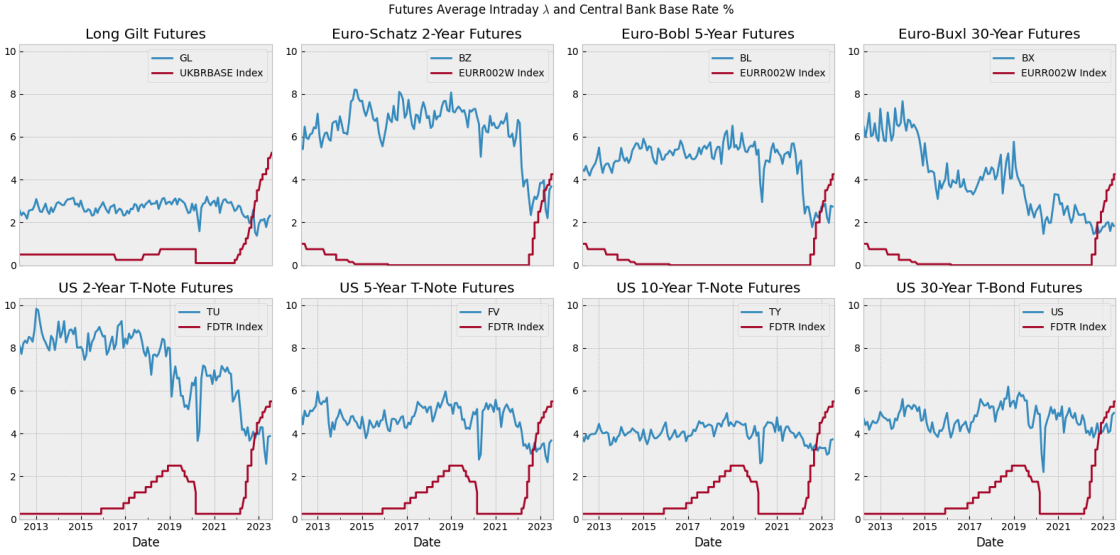


Figure 4.8: Average intraday λ for interest rate futures with $\delta = 0.1$, 1 month training data.

market, opting to trade longer dated futures may be more advantageous as Figure 4.8 implies their trades would have less market impact.

In Figure 4.8 we also plot the central bank interest rate for the Federal Reserve, Bank of England and European Central Bank. For interest rate futures with a tenure of less than 30 years, we see that λ is extremely sensitive to sharp changes in interest rates. This can be seen most prominently in 2022 when central banks began to raise interest rates in an attempt to reduce high inflation, and during this time the value of λ decreased dramatically. During this period, the price of bonds could have been dictated to a greater extent by investor expectations of central bank rate changes, and to a lesser extent by an imbalance of supply and demand. Whilst we also see a drop

in λ in early 2021, it is not clear if this is caused by the sharp reduction in central bank interest rates (UK & US), because a similar drop is seen in all Euro futures where there was no change in interest rates.

4.5 Recalibration Optimisation

Our analysis thus far has mainly consisted of using a training period of 6 months and a testing period of 3 months. This decision was largely arbitrary, and in Figure 4.5 we saw examples of some backtest windows having much worse performance in the testing period compared to the training period, which implies that the data in the testing window has little resemblance to that in the training. In an attempt to mitigate these examples of poor out-of-sample performance we can reduce the length of both test and train windows.

Train months	Test months	Avg Train R^2	Avg Test R^2	Test Train Ratio	Test Regret
2	1	0.6562	0.6504	0.9911	0.0029
1	1	0.6646	0.6495	0.9774	0.0028
4	2	0.6481	0.6439	0.9936	0.0040
6	3	0.6426	0.6363	0.9903	0.0042

Table 4.2: Comparison of model fit for varying lengths of training and testing.

In Table 4.2 we compare a number of metrics for the best model (as defined in Section 4.3) averaged over all the futures and backtest windows. It can be seen that decreasing the train/test period from (6, 3) allows for better in and out-of-sample performance, with $(2, 1)$ giving the best test R^2. We also reduce the training window further by attempting (1, 1), and whilst this increases the in-sample R^2, the out-of-sample performance is worse. We also find the 'Test Regret' to be smallest for (2, 1) and (1, 1) at a value of 0.003, meaning the model selection method of using the best train R^2 for the test period is on average 0.003 R^2 worse than the theoretically best performing model.

Chapter 5

Optimal Execution

Returning to the problem of optimal execution, we aim to find the optimal trading schedule for a trader wanting to buy a given volume X in a specific period of time. To do this we need to understand how price reacts to volume traded, and to both formalise and calibrate a model to market data. Assuming linear market impact, this problem has been solved under a number of different frameworks. Under the continuous transient impact model with power law decay, Curato et al. [30, Equation 2.8] derive an explicit formula for the optimal trading speed. Assuming exponential decay, Obizhaeva & Wang [6, Proposition 2] also solve for the optimal execution plan, with the continuous solution covered in Section 1.3. Both these solutions have similar characteristics with higher speed of trading at the beginning and end of the trading session, and a symmetric trading schedule.

Attempts have been made to further these models by including nonlinear market impact but, as covered in Section 1.5, introducing nonlinear impact will yield a solution that induces price manipulation. Dang [31, Section 5] proposed a numerical scheme and method for finding the optimal trading schedule, but the absence of price manipulation was dependent on the parameters used in both the decay and impact function. Curato et al. [30, Section 2.3, 3] also explored a number of methods to derive an optimal trading schedule: a perturbative approach where the linear impact solution was perturbed to find a solution for a slightly concave impact function ($f(v) = v^{1-\epsilon}, 0 < \epsilon \ll 1$), a 'Homotopy Analysis Method' where an initial guess is iteratively deformed, and a numerical cost optimisation method using a Sequential Quadratic Programming (SQP) algorithm. In Figure 5.1a we can observe the effect that nonlinear market impact has on the optimal trading schedule compared to linear impact. Whilst the linear solution is symmetrical, the nonlinear solution shows asymmetry with faster trading at the beginning of the interval followed by comparatively slower trading in the middle. There is a final faster burst at the end, which is slower than the initial nonlinear speed.

SQP is a framework for solving nonlinear constrained optimisation problems using an initial guess and an iterative scheme. With this method, convergence to local optimal values is guaranteed providing they exist, but not convergence to global solutions. We do not go into full detail on the theory of SQP, but we direct the reader to [32] for further detail. The implementation of SQP that we use is from the python package 'SciPy', which implements Sequential Least SQuares Programming (SLSQP) [33].

5.1 Optimal Execution for DPM

5.1.1 Problem Definition

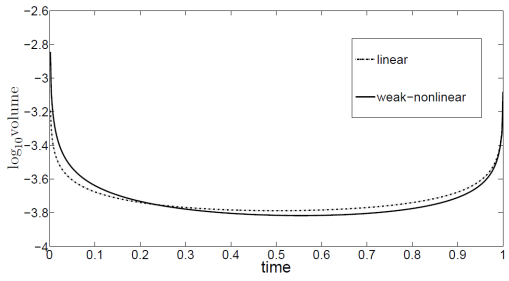
We assume the price of an asset follows the Discrete Propagator Model (Equation 1.6.1), and the trader has a mandate to buy X units of a security by time T. We look to minimise the excess cost of the trade which is equivalent to:

$$\begin{aligned} \min_{\Delta Q_n} \quad & \mathbb{E}\left[\sum_{n=1}^{N} I_n \Delta Q_n\right] \\ \text{s.t.} \quad & \sum_{n=1}^{N} \Delta Q_n = X \\ \text{where} \quad & I_n = (1-\beta\Delta t) I_{n-1} + \sigma\lambda_n \text{sign}\left(\Delta Q_n\right) |\Delta Q_n|^{\delta} \end{aligned} \tag{5.1.1}$$

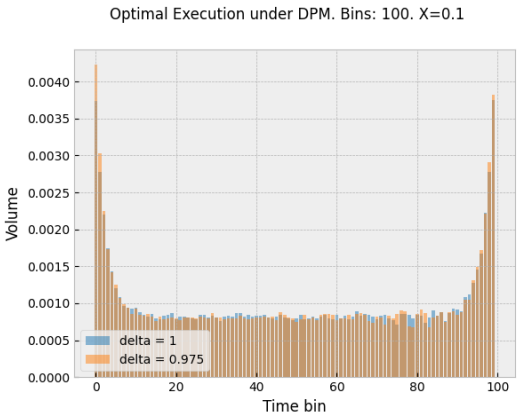
Note, we remove references to ADV and assume for this chapter that all volumes are quoted as a proportion of their average daily volume. We also add a subscript of n to λ to express its dependence on the time of day. Minimising the expected excess cost is equivalent to minimising the total cost, because $\mathbb{E}\left[\sum_{n=1}^{N} I_n \Delta Q_n\right] = \mathbb{E}\left[\sum_{n=1}^{N} \mathbb{E}\left[S_n/S_0\right] \Delta Q_n\right] = \mathbb{E}\left[\sum_{n=1}^{N} S_n \Delta Q_n\right]/S_0$, and so the two optimisation problems are equivalent up to the scaling constant S_0.

5.1.2 Slightly Concave Impact

To our knowledge there are no papers exploring optimal execution using the DPM with nonlinear price impact and, therefore, no equivalent benchmark to which we can reference our results. However, we can look to the results found in [30] due to the similarities between the continuous TIM and the DPM. We compute using SQP the optimal volume schedule for both linear and slightly concave instantaneous market impact in Figure 5.1b. We can see that despite the differences between the models, there are many shared characteristics between the optimal execution schedules. For linear market impact, both solutions show a symmetrical output with high trading speeds at the beginning and end of the trading and a steady speed in the middle. For slightly concave impact, the DPM trades faster than its linear counterpart at the beginning, followed by slower trading in the middle of the session, and finally a fast burst at the end but still slower than the opening speed. Whilst the two models are different, they both attempt to capture the nonlinear and transient nature of market impact, and so it is encouraging that the optimal volume schedules are similar for slightly concave market impact.



(a) Optimal volume schedule for TIM with linear and slightly concave impact under power law decay [30, Figure 2].



(b) Optimal volume schedule under DPM.

Figure 5.1: Optimal volume schedules for linear and slightly concave under TIM and DPM.

5.2 Concave Impact

In this section we explore the different output and limitations of using SLSQP under the DPM when varying a number of the input parameters. To best align this analysis with the intention of using the solutions for real world trading, we will use parameters similar to those we have found during the calibration in Section 4. We chose to model the future NQ using $\sigma = 0.0002$, and for each respective δ we will use the average intraday λ that has been fit via regression.

5.2.1 Delta

Whilst Figure 5.1b shows sensible convergence for slightly concave impact ($\delta = 0.975$), there is strong evidence to suggest the convergence of such optimisation problems and absence of price manipulation depends heavily on the non linearity of impact and the speed of impact decay. In Figure 5.2 we plot the optimal volume schedule for $\delta \in \{0.1, 0.3, 0.5, 0.7\}$.

For all values of δ the optimal strategy involves both buy and sell orders, exhibiting price manipulation by using sell orders to force the price down. For $\delta = 0.7$ there are still remnants of some structure, similar to that in Figure 5.1b, with stronger buying at the beginning of the trading session. For $\delta = 0.5$, there is a clear structure to the trading strategy with periodic bursts of buying with small but increasing sell orders between, similar to that of [30] (Figure 6). Both $\delta \in \{0.3, 0.1\}$ show no real signs of structure with individual trades being larger than the intended total volume of 0.1. Clearly none of these 'optimal' trading strategies are sensible for real world trading, and the solutions display some of the limitations and drawbacks of both the DPM and the method of solving for optimal solutions via SQP.

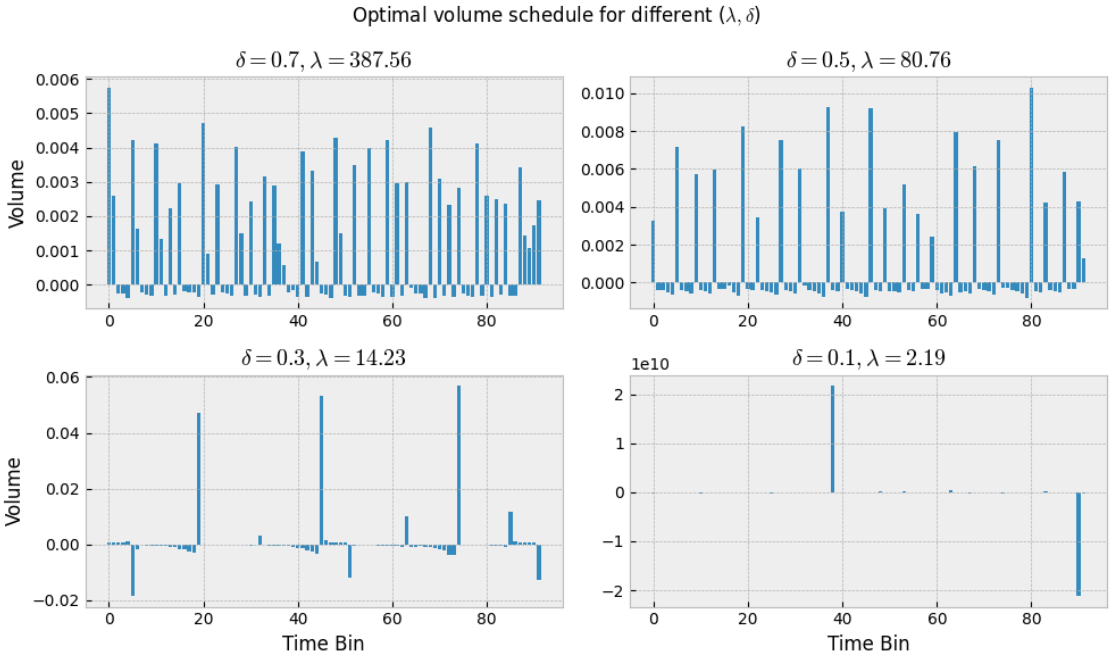


Figure 5.2: Optimal Execution for a range of δ.

5.2.2 Bid-Ask Spread

One unrealistic aspect of the model is the assumption that there is one price at which traders can both buy and sell at. In reality there is a gap in the LOB separating the best and best ask limit orders, called the bid-ask spread. A trader who buys then immediately sells an asset will make a loss equal to the bid-ask spread, and including this in the model could help to penalise the selling seen in Figure 5.2. To integrate this feature into the model, we can add a cost to the optimisation problem (Equation 5.1.1) when the trader executes an order, regardless of the direction. This idea has been implemented in a number of papers [30, Section 5.3][16, Section 6.2], and equates to adding a Lasso penalty to the optimisation problem

$$\begin{aligned} \min_{\Delta Q_n} \quad & \mathbb{E}\left[\sum_{n=1}^{N} I_n \Delta Q_n\right] + \frac{s}{2}\sum_{n=1}^{N} |\Delta Q_n| \\ \text{s.t.} \quad & \sum_{n=1}^{N} \Delta Q_n = X, \end{aligned} \tag{5.2.1}$$

where s is the bid-ask spread. In Figure 5.3 we display the effect of adding spread cost on the optimal volume schedule, where all plots use the same λ and δ of 0.5 and 80.76 respectively. The

0bps plot is identical to that in Figure 5.2 with periodic large buy orders followed by longer periods of weak selling. Adding a 10bps spread reduces the magnitude of the sell orders as we had hoped, but does not completely eliminate them. Adding 25 and 50bps further reduces the sell orders, and 100bps spread looks as if it completely eliminated the problem. However, in Table 5.1 we display the negative volume (ie sell orders) for each execution plan, and there are still some negative trades with 100bps of spread. One option when using the SLSQP solver is to restrict the solution space to only positive values, which is a strategy used in [30, Section 4.3.3], and we display this solution in the final plot of Figure 5.3. The solution involves large buy orders at the start of the trading session, followed by semi-frequent buy orders and ending with a period of consistent buying.

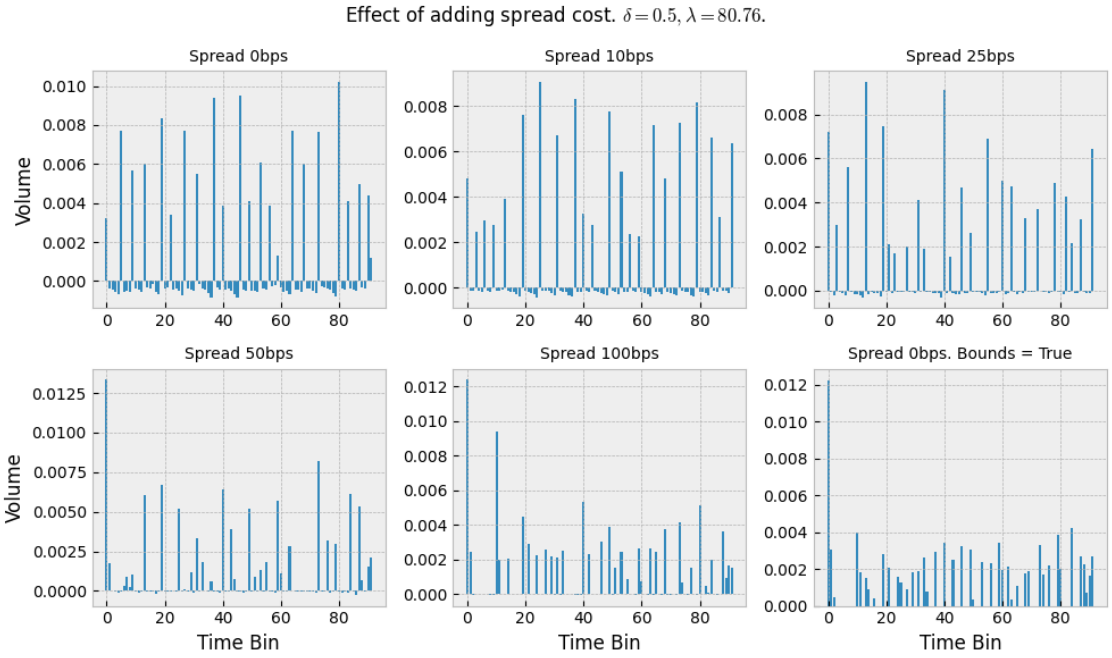


Figure 5.3: Optimal Execution for a range of spreads.

Whilst adding a spread cost is useful in reducing the amount of negative trading, it does not completely eliminate the problem. It requires a 100bps spread to reduce the amount of negative trading to roughly 1% of the total traded volume, but is this amount of spread realistic for such a liquid instrument such as NQ futures? At the time of writing, the most active NQ contract trades with a bid-ask spread of one tick size, which equates to 0.17bps; this is much less than the values used for Figure 5.3.

Adding a bid-ask spread makes the model more realistic to real life trading, but the constant spread is somewhat unrealistic. After submitting a market buy order only the offer limit orders will have been matched causing an increase in ask price but leaving the bid price unchanged. After this, market makers can quickly react by increasing their best bid price, replenishing offer market orders, or a combination of the two. Regardless, the spread will temporarily increase - a feature that is not represented in the optimisation problem 5.2.1. In fact [1, Figure 4.2] shows that the bid-ask spread for a number of stocks exhibits a pattern of higher spread at the beginning of the trading day and lower towards the end.

5.2.3 Trading Frequency

Thus far we have used a trading frequency (Δt) of 15 minutes. As $\Delta t \to 0$ we would hope that the optimal solution would tend also to a finite limit, as opposed to the individual trades 'blowing up' to infinity. In Figure 5.4 we can observe the effect of the trading frequency on the optimal solution from a range of 5 to 60 minutes. The solution for 60 minutes looks to have similar characteristics to the slightly concave example in Figure 5.1b, with 45 minutes being a more pronounced version of this. However at 30 minutes we see the negative trades appearing in the optimal solution, inducing price manipulation, and 20 and 10 minutes take this further with more structured patterns in the

Spread (bps)	Negative Volume
0	-0.031966
10	-0.015618
25	-0.007091
50	-0.002847
100	-0.000819

Table 5.1: Total negative volume in the optimal execution schedule.

selling periods. At the 5 minute trading frequency, we see the optimal solution behaving very erratically with buy and sell orders multiple times larger than the intended total volume traded of 0.1.

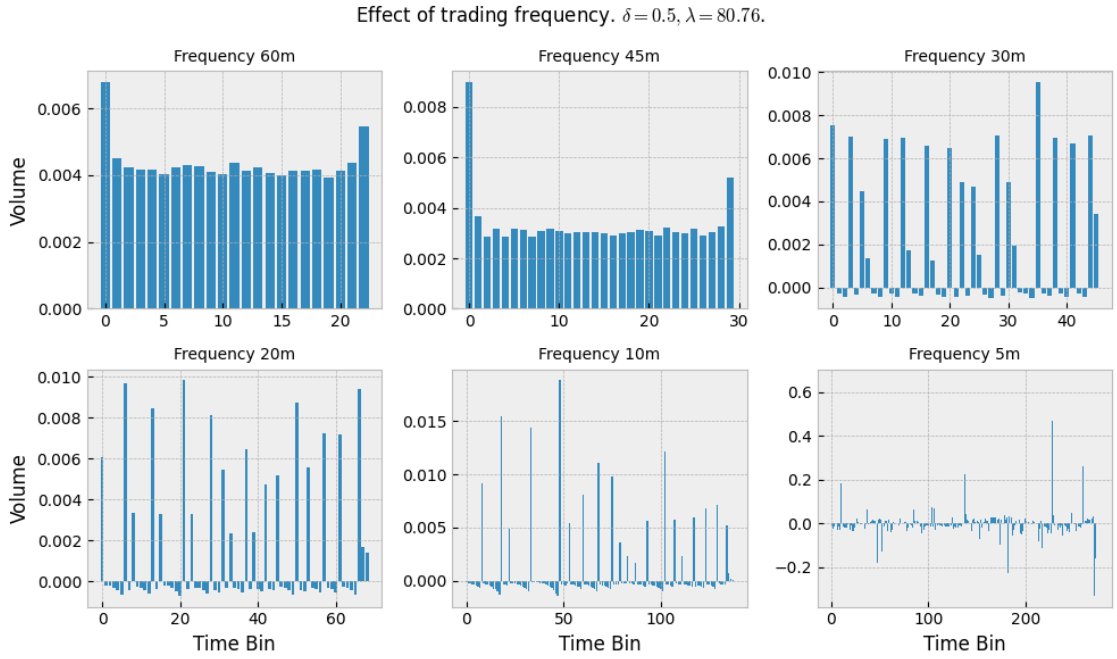


Figure 5.4: Optimal Execution for a range of frequencies.

It is worth questioning at this point the practical limit of trading speed for buy side institutions using this model for trading purposes. We do not claim that this model would be optimal at the high frequency trading limit, where both market and limit orders could be dynamically submitted at the millisecond frequency. The data used to train the model has been 10 second binned data for both price and volume, a dataset of which any institution running intraday strategies could already have in house. Institutions executing these strategies may not have the ability to submit orders electronically, and could still rely on phone conversations with brokers and dealers making the idea of communicating orders every 5 minutes impractical and unrealistic. With this in mind, we argue that using a value for Δt between 10 and 30 minutes could be realistic value for the frequency of order execution for some traders.

5.2.4 Tolerance

SLSQP and other minimisation algorithms often work by evaluating the function to optimise $f(x_n)$ and then trying to find a nearby x_{n+1} such that $f(x_{n+1}) < f(x_n)$. Whilst this process could continue forever as the values of x_n slowly converge to a local minimum, there reaches a point where the improvement gained from the next iteration is of little practical use. One common method to counteract this is to use a 'tolerance' ϵ such that the minimisation algorithm ends when $|f(x_{n+1}) - f(x_n)| < \epsilon$, and often this can be defined by the user. SciPy's *minimize* function

provides this option, and so we compute the optimal volume schedule for a range of tolerances in Figure 5.5.

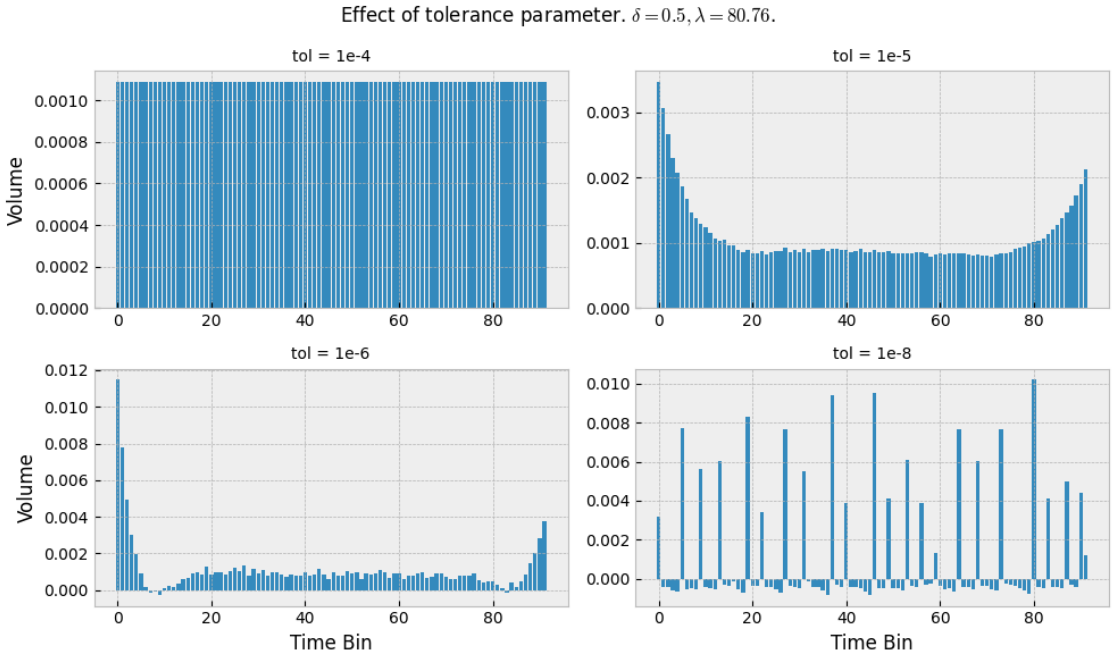


Figure 5.5: Optimal Execution for a range of tolerances.

For a tolerance of 10^{-4} the optimiser does not even take one step in search for a better function value, presumably because the relative decrease in cost is so small in comparison to the tolerance. The resultant volume schedule is therefore the initial guess inputted into the optimiser, which in our case is a TWAP strategy. Reducing the tolerance by an order of magnitude, the optimal execution for a tolerance of 10^{-5} is a more extreme version of Figure 5.1b, with lower trading speeds after the initial and before the last burst of trading. It is worth noting the kinks in this solution are somewhat similar to the optimal volume schedule found in [16, Figure 5], which uses a slightly different discrete propagator model with the addition of linear market impact and a risk aversion parameter. 10^{-6} takes this a step further with the slower trading speed from before becoming negative, and the solution for a tolerance of 10^{-8} is periodic buying followed by periods of weak selling. Tolerance values outside this range were also tried, but values above 10^{-4} all resulted in the TWAP strategy, and values below 10^{-8} showed no real change to that of the schedule for 10^{-8}.

Whilst the strategy of restricting the tolerance to yield sensible results from the optimiser is clearly effective, it is worth asking whether this is a reasonable approach or just numeric trickery. In [30, Section 2.3], Curato et al. use a perturbative approach by slightly deforming an explicit optimal solution for linear market impact (GSS) to find a solution for slightly concave impact, which is shown in Figure 5.1b. They also use a homotopy approach, where a final trading speed is found by the sum of many 'homotopy derivatives' using the GSS solution as initial input. To implement this, the 7th order homotopy derivative was used, but this is a somewhat arbitrary choice and any number of derivatives could have been used to further deform the initial guess. Both of these methods are examples of beginning with an initial guess of the solution, and modifying it up to a somewhat arbitrarily point. This places restrictions on the type output that the final optimal output can have, and we argue that increasing the tolerance in the SLSQP function can achieve the same objective of restricting the output to 'sensible' forms.

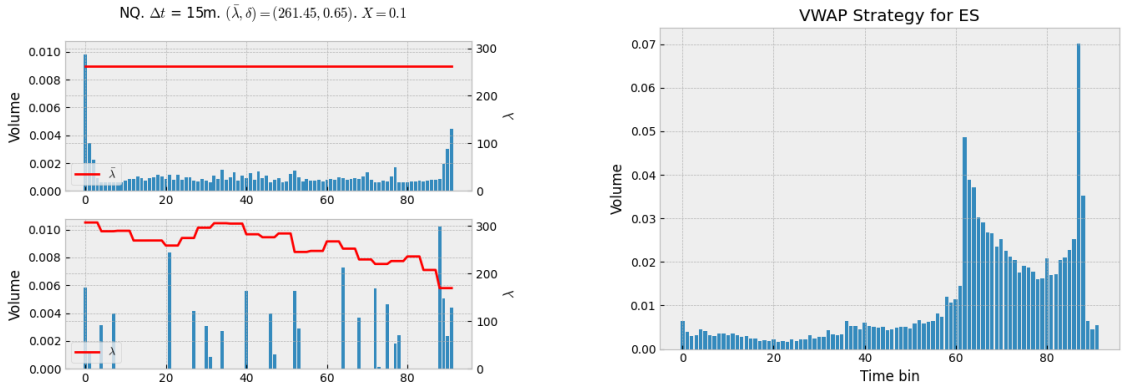
5.2.5 Summary

In this subsection we have thoroughly explored the limits of using SLSQP to solve for optimal volume schedules using the DPM. We have seen similarities in the trading output to other studies [30, 16] for both the 'sensible' and 'non sensible' outputs, despite the differences in the modelling approach. We have found a number of effective ways to restrict the solver to yield results that

could be acted upon by traders, and to avoid the price manipulation involving negative sells in a buy order. These methods include changing the frequency of trading, adding a spread cost to the model, and limiting the tolerance of the solver, all of which we have argued to be reasonable methods.

5.3 Intraday λ

Until now, we have studied the optimal execution problem assuming that the coefficient of impact is constant throughout the trading day. In Section 4 we found strong evidence against this assumption, showing intraday patterns in market impact that persisted throughout time. In this section we incorporate this feature into the optimisation problem and analyse the potential cost savings compared to other industry benchmarks. To achieve this we use a combination of tolerance limitation to give sensible trading output, and a reduction in the solution space to positive values only to avoid price manipulation.



(a) Optimal Execution with intraday λ.

(b) Example VWAP strategy for ES.

Figure 5.6: Optimal Execution with intraday λ & example VWAP schedule.

In Figure 5.6a we plot the optimal execution for fit values of (λ, δ) the future NQ. We use a constant λ throughout the day by averaging the intraday values. The optimal execution schedule has the familiar characteristics of asymmetry and kinks near the beginning and end of the trading session. We also plot the optimal execution schedule after accounting for the intraday changes in the coefficient of impact, which results in a much different pattern. The new schedule often avoids trading during periods of relatively high impact, opting to trade more both before and after the high impact period. A lot of volume is traded at the end of the session when the impact is lowest. This period also has the added benefit of removing the need to trade immediately after the price has been driven higher. Whilst there is trading at the beginning of the session, the activity is much smaller compared to the baseline due to the the beginning of the session having the highest impact in the day.

5.3.1 Cost Saving

One of the aims of this thesis is to find a volume schedule that minimises a trader's cost of execution, and so it is important that we can exactly quantify the costs incurred for a given volume schedule. Ideally we would like to find a volume schedule that materially affects the cost incurred compared to a standard benchmark, and saves sufficient money to make the cost of calibrating and retraining the DPM worthwhile. To evaluate this, we will introduce a number of industry benchmarks that can be used to execute a large order over a given period of time.

- TWAP - A Time Weighted Average Price (TWAP) strategy involves splitting the meta-order up into identically sized child orders which are executed linearly throughout time. This is the strategy we have used as our 'initial guess' thus far, and can be seen in the first plot of Figure 5.5.

- VWAP - A Volume Weighted Average Price (VWAP) strategy involves splitting the meta-order up into child orders which mirror the shape of the expected gross volume traded in each time bin. In Figure 5.6b we show an example VWAP strategy for ES using 15 minute trading frequency, where our estimate for the volume traded in each bin is the average of the gross volume traded for that bin for the previous 2 months.
- Market Open - A Market Open strategy is used by traders with a time urgency for the execution of the meta-order. This strategy involves executing the entire meta-order at the start of the exchange's trading session, and so for our purpose the whole volume will be traded in the first 15 minute bin.

In Section 4 we found strong evidence for λ changing throughout the day, a feature which models do not often include. We would like to understand the cost of assuming λ is constant throughout the day, and to do this we also consider the optimal execution plan assuming constant λ. The λ we use is the average of intraday values found, and whilst this may not be the exact value of λ that would be fit if using all the data, it should not matter as the optimisation problem (5.1.1) will just be scaled by a constant, and therefore the optimal schedule would not change.

Spot	X	δ	$\bar{\lambda}$	TWAP	VWAP	Market Open	Constant	Intraday
TU	0.025	0.1	6.05	0.0419	0.0424	0.0126	0.0125	0.0125
TU	0.05	0.1	6.05	0.0898	0.0908	0.027	0.0266	0.0261
TU	0.1	0.1	6.05	0.1924	0.1947	0.0579	0.0579	0.0566
TU	0.25	0.1	6.05	0.5273	0.5335	0.1587	0.1586	0.1575
GC	0.025	0.45	44.41	0.1412	0.1615	0.2833	0.1211	0.1109
GC	0.05	0.45	44.41	0.3859	0.4412	0.774	0.3799	0.3488
GC	0.1	0.45	44.41	1.0542	1.2055	2.1145	1.1134	0.7534
GC	0.25	0.45	44.41	3.9806	4.5516	7.9842	3.7857	3.3708
NQ	0.025	0.65	256.8	0.1722	0.2014	0.6692	0.1366	0.1315
NQ	0.05	0.65	256.8	0.5404	0.632	2.1002	0.4392	0.4114
NQ	0.1	0.65	256.8	1.6959	1.9835	6.5912	1.4446	1.2962
NQ	0.25	0.65	256.8	7.6911	8.9955	29.8927	7.3162	5.8822

Table 5.2: Cost of different execution strategies (bps).

We display the cost of different strategies in Table 5.2, modelling the execution of different volume sizes $X \in \{0.025, 0.05, 0.1, 0.25\}$ using the best fit of (λ, δ) found previously for TU, GC, and NQ. These futures allow us to see the effect of different benchmarks for a large range of deltas. One thing that can be noticed immediately is the superior performance of a TWAP strategy compared to a VWAP strategy, a result which could be due to the VWAP strategy trading too much in the high impact periods of US spot open, and not enough in the overnight hours. We see that the cost of a single strategy increases as the total volume X increases, a result that should be expected as more volume traded will cause more market impact. Whist the cost of VWAP and TWAP are often similar, the Market Open strategy is often much larger or much smaller depending on the value of δ. For $\delta = 0.1$ the Market Open strategy is substantially better due to the strong concavity of instantaneous impact penalising small but positive trades, whereas for $\delta = 0.65$ the Market open strategy consistently causes 3 times more impact cost.

For $\delta = 0.1$, the optimal schedules found are only slightly better than following the Market Open strategy. In fact the optimal volume schedules found in this case trade almost all the total volume in the first time bin, and then the remaining amount near the end of the trading session.

Table 5.3 shows for each benchmark the relative savings of using the 'optimal' strategy compared to using that benchmark, with negative savings indicating the optimiser output did not converge to a solution better than the benchmark. For extremely low values of δ we see that the optimal solution is not much better than a Market Open strategy, and in some cases a Market Open strategy may be close to the true optimal strategy. For higher values of δ, a better benchmark to compare to is the TWAP strategy, where the optimal solution is often 20%+ better than the benchmark.

We can also see the effectiveness of including intraday λ into the optimisation problem, with the intraday optimal solution outperforming the constant $\bar{\lambda}$ solution by up to 32.33%. This is a

Spot	X	δ	$\bar{\lambda}$	TWAP	VWAP	Market Open	Constant
TU	0.025	0.1	6.05	70.22%	70.57%	1.03%	0.4%
TU	0.05	0.1	6.05	70.91%	71.25%	3.31%	1.89%
TU	0.1	0.1	6.05	70.61%	70.96%	2.33%	2.3%
TU	0.25	0.1	6.05	70.12%	70.47%	0.71%	0.69%
GC	0.025	0.45	44.41	21.5%	31.34%	60.86%	8.45%
GC	0.05	0.45	44.41	9.61%	20.95%	54.94%	8.19%
GC	0.1	0.45	44.41	28.53%	37.5%	64.37%	32.33%
GC	0.25	0.45	44.41	15.32%	25.94%	57.78%	10.96%
NQ	0.025	0.65	256.8	23.63%	34.7%	80.35%	3.74%
NQ	0.05	0.65	256.8	23.87%	34.91%	80.41%	6.33%
NQ	0.1	0.65	256.8	23.57%	34.65%	80.33%	10.27%
NQ	0.25	0.65	256.8	23.52%	34.61%	80.32%	19.6%

Table 5.3: Savings from using the optimal strategy compared to benchmarks.

substantial amount of cost saving, and for funds with a high turnover who are fully entering and exiting large positions on a regular basis, this cost saving can compound over time to a relatively meaningful amount.

5.3.2 Summary

Using calibrated real world values to parameterise our market impact model, we have found the optimal trading strategy in a number of different environments. In all but one case, we have found that using the SLSQP output using intraday λ can provide costs savings to standard benchmarks. We have quantified this cost saving, finding 80%+ savings compared to some benchmarks and also found that including intraday λ saves on average 12.15% for higher δ values, compared to assuming constant λ. We have also explored under which conditions different industry benchmarks are better to use, finding the Market Open strategy to be better for lower values of δ, and the TWAP strategy to outperform for higher δ. This can be a useful output alone as these benchmarks can be easily communicated over the phone to brokers and can provide savings to institutions without needing to trade a more complex SLSQP volume schedule.

Conclusion

In this thesis, we explored two vital aspects of financial markets: price impact modelling and optimal execution. After reviewing the empirical characteristics of nonlinear instantaneous impact, intraday liquidity changes, and impact decay, we calibrated the Discrete Propagator Model on futures data using aggregated tick level trade data, covering asset classes such as Equity Indices, Commodities, Interest Rates and Currencies.

We calibrated the model using sliding windows over a period of 10 years and found the DPM was effective at explaining price changes due to traded volume, achieving average in sample R^2 of up to 0.795. Strong evidence was found for intraday liquidity changes with a high coefficient of impact in overnight hours, and examples of lower impact in spot trading hours falling to a minimum at the end of the spot trading session. Whilst there were some comparable characteristics seen between futures within the same asset class, we found significant differences between others such as Interest Rate futures exhibiting highly concave ($\delta = 0.1$) instantaneous impact whereas other asset classes were generally much less concave. It is evident that a universal model (as suggested in previous literature for single stocks) would not be appropriate for all futures.

We provided novel analysis of the stability of this calibration throughout time, showing common occurrences of the scale of impact changing over time. There were also instances of the intraday impact profile changing over time, with S&P E-mini futures showing a steep relative reduction in overnight impact in 2018. We also found the concavity of instantaneous impact to change throughout time, with the changes best fitting δ indicating regime changes in the market microstructure, suggesting changes to the general shape of the LOB. We provided an in-depth analysis into the change in impact throughout time in the Interest Rate futures asset class, revealing that impact was negatively correlated with bond tenure and showing the effect of central bank interest rate hikes on the respective bond futures. The effect of the pandemic crash on impact was also observed, with less impact during this time for Interest Rate futures and higher impact for Equity Indices and Commodities. To the best of my knowledge, the cross-sectional analysis of price impact between futures of different asset classes, and the study of calibration changes over a decade, is a novel contribution to the field.

Given the observed changes in impact throughout time, we looked to understand the optimal train and test window length. We found that using a training period of two months was the optimal balance between having a large training dataset and having a more recent and relevant training set. Using this training length, a one month testing period, and a one month sliding window, we were able to attain an average in and out-of-sample R^2 of 0.6562 and 0.6504 respectively.

Using calibrated values of λ and δ, we explored the problem of optimal execution within the framework of the DPM. Our findings closely paralleled those of earlier literature, albeit with the application of different models. Notably, some instances revealed that the optimal trading schedule could lead to price manipulation. We thoroughly investigated the sensitivity of input parameters to the output schedule, and proposed effective solutions to remove price manipulation including adding a bid-ask spread, reducing the solver tolerance and trading frequency, and restricting the solver to find solutions only over the set of positive values. We quantified the impact of including intraday liquidity dynamics in the optimal strategy on execution costs, comparing to common industry benchmarks of TWAP, VWAP and Market Open. Using the optimal execution strategy showed significant savings, with an average of 21% cost reduction compared to industry benchmarks for higher values of δ. We found savings of 8.76% on average compared to assuming a constant intraday impact coefficient, and showed that different benchmarks were better depending on the concavity of impact.

Our insights hold significance for refining execution strategies, especially for institutions with limited access to LOB data or exchange connectivity, and high turnover trading strategies.

Appendix A

Futures

Table A.1 provides a lookup between the expiration rule used for futures contract expiration date in Table A.3. Table A.2 provides a lookup between the month key used in futures tickers (ie. ESU3).

Expiration Rule	Rule Key
third friday	0
last business day	1

Table A.1: Expiration rules.

Month Key	Month
F	January
G	February
H	March
J	April
K	May
M	June
N	July
Q	August
U	September
V	October
X	November
Z	December

Table A.2: Month key lookup.

Table A.3 proves further detail into the futures we cover in the analysis in this thesis. The columns 'Expiry Rule', 'Active Futures Month', and 'Roll Lookback Weeks' are used to establish how multiple futures are used to create one synthetic future. A contract is used in the synthetic future if the expiration month is in 'Active Futures Month', in the time period prior to the expiry date (as defined by the expiry rule) less the number of roll lookback weeks.

Table A.3: Futures used.

tick code	bbg ticker	future months	roll lookback weeks	expiry rule	description	asset class
AD	AD Curncy	HMUZ	1.0	0	Australian Dollar Futures	Currency Futures
BP	BP Curncy	HMUZ	1.0	0	British Pound Futures	Currency Futures
CD	CD Curncy	HMUZ	1.0	0	Canadian Dollar Futures	Currency Futures
DX	DX Curncy	HMUZ	1.0	0	Dollar Index Futures ICE	Currency Futures
EC	EC Curncy	HMUZ	1.0	0	Euro FX Futures	Currency Futures
JY	JY Curncy	HMUZ	1.0	0	Japanese Yen Futures	Currency Futures
NZ	NV Curncy	HMUZ	1.0	0	New Zealand Dollar Futures	Currency Futures
SF	SF Curncy	HMUZ	1.0	0	Swiss Franc Futures	Currency Futures
CL	CL Comdty	FGHJKM NQUVXZ	4.5	0	Light Crude Oil Futures NYMEX	Energy Futures
CO	CO Comdty	FGHJKM NQUVXZ	8.0	0	Brent Crude Oil Futures ICE	Energy Futures
GO	QS Comdty	FGHJKM NQUVXZ	4.0	0	Low Sulphur Gasoil Futures ICE	Energy Futures
HO	HO Comdty	FGHJKM NQUVXZ	4.5	0	Heating Oil #2 Futures NYMEX	Energy Futures
NG	NG Comdty	FGHJKM NQUVXZ	4.0	0	Natural Gas Futures NYMEX	Energy Futures
XB	XB Comdty	FGHJKM NQUVXZ	4.5	0	RBOB Gasoline Futures NYMEX	Energy Futures
EI	MES Index	HMUZ	1.0	0	MSCI Emerging Markets Mini Futures	Equity Index Futures
EN	NI Index	HMUZ	1.5	0	Nikkei 225 Futures SGX	Equity Index Futures
ES	ES Index	HMUZ	1.0	0	S&P 500 E-Mini Futures	Equity Index Futures
FT	Z Index	HMUZ	1.0	0	FTSE 100 Index Futures	Equity Index Futures
HI	HI Index	FGHJKM NQUVXZ	0.5	1	Hang Seng Index Futures	Equity Index Futures
MI	FA Index	HMUZ	1.0	0	S&P 400 MidCap E-Mini Futures	Equity Index Futures
NE	NK Index	HMUZ	1.5	0	Nikkei 225 Futures JPX	Equity Index Futures
NK	NX Index	HMUZ	1.0	0	Nikkei 225 Futures CME	Equity Index Futures
NQ	NQ Index	HMUZ	1.0	0	NASDAQ 100 E-Mini Futures	Equity Index Futures
NY	NH Index	HMUZ	1.0	0	Nikkei 225 Futures Yen-Denominated CME	Equity Index Futures
PT	PT Index	HMUZ	1.0	0	S&P Canada 60 Futures	Equity Index Futures
SW	SM Index	HMUZ	0.0	0	Swiss Market Index Futures	Equity Index Futures
TP	TP Index	HMUZ	1.5	0	TOPIX Futures JPX	Equity Index Futures
VX	UX Index	FGHJKM NQUVXZ	5.0	0	VIX Futures	Equity Index Futures
XP	XP Index	HMUZ	0.5	0	ASX SPI 200 Index Futures	Equity Index Futures
XX	VG Index	HMUZ	0.0	0	EURO STOXX 50 Index Futures	Equity Index Futures
CC	CC Comdty	HKNUZ	6.0	0	Cocoa Futures	Food & Fiber Futures
CT	CT Comdty	HKNZ	5.0	0	Cotton #2 Futures	Food & Fiber Futures
KC	KC Comdty	HKNUZ	5.0	0	Coffee C Futures	Food & Fiber Futures

Continued on next page

Table A.3: Futures used

tick code	bbg ticker	future months	roll lookback weeks	expiry rule	description	asset class
SB	SB Comdty	HKNV	5.0	0	Sugar #11 Futures	Food & Fiber Futures
BO	BO Comdty	FHKNZ	4.0	0	Soybean Oil Futures	Grain Futures
CN	C Comdty	HKNZ	3.0	0	Corn Futures	Grain Futures
KW	KW Comdty	HKNUZ	4.0	0	Hard Red Winter Wheat Futures	Grain Futures
SM	SM Comdty	FHKNZ	4.0	0	Soybean Meal Futures	Grain Futures
SY	S Comdty	FHKNX	4.0	0	Soybean Futures	Grain Futures
WC	W Comdty	HKNUZ	4.0	0	Wheat Futures CBOT	Grain Futures
AX	XM Comdty	HMUZ	1.0	0	Australian 10-Year Bond Futures	Interest Rate Futures
AY	YM Comdty	HMUZ	1.0	0	Australian 3-Year Bond Futures	Interest Rate Futures
BF	OAT Comdty	HMUZ	1.5	0	Euro-OAT Futures	Interest Rate Futures
BL	OE Comdty	HMUZ	1.5	0	Euro-Bobl 5-Year Futures	Interest Rate Futures
BN	RX Comdty	HMUZ	1.5	0	Euro-Bund 10-Year Futures	Interest Rate Futures
BT	IK Comdty	HMUZ	1.5	0	Long-Term Euro-BTP Futures	Interest Rate Futures
BX	UB Comdty	HMUZ	1.5	0	Euro-Buxl 30-Year Futures	Interest Rate Futures
BZ	DU Comdty	HMUZ	1.5	0	Euro-Schatz 2-Year Futures	Interest Rate Futures
CB	CN Comdty	HMUZ	3.0	0	Canadian 10-Year Futures	Interest Rate Futures
FV	FV Comdty	HMUZ	3.0	0	US 5-Year T-Note Futures	Interest Rate Futures
GL	G Comdty	HMUZ	3.0	0	Long Gilt Futures	Interest Rate Futures
JB	JB Comdty	HMUZ	1.0	0	Japanese 10-Year Bond Futures JPX	Interest Rate Futures
JBM	BJ Comdty	HMUZ	1.5	0	Mini 10-Year Japanese Government Bond Futures	Interest Rate Futures
TU	TU Comdty	HMUZ	3.0	0	US 2-Year T-Note Futures	Interest Rate Futures
TY	TY Comdty	HMUZ	3.0	0	US 10-Year T-Note Futures	Interest Rate Futures
UB	WN Comdty	HMUZ	3.0	0	Ultra T-Bond Futures	Interest Rate Futures
US	US Comdty	HMUZ	3.0	0	US 30-Year T-Bond Futures	Interest Rate Futures
LC	LC Comdty	GJMQVZ	5.0	0	Live Cattle Futures	Meat Futures
GC	GC Comdty	GJMQZ	3.0	0	Gold Futures COMEX	Metal Futures
HG	HG Comdty	HKNUZ	3.0	0	Copper High Grade Futures COMEX	Metal Futures
PA	PA Comdty	HMUZ	3.0	0	Palladium Futures NYMEX	Metal Futures
PL	PL Comdty	FJNV	3.0	0	Platinum Futures NYMEX	Metal Futures
SV	SI Comdty	HKNUZ	3.0	0	Silver Futures COMEX	Metal Futures

Appendix B

Results

Table B.1 shows a summary of R^2 results for all futures studied. In each backtest window we use 6 months for training the model and 3 months for testing. The values for R^2 used in each backtest windows are that of λ^*, δ^*, as defined in Section 4.3, and we average these values over the whole backtest period. As we are unable to produce plots for all 61 futures, we provide further statistics to understand the most common δ over time:

- $\hat{\delta}$ - the modal value of δ^*
- δ Freq - the frequency of $\hat{\delta}$ over the backtest window, with a value of 1 implying the optimal value of δ never changed. Formally, $\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{\delta_i^*=\hat{\delta}\}}$
- Stability - The proportion of time the current best δ was the best in the next backtest window. Formally, $\frac{1}{N-1}\sum_{i=1}^{N-1} \mathbb{1}_{\{\delta_i^*=\delta_{i+1}^*\}}$.

Table B.1: Average R^2 results for best model with 6 months train, 3 months test.

Spot	Description	Train R^2	Test R^2	TT Ratio	Test Regret	$\hat{\delta}$	$\hat{\delta}$ Freq	Stability
ES	S&P 500 E-Mini Futures	0.403	0.389	0.972	0.008	0.100	0.419	0.524
HI	Hang Seng Index Futures	0.442	0.436	0.991	0.003	0.350	0.326	0.595
XX	EURO STOXX 50 Index Futures	0.459	0.438	0.965	0.017	0.100	0.884	0.833
VX	VIX Futures	0.496	0.484	0.978	0.004	0.100	0.683	0.725
MI	S&P 400 MidCap E-Mini Futures	0.507	0.495	0.978	0.003	0.450	0.302	0.452
XB	RBOB Gasoline Futures NYMEX	0.511	0.506	0.992	0.001	0.450	0.442	0.619
BT	Long-Term Euro-BTP Futures	0.521	0.520	1.017	0.005	0.150	0.302	0.357
HO	Heating Oil #2 Futures NYMEX	0.528	0.523	0.993	0.001	0.500	0.512	0.571
CL	Light Crude Oil Futures NYMEX	0.529	0.520	0.985	0.007	0.350	0.209	0.429
NQ	NASDAQ 100 E-Mini Futures	0.541	0.527	0.975	0.003	0.650	0.302	0.548
GC	Gold Futures COMEX	0.542	0.536	0.990	0.005	0.450	0.209	0.405
PA	Palladium Futures NYMEX	0.553	0.554	1.004	0.003	0.450	0.488	0.500
PL	Platinum Futures NYMEX	0.554	0.553	1.000	0.001	0.400	0.488	0.714
CT	Cotton #2 Futures	0.568	0.572	1.007	0.001	0.400	0.372	0.738
EC	Euro FX Futures	0.571	0.567	0.998	0.004	0.300	0.256	0.452
BX	Euro-Buxl 30-Year Futures	0.584	0.578	0.996	0.004	0.100	0.302	0.524
FT	FTSE 100 Index Futures	0.594	0.597	1.010	0.003	0.450	0.214	0.463
GL	Long Gilt Futures	0.595	0.594	1.006	0.001	0.100	0.837	0.833
JY	Japanese Yen Futures	0.596	0.592	0.999	0.006	0.200	0.279	0.333
CO	Brent Crude Oil Futures ICE	0.599	0.604	1.013	0.002	0.250	0.279	0.571
SW	Swiss Market Index Futures	0.602	0.597	0.995	0.003	0.200	0.302	0.524
BF	Euro-OAT Futures	0.606	0.595	0.991	0.003	0.100	0.786	0.780
BP	British Pound Futures	0.617	0.614	1.000	0.004	0.250	0.442	0.429
TY	US 10-Year T-Note Futures	0.620	0.616	0.997	0.000	0.100	1.000	1.000
TP	TOPIX Futures JPX	0.632	0.623	0.994	0.004	0.100	0.814	0.833
NG	Natural Gas Futures NYMEX	0.634	0.625	0.990	0.012	0.300	0.186	0.381
LC	Live Cattle Futures	0.654	0.641	0.980	0.004	0.200	0.286	0.154
HG	Copper High Grade Futures COMEX	0.665	0.661	0.997	0.004	0.400	0.233	0.405
JB	Japanese 10-Year Bond Futures JPX	0.669	0.657	0.988	0.003	0.100	0.907	0.881
FV	US 5-Year T-Note Futures	0.670	0.661	0.990	0.001	0.100	1.000	1.000
EN	Nikkei 225 Futures SGX	0.670	0.662	0.994	0.010	0.100	0.209	0.476
AD	Australian Dollar Futures	0.677	0.670	0.994	0.006	0.100	0.302	0.571
DX	Dollar Index Futures ICE	0.677	0.673	0.995	0.004	0.400	0.302	0.405
KC	Coffee C Futures	0.678	0.685	1.012	0.003	0.350	0.308	0.583
PT	S&P Canada 60 Futures	0.678	0.632	0.942	0.016	0.450	0.186	0.500

Continued on next page

Table B.1: Average R^2 results for best model with 6m train, 3m test.

Spot	Description	Train R^2	Test R^2	TT Ratio	Test Regret	$\hat{\delta}$	$\hat{\delta}$ Freq	Stability
CB	Canadian 10-Year Futures	0.678	0.674	0.998	0.002	0.100	0.581	0.690
JBM	Mini 10-Year Japanese Government Bond Futures	0.680	0.666	0.986	0.003	0.100	0.480	0.500
BL	Euro-Bobl 5-Year Futures	0.684	0.667	0.979	0.004	0.100	1.000	1.000
NE	Nikkei 225 Futures JPX	0.690	0.673	0.981	0.006	0.100	0.953	0.952
KW	Hard Red Winter Wheat Futures	0.694	0.686	0.994	0.002	0.150	0.476	0.600
CC	Cocoa Futures	0.695	0.702	1.011	0.002	0.300	0.500	0.444
NZ	New Zealand Dollar Futures	0.697	0.692	0.996	0.004	0.100	0.279	0.595
SY	Soybean Futures	0.699	0.710	1.016	0.002	0.350	0.556	0.625
SV	Silver Futures COMEX	0.701	0.697	0.998	0.007	0.100	0.233	0.405
SF	Swiss Franc Futures	0.705	0.701	1.004	0.007	0.100	0.209	0.476
BZ	Euro-Schatz 2-Year Futures	0.710	0.696	0.982	0.000	0.100	1.000	1.000
CD	Canadian Dollar Futures	0.712	0.705	0.992	0.004	0.150	0.488	0.595
NY	Nikkei 225 Futures Yen-Denominated CME	0.715	0.693	0.980	0.017	0.100	0.349	0.548
UB	Ultra T-Bond Futures	0.717	0.710	0.994	0.005	0.100	0.581	0.738
US	US 30-Year T-Bond Futures	0.721	0.716	0.995	0.003	0.100	0.953	0.952
BO	Soybean Oil Futures	0.721	0.732	1.015	0.001	0.300	0.667	0.750
SM	Soybean Meal Futures	0.723	0.738	1.022	0.001	0.450	0.857	0.833
NK	Nikkei 225 Futures CME	0.723	0.704	0.977	0.013	0.200	0.326	0.429
XP	ASX SPI 200 Index Futures	0.731	0.725	0.994	0.006	0.250	0.209	0.524
SB	Sugar #11 Futures	0.748	0.750	1.004	0.000	0.100	0.714	0.833
WC	Wheat Futures CBOT	0.749	0.756	1.010	0.003	0.300	0.444	0.500
CN	Corn Futures	0.755	0.755	1.003	0.000	0.100	0.692	0.750
GO	Low Sulphur Gasoil Futures ICE	0.768	0.761	0.993	0.003	0.150	0.279	0.381
AX	Australian 10-Year Bond Futures	0.771	0.765	0.994	0.000	0.100	1.000	1.000
TU	US 2-Year T-Note Futures	0.775	0.759	0.980	0.000	0.100	1.000	1.000
AY	Australian 3-Year Bond Futures	0.795	0.790	0.994	0.000	0.100	1.000	1.000

Bibliography

[1] Jean-Philippe Bouchaud and Julius Bonart. *Trades, quotes and prices: Financial Markets under the Microscope*. Cambridge University Press, 2018.

[2] Eleni Gousgounis and Esen Onur. The effect of the pit closure on futures trading. *SSRN Electronic Journal*, 2017.

[3] Lucjan T. Orlowski. From pit to electronic trading: Impact on price volatility of u.s. treasury futures. *Review of Financial Economics*, 25:3–9, 2015.

[4] Albert S. Kyle. Continuous auctions and insider trading. *Econometrica*, 53(6):1315–1335, 1985.

[5] Robert Almgren and Neil Chriss. Optimal execution of portfolio transactions. *The journal of risk*, 3(2):5–39, 2001.

[6] Anna Obizhaeva and Jiang Wang. Optimal trading strategy and supply/demand dynamics, 2005.

[7] Fabrizio Lillo, J Doyne Farmer, † Mckinsey, and Rosario N Mantegna. Single curve collapse of the price impact function for the new york stock exchange, 2002.

[8] Jean-Philippe Bouchaud, Yuval Gefen, Marc Potters, and Matthieu Wyart. Fluctuations and response in financial markets: the subtle nature of 'random' price changes, 2003.

[9] Damian Eduardo Taranto, Giacomo Bormetti, Jean-Philippe Bouchaud, Fabrizio Lillo, and Bence Tóth. Linear models for the impact of order flow on prices (I). history dependent impact models. *Quantitative Finance*, 18(6):903–915, 2018.

[10] Rama Cont, Arseniy Kukanov, and Sasha Stoikov. The price impact of order book events. *SSRN Electronic Journal*, Apr 2012.

[11] Luca Philippe Mertens, Alberto Ciacci, Fabrizio Lillo, and Giulia Livieri. Liquidity fluctuations and the latent dynamics of price impact. *Quantitative Finance*, 22:149–169, 2022.

[12] Qingxue Wang, Bin Teng, Qi Hao, and Yufeng Shi. High-frequency statistical arbitrage strategy based on stationarized order flow imbalance. *Procedia Computer Science*, 187:518–523, 2021.

[13] Johannes Muhle-Karbe, Zexin Wang, and Kevin Webster. Stochastic liquidity as a proxy for nonlinear price impact, 2022.

[14] Natascha Hey, Jean-Philippe Bouchaud, Iacopo Mastromatteo, Johannes Muhle-Karbe, and Kevin Webster. The cost of misspecifying price impact. *SSRN Electronic Journal*, 2023.

[15] Jim Gatheral. No-dynamic-arbitrage and market impact. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1292353`, 2010.

[16] Enzo Busseti and Fabrizio Lillo. Calibration of optimal execution of financial transactions in the presence of transient market impact. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(09), 2012.

[17] Futures Sales Strats. Futures liquidity analysis: Order book sweep / cost-to-trade. Technical report, Goldman Sachs, July 2023.

[18] Jim Gatheral and Alexander Schied. Dynamical models of market impact and algorithms for order execution, 2011.

[19] Robert I. Webb, Doojin Ryu, Doowon Ryu, and Joongho Han. The price impact of futures trades and their intraday seasonality. *Emerging Markets Review*, 26:80–98, 2016.

[20] The Investopedia Team. What is the history of futures? `https://www.investopedia.com/ask/answers/031015/what-history-futures.asp`, Oct 2022. Accessed: 2023-07-16.

[21] Group product slate - cme group. `https://www.cmegroup.com/markets/products.html`. Accessed: 2023-07-16.

[22] Kushal Agarwal. Advantages of trading futures vs. stocks. `https://www.investopedia.com/articles/active-trading/032515/advantages-trading-futures-over-stocks.asp`, Jan 2023. Accessed: 2023-07-16.

[23] The benefits of liquidity. `https://www.cmegroup.com/education/courses/understanding-the-benefits-of-futures/the-benefits-of-liquidity.html`. Accessed: 2023-07-16.

[24] Historical data products - futures data. `https://www.tickdata.com/product/historical-futures-data/`, Jan 2023. Accessed: 2023-07-27.

[25] Thomas Neal Falkenberry. High frequency data filtering. `https://s3-us-west-2.amazonaws.com/tick-data-s3/pdf/Tick_Data_Filtering_White_Paper.pdf`.

[26] Lobster data structure. =https://lobsterdata.com/info/DataStructure.php.

[27] Robert W. Holthausen, Richard W. Leftwich, and David Mayers. The effect of large block transactions on security prices: A cross-sectional analysis. *Journal of Financial Economics*, 19(2):237–267, 1987.

[28] Charles M. Lee and Mark J. Ready. Inferring trade direction from intraday data. *The Journal of Finance*, 46(2):733–746, 1991.

[29] Vitor Cerqueira, Luis Torgo, and Igor Mozetič. Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109(11):1997–2028, 2020.

[30] Gianbiagio Curato, Jim Gatheral, and Fabrizio Lillo. Optimal execution with nonlinear transient market impact. `https://arxiv.org/abs/1412.4839`, Dec 2014.

[31] Ngoc-Minh Dang. Optimal execution with transient impact, 2014.

[32] Paul T Boggs and Jon W Tolle. Sequential quadratic programming, 1996.

[33] scipy.optimize.minimize - SciPy v1.11.1 Manual. `https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html#scipy.optimize.minimize`. Accessed: 2023-08-09.