Spring 2015

A pure-jump market-making model for high-frequency trading

Chi Wai Law
Purdue University

Graduate School Form 30
Updated 1/15/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Chi Wai Law

Entitled

A Pure-Jump Market-Making Model for High-Frequency Trading

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Frederi G. Viens
Chair

Fabrice Baudoin

Hao Zhang

Jose E. Figueroa-Lopez

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Frederi Viens

Approved by: Jun Xie 4/21/2015

Head of the Departmental Graduate Program Date

A PURE-JUMP MARKET-MAKING MODEL FOR HIGH-FREQUENCY TRADING

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Chi Wai Law

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2015

Purdue University

West Lafayette, Indiana

Dedicated to my parents

Tat Hung Law & *Kit Chun Li*

and siblings

Dikman Law & *Anthea Law*

ACKNOWLEDGMENTS

This thesis could not have been completed without the tremendous help from my advisor Prof. Frederi G. Viens, who has been my mentor since day one when I came to Purdue. I am indebted to his enthusiastic support, guidance and encouragement throughout the four years of my PhD study. Under the enjoyable supervision of Prof. Viens, I have the freedom to discover my niche without the burden of following any prescribed route. Theoretical research is elegant, refined and polished but I also adore the powerful impact that applied research can create. However, without the profound knowledge of Prof. Viens, I could have fallen into the trap of endless exploration before settling down to a realistic agenda. High-frequency algorithmic trading is a new topic that has so many exciting areas and I am so delighted to have had the opportunity to work with Prof. Viens on this subject.

I would also like to thank my committee, Prof. Hao Zhang, Prof. Jose E. Figueroa-Lopez and Prof. Fabrice Baudoin, for their inspiration, backing as well as challenging questions, in addition to our beloved Department Head Prof. Rebecca W. Doerge, and Graduate Chair Prof. Jun Xie, for their unlimited counseling during my academic journey.

Finally, I would give my full appreciation to the love and consideration from my parents, who always stand by me to navigate through difficult times. Without their countless sacrifices, I would not have been able to fulfil my dream.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

SYMBOLS

$\mathbb{R}$	$(-\infty,\infty)$
$\mathbb{R}_+$	$[0,\infty)$
$\mathbb{N}$	$\{0,1,..\}$
$\mathbb{Z}_+$	$\{1,2,...\}$
$\overline{\mathbb{Z}}_+$	$\{1,2,...\}\cup\{\infty\}$
$\mathbb{I}$	finite regime space $\{(0,0),(0,1),(1,0),(1,1)\}$
$\mathbb{J}$	compact impulse space $[-C,C]$
$\mathbb{K}$	generic mark space
h_1	cost of switching
h_2	cost of impulse
h	h_1+h_2
B_t	cash
Q_t	inventory (quantity)
S_t^a	ask price
S_t^b	bid price
S_t	mid price
I_t	regime
$M_a(t)$	buy market order
$M_b(t)$	sell market order
N_t	point process
λ_i	intensity of type i order
u	control
$\mathscr{U}(\bullet)$	utility function
μ	base rate of intensity
α	excitation coefficient in exponential Hawkes kernel
β	decay coefficient in exponential Hawkes kernel
v	volume of a order
V	value function of the control problem
i	regime
ζ	volume of an impulse market order
ζ^+	$\max(0,\zeta)$
ζ^-	$\max(0,-\zeta)$
δ	tick size
Δ_t	bid-ask spread
ε	rebate of limit order
η	fee of market order
N	number of intervals
M	number of sample paths
K	number of basis functions

ABBREVIATIONS

càdlàg	right continuous with left limit
BSDE	Backward Stochastic Differential Equation
CBSDE	Constrained Backward Stochastic Differential Equation
CFBSDE	Constrained Forward Backward Stochastic Differential Equation
EM	Expectation Maximization
FBSDE	Forward Backward Stochastic Differential Equation
HF	High-Frequency
HFT	High-Frequency Trading
HJBQVI	Hamilton-Jacobi-Bellman quasi-variational inequality
LOB	Limit Order Book
MPP	Marked Point Process
PDE	Partial Differential Equation
PIDE	Partial Integro-Differential Equation
SDE	Stochastic Differential Equation

ABSTRACT

Law, Chi Wai PhD, Purdue University, May 2015. A Pure-Jump Market-Making Model for High-Frequency Trading. Major Professor: Frederi G. Viens.

We propose a new market-making model which incorporates a number of realistic features relevant for high-frequency trading. In particular, we model the dependency structure of prices and order arrivals with novel self- and cross-exciting point processes. Furthermore, instead of assuming the bid and ask prices can be adjusted continuously by the market maker, we formulate the market maker's decisions as an optimal switching problem. Moreover, the risk of overtrading has been taken into consideration by allowing each order to have different size, and the market maker can make use of market orders, which are treated as impulse control, to get rid of excessive inventory. Because of the stochastic intensities of the cross-exciting point processes, the optimality condition cannot be formulated using classical Hamilton-Jacobi-Bellman quasi-variational inequality (HJBQVI), so we extend the framework of constrained forward backward stochastic differential equation (CFBSDE) to solve our optimal control problem.

1. INTRODUCTION

1.1 Background

Market makers provide liquidity to the market by posting buy and sell orders simultaneously on both sides of the limit order book (LOB). They earn the profit from the bid-ask spread in each round-trip buy and sell transaction in return for bearing the risks of adverse price movements, uncertain executions and adverse selections [1, 2]. In the US equity market, market makers also receive a special form of income called *rebates* from the stock exchanges due to keen competition of the exchange marketplace.

Table 1.1.
Simplified fee structure of US stock exchanges as of 2/19/2015

Exchange	Limit Order (Rebate)	Market Order (Fee)
NYSE	0.0022	0.0027
NYSE Arca	0.0030	0.0030
NYSE MKT	0.0016	0.0028
Nasdaq	0.00295	0.0030
Nasdaq BX	-0.0014	-0.0015
Nasdaq PSX	0.0025	0.0026
BZX	0.0020	0.0030
BYX	-0.0018	-0.0016
EDGX	0.0020	0.0030
EDGA	-0.0005	-0.0002
CHX	0.0020	0.0030

In 2005, New York Stock Exchange (NYSE) had about 80% market share (by volume) of the US equity market [3]. However, after the introduction of Regulation ATS in 1998 and Regulation NMS in 2005, its market share plunged to 25% in 2009. To attract liquidity among fierce competition, exchanges adopt the so-called *maker-taker* fee structure [4], where exchanges reward participants adding liquidity (limit orders) while charging players removing liquidity (market orders) (see Table 1.1). As a consequence, the market-making

business becomes more lucrative and research in market making draws people's attention again.

To give a ballpark estimate, assuming daily trading volume is 36 million shares (e.g. MSFT), the market maker can capture 5% of the order flows, the rebate is \$0.002 per share, tick size is \$0.01, there are 250 trading days per year, then the annual profit of market-making this security is 3.15 million with 0.9 million coming from the rebate and 2.25 million from the bid-ask spread. Needless to say, high-frequency trading (HFT) firms often operate on thousands of stocks driven by fully automated computer algorithms.

However, the above calculation assumes an unrealistic scenario that the price does not move; in fact, the volatility of stock can be so large that market maker may suffer huge loss. To understand the behavior of a rational market maker, we need to figure out how he controls the inventory risk[1] while maximizing the expected profit. In the next section, we will look at some classical market-making models.

1.2 Review of Market-Making Models

The early literature on market making appears mostly in the field of market microstructure in finance where researchers study the behavior of various market participants in the financial exchanges. The early models [5–8] are commonly called inventory models where a monopolistic market maker adjusts his bid and ask prices in order to control his inventory level. Such models provide a lucid framework to understand the interactions between market players as well as their impact on the market. However, the models often depend on the hard-to-estimate demand/supply functions and the setting of the market environment are unrealistic (e.g. bid/ask price is continuous, all orders have the same size, trade ratio of uninformed/uninformed traders is fixed etc)

Another type of market-making models are the pure stochastic models as in [9–12]. In those models, the market maker is assumed to be so tiny that he has negligible influence on the prices and order arrivals, which follow some stochastic processes with model pa-

[1] In this thesis, we will not consider adverse selection risk as in [1, 2].

rameters estimated from historical data. The goal of the market maker is to maximize his risk-adjusted profit under the given state dynamics.

1.2.1 Garman (1976)

Garman's [5] model is often regarded as one of the earliest model of market making, and the title of his paper, market microstructure, develops into a discipline of rigourous study of market mechanism in the field of finance. In Garman's model, there is only one monopolistic market maker for the whole market and all trades must go through this market maker; in other words, no direct exchange of buyer and seller is allowed. As a result, the market maker has the full price control. However, the rate of incoming Poisson buy and sell order λ_a, λ_b will depend on the ask and bid price S_a, S_b which he sets at time 0 and the prices will remain the same throughout the whole trading period. At time 0, he has cash B_0 and inventory Q_0 and he will go bankrupt when either of them drops to zero. In Garman's setting, the market maker is risk-neutral and he seeks only to maximize the expected profit while avoiding bankruptcy.

Assuming a linear rate function $\lambda_b(s) = \alpha + \beta s,\ \lambda_a(s) = \gamma - \delta s$ with $\gamma > \alpha \geq 0,\ \beta, \delta > 0$, in order to avoid running out of inventory or holding infinite amount of stock, the market maker will set the bid and ask prices S_b, S_a such that $\lambda_b = \lambda_a$, so the market maker seeks to maximize the profit by solving the static optimization problem

$$\max_{S_b, S_a} (S_a - S_b)(\alpha + \beta S_b) \text{ s.t.} \tag{1.1}$$

$$\alpha + \beta S_b = \gamma - \delta S_a \tag{1.2}$$

The solution is $\lambda^* = (\alpha\delta + \gamma\beta)/(2(\beta + \delta)),\ S_b = (\lambda^* - \alpha)/\beta,\ S_a = (\gamma - \lambda^*)/\delta$.

Under Garman's setting, the inventory Q_t can be shown to be a birth and death process with birth rate $\lambda_{i,i+1} = \lambda_b$ and death rate $\lambda_{i,i-1} = \lambda_a$. From the theory of continuous time Markov chain, when $\lambda_a = \lambda_b$, the stock ruin probability $P(Q_t = 0\ \exists t \geq 0 | Q_0 = i) = 1$. In other words, setting the bid and price only once at $t = 0$ is not viable as the market market will run out of inventory with probability one.

1.2.2 Ho and Stoll (1981)

Ho and Stoll [8] extend Garman's model by allowing the bid and ask prices to change over time and use stochastic optimal control technique to solve the market-making problem. Same as Garmen, the authors use linear demand/supply functions for the Poisson process of buy and sell orders N_t^a, N_t^b. Moreover, they assume the inventory value I_t follows geometric Brownian motion and the market maker is risk-averse with quadratic utility $U(w)$. The optimal control problem is as follows (B_t is cash, S is market maker's own constant fair price).

$$\max_{S_t^b, S_t^a} \mathbb{E}(U(B_T + I_T)) \tag{1.3}$$

$$dB_t = r_B B_t dt - S_t^b dN_t^b + S_t^a dN_t^a,\ B_0 = 0 \tag{1.4}$$

$$dI_t = r_I I_t dt + S(dN_t^b - dN_t^a) + \sigma_I I_t dW_t^I,\ I_0 = 0 \tag{1.5}$$

$$\lambda_t^a = \alpha - \beta(S_t^a - S) \tag{1.6}$$

$$\lambda_t^b = \alpha - \beta(S - S_t^b) \tag{1.7}$$

1.2.3 Avellaneda and Stoikov (2008)

27 years after Ho and Stoll [8], Avellaneda and Stoikov [9] propose another model from a mathematical finance perspective. Instead of assuming a monopolistic market maker, the authors consider a small market maker who has no pricing power. Based on some empirical studies [13–17], Avellaneda and Stoikov claim that the arrival intensity is in the form $\lambda(\delta) = A\exp(-k\delta)$ where δ is the distance from the mid price S_t. Also, they use the mid-price S_t, which follows Brownian motion, as the reference price rather than the fair price as in Ho and Stoll [8]. Instead of describing the dynamics of inventory value, they directly use the accounting equation of the inventory quantity, which seems to be much more intuitive. Finally, they use exponential rather quadratic utility as in Ho and Stoll [8].

$$\max_{S_t^a, S_t^b} \mathbb{E}(U(B_T + Q_T S_T)) \tag{1.8}$$

$$dB_t = S_t^a dN_t^a - S_t^b dN_t^b \tag{1.9}$$

$$dQ_t = dN_t^b - dN_t^a \tag{1.10}$$

$$dS_t = \sigma dW_t \tag{1.11}$$

$$\lambda_b(S_t - S_t^b) = A\exp(-k(S_t - S_t^b)) \tag{1.12}$$

$$\lambda_a(S_t^a - S_t) = A\exp(-k(S_t^a - S_t)) \tag{1.13}$$

1.2.4 Guilbaud and Pham (2013)

Guilbaud and Pham [12] is the latest stochastic market-making model and the authors pioneer a number of modern features not seen in previous papers. First, the market maker's limit orders are either pegged to the best bid/ask or *one tick better*. When the bid-ask spread is only one tick, a one-tick-better limit order means market order. Second, the mid-price S_t is extended to jump diffusion (Lévy process) and the bid-ask spread Δ_t is modeled by a continuous time Markov chain. Besides, market maker can choose the size of limit orders L_t^a, L_t^b posted to the limit order book as well as the time τ_n and size ζ_n of market orders, which are used to remove excessive inventory. Lastly, the final liquidation value includes the cost of crossing the spread and a non-proportional exchange fee η.

$$\max_{S_t^a, S_t^b, L_t^a, L_t^b, \tau_n, \zeta_n} \mathbb{E}\Big(U\left(B_T + Q_T S_T - |Q_T|\Delta_T/2 - \eta\right)\Big) \tag{1.14}$$

$$B_t = \int_0^t S_s^a L_s^a dN_s^a - \int_0^t S_s^b L_s^a dN_s^b - \sum_{\tau_n \le t} \left(\zeta_n S_{\tau_n} + |\zeta_n|\Delta_{\tau_n}/2 + \eta\right) \tag{1.15}$$

$$Q_t = \int_0^t L_s^b dN_s^b - \int_0^t L_s^a dN_s^a + \sum_{\tau_n \le t} \zeta_n \tag{1.16}$$

$$S_t = \int_0^t \mu_s ds + \int_0^t \sigma_s dW_s + \int_0^t \gamma_s d\widetilde{N}_s \tag{1.17}$$

1.3 Issues of Existing Market-Making Models

In this section, we highlight some issues of existing market-making models in the context of high-frequency trading.

1. In modern financial exchanges, prices are only allowed on a predefined fixed grid called price ticks. As a result, price is a pure-jump process and it has two dimensions,

namely times and magnitudes of the jumps. Diffusion can only approximate the magnitudes of the jumps but cannot describe the properties related to timing of the jumps such as jump clustering.

2. The common assumption of Poisson order arrivals is often rejected in empirical literature as order arrivals depict strong self-excitation behavior [18–20].
3. All models assume that price and order arrivals are independent, which is far from the truth by realizing that price rises with large buy market order and falls with large sell market order. Because of adverse selection [1, 2], the absence of this crucial dependency structure will generate large phantom profit for the market maker and cause the average profit of the market-making strategy to be overstated.
4. Since almost all exchanges nowadays use the price-time[2] priority, changing price or quantity of limit orders means loss of priority. Nonetheless, existing models all use regular control to continuously adjust the quotes without any penalty.
5. A critical component of many existing models is the demand/supply rate function. For example, in [8], it is in the form $\alpha - \beta\delta$ while in [9], it can be expressed as $A\exp(-k\delta)$. Yet the parameters are hard to estimate since when the limit order is more than one tick from the best quote, the execution probability is minuscule (e.g. less than 3% for E-mini S&P future [21]). In addition, the quoted price is not continuous but only allowed in a fixed grid of price ticks.
6. For the sake of simplicity, existing models assume all orders are of the same size. Such an assumption will mask the risk of overtrading of the market maker. For instance, to continuously maintain priority in the queue, the market maker may post more limit orders in the order book than his risk tolerance. However, the arrival of one giant market order may raise his inventory to an unacceptable level, which can potentially lead to bankruptcy. Such kind of risk cannot be modeled with all orders having the same size.

[2]Limit order having better price and then earlier time-stamp will have higher execution priority.

2. A NEW PURE-JUMP MARKET-MAKING MODEL FOR HIGH-FREQUENCY TRADING

In view of the drawbacks of existing models, the main theme of this thesis to construct a new market-making model under a realistic trading environment, such that the market maker can compensate the inventory risk by adequate profit. In particular, we focus on instruments trading on modern electronic order-driven exchanges while the market maker is small enough so that his decision will not have significant impact on the market. Before we go into the details of the new model, this chapter provides an executive summary of our new ideas.

2.1 Prices and Order Arrivals

In our new market-making model, prices are now pure-jump processes dependent on the order arrivals but the dependency structure is remarkably simple and intuitive.

We classify each order into one of the twelve order types according to its type (limit, market, cancellation), direction (buy, sell) and aggressiveness (whether the order moves price or not) (see Table 4.1). The twelve type of orders are modeled as cross-exciting point processes and in particular, we will use the Hawkes process representation, which will be discussed in Chapter 3.

Similar classification schemes have been used in other papers [18, 22] but our key contribution is that we discover a simple relation between prices and order arrivals under this classification. Let $N(t) = (N_1(t), ..., N_{12}(t))$ denotes the multivariate point process of the twelve types of order, and $M_a(t)$, $M_b(t)$, $S_a(t)$, $S_b(t)$ denote the *buy* market orders, *sell* market orders, *ask* price and *bid* price respectively. Assuming the tick size δ of the stock

is fixed and each price jump is of size one tick[1]. It is not hard to realize the following association.

$$M_a(t) = N_1(t) + N_5(t) \tag{2.1}$$

$$M_b(t) = N_2(t) + N_6(t) \tag{2.2}$$

$$S_a(t) = S_a(0) + (N_1(t) + N_4(t) - N_{12}(t))\delta \tag{2.3}$$

$$S_b(t) = S_b(0) + (N_{11}(t) - N_2(t) - N_3(t))\delta \tag{2.4}$$

Such a set of simple equations provides the dependency structure between prices S_a, S_b and market orders M_a, M_b via two linkages, namely the common components N_1, N_2 and the cross-excitations among $(N_1, .., N_{12})$.

Under further assumptions and regularity conditions, we have shown that the change in mid-price ΔS in our framework converges to a Brownian motion using functional central limit theorem for Hawkes process [23, Corollary 1],

$$\sqrt{n}\left(\Delta S(\bullet n)/n - \bullet a^\top (I_{10} - \Gamma)^{-1}\mu\right) \xrightarrow[n\to\infty]{\text{weak}} a^\top (I_{10} - \Gamma)^{-1}\Sigma^{1/2} W(\bullet) \tag{2.5}$$

where $a = (\delta/2)[1,\ -1,\ -1,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0]^\top$, $N(t) = [N_1(t), ..., N_{10}(t)]^\top$, $\Gamma = [\int_0^\infty \gamma_{ij}(t)dt]_{i,j}$, γ is the Hawkes kernel, $\Sigma = \text{diag}((I_{10} - \Gamma)^{-1}\mu)$ and μ is background arrival rate.

The details of the joint model as well as the result of some empirical experiments using Nasdaq tick data will be presented in Chapter 4.

2.2 Trading Features

In addition to the dependency of price and order arrivals, we also introduce other pragmatic trading features. For example, quotes of market maker are no longer changed continuously; instead, we formulate the control problem as an optimal switching where the market maker will be penalized every time he revises his quotes due to the price-time priority of modern exchanges.

[1] We will look at the general model where price can jump more than one tick in Section 4.6.

For the switching regimes, we only consider either pegging the limit orders to the best quotes or withdrawing from the market. The effective arrival rate of the market orders hitting the market maker is determined from the target market share parameter ρ chosen in advance by the market maker. Under this setting, there is no demand/supply function to estimate.

The risk of overtrading is modeled via marked point process where the mark corresponds to volume of each order. Also, we follow Guilbaud and Pham [12] to allow the market maker to use market orders to liquidate excessive inventory.

The control problem is now a combined (simultaneous) optimal switching and impulse control problem under a pure-jump environment (see definition 5.2.1 for full specification) and we have proposed some simplifying assumption to make the problem more tractable (see Section 5.2.2).

2.3 Constrained Forward Backward Stochastic Differential Equation

Though the above enhancements make the model more realistic, they come with a price tag. The main difficulty lies in the stochastic intensity λ of the self-exciting point process N.

Suppose we have a control problem in the following form.

$$V(s,x,i) = \max_{\{\tau_n, i_n, \zeta_n\}} \mathbb{E}\Bigg(g(T, X_T, I_T) + \int_s^T f(t, X_t, I_t)dt - \sum_{\tau_n \in (s,T]} \Big(h_1(\tau_n, X_{\tau_n^-}, i_{n-1}, i_n) + h_2(\tau_n, X_{\tau_n^-}, \zeta_n) \Big) \Bigg| \mathscr{F}_s \Bigg) \tag{2.6}$$

$$X_t = x + \int_s^t b(r, X_r, I_r)dr + \int_{(s,t]\times\mathbb{K}} \gamma(r, X_{r^-}, I_{r^-}, k) N(dr \times dk) + \sum_{\tau_n \in (s,t]} \Gamma(\tau_n, X_{\tau_n^-}, \zeta_n) \tag{2.7}$$

$$I_t = i\mathbb{1}_{[s,\tau_1)}(t) + \sum_{n=1}^{\infty} i_n \mathbb{1}_{[\tau_n, \tau_{n+1})}(t) \tag{2.8}$$

If the intensity $\lambda(t)$ of the marked point process $N(dt,dk)$ is deterministic, the value function V is the viscosity solution [24] of the Hamilton-Jacobi-Bellman quasi-variational inequality (HJBQVI) (A.4-A.5) [25, 26].

$$\begin{aligned}\max\Bigg\{ & f(t,x,i)+V_t(t,x,i)+V_x(t,x,i)^\top b(t,x,i) \\ & \quad +\int_{\mathbb{K}}\Big(V(t,x+\gamma(t,x,i,k),i)-V(t,x,i)\Big)\lambda(t)\mu(t,dk), \\ & \max_{j,\zeta}\Big\{V(t,x+\Gamma(t,x,\zeta),j)-h_1(t,x,i,j)-h_2(t,x,\zeta)\Big\}-V(t,x,i)\Bigg\}=0 \qquad (2.9)\end{aligned}$$

$$V(T,x,i)=g(T,x,i) \qquad (2.10)$$

However, when the intensity $\lambda(t)$ is stochastic but we still apply the same method naively, the resulting equation will be a partial integro-differential equation (PIDE) with *random* coefficients. Even we can solve the PIDE for each ω, the solution will not equal the value function V of the control problem as the value function is non-random.

In 2010, Kharroubi et al. [27] establish the connection of constrained forward backward stochastic differential equation (CFBSDE) to impulse control problem and later in 2014, Elie and Kharroubi [28] apply CFBSDE to solve optimal switching. While Kharroubi et al. [27], Elie and Kharroubi [28] focus on state variable driven by Brownian motion, we have extended the formulation to include state variable driven by marked point process with stochastic intensity and enrich the framework to handle the combined optimal switching and impulse control problem, where switching and impulse can happen at the same time.

We have shown that the value function V of the above control problem (2.6-2.8) is given by the Y component of the unique minimal solution (Y,U,U',K) of the following CFBSDE (2.11-2.14), where N' is the marked point process associated with the control events after some change of probability measure. The constrain that forces component U' to be below the sum of switching cost and impulse cost h pushes the component Y towards the value function V of the control problem.

$$\begin{aligned}X_t=X_s+\int_s^t b(r,X_r,I_r)dr+\int_{(s,t]\times\mathbb{K}}\gamma(r,X_{r^-},I_{r^-},k)N(dr\times dk) \\ +\int_{(s,t]\times\mathbb{I}\times\mathbb{J}}\Gamma(r,X_{r^-},\zeta)N'(dr\times di\times d\zeta) \qquad (2.11)\end{aligned}$$

$$I_t = I_s + \int_{(s,t]\times\mathbb{I}\times\mathbb{J}} (i - I_{r^-}) N'(dr \times di \times d\zeta) \tag{2.12}$$

$$Y_t = g(T, X_T, I_T) + \int_t^T f(r, X_r, I_r) dr - \int_{(t,T]\times\mathbb{K}} U(r,k)\widetilde{N}(dr \times dk) - \int_{(t,T]\times\mathbb{I}\times\mathbb{J}} U'(r,i,\zeta) N'(dr \times di \times d\zeta) + K_T - K_t \tag{2.13}$$

$$U'(t,i,\zeta) \leq h_1(t, X_{t^-}, I_{t^-}, i) + h_2(t, X_{t^-1}, \zeta) \; \forall t \in (s,T] \tag{2.14}$$

Some numerical schemes of solving the CFBSDE and simulation results will be examined in the Appendix.

2.4 Thesis Layout

The layout of the remaining part of the thesis will be as follows. Chapter 3 will give a survey of Hawkes processes and their application to high-frequency data modeling in finance. Chapter 4 will describe our new joint price and order arrival model. The mathematical formulation of the new market-making model will be presented in Chapter 5, followed by the summary of our contributions and conclusion in Chapter 6. The Appendix will give a brief introduction to optimal switching and impulse control, followed by the details of our extension to constrained forward backward stochastic differential equation.

3. HAWKES PROCESSES

3.1 Introduction

This chapter introduces and surveys an emerging class of stochastic point processes used in modeling the evolution of high-frequency data on stock markets at a high level of quantitative detail.

The information contained in a stock market's *Limit Order Book (LOB)* is a multivariate time series which records the order arrival times and volumes at each price level of thousands of stocks trading on the exchange. A LOB exhibits a number of distinctive characteristics [29–31] including

1. irregular time interval between arrivals
2. discrete state space of price ticks and volume lot sizes
3. intraday seasonality (more activities around market open and close)
4. arrival clustering
5. self-excitation from its own history
6. cross-excitation from the history of other assets
7. long memory of excitation effect

Consequently, classical time series models with fixed time intervals such as ARIMA and GARCH are not suitable to model High-Frequency (HF) financial data. A standard approach commonly used in practice is to re-sample the data in 5-minutes intervals [32, 33], thereby avoiding the time scale for liquid stocks where many of the characteristics listed above can be observed, but this may amount to discarding about 99% of the data for such stocks. On the other hand, Poisson processes, which are widely used in the market microstructure literature [5, 8], fail to depict the above features prevalent in HF data.

This chapter, on the current research in HF financial data modeling, concentrates on the use of the so-called with Hawkes processes, a family of point processes designed to model

self- and cross-excitation. In Section 3.2, we offer an introduction to point processes and the material is thoroughly covered in major textbooks such as [34–38]. Sections 3.3 and 3.4 introduce Hawkes processes and their statistical inference. The applications of Hawkes processes to HF data modeling is presented in Section 3.5 and a brief history of Hawkes processes is contained in 3.6.

3.2 Point Processes

3.2.1 Definition

Let X (state space) be a locally compact Hausdorff second countable topological space[1], $\mathscr{B}_X$ be the Borel sets on X and $\mathscr{B}$ be the collection of bounded (relatively compact) sets on X. A Borel measure μ on $(X,\mathscr{B}_X)$ is called locally finite if $\mu(B) < \infty \; \forall B \in \mathscr{B}$. Let $\mathfrak{N}(X)$[2] be the set of (positive) locally finite Borel counting (integer-valued) measure on $(X,\mathscr{B}_X)$ and $\mathscr{N}(X)$[3] be the σ-algebra of $\mathfrak{N}(X)$ generated by the set of evaluation functionals $\{\Phi_B : \mathfrak{N}(X) \longrightarrow \mathbb{N} |\; B \in \mathscr{B}\}$ where $\Phi_B(\mu) = \mu(B)$ and $\mathbb{N} = \{0,1,2,...\}$.

A point process N on X is defined as a measurable mapping from a probability space $(\Omega,\mathscr{F},\mathbb{P})$ to $(\mathfrak{N}(X),\mathscr{N}(X))$; thus a point process is formally a measure-valued random element. However, for any point process N, there exists random variables $b_i \in \mathbb{Z}_+ = \{1,2,..\}$, $x_i \in X$, $n \in \overline{\mathbb{Z}}_+ = \mathbb{Z}_+ \cup \{\infty\}$ such that $N(\bullet) = \sum_{i=1}^{n} b_i \delta_{x_i}(\bullet)$ where δ_x is the Dirac measure $(\delta_x(A) = \mathbb{1}(x \in A))$ [34, p.20]. If we think of b_i as the number of points at x_i, we can see that the point process N is indeed the random counting measure showing the total number of points in any given region and this matches our intuition that a point process is a random set of points $\{x_i\}$ on X.

The point process N is called simple if $\mathbb{P}(N(\{x\}) > 1) = 0 \; \forall x \in X$[4]; that is, each location has at most one point. In this case, $N(\bullet) = \sum_{i=1}^{n} \delta_{x_i}(\bullet)$.

[1] Some textbooks use complete separable metric space, but locally compact Hausdorff second countable space has a complete separable metrization and all the results do not depend on any particular choice of metric [34, p.11]. In most cases, $X = \mathbb{R}^m$.

[2] On locally compact Hausdorff second countable space, all locally finite Borel measures are Radon measures.

[3] $\mathscr{N}(X)$ is the same as the Borel σ-algebra generated by the vague topology of $\mathfrak{N}(X)$ [34, p.32].

[4] X is Hausdorff, so all singletons are closed and thus measurable.

Suppose X is also a topological vector space (e.g. $\mathbb{R}^m$), the shift operator $S_t : \mathscr{B}_X \longrightarrow \mathscr{B}_X$ is defined as $S_t(A) = A + t = \{(s+t) \in X | s \in A\}$. A point process N is called stationary if the shifted process $N \circ S_t$[5] has the same distribution as N $\forall t \in X$.

3.2.2 Moments

Let $k \in \mathbb{Z}_+$, the k^{th} moment measure[6] $M^k : \mathscr{B}_X^{\otimes k} \longrightarrow [0,\infty]$ of a point process N is defined as

$$M^k(A_1,..,A_k) = \mathbb{E}(N(A_1)...N(A_k)) = \mathbb{E}\left(\sum_{x_1}..\sum_{x_k}\delta_{(x_1,..,x_k)}(A_1 \times ... \times A_k)\right) \tag{3.1}$$

The first moment measure is also called mean (intensity) measure and denoted as $M(\bullet)$. The covariance measure is defined as

$$C^2(A_1,A_2) = \mathrm{Cov}(N(A_1),N(A_2)) = M^2(A_1,A_2) - M(A_1)M(A_2) \tag{3.2}$$

The second and higher moment measures have concentration along diagonals, so we also have the k^{th} *factorial* moment measure.

$$M^{(k)}(A_1,..,A_k) = \mathbb{E}\left(\sum_{x_1\neq..\neq x_k}..\sum\delta_{(x_1,..,x_k)}(A_1 \times ... \times A_k)\right) \tag{3.3}$$

The name factorial comes from the fact that $M^{(k)}(A,..,A) = \mathbb{E}(N(A)(N(A)-1)...(N(A)-k+1))$. Obviously, $M(A) = M^{(1)}(A)$ and for $k = 2$, we have $M^2(A_1,A_2) = M^{(2)}(A_1,A_2) + M(A_1 \cap A_2)$.

If $X = \mathbb{R}^m$ and N is stationary, it can be shown that $M(A) = \lambda|A|$ where $\lambda = M((0,1]^m)$ and $|\bullet|$ is the Lebesgue measure. That implies the mean measure M of a stationary point process is absolutely continuous with respect to Lebesgue measure with constant density $M((0,1]^m)$. If the covariance factorial moment measure $C^{(2)}$ is also absolutely continuous, we denote its density function as $c^{(2)}(x,y)$. Since N is stationary, $c^{(2)}(x,y) = c^{(2)}(y-x)$ and $c^{(2)}(\bullet)$ is called reduced covariance density. The covariance measure C^2 is usually

[5] $N \circ S_t : \Omega \longrightarrow (\mathfrak{N}(X), \mathscr{N}(X))$, $((N \circ S_t)(\omega))(A) = (N(\omega))(S_t(A))$, that is N is shifted t unit to the left when $X = \mathbb{R}$.

[6] The notations of moment, covariance, factorial moment, reduced moment vary between authors.

not absolutely continuous but for simple point process N on $\mathbb{R}_+$, the quantity below is still called (reduced) covariance density, and is useful in estimation:

$$c^2(dx) = \mathbb{E}(N(x+dx)N(x))/dx^2 - \lambda^2 = \lambda\delta(dx) + c^{(2)}(dx) \quad \left(\int_{-\infty}^{\infty} \delta(x)dx = 1\right) \tag{3.4}$$

3.2.3 Marked Point Processes

When an event happens, it may carry an additional information (mark). For instance, each order arrival is associated with an order quantity (volume) and each earthquake is reported with a magnitude. A point process with marks is called marked point process.

Let Y (mark space) be a locally compact Hausdorff second countable space, $(Y, \mathscr{B}_Y)$ be a measurable space and ν (mark distribution) be a probability measure on $(Y, \mathscr{B}_Y)$. A marked point process (MPP) N is a measurable mapping $N : \Omega \longrightarrow (\mathfrak{N}(X \times Y), \mathscr{N}(X \times Y))$ such that the ground measure $N_g(\bullet) = N(\bullet \times Y)$ is a point process (i.e. locally finite)[7]. Hence a marked point process is nothing but a point process on a product space, but usually we treat the location x and mark y differently and we have a few more definitions.

N is called a multivariate point process if $Y = \{1, .., d\}$. In this case, $N_i(\bullet) = N(\bullet \times \{i\})$ is called the marginal process of type i points. A MPP N is called simple if N_g is simple[8]. The marks of a MPP are called *unpredictable* if y_n is independent of $\{(x_i, y_i)\}_{i<n}$ and they are called *independent* if y_n is independent of $\{(x_i, y_i)\}_{i\neq n}$[9].

3.2.4 Stochastic Intensity

In this section, $X = \mathbb{R}_+$[10] and $N_t = N((0,t])$. Let $(\Omega, \mathscr{F}, \mathscr{F}_t, \mathbb{P})$ be a filtered complete probability space. A stochastic process $Z : \mathbb{R}_+ \times \Omega \longrightarrow \mathbb{R}$ is called $\mathscr{F}$-predictable if it is measurable with respect to the predictable σ-algebra $\mathscr{P} = \sigma(\{(s,t] \times A | 0 \leq s < t, A \in \mathscr{F}_s\})$. If Z_t is adapted and left-continuous, then Z_t is predictable [36, p.9]. In practice, all the

[7] From this definition, Poisson random measure N on $\mathbb{R}^2$ is not MPP on $\mathbb{R} \times \mathbb{R}$ as $N(A \times \mathbb{R}) = \infty$.

[8] Any point process can be treated as simple MPP with the mark being the number of points at x_i.

[9] Notice that if marks are independent, future location x_{n+1} cannot depends on previous mark y_n.

[10] The stochastic intensity of point process on $\mathbb{R}_+$ is extended to higher dimension in [39].

predictable processes we use are in this category. Also if Z_t is predictable, then $Z_t \in \mathscr{F}_{t^-}$; in other words, the predictable process Z_t is "known" just before time t.

We assume the filtration $\{\mathscr{F}_t\}$ satisfies the usual condition (complete and right-continuous) and $\{N_t\}$ is adapted and simple. A stochastic process $A : \mathbb{R}_+ \times \Omega \longrightarrow \mathbb{R}_+$ is called a $\mathscr{F}$-compensator of a point process N if it is increasing, right-continuous, $\mathscr{F}$-predictable, $A_0 = 0$ a.s. and $(N_t - A_t)$ is a $\mathscr{F}$-local martingale. If $A_t = \int_0^t \lambda_s ds$ a.s., λ_t is non-negative and $\mathscr{F}$-predictable, then λ_t is called the stochastic or conditional $\mathscr{F}$-intensity of N[11,12,13]. In other words, intensity exists if and only if the compensator is absolutely continuous. A defining properties of λ_t is that

$$\mathbb{E}\left(\int_s^t \lambda_u du \middle| \mathscr{F}_s\right) = \mathbb{E}(N_t - N_s|\mathscr{F}_s) \quad \text{a.s.} \quad \forall s < t \tag{3.5}$$

When $s \to t$, this becomes $\lambda_t dt = \mathbb{E}(N(dt)|\mathscr{F}_{t^-})$. We can see that the stochastic intensity λ_t is the instantaneous rate of arrival conditioned on all information just before time t. For a multivariate point process, $\lambda_i(t)$ is the intensity of the marginal process $N_i(t)$.

On the other hand, the compensator of a point process can be expressed in term of the conditional inter-arrival time $(t_n - t_{n-1})|\mathscr{F}_{t_{n-1}}$ if the conditional distribution has support over $\mathbb{R}_+$. Under this condition, the intensity exists if and only if the conditional inter-arrival time is absolutely continuous. In this case, the intensity is given by [35, p.70]

$$\lambda_t = h_n(t - t_{n-1}) \text{ if } t \in (t_{n-1}, t_n] \tag{3.6}$$

$$h_n(t) = \frac{g_n(t)}{1 - G_n(t^-)},\ (t_n - t_{n-1})|\mathscr{F}_{t_{n-1}} \sim G_n \tag{3.7}$$

Once we know the intensity, we know the conditional distributions of all inter-arrival times and hence the complete distribution of the point process [37, p.233].

In the same way, we can define compensator and intensity for MPP. $A : \mathbb{R}_+ \times \mathscr{B}_Y \times \Omega \longrightarrow \mathbb{R}_+$ is called a compensator of the MPP N if $A(\bullet, B)$ is a compensator of $N(\bullet \times$

[11] Stochastic intensity is unique up to modification [36, p.31].

[12] Notice that stochastic intensity depends on the underlying filtration, so some text use the notation $\lambda(t|\mathscr{F}_t)$ but we will simply use $\lambda(t)$ and call it stochastic intensity or intensity when there is no confusion about the filtration.

[13] If A_t is absolutely continuous with respect to Lebesgue measure and λ_t is the Radon-Nikodym derivative (may not be predictable) then $\mathbb{E}(\lambda_t|F_{t^-})$ is a version of the stochastic intensity. Some authors only require the intensity to be adapted, but using the conditional expectation, one can always find a predictable version of intensity provided that the intensity has finite first moment.

$B)\ \forall B \in \mathscr{B}_Y$ and $A(t,\bullet)$ is a measure on $(Y,\mathscr{B}_Y)\ \forall t \in \mathbb{R}_+$. If $A(t,B) = \int_0^t \int_B \lambda(s)\nu(s,dy)ds$ a.s. where $\lambda(t)$ is non-negative and predictable, then $\lambda(t)$ is the stochastic intensity of the ground process of the MPP N and $\nu(t_n,dy) = \mathbb{P}(y_n \in dy|\mathscr{F}_{t_n^-})$ is the conditional mark distribution.

3.2.5 Random Time Change

If the filtration is usual, a point process N on $\mathbb{R}_+$ is simple and adapted, its intensity $\lambda(t)$ exists and $\int_0^\infty \lambda(s)ds = \infty$ a.s., then $\{\tilde{t}_n = \int_0^{t_n} \lambda(s)ds\}$ is a standard Poisson process (rate=1). The above theorem is called random time change theorem [40, 41] and is extremely useful in testing the goodness-of-fit of a stochastic intensity model. The random time change can be also used on MPP by focusing on the intensity of its ground process.

3.3 Hawkes Processes

A Hawkes process [42] is a point process where the stochastic intensity has an autoregressive form. For a nonlinear multivariate marked Hawkes process, the intensity $\lambda(t) = (\lambda_1(t),..,\lambda_d(t))$ of the ground process $N(t) = (N_1(t),..,N_d(t))$ is given by [14,15]

$$\lambda_i(t) = \Phi_i\left(\sum_{j=1}^{d}\int_{(-\infty,t)\times Y}\gamma_{ij}(t-s,y)N_j(ds\times dy),\ t\right) = \Phi_i\left(\sum_{t_n<t}\gamma_{i,w_n}(t-t_n,y_n),\ t\right) \tag{3.8}$$

$$\Phi_i : \mathbb{R}\times\mathbb{R}_+ \longrightarrow \mathbb{R}_+, \quad \gamma_{ij} : \mathbb{R}_+\times Y \longrightarrow \mathbb{R}, \quad N_j : \mathscr{B}(\mathbb{R}_+\times Y) \longrightarrow \mathbb{N}$$

where $w_n \in \{1,..,d\}$ denotes the type of t_n and Φ_i is known as a rate function. Consider the special case

$$\lambda_i(t) = \mu_i(t) + \sum_{j=1}^{d}\int_{(-\infty,t)\times Y}\gamma_{ij}(t-s,y)N_j(ds\times dy) = \mu_i(t) + \sum_{t_n<t}\gamma_{i,w_n}(t-t_n,y_n) \tag{3.9}$$

[14]Notice that some authors use γ_{ji}, so that the first index is the source type and the second index is the destination type.

[15]The Hawkes process only specifies the intensity for the ground process without any restriction on the mark distribution.

$$\mu_i : \mathbb{R}_+ \longrightarrow \mathbb{R}_+, \quad \gamma_{ij} : \mathbb{R}_+ \times Y \longrightarrow \mathbb{R}_+, \quad N_j : \mathscr{B}(\mathbb{R}_+ \times Y) \longrightarrow \mathbb{N}$$

i.e. $\Phi_i(x,t) = \mu_i(t) + x$. Such a Hawkes process determined by (3.9) is called linear, and $\mu_i(t)$ is called the base or background rate. The function γ_{ij} is called (marked) decay/exciting/fertility kernel and often $\gamma_{ij}(t,y)$ takes the separable form $\gamma_{ij}(t)g_{ij}(y)$ where g_{ij} is called mark impact kernel. Popular choices of decay kernel $\gamma_{ij}(t)$ include exponential $(\alpha_{ij}\mathrm{e}^{-\beta_{ij}t})$ [42], power law $(\alpha_{ij}(c_{ij}+t)^{-\beta_{ij}})$ [43] or Laguerre-type polynomial $(\sum_{k=0}^{K}\alpha_{ijk}t^k\mathrm{e}^{-\beta_{ijk}t})$ [44].

If the decay function is exponential with $\beta_{ij} = \beta_i$, the intensity $\lambda(t)$ and the vector $(N(t),\lambda(t))$ are both Markov processes[16] [20, 45]. Moreover, provided that $\mu_i(t) = \mu_i$, then $(\lambda_1(t),..,\lambda_d(t))$ satisfies the system of stochastic differential equations (SDE)

$$d\lambda_i(t) = \beta_i(\mu_i - \lambda_i(t))dt + \sum_{j=1}^{d}\alpha_{ij}dN_j(t) \tag{3.10}$$

This specification has the simple interpretation that the events of N_j which happened just before time t increase the intensity $\lambda_i(t)$ by $\alpha_{ij} \geq 0$ and thus trigger further events. Yet if the intensity $\lambda_i(t)$ is higher than μ_i, the first term becomes negative $(\beta_i > 0)$ and prevents the intensity from exploding, drawing it back to the mean level μ_i. In other words, the intensity $\lambda_i(t)$ is a mean-reverting process driven by its own point process. The Markov property and this intuitive interpretation may explain why the exponential decay kernels are so widely used.

For linear Hawkes processes, $\mu_i(t), \gamma_{ij}(t), g_{ij}(t)$ must be non-negative for all t, in order to ensure the positivity of $\lambda_i(t)$. As a result, unlike nonlinear Hawkes processes, linear Hawkes processes cannot model inhibitory effect (negative excitation). Nonetheless, the linear Hawkes processes are easier to handle, their properties are better understood and most importantly, they have a branching structure representation, which is extremely useful in simulation, estimation and interpretation of the models.

[16] N itself is not a Markov process as its intensity at time t depends on its full history before time t.

3.3.1 Branching Structure Representation

Linear Hawkes processes have a very elegant branching structure representation [46]. We describe here the version for the multivariate Hawkes processes with unpredictable marks (see [47]).

There are d types of immigrants arriving according to Poisson processes with rates $\mu_1, ..., \mu_d$. Each individual (descendant or immigrant) will carry an unpredictable mark when born or arrived. An individual of type j born at time t_n with mark y_n will give birth to an individual of type i according to a non-homogeneous Poisson process with rate $\gamma_{ij}(t - t_n, y_n)$. All the non-homogeneous Poisson processes are independent of each other.

Let $N_i(t)$ be the total number of individuals of type i born/arrived at or before time t under the above scenario, then $N(t) = (N_1(t), .., N_d(t))$ will follow the linear marked Hawkes process (3.9). This representation forms the basis of the Expectation Maximization (EM) algorithm in Section 3.4.2 and we will also see how it is used to measure the endogeneity of a point process in Section 3.5.4.

3.3.2 Stationarity

Considering a Hawkes process N with intensity (3.8) such that $\Phi_i(x,t) = \Phi_i(x)$, N has an unique stationary version[17] if either of the following conditions are satisified [46, 48]:

1. $\Phi_i(x)$ is k_i-Lipschitz[18] and the spectral radius[19] $\rho(A) < 1$ for the $d \times d$ matrix $A = [k_i \int_0^\infty |\gamma_{ij}(t)|dt]_{i,j}$
2. $\Phi_i(x)$ is Lipschitz, $\Phi_i(x) \leq M$, $\int_0^\infty |\gamma_{ij}(t)|dt < \infty$ and $\int_0^\infty t|\gamma_{ij}(t)|dt < \infty$

Technically speaking, N may have other non-stationary versions together with the stationary one; however, the non-stationary version will converges weakly to the stationary version when $t \to \infty$ (see [49] for exact meaning). Since the Hawkes process starts at $-\infty$, $N((0,t])$ will have the stationary distribution for all $t > 0$.

[17] See Appendix A for definition of stationarity of point processes.

[18] $f : \mathbb{R} \longrightarrow \mathbb{R}$ is called k-Lipschitz (k>0) if $|f(x) - f(y)| \leq k|x - y|$ $\forall x, y \in \mathbb{R}$.

[19] $\rho(A) = \max_i\{|\pi_i|\}$, $\{\pi_i\}$ are eigenvalues of A.

For the case of an exponential decay kernel ($\alpha_{ij}e^{-\beta_{ij}t}$), we have a simpler result. Let $A = [\int_0^\infty \alpha_{ij}e^{-\beta_{ij}t}dt]_{i,j} = [\alpha_{ij}/\beta_{ij}]_{i,j}$, then N has an unique stationary version under either of the following conditions [50]:

1. $\Phi_i(x) = \mu_i + x,\ \alpha_{ij} \geq 0,\ \beta_{ij}, \mu_i > 0,\ \rho(A) < 1$ (linear Hawkes process)
2. $\Phi_i(x) = \max(\mu_i + x, \varepsilon_i),\ \alpha_{ij} \in \mathbb{R},\ \beta_{ij}, \mu_i > 0,\ \varepsilon_i > 0,\ \rho(A) < 1$ (T-Hawkes process)
3. $\Phi_i(x) = \min(\mu_i + \exp(x), M_i),\ \alpha_{ij} \in \mathbb{R},\ \beta_{ij} > 0,\ M_i > \mu_i > 0$ (E-Hawkes process)

For the univariate linear case with $\mu = 0$, if there exists $r, R > 0, c \in (0, 1/2)$ such that $\int_0^\infty \gamma(t)dt = 1,\ \sup_{t\geq 0} t^{1+c}\gamma(t) \leq R,\ \lim_{t\to\infty} t^{1+c}\gamma(t) = r$, Brémaud and Massoulié [51] show that there exists a unique stationary non-trivial Hawkes process having such an intensity and he calls it critical Hawkes process or Hawkes process without ancestors ($\mu = 0$).

3.3.3 Convergence

In this section, we state the results about the convergence of Hawkes processes. A properly scaled linear Hawkes process will converge weakly to a Brownian diffusion when the spectral radius of decay functions' L^1-norm is less than one [23]. When the spectral radius is close to one in a certain sense, it converges to the integrated Cox-Ingersoll-Ross (CIR) process [52]. For the non-linear Hawkes processes, we only have the result for the univariate case and the sufficient conditions depends on the Lipschitz constant of Φ [53].

Law of Large Numbers for Multivariate Linear Hawkes processes

Assuming the model (3.9) without marks, if the spectral radius $\rho(A) < 1$ where $A = [\int_0^\infty \gamma_{ij}(t)dt]_{i,j}$, then [23]

$$\sup_{t\in[0,1]} \left\| \frac{N(nt)}{n} - t(I_d - A)^{-1}\mu \right\| \xrightarrow[n\to\infty]{\text{a.s./}L^2} 0^{20,21} \tag{3.11}$$

where $\mu = (\mu_1, .., \mu_d)$. When $d = 1$ and we take $t = 1$, it implies

$$\frac{N(T)}{T} \xrightarrow[T\to\infty]{\text{a.s./}L^2} \frac{\mu}{1 - \int_0^\infty \gamma(t)dt} \tag{3.12}$$

Functional Central Limit Theorem for Multivariate Linear Hawkes processes

Assuming the model (3.9) without marks, $N = (N_1, .., N_d)$, if the spectral radius $\rho(A) < 1$ where $A = [\int_0^\infty \gamma_{ij}(t)dt]_{i,j}$ and $\int_0^\infty \sqrt{t}\gamma_{ij}(t)dt < \infty \; \forall i, j$, then [23]

$$\sqrt{n}\left(N(\bullet n)/n - \bullet(I_d - A)^{-1}\mu\right) \xrightarrow[n\to\infty]{\text{weak}} (I_d - A)^{-1}\Sigma^{1/2}W(\bullet)^{22} \tag{3.13}$$

$$\Sigma = \text{diag}((I_d - A)^{-1}\mu)^{23} \tag{3.14}$$

$W(\bullet)$ is standard $d-$dimensional Brownian Motion

Functional Central Limit Theorem for Univariate Non-linear Hawkes processes

Assuming the model (3.8) without marks and $d = 1$, if $\gamma(t)$ is decreasing, $\int_0^\infty t\gamma(t)dt < \infty$, $\Phi(x,t) = \Phi(x)$ is increasing and k-Lipschitz, $\int_0^\infty k\gamma(t)dt < 1$ then [53]

$$\sqrt{n}\,(N(\bullet n)/n - \bullet\nu) \xrightarrow[n\to\infty]{\text{weak}} \sigma W(\bullet) \tag{3.15}$$

$$\sigma^2 = \mathbb{E}((N([0,1]) - \nu)^2) + 2\sum_{n=1}^{\infty} \mathbb{E}((N([0,1]) - \nu)(N([n,n+1]) - \nu) \tag{3.16}$$

$$\nu = \mathbb{E}(N([0,1])) \tag{3.17}$$

Convergence of Nearly Unstable Univariate Linear Hawkes processes

Considering the linear model (3.9) without marks and $d = 1$, $N(T)/T \longrightarrow \mu/(1 - \int_0^\infty \gamma(t)dt)$ when $\int_0^\infty \gamma(t)dt < 1$ by (3.12) while $N(T)/T$ explodes when $\int_0^\infty \gamma(t)dt = 1$.

[20] A sequence of random variables $X_n \xrightarrow[n\to\infty]{a.s.} X$ if $\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$

[21] A sequence of random variables $X_n \xrightarrow[n\to\infty]{L^2} X$ if $\lim_{n\to\infty} E(|X_n - X|^2) = 0$

[22] A sequence of probability measure P_n converges weakly to P if $\int_\Omega f dP_n \to \int_\Omega f dP$ for all bounded continuous function f. A sequence of stochastic process $X_n : \Omega \to D[0,1]$ converges weakly (in distribution) to X if the law of $X_n(P_n \circ X_n^{-1})$ converges weakly to law of $X(P \circ X^{-1})$ in the sense of probability measure, D[0,1] is the Skorokhod space of càdlàg (right continuous with left limits) functions (see [54, 55]).

[23] $v \in \mathbb{R}^d$, $\text{diag}(v) = [a_{ij}]_{d\times d}$, $a_{ii} = v_i$, $a_{ij} = 0 \; \forall i \neq j$.

However, Jaisson and Rosenbaum [52] find that the properly scaled Hawkes process converges to the integrated Cox-Ingersoll-Ross (CIR) process when one has a sequence of decay kernel $\gamma^{(n)}(t)$ whose integral converges to 1 at the speed of n^{-1} (see 3.21). More precisely, let

$$\lambda^{(n)}(t) = \mu + \int_{(0,t)} \gamma^{(n)}(t-s) dN^{(n)}(s), \quad \mu > 0, \quad \gamma^{(n)}(t) = \alpha^{(n)} \gamma(t) \tag{3.18}$$

$$\gamma : \mathbb{R}_+ \longrightarrow \mathbb{R}_+, \quad \int_0^\infty \gamma(t) dt = 1, \quad \int_0^\infty t\gamma(t) dt = m < \infty \tag{3.19}$$

$$\int_0^\infty |\gamma\prime(t)| dt < \infty, \ \sup_{t \in [0,\infty)} |\gamma\prime(t)| < \infty \tag{3.20}$$

$$\alpha^{(n)} \in [0,1), \quad \lim_{n\to\infty} \alpha^{(n)} = 1, \quad \lim_{n\to\infty} n(1-\alpha^{(n)}) = c > 0 \tag{3.21}$$

$$\psi^{(n)}(t) = \sum_{k=1}^{\infty} \gamma^{(n)\otimes k}(t), \quad \rho^{(n)}(t) = \frac{n\psi^{(n)}(nt)}{\int_0^\infty \psi^{(n)}(t) dt}, \quad |\rho^{(n)}(t)| \leq M \ \forall n \ \forall t \tag{3.22}$$

where $f^{\otimes k}$ denotes the k-fold self-convolution of f. If the sequence of Hawkes process $N^{(n)}$ has intensity $\lambda^{(n)}$ satisfying (3.18-3.22), then the scaled intensity converges to the CIR process and the scaled Hawkes process converges to the integrated CIR process [52] as follows:

$$(1-\alpha^{(n)}) \lambda^{(n)}(n\bullet) \xrightarrow[n\to\infty]{\text{weak}} X(\bullet) \tag{3.23}$$

$$(1-\alpha^{(n)}) \frac{N^{(n)}(n\bullet)}{n} \xrightarrow[n\to\infty]{\text{weak}} \int_0^\bullet X(s) ds \tag{3.24}$$

$$dX_t = \frac{c}{m}(\mu - X_t) dt + \frac{\sqrt{c}}{m} \sqrt{X_t} dW_t, \quad X_0 = 0 \tag{3.25}$$

3.4 Statistical Inference of Hawkes Processes

3.4.1 Simulation

In this section, we will give an overview of the algorithms to simulate Hawkes processes. Assume we know all the parameters in the functional form of $\mu(t)$ and $\gamma(t,y)$, our goal is to simulate the points $(t_1,y_1),(t_2,y_2),...$ on the interval [0,T].

If the marks distribution depends only on t_n, we can simply generate y_n conditioned on the generated t_n. Next, t_{n+1} can be generated from the intensity $\lambda(t)$ for $t > t_n$ which

depends on $\{(t_1,y_1),...,(t_n,y_n)\}$. If the distribution of y_n also depends on $\{(t_{n-1},y_{n-1}),(t_{n-2},y_{n-2}),...\}$, the algorithms can be modified accordingly.

Inverse CDF Transform

The first simulation algorithm for Hawkes processes appears in [56]. Suppose the intensity is governed by the univariate Hawkes model in (3.9). Let t_i be the arrival time and $\tau_n = t_n - t_{n-1}$ be the inter-arrival time. By (3.6), $\lambda(t) = h_n(t-t_{n-1})$ for $t \in (t_{n-1},t_n]$ where $h_n(t) = g_n(t)/(1-G_n(t^-))$ and g, G are the conditional pdf, cdf of τ_n given $\mathscr{F}_{t_{n-1}}$. If $G_n(t)$ is continuous, $h_n(t)$ is simply the hazard function, and it can be shown that

$$G_n(\tau_n) = 1 - \exp\left(-\int_0^{\tau_n} h_n(s)ds\right) = 1 - \exp\left(-\int_{t_{n-1}}^{t_{n-1}+\tau_n} \lambda(s)ds\right) \tag{3.26}$$

Given t_{n-1}, we can generate $t_n = t_{n-1} + \tau_n$ by inverse cdf transform $\tau_n = G_n^{-1}(U)$, $U \sim \text{Unif}(0,1)$. However, the inversion needs to be done numerically, so this method is largely superceded by Ogata's modified thinning which we now discuss.

Ogata's Modified Thinning

Ogata [57] introduces the modified thinning method which does not require numerical inversion. The algorithm is based on the following theorem. Let $N = (N_1,..,N_d)$ be a multivariate point process with intensity $(\lambda_1,..,\lambda_d)$ such that $\sum_{i=1}^d \lambda_i(t) \leq \lambda^*(t)\ \forall t\ a.s.$ ($\lambda^*(t)$ is an exogenously chosen deterministic rate function) and N^* is the univariate non-homogeneous Poisson process with intensity $\lambda^*(t)$. If each point t_n in N^* is given a mark y_n such that $\mathbb{P}(y_n = i) = \lambda_i(t_n)/\lambda^*(t_n),\ i = 1,..,d$, then $(N_1^*,..,N_d^*)$ has the same distribution as $(N_1,..,N_d)$.

The following algorithm generates a d-dimensional multivariate Hawkes process such that $\lambda_i(t)$ is decreasing between points and $|\lambda_i(t) - \lambda_i(t^-)| \leq \alpha_i\ \forall t$.

Ogata's Modified Thinning [57]

1. $n = 1,\ t_0 = 0$

2. Generate $\tau_n \sim \text{Exp}(\lambda_n^*)$ for some $\lambda_n^* \geq \sum_{i=1}^{d}(\lambda_i(t_{n-1}) + \alpha_i)$
   (Exp(λ) is exponential distribution with rate λ)
3. Let $t_n = t_{n-1} + \tau_n$
4. Generate $U_n \sim \text{Unif}(0,1)$
5. if $U_n \in (\sum_{i=0}^{k-1} \lambda_i(t_n)/\lambda_n^*, \sum_{i=0}^{k} \lambda_i(t_n)/\lambda_n^*]$ for some $k \in \{1,..,d\}$ return t_n and the point is of type k (also generate $y_n|t_n$ for MPP) , else discard t_n (but keep the value for use in next generation)
6. $n = n+1$, goto step 2

Simulation by Branching Structure

This method generates points using the branching structure representation of linear marked Hawkes processes. Type j immigrants arrive according to a non-homogeneous Poisson with rate $\mu_j(t)$. Next the type-j parent arriving at t_n produces type-i descendants according to non-homogeneous Poisson with rate $\gamma_{ij}(t-t_n, y_n)$ and the generation is repeated for each descendant until all of them exceed the pre-defined time T. Since all the non-homogeneous Poisson processes are independent, the generations can be done in parallel, making this algorithm very suitable for parallel implementation.

Simulation by the Branching Structure [58]

1. Generate non-homogeneous Poisson processes with intensities $\mu_i(t),\ i = 1,..,d$ on $[0,T]$
2. For each points t_n, generate $y_n|t_n$
3. Suppose t_n is of type j, generates type-i descendants according to non-homogeneous Poisson process with intensity $\gamma_{ij}(t-t_n, y_n)$ on $[t_n, T],\ i = 1,..,d$
4. repeat step 2, 3 for all descendants

The non-homogeneous Poisson process with intensity $\mu(t)$ on $[0,T]$ can be generated using Lewis' thinning algorithm [59]

1. generate $N \sim \text{Poisson}(\mu^*)$ for some $\mu^* \geq \max_{t\in(0,T]}\mu(t)$
2. generate $U_n \sim \text{Unif}(0,1),\ n = 1,..,N$

3. $T_n = U_{(n)}T,\ n = 1,..,N$ ($\{U_{(n)}\}$ is the order statistics of $\{U_n\}$)
4. generate $V_n \sim \text{Unif}(0,1),\ i = 1,..,N$
5. return T_n if $V_n \leq (\mu(T_n)/\mu^*),\ n = 1,..N$; otherwise discard T_n

3.4.2 Estimation

Suppose we observe a point process from 0 to T and collect the event times and marks $\{(t_1,y_1),..,(t_N,y_N)\}$, now we would like to estimate the functions $\mu(t)$ and $\gamma(t,y)$ in the intensity $\lambda(t)$ which drives the process $N(t)$. We will summarize the various methods appearing in the literature, but so far the focus is on unmarked process. In the special case where the marks are independent and identically distributed (IID), the mark distribution can be estimated separately from the point process.

If we assume $\mu(t)$ and $\gamma(t)$ have some parametric representations, we can use Maximum Likelihood Estimation (MLE), Expectation Maximization (EM), or Generalized Method of Moments (GMM) to estimate the parameters. Otherwise, we need to rely on some advanced non-parametric techniques to estimate the whole function curves.

Maximum Likelihood Estimation (MLE)

The log likelihood of a Hawkes process is given by [56]

$$\log(L(\theta)) = \sum_{i=1}^{d} \left(-\int_0^T \lambda_i(t;\theta)dt + \int_0^T \log(\lambda_i(t;\theta))dN_i(t) \right) \tag{3.27}$$

In the case of multivariate linear Hawkes process, it becomes

$$\begin{aligned}\log(L(\theta)) = &-\int_0^T \left(\sum_{i=1}^{d} \mu_i(t;\theta) \right) dt - \sum_{n=1}^{N} \int_{t_n}^{T} \left(\sum_{i=1}^{d} \gamma_{i,w_n}(t-t_n;\theta) \right) dt \\ &+ \sum_{n=1}^{N} \log \left(\mu_{w_n}(t_n;\theta) + \sum_{t_m<t_n} \gamma_{w_n,w_m}(t_n-t_m;\theta) \right)\end{aligned} \tag{3.28}$$

The parameters in the Hawkes process can be estimated by maximizing the log-likelihood. However, the numerical optimization is problematic as the log likelihood function is usually quite flat (see [60, fig.2,3]) and may have a lot of local maxima (see [60, fig.4]).

Expectation Maximization (EM)

For linear Hawkes process, the estimation can also be done via Expectation Maximization (EM) [61, 62] as in [60, 63–66]. EM is a variant of MLE where part of the data is missing. In the branching structure representation, the missing data is the parents which produce the descendants. Let z_n denotes the index of the parent of t_n and w_{z_n} represents the type of the parent of t_n. If $z_n = m$ and $w_{z_n} = j$, that means t_n is produced by the type j point t_m. When z_n is 0, t_n is an immigrant. Also we define $w_0 = 0$, $\gamma_{i,0}(t) = \mu_i(t)$ and $t_0 = 0$ to simplify the expression. Suppose $\{t_n, w_n, z_n\}$ are known, since each generation is an independent Poisson process, the complete data log likelihood is

$$\log(L(\theta)) = \sum_{n=0}^{N}\sum_{i=1}^{d}\Bigg\{ -\int_{t_n}^{T}\gamma_{i,w_n}(t-t_n;\theta)dt + \sum_{t_m>t_n}\log(\gamma_{i,w_n}(t_m-t_n;\theta))\mathbb{1}(z_m=n)\mathbb{1}(w_m=i)\Bigg\} \tag{3.29}$$

$$\begin{aligned} Q(\theta|\theta^{(k)}) =& \mathbb{E}^{\theta^{(k)}}(\log(L(\theta))|\{(t_k,w_k)\}) \\ =& \sum_{n=0}^{N}\sum_{i=1}^{d}\Bigg\{ -\int_{t_n}^{T}\gamma_{i,w_n}(t-t_n;\theta)dt + \sum_{t_m>t_n}\log(\gamma_{i,w_n}(t_m-t_n;\theta)) \\ & \mathbb{P}^{\theta^{(k)}}(z_m=n|\{(t_k,w_k)\})\mathbb{1}(w_m=i)\Bigg\} \end{aligned} \tag{3.30}$$

$$\mathbb{P}^{\theta^{(k)}}(z_m=n|\{(t_k,w_k)\})\mathbb{1}(w_m=i) = \frac{\gamma_{i,w_n}(t_m-t_n;\theta^{(k)})\mathbb{1}(w_m=i)}{\sum_{l=0}^{m-1}\gamma_{i,w_l}(t_m-t_l;\theta^{(k)})} \tag{3.31}$$

The EM algorithm can be implemented as follows:

1. $k = 0$ and choose an initial guess $\theta^{(0)}$
2. E-step: compute $Q(\theta|\theta^{(k)}) = \mathbb{E}^{\theta^{(k)}}(\log(L(\theta))|\{(t_k,w_k)\})$
3. M-step: compute $\theta^{(k+1)} = \text{argmax}_\theta\ Q(\theta|\theta^{(k)})$
4. $k = k+1$, repeat E-step and M-step until $\theta^{(k)}$ converges (e.g. $\|\theta^{(k+1)} - \theta^{(k)}\| < \varepsilon$)

In general, the optimization in M-step need to be solve numerically but when the decay kernel has the exponential form $\alpha_{ij}\beta_{ij}\exp(-\beta_{ij}t)$, Olson and Carley [66] suggest a closed form approximate iteration.

$$\mu_i^{(k+1)} = \frac{\sum_{m=1}^{N}\mathbb{P}^{\theta^{(k)}}(z_m=0|\{(t_k,w_k)\})\mathbb{1}(w_m=i)}{T} \tag{3.32}$$

$$\alpha_{ij}^{(k+1)} = \frac{\sum_{n=1}^{N}\sum_{m=n+1}^{N}\mathbb{P}^{\theta^{(k)}}(z_m = n|\{(t_k,w_k)\})\mathbb{1}(w_m = i, w_n = j)}{\sum_{n=1}^{N}\mathbb{1}(w_n = j)} \tag{3.33}$$

$$\beta_{ij}^{(k+1)} = \frac{\sum_{n=1}^{N}\sum_{m=n+1}^{N}\mathbb{P}^{\theta^{(k)}}(z_m = n|\{(t_k,w_k)\})\mathbb{1}(w_m = i, w_n = j)}{\sum_{n=1}^{N}\sum_{m=n+1}^{N}(t_m - t_n)\mathbb{P}^{\theta^{(k)}}(z_m = n|\{(t_k,w_k)\})\mathbb{1}(w_m = i, w_n = j)} \tag{3.34}$$

$$\mathbb{P}^{\theta^{(k)}}(z_m = n|\{(t_k,w_k)\})\mathbb{1}(w_m = i, w_n = j) = \frac{\alpha_{ij}^{(k)}\beta_{ij}^{(k)}\exp(-\beta_{ij}^{(k)}(t_m - t_n))\mathbb{1}(w_m = i, w_n = j)}{\mu_i^{(k)} + \sum_{l=1}^{m-1}\alpha_{i,w_l}^{(k)}\beta_{i,w_l}^{(k)}\exp(-\beta_{i,w_l}^{(k)}(t_m - t_l))} \tag{3.35}$$

$$\mathbb{P}^{\theta^{(k)}}(z_m = 0|\{(t_k,w_k)\})\mathbb{1}(w_m = i) = \frac{\mu_i^{(k)}\mathbb{1}(w_m = i)}{\mu_i^{(k)} + \sum_{l=1}^{m-1}\alpha_{i,w_l}^{(k)}\beta_{i,w_l}^{(k)}\exp(-\beta_{i,w_l}^{(k)}(t_m - t_l))} \tag{3.36}$$

In addition, the summation $\sum_{l=1}^{m-1}\alpha_{i,w_l}^{(k)}\beta_{i,w_l}^{(k)}\exp(-\beta_{i,w_l}^{(k)}(t_m - t_l))$ can be truncated after $\exp(-\beta_{i,w_l}^{(k)}(t_m - t_l))$ has decayed to a small value. The speed of EM is reported to be 10-100 times faster than MLE using Nelder-Mead and more importantly, MLE does not converges within 500 iterations in practically all test cases while EM does [66].

Generalized Method of Moments (GMM)

Another method for statistical estimation beyond MLE is the *Generalized Method of Moments*[24] [67]. The idea is to find the parameters which minimize the difference between theoretically moments in term of the unknown parameters and the empirical moments computed directly from the data. If we have more moments than the number of parameters, the method involves solving a weighted least squares problem.

Da Fonseca and Zaatour [68] obtain analytic moment expressions by restricting the process to be univariate with exponential kernel and making use of the Markov property in this special case. The authors claim that this method is extremely fast but no speed comparison result is provided.

[24] Although GMM is consistent under some mild regularity conditions, unlike MLE, it is not asymptotic efficient among the class of consistent estimators.

Nonparametric Estimation

Without assuming any parametric form for $\mu(t)$ nor $\gamma(t)$, some nonparametric methods are developed recently to estimate the whole base rate and decay kernel functions. Similar to parametric estimation, penalized MLE or GMM is used to find the function with desirable characteristics (e.g. smooth functions or sparse coefficients). Nonetheless, the nonparametric method, which involves finding the unknown functions in infinite-dimensional spaces, requires extensive computational effort and the underlying statistical construction is usually much more involved than any parametric counterpart.

To the best of our knowledge, the first attempt in nonparameteric estimation of Hawkes process is by Gusto and Schbath [69] in 2005. The authors express the kernel function of the multivariate Hawkes process using B-splines [70] with equally spaced knots. The log likelihood function involving the basis coefficients are then maximized numerically and the optimal order for the B-splines basis as well as number of knots are determined using AIC criteria [71].

Instead of B-splines, Reynaud-Bouret and Schbath [72] find the function within the space of piecewise constant functions which minimizes the empirical L^2-norm between the true and estimated kernel functions. The method is later extended to multivariate cases [73] with a Lasso type penalty [74] in the minimization objective.

On the other hand, the base and kernel functions can be estimated nonparametrically in each M-step of EM as in [75], within the space $C^1(\mathbb{R}_+)$ with Good's penalty $\|(\sqrt{\gamma})'\|_2$ [76]. Using calculus of variations, the solution of each penalized maximization in M-step can be found by solving the Euler-Lagrange equation numerically.

Instead of EM, Zhou et al. [77] use Minorize-Maximization (MM) algorithm [78], in which EM is a special case. In the E-step of MM algorithm, $Q(\bullet|\theta^{(k)})$ is any lower bound of the objective function $\log(L(\bullet))$ such that $Q(\theta^{(k)}|\theta^{(k)}) = \log(L(\theta^{(k)}))$. It is then iteratively maximized in the M-steps until convergence. In [77], the kernel functions are expressed using a finite number of basis functions which are estimated nonparametrically in M-step by solving the Euler-Lagrange equation.

Another approach is to use moment matching to find the kernel function as in [79–81]. In Bacry and Muzy [81], the authors derive the conditional moment density $\mathbb{E}(dN_i(t)|dN_j(0)=1,dy)$ of multivariate marked Hawkes process as the solution of Wiener-Hopf equation [82] involving $\mu_i, \gamma_{ij}(t), g_{ij}(y)$ for the case that the mark impact kernel is piecewise constant. The conditional moment density can be estimated by any kernel density estimation technique and the Wiener-Hopf equation can be solved numerically via the Nyström method [83].

3.4.3 Hypothesis Testing

Random Time Change

The classical method to test the goodness-of-fit of a point process model on $\mathbb{R}_+$ is Ogata's residual analysis [43]. Ogata calls $\{\tilde{t}_n = \int_0^{t_n} \hat{\lambda}(s)ds\}$ the residual process[25] and according to the random time change theorem, the residual process should be close to a standard Poisson process if the estimated intensity $\hat{\lambda}(t)$ is close to the true intensity $\lambda(t)$. The hypothesis that $\{\tilde{t}_n\}$ is a standard Poisson process can be tested by the following methods:

1. QQ Plot [85] of $\{\tilde{\tau}_n = \tilde{t}_n - \tilde{t}_{n-1}\}$ vs Exp(1).
2. Kolmogorov-Smirnov Test [86–88] to test $\tilde{\tau}_n \sim$ Exp(1)
3. Ljung-Box Test [89] to test the lack of serial correlation of $\{\tilde{\tau}_n\}$

Approximate Thinning

Another method to test goodness-of-fit is by thinning, which does not require integration of the intensity function. It is useful if the intensity function is estimated non-parametrically. However, the thinned residual process is only approximately a Poisson process.

[25] The terminology is not standard, Baddeley et al. [84] refer $\{N(t_n) - \int_0^{t_n} \hat{\lambda}(s)ds\}$ as residual in order to extend the concept to higher dimension.

By Ogata's modified thinning [57], we know that if there exists $b > 0$ such that $b \leq \lambda(t) \forall t$ and we keep point t_n with probability $b/\lambda(t_n)$, the thinned point process is a homogeneous Poisson process with rate b. However, the infimum b of the intensity function is often close to 0, making the number of points in the thinned process very small and the test to have little power. A remedy is to use approximate thinning [90] as follows: choose an integer $k \ll N$, select a point from $\{t_1, .., t_N\}$ with probability of selecting t_n proportional to $\lambda(t_n)^{-1}$. Repeat the selection (without replacement) until k points are selected. The resulting k points will be approximately a homogeneous Poisson process.

3.5 Applications of Hawkes processes

After the groundwork of basic theory and statistical inference for Hawkes processes, we now unleash their power to model HF data. First, the readers are reminded about how diverse the notion of stock trading frequency can be. According to the Trade And Quote database (TAQ), between 9:30am to 4:00pm on May 2, 2014, there were 11 million quote changes (limit + cancellation + market orders) and 0.3 million trades (market orders) for SPDR S&P 500 ETF (SPY). In other words, on average there are 460 quote changes and 13 trades per second. If we take a snapshot every 5 minutes as in [32, 33], we will only use 0.03% of trade data and 0.0007% of quote data. In comparison, Pathfinder Bancorp (PBHC) only has 306 quote changes and 11 trades on the the same day, which means there is a 35 minutes lag between trades on average and thus the 5 minutes snapshots will just give a series of repeated information. Regardless of the sampling frequency, we are likely to get some misleading result if we analyze the asynchronous data from a portfolio of liquid and illiquid stocks using models with fixed intervals.

The construction of multivariate point processes shows that each variate can have a completely different arrival intensity $\lambda_i(t)$ from its peers'. Nonetheless, the multivariate Hawkes process can still model the dependence structure easily via the $\gamma_{ij}(t)$'s, which are estimated by duly considering all the asynchronous data of highest frequency without any re-sampling.

Order arrivals and price changes are unarguably two of the most important elements in high frequency trading. Using Hawkes processes, we can estimate their distributions conditioned on all the historical HF asynchronous data, enabling us to give a more accurate real time prediction of future event occurrences. In the following subsections, we are going to highlight some of the literature which take advantage of Hawkes processes to model HF data.

3.5.1 Modeling Order Arrivals

Bowsher [50][26] is the first to use Hawkes processes to model order arrivals. He uses nonlinear Hawkes processes to allow for inhibitory effect and he considers two rate functions $\Phi_i(x,t) = \mu_i(t) + \exp(x)$ and $\Phi_i(x,t) = \max(\mu_i(t) + x, \varepsilon_i),\ \varepsilon_i > 0$, where both of them guarantee that the stochastic intensity will be strictly positive at all times. For the deterministic base rate $\mu_i(t)$, he exploits a piecewise linear function with knots at 9:30, 10:00, 11:00,...,16:00 while the decay kernel is the exponential function without marks. In addition, an extra term is included to represent the spillover effects from the previous trading day.

Bowsher uses Maximum Likelihood Estimation (MLE) to estimate the parameters for the bivariate point process of trade and quote of General Motor (GM), trading on NYSE between 5 July 2000 to 29 August 2000. The model is found to be decent according to the goodness-of-fit test using random time change.

Instead of modeling arrivals of all trades and quotes, Large [22] uses Hawkes processes to model only the arrivals of aggressive orders, which are market orders depleting the queue and limit orders falling inside the bid-ask spread, in order to study the *resiliency* of the LOB. A LOB is called resilient if it reverts to its generic shape promptly after large trades. The idea is that when a large trade causes the bid-ask spread to widen, the arrival intensity of aggressive limit orders in a resilient LOB will surge so that the gap will be filled very

[26]Though Bowsher's paper was published in 2007, the first draft appeared in 2002.

quickly. In order words, the cross-excitation effect $\gamma_{ij}(t)$ from aggressive market orders to aggressive limit orders should be reasonably large for a resilient LOB.

In addition to market orders and limit orders, Large also includes the cancellations of limit orders as well as limit orders falling outside the best quotes. Therefore, he builds a 10-variate linear marked Hawkes process with exponential decay and mark impact kernel to fit the HF data of Barclays (BARC), trading on LSE between 2 Jan 2002 to 31 Jan 2002. The result shows that the widening of bid-ask spread indeed pumps up the intensities of aggressive limit orders, causing the gap to be filled very quickly and hence making the LOB resilient.

More examples of applications of Hawkes processes to order arrivals include the following papers: Muni Toke and Pomponio [91] use similar approach as Large [22] to model trades-through, namely market orders which deplete the best queues and consume at least one share in the second best. Muni Toke [92] designs a more realistic market simulator using Hawkes processes with exponential kernel for order arrivals. Shek [93], Fauth and Tudor [94] apply the Hawkes processes with volume mark on stock and FX market respectively. Hewlett [95] models the arrival of market orders with Hawkes processes for single period market making. Finally, Alfonsi and Blanc [96], Jaisson [97] tackle the problem of optimal execution with market orders coming from multivariate Hawkes processes.

3.5.2 Modeling Price Jumps

Single Asset

Traditionally the events of price jumps are modeled by Poisson processes, which suffer from the drawbacks mentioned in the introduction section. Again, Hawkes processes can be applied to model price jumps, which often delineate clustering, self- and cross-excitation behavior.

Bacry et al. [98] use Hawkes processes to model the price jumps, resulting in a model which can reproduce the microstructure noise [99], Epps effect [100] and jump clustering,

while maintaining the coarse scale limit of Brownian diffusion. In their model, the trade price $X(t)$ has the dynamics

$$X(t) = N_1(t) - N_2(t) \tag{3.37}$$

where $N(t) = (N_1(t), N_2(t))$ is a bivariate linear Hawkes process with exponential decay kernel. $N_1(t), N_2(t)$ represents the total number of upward and downward jumps respectively. The authors make additional assumptions that the Hawkes process N has only cross-excitation and coefficients are symmetric in order to simplify computation:

$$\lambda_1(t) = \mu + \int_{(0,t)} \gamma(t-s) dN_2(t), \quad \lambda_2(t) = \mu + \int_{(0,t)} \gamma(t-s) dN_1(t) \tag{3.38}$$

$$\gamma(t) = \alpha \mathrm{e}^{-\beta t} \tag{3.39}$$

According to the model, when X jumps up(down), λ_2 (resp. λ_1) increases, causing the probability of jumping down (resp. up) to increase. Such cross-linkage generates the effect of microstructure noise where the trade price is bouncing between best bid and best ask.

Due to the bid-ask bounce, it is well-known that the realized variance (annualized) increases when the sampling frequency increases [101].

$$V(\tau) = \mathbb{E}\left(\frac{1}{T} \sum_{n=0}^{T/\tau} (X((n+1)\tau) - X(n\tau))^2 \right) \tag{3.40}$$

$$= \frac{2\mu}{1-\alpha/\beta} \left(\frac{1}{(1+\alpha/\beta)^2} + \left(1 - \frac{1}{(1+\alpha/\beta)^2}\right) \frac{1-\mathrm{e}^{-(\alpha+\beta)\tau}}{(\alpha+\beta)\tau} \right) \tag{3.41}$$

Such an effect can be easily demonstrated by computing the expected realized variance (3.41) of the jump model (3.37) and the result with $\mu = 0.16, \alpha = 0.024, \beta = 0.11$ is shown in Fig.3.1. The authors apply the model to Euro-Bund futures and find a very good fit between the observed and theoretical realized variance under this highly simplified model.

Let $Y(t) = X(nt)$, then $Y(t)$ is a coarse scale version of $X(t)$. For example, if t in X is in micro second and $n = 60000$, then t in Y will be in minute. When we look at the trade price in a low frequency setting, Bacry et al. [23] show that the macroscopic Hawkes jump model goes back to the classical model of Brownian motion due to the functional central limit theorem for linear Hawkes process (3.13). Assuming that $\int_0^\infty \gamma(t)dt < 1$, then

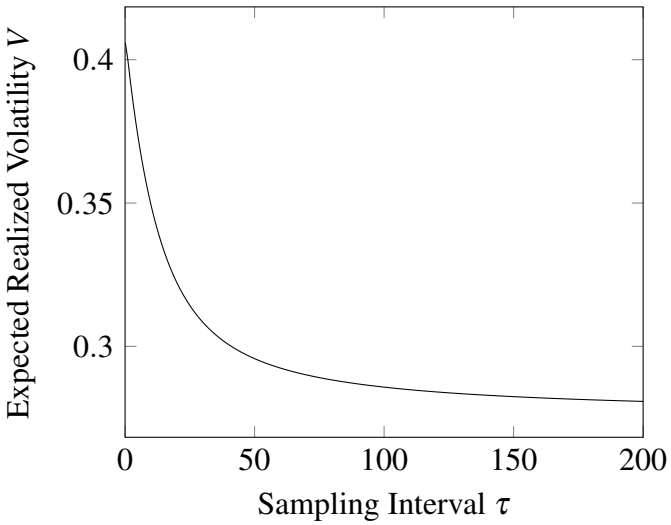

Figure 3.1. Volatility Signature Plot of Hawkes Jump Model

$$\frac{X(n\bullet)}{\sqrt{n}} \xrightarrow[n\to\infty]{\text{weak}} \sigma W(\bullet),\ \sigma^2 = \frac{2\mu}{(1-\int_0^\infty \gamma(t)dt)(1+\int_0^\infty \gamma(t)dt)^2} \tag{3.42}$$

It is interesting to see how the macroscopic variance σ^2 is related to the microscopic base rate μ and cross-excitation $\gamma(t)$. As $\int_0^\infty \gamma(t)dt$ approaches 1, the variance goes to ∞.

Jaisson and Rosenbaum [52] extend the model of Bacry et al. [98] for the case of nearly unstable Hawkes process, where $\int_0^\infty \gamma(t)dt \simeq 1$, by constructing a sequence of kernel functions whose integrals converge to 1 at the speed of n^{-1}. They show that the properly scaled price process converges to Brownian diffusion with Heston stochastic volatility [102]. The full result is stated below.

$$X^{(n)}(t) = N_1^{(n)}(t) - N_2^{(n)}(t) \tag{3.43}$$

$$\lambda_1^{(n)}(t) = \mu + \int_0^t \gamma_1^{(n)}(t-s)dN_1^{(n)}(s) + \int_0^t \gamma_2^{(n)}(t-s)dN_2^{(n)}(s) \tag{3.44}$$

$$\lambda_2^{(n)}(t) = \mu + \int_0^t \gamma_2^{(n)}(t-s)dN_1^{(n)}(s) + \int_0^t \gamma_1^{(n)}(t-s)dN_2^{(n)}(s) \tag{3.45}$$

$$\left(\int_0^\infty \gamma_1^{(n)}(t)dt + \int_0^\infty \gamma_2^{(n)}(t)dt\right) < 1, \quad \gamma_i^{(n)}(t) = \alpha^{(n)}\gamma_i(t) \tag{3.46}$$

$$\gamma_i : \mathbb{R}_+ \longrightarrow \mathbb{R}_+, \quad \int_0^\infty (\gamma_1(t) + \gamma_2(t))dt = 1 \tag{3.47}$$

$$\int_0^\infty t(\gamma_1(t)+\gamma_2(t))dt = m < \infty, \quad \int_0^\infty |\gamma_i\prime(t)|dt < \infty, \quad \sup_{t\in[0,\infty)} |\gamma_i\prime(t)| < \infty \tag{3.48}$$

$$\alpha^{(n)} \in [0,1), \quad \lim_{n\to\infty} \alpha^{(n)} = 1, \quad \lim_{n\to\infty} n(1-\alpha^{(n)}) = c > 0 \tag{3.49}$$

$$\psi^{(n)}(t) = \sum_{k=1}^{\infty} \left(\gamma_1^{(n)} + \gamma_2^{(n)}\right)^{\otimes k}(t), \quad \rho^{(n)}(t) = \frac{n\psi^{(n)}(nt)}{\int_0^\infty \psi^{(n)}(t)dt}, \quad |\rho^{(n)}(t)| \le M \ \forall n \ \forall t \tag{3.50}$$

Under the conditions of (3.43 - 3.50),

$$\frac{X^{(n)}(n\bullet)}{n} \xrightarrow[n\to\infty]{\text{weak}} Y(\bullet) \tag{3.51}$$

$$dY_t = \frac{\sqrt{V_t}}{1-\int_0^\infty |\gamma_1(t)-\gamma_2(t)|dt} dW_t^1, \quad Y_0 = 0 \tag{3.52}$$

$$dV_t = \frac{c}{m}\left(\frac{2\mu}{c} - V_t\right)dt + \frac{\sqrt{V_t}}{m} dW_t^2, \quad V_0 = 0 \tag{3.53}$$

Conditions (3.43 - 3.45) is just a bivariate Hawkes model (with both self- and cross-excitation) but now we have a different $\gamma_i^{(n)}(t)$ for each n that use to scale the time. The rest are the regularity conditions similar to the univariate nearly unstable Hawkes process (3.25) and the most important one is (3.49) which states that $\alpha^{(n)}$ converges to one at the speed of n^{-1}. However, the interesting result is that instead of converging to an integrated CIR, the price dynamics formed by the difference between two Hawkes processes converges to a stochastic volatility model.

Two Assets

To model the Epps effect, Bacry et al. [98] consider the two-asset case with prices $(X_1(t), X_2(t))$ given by

$$X_1(t) = N_1(t) - N_2(t), \quad X_2(t) = N_3(t) - N_4(t) \tag{3.54}$$

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{4} \int_{(0,t)} \alpha_{ij} \exp(-\beta(t-s))dN_j(s), \quad i = 1,..,4 \tag{3.55}$$

$(N_1(t),..,N_4(t))$ is a 4-variate Hawkes process with exponential kernel where $\beta_{ij}=\beta$. The coupling of excitation effects is constrained to have the form

$$\alpha=\begin{pmatrix}0 & \alpha_{12} & \alpha_{13} & 0\\ \alpha_{12} & 0 & 0 & \alpha_{13}\\ \alpha_{31} & 0 & 0 & \alpha_{34}\\ 0 & \alpha_{31} & \alpha_{34} & 0\end{pmatrix} \tag{3.56}$$

In this case, there is a closed form representation for the realized correlation, which vanishes when the sampling interval goes to zero (Epps effect).

If we assume $\mu_1=\mu_2,\ \mu_3=\mu_4,\ \alpha_{12}=\alpha_{34}=0,\ (\int_0^\infty \gamma_{13}(t)dt)(\int_0^\infty \gamma_{31}(t)dt)<1$, then the macroscopic bivariate asset prices converges to correlated Brownian diffusion [23]

$$\frac{1}{\sqrt{n}}\begin{pmatrix}X_1(n\bullet)\\ X_2(n\bullet)\end{pmatrix}\xrightarrow[n\to\infty]{\text{weak}}\frac{\sqrt{2}\begin{pmatrix}\sqrt{\nu_1}W_1(\bullet)+\sqrt{\nu_2}\int_0^\infty \alpha_{13}(t)dtW_2(\bullet)\\ \sqrt{\nu_1}\int_0^\infty \alpha_{31}(t)dtW_1(\bullet)+\sqrt{\nu_2}W_2(\bullet)\end{pmatrix}}{(1-(\int_0^\infty \gamma_{13}(t)dt)(\int_0^\infty \gamma_{31}(t)dt))^{3/2}} \tag{3.57}$$

$$\nu_1=\mu_1+\left(\int_0^\infty \gamma_{13}(t)dt\right)\mu_3,\ \nu_2=\mu_3+\left(\int_0^\infty \gamma_{31}(t)dt\right)\mu_1 \tag{3.58}$$

$(W_1(\bullet),W_2(\bullet))$ is standard 2-dimensional Brownian motion

This convergence result gives us an explicit formula to estimate the macroscopic correlation from the asynchronous high frequency data.

As a final remark, under this jump representation, the observed trade price is not some hidden continuous fair value plus some microstructure noise as in [103]. It is the result of the trading actions between buyers (N_1,N_3) and sellers (N_2,N_4) on a fixed price grid. There is no such thing as high frequency volatility or correlation since prices are not diffusions but pure-jump processes in the HF scale. Volatility and correlation are only meaningful when we look at the coarse scale diffusion approximation at low frequency, but those low frequency representation parameters can be computed directly from the high frequency jump model characteristics.

3.5.3 Modeling Jump-Diffusion

Duffie et al. [104, 105] propose the affine jump-diffusion $X(t)$, which has the following structure.[27]

$$dX(t) = (k_0(t) + k_1(t)X(t))dt + (h_0(t) + h_1(t)X(t))dW(t) + \zeta dN(t) \tag{3.59}$$

$$\lambda(t) = a_0(t) + a_1(t)X(t) \tag{3.60}$$

The jump intensity $\lambda(t)$ of $N(t)$ is an affine function of $X(t)$, which depends on the Brownian motion $W(t)$ and the jump process $N(t)$, with jump size ζ drawn from a fixed distribution. When $k_0 = \beta\theta, k_1 = -\beta, h_0 = h_1 = 0, a_0 = 0, a_1 = 1, \zeta = \alpha$, we can see that $\lambda(t) = X(t)$ and $d\lambda(t) = \beta(\theta - \lambda(t))dt + \alpha dN(t)$. Hence in this case $N(t)$ is the Hawkes process with exponential kernel.

Zhu [106] derives some convergence results when ζ is a constant and the diffusion part is a CIR process.

$$dX(t) = \beta(\mu - X(t))dt + \sigma\sqrt{X(t)}dW(t) + \alpha dN(t) \tag{3.61}$$

$$\lambda(t) = a_0 + a_1X(t) \tag{3.62}$$

Aït-Sahalia et al. [107] model the contagion of financial crisis with the Hawkes jump-diffusion where the price dynamic $X_i(t)$ is given by

$$dX_i(t) = \mu_i dt + \sqrt{V_i(t)}dW_i^X(t) + Z_i(t)dN_i(t) \tag{3.63}$$

$$dV_i(t) = \kappa_i(\theta_i - V_i(t))dt + \eta_i\sqrt{V_i(t)}dW_i^V(t) \tag{3.64}$$

The diffusion part is the Heston stochastic volatility model and the jump part is a multivariate Hawkes process modeling the clustering and propagation of jumps among multiple assets. $Z_i(t)$ corresponds to the jump size and direction.

3.5.4 Measuring Endogeneity (Reflexivity)

In term of the Hawkes branching structure representation of events arrivals, Filimonov et al. [108, 109] portray immigrants as exogenous news whereas the descendants are en-

[27] We show the one dimensional case for simplicity.

dogenous incidents. In the context of price movements in the stock or commodity market, immigrants are the price discovery due to orders from informed traders, who react to external information, whereas the descendants are the destabilizing ripples created by noise traders, who engage in herding [110], momentum trading [111] and parasite trading [112] etc.

Under the univariate linear Hawkes model with exponential decay kernel and constant base rate, the expected number of direct descendants per individual (branching coefficient) is given by

$$n = \int_0^\infty \gamma(s)ds = \int_0^\infty \alpha e^{-\beta s} ds = \alpha/\beta \tag{3.65}$$

For a given immigrant, the expected number of descendants in all generations is $n + n^2 + n^3 + ... = n/(1-n)$ if $n < 1$, so the ratio of descendants (non-immigrants) vs total population is

$$\frac{\text{descendants}}{\text{descendants + immigrant}} = \frac{n/(1-n)}{n/(1-n)+1} = n \tag{3.66}$$

Therefore, the branching coefficient n characterizes the degree of endogenous feedback activities while the base rate μ measures the arrival rate of exogenous information.

Using E-mini S&P futures as proxy, Filimonov and Sornette [108] find that the level of endogeneity (reflexivity[28]) n in the US market has gone from 0.3 in 1998 to 0.7 in 2007. Moreover, for the flash crash of May 6, 2010, n reached a peak of 0.95.

Nonetheless, using the power law decay kernel, Hardiman et al. [114] challenge the result of Filimonov and Sornette [108] by reporting that the branching ratio n has always been close to one since 1998 and that the market could be a critical Hawkes process [51], but Filimonov and Sornette [115] refute that the power law kernel is sensitive to outliers in addition to other counter arguments. Later Hardiman and Bouchaud [116] devised a nonparametric estimation of the branching ratio in term of moments, but the result depends heavily on the window size used in the empirical moment computation.

[28] Filimonov and Sornette [108] borrow this term from Soros [113].

3.6 A Brief History of Hawkes processes

Hawkes processes are proposed by Hawkes [42] in 1971 in order to model contagious processes like epidemics, neuron firing and particle emission, where the occurrences of events trigger further events. Although the intensity of Cox processes [117], introduced in 1955, are stochastic, they are determined before the events are unfolded[29]. In order to portray the excitation behavior in contagious processes, Hawkes extends the model in such a way that the intensity is a predictable stochastic process with an intuitive autoregressive form, which allows it to adapt to events that happen over time.

After Hawkes' seminal paper, a number of theoretical developments include the branching structure representation by Hawkes and Oakes [46] in 1974, the Markov property for the intensity with exponential decay kernel by Oakes [45] in 1975, the MLE for Hawkes processes by Ozaki [56] in 1979, Ogata's modified thinining simulation algorithm [57] in 1981, nonlinear Hawkes processes by Brémaud and Massoulié [48] in 1996, Nonparametric Estimation by Gusto and Schbath [69] in 2005, EM for Hawkes processes by Veen and Schoenberg [60] in 2008 and the functional central limit theorem for Hawkes processes by Bacry et al. [23], Jaisson and Rosenbaum [52], Zhu [53] in 2013.

Although the first application of Hawkes processes to earthquake occurrences appeared in 1982 [44], it was not until 1988 [43] that Hawkes processes received much attention. Since then, the versatility of Hawkes model was leveraged in seismology [43, 118], finance (risk and credit default modeling) [119, 120], social networks [121, 122], neuroscience [123, 124] etc. (see [125, 126] for more applications).

The use of Hawkes processes in HF financial data modeling starts with Bowsher [50] in 2007[30] and then Large [22] in the same year. Both authors exploit Hawkes processes so as to describe the interactions between order arrivals of different types. Later, Aït-Sahalia et al. [107] and Bacry et al. [98] employ Hawkes processes to reproduce clustering in jump diffusion and pure-jump process representation of stock prices in 2010 and 2013 respectively. On the other hand, an interesting idea appears in Filimonov and Sornette

[29] For processes on $\mathbb{R}_+$, it means their intensity is $\mathscr{F}_0$ measurable.

[30] Though Bowsher's paper was published in 2007, the first draft appeared in 2002.

[108] in 2012 which utilizes the branching coefficient of linear Hawkes model to measure the level of endogenous activities in the US stock market, though the debate about the validity of the result is still going on [114–116].

4. JOINT MODELING OF PRICES AND ORDER ARRIVALS

4.1 Introduction

The joint modeling of prices and order arrivals has not yet received much attention in the mathematical finance literature. All the existing market-making models assume that prices are independent from order arrivals, which is clearly far from the truth. For example, when a large buy market order depletes the best ask queue, the ask price can only move up. However, when you model the price and order arrival as two independent processes, roughly half of the time price will go down with the arrival of buy order. Such an unrealistic scenario produces a large phantom profit for the market maker and cause the average profit of the market-making model to be overstated.

Besides, price is commonly modeled as diffusion whereas in high-frequency setting price is a pure-jump process living on a fixed price grid. Moreover, each jump has two components, namely time and magnitude. Diffusion can only describe the magnitude but cannot model the timing of the jump and its dependency on the order arrival history. The simplest way is to replace the diffusion with a Poisson process. However, it has been documented that order flow has clustering and long memory properties [18, 19] and so a self-exciting point process may be a better alternative.

Bacry and Muzy [127] model the mid-price and market buy and sell order as a 4-variate Hawkes processes (T^+, T^-, N^+, N^+) where $T^\pm, N^\pm$ denotes the buy/sell market order and upward/downward jump of mid-price respectively. Although the price jump and market order are now correlated via the cross-excitation, the problem of price goes down with a buy market order still remains. In addition, multivariate Hawkes processes is by definition a *simple* point process, meaning that price and market order cannot jump at the same time, but this assumption contradicts the fact that price and market order *always* jump together upon the arrival of aggressive market order that depletes the whole best bid/ask queue.

The authors work around the issue with some tricks that introducing a small delay Δt to the price jump and describe the excitation effect from market order to price jump as Dirac delta function as Δt goes to 0.

In this chapter, we make use of a simple technique developed in 60's [128] for Poisson random variables to model the dependency structure of jump processes. Namely if $N_1 = N_1' + U',\ N_2 = N_2' + U'$, where (N_1', N_2', U') are independent Poisson random variables, then (N_1, N_2) have the so-called *bivariate* Poisson distribution. Co-jumps as well as the dependency of (N_1, N_2) are modeled by the common component U'. In our model, (N_1', N_2', U') are not some latent processes but in fact some directly observable order book events which we are going to describe shortly.

As a remark, our model does not aim to describe the full dynamics of a limit order book, instead we strive to depict the relationship of the best bid/ask price with the order arrivals using a tractable point process model. In many trading applications, traders will only consider limit orders in the best bid/ask as the chance of execution beyond the best quotes is simply too low (e.g. less than 3% for E-mini S&P future [21]). The utmost concern of the traders is when the market orders arrive to fill their limit orders, so the activities outside the best quotes have little value to them. Moreover, the practice of quote stuffing[1] [129] and spoofing[2] [130] render the information content of the limit order book questionable, especially beyond the best quotes. As a result, the benefit of a full order book model may not justify the added nontrivial complexity and this may explain the emergence of reduced-form models which focus only on the top of the book [131, 132].

4.2 Joint Modeling of Prices and Order Arrivals

We classify all orders which affect the top of book into twelve categories according to type (limit, market, cancellation), direction (buy, sell) and aggressiveness similar to [18, 22]. We follow the definition of [22] that aggressive orders are the ones which move the bid or ask price. To be more precise, aggressive market order is the one which completely

[1] Rapid placement and cancellation of large amount of limit orders.

[2] submission of limit orders to create an illusion of demand/supply imbalance

Table 4.1.
Classification of orders

Type	Order Arrival Event	Bid Price	Ask Price
1	aggressive market buy	0	+
2	aggressive market sell	-	0
3	aggressive limit buy cancellation	-	0
4	aggressive limit sell cancellation	0	+
5	non-aggressive market buy	0	0
6	non-aggressive market sell	0	0
7	non-aggressive limit buy	0	0
8	non-aggressive limit sell	0	0
9	non-aggressive limit buy cancellation	0	0
10	non-aggressive limit sell cancellation	0	0
11	aggressive limit buy	+	0
12	aggressive limit sell	0	-

depletes the best bid or ask queue, aggressive limit order is the one with limit price inside the bid-ask spread and aggressive cancellation is the one which cancel the whole bid or ask queue.[3] As one can see, the best bid or best ask will move at the exact instant of the execution or placement of an aggressive order.

Let $N(t) = (N_1(t), ..., N_{12}(t))$ denotes the multivariate *simple*[4] point process of the twelve types of order, and $M_a(t)$, $M_b(t)$, $S_a(t)$, $S_b(t)$ denote the *buy* market orders, *sell* market orders, *ask* price and *bid* price respectively. Assuming the tick size δ of the stock is fixed and each price jump is of size one tick[5]. It is not hard to realize the following straight forward but important relation.

$$M_a(t) = N_1(t) + N_5(t) \tag{4.1}$$

$$M_b(t) = N_2(t) + N_6(t) \tag{4.2}$$

$$S_a(t) = S_a(0) + (N_1(t) + N_4(t) - N_{12}(t))\delta \tag{4.3}$$

$$S_b(t) = S_b(0) + (N_{11}(t) - N_2(t) - N_3(t))\delta \tag{4.4}$$

[3]Notice that an aggressive order can be a small order if the size of queue is small at the time of execution.

[4]For simplicity, we assume no two types of orders can arrive at the same time, but of course when the exchange has multiple servers, it can accept buy and sell orders at the same time. However, the probability that two orders arrive at the exact same instant (Nasdaq timestamp are down to nanosecond) is close to zero.

[5]We will look at the general model where price can jump more than one tick in Section 4.6.

Through those remarkably simple equations (4.1)-(4.4), we can observe the dependency of price and order arrivals via the common components N_1 and N_2. For instance, when there is aggressive buy market order (type 1), both the buy market order point process $M_a(t)$ and ask price $S_a(t)$ will jump at the same time (co-jump), but they can also jump separately upon the arrival of other order types. From the equations, we can also recognize ask price cannot go down with a buy market order (type 1 or 5).

This is in sharp contrast with the approach in [127] where the prices and market orders are modeled as a *simple* multivariate point process. The first issue is that the prices *always* jump at the instant of aggressive orders, but under the *simple* point process assumption in [127], price jump and order arrival happen at the same time with probability zero. Second, even if the ask price can be described as positively correlated with the buy market order, there is still a non-zero probability that ask price goes *down* after a *buy* market order.

The bid-ask spread $\Delta(t) = S_a(t) - S_b(t)$ in our model is given by

$$\Delta(t) = (S_a(0) - S_b(0)) + (N_1(t) + N_2(t) + N_3(t) + N_4(t) - N_{11}(t) - N_{12}(t))\delta \quad (4.5)$$

If we do not put any constrain on the point processes $(N_1(t), .., N_{12}(t))$, there is no guarantee that $\Delta(t)$ will be always greater than or equal to δ. Nonetheless if we look carefully at type 11 and type 12 orders (limit orders inside the spread), they can only appear when $\Delta(t^-) > \delta$ and as a result $\Delta(t)$ will not shrink below δ. To put this constrain in the model, the simplest way to restrict the intensity of the point process, using the fact that when the intensity is 0 at time t, the probability that an event happens at time t is zero [36, T12, p.31]. Let $\lambda_i(t)$ denote the stochastic intensity of the point process $N_i(t)$, we thus impose the condition[6] that

$$\lambda_{11}(t) = \mu_{11}(t)\mathbb{1}(\Delta(t^-) > \delta) \quad (4.6)$$

$$\lambda_{12}(t) = \mu_{12}(t)\mathbb{1}(\Delta(t^-) > \delta) \quad (4.7)$$

where $\mu_{11}(t)$ and $\mu_{12}(t)$ is any predictable non-negative stochastic processes.

[6]We can also use the equivalent form $\lambda_i(t) = \mu_i(t)\mathbb{1}(\Delta(t) > \delta)$, see [36, T10, p.29] for proof.

4.3 One-Tick Bid-Ask Spread Model

For highly liquid stocks, the bid-ask spread $\Delta(t)$ can be one tick 99% of the time [133]. This kind of order book *resiliency* can be easily reproduced by a large $\mu_{11}(t)$ or $\mu_{12}(t)$. N_{11} and N_{12} will be *activated* right after the bid-ask spread is widen by the shock N_1, N_2, N_3 or N_4. Provided that $\mu_{11}(t)$ or $\mu_{12}(t)$ is large enough, the aggressive limit order will arrive quickly to fill the gap in the bid-ask spread. In fact, $\mu_{11}(t) + \mu_{12}(t)$ may be used as a first order measure[7] of the resiliency of the order book.

In this case, bid and ask prices will always move in lockstep, so we only need to model the mid-price $S(t)$, which is given by (4.9)

$$S(t) = (S_a(t) + S_b(t))/2 \tag{4.8}$$

$$= S(0) + (N_1(t) - N_2(t) - N_3(t) + N_4(t) + N_{11}(t) - N_{12}(t))\delta/2 \tag{4.9}$$

If we assume that $N_{11} \simeq N_{12}$, the model enjoys a further simplification (4.10) as N_{11}, N_{12} cancel each other. We would like to stress that since N_{11}, N_{12} can only jump when the bid-ask spread is larger than one tick, the overall impact on the price dynamics for liquid stocks is limited even if the approximation $N_{11} \simeq N_{12}$ is crude.

$$S(t) = S(0) + (N_1(t) - N_2(t) - N_3(t) + N_4(t))\delta/2 \tag{4.10}$$

4.4 Multivariate Hawkes Process

The simplest way to complete the specification of the model is to assume $N_1, .., N_{12}$ are independent Poisson processes; nonetheless the assumption of independent arrival is often rejected in the literature. In the classical study, Biais et al. [18] document the so-called *diagonal effect*, which means the next order is more likely to have same type as the previous one, and similar findings are also reported in [19, 134, 135]. In particular, Tóth et al. [19] show that the dominant reason for such kind of persistent order flow is due to order splitting, rather than herding. Since large institutions often split large orders into

[7] A better measure of resiliency is the time for the queue to return back to its *normal* size after a shock, but $\mu_{11}(t), \mu_{12}(t)$ will be irrelevant after the *first* order is placed inside the spread.

small pieces and then continuously trade for hours or even days, the observed long memory characteristics makes perfect sense.

Because of the above reasons, we will use a point process with stochastic intensity to represent our order arrivals. As described in Chapter 3, multivariate Hawkes process [42] is a popular self-exciting point process with intensity $\lambda_i(t)$ depends on its own history in the following form.

$$\lambda_i(t) = \mu_i(t) + \sum_{j\geq 1} \int_{(-\infty,t)} \gamma_{ij}(t-s) dN_j(s) = \mu_i(t) + \sum_{t_n<t} \gamma_{i,w_n}(t-t_n) \tag{4.11}$$

where w_n is the type of the point t_n and $\mu_i(t), \gamma_{ij}(t)$ are some non-negative functions. Using the Hawkes model, the diagonal effect can be generated by large $\gamma_{ii}(t)$, which increases the intensity of type i order upon its own arrivals.

For type $11,12$ orders, their intensities are constrained to be 0 when $\Delta(t^-) = \delta$. One possibility is to multiply (4.11) with the indicator function $\mathbb{1}(\Delta(t^-) > \delta)$, resulting in the so-called constrained Hawkes process [136]. The constrained Hawkes process will consist of a latent unconstrained Hawkes process and the observed process will be the thinned unconstrained Hawkes process with probability $\mathbb{1}(\Delta(t^-) > \delta)$. However, we opt to avoid this complicated structure and we will assume the intensities of type $11,12$ are of the form.

$$\lambda_{11}(t) = \mu_{11}\mathbb{1}(\Delta(t^-) > \delta) \tag{4.12}$$

$$\lambda_{12}(t) = \mu_{12}\mathbb{1}(\Delta(t^-) > \delta) \tag{4.13}$$

where μ_{11}, μ_{12} are non-negative constants. $N(t) = (N_1(t), ..., N_{10}(t))$ will be modeled as a 10-variate (unconstrained) Hawkes process.

The reason for not using constrained Hawkes process is that since only the first limit order placed inside the spread are aggressive and the bid-ask spread is one tick most of the time for liquid stocks, the total number of type $11,12$ orders is tiny and hence their excitation effects on other order types are insignificant. On the other hand, the intensities of aggressive limit order N_{11}, N_{12} are induced by the presence of gap in the bid-ask spread rather than the arrivals of other orders shortly before; therefore the cross-excitation effects on N_{11}, N_{12} from other orders will be negligible compared with the parameters μ_{11} or μ_{12} (see [22]).

Without the cross-linkage of the constrained components, the estimation will be much easier. The multivariate Hawkes process (without $11, 12$) can be estimated using Maximum Likelihood Estimation (MLE) or Expectation Maximization (EM) and μ_{11} and μ_{12} can be simply estimated by the inverse of average arrival times of the type $11, 12$ orders during the *active* period.

In the sequel, we will focus on the Markovian exponential kernel with a constant base rate.

$$\gamma_{ij}(t) = \alpha_{ij}\exp(-\beta_i t) \qquad (\alpha_{ij} \geq 0,\ \beta_i > 0) \tag{4.14}$$

$$\mu_i(t) = \mu_i \tag{4.15}$$

The reason is that under this condition, the intensity $\lambda(t) = (\lambda_1(t), .., \lambda_{10}(t))$ is a Markov process [20, 45] (see also 5.2.2) and can be expressed in the form of SDEs (4.16). Hence, $(\lambda(t), N(t))$ is also a Markov process even though $N(t)$ depends on its whole history.

$$d\lambda_i(t) = \beta_i(\mu_i - \lambda_i(t))dt + \sum_{j=1}^{d} \alpha_{ij} dN_j(t) \tag{4.16}$$

4.5 Scaling Limit

Assuming the one-tick spread model with $N_{11} \simeq N_{12}$ (4.10), the change in mid-price equals

$$\Delta S(t) = S(t) - S(0) = (N_1(t) - N_2(t) - N_3(t) + N_4(t))\delta/2 = a^\top N(t) \tag{4.17}$$

where $a = (\delta/2)[1,\ -1,\ -1,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0]^\top$ and $N(t) = [N_1(t), ..., N_{10}(t)]^\top$. The following theorem shows that under certain regularity conditions, the pure-jump mid-price converges weakly to a Brownian motion.

Theorem 4.5.1 (Scaling Limit of Mid-price)
Let $\Gamma = [\int_0^\infty \gamma_{ij}(t)dt]_{i,j}$ *and* $\Sigma = diag((I_{10} - \Gamma)^{-1}\mu)$. *If the spectral radius of* $\Gamma < 1$ *and* $\int_0^\infty \sqrt{t}\gamma_{ij}(t)dt < \infty\ \forall i, j$, *then*

$$\sqrt{n}\left(\Delta S(\bullet n)/n - \bullet a^\top (I_{10} - \Gamma)^{-1}\mu\right) \xrightarrow[n\to\infty]{\text{weak}} a^\top (I_{10} - \Gamma)^{-1}\Sigma^{1/2} W(\bullet) \tag{4.18}$$

Proof Since the spectral radius of $\Gamma < 1$ and $\int_0^\infty \sqrt{t}\gamma_{ij}(t)dt < \infty$, by the functional central limit theorem for Hawkes process [23, Corollary 1], we have

$$\sqrt{n}\Big(N(\bullet n)/n - \bullet(I_{10} - \Gamma)^{-1}\mu\Big) \xrightarrow[n\to\infty]{\text{weak}} (I_{10} - \Gamma)^{-1}\Sigma^{1/2}W(\bullet) \tag{4.19}$$

Notice that $\Delta S(t) = a^\top N(t)$ and $f(N) = a^\top N$ is continuous (a is a constant vector). Hence result follows from the continuous mapping theorem. ■

In other words,

$$\Delta S(nt) \simeq (a^\top(I_{10} - \Gamma)^{-1}\mu)(nt) + (a^\top(I_{10} - \Gamma)^{-1}\Sigma^{1/2})(\sqrt{n}W(t)) \tag{4.20}$$

On a macroscopic scale, $\Delta S(nt)$ behaves like a diffusion with variance $(a^\top(I_{10} - \Gamma)^{-1}$ $\Sigma(I_{10} - \Gamma^\top)^{-1}a)(nt)$. Therefore under a coarser time scale, our pure-jump model agrees with the diffusion model commonly used to characterize longer term price movement.

4.6 General Model with Volume and Jump Size

Though less common, the price jumps caused by the aggressive orders can be larger than one tick. Therefore in addition to the random jump times $\tau_n \in \mathbb{R}_+$, we add the random marks $\xi_n \in \mathbb{N}$ which correspond to the jump sizes (in ticks) of the aggressive orders ($\xi_n = 0$ for non-aggressive orders). Moreover, we add another mark v_n corresponds to the volumes of the orders.

The multivariate marked Hawkes process now becomes $N_i(dt \times dv \times d\xi)$ and the compensator will be of the form $\lambda_i(t)\mu_i(t, dv \times d\xi)dt$ where $\lambda_i(t)$ is the intensity of the ground process $N_i(dt \times \mathbb{R}_+ \times \mathbb{N})$ and $\mu_i(t, dv \times d\xi)$ is the conditional mark (jump and volume) distribution. To simplify the presentation, we will use the notation $N_i(dt \times dv) = N_i(dt \times dv \times \mathbb{N})$ and $(N_i + N_j)(dt \times dv \times d\zeta) = N_i(dt \times dv \times d\zeta) + N_j(dt \times dv \times d\zeta)$. The joint model of prices and market orders now becomes:

$$M_a(dt \times dv) = (N_1 + N_5)(dt \times dv) \tag{4.21}$$

$$M_b(dt \times dv) = (N_2 + N_6)(dt \times dv) \tag{4.22}$$

$$S_a(t) = S_a(0) + \delta \int_{(0,t]\times\mathbb{R}_+\times\mathbb{N}} \xi (N_1 + N_4 - N_{12})(dr \times dv \times d\xi) \tag{4.23}$$

$$S_b(t) = S_b(0) + \delta \int_{(0,t]\times\mathbb{R}_+\times\mathbb{N}} \xi (N_{11} - N_2 - N_3)(dr \times dv \times d\xi) \tag{4.24}$$

For the one-tick spread model with $N_{11} \simeq N_{12}$, it becomes

$$M_a(dt \times dv) = (N_1 + N_5)(dt \times dv) \tag{4.25}$$

$$M_b(dt \times dv) = (N_2 + N_6)(dt \times dv) \tag{4.26}$$

$$S(t) = S(0) + (\delta/2) \int_{(0,t]\times\mathbb{R}_+\times\mathbb{N}} \xi (N_1 - N_2 - N_3 + N_4)(dr \times dv \times d\xi) \tag{4.27}$$

4.7 Numerical Illustration

In this section, we provide some numerical results in fitting the Hawkes model to the tick data from Nasdaq TotalView-ITCH. TotalView is a message level database where all the order additions, cancellations and executions[8] are recorded in an amazing nano-second precision and this allows us to observe the extreme short burst of arrivals not possible in database time-stamped in seconds. However, we would like to point out there is a major issue in our analysis with only Nasdaq data. *Nasdaq exchange only matches about 20% of the total volume in the US equity market, so the classification of orders using only Nasdaq data will potentially overstate the number of aggressive orders and understate the number of non-aggressive orders.* We remind our readers the result in this section is more of an illustration in fitting the Hawkes model rather than an empirical study of US equity market.

4.7.1 Summary Statistics

The background activities of the stock market is well-known to be higher around open (9:30am) and close (4:00pm) than that during the mid-day [22, 50]. One way is to fit a spline [70] to the base rate $\mu(t)$; however we would like to make our lives easier by focusing on the period 12:00 - 2:00pm and assuming the $\mu(t)$ is constant over this period.

[8]Nasdaq splits a market order into pieces (with same time-stamp) when it is executed against several limit orders but we combine them as one single market order during our processing. Also, market orders executed against hidden limit orders are ignored in this study.

Table 4.2.
Summary statistics of QQQ on June 2, 2014 (12pm-2pm)

Type	Description	Count	% Count	Avg Rate (/s)	Avg Size
1	aggressive market buy	190	0.08%	0.0264	1,040
2	aggressive market sell	171	0.08%	0.0238	667
3	aggressive limit buy cancellation	164	0.07%	0.0228	685
4	aggressive limit sell cancellation	106	0.05%	0.0147	544
5	non-aggressive market buy	312	0.14%	0.0433	670
6	non-aggressive market sell	210	0.09%	0.0292	874
7	non-aggressive limit buy	54,187	24.19%	7.5260	1,006
8	non-aggressive limit sell	53,332	23.80%	7.4072	1,061
9	non-aggressive limit buy cancellation	58,481	26.10%	8.1224	897
10	non-aggressive limit sell cancellation	56,265	25.11%	7.8146	927
11	aggressive limit buy	316	0.14%	0.0439	842
12	aggressive limit sell	315	0.14%	0.0438	681

Figure 4.2 shows the summary statistics of Powershares QQQ Trust (Nasdaq 100 index ETF) (QQQ). It is well-known that the activities of limit order revisions dominate the market and it is reflected clearly in the table that 99.2% of the orders are type 7-10.

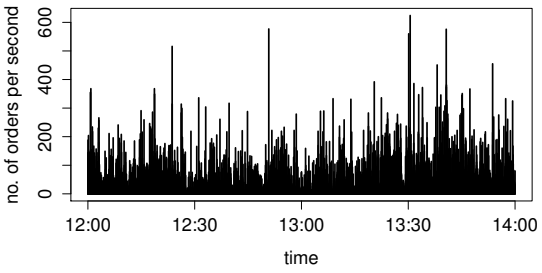

Figure 4.1. Activities of QQQ (all order types) on June 2, 2014

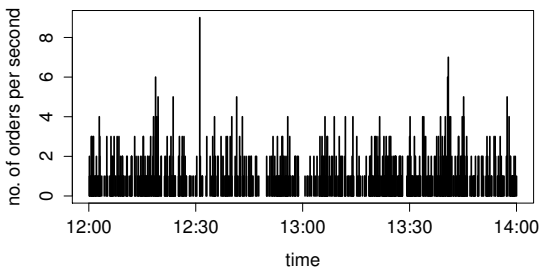

Figure 4.2. Activities of QQQ (type 1-6) on June 2, 2014

Figure 4.1 shows the total no. of orders (all types) per second of QQQ during 12-2pm, June 2, 2014, but the chart is not very useful as it is masked by the activities of limit order revisions. Figure 4.2 shows that QQQ has less than 10 type 1-6 orders per second albeit QQQ is one of the most liquid stocks in US.

4.7.2 Volume Distribution

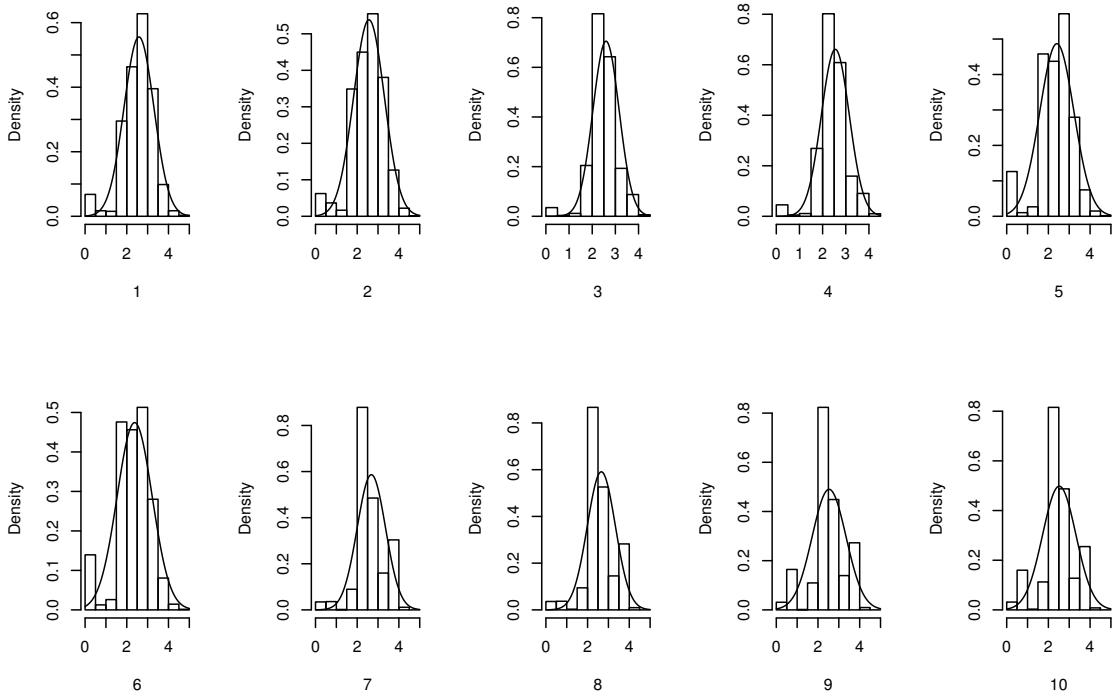

Figure 4.3. Histogram of $\log_{10}$(volume)

Before we study the arrivals timing of the orders, let's look at the histograms of the $\log_{10}$(volume) in Figure 4.3. The volume distributions of all order types are similar, with $\log_{10}$(volume) roughly normal with mean 2.5 (300 shares) and standard deviation 0.7. However, there is a higher-than-expected concentration of small orders in type 1,2,5,6,9,10. We believe they are the so-called *pinging* orders [137] that HFT firms use to detect hidden orders. If we remove those tiny orders, the mean and standard deviation of the $\log_{10}$(volume) change to around 2.7 (500 shares) and 0.5 and the histogram and QQ plots without tiny orders are shown in Figure 4.4-4.5.

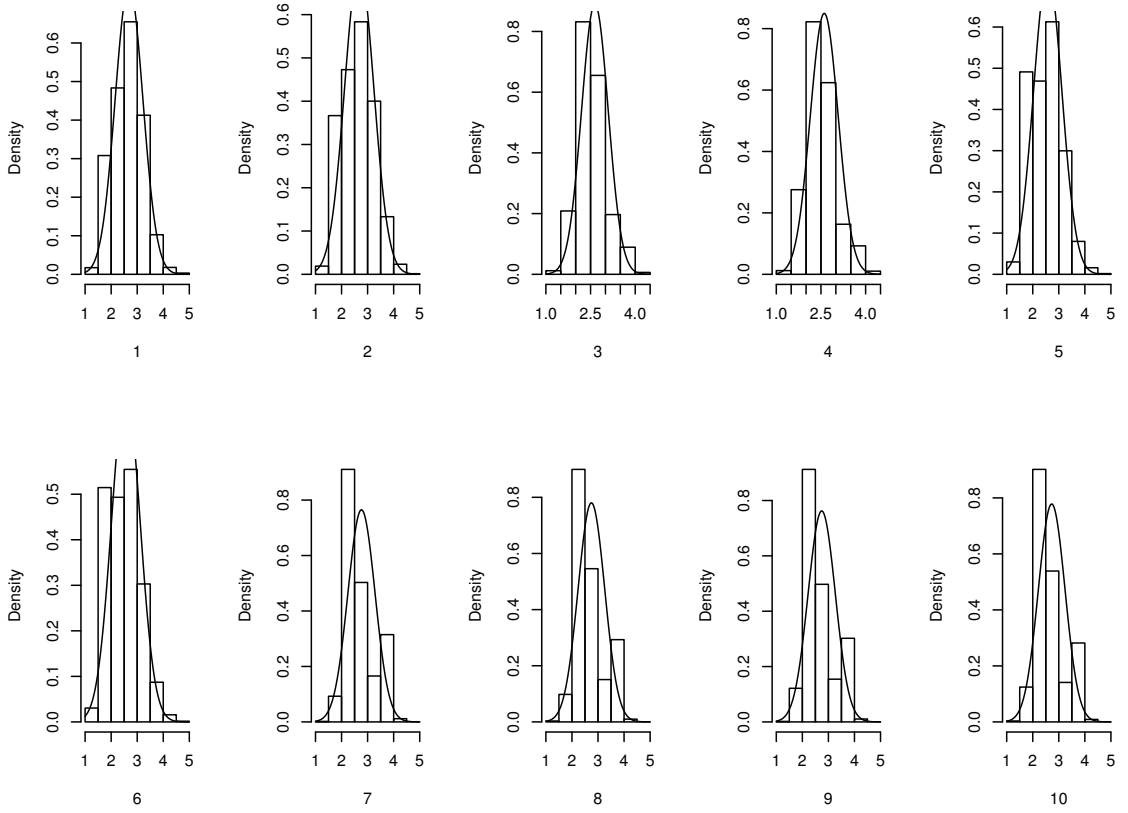

Figure 4.4. Histogram of $\log_{10}$(volume) (without tiny orders)

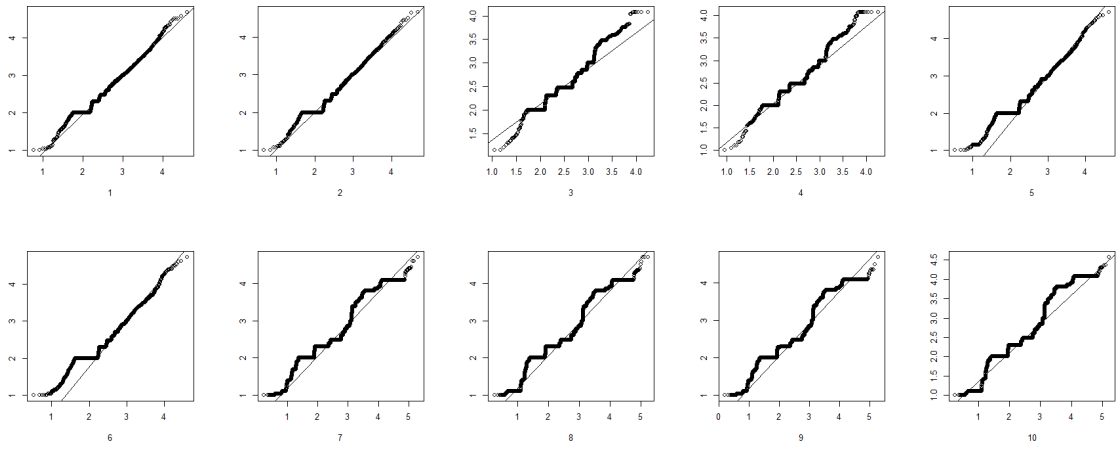

Figure 4.5. QQ plots of $\log_{10}$(volume) vs normal distribution (without tiny orders)

4.7.3 Timing Distribution

In this section, we will look at the distribution of order arrival times. However, before we look at the real data, we would like to try out the classical QQ plots on simulated scenarios. In Figure 4.6 and 4.7, we show the inter-arrivals time of a simulated Hawkes process ($\mu = 0.01, \alpha_{ij} = 100, \beta_j = 1000, N = 60,000$) as well as the fitted residuals. Visually there is not much difference, but from the p-value of the Kolmogorov–Smirnov goodness-of-fit test in Table 4.8, you can see the inter-arrival times fit the exponential distribution very badly while the fitted residuals match the exponential distribution almost perfectly. The key takeaway is that although QQ plot provides a very intuitive visual interface to see the fit of data, some subtle difference may not be able to see in the graph, especially when the amount of data is large.

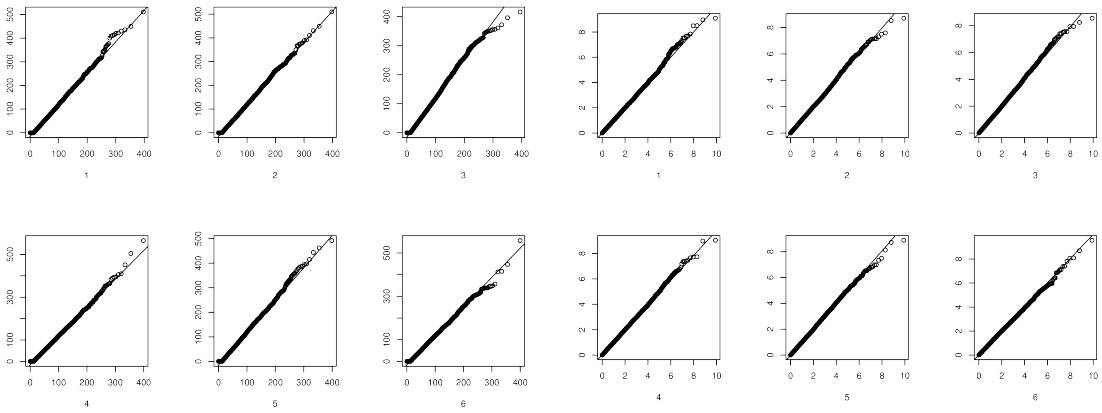

Figure 4.6. QQ plots of inter-arrivals from simulated Hawkes process (N=60,000)

Figure 4.7. QQ plots of fitted residuals from simulated Hawkes process (N=60,000)

Figure 4.8. p-value of Kolmogorov–Smirnov test on simulated Hawkes process (N=60,000)

type	1	2	3	4	5	6
inter-arrivals	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
fitted residuals	0.9891	0.8653	0.8105	0.6587	0.8934	0.6146

Table 4.3.
Fitted alpha (excitation coefficient) for type 1-10 (Jun 2014 (12-2pm))

	α_{ij}									
i\j	1	2	3	4	5	6	7	8	9	10
1	0	0	0	6	215	0	4	0	0	0
2	0	0	0	0	0	200	0	3	0	0
3	0	0	0	0	0	328	0	6	3	0
4	0	0	0	11	445	0	4	0	0	2
5	0	2	2	0	140	0	2	0	0	1
6	6	116	0	0	0	208	0	4	1	0
7	101,309	124	0	75,444	27,041	0	6,460	1	93	977
8	212	102,295	71,799	0	0	25,808	1	6,376	1,037	65
9	35,124	7,965	9,334	54,329	240	6,491	2,521	239	3,076	4
10	6,780	36,407	57,023	7,699	6,172	251	277	2,458	4	3,109

$$\lambda_i(t) = \mu_i + \sum_{j\geq 1} \int_{(-\infty,t)} \alpha_{ij} \exp(-\beta_{ij}(t-s)) dN_j(s) \tag{4.28}$$

As mentioned in Section 5.2.2, our model (4.23-4.24) does not directly depend on type 7-10, but we would like to see if the excitation effects from type 7-10 are also negligible. Therefore, we fit a 10-variate Hawkes process with a non-Markovian exponential kernel (4.28) using approximate expectation maximization (EM) [66], with one month's data (21 trading days) in June 2014 (12-2pm). The non-Markovian kernel allows a more flexible model so that we can get an unbiased assessment about the sizes of excitation coefficients. From Table 4.3, we can rest assured that type 7-10 won't have any significant influence to our model[9].

From the trading perspective, this result won't be any surprise as it is well-known that HFT firms keep revising their quotes in the limit order book [138] and some may even engage in *spoofing* [130] (see also the case of United States v. Michael Coscia 2014). These kinds of quote stuffing [129] have no information content and should not affect the decision of a bona fide market maker, who provides liquidity to the market.

[9] α_{ij} is the excitation effect from type j to type i. From Table 4.3, α_{ij} is small for $i \in \{1,..,6\},\ j \in \{7,8,9,10\}$

$$\lambda_i(t) = \mu_i + \sum_{j\geq 1} \int_{(-\infty,t)} \alpha_{ij} \exp(-\beta_i(t-s)) dN_j(s) \tag{4.29}$$

Under the assumption of the one-tick spread and irrelevance of type 7-10, we can reduce the complexity of the model to 6 dimensions. In the sequel, we will also assume $\beta_{ij} = \beta_i$ (4.29); that means the impact decay to type i orders are the same regardless of the triggering type j. This assumption may not seems plausible in all cases; however, such an assumption will simplify the control problem, as the system is now Markovian.

Table 4.4.
Fitted parameters (Markovian kernel) for type 1-6 (Jun 2014 (12-2pm))

	μ_i	α_{ij}						β_i
i\j		1	2	3	4	5	6	
1	0.0152	5	0	0	3	136	0	877
2	0.0139	0	3	2	0	0	127	929
3	0.0151	0	8	10	2	0	240	2,226
4	0.0121	14	0	0	8	324	0	3,586
5	0.0269	7	12	5	2	48	0	182
6	0.0254	14	4	1	6	0	48	191

The fitted parameters of the 6-dimensional Markovian model are shown in Table 4.4 and the corresponding QQ plots are shown in Figure 4.9-4.10. The base rates μ in Table 4.4 are roughly 60% of the average arrival rates in Table 4.2; in other words, around 40% of the arrivals are caused by excitations. From the magnitude of the coefficient α_{ij}, it is clear that the major triggering events are the non-aggressive market orders. In particular, non-aggressive market buy will have large excitation on itself as well as aggressive market buy and aggressive sell cancellation and the case for non-aggressive market sell is similar. It is a bit surprise that aggressive market orders have little excitation effect, which may be caused by misclassification due to incomplete market data. Also, the excitation pattern seems to be stock specific.

Though the QQ plots of the fitted residuals in Figure 4.10 show some improvement with respect to the raw inter-arrival times in Figure 4.9, the result is not completely satisfactory, in particular when compared with the QQ plots in Large [22], where the author

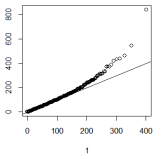
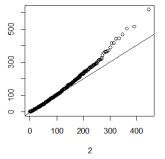
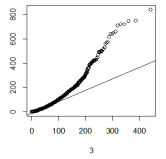
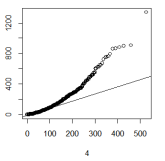
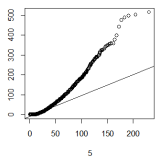
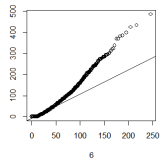

Figure 4.9. QQ plots of inter-arrival times (Jun 2014 (12-2pm))

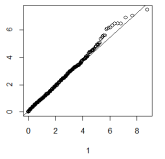
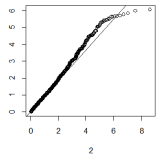
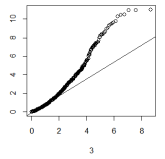
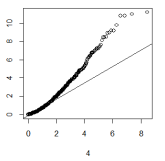
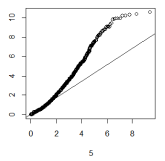
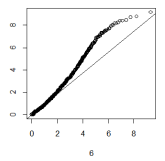

Figure 4.10. QQ plots of fitted residuals (Jun 2014 (12-2pm))

fits a Hawkes model very similar to ours. A couple of reasons may contribute to the difference. First, the Hawkes model in [22] is more sophisticated as he does not assume $\beta_{ij} = \beta_i$ and the self-excitation kernel $\gamma_{ii}(t)$ has two exponential terms. Second, the classification in Large's dataset should be perfect as London Stock Exchange is the only exchange in United Kingdom as of 2002 while Nasdaq matches only 21.6% of the trades for QQQ according to the Nasdaq monthly statistics. In addition, the Large's dataset is time-stamped in seconds while ours is in nano-seconds.

In the high-frequency trading environment with latency in tenths of micro-second, the second-timestamp completely masks the intricate nano-second dynamics, which seems not well described by a simple autoregressive model. In order to better understand the microscopic activities of the order arrivals, we zoom into the data for an one second interval. In Figure 4.11, each bar represents the number of orders in one millisecond interval. From the graph, we can see the orders arrive in clusters. Removing those noisy limit order revisions, the scenario can be more extreme. In Figure 4.12, we see a short burst of 3 orders within 10 millisecond followed by a long period of silence. In view of these two figures, it is not hard to understand why the decay coefficients β_i in Table 4.4 are in the order of thousands, which translates into a half life of less than one millisecond. This is in sharp contrast with

[22] where the half lives of β's, estimated from a second-stamped database, are in the order of seconds.

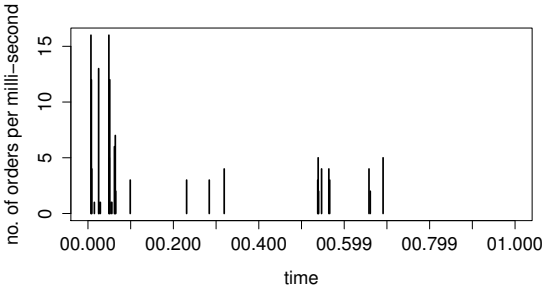

Figure 4.11. One second activities of QQQ (all order types) on June 2, 2014, 12:00:00-12:00:01pm

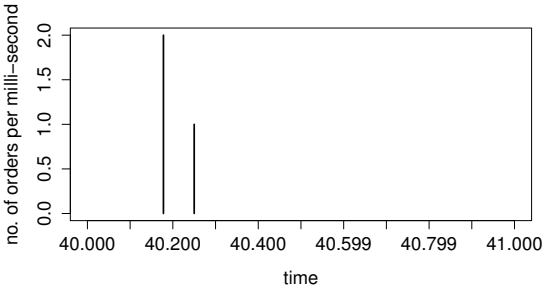

Figure 4.12. One second activities of QQQ (type 1-6) on June 2, 2014, 12:00:40-12:00:41pm

5. THE MARKET-MAKING MODEL

This chapter is the core of the thesis as we are going to describe in details our new market-making model. We will make use of the frameworks of optimal switching and impulse control to formulate the market maker's decision problem and solve it using constrained forward backward stochastic differential equation. However, in order to streamline the presentation of our new model, we have put the technical details in the appendix.

5.1 Trading Environment

The first assumption is that the market maker has only a small market share ρ among all the transactions, so his limit and market orders have negligible influence on the market. For example, when ρ =5%, that means the effective arrival rate of market orders hitting his limit orders is around 5% of all market orders. He may achieve this by continuously adjusting his limit order quantity to be 5% of the total quantity in the queue[1] but the detailed mechanism is outside the scope of this study. The market share ρ will be multiplied to the cash and quantity dynamics of the market maker while the price will evolve according to all orders in the market.

The chance of execution outside the best quotes is deemed to be zero [21]; therefore the market maker will only post limit orders at the best bid and best ask or withdraw from one or both sides of the market. The regime where the market maker is operating is indicated by $I_t \in \mathbb{I}$, $\mathbb{I} = \{(0,0),(0,1),(1,0),(1,1)\}$. For example, under regime (1,0), he will only post limit buy orders at the best bid. In addition, the market maker can also issue market order (impulse) of volume ζ to adjust his inventory, subject to the cost of crossing the bid-ask spread Δ_{t^-} and exchange fee η.

[1]This simple strategy is vulnerable to quote stuffing and spoofing [129, 130].

We will not consider aggressive limit orders (limit orders inside spread) as in [12]. The effect of switching is to change the effective arrival rate of market orders hitting our market maker. However, once you post the first limit order inside the spread, your newly placed order will be on the new best bid/ask queue and is no longer aggressive. In our model, the market orders hitting the best bid/ask always follow the prescribed Hawkes point process and unlike [12] our model does not have any limit order which has effective fill rate higher than that of best quotes. In our framework, placing order inside the spread is like an impulse control. The gain is the priority in the queue over existing orders. However, since we have not yet incorporated a model for the value of queue position. We will leave this kind of strategy in future works. On the other hand, we would like to point out that in [12] an aggressive limit order is simply treated as market order when the bid-ask spread is just one tick, without realizing that market order pays a fee while limit order receives a rebate in a regular (non-inverted) exchange.

Since almost all exchanges implement the price-time[2] priority, withdrawal from the market involves loss of priority of the current limit orders. Hence we penalize the switching of regime by imposing a constant cost $c > 0$. The constant penalization cost is purely a simplifying assumption as the true cost of switching depends on the status of the order book such as how the market maker's limit orders are distributed within the book, the arrival rate of market orders, the bid-ask spread etc at the time of switching.

5.2 Optimal Control Problem

5.2.1 General Model

The evolutions of the cash holding B_t and inventory Q_t of the market maker depend on the regime I_{t^-}. For instance, when the market maker does not post any limit order, the change in B_t, Q_t will be 0. When he has limit orders in the bid queue, the increase in quantity will be the number of shares v of the market order hitting the bid multiplied by the market share ρ (assume market maker's limit orders are distributed evenly within the

[2]Limit orders having better price and then earlier time-stamp will have higher execution priority.

queue). The cash paid out (decrease) will be the share quantity v multiplied by bid price $S^b_{t^-}$, adjusted for rebate ε. The logic on the ask size is similar.

In short, the accounting equation for the cash and inventory B_t, Q_t can be written as

$$\begin{aligned} B_t = b &+ \int_{(s,t]\times\mathbb{R}_+} \mathbb{1}_{\{(0,1),(1,1)\}}(I_{r^-})\Big((S^a_{r^-}+\varepsilon)\rho v(N_1+N_5)(dr\times dv)\Big) \\ &+ \int_{(s,t]\times\mathbb{R}_+} \mathbb{1}_{\{(1,0),(1,1)\}}(I_{r^-})\Big((-S^b_{r^-}+\varepsilon)\rho v(N_2+N_6)(dr\times dv)\Big) \\ &+ \sum_{\tau_n\in(s,t]}\Big((-S^a_{\tau_n^-}-\eta)\zeta_n^+ + (S^b_{\tau_n^-}-\eta)\zeta_n^-\Big) \end{aligned} \tag{5.1}$$

$$\begin{aligned} Q_t = q &+ \int_{(s,t]\times\mathbb{R}_+} \mathbb{1}_{\{(0,1),(1,1)\}}(I_{r^-})\Big(-\rho v(N_1+N_5)(dr\times dv)\Big) \\ &+ \int_{(s,t]\times\mathbb{R}_+} \mathbb{1}_{\{(1,0),(1,1)\}}(I_{r^-})\Big(\rho v(N_2+N_6)(dr\times dv)\Big) + \sum_{\tau_n\in(s,t]}\zeta_n \end{aligned} \tag{5.2}$$

The control u is a sequence of ordered triples $\{(\tau_n, i_n, \zeta_n)\}_{n\geq 1}$, where τ_n is the stopping time of the switching and/or impulse (market order). $i_n \in \mathscr{F}_{\tau_n^-}$, $i_n \in \mathbb{I}$ is the new regime and $\zeta_n \in \mathscr{F}_{\tau_n^-}$, $\zeta_n \in \mathbb{J} \subset \mathbb{R}$ is the signed impulse strength (positive(negative) for buy(sell) and $\mathbb{J}$ is compact). If $i_n = i_{n-1}$, it indicates no change of regime. If $\zeta_n = 0$, it means there is only switching but no market order. When $i_n \neq i_{n-1}$ and $\zeta \neq 0$, the market maker switches the regime and issues market order at the same time. Since we assume the cost of market order does not depends on the state of the regime I_{t^-}, the order of the execution does not matter. $\tau_0 = 0$ and i_0 is the initial regime.

The market-making optimal control problem is to maximize the expected utility of total wealth (cash + inventory - liquidation cost) at the end of period T minus of the expected total cost of switching (the impulse cost is already reflected in B_t) by choosing an optimal control $u = \{(\tau_n, i_n, \zeta_n)\}_{n\geq 1}$ subject to the dynamics of price and order arrivals. Moreover since the Hawkes intensity may not be Markov, the value function at time s may depends on the whole path of the arrivals $N^{0:s}$ up to time s.

The value function $V(s, b, q, s^b, s^a, i, j, N^{0:s})$ is the optimal expected utility when we start the system at time s with $B_s = b, Q_s = q, S^b_s = s^b, S^a_s = s^a, I_{s^-} = (i, j), N(t) = N^{0:s}(t)\ \forall t \in [0, s]$[3]. The complete specification of the control problem is stated below.

[3]The intensity $\lambda_i(t)$ is assumed to start with μ_i at time 0.

Definition 5.2.1 (General Market-Making Model)

$$V(s,b,q,s^b,s^a,i,j,N^{0:s}) = \max_{u\in\mathbb{U}_{ad}} J(s,b,q,s^b,s^a,i,j,N^{0:s},u) \tag{5.3}$$

$$\begin{aligned} &J(s,b,q,s^b,s^a,i,j,N^{0:s},u) \\ &= \mathbb{E}\Big\{\mathscr{U}\Big(B_T + (S_T^b-\eta)Q_T^+ - (S_T^a+\eta)Q_T^-\Big) - \sum_{\tau_n\in(s,T]} c\mathbb{1}(i_n\neq i_{n-1})\Big|\mathscr{F}_s\Big\} \end{aligned} \tag{5.4}$$

$$\begin{aligned} B_t = b &+ \int_{(s,t]\times\mathbb{R}_+} \mathbb{1}_{\{(0,1),(1,1)\}}(I_{r^-})\Big((S_{r^-}^a+\varepsilon)\rho\nu(N_1+N_5)(dr\times dv)\Big) \\ &+ \int_{(s,t]\times\mathbb{R}_+} \mathbb{1}_{\{(1,0),(1,1)\}}(I_{r^-})\Big((-S_{r^-}^b+\varepsilon)\rho\nu(N_2+N_6)(dr\times dv)\Big) \\ &+ \sum_{\tau_n\in(s,t]}\Big((-S_{\tau_n^-}^a-\eta)\zeta_n^+ + (S_{\tau_n^-}^b-\eta)\zeta_n^-\Big) \end{aligned} \tag{5.5}$$

$$\begin{aligned} Q_t = q &+ \int_{(s,t]\times\mathbb{R}_+} \mathbb{1}_{\{(0,1),(1,1)\}}(I_{r^-})\Big(-\rho\nu(N_1+N_5)(dr\times dv)\Big) \\ &+ \int_{(s,t]\times\mathbb{R}_+} \mathbb{1}_{\{(1,0),(1,1)\}}(I_{r^-})\Big(\rho\nu(N_2+N_6)(dr\times dv)\Big) + \sum_{\tau_n\in(s,t]}\zeta_n \end{aligned} \tag{5.6}$$

$$S_t^a = s^a + \int_{(s,t]\times\mathbb{R}_+\times\mathbb{N}} \delta\xi(N_1+N_4-N_{12})(dr\times dv\times d\xi) \tag{5.7}$$

$$S_t^b = s^b + \int_{(s,t]\times\mathbb{R}_+\times\mathbb{N}} \delta\xi(N_{11}-N_2-N_3)(dr\times dv\times d\xi) \tag{5.8}$$

$$I_t = (i,j)\mathbb{1}_{[s,\tau_1)}(t) + \sum_{n\geq 1} i_n\mathbb{1}_{[\tau_n,\tau_{n+1})}(t) \tag{5.9}$$

$$\lambda_i(t) = \mu_i(t) + \sum_{j=1}^{10}\left(\int_{[0,s)}\gamma_{ij}(t-r)dN_j^{0:s}(r) + \int_{[s,t)}\gamma_{ij}(t-r)dN_j(r)\right) \quad i=1,\ldots,10 \tag{5.10}$$

$$\lambda_{11}(t) = \mu_{11}\mathbb{1}\Big((S_{t^-}^a - S_{t^-}^b) > \delta\Big) \tag{5.11}$$

$$\lambda_{12}(t) = \mu_{12}\mathbb{1}\Big((S_{t^-}^a - S_{t^-}^b) > \delta\Big) \tag{5.12}$$

5.2.2 Simplified Model

In the general model, the point processes are non-Markovian, so the resulting optimal control problem is quite hard to solve. However, if the following approximations hold, the model will become much more tractable.

Assumption 5.2.1 *Simplifying Assumptions*

1. *the bid-ask spread is always one tick*
2. $N_{11}(t) \simeq N_{12}(t)$
3. *the price jump is always one tick*
4. *the cross-excitation on* $N_1 - N_6$ *from* $N_7 - N_{10}$ *is negligible*
5. *the stochastic intensities* $(\lambda_1(t), .., \lambda_6(t))$ *are Markov processes*
6. *the distribution of terminal state* B_T, Q_T, S_T *does not depends on the initial history of order arrivals*
7. *the market maker is risk-neutral or has exponential utility with small risk-aversion*

When μ_{11} and μ_{12} are large, the gap in the bid-ask spread will be filled quickly. As a result, the spread will be one tick most of the time. In this case, we need only to deal with the mid-price which has dynamics

$$S(t) = \bar{s} + \int_{(s,t]\times\mathbb{R}_+\times\mathbb{N}} (\delta/2)\xi(N_1 - N_2 - N_3 + N_4 + N_{11} - N_{12})(dt \times dv \times d\xi) \quad (5.13)$$

and the terminal wealth can be expressed directly in mid-price S_t and tick size δ as

$$B_T + S_T Q_T - \delta|Q_T|/2 \quad (5.14)$$

If we assume there is no upward or downward bias, N_{11} and N_{12} should be roughly equal and the mid-price can be simplified to

$$S(t) = \bar{s} + \int_{(s,t]\times\mathbb{R}_+\times\mathbb{N}} (\delta/2)\xi(N_1 - N_2 - N_3 + N_4)(dt \times dv \times d\xi) \quad (5.15)$$

Furthermore, if the price always jumps one tick on aggressive order, we can then eliminate the mark of jump size.

$$S(t) = \bar{s} + (\delta/2)(N_1 - N_2 - N_3 + N_4)((s,t]) \quad (5.16)$$

The state variables in our model in fact do not depend on orders of type 7-10, but they may have excitation effect on type 1-6. However, in the empirical analysis, we find that the excitation effects of those limit order revisions are negligible. As a result we can remove them from the model and the intensities dynamics become

$$\lambda_i(t) = \mu_i(t) + \sum_{j=1}^{6} \left(\int_{[0,s)} \gamma_{ij}(t-r) dN_j^{0:s}(r) + \int_{[s,t)} \gamma_{ij}(t-r) dN_j(r) \right) \quad i = 1, ..., 6 \quad (5.17)$$

When we solve the control problem using forward scheme (see Appendix A.5.2), we need to compute the following conditional expectations

$$U'^{q+1}(t_n,i,\zeta)=\frac{\mathbb{E}\Big(\Big(g(T,X_T,I_T)+\sum_{k=n}^{N-1}f_p(t_k,U'^{q}(t_k,\bullet))\Delta t\Big)\widetilde{N}'((t_n,t_{n+1}],i,\zeta)\Big|\mathscr{F}_n\Big)}{\lambda'\mu(i,\zeta)\Delta t} \tag{5.18}$$

$$Y_n^{q+1}=\mathbb{E}\Big(g(T,X_T,I_T)+\sum_{k=n}^{N-1}f_p(t_k,U'^{q}(t_k,\bullet))\Delta t\Big|\mathscr{F}_n\Big) \tag{5.19}$$

In general the conditional expectation depends on the arrival history $N^{0:t_n}$. Nonetheless, if λ is a Markov process, the computation of the conditional expectation will be tremendously simplified as now it only depends on B_n,Q_n,S_n,I_n and $\lambda(t_n)$. Moreover, for non-Markovian intensity, the computation involves a double summation that amounts to $O(N^2)$ complexity. If we use a Markovian intensity like the exponential kernel (4.14-4.15), there are some recursive formulae that reduce the computation effort to $O(N)$.

Hawkes process is a stationary model in the sense that under some regularity conditions, the distribution of the point process will converge to the stationary distribution as $t\longrightarrow\infty$ (see Section 3.3.2). Therefore the initial history of order arrivals only have limited impact on the distribution of the terminal states B_T,Q_T,S_T, provided that the terminal time T is far enough. However, this is not the same as saying the order arrivals become independent Poisson. The point processes $N_1,..,N_d$ are still interdependent throughout the period, but just the joint distribution converges to the equilibrium distribution.

Under this assumption, the value function will have the form $V(s,b,q,\bar{s},i,j)$. Also when we compute the regression for Y_t and $U'(t,i,\zeta)$, the basis functions only need to involve B_t,Q_t,S_t,I_t. The accuracy of this approximation depends on the decay speed of excitation. If we have an exponential kernel with large β, the approximation should be reasonably good. However, if the kernel has long memory (e.g. power kernel), then the result may be compromised.

Intuitively an initial cash at time s should not affect the trading strategy as cash is risk-free and there is no discounting in the model. Nevertheless, the utility function heavily penalizes loss while only modestly rewards gain, so the initial cash does help to buffer the loss and allows the market maker to engage in more risky strategy. Yet when the market

maker is risk-neutral ($\mathscr{U}(w) = w$), the initial cash b will not affect the choice of optimal control as

$$V(s,b,\bullet) = b + V(s,0,\bullet) \tag{5.20}$$

On the other hand, if the utility function is of the form $\mathscr{U}(w) = -\exp(-\theta w)$ (θ is absolute risk aversion) with θ small, we have

$$-\exp\bigg(-\theta\Big(B_T + S_T Q_T - \delta|Q_T|/2\Big)\bigg) - \sum_{\tau_n\in(s,T]} c\mathbb{1}(i_n \neq i_{n-1}) \tag{5.21}$$

$$= -\mathrm{e}^{-\theta b}\exp\bigg(-\theta\Big(B_T - b + S_T Q_T - \delta|Q_T|/2\Big)\bigg) - \sum_{\tau_n\in(s,T]} c\mathbb{1}(i_n \neq i_{n-1}) \tag{5.22}$$

$$\simeq \mathrm{e}^{-\theta b}\bigg\{-\exp\bigg(-\theta\Big(B_T - b + S_T Q_T - \delta|Q_T|/2\Big)\bigg) - \sum_{\tau_n\in(s,T]} c\mathbb{1}(i_n \neq i_{n-1})\bigg\} \tag{5.23}$$

Therefore, we have $V(s,b,\bullet) \simeq \mathrm{e}^{-\theta b}V(s,0,\bullet)$ when $\theta b \ll 1$.

Definition 5.2.2 (Simplified Market-Making Model)

$$V(s,b,q,\bar{s},i,j) = \max_{u\in\mathbb{U}_{ad}} J(s,b,q,\bar{s},i,j,u) \tag{5.24}$$

$$J(s,b,q,\bar{s},i,j,u)$$

$$= \mathbb{E}\bigg\{\mathscr{U}\Big(B_T + S_T Q_T - (\delta/2+\eta)|Q_T|\Big) - \sum_{\tau_n\in(s,T]} c\mathbb{1}(i_n \neq i_{n-1})\bigg|\mathscr{F}_s\bigg\} \tag{5.25}$$

$$\begin{aligned} B_t = b &+ \int_{(s,t]\times\mathbb{R}_+} \mathbb{1}_{\{(0,1),(1,1)\}}(I_{r^-})\Big((S_{r^-} + \delta/2 + \varepsilon)\rho v(N_1+N_5)(dr\times dv)\Big) \\ &+ \int_{(s,t]\times\mathbb{R}_+} \mathbb{1}_{\{(1,0),(1,1)\}}(I_{r^-})\Big((-S_{r^-} + \delta/2 + \varepsilon)\rho v(N_2+N_6)(dr\times dv)\Big) \\ &+ \sum_{\tau_n\in(s,t]} \Big((-S_{\tau_n^-} - \delta/2 - \eta)\zeta_n^+ + (S_{\tau_n^-} - \delta/2 - \eta)\zeta_n^-\Big) \end{aligned} \tag{5.26}$$

$$\begin{aligned} Q_t = q &+ \int_{(s,t]\times\mathbb{R}_+} \mathbb{1}_{\{(0,1),(1,1)\}}(I_{r^-})\Big(-\rho v(N_1+N_5)(dr\times dv)\Big) \\ &+ \int_{(s,t]\times\mathbb{R}_+} \mathbb{1}_{\{(1,0),(1,1)\}}(I_{r^-})\Big(\rho v(N_2+N_6)(dr\times dv)\Big) + \sum_{\tau_n\in(s,t]} \zeta_n \end{aligned} \tag{5.27}$$

$$S(t) = \bar{s} + (\delta/2)(N_1 - N_2 - N_3 + N_4)((s,t]) \tag{5.28}$$

$$I_t = (i,j)\mathbb{1}_{[s,\tau_1)}(t) + \sum_{n\geq 1} i_n \mathbb{1}_{[\tau_n,\tau_{n+1})}(t) \tag{5.29}$$

$$\lambda_i(t) = \lambda_i(s) + \int_s^t \beta_i(\mu_i - \lambda_i(t))dt + \sum_{j=1}^{6} \int_{[s,t)} \alpha_{ij} dN_j(t) \quad i = 1,...,6 \tag{5.30}$$

5.3 Solving the Optimal Control Problem

The stochastic optimal control problem can be solved via constrained forward backward stochastic differential equation (CFBSDE) as in Section A.4 and we state the representation for the simplified model 5.2.2 here.

Theorem 5.3.1 *The value function V of the simplified market-making model 5.2.2 is given by $V(s,b,q,\bar{s},i,j) = Y_s$ where $(Y,U,U',K) \in \mathbb{S}^2 \times \mathbb{L}^2_N \times \mathbb{L}^2_{N'} \times \mathbb{A}^2$ is the unique minimal solution of the following CFBSDE.*

$$\begin{aligned} B_t = b &+ \int_{(s,t]\times\mathbb{R}_+} \mathbb{1}_{\{(0,1),(1,1)\}}(I_{r^-})\Big((S_{r^-} + \delta/2 + \varepsilon)\rho v(N_1+N_5)(dr\times dv)\Big) \\ &+ \int_{(s,t]\times\mathbb{R}_+} \mathbb{1}_{\{(1,0),(1,1)\}}(I_{r^-})\Big((-S_{r^-} + \delta/2 + \varepsilon)\rho v(N_2+N_6)(dr\times dv)\Big) \\ &+ \int_{(s,t]\times\mathbb{I}\times\mathbb{J}} \Big((-S_{r^-} - \delta/2 - \eta)\zeta^+ + (S_{r^-} - \delta/2 - \eta)\zeta^-\Big)N'(dr\times di\times d\zeta) \end{aligned} \tag{5.31}$$

$$\begin{aligned} Q_t = q &+ \int_{(s,t]\times\mathbb{R}_+} \mathbb{1}_{\{(0,1),(1,1)\}}(I_{r^-})\Big(-\rho v(N_1+N_5)(dr\times dv)\Big) \\ &+ \int_{(s,t]\times\mathbb{R}_+} \mathbb{1}_{\{(1,0),(1,1)\}}(I_{r^-})\Big(\rho v(N_2+N_6)(dr\times dv)\Big) \\ &+ \int_{(s,t]\times\mathbb{I}\times\mathbb{J}} \zeta N'(dr\times di\times d\zeta) \end{aligned} \tag{5.32}$$

$$S(t) = \bar{s} + (\delta/2)(N_1 - N_2 - N_3 + N_4)((s,t]) \tag{5.33}$$

$$I_t = (i,j) + \int_{(s,t]\times\mathbb{I}\times\mathbb{J}} (i - I_{r^-})N'(dr\times di\times d\zeta) \tag{5.34}$$

$$\lambda_i(t) = \mu_i + \int_s^t \beta_i(\mu_i - \lambda_i(t))dt + \sum_{j=1}^{6} \int_{[s,t)} \alpha_{ij} dN_j(t) \quad i = 1,...,6 \tag{5.35}$$

$$\begin{aligned} Y_t = \mathscr{U}\Big(B_T + S_T Q_T - \delta|Q_T|/2\Big) &- \sum_{j=1}^{d} \int_{(t,T]\times\mathbb{R}_+} U_j(r,v)\widetilde{N}_j(dr\times dv) \\ &- \int_{(t,T]\times\mathbb{I}\times\mathbb{J}} U'(r,i,\zeta)N'(dr\times di\times d\zeta) + K_T - K_t \end{aligned} \tag{5.36}$$

$$U'(t,i,\zeta) \le c \ \forall t \in (s,T] \tag{5.37}$$

Proof By theorem A.4.1 and A.4.2. ■

6. CONCLUSION

6.1 Summary of Contributions

- A new joint model of price and order arrivals is proposed and it has following features:
  - Prices and order arrivals are dependent via two linkages: common components and cross-excitations of the underlying marked point processes.
  - Prices are pure-jump processes living on a pre-defined price grid.
  - Price jumps at the exact instant of aggressive order.
  - The underlying processes driving the price movements are directly observable and thus can be estimated using simple statistical methods.
  - The order arrivals exhibit self- and cross-excitation behavior. Hawkes process is used as an example but any marked point process with stochastic intensity is compatible with our framework.
- The framework of solving optimal switching [28] and impulse control [27] by CFBSDE is extended in two ways
  - The state variables can include marked point processes with stochastic intensities.
  - The switching and impulse can occur at the same time.
- A new market-making model is put forward, which incorporate the enhancements below:
  - Prices and order arrivals are modeled using our new joint model.
  - The quotes of market maker are switched in a discrete manner rather than continuously as in [9].

- The hard-to-estimate demand/supply rate function is no longer needed as the market maker will either peg to the best quotes or withdraw from the market similar to [12].
- Order volume is modeled by the random mark of the arrival point process.
- Market order is allowed as in [12] and it can be executed at the same time as the quote switching.
- The stochastic control problem can be solved via Monte Carlo regression of CFBSDE, which is more efficient than finite difference for state variables of high dimension.

6.2 Future Works

Our new market-making model represents only the first attempt to provide a sensible framework for real-world trading and it is definitely far from complete. In fact, it may signify the start, rather than the end, of new series of research to be conducted under the new structure of dependent price and arrivals in a pure-jump environment.

6.2.1 Point Process Modeling

From the result of the numerical experiment, a simple Markovian exponential kernel seems not good enough to describe the subtle nature of high-frequency data. A thorough examination of different types of kernels, such as power kernel [43], double exponential [22] and Laguerre-type polynomial [44], in addition to a systematic study on the effect of volume on excitation, is needed to improve the explanatory power of the model. Besides, the non-linear Hawkes process (3.8) [48] have been developed for quite some time but its use has been very limited. It will be very interesting to see if such sophistication can help to portray the data more faithfully.

6.2.2 Portfolio Extension and Dimension Reduction

A natural progression for the joint price and order model is to extend it to multiple assets. In a straight-forward manner, we can build a Hawkes process with $n \times d$ variates. However, if we take the S&P 500 index with $d = 6$ and use the Markovian exponential kernel, just the α_{ij} matrix will have 9 million entries! Undoubtedly some kinds of dimension reduction techniques are needed. One way is to model dimensions as nodes, cross-excitations as links between nodes and use graphical model to express the sparse dependency structure [139].

The multi-asset joint model can help to develop a market-making strategy for portfolio which explores the correlation structure across multiple assets. For example, if the inventory of a market maker is $+\$1,000,000$ of A and $-\$1,000,000$ of B, his risk may not so high if the prices of A and B are highly correlated. However, a single asset model may opt for a suboptimal decision to reduce the seemingly excessive risk by liquidating both inventory on A and B.

6.2.3 Queue Modeling

In our model, the penalization cost is simply a constant c and it is of course far from satisfactory. The true cost of switching depends on how the limit orders are distributed within the order book, which in turn depends on the strategy that the market maker uses to secure the target order flow.

If the market maker has the subscription to the data feed from the exchange, he may be able to compute the queue positions of his limit orders in real-time [140]. Armed with such information, if we can develop a framework which estimates the intrinsic value of an array of queue positions, the market-making model can adjust the switching decision based on this dynamic switching cost.

6.2.4 Numerical Methods

The numerical methods to solve CFBSDE are still in their infancy. As see in the numerical examples, the numerical schemes are slow to converge and memory-intensive. Besides, they become numerically unstable when penalization is used and the accuracy is further deteriorates for point process with volatile marks.

Moreover the majority of the research effort concentrates on equation driven by Brownian motion and we have not seen any numerical method specifically designed for pure-jump process with stochastic intensity.

In addition, we would like to mention there are some numerical methods involving Malliavin calculus [141] but they cannot be used in our framework as there is not yet a version of Malliavin calculus for point process with stochastic intensity.

6.2.5 Adverse Selection

Since the market maker gives free options to all market participants in order to earn the spread, he will always lose to informed traders. Adverse selection [1, 2] is a well-known issue in market making. One way to minimize the loss is to temporarily withdraw from the market if the market maker detects the heavy trading of insiders.

The Volume-synchronized Probability of Informed Trading (VPIN) [142, 143] is one such measure, but more work is needed to see how we can incorporate this signal into the control problem and backtest its performance on real-world trading.

6.3 Conclusion

In this thesis, we develop a new market-making model that try to incorporate a number of realistic assumptions relevant for high-frequency trading. The job is tough and our result is far from perfect. We have discussed some potential future works and we wish our initial study will attract the attention of other researchers to work on this intriguing problem.

APPENDIX

Appendix: Control Problem and CFBSDE

A.1 Introduction

Assuming the evolution of state X_t depends on the regime I_t, at each user chosen event time τ_n we can either change the regime from i_{n-1} to i_n or fire an impulse ζ_n, which causes X_t to change by $\Gamma(\tau_n, X_{\tau_n^-}, \zeta_n)$. Our objective is to find a sequence of switches and impulses $\{(\tau_n, i_n, \zeta_n)\}$ so as to maximize the sum of terminal value $g(T, X_T, I_T)$ and running gain $\int_s^T f(t, X_t, I_t)dt$ subject to the switching cost $h_1(\bullet)$ and impulse cost $h_2(\bullet)$. Such kind of optimization problem (A.1-A.3) is called optimal switching and impulse control problem.

$$V(s,x,i) = \max_{\{\tau_n, i_n, \zeta_n\}} \mathbb{E}\Bigg(g(T, X_T, I_T) + \int_s^T f(t, X_t, I_t)dt - \sum_{\tau_n \in (s,T]} \Big(h_1(\tau_n, X_{\tau_n{}^-}, i_{n-1}, i_n) + h_2(\tau_n, X_{\tau_n{}^-}, \zeta_n) \Big) \Bigg| \mathscr{F}_s \Bigg) \tag{A.1}$$

$$X_t = x + \int_s^t b(r, X_r, I_r)dr + \int_{(s,t]\times\mathbb{K}} \gamma(r, X_{r^-}, I_{r^-}, k) N(dr \times dk) + \sum_{\tau_n \in (s,t]} \Gamma(\tau_n, X_{\tau_n^-}, \zeta_n) \tag{A.2}$$

$$I_t = i\mathbb{1}_{[s,\tau_1)}(t) + \sum_{n=1}^{\infty} i_n \mathbb{1}_{[\tau_n, \tau_{n+1})}(t) \tag{A.3}$$

If the intensity $\lambda(t)$ of the marked point process $N(dt, dk)$ is deterministic, the value function V is the viscosity solution [24] of the Hamilton-Jacobi-Bellman quasi-variational inequality (HJBQVI) (A.4-A.5) [25, 26].

$$\max\Bigg\{ f(t,x,i) + V_t(t,x,i) + V_x(t,x,i)^\top b(t,x,i) + \int_{\mathbb{K}} \Big(V(t, x+\gamma(t,x,i,k), i) - V(t,x,i) \Big) \lambda(t) \mu(t, dk), \max_j \{ V(t,x,j) - h_1(t,x,i,j) - V(t,x,i) \}, \max_\zeta \{ V(t, x+\Gamma(t,x,\zeta), i) - h_2(t,x,\zeta) \} - V(t,x,i) \Bigg\} = 0 \tag{A.4}$$

$$V(T,x,i) = g(T,x,i) \tag{A.5}$$

However, when the intensity $\lambda(t)$ is stochastic but we still apply the same method naively, the resulting optimality condition will become a partial integro-differential equation (PIDE) with *random* coefficients. Even if we can solve the PIDE for each ω, the solution will not equal the value function of the control problem as the value function is non-random.

The trouble lies on the fact that the intensity now contains information about the current state of the system. Suppose the intensity λ_t is bounded below, there exists a Girsanov change of measure [55] such that under some equivalent probability measure $\mathbb{Q}$, N is a marked Poisson process. The stochastic intensity $\lambda(\bullet)$ will appear in the Girsanov kernel L_t when we compute the value function under $\mathbb{Q}$.

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = L_t \tag{A.6}$$

$$L_t = \exp\Big(\int_{(0,t]} -\log(\lambda(r))N(dr) + \int_{(0,t]\times\mathbb{K}} (\lambda(r)-1)\mu(r,dk)dr\Big) \tag{A.7}$$

$$\begin{aligned} &V(s,x,i) \\ &= \max_{\{\tau_n,i_n,\zeta_n\}} \mathbb{E}^{\mathbb{P}}\Bigg(g(T,X_T,I_T) + \int_s^T f(t,X_t,I_t)dt \\ &\qquad\qquad - \sum_{\tau_n\in(s,T]} \Big(h_1(\tau_n,X_{\tau_n{}^-},i_{n-1},i_n) + h_2(\tau_n,X_{\tau_n{}^-},\zeta_n)\Big)\Bigg|\mathscr{F}_s\Bigg) \end{aligned} \tag{A.8}$$

$$\begin{aligned} &= \max_{\{\tau_n,i_n,\zeta_n\}} L_s\,\mathbb{E}^{\mathbb{Q}}\Bigg(g(T,X_T,I_T)/L_T + \int_s^T f(t,X_t,I_t)/L_T dt \\ &\qquad\qquad - \sum_{\tau_n\in(s,T]} \Big(h_1(\tau_n,X_{\tau_n{}^-},i_{n-1},i_n)/L_T + h_2(\tau_n,X_{\tau_n{}^-},\zeta_n)/L_T\Big)\Bigg|\mathscr{F}_s\Bigg) \end{aligned} \tag{A.9}$$

However, after the change of measure, the control problem is no longer in the standard form that can be solved by HJBQVI as the driver f and cost h_1, h_2 involve a random quantity measurable at the terminal time T.

Regular stochastic optimal control problem can be solved via *stochastic maximum principle* which describes the optimality condition in the form of a backward stochastic differential equation (BSDE) [144]. However, it is not until 2010[1] that Kharroubi et al. [27] establish the connection of constrained forward backward stochastic differential equation

[1]Peng and Xu [145] (2007) also discuss the connection of CBSDE with QVI but the paper remains unpublished as of Apr 2015.

(CFBSDE) to impulse control problem. Later in 2014[2], Elie and Kharroubi [28] apply CFBSDE to solve optimal switching.

While Kharroubi et al. [27], Elie and Kharroubi [28] focus on state variable driven by Brownian motion, we have extended the formulation to include state variable driven by marked point process with stochastic intensity and enrich the framework to handle the combined optimal switching and impulse control problem, where switching and impulse can happen at the same time.

Unlike HJBQVI, the structure of CFBSDE is by design random, hence the stochastic intensity does not add much complexity to the system. Moreover, the CFBSDE can be solved via Monte Carlo simulation, which is much more efficient than finite difference when the dimension of X_t is high.

A nice introduction of impulse control with jumps can be found in [26] and the classic reference for HJBQVI is [146]. For an excellent introduction of FBSDE with jumps and CFBSDE, readers can refer to [147] and [27, 28, 148, 149] respectively.

A.2 Notation

Let $T > 0$ be a fixed terminal time and $(\Omega, \mathscr{F}, \mathbb{P})$ be a complete probability space endowed with a d-variate simple marked point process (MPP) N on $[0,T] \times \mathbb{K}$, as well as another MPP N' on $[0,T] \times \mathbb{I} \times \mathbb{J}$ where $\mathbb{K}$ is the mark space of N and $\mathbb{I}, \mathbb{J}$ are respectively the regime and impulse space. $\mathbb{I}, \mathbb{J}, \mathbb{K}$ are subsets of Euclidean space and N, N' are independent from each others. $\{\mathscr{F}_t\}$ denotes the augmented right-continuous filtration generated by N, N' and $\mathscr{P}$ is the predictable σ-field. Each variate of N is assumed to have an absolutely continuous compensator in the form of $\lambda_j(t)\mu_j(t,dk)dt$ where $\lambda_j(t)$ is the stochastic intensity, $\mu_j(t,dk)$ is the conditional mark distribution and $\widetilde{N}_j(dt,dk) = N_j(dt,dk) - \lambda_j(t)\mu_j(t,dk)dt$ is the compensated marked point process. N' is assumed to a marked Poisson process with intensity λ'.

[2]Though the paper was published in 2014, the preprint appeared in 2009.

The dimensions of various functions to be used in the coming sections are specified here.

$$b : [0,T] \times \mathbb{R}^n \times \mathbb{I} \longrightarrow \mathbb{R}^n, \quad \gamma_j : [0,T] \times \mathbb{R}^n \times \mathbb{I} \times \mathbb{K} \longrightarrow \mathbb{R}^n \tag{A.10}$$

$$f : [0,T] \times \mathbb{R}^n \times \mathbb{I} \longrightarrow \mathbb{R}, \quad g : [0,T] \times \mathbb{R}^n \times \mathbb{I} \longrightarrow \mathbb{R} \tag{A.11}$$

$$h_1 : [0,T] \times \mathbb{R}^n \times \mathbb{I}^2 \longrightarrow \mathbb{R}_+, \quad h_2 : [0,T] \times \mathbb{R}^n \times \mathbb{J} \longrightarrow \mathbb{R}_+ \tag{A.12}$$

$$\Gamma : [0,T] \times \mathbb{R}^n \times \mathbb{J} \longrightarrow \mathbb{R}^n \tag{A.13}$$

We also define a few standard function spaces and their norms for the ease of exposition.

$$\mathbb{S}^2 = \left\{ Y : \Omega \times [0,T] \longrightarrow \mathbb{R} \middle| Y_t \in \mathscr{F}_t, \text{ càdlàg}, \mathbb{E}\left(\sup_{t\in[0,T]} |Y_t|^2 \right) < \infty) \right\} \tag{A.14}$$

$$\|Y\|^2_{\mathbb{S}^2} = \mathbb{E}\left(\sup_{t\in[0,T]} |Y_t|^2 \right) \tag{A.15}$$

$$\mathbb{A}^2 = \left\{ K \in \mathbb{S}^2 \middle| K_t \text{ is non-decreasing a.s., } K_0 = 0 \right\} \tag{A.16}$$

$$\mathbb{L}^2_N = \left\{ U : \Omega \times [0,T] \times \mathbb{K} \longrightarrow \mathbb{R}^d \middle| U(\bullet,k) \in \mathscr{P}, \right.$$
$$\left. \mathbb{E}\left(\sum_{j=1}^{d} \int_0^T \int_{\mathbb{K}} |U_j(t,k)|^2 \lambda_j(t) \mu_j(t,dk) dt \right) < \infty) \right\} \tag{A.17}$$

$$\|U\|^2_{\mathbb{L}^2_N} = \mathbb{E}\left(\sum_{j=1}^{d} \int_0^T \int_{\mathbb{K}} |U_j(t,k)|^2 \lambda_j(t) \mu_j(t,dk) dt \right) \tag{A.18}$$

A.3 Problem Formulation

The combined optimal switching and impulse control problem consists of finding a sequence of stopping time to switch the regime of the state X_t in order to change its dynamics and/or to initiate an impulse which change the value of X_t immediately. We extend the setting in [25] where the state is driven by a Brownian motion and the optimal switching and impulse control is triggered one at a time. When the state is a diffusion, usually one of the switching or impulse triggering boundaries will be hit before the other, so this setting is appropriate.

However, in our case of pure-jump process, a large jump may push the state far out of the optimal region and it may be optimal to switch regime and fire impulse right af-

ter this large jump. Considering only one action at a time may lead to suboptimal result. For example, suppose the impulse is costly but can create immediate impact on the state, the *one-at-a-time* setting may choose to trigger only a large impulse while the *simultaneous* setting may opt for a switch to a defensive regime together with a smaller impulse. Anyways, since the *simultaneous* setting will consider switching only, impulse only and switching + impulse, the result will always outperform the *one-at-a-time* framework.

The switching and impulse are assumed to be activated shortly[3] after the uncontrolled jump N, so the new regime i and impulse strength ζ can depend on the state after the uncontrolled jump. An impulse at time τ with strength ζ will advance the state by $\Gamma(\tau, X_{\tau^-}, \zeta)$ where X_{τ^-} is the state just before time τ. I_t is the active regime at time t after the switching. To simplify the model, we assume that the impulse transformation Γ does not depend on the regime; otherwise we need to take into consideration the order of switching and impulse. The dynamics of the state variable and regime can be written as follows.

$$X_t = y + \int_s^t b(r, X_r, I_r)dr + \sum_{j=1}^{d} \int_{(s,t]\times\mathbb{K}} \gamma_j(r, X_{r^-}, I_{r^-}, k) N_j(dr \times dk) + \sum_{\tau_n \in (s,t]} \Gamma(\tau_n, X_{\tau_n^-}, \zeta_n) \tag{A.19}$$

$$I_t = i\mathbb{1}_{[s,\tau_1)}(t) + \sum_{n\geq 1} i_n \mathbb{1}_{[\tau_n, \tau_{n+1})}(t) \tag{A.20}$$

We impose the following standard Lipschitz and linear growth assumptions to ensure the existence and regularity of the solution of the (forward) SDE.

Assumption A.3.1

$$\|b(t,x,i) - b(t,x',i)\| + \sum_{j=1}^{d} \int_{\mathbb{K}} \|\gamma_j(t,x,i,k) - \gamma_j(t,x',i,k)\| \lambda_j(t)\mu_j(t,dk) + \sup_{\zeta\in\mathbb{J}} \|\Gamma(t,x,\zeta) - \Gamma(t,x',\zeta)\| \leq C\|x - x'\| \quad a.s. \tag{A.21}$$

$$\|b(t,x,i)\| + \sum_{j=1}^{d} \int_{\mathbb{K}} \|\gamma_j(t,x,i,k)\| \lambda_j(t)\mu_j(t,dk) + \sup_{\zeta\in\mathbb{J}} \|\Gamma(t,x,\zeta)\| \leq C(1 + \|x\|) \quad a.s. \tag{A.22}$$

[3]For technical reason, we assume the switching and impulse cannot occur at the same time as the uncontrolled jump N.

A control $u = \{(\tau_n, i_n, \zeta_n)\}_{n\geq 1}$ consists of an increasing sequence of $\mathscr{F}$-stopping time τ_n, new regime $i_n \in \mathbb{I} = \{1, ..., R\}$ and impulse strength $\zeta_n \in \mathbb{J} = \mathbb{C} \cup \{\varnothing\}, \mathbb{C}$ compact. $i_n = i_{n-1}$ means no switching while $\zeta_n = \varnothing$ indicates no impulse.

There is a technical condition that the control time τ_n cannot coincide with the jump time of the point processes $\{N_j\}$ as we will later use change of measure to change the intensity of point process associated with the control. Without this technical condition, the other point processes will also be affect by this change of measure. In practice, there is always a non-zero processing time, so the switching and impulse will never happen right after the order arrivals.

The control problem is to find the optimal decision u such that the value function $V(s,y,i)$, composed of the terminal gain g, running gain f minus the switching cost h_1 and impulse cost h_2, is maximized. We define $h = h_1 + h_2$ in order to simplify the notation. $X_t^{s,y,i}$ stands for value of X at time t starting with $X_s = y$ and $I_s = i$ but we will omit the superscript if the meaning is clear.

$$V(s,y,i) = \sup_{u\in\mathbb{U}_{ad}} \mathbb{E}\Bigg(g(T, X_T^{s,y,i}, I_T^{s,y,i}) + \int_s^T f(t, X_t^{s,y,i}, I_t^{s,y,i})dt \\ - \sum_{\tau_n \in (s,T]} h(\tau_n, X_{\tau_n^-}^{s,y,i}, I_{\tau_n^-}^{s,y,i}, i_n, \zeta_n) \Bigg| \mathscr{F}_s \Bigg) \tag{A.23}$$

$$\mathbb{U}_{ad} = \Bigg\{ (\tau_n, i_n, \zeta_n) \in \mathbb{R}_+ \times \mathbb{I} \times \mathbb{J}\} | \{\tau_n\} \text{ are } \mathscr{F}_t \text{ stopping times}, \\ s < \tau_n < \tau_{n+1} \leq T,\ i_n \in \mathscr{F}_{\tau_n^-},\ \zeta_n \in \mathscr{F}_{\tau_n^-},\ i_n \neq i_{n-1} \text{ or } \zeta_n \neq \varnothing, \\ |\tau_n - \sigma_{n,j}| \geq \varepsilon\ \forall n, j \text{ where } \sigma_{n,j} \text{ is } n^{th} \text{ jump of } N_j \Bigg\} \tag{A.24}$$

$$h(t,x,i,j,\zeta) = h_1(t,x,i,j) + h_2(t,x,\zeta) \tag{A.25}$$

The following set of assumption makes sure the control problem is well-posed. The Lipschitz and linear growth conditions of f, g, h_1, h_2 (A.26-A.27) ensure that the value function is well-defined. The strict sublinearity conditions of h_1, h_2 (A.28-A.29) remove the possibility of consecutive switches or impulses. The lower bounds of h_1, h_2 (A.30-A.31) guarantee that the optimal control is a finite sequence. (A.32) mandates that there is no switching or impulse at the terminal time T so as to simplify the terminal condition.

Finally, (A.33-A.34) require that γ_j are predictable and Γ, h_1, h_2 are left-continuous so that the stochastic integrals with respect to point processes are well-defined.

Assumption A.3.2

$$|f(t,x,i)-f(t,x',i)|+|g(t,x,i)-g(t,x',i)|$$
$$+|h_1(t,x,i,j)-h_1(t,x',i,j)|+|h_2(t,x,\zeta)-h_2(t,x',\zeta)|\leq C\|x-x'\| \tag{A.26}$$

$$f(t,x,i)\leq C(1+\|x\|),\ g(t,x,i)\leq C(1+\|x\|),\ h(t,x,i,j,\zeta)\leq C(1+\|x\|) \tag{A.27}$$

$$h_1(t,x,i,k)+C\leq h_1(t,x,i,j)+h_1(t,x,j,k)\ \exists C>0 \tag{A.28}$$

$$h_2(t,x,\zeta+\zeta')+C\leq h_2(t,x,\zeta)+h_2(t,x,\zeta')\ \exists C>0 \tag{A.29}$$

$$h_1(t,x,i,i)=0,\ h_1(t,x,i,j)\geq C\ \exists C>0\ \forall i\neq j \tag{A.30}$$

$$h_2(t,x,\varnothing)=0,\ h_2(t,x,\zeta)\geq C\ \exists C>0\ \forall \zeta\neq\varnothing \tag{A.31}$$

$$g(T,x,i)\geq \sup_{j\in\mathbb{I},\zeta\in\mathbb{J}}\{g(T,x+\zeta,j)-h_1(T,x,i,j)-h_2(T,x,\zeta)\} \tag{A.32}$$

$$\gamma_j\in\mathscr{P},\ \Gamma(t,x,\zeta)=\Gamma(t^-,x,\zeta) \tag{A.33}$$

$$h_1(t,x,i,j)=h_2(t^-,x,i,j),\ h_2(t,x,\zeta)=h_2(t^-,x,\zeta) \tag{A.34}$$

A.4 Solution via CFBSDE

The following constrained forward backward stochastic differential equation (CFBSDE) is the key to finding the value function $V(s,y,i)$ of the optimal control problem. Since the forward (A.35-A.36) and backward (A.37-A.38) equations are uncoupled (forward equation does not depend on (Y,U,U',K)), the forward equation can be solved separately using classical methods. Hence our focus will be on the constrained backward equation.

$$X_t=X_s+\int_s^t b(r,X_r,I_r)dr+\sum_{j=1}^{d}\int_{(s,t]\times\mathbb{K}}\gamma_j(r,X_{r^-},I_{r^-},k)N_j(dr\times dk)$$
$$+\int_{(s,t]\times\mathbb{I}\times\mathbb{J}}\Gamma(r,X_{r^-},\zeta)N'(dr\times di\times d\zeta) \tag{A.35}$$

$$I_t=I_s+\int_{(s,t]\times\mathbb{I}\times\mathbb{J}}(i-I_{r^-})N'(dr\times di\times d\zeta) \tag{A.36}$$

$$Y_t = g(T,X_T,I_T) + \int_t^T f(r,X_r,I_r)dr - \sum_{j=1}^{d} \int_{(t,T]\times\mathbb{K}} U_j(r,k)\widetilde{N}_j(dr\times dk) - \int_{(t,T]\times\mathbb{I}\times\mathbb{J}} U'(r,i,\zeta)N'(dr\times di\times d\zeta) + K_T - K_t \quad (A.37)$$

$$U'(t,i,\zeta) \le h(t,X_{t^-},I_{t^-},i,\zeta)\ \forall t \in (s,T] \quad (A.38)$$

In the CBSDE, $N'(dt\times di\times d\zeta)$ is the marked point process associated with the control but it can be chosen to have any distribution provided that the intensity of N' is strictly positive and the support of the conditional mark distribution of N' is the whole mark space $\mathbb{I}\times\mathbb{J}$. For the ease of computation, N' is usually taken to be a Poisson process independent from N. Such a feature is called control randomization in [150] and in the proof of theorem A.4.2, you can see that it is a result of change of probability measure.

As its name implied, the solution component U' of the CBSDE must be less than h at all time between s and T. It is this restriction that force the component Y to equal to the value function V of the associated control problem.

The following condition gives a way to check if the CBSDE (A.37-A.38) is well-posed.

Theorem A.4.1 *If $f(t,x,i) \le C(1+\|x\|)$, $g(t,x,i) \le C(1+\|x\|)$, $0 \le h(t,x,i,j,\zeta) \le C(1+\|x\|)$, then the CBSDE (A.37-A.38) admits a solution.*

Proof Let

$$U_j(t,k) = 0,\ U'(t,i,\zeta) = h(t,X_{t^-},I_{t^-},i,\zeta) \quad (A.39)$$

$$K_t = \begin{cases} \int_0^t \Big(C(1+\|X_r\|) - f(r,X_r,I_r)\Big)dr \\ \quad + \int_{(0,t]\times\mathbb{I}\times\mathbb{J}} h(r,X_{r^-},I_{r^-},i,\zeta)N'(dr\times di\times d\zeta) & t\in[0,T) \\ K_{T^-} + C(1+\|X_T\|) - g(T,X_T,I_T) & t = T \end{cases} \quad (A.40)$$

$$Y_t = \begin{cases} C(1+\|X_T\|) + \int_t^T C(1+\|X_r\|)dr & t\in[0,T) \\ g(T,X_T,I_T) & t = T \end{cases} \quad (A.41)$$

By construction (Y,U,U',K) satisfies the equation (A.37-A.38). From the given condition, K_t is non-decreasing. Using $X \in \mathbb{S}^2$, Jensen's and other simple inequalities, we get

$$\mathbb{E}\left(\sup_{t\in[0,T]} |Y_t|^2\right) \quad (A.42)$$

$$\leq \mathbb{E}\bigg(\sup_{t\in[0,T]} \Big(C(1+\|X_T\|) + \int_t^T C(1+\|X_r\|)dr + g(T,X_T,I_T)\Big)^2\bigg) \tag{A.43}$$

$$\leq \mathbb{E}\bigg(\sup_{t\in[0,T]} 3\Big(C^2(1+\|X_T\|)^2 + \int_t^T C^2(1+\|X_r\|)^2 dr + g^2(T,X_T,I_T)\Big)\bigg) \tag{A.44}$$

$$\leq \mathbb{E}\bigg(\sup_{t\in[0,T]} 3\Big(2C^2(1+\|X_T\|)^2 + \int_0^T C^2(1+\|X_r\|)^2 dr\Big)\bigg) \tag{A.45}$$

$$\leq \mathbb{E}\bigg(\sup_{t\in[0,T]} 3\Big(4C^2(1+\|X_T\|^2) + 2TC^2(1+\|X_t\|^2)\Big)\bigg) \tag{A.46}$$

$$\leq \mathbb{E}\bigg(12C^2(1+\sup_{t\in[0,T]}\|X_t\|^2) + 6TC^2(1+\sup_{t\in[0,T]}\|X_t\|^2)\bigg) \leq \infty \tag{A.47}$$

$$\mathbb{E}\bigg(\sup_{t\in[0,T]} |K_t|^2\bigg) = \mathbb{E}(|K_T|^2) \tag{A.48}$$

$$\leq 4\mathbb{E}\bigg(\int_0^T \Big(C(1+\|X_r\|) - f(r,X_r,I_r)\Big)^2 dr + C^2(1+\|X_T\|)^2 + g^2(T,X_T,I_T) + \int_{(0,T]\times\mathbb{I}\times\mathbb{J}} h^2(r,X_{r^-},I_{r^-},i,\zeta)N'(dr\times di\times d\zeta)\bigg) \tag{A.49}$$

$$\leq 4\mathbb{E}\bigg(\int_0^T 4\Big(C^2(1+\|X_r\|)^2\Big)dr + C^2(1+\|X_T\|)^2 + C^2(1+\|X_T\|)^2 + \int_{(0,T]\times\mathbb{I}\times\mathbb{J}} C^2(1+\|X_r\|)^2\lambda'\mu'(r,di\times d\zeta)dr\bigg) \tag{A.50}$$

$$\leq 4\mathbb{E}\bigg(\int_0^T 8\Big(C^2(1+\|X_r\|^2)\Big)dr + 4C^2(1+\|X_T\|^2) + \int_{(0,T]\times\mathbb{I}\times\mathbb{J}} 2C^2(1+\|X_r\|^2)\lambda'\mu'(r,di\times d\zeta)dr\bigg) \tag{A.51}$$

$$\leq 4\mathbb{E}\bigg(8TC^2(1+\sup_{t\in[0,T]}\|X_t\|^2) + 4C^2(1+\sup_{t\in[0,T]}\|X_t\|^2) + 2T\lambda'C^2(1+\sup_{t\in[0,T]}\|X_t\|^2)\bigg) \leq \infty \tag{A.52}$$

$$\mathbb{E}\bigg(\int_0^T \int_{\mathbb{K}} |U'(t,i,\zeta)|^2\lambda'\mu'(t,di\times d\zeta)dt\bigg) \tag{A.53}$$

$$= \mathbb{E}\bigg(\int_0^T \int_{\mathbb{K}} |h'(t,X_{t^-},I_{t^-},i,\zeta)|^2\lambda'\mu'(r,di\times d\zeta)dt\bigg) \tag{A.54}$$

$$\leq \mathbb{E}\bigg(\int_0^T \int_{\mathbb{K}} C^2(1+\|X_t\|)^2\lambda'\mu'(t,di\times d\zeta)dt\bigg) \tag{A.55}$$

$$\leq \mathbb{E}\bigg(2C^2\lambda'T(1+\sup_{t\in[0,T]}\|X_t\|^2)\bigg) \leq \infty \tag{A.56}$$

Thus $(Y,U,U',K) \in \mathbb{S}^2 \times \mathbb{L}^2_N \times \mathbb{L}^2_{N'} \times \mathbb{A}^2$ ■

The following theorem is the main result of this section.

Theorem A.4.2 *If assumptions* (A.3.1)*,* (A.3.2) *hold and there exists a solution to the CB-SDE (A.37-A.38), then there exists an unique minimal solution*[4] $(Y,U,U',K) \in \mathbb{S}^2 \times \mathbb{L}^2_N \times \mathbb{L}^2_{N'} \times \mathbb{A}^2$ *with K predictable.*

Moreover, if $(Y,U,U',K) \in \mathbb{S}^2 \times \mathbb{L}^2_N \times \mathbb{L}^2_{N'} \times \mathbb{A}^2$ *is a minimal solution of (A.37-A.38), then*

$$V(s,y,i) = Y_s^{s,y,i} \ \forall s \in [0,T] \tag{A.57}$$

and the optimal control is

$$\begin{aligned}\tau_n = \inf\big\{t > \tau_{n-1} \big| & V(t,X_{t^-}^{s,y,i},I_{t^-}^{s,y,i}) = V(t,X_{t^-}^{s,y,i} + \Gamma(t,X_{t^-}^{s,y,i},\zeta),j) \\ & - h(t,X_{t^-}^{s,y,i},I_{t^-}^{s,y,i},j,\zeta) \text{ for some } j \neq I_{t^-} \text{ or } \zeta \neq \varnothing\big\}\end{aligned} \tag{A.58}$$

$$(i_n,\zeta_n) = \underset{j\in\mathbb{I},\zeta\in\mathbb{J}}{\operatorname{argmax}} \left\{V(\tau_n,X_{\tau_n^-}^{s,y,i} + \Gamma(\tau_n,X_{\tau_n^-}^{s,y,i},\zeta),j) - h(\tau_n,X_{\tau_n^-}^{s,y,i},I_{\tau_n^-}^{s,y,i},j,\zeta)\right\} \tag{A.59}$$

Proof Let $\mathbb{V}$ be the set of bounded predictable random field $\nu(t,i,\zeta)$ on $[0,T]\times\mathbb{I}\times\mathbb{J}$ and we define L_t and the probability measure $\mathbb{P}^\nu$ as

$$dL_t = L_{t^-}\int_{\mathbb{I}\times\mathbb{J}} (\nu(t,i,\zeta)-1)\widetilde{N}'(dt\times di\times d\zeta), \quad L_0 = 1 \tag{A.60}$$

$$\mathbb{P}^\nu(A) = \int_A L_t d\mathbb{P} \tag{A.61}$$

$\mathbb{E}^\nu$ denotes the expectation under $\mathbb{P}^\nu$. By Girsanov theorem for marked point process [55], the compensator measure of N' changes from $\lambda'\mu'(t,di\times d\zeta)dt$ in $\mathbb{P}$ to

$$\nu(t,i,\zeta)\lambda'\mu'(t,di\times d\zeta)dt \tag{A.62}$$

in $\mathbb{P}^\nu$. Noticing that the value function has another representation.

$$\begin{aligned}V_t = \sup_{\nu\in\mathbb{V}} \mathbb{E}^\nu \Bigg(& g(T,X_T,I_T) + \int_t^T f(r,X_r,I_r)dr \\ & - \int_{(t,T]\times\mathbb{I}\times\mathbb{J}} h(r,X_{r^-},I_{r^-},i,\zeta)N'(dr\times di\times d\zeta)\Bigg| \mathscr{F}_t\Bigg)\end{aligned} \tag{A.63}$$

[4] Y_t is called a minimal solution if Y_t is a solution and $Y_t \leq \tilde{Y}_t$ for all solution $\tilde{Y}_t$.

Let $(Y,U,U',K) \in \mathbb{S}^2 \times \mathbb{L}^2_N \times \mathbb{L}^2_{N'} \times \mathbb{A}^2$ be any solution of the CBSDE (A.37-A.38). Taking the essential supremum of conditional expectation under $\mathbb{P}^\nu$ on Y_t (A.37) ($\widetilde{N}$ is not affected by the change of measure), we have

$$\begin{aligned} Y_t = \sup_{\nu\in\mathbb{V}} \mathbb{E}^\nu \Bigg(& g(T,X_T,I_T) + \int_t^T f(r,X_r,I_r)dr \qquad (A.64) \\ & - \int_{(t,T]\times\mathbb{I}\times\mathbb{J}} U'(r,i,\zeta)N'(dr\times di\times d\zeta) + K_T - K_t \Big| \mathscr{F}_t \Bigg) \\ \geq \sup_{\nu\in\mathbb{V}} \mathbb{E}^\nu \Bigg(& g(T,X_T,I_T) + \int_t^T f(r,X_r,I_r)dr \qquad (A.65) \\ & - \int_{(t,T]\times\mathbb{I}\times\mathbb{J}} h(r,X_{r^-},I_{r^-},i,\zeta)N'(dr\times di\times d\zeta) \Big| \mathscr{F}_t \Bigg) \\ = & V_t \qquad (A.66) \end{aligned}$$

Conversely, it can well-known that

$$\begin{aligned} Q_t = \sup_{\nu\in\mathbb{V}} \mathbb{E}^\nu \Bigg(& g(T,X_T,I_T) + \int_0^T f(r,X_r,I_r)dr \qquad (A.67) \\ & - \int_{(0,T]\times\mathbb{I}\times\mathbb{J}} h(r,X_{r^-},I_{r^-},i,\zeta)N'(dr\times di\times d\zeta) \Big| \mathscr{F}_t \Bigg) \\ = & V_t + \int_0^t f(r,X_r,I_r)dr - \int_{(0,t]\times\mathbb{I}\times\mathbb{J}} h(r,X_{r^-},I_{r^-},i,\zeta)N'(dr\times di\times d\zeta) \qquad (A.68) \end{aligned}$$

is a càdlàg $\mathbb{P}^\nu$-supermartingale $\forall \nu \in \mathbb{V}$ (see [151, proposition 4.2] or [152, theorem 2.1.1]). Since $V_t \leq Y_t \in \mathbb{S}^2$ and $f(t,x,i), h(t,x,i,j,\zeta)$ grows sublinearly with x, we can see that $Q \in \mathbb{S}^2$. Take $\nu(t,i,\zeta) = 0$ and by Doob-Meyer decomposition under $\mathbb{P}^0$,

$$Q_t = V_0 + M^0_t - K^0_t \qquad (A.69)$$

where M^0_t is a càdlàg P^0-martingale with $M^0_0 = 0$ and K^0_t is a càdlàg, predictable and increasing process with $K^0_0 = 0$. Since $Q \in \mathbb{S}^2$, we have $M^0 \in \mathbb{S}^2$ and $K^0 \in \mathbb{A}^2$. By martingale representation theorem, we can express M^0_t in $\widetilde{N}_j$ and $\widetilde{N}'^0$. However, when $\nu(t,i,\zeta) = 0$, $N'^\nu(dt\times di\times d\zeta) = 0$. As $M^0 \in \mathbb{S}^2$, there exists $U_j \in \mathbb{L}^2_N$ such that

$$Q_t = V_0 + \sum_{j=1}^d \int_{(0,t]\times\mathbb{K}} U_j(r,k)\widetilde{N}_j(dr\times dk) - K^0_t \qquad (A.70)$$

Substitute Q_t in (A.68)

$$V_t = g(T,X_T,I_T) + \int_t^T f(r,X_r,I_r)dr - \sum_{j=1}^{d} \int_{(t,T]\times\mathbb{K}} U_j(r,k)\widetilde{N}_j(dr\times dk) - \int_{(t,T]\times\mathbb{I}\times\mathbb{J}} h(r,X_{r^-},I_{r^-},i,\zeta)N'(dr\times di\times d\zeta) + K_T^0 - K_t^0 \quad \text{(A.71)}$$

Hence $(V,U,h,K^0) \in \mathbb{S}^2 \times \mathbb{L}^2_N \times \mathbb{L}^2_{N'} \times \mathbb{A}^2$ is a minimal solution of the CBSDE (A.37-A.38).

Y is unique by definition of minimal solution. The uniqueness of (U,U',K) follows by identifying the predictable and totally inaccessible parts as well as from the fact that N and N' have no common jump.

■

A.5 Numerical Scheme

One way to solve the CFBSDE (A.35-A.38) is via solving the associated penalized (unconstrained) FBSDE (A.72) which converges weakly to the CFBSDE as $p \longrightarrow \infty$ [27].

$$Y_t = g(T,X_T,I_T) + \int_t^T f_p(r,X_r,I_r,U'(r,\bullet))dr - \sum_{j=1}^{d} \int_{(t,T]\times\mathbb{K}} U_j(r,k)\widetilde{N}_j(dr\times dk) - \int_{(t,T]\times\mathbb{I}\times\mathbb{J}} U'(r,i,\zeta)\widetilde{N'}(dr\times di\times d\zeta) \quad \text{(A.72)}$$

$$f_p(r,X_r,I_r,U'(r,\bullet)) = f(r,X_r,I_r) + \int_{\mathbb{I}\times\mathbb{J}} \Big(p\Big(U'(r,i,\zeta) - h(r,X_r,I_r,i,\zeta)\Big)^+ - U'(r,i,\zeta)\Big)\lambda'\mu'(r,di\times d\zeta) \quad \text{(A.73)}$$

Therefore, the methods for solving CFBSDE boil down to the ones for solving unconstrained FBSDE. We recommend using the forward scheme [153] but we will first discuss the more intuitive backward scheme and its drawbacks.

For simplicity, we will assume the mark distributions of both N and N' are non-random and independent of time.

A.5.1 Backward Scheme

The period $[0,T]$ is divided into a time grid $0 = t_0 < t_1 < ... < t_N = T$ of N[5] uniform intervals of length $\Delta t = T/N$ and M sample paths are generated. For simplicity, we denote the value of $X(t_n)$ as X_n and the filtration $\mathscr{F}_{t_n}$ as $\mathscr{F}_n$.

We assume the forward equation is a pure-jump process, and it can be simulated exactly without any discretization error. For details about simulation of multivariate marked Hawkes process, please refer to Section 3.4.1.

It is easy to see that the backward equation can be approximated by

$$Y_n = Y_{n+1} + f_p(t_n, X_n, I_n, U'(t_n, \bullet))\Delta t - \sum_{j=1}^{d} \int_{\mathbb{K}} U_j(t_n, k)\widetilde{N}_j((t_n, t_{n+1}] \times dk) \\ - \int_{\mathbb{I}\times\mathbb{J}} U'(t_n, i, \zeta)\widetilde{N}'((t_n, t_{n+1}] \times di \times d\zeta) \quad \text{(A.74)}$$

Since the terminal value Y_N is known to be $g(T, X_N, I_N)$, the backward equation can be computed backward from time T. However, such a solution will not be adapted, so we take conditional expectation given $\mathscr{F}_n$. Since integral of $\widetilde{N}_j$ is a martingale and f_p does not depend on U_j, this relieves our burden to compute U_j. Taking conditional expectation on (A.74) given $\mathscr{F}_n$, we get

$$Y_n = \mathbb{E}(Y_{n+1}|\mathscr{F}_n) + f_p(t_n, X_n, I_n, U'(t_n, \bullet))\Delta t \quad \text{(A.75)}$$

$U'(t_n, i, \zeta)$ is determined by discretizing the mark (i, ζ) and then taking conditional expectation on (A.74) after multiplying the martingale $\widetilde{N}'((t_n, t_{n+1}], i, \zeta)$ to both side of the equation (N' and N_j are independent).

$$U'(t_n, i, \zeta) = \frac{\mathbb{E}(Y_{n+1}\widetilde{N}'((t_n, t_{n+1}], i, \zeta)|\mathscr{F}_n)}{\lambda'\mu(i, \zeta)\Delta t} \quad \text{(A.76)}$$

$$f_p(t_n, X_n, I_n, U'(t_n, \bullet)) = f(t_n, X_n, I_n) \\ + \sum_{i,\zeta} \left(p\Big(U'(t_n, i, \zeta) - h(r, X_n, I_n, i, \zeta)\Big)^+ - U'(t_n, i, \zeta) \right) \lambda'\mu'(i, \zeta) \quad \text{(A.77)}$$

[5]We use the standard symbol N for the number of time steps and readers are reminded not to confuse with the point process N, which will be denoted as $N_j(t)$.

Assuming Y_n and $U(t_n, i, \zeta)$ do not depend on $\lambda_j(t_n)$ (see Section 5.2.2), the conditional expectation given $\mathscr{F}_n$ is simply the conditional expectation given X_n, I_n and it can be estimated via Monte Carlo regression on a set of user chosen basis functions of size K with the M generated sample paths as input data. For details, readers can refer to [154, 155].

In addition to the choice of basis functions, there are five parameters in the backward scheme, namely p, λ', N, M, K, where p, λ' control the penalization error. As the unconstrained FBSDE converges to the CFBSDE when $p \longrightarrow \infty$, p needs to be reasonably large. However, a large p will introduce numerically instability to the algorithm (see numerical example 2). Also, if λ' is too small, there is simply no control event and U' will never be penalized. When both λ' and p are large, the discretization error will increase and we need a finer grid to better approximate f_p and U' due to more frequent control events. For Brownian motion, readers can refer to [156] for the formula of the error bound.

In the backward scheme, the output of the regression at t_{n+1} is used as an input for the next regression at t_n (nested regressions), so the error accumulates along the backward iteration. When we increase the number of time steps N, the number of regressions increases and so is the regression error. In order to keep the final error under control, we need to increase the number of sample paths M and the number of basis functions K. For the case of FBSDE drive by Brownian motion and using hypercube basis functions, Gobet [157] has the following suggestion: $M = O(N^{3+d})$, $K = O(N^d)$ where d is the dimension of state variable X. For example, when $N = 10^3,\ d = 4$, we have $M = O(10^{21}),\ K = O(10^{12})$, which is clearly impractical.

A.5.2 Forward Scheme

In the so-called forward scheme [153], in each iteration, all the $\{Y_n,\ U'_n\}_{n=0}^N$ are computed at the same time using the following well-known Picard type iteration for BSDE.

Theorem A.5.1 *If assumptions* (A.3.1), (A.3.2) *hold, then the solution* $(Y^{q+1}, U^{q+1}, U'^{q+1})$ *of the BSDE*

$$Y_t^{q+1} = g(T,X_T,I_T) + \int_t^T f_p(r,X_r,I_r,U'^q(r,\bullet))dr - \sum_{j=1}^{d} \int_{(t,T]\times\mathbb{K}} U_j^{q+1}(r,k)\widetilde{N}_j(dr\times dk) - \int_{(t,T]\times\mathbb{I}\times\mathbb{J}} U'^{q+1}(r,i,\zeta)\widetilde{N'}_j(dr\times di\times d\zeta) \quad \text{(A.78)}$$

converges to the solution (Y,U,U') *of the BSDE*

$$Y_t = g(T,X_T,I_T) + \int_t^T f_p(r,X_r,I_r,U'(r,\bullet))dr - \sum_{j=1}^{d} \int_{(t,T]\times\mathbb{K}} U_j(r,k)\widetilde{N}_j(dr\times dk) - \int_{(t,T]\times\mathbb{I}\times\mathbb{J}} U'(r,i,\zeta)\widetilde{N'}_j(dr\times di\times d\zeta) \quad \text{(A.79)}$$

as $q \longrightarrow \infty$ *in* $\mathbb{S}^2 \times \mathbb{L}_N^2 \times \mathbb{L}_{N'}^2$.

Proof See [147, Theorem 3.1.1] ■

Formulae for $\{(Y_n^{q+1}, U'^{q+1}_n)\}$ in the $(q+1)^{th}$ Picard iteration are derived using the similar method as the backward scheme except (Y_n^{q+1}, U'^{q+1}_n) is calculated directly from $(g(T,X_T,I_T), U'^q)$ in the previous Picard iteration instead of $(Y_{n+1}^{q+1}, U'^{q+1}_{n+1})$ from the next period.

$$U'^0_n = 0 \ \forall n \quad \text{(A.80)}$$

$$U'^{q+1}(t_n,i,\zeta) = \frac{\mathbb{E}\Big(\Big(g(T,X_T,I_T) + \sum_{k=n}^{N-1} f_p(t_k,X_k,I_k,U'^q(t_k,\bullet))\Delta t\Big)\widetilde{N'}((t_n,t_{n+1}],i,\zeta)\Big|\mathscr{F}_n\Big)}{\lambda'\mu(i,\zeta)\Delta t} \quad \text{(A.81)}$$

$$f_p(t_n,X_n,I_n,U'^{q+1}(t_n,\bullet)) = f(t_n,X_n,I_n) + \sum_{i,\zeta}\Big(p\big(U'^{q+1}(t_n,i,\zeta) - h(r,X_n,I_n,i,\zeta)\big)^+ - U'^{q+1}(t_n,i,\zeta)\Big)\lambda'\mu'(i,\zeta) \quad \text{(A.82)}$$

$$Y_n^{q+1} = \mathbb{E}\Big(g(T,X_T,I_T) + \sum_{k=n}^{N-1} f_p(t_k,X_k,I_k,U'^q(t_k,\bullet))\Delta t\Big|\mathscr{F}_n\Big) \quad \text{(A.83)}$$

The conditional expectation is again computed using Monte Carlo regression [154, 155].

Compared with the backward scheme, which needs only one iteration to compute Y_0, the forward scheme need several iterations albeit the Picard iteration usually converges very fast. However, the regressions at different times t_n in the forward scheme can be computed in parallel as the input is X, I and U_n^q in the previous Picard iteration whereas

in the backward scheme, the regression at time t_n depends on the regression result at time t_{n+1}.

The second and more important difference is that the regression error does not accumulate over time. In each regression, $g(T, X_T, I_T)$ can be simulated exactly in our pure-jump model and U'^q will be close to the true U' when q is sufficiently large, so the regression error will mostly depend on the choice of basis functions and the number of sample paths M. Although the regression error will accumulate over Picard iterations, the number of iterations required is usually very small.

In short, we can choose a relatively large p, λ' and N such that the penalization error and discretization error is sufficiently small. Then the final error will only depend on the goodness-of-fit of the worst regression. Since the regression error does not accumulate over time, it will only depends on M, K and the choice of basis functions but not p, λ', N (see [153] for error bound in the case of Brownian motion).

A.5.3 Numerical Examples

Example 1

Let N_t be a Poisson process with intensity λ. The solution of the following FBSDE

$$X_t = N_t \tag{A.84}$$

$$Y_t = \left(X_T + 2X_T^2\right) + \int_t^T \left(\exp\left(-(U_r - (3+4X_r))^2\right) - \lambda U_r\right) dr - \int_t^T U_r d\widetilde{N}_r \tag{A.85}$$

is given by $Y_t = (1-t) + X_t + 2X_t^2$, $U_t = 3 + 4X_{t^-}$ and in particular, $Y_0 = 1$.

We solve the FBSDE numerically using forward scheme with hypercube basis of 10 intervals ($K = 10$) and 5 Picard iterations. Because of memory limitation[6], we fix the number of time step N to 10 and use different values of λ to study the effect of discretization. The values of Y_0 for different intensities λ and numbers of sample paths M are shown in Table A.1.

[6]We already run our program on a special cluster of Purdue's supercomputer *Carter* with 256G memory.

Table A.1.

Value of Y_0 by solving the FBSDE (A.85) numerically with $N = 10, K = 10$ and 5 Picard iterations. True $Y_0 = 1$.

$M \backslash \lambda$	0.01	0.1	1
10^4	0.6714	0.7848	0.5916
10^5	0.7938	0.9253	0.6685
10^6	0.8276	0.9962	0.6712
10^7	0.8610	0.9969	0.6464

Although the regression error does not accumulate over time in the forward scheme, we still needs more than a few million sample paths M in order to get a reasonable estimate of Y_0. On the other hand, as the intensity λ increases, the events arrive more often and we need a finer grid in order to describe $U(t)$ more accurately. From Table A.1, it seems that we need something around $N \simeq 100\lambda T$ for the algorithm to converge. However, unlike the case for M, larger N may not always mean better result in the pure-jump model. The reason is that when the interval is too small relative to λ, there is simply no arrivals in most of the intervals. More regressions just increase the volatility of $U(t)$ and the rounding error when we compute $\sum_{n=0}^{N-1} f(X_n, U_n)\Delta t$.

Example 2

In this example, we introduce a penalization term $p(U_t - (3+4X_t))^+$ to the driver f, but since $U_t = 3 + 4X_{t^-}$, the penalization term is indeed 0 and thus the solution is the same as example 1.

$$X_t = N_t \tag{A.86}$$

$$\begin{aligned} Y_t = &\left(X_T + 2X_T^2\right) - \int_t^T U_r d\widetilde{N}_r \\ &+ \int_t^T \left(p(U_r - (3+4X_r))^+ + \exp\left(-(U_r - (3+4X_r))^2\right) - \lambda U_r\right) dr \end{aligned} \tag{A.87}$$

However, because of the discretization, regression and Picard iteration error, $U(t)$ will be larger than $3 + 4X_{t^-}$ in some cases and a large p will cause the error to blow up (see

[156]) as shown in Table A.2. In this example, we fix $N = 10, M = 10^7, \lambda = 0.1, T = 1$ and use 5 Picard iterations.

Table A.2.

Effect of penalization on the forward scheme with $N = 10, M = 10^7, K = 10, \lambda = 0.1, T = 1$ and 5 Picard iterations. True $Y_0 = 1$.

p	0.01	0.10	0.25	0.50	0.75	1.00
Y_0	0.9972	0.9998	1.0058	1.1051	3.3716	14.6054

Example 3

In this example, we illustrate the effect of marks to the accuracy of the numerical solution.

$$X_t = \int_{(0,t]\times\mathbb{K}} kN(dr \times dk) \tag{A.88}$$

$$Y_t = X_T + \int_t^T \int_{\mathbb{K}} \Big(\exp\big(-(U(r,k)-k)^2\big) - \lambda U(r,k) \Big) \mu(dk) dr - \int_{(t,T]\times\mathbb{K}} U(r,k) \widetilde{N}(dr \times dk) \tag{A.89}$$

The solution of the FBSDE is $Y_t = (T-t) + X_t,\ U(t,k) = k$ for any mark distribution $\mu(dk)$. Similar to previous cases, we fix $N = 10, M = 10^7, \lambda = 0.1, T = 1$ and use 5 Picard iterations. The mark distribution $\mu(dk)$ is discrete uniform on $\{0, 1C, 2C, .., 9C\}$ with $C = 1, 10, ..., 100000$.

Table A.3.

Effect of marks on the forward scheme with $N = 10, M = 10^7, K = 10, \lambda = 0.1, T = 1$ and 5 Picard iterations. True $Y_0 = 1$.

C	1	10	100	1000	10000	100000
$Y_0(M = 10^7)$	0.9715	0.8100	0.2389	-0.1308	-1.5024	-15.0239
$Y_0(M = 10^8)$	0.9967	0.8996	0.4324	0.0828	0.3829	3.8290

From Table A.3, we can see that the error increases as the marks become more volatile. Moreover the computation time is 10 times longer as we need run a regression for each value of k. Also the memory requirement drastically increases we need to hold more immediate results during the Monte Carlo regression. The accuracy can be improved by having more sample paths M but it will further aggravate the memory requirement and the computational speed of the numerical scheme.

REFERENCES

REFERENCES

[1] L. Glosten and P. Milgrom. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Fiancial Economics*, 14(1):71–100, 1985.

[2] A. S. Kyle. Continuous auctions and insider trading. *Econometrica*, 53(6):1315, 1985.

[3] Securities Exchange Commission. Concept release on equity market structure. *Federal Register*, 75(13):3594–3614, 2010.

[4] The Board of the International Organization of Securities Commissions. Trading fee models and their impact on trading behaviour. Technical report, International Organization of Securities Commissions, 2013.

[5] M. Garman. Market microstructure. *Journal of Fiancial Economics*, 3(3):257–275, 1976.

[6] Y. Amihud and H. Mendelson. Dealership market: Market-making with inventory. *Journal of Fiancial Economics*, 8(1):31–53, 1980.

[7] H. Stoll. The supply of dealer services in securities markets. *The Journal of Finance*, 33(4):1133–1151, 1978.

[8] T. Ho and H. Stoll. Optimal dealer pricing under transactions and return uncertainty. *Journal of Fiancial Economics*, 9(1):47–73, 1981.

[9] M. Avellaneda and S. Stoikov. High-frequency trading in a limit order book. *Quantitative Finance*, 8(3):217–224, 2008.

[10] P. Fodra and M. Labadie. High-frequency market-making with inventory constraints and directional bets. *arXiv preprint arXiv:1206.4810*, 2012.

[11] O. Guéant, C.-A. Lehalle, and J. Fernandez-Tapia. Dealing with the inventory risk: A solution to the market making problem. *Mathematics and financial economics*, 7 (4):477–507, 2013.

[12] F. Guilbaud and H. Pham. Optimal high-frequency trading with limit and market orders. *Quantitative Finance*, 13(1):79–94, 2013.

[13] P. Gopikrishnan, V. Plerou, X. Gabaix, and H. E. Stanley. Statistical properties of share volume traded in financial markets. *Physical Review E*, 62(4):R4493, 2000.

[14] S. Maslov and M. Mills. Price fluctuations from the order book perspective—empirical facts and a simple model. *Physica A: Statistical Mechanics and its Applications*, 299(1):234–246, 2001.

[15] X. Gabaix, P. Gopikrishnan, V. Plerou, and H. E. Stanley. Institutional investors and stock market volatility. *The Quarterly Journal of Economics*, 121(2):461–504, 2006.

[16] P. Weber and B. Rosenow. Order book approach to price impact. *Quantitative Finance*, 5(4):357–364, 2005.

[17] M. Potters and J.-P. Bouchaud. More statistical properties of order books and price impact. *Physica A: Statistical Mechanics and its Applications*, 324(1):133–140, 2003.

[18] B. Biais, P. Hillion, and C. Spatt. An empirical analysis of the limit order book and the order flow in the paris bourse. *The Journal of Finance*, 50(5):1655–1689, 1995.

[19] B. Tóth, I. Palit, F. Lillo, and J. D. Farmer. Why is order flow so persistent? *arXiv preprint arXiv:1108.1632*, 2011.

[20] J. Da Fonseca and R. Zaatour. Clustering and mean reversion in a Hawkes microstructure model. *Journal of Futures Markets*, 2014.

[21] F. Pomponio and F. Abergel. *Trade-throughs: Empirical facts and application to lead-lag measures*, pages 3–16. Springer, 2011.

[22] J. Large. Measuring the resiliency of an electronic limit order book. *Journal of Financial Markets*, 10(1):1–25, 2007.

[23] E. Bacry, S. Delattre, M. Hoffmann, and J. F. Muzy. Some limit theorems for Hawkes processes and application to financial statistics. *Stochastic Processes and their Applications*, 2013.

[24] M. G. Crandall and P.-L. Lions. Viscosity solutions of Hamilton-Jacobi equations. *Transactions of the American Mathematical Society*, 277(1):1–42, 1983.

[25] S. Tang and J. Yong. Finite horizon stochastic optimal switching and impulse controls with a viscosity solution approach. *Stochastics: An International Journal of Probability and Stochastic Processes*, 45(3-4):145–176, 1993.

[26] B. K. Øksendal and A. Sulem. *Applied stochastic control of jump diffusions*. Springer, 2005.

[27] I. Kharroubi, J. Ma, H. Pham, and J. Zhang. Backward SDEs with constrained jumps and quasi-variational inequalities. *The Annals of Probability*, 38(2):794–840, 2010.

[28] R. Elie and I. Kharroubi. BSDE representations for optimal switching problems with controlled volatility. *Stochastics and Dynamics*, 14(03), 2014.

[29] J.-P. Bouchaud, J. D. Farmer, and F. Lillo. How markets slowly digest changes in supply and demand. *Handbook of financial markets: dynamics and evolution*, 1:57, 2009.

[30] R. Cont. Statistical modeling of high-frequency financial data. *Signal Processing Magazine, IEEE*, 28(5):16–25, 2011.

[31] A. Chakraborti, I. M. Toke, M. Patriarca, and F. Abergel. Econophysics review: I. empirical facts. *Quantitative Finance*, 11(7):991–1012, 2011.

[32] T. Andersen, T. Bollerslev, F. Diebold, and P. Labys. The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, 96(453): 42–55, 2001.

[33] O. E. Barndorff-Nielsen. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):253–280, 2002.

[34] O. Kallenberg. *Random measures*. Akademie-Verlag, 1983.

[35] A. Karr. *Point processes and their statistical inference*. CRC press, 1991.

[36] P. Brémaud. *Point processes and queues*. Springer, 1981.

[37] D. Daley and D. Vere-Jones. *An introduction to the theory of point processes: Volume I: Elementary theory and methods*. Springer, 2003.

[38] D. Daley and D. Vere-Jones. *An introduction to the theory of point processes: Volume II: General theory and structure*. Springer, 2007.

[39] F. Papangelou. The conditional intensity of general point processes and an application to line processes. *Probability Theory and Related Fields*, 28(3):207–226, 1974.

[40] P. A. Meyer. Démonstration simplifiée d'un théorème de knight. In *Séminaire de probabilités v université de strasbourg*, pages 191–195. Springer, 1971.

[41] F. Papangelou. Integrability of expected increments of point processes and a related random change of scale. *Transactions of the American Mathematical Society*, 165: 483–506, 1972.

[42] A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

[43] Y. Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27, 1988.

[44] Y. Ogata and H. Akaike. On linear intensity models for mixed doubly stochastic poisson and self-exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 102–107, 1982.

[45] D. Oakes. The markovian self-exciting process. *Journal of Applied Probability*, pages 69–77, 1975.

[46] A. G. Hawkes and D. Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, pages 493–503, 1974.

[47] J. Møller and J. G. Rasmussen. Perfect simulation of Hawkes processes. *Advances in applied probability*, pages 629–646, 2005.

[48] P. Brémaud and L. Massoulié. Stability of nonlinear Hawkes processes. *The Annals of Probability*, pages 1563–1588, 1996.

[49] P. Brémaud, G. Nappo, and G. Torrisi. Rate of convergence to equilibrium of marked Hawkes processes. *Journal of Applied Probability*, 39(1):123–136, 2002.

[50] C. G. Bowsher. Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, 141(2):876–912, 2007.

[51] P. Brémaud and L. Massoulié. Hawkes branching point processes without ancestors. *Journal of Applied Probability*, 38(1):122–135, 2001.

[52] T. Jaisson and M. Rosenbaum. Limit theorems for nearly unstable Hawkes processes. *arXiv preprint arXiv:1310.2033*, 2013.

[53] L. Zhu. Central limit theorem for nonlinear Hawkes processes. *Journal of Applied Probability*, 50(3):760–771, 2013.

[54] P. Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2009.

[55] J. Jacod and A. N. Shiryaev. *Limit theorems for stochastic processes*. Springer-Verlag Berlin, 1987.

[56] T. Ozaki. Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31(1):145–155, 1979.

[57] Y. Ogata. On Lewis' simulation method for point processes. *Information Theory, IEEE Transactions on*, 27(1):23–31, 1981.

[58] J. Zhuang, Y. Ogata, and D. Vere-Jones. Analyzing earthquake clustering features by using stochastic reconstruction. *Journal of Geophysical Research*, 109(B5), 2004.

[59] P. A. Lewis and G. S. Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly*, 26(3):403–413, 1979.

[60] A. Veen and F. P. Schoenberg. Estimation of space–time branching process models in seismology using an EM–type algorithm. *Journal of the American Statistical Association*, 103(482):614–624, 2008.

[61] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[62] G. Mclachlan and T. Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, 2007.

[63] D. Marsan and O. Lengline. Extending earthquakes' reach through cascading. *Science*, 319(5866):1076–1079, 2008.

[64] P. Halpin and P. Boeck. Modelling dyadic interaction with Hawkes processes. *Psychometrika*, 78(4):793–814, 2013.

[65] P. F. Halpin. *A scalable EM algorithm for Hawkes processes*. New developments in quantitative psychology. Springer, 2013.

[66] J. F. Olson and K. M. Carley. Exact and approximate EM estimation of mutually exciting Hawkes processes. *Statistical Inference for Stochastic Processes*, 16(1): 63–80, 2013.

[67] L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, pages 1029–1054, 1982.

[68] J. Da Fonseca and R. Zaatour. Hawkes process: Fast calibration, application to trade clustering, and diffusive limit. *Journal of Futures Markets*, 2013.

[69] G. Gusto and S. Schbath. Fado: A statistical method to detect favored or avoided distances between occurrences of motifs using the Hawkes' model. *Statistical Applications in Genetics and Molecular Biology*, 4(1):1119, 2005.

[70] C. De Boor. *A practical guide to splines*. Springer-Verlag New York, 1978.

[71] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory*, pages 267–281. Akademinai Kiado, 1973.

[72] P. Reynaud-Bouret and S. Schbath. Adaptive estimation for Hawkes processes; application to genome analysis. *The Annals of Statistics*, 38(5):2781–2822, 2010.

[73] N. R. Hansen, P. Reynaud-Bouret, and V. Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *arXiv preprint arXiv:1208.0570*, 2012.

[74] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[75] E. Lewis and G. Mohler. A nonparametric EM algorithm for multiscale Hawkes processes. *Preprint*, 2011.

[76] I. Good and R. Gaskins. Nonparametric roughness penalties for probability densities. *Biometrika*, 58(2):255–277, 1971.

[77] K. Zhou, H. Zha, and L. Song. Learning triggering kernels for multi-dimensional Hawkes processes. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1301–1309, 2013.

[78] D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, 2004.

[79] E. Bacry, K. Dayri, and J. F. Muzy. Non-parametric kernel estimation for symmetric Hawkes processes. application to high frequency financial data. *The European Physical Journal B*, 85(5), 2012.

[80] E. Bacry and J. F. Muzy. Hawkes model for price and trades high-frequency dynamics. *Quantitative Finance*, 14(7):1147–1166, 2014.

[81] E. Bacry and J. F. Muzy. Second order statistics characterization of Hawkes processes and non-parametric estimation. *arXiv preprint arXiv:1401.0903*, 2014.

[82] B. Noble. *Methods based on the Wiener-Hopf technique*. Pergamon Press New York, 1958.

[83] G. Chandlen and I. G. Graham. The convergence of Nyström methods for Wiener-Hopf equations. *Numerische Mathematik*, 52(3):345–364, 1987.

[84] A. Baddeley, R. Turner, J. Møller, and M. Hazelton. Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(5):617–666, 2005.

[85] M. Wilk and R. Gnanadesikan. Probability plotting methods for the analysis of data. *Biometrika*, 55(1):1–17, 1968.

[86] A. N. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4(1):83–91, 1933.

[87] N. V. Smirnov. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou*, 2(2), 1939.

[88] N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, pages 279–281, 1948.

[89] G. M. Ljung and G. E. Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 1978.

[90] F. P. Schoenberg. Multidimensional residual analysis of point process models for earthquake occurrences. *Journal of the American Statistical Association*, 98(464), 2003.

[91] I. Muni Toke and F. Pomponio. Modelling trades-through in a limited order book using Hawkes processes. *Economics: The Open-Access, Open-Assessment E-Journal*, 6(2012-22), 2012.

[92] I. Muni Toke. "Market Making" in an order book model and its impact on the spread. In *Econophysics of order-driven markets*, pages 49–64. Springer, 2011.

[93] H. Shek. Modeling high frequency market order dynamics using self-excited point process. *Available at SSRN 1668160*, 2011.

[94] A. Fauth and C. Tudor. Modeling first line of an order book with multivariate marked point processes. *arXiv preprint arXiv:1211.4157*, 2012.

[95] P. Hewlett. Clustering of order arrivals, price impact and trade path optimisation. In *Workshop on Financial Modeling with Jump processes, Ecole Polytechnique*, pages 6–8, 2006.

[96] A. Alfonsi and P. Blanc. Dynamic optimal execution in a mixed-market-impact Hawkes price model. *arXiv preprint arXiv:1404.0648*, 2014.

[97] T. Jaisson. Market impact as anticipation of the order flow imbalance. *arXiv preprint arXiv:1402.1288*, 2014.

[98] E. Bacry, S. Delattre, M. Hoffmann, and J. F. Muzy. Modelling microstructure noise with mutually exciting point processes. *Quantitative Finance*, 13(1):65–77, 2013.

[99] Y. Aït-Sahalia, P. A. Mykland, and L. Zhang. How often to sample a continuous-time process in the presence of market microstructure noise. *Review of Financial Studies*, 18(2):351–416, 2005.

[100] T. W. Epps. Comovements in stock prices in the very short run. *Journal of the American Statistical Association*, 74(366a):291–298, 1979.

[101] T. Andersen, T. Bollerslev, F. Diebold, and P. Labys. Great realizations. *Risk Magazine*, 2000.

[102] S. L. Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of financial studies*, 6(2):327–343, 1993.

[103] F. M. Bandi and J. R. Russell. Separating microstructure noise from volatility. *Journal of Financial Economics*, 79(3):655–692, 2006.

[104] D. Duffie and R. Kan. A yield-factor model of interest rates. *Mathematical finance*, 6(4):379–406, 1996.

[105] D. Duffie, J. Pan, and K. Singleton. Transform analysis and asset pricing for affine jump-diffusions. *Econometrica*, 68(6):1343–1376, 2000.

[106] L. Zhu. Limit theorems for a Cox-Ingersoll-Ross process with Hawkes jumps. *arXiv preprint arXiv:1309.5625*, 2013.

[107] Y. Aït-Sahalia, J. Cacho-Diaz, and R. Laeven. Modeling financial contagion using mutually exciting jump processes. *NBER Working Paper*, 2010.

[108] V. Filimonov and D. Sornette. Quantifying reflexivity in financial markets: Toward a prediction of flash crashes. *Physical Review E*, 85(5):056108, 2012.

[109] V. Filimonov, D. Bicchetti, N. Maystre, and D. Sornette. Quantification of the high level of endogeneity and of structural regime shifts in commodity markets. *Journal of International Money and Finance*, 2013.

[110] R. J. Shiller. *Irrational exuberance*. Princeton University Press, 2005.

[111] J. B. De Long, A. Shleifer, L. H. Summers, and R. J. Waldmann. Positive feedback investment strategies and destabilizing rational speculation. *Journal of Finance*, pages 379–395, 1990.

[112] L. Harris. Order exposure and parasitic traders. *Preprint*, 1997.

[113] G. Soros. *The alchemy of finance*. John Wiley Sons, 2003.

[114] S. J. Hardiman, N. Bercot, and J.-P. Bouchaud. Critical reflexivity in financial markets: A Hawkes process analysis. *The European Physical Journal B*, 86(10):1–9, 2013.

[115] V. Filimonov and D. Sornette. Apparent criticality and calibration issues in the Hawkes self-excited point process model: Application to high-frequency financial data. *arXiv preprint arXiv:1308.6756*, 2013.

[116] S. J. Hardiman and J.-P. Bouchaud. Branching ratio approximation for the self-exciting Hawkes process. *arXiv preprint arXiv:1403.5227*, 2014.

[117] D. R. Cox. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(2):129–164, 1955.

[118] Y. Ogata. Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402, 1998.

[119] V. Chavez-Demoulin, A. Davison, and A. Mcneil. Estimating value-at-risk: A point process approach. *Quantitative Finance*, 5(2):227–234, 2005.

[120] E. Errais, K. Giesecke, and L. R. Goldberg. Affine point processes and portfolio credit risk. *SIAM Journal on Financial Mathematics*, 1(1):642–665, 2010.

[121] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, 2008.

[122] C. Blundell, J. Beck, and K. A. Heller. Modelling reciprocating relationships with Hawkes processes. In *Advances in Neural Information Processing Systems*, pages 2609–2617, 2012.

[123] M. Krumin, I. Reutsky, and S. Shoham. Correlation-based analysis and generation of multiple spike trains using Hawkes models with an exogenous input. *Frontiers in computational neuroscience*, 4, 2010.

[124] V. Pernice, B. Staude, S. Cardanobile, and S. Rotter. How structure determines correlations in neuronal networks. *PLoS computational biology*, 7(5):e1002059, 2011.

[125] T. Liniger. *Multivariate Hawkes processes*. PhD thesis, Swiss Federal Institute of Technology, 2009.

[126] L. Zhu. *Nonlinear Hawkes processes*. PhD thesis, New York University, 2013.

[127] E. Bacry and J. F. Muzy. Hawkes model for price and trades high-frequency dynamics. *Quantitative Finance*, 14(7):1147–1166, 2014.

[128] P. Holgate. Estimation for the bivariate Poisson distribution. *Biometrika*, 51(1-2): 241–287, 1964.

[129] J. F. Egginton, B. F. Van Ness, and R. A. Van Ness. Quote stuffing. *Available at SSRN 1958281*, 2014.

[130] E. J. Lee, K. S. Eom, and K. S. Park. Microstructure-based manipulation: Strategic behavior and performance of spoofing traders. *Journal of Financial Markets*, 16(2): 227–252, 2013.

[131] R. Cont and A. De Larrard. Price dynamics in a markovian limit order market. *SIAM Journal on Financial Mathematics*, 4(1):1–25, 2013.

[132] M. Avellaneda, J. Reed, and S. Stoikov. Forecasting prices from level-i quotes in the presence of hidden liquidity. *Algorithmic Finance*, 1(1):35–43, 2011.

[133] R. Cont and A. De Larrard. Price dynamics in a markovian limit order market. *SIAM Journal on Financial Mathematics*, 4(1):1–25, 2013.

[134] F. Lillo and J. D. Farmer. The long memory of the efficient market. *Studies in Nonlinear Dynamics & Econometrics*, 8(3), 2004.

[135] A. Ellul, C. W. Holden, P. Jain, and R. Jennings. Order dynamics: Recent evidence from the NYSE. *Journal of Empirical Finance*, 14(5):636–661, 2007.

[136] B. Zheng, F. Roueff, and F. Abergel. Modelling bid and ask prices using constrained Hawkes processes: Ergodicity and scaling limit. *SIAM Journal on Financial Mathematics*, 5(1):99–136, 2014.

[137] G. Scopino. The (questionable) legality of high-speed "pinging" and "front running" in the futures market. *Connecticut law review*, 47(3):607, 2015.

[138] K. Y. L. Fong and W.-M. Liu. Limit order revisions. *Journal of Banking and Finance*, 34(8):1873–1885, 2010.

[139] S. Delattre, N. Fournier, and M. Hoffmann. High dimensional Hawkes processes. *arXiv preprint arXiv:1403.5764*, 2014.

[140] R. West, R. Andrews, J. Schluetter, D. Garrison, and M. Burns. System and method for estimating order position (US Patent 8762254 B2), 2014.

[141] B. Bouchard and N. Touzi. Discrete-time approximation and monte-carlo simulation of backward stochastic differential equations. *Stochastic Processes and their Applications*, 111(2):175–206, 2004.

[142] D. Easley, N. M. Kiefer, M. O'hara, and J. B. Paperman. Liquidity, information, and infrequently traded stocks. *The Journal of Finance*, 51(4):1405–1436, 1996.

[143] D. Easley, M. M. L. De Prado, and M. O'hara. Flow toxicity and liquidity in a high-frequency world. *Review of Financial Studies*, page hhs053, 2012.

[144] H. Kushner. Necessary conditions for continuous parameter stochastic optimization problems. *SIAM Journal on Control*, 10(3):550–565, 1972.

[145] S. Peng and M. Xu. Constrained BSDE and viscosity solutions of variation inequalities. *arXiv preprint arXiv:0712.0306*, 2007.

[146] A. Bensoussan and J. L. Lions. *Impulse control and quasi-variational inequalities*. Gauthier-Villars, 1984.

[147] L. Delong. *Backward stochastic differential equations with jumps and their actuarial and financial applications*. Springer, 2013.

[148] R. Elie and I. Kharroubi. Probabilistic representation and approximation for coupled systems of variational inequalities. *Statistics & Probability Letters*, 80(17):1388–1396, 2010.

[149] R. Elie and I. Kharroubi. Adding constraints to BSDEs with jumps: An alternative to multidimensional reflections. *ESAIM: Probability and Statistics*, 18:233–250, 2014.

[150] I. Kharroubi, N. Langrené, and H. Pham. A numerical algorithm for fully nonlinear HJB equations: An approach by control randomization. *Monte Carlo Methods and Applications*, 20(2):145–165, 2014.

[151] D. O. Kramkov. Optional decomposition of supermartingales and hedging contingent claims in incomplete security markets. *Probability Theory and Related Fields*, 105(4):459–479, 1996.

[152] N. El Karoui and M. C. Quenez. Dynamic programming and pricing of contingent claims in an incomplete market. *SIAM journal on Control and Optimization*, 33(1): 29–66, 1995.

[153] C. Bender and R. Denk. A forward scheme for backward sdes. *Stochastic Processes and their Applications*, 117(12):1793–1812, 2007.

[154] E. Gobet, J.-P. Lemor, and X. Warin. A regression-based monte carlo method to solve backward stochastic differential equations. *The Annals of Applied Probability*, 15(3):2172–2202, 2005.

[155] J.-P. Lemor, E. Gobet, and X. Warin. Rate of convergence of an empirical regression method for solving generalized backward stochastic differential equations. *Bernoulli*, 12(5):889–916, 2006.

[156] M. Bernhart, H. Pham, P. Tankov, and X. Warin. Swing options valuation: A BSDE with constrained jumps approach. In *Numerical methods in finance*, pages 379–400. Springer, 2012.

[157] E. Gobet. Introduction to stochastic calculus and to the resolution of pdes using monte carlo simulations. In *Advances in numerical simulation in physics and engineering*, pages 107–178. Springer, 2014.

VITA

VITA

Chi Wai, also known as Baron, was born and grown up in Hong Kong but he had been living in Japan before moving to United States to pursue his PhD.

Prior to working with Prof. Frederi G. Viens on the topic of high-frequency data modeling and algorithmic trading, Baron was a vice president in the Quantitative and Derivative Strategies (QDS) of Morgan Stanley (Japan), where he conducted quantitative research on the Japanese equity market. Before that, he worked in Credit Suisse (Japan) as a vice president undertaking research on quantitative equity valuation methodologies.

In addition to the CFA charter, Baron holds a MSc Statistics from Purdue University, MSc Financial Engineering from Columbia University, MSc Finance from Chinese University of Hong Kong and BEng Electrical & Electronic Engineering from University of Hong Kong.