The Fundamentals of Market Liquidity: Optimal Execution and Market Making

Karim Layoun

MS QCF - Georgia Institute of Technology

ABSTRACT

We present an overview of the formulations and solutions to problems arising from the organization of electronic exchanges around a limit order book, namely to the problems of optimal execution and market making.

Keywords: Microstructure, Limit Order Book, Optimal Execution, Market Making, High-Frequency Trading

Introduction

Initially, microstrucutre was the prerogative of economists, attempting to understand the price formation process of financial assets, and the impact of different market structures and their inefficiencies on the process. The electronification of markets around the 2000's brought microstructure into the purview of "quants", solving the problem of optimal scheduling to buy and sell a large number of stocks; however, they largely ignored the many sources of market frictions. And so, the early beginnings of optimal execution may be seen as an extension of the classical microstructure literature.

Before the 2007-2008 crisis, quantitative finance was more preoccupied with financial assets with increasingly complex payoffs and less so with financial markets. The crisis has exposed the limitations of risk-neutral pricing approaches, freeing "quants" to pursue other research topics. Moreover, the rise of microstructure in the research literature is largely attributed to regulatory changes affecting the trading environment; namely Reg NMS and MiFID, fragmenting the liquidity of stocks over multiple trading venues.

The new literature on microstructure can be broadly separated into two research topics: dealing with the execution of large orders and designing strategies for market making. Furthermore, optimal execution models can be used to solve classical problems in quantitative finance, such as pricing and hedging derivatives, and market making is of great relevance to high-frequency trading strategies.

The exercise of hedging and pricing derivatives in illiquid markets cannot escape the mechanics of exchange as the fundamental assumption of continuous trading of the underlying is violated, rendering the classical risk-neutral approaches inadequate. Also, the rise of electronic markets has incentivized arbitrageurs to test the extent of the word instantaneous in the classical statement on no-arbitrage - there are never opportunities to make an instantaneous risk-free profit - in their pursuit of profit, placing trades at ever increasing speeds. High-frequency trading is inextricably intertwined with microstructure [9].

The project is an attempt to understand and implement foundational models in the modern microstructure literature.

THE LIMIT ORDER BOOK

No discussion on market microstructure can be complete without a description of the Limit Order Book, the most important feature of trading venues. Exchanges are organized around a visible LOB, a transparent system matching buy and sell orders of market participants on a price/time priority basis or other matching algorithms. A limit order may be seen as a triplet, comprising the order side (buy or sell/bid or ask), the order size, and the order price. The best bid price is the highest bid price, and the best ask price is the lowest ask price. And so, the best bid price is strictly less than the best ask price, by construction, as overlapping orders on opposite sides of the LOB would be matched and eliminated from the LOB. Also, the mid-price the average of the best bid and ask prices, and the bid-ask spread is the difference between the best bid and the best ask. Conceptually, exchanges organized around a LOB are a decentralized double auctions where market participants compete by submitting bid and ask orders, hopefully in search of the true price of financial assets [8].

Tick Size

Tick sizes vary across exchanges and asset prices, affecting the trading environment. There exists optimal tick sizes for a given stock [4].

Fees

Competition amongst exchanges enabled by regulation has driven fees down, as exchanges have adopted different fee structures, where pricing only accounts for executed orders. There exists optimal fee structure, subsidizing liquidity makers or liquidity takers depending on market characteristics [5].

High-Frequency Trading

A diverse field, sometimes detrimental, but mostly beneficial to the price formation process. HFT may be seen as a necessary evil to achieve a low fee environment and an efficient market. Market makers are most relevant to the project. Indeed, high-frequency market makers are essential to solve the problem arising from the competition between traders and the competition between platforms to organize the competition between traders [11].

Optimal Execution

Suppose a market participant has a portfolio of assets that he/she wishes to liquidate. What is the best strategy to unwind a portfolio during a set amount of time? In 1998, Bertsimas and Lo published the first paper aiming to solve the problem of optimal execution with market orders, minimizing execution price while ignoring various sources of risk [3]. Mainly, they did not consider that the market price may vary during the liquidation of a portfolio. Their model often results in a strategy executing small and equal orders at constant speed.

Almgren and Chriss presented a more realistic model capturing the tradeoff between execution costs and price risk, leading to non-trivial optimal execution strategies. They proposed a simple way to model market impact, consisting of two elements: instantaneous and permanent impact. Indeed, the price of execution of a market order depends on the depth of the LOB and the price of assets must be impacted by trades. However, their simple model fails to capture the complexity of real market impact [1].

Modern execution algorithms do not merely use market orders for optimal liquidation. They generally consist of two parts, a schedule and a strategy. The strategic part uses all types of orders across different trading venues to obtain good execution prices, orchestrated by a schedule, an optimal trading curve computed by generalized Almgren-Chriss models.

SINGLE-ASSET GENERALIZED ALMGREN-CHRISS MODEL

The model can be presented in continuous and discrete time. However, we shall present the model in discrete time as it is more amenable to numerical methods. Moreover, Almgren and Chriss initially presented a discrete-time model [1].

Let us introduce some notation:

- q_0, the initial single asset portfolio to be unwinded
- $[0,\ T]$, the execution time window
- $\delta t = \frac{T}{N}$, a time interval such that $[0,T] \equiv \{t_0,\ t_1,\ \ldots,\ t_{N-1},\ t_N\}$
- $\delta t \cdot v_{n+1}$, the number of shares purchased during interval $[t_n,\ t_{n+1}]$ such that $v_{n+1} < 0$ corresponds to a sale of shares
- $\delta t \cdot V_{n+1}$, the market volume during interval $[t_n,\ t_{n+1}]$

The model for the evolution of the portfolio is

$$q_{n+1} = q_n + \delta t \cdot v_{n+1}, \quad 0 \leq n < N$$

The model for the evolution of the asset mid-price is

$$S_{n+1} = S_n + \sigma\sqrt{\delta t}\varepsilon_{n+1} + k\delta t \cdot v_{n+1}, \quad 0 \leq n < N$$

where

- $\varepsilon_0,\ \ldots,\ \varepsilon_N$ are i.i.d $N(0,1)$

- σ is the asset volatility

- $k \geq 0$ represents the magnitude of permanent market impact. Indeed, buying (selling) shares of the asset increases (decreases) its price

The model for the evolution of the cash account is

$$X_{n+1} = X_n - \delta t \cdot v_{n+1} \cdot S_n - L\left(\frac{v_{n+1}}{V_{n+1}}\right) \delta t \cdot V_{n+1}$$

where L is the execution cost function.

Before introducing the optimization problem, we must discuss our modeling choices and assumptions:

- In the optimal execution literature, prices are assumed to be normally distributed as opposed to log-normally distributed, for simplicity. The probability of non-positive prices during typical liquidation timeframes are very low, and the dynamics of normally and log-normally distributed prices are very similar as $\delta t \longrightarrow 0$ for reasonable values of σ

- k is a constant coefficient, implying that the permanent market impact is linear, and guaranteeing the absence of dynamic arbitrage [7] [13].

- $L: \ \mathbb{R} \longrightarrow \mathbb{R}$ has the following properties

  - $L(0) = 0$

  - L is strictly convex, $L(\lambda x + (1-\lambda y)) < \lambda L(x) + (1-\lambda)L(y), \ \forall\, x, y, x \neq y, \lambda \in (0,1)$

  - L is asymptotically super-linear, $\lim_{|\rho| \longrightarrow +\infty} \frac{L(\rho)}{|\rho|} = +\infty$

  - L is commonly set to be a convex power function of the form

    $$L(\rho) = \eta |\rho|^{1+\phi} + \psi|\rho|, \ \phi, \ \psi > 0$$

    where the $\psi|\rho|$ term corresponds to the bid-ask spread cost

The Optimization Problem

Bertsimas and Lo, ignoring measures of risk, chose to maximize $E\left[X_N\right]$ [3]. As is common in portfolio optimization, Almgren and Chriss have chosen a mean-variance model maximizing $E\left[X_N\right] - \frac{\gamma}{2} V\left[X_N\right]$, $\gamma \geq 0$ [1]. We consider a CARA (Constant Absolute Risk Aversion) objective function

$$E\left[-e^{-\gamma X_N}\right], \quad \gamma \geq 0$$

where γ is the absolute risk aversion of the trader. A solution to the optimization problem is a strategy $(v_n)_{n \in [0,N]} \in \mathscr{A}$, where $\mathscr{A}$ is the set of admissible control processes [10].

Deterministic vs Stochastic Strategies

A deterministic strategy would enable a trader to compute the optimal trading curve before execution, a desirable characteristic. In addition, an optimal deterministic control process implies that the strategy is independent of price changes. Let $\mathscr{A}_{\text{det}}$ be the restricted set of deterministic control processes,

$$\mathscr{A}_{\text{det}} = \left\{ (v_1, \ldots, v_N) \in \mathbb{R}^N, \sum_{n=0}^{N-1} \delta t \cdot v_n = -q_0 \right\}$$

Guéant [9] and others have proven that no liquidation strategy can do better than the optimal deterministic strategy

$$\sup_{v \in \mathscr{A}} E\left[-e^{-\gamma X_N}\right] = \sup_{v \in \mathscr{A}_{\text{det}}} E\left[-e^{-\gamma X_N}\right]$$

And so, we can restrict the set of feasible controls to $\mathscr{A}_{\text{det}}$, further justifying the use of generalized Almgren-Chriss models as the first part of modern liquidation strategies.

A Unique Solution

To maximize the objective function, we must start with computing the final value of the cash account, the final proceeds of the liquidation strategy

$$X_N = X_0 - \sum_{n=0}^{N-1} v_{n+1} S_n \cdot \delta t - \sum_{n=0}^{N-1} L\left(\frac{v_{n+1}}{V_{n+1}}\right) V_{n+1} \cdot \delta t$$

We then expand it with the dynamics of q_n and S_n

$$X_0 + q_0 S_0 + \sigma \sqrt{\delta t} \sum_{n=0}^{N-1} q_{n+1} \varepsilon_{n+1} + k \sum_{n=0}^{N-1} q_{n+1} v_{n+1} \delta t$$
$$- \sum_{n=0}^{N-1} L\left(\frac{v_{n+1}}{V_{n+1}}\right) V_{n+1} \delta t,$$

use the rearranged equation

$$v_{n+1} \cdot \delta t = q_{n+1} - q_n$$

and the algebraic trick

$$q_{n+1} = \frac{q_{n+1} + q_n}{2} + \frac{q_{n+1} - q_n}{2}$$

resulting in the form

$$X_N = X_0 + q_0 S_0 - \frac{k}{2} q_0^2 + \sigma \sqrt{\delta t} \sum_{n=0}^{N-1} q_{n+1} \varepsilon_{n+1}$$
$$+ \frac{k}{2} \sum_{n=0}^{N-1} v_{n+1}^2 \delta t^2 - \sum_{n=0}^{N-1} L\left(\frac{v_{n+1}}{V_{n+1}}\right) V_{n+1} \delta t.$$

The δt^2 term vanishes in the limit $\delta t \to 0$. Since we have restricted the set of admissible controls to $\mathscr{A}_{\text{det}}$, $X_N \sim N(E[X_N], V[X_N])$, $(S_n)_n$ being the only random process

$$E\left[X_N\right] = X_0 + q_0 S_0 - \frac{k}{2} q_0^2 - \sum_{n=0}^{N-1} L\left(\frac{v_{n+1}}{V_{n+1}}\right) V_{n+1} \delta t$$
$$V\left[X_N\right] = \sigma^2 \delta t \sum_{n=0}^{N-1} q_{n+1}^2,$$

and the objective function is

$$
\begin{aligned}
E\left[-\exp\left(-\gamma X_N\right)\right] &= -\exp\left(-\gamma E\left[X_N\right] + \frac{1}{2}\gamma^2 V\left[X_N\right]\right) \\
&= -\exp\left(-\gamma\left(X_0 + q_0 S_0 - \frac{k}{2}q_0^2\right)\right) \\
&\quad \times \exp\left(\gamma\left(\sum_{n=0}^{N-1} L\left(\frac{v_{n+1}}{V_{n+1}}\right) V_{n+1}\delta t + \frac{\gamma}{2}\sigma^2 \delta t \sum_{n=0}^{N-1} q_{n+1}^2\right)\right).
\end{aligned}
$$

The function e^x is monotonic non-decreasing. Consequently, minimizing the above objective function is equivalent to minimizing

$$
\sum_{n=0}^{N-1} L\left(\frac{v_{n+1}}{V_{n+1}}\right) V_{n+1}\delta t + \frac{\gamma}{2}\sigma^2 \delta t \sum_{n=0}^{N-1} q_{n+1}^2,
$$

which we can express as a function of q to directly solve for the optimal trading curve

$$
\begin{aligned}
P(q) : & \sum_{n=0}^{N-1} L\left(\frac{q_{n+1} - q_n}{V_{n+1}\delta t}\right) V_{n+1}\delta t + \frac{\gamma}{2}\sigma^2 \delta t \sum_{n=0}^{N-1} q_{n+1}^2 \\
& q \in \mathscr{C}_{\text{det}} = \{(q_0,\, \ldots,\, q_N) : q_0 = q_0,\; q_N = 0\}
\end{aligned}
$$

The resulting form of the problem is a convex optimization problem, as the objective function is strictly convex, and the feasible region $\mathscr{C}_{\text{det}}$ is convex. And so, the problem has a unique solution q^*, characterized by the Hamiltonian system [9]

$$
\begin{cases}
p_{n+1} = p_n + \delta t \gamma \sigma^2 q_{n+1}^*, & 0 \leq n < N-1 \\
q_{n+1}^* = q_n^* + \delta t V_{n+1} H'\left(p_n\right), & 0 \leq n < N
\end{cases}
\qquad q_0^* = q_0, \quad q_N^* = 0
$$

where H is the Legendre-Fenchel transform of L

$$
H(p) = \sup_{\rho} \rho p - L(p)
$$

H is always convex, and is differentiable since L is strictly convex, enabling us to write the system as a series of equations and directly express the unique solution q^* as a function of the dual variable p. Moreover, if

$$
L(\rho) = \eta |\rho|^{1+\phi} + \psi |\rho|
$$

then

$$
H(p) = \sup_{\rho} p\rho - \eta|\rho|^{1+\phi} - \psi|\rho| =
\begin{cases}
0 & \text{if } |p| \leq \psi \\
\phi \eta \left(\frac{|p|-\psi}{\eta(1+\phi)}\right)^{1+\frac{1}{\phi}} & \text{otherwise.}
\end{cases}
$$

A Brief Tangent

Many a method in convex optimization uses gradients to arrive at a solution. However, some convex functions are not everywhere differentiable. Consequently, we resort to using subdifferentials or supporting lines. A supporting line of a function f at a point x everywhere underestimates f at x. Let f^* be the Legendre-Fenchel transform of f. If f admits a supporting line at x with slope k, then f^* at k admits a supporting line with slope x [14]. This supporting line duality is essential to the correctness of the above Hamiltonian system.

Towards the Original Almgren-Chriss Model

Let $L(\rho) = \eta\rho^2$ and $V_n = V \ \forall \ n$. The objective function becomes

$$\sum_{n=0}^{N-1}\left(\frac{\eta - \frac{k}{2}V\delta t}{V}\right)v_{n+1}^2\delta t + \frac{\gamma}{2}\sigma^2\delta t\sum_{n=0}^{N-1}q_{n+1}^2$$

and the Hamiltonian system simplifies to

$$\begin{cases} p_{n+1} = p_n + \delta t\gamma\sigma^2 q_{n+1}^*, & 0 \leq n < N-1 \\ q_{n+1}^* = q_n^* + \frac{V}{2\eta}\delta t p_n, & 0 \leq n < N \end{cases}$$

which further simplifies to a second-order recursive equation

$$q_{n+2}^* - \left(2 + \frac{\gamma\sigma^2 V}{2n}\delta t^2\right)q_{n+1}^* + q_n^* = 0$$

The equation admits an explicit solution

$$q_n^* = q_0\frac{\sinh\left(\alpha\left(T - t_n\right)\right)}{\sinh(\alpha T)}$$

where α is defined by

$$2(\cosh(\alpha\delta t) - 1) = \frac{\gamma\sigma^2 V}{2\eta}\delta t^2.$$

When $\delta t \to 0$, we get the optimal trading curve

$$q_n^* = q_0\frac{\sinh\left(\sqrt{\frac{\gamma\sigma^2 V}{2\eta}}(T - t_n)\right)}{\sinh\left(\sqrt{\frac{\gamma\sigma^2 V}{2\eta}}T\right)}, \quad 0 \leq n \leq N$$

and optimal deterministic strategy $(v_n)_n$

$$v_n^* = q_n^{*'} = -q_0\sqrt{\frac{\gamma\sigma^2 V}{2\eta}}\frac{\cosh\left(\sqrt{\frac{\gamma\sigma^2 V}{2\eta}}(T - t_n)\right)}{\sinh\left(\sqrt{\frac{\gamma\sigma^2 V}{2\eta}}T\right)}$$

Sensitivity to Parameters

In our efforts to better understand the tradeoff between execution costs and price risk, we present the sensitivity of the optimal trading curve to the liquidity, volatility, and risk aversion parameters.

$\underline{\gamma}$

The risk aversion parameter represents a trader's sensitivity to price risk. And so, the greater the γ, the faster the unwinding of the portfolio.

σ

Asset volatility directly quantifies price risk. And so, the higher the volatility, the greater the exposure to price risk over time, and the faster the unwinding of the portfolio.

η & V

The execution cost function L is increasing in η and decreasing in V (increasing in the participation rate). And so, the higher the execution cost, the slower the unwinding of the portfolio.

IMPLEMENTATION

We have implemented the basis of modern optimal execution algorithms, the trading schedule, to be computed before portfolio liquidation as the optimal strategy is deterministic and independent of price fluctuations

Algorithm 1 The Optimal Trading Curve

```
Input: $\gamma,\ \sigma,\ V,\ T,\ N,\ q_0$
Output: $Q$
$Q[0] = q_0$
for $n = 1$ to $N$ do
    $Q[n] = q_0 \frac{\sinh\left(\sqrt{\frac{\gamma\sigma^2 V}{2\eta}}(T - t_n)\right)}{\sinh\left(\sqrt{\frac{\gamma\sigma^2 V}{2\eta}}T\right)}$
end for
```

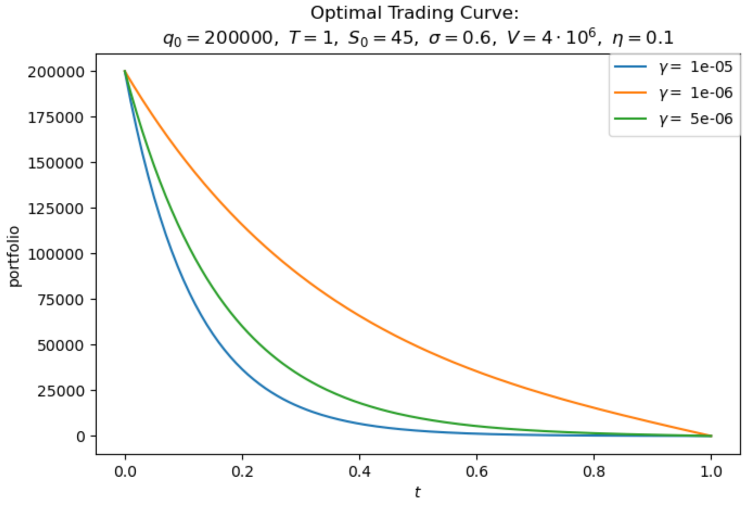


Figure 1. Optimal Trading Curve for Different Risk Aversion Parameters

Market Making

While optimal execution is generally a liquidity taking endeavor, market making is a liquidity providing endeavor, essential in the price formation process. A market maker is a trader who continuously submits buy and sell limit orders, expecting to make a profit on the bid-ask spread. Market making models provide solutions to the optimal quoting problem. The first model for market making is presented by Garman [6] in 1976, and extended by Ho and Stoll [12] in 1981. The most famous market making model is the Avellanada-Stoikov solution [2], the basis of all subsequent market making models, and in reality better suited for quote-driven markets. Guéant et al. [10] proved that the equations in the Avellanada-Stoikov model could be transformed into a linear system of ODEs, and further simplified with the asymptotic expansion of the optimal bid/ask quotes [9].

The optimal market making problem is twofold; first, a tradeoff between large spreads thus low volume and small spreads and high volume; second, a tradeoff between conservative quoting thus a persistent inventory and aggressive quoting thus an increased exposure to price risk. And so, to control price risk, a market maker (MM) adopts a skewed strategy. For example, a MM with a long inventory should quote aggressively on the ask side and conservatively on the bid side, to maximize the probability of selling assets.

In the following sections, we present the framework and general sketch of the derivation for the Avellanada-Stoikov model. Moreover, we present the closed form approximation of Gueant et al. for the original model and its extensions to more realistic asset price dynamics. Finally, we detail the implementation of the simulation to test the Avellanada-Stoikov market making strategy. The derivation of the optimal strategies and their closed form is lengthier and more involved than that of optimal execution, comprising a series of variable changes and algebraic manipulations reducing the HJB equation associated with the problem to a system of ordinary differential equation, to be solved with numerical methods, i.e. a Euler scheme. The interested reader should refer to the original texts [2] [10].

THE AVELLANADA-STOIKOV MODEL

The model is a solution to the market making problem for a single asset, where each transaction is a trade for one unit of the asset. Let us introduce some notation:

- $\left(S_t^b\right)_t$, the stochastic process modeling bid prices
- $(S_t^a)_t$, the stochastic process modeling ask prices
- $\left(N_t^b\right)_t$, the point process modeling the number of filled buy limit orders
- $(N_t^a)_t$, the point process modeling the number of filled sell limit orders
- $(q_t)_t$, the trader inventory such that

  $$q_t = N_t^b - N_t^a$$

- $\left(\lambda_t^b\right)_t$ and $(\lambda_t^a)_t$, the intensity processes of $\left(N_t^b\right)_t$ and $(N_t^a)_t$ respectively
  - Note that the intensity process $(\lambda_t)_t$of a point process $(N_t)_t$ is defined as

    $$\lim_{dt \longrightarrow 0} \frac{P\{N(t,t+dt] > 0\}}{dt}$$

The model for the evolution of the asset mid-price (reference price) $(S_t)_t$ is

$$dS_t = \sigma dW_t, \quad W_t \sim N(0, dt)$$

The main assumption of the Avellaneda-Stoikov model is symmetric, exponential arrival rates

$$\lambda_t^b = \Lambda^b(\delta_t^b) = Ae^{-\kappa\delta_t^b}$$
$$\lambda_t^a = \Lambda^a(\delta_t^a) = Ae^{-\kappa\delta_t^a}$$
$$A > 0, \kappa > 0$$

where

$$\delta_t^b = S_t - S_t^b$$
$$\delta_t^a = S_t^a - S_t$$

and A represents the liquidity of the asset, κ represents the price sensitivity of market participants. Consequently, the instantaneous probability of an order being filled is a function of the distance between the reference price and quoted price, adequately capturing the described tradeoff between conservative and aggressive quoting; i.e., the greater the distance between the quote and market price, the smaller the probability to trade.

We impose a constraint on the size of the market maker's inventory $q_t \in [-Q,\ Q]$, resulting in the updated intensities

$$\lambda_t^b = \Lambda^b(\delta_t^b) 1_{q_{t-}<Q}$$
$$\lambda_t^a = \Lambda^a\left(\delta_t^a\right) 1_{q_{t-}>-Q}$$

Hence we arrive at the model for the dynamics of the market maker's cash account

$$dX_t = S_t^a dN_t^a - S_t^b dN_t^b = (S_t + \delta_t^a)\, dN_t^a - (S_t - \delta_t^b) dN_t^b$$

The Optimization Problem

The market maker aims to optimize the expected utility of the Mark-to-Market value of his porfolio at time T. Namely, his Pnl is equal to the value of his cash account and that of his remaining inventory (liquidated at reference price S_T)

$$X_T + q_T S_T$$

and we use CARA utility functions as in optimal execution. Therefore, the optimal market making problem is

$$\sup_{(\delta_t^a)_t, \left(\delta_t^b\right)_t \in \mathscr{A}} \mathbb{E}\left[-\exp\left(-\gamma(X_T + q_T S_T)\right)\right]$$

where

- $\mathscr{A}$ is the set of below-bounded predictable processes
- γ represents the absolute risk aversion of the market maker

A Solution Sketch

A series of algebraic manipulations and variable changes transforms the associated Hamilton-Jacobi-Bellman equation

$$
\begin{aligned}
|q| < Q: \qquad & \partial_t u(t,x,q,s) + \frac{1}{2}\sigma^2 \partial_{ss}^2 u(t,x,q,s) \\
& + \sup_{\delta^b} \lambda^b\left(\delta^b\right)\left[u\left(t, x-s+\delta^b, q+1, s\right) - u(t,x,q,s)\right] \\
& + \sup_{\delta^a} \lambda^a\left(\delta^a\right)\left[u\left(t, x+s+\delta^a, q-1, s\right) - u(t,x,q,s)\right] = 0 \\
q = Q: \qquad & \partial_t u(t,x,Q,s) + \frac{1}{2}\sigma^2 \partial_{ss}^2 u(t,x,Q,s) \\
& + \sup_{\delta^a} \lambda^a\left(\delta^a\right)\left[u\left(t, x+s+\delta^a, Q-1, s\right) - u(t,x,Q,s)\right] = 0 \\
q = -Q: \qquad & \partial_t u(t,x,-Q,s) + \frac{1}{2}\sigma^2 \partial_{ss}^2 u(t,x,-Q,s) \\
& + \sup_{\delta^b} \lambda^b\left(\delta^b\right)\left[u\left(t, x-s+\delta^b, -Q+1, s\right) - u(t,x,-Q,s)\right] = 0
\end{aligned}
$$

with the boundary condition

$$
u(T,x,q,s) = -\exp(-\gamma(x+qs)), \quad \forall q \in \{-Q, \ldots, Q\}
$$

to a linear system of ordinary differential equations

$$
M = \begin{pmatrix}
\alpha Q^2 & -\eta & 0 & \cdots & \cdots & \cdots & 0 \\
-\eta & \alpha(Q-1)^2 & -\eta & 0 & \ddots & \ddots & \vdots \\
0 & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\
\vdots & \ddots & \ddots & 0 & -\eta\alpha(Q-1)^2 & -\eta & \\
0 & \cdots & \cdots & \cdots & 0 & -\eta & \alpha Q^2
\end{pmatrix}
$$

where $\alpha = \frac{k}{2}\gamma\sigma^2$, $\eta = A\left(1+\frac{\gamma}{k}\right)^{-\left(1+\frac{k}{\gamma}\right)}$, and

$$
\begin{aligned}
v(t) &= (v_{-Q}(t), v_{-Q+1}(t), \ldots, v_0(t), \ldots, v_{Q-1}(t), v_Q(t))' \\
&= \exp(-M(T-t)) \times (1, \ldots, 1)'
\end{aligned}
$$

The resulting form is a stochastic control problem with the following solution, as derived by Guéant et al. [10]

$$
\begin{aligned}
s - s^{b*}(t,q,s) = \delta^{b*}(t,q) &= \frac{1}{k}\ln\left(\frac{v_q(t)}{v_{q+1}(t)}\right) + \frac{1}{\gamma}\ln\left(1+\frac{\gamma}{k}\right), \quad q \neq Q \\
s^{a*}(t,q,s) - s = \delta^{a*}(t,q) &= \frac{1}{k}\ln\left(\frac{v_q(t)}{v_{q-1}(t)}\right) + \frac{1}{\gamma}\ln\left(1+\frac{\gamma}{k}\right), \quad q \neq -Q
\end{aligned}
$$

resulting in the following bid-ask spread

$$
\psi^*(t,q) = -\frac{1}{k}\ln\left(\frac{v_{q+1}(t)v_{q-1}(t)}{v_q(t)^2}\right) + \frac{2}{\gamma}\ln\left(1+\frac{\gamma}{k}\right), \quad |q| \neq Q
$$

GUÉANT ET AL. CLOSED FORM APPROXIMATIONS

Guéant et al. [10] have shown that when $T \to +\infty$, we can approximate an explicit solution for the optimal bid/ask quotes

$$\delta_\infty^{b*}(q) \simeq \frac{1}{\gamma}\ln\left(1+\frac{\gamma}{\kappa}\right) + \frac{2q+1}{2}\sqrt{\frac{\sigma^2\gamma}{2\kappa A}\left(1+\frac{\gamma}{\kappa}\right)^{1+\frac{\kappa}{\gamma}}},$$
$$\delta_\infty^{a*}(q) \simeq \frac{1}{\gamma}\ln\left(1+\frac{\gamma}{\kappa}\right) - \frac{2q-1}{2}\sqrt{\frac{\sigma^2\gamma}{2\kappa A}\left(1+\frac{\gamma}{\kappa}\right)^{1+\frac{\kappa}{\gamma}}}.$$

resulting in the following bid-ask spread

$$\psi_\infty^{*}(q) = \delta_\infty^{b*}(q) + \delta_\infty^{a*}(q) \simeq \frac{2}{\gamma}\ln\left(1+\frac{\gamma}{\kappa}\right) + \sqrt{\frac{\sigma^2\gamma}{2\kappa A}\left(1+\frac{\gamma}{\kappa}\right)^{1+\frac{\kappa}{\gamma}}}.$$

and ”skewness” of the market strategy

$$\delta_\infty^{b*}(q) - \delta_\infty^{a*}(q) \simeq 2q\sqrt{\frac{\sigma^2\gamma}{2\kappa A}\left(1+\frac{\gamma}{\kappa}\right)^{1+\frac{\kappa}{\gamma}}}$$

Model Extensions

Guéant et al. have generalized the model to different asset price dynamics, including drift and market impact.

Asset Price Drift

Avellanada and Stoikov assumed that the market maker has no opinion on the drift [2]. Let μ be the average rate of growth of the asset price. The reference price of the asset evolves according to

$$dS_t = \mu dt + \sigma dW_t$$

The solution to the optimal market making problem and optimal bid-ask spread are

$$\delta_\infty^{b*}(q) \simeq \frac{1}{\gamma}\ln\left(1+\frac{\gamma}{k}\right) + \left[-\frac{\mu}{\gamma\sigma^2} + \frac{2q+1}{2}\right]\sqrt{\frac{\sigma^2\gamma}{2kA}\left(1+\frac{\gamma}{k}\right)^{1+\frac{k}{\gamma}}}$$
$$\delta_\infty^{a*}(q) \simeq \frac{1}{\gamma}\ln\left(1+\frac{\gamma}{k}\right) + \left[\frac{\mu}{\gamma\sigma^2} - \frac{2q-1}{2}\right]\sqrt{\frac{\sigma^2\gamma}{2kA}\left(1+\frac{\gamma}{k}\right)^{1+\frac{k}{\gamma}}}$$
$$\psi_\infty^{*}(q) \simeq \frac{2}{\gamma}\ln\left(1+\frac{\gamma}{k}\right) + \sqrt{\frac{\sigma^2\gamma}{2kA}\left(1+\frac{\gamma}{k}\right)^{1+\frac{k}{\gamma}}}$$

Market Impact

Avellanada and Stoikov did not consider the market impact of market making. A simple way to model such dynamics is

$$dS_t = \sigma dW_t + \xi dN_t^a - \xi dN_t^b, \quad \xi > 0$$

where the reference price decreases when a bid limit order is filled, and decreases when an ask limit order is filled. As discussed in the optimal execution section, market impact should be linear to guarantee the absence of dynamic arbitrage. The solution to the optimal market making problem and optimal bid-ask spread are

$$\delta_\infty^{b*}(q) \simeq \frac{1}{\gamma}\ln\left(1+\frac{\gamma}{k}\right) + \frac{\xi}{2} + \frac{2q+1}{2}e^{\frac{k}{4}\xi}\sqrt{\frac{\sigma^2\gamma}{2kA}\left(1+\frac{\gamma}{k}\right)^{1+\frac{k}{\gamma}}}$$

$$\delta_\infty^{a*}(q) \simeq \frac{1}{\gamma}\ln\left(1+\frac{\gamma}{k}\right) + \frac{\xi}{2} - \frac{2q-1}{2}e^{\frac{k}{4}\xi}\sqrt{\frac{\sigma^2\gamma}{2kA}\left(1+\frac{\gamma}{k}\right)^{1+\frac{k}{\gamma}}}$$

$$\psi_\infty^{*}(q) \simeq \frac{2}{\gamma}\ln\left(1+\frac{\gamma}{k}\right) + \xi + e^{\frac{k}{4}\xi}\sqrt{\frac{\sigma^2\gamma}{2kA}\left(1+\frac{\gamma}{k}\right)^{1+\frac{k}{\gamma}}}$$

IMPLEMENTATION

The basis of our implementation are three classes representing the different components of the simulation:

- An asset class (Asset), with class attributes
  - S, the price of the asset at $t = 0$
  - σ, the volatility of the asset
  - ξ, the coefficient of market impact
  - μ, the drift component of asset price dynamics
  - A, the asset liquidity
  - **process**, a standard normal random variable generator
- A trader class (Trader), with class attributes
  - γ, the risk aversion coefficient
  - Q, the current inventory/portfolio
  - X, the current value of the cash account
- A strategy class (OptimalMarketMaking), with class attributes and methods
  - **trader**, an instance of the Trader class
  - **asset**, an instance of the Asset class
  - κ, the price sensitivity of market participants
  - T, the time horizon of the strategy
  - dt, the time step. Note that dt must be small enough to minimize the probability that the market maker receives multiple simultaneous orders, yet larger than the typical tick time. Indeed, if dt is shorter than the tick time, the agent will submit quotes faster than the market can fill them
  - $N = \frac{T}{dt}$, the number of time steps
  - M, the constraint on the size of the trader's inventory
  - $Q[\,]$, the trader's inventory over time
  - $X[\,]$, the trader's cash account over time
  - $P[\,]$, the trader's PnL over time
  - $S[\,]$, the asset price random walk

- $dSa[\,]$, the optimal ask quotes
- $dSb[\,]$, the optimal bid quotes
- $Na[\,]$, the number of filled sell limit orders
- $Nb[\,]$, the number of filled buy limit orders
- **simulator**, a Bernoulli random variable generator
- **generate_quotes**, the function for generating optimal quotes, described below in Algorithm 2
- **simulate_strategy**, the test function for the market making strategy, described below in Algorithm 3

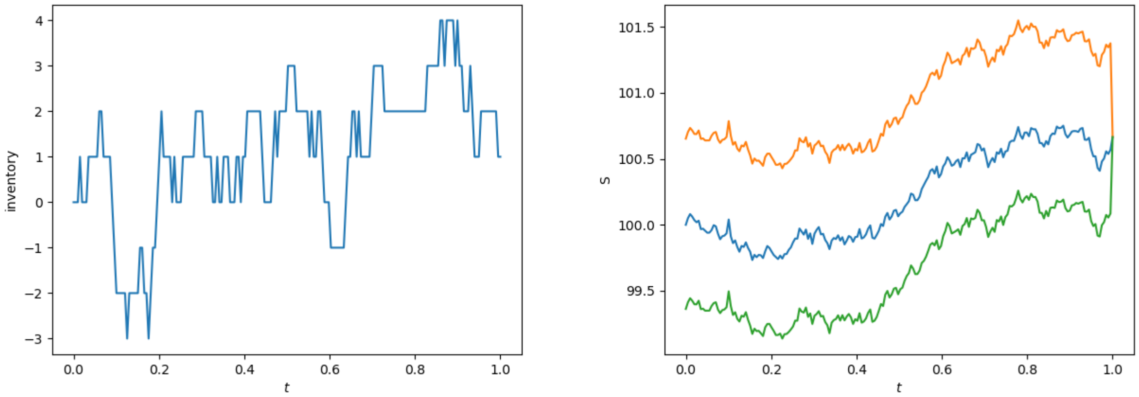
Algorithm 2 Generate Optimal Quotes

```
Input: self
Output: quotes (a tuple of two terms necessary to compute the optimal quotes)

term1 = $\frac{1}{\gamma}\ln\left(1+\frac{\gamma}{k}\right)+\frac{\xi}{2}$
term2 = $\left[-\frac{\mu}{\gamma\sigma^2}+\frac{2q+1}{2}\right]e^{\frac{k}{4}\xi}\sqrt{\frac{\sigma^2\gamma}{2kA}\left(1+\frac{\gamma}{k}\right)^{1+\frac{k}{\gamma}}}$

quotes = (term1, term2)
```

Algorithm 3 Testing the Market Making Strategy

```
Input self
Output: $P[\,]$

for $n = 1$ to $N$ do

    quotes = generate_quotes(self)
    $dSa[n-1]$ = quotes[0] + quotes[1]
    $dSb[n-1]$ = quotes[0] - quotes[1]

    if (simulator($Ae^{-\kappa \cdot dSa[n-1]} \cdot dt$) == 1 & $Q[n-1] > -M$) then
        $Na[n] = 1$
    end if
    if (simulator($Ae^{-\kappa \cdot dSb[n-1]} \cdot dt$) == 1 & $Q[n-1] < M$) then
        $Nb[n] = 1$
    end if

    $S[n] = \mu \cdot dt + \sigma \cdot \sqrt{dt} \cdot$ asset.process() $+ \xi \cdot Na[n-1] - \xi \cdot Nb[n-1]$
    $Q[n] = Nb[n] - Na[n]$
    $X[n] = X[n-1] + (S[n] + dSa[n-1]) \cdot Na[n] - (S[n] - dSb[n-1]) \cdot Nb[n]$
    $P[n] = X[n] + Q[n] \cdot S[n]$
end for
```

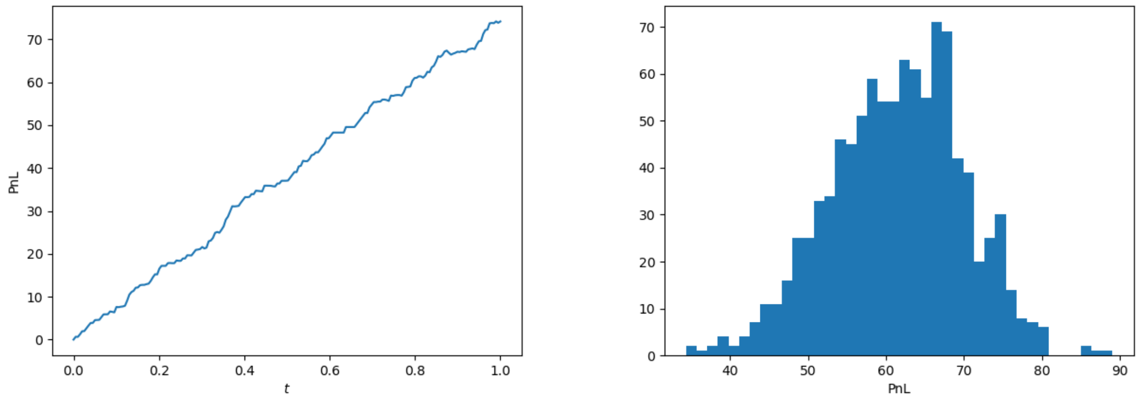
We ran a 1000 simulations for the market making strategy with parameters set in the seminal paper by Avellanada and Stoikov [2], obtaining very similar results: $S_0 = 100,\ T = 1,\ \sigma = 0.6,\ dt = 0.005,\ Q_0 = 0,\ \gamma = 0.1,\ \kappa = 1.5,\ A = 140$



(a) Inventory

(b) Asset Price

Figure 2. Evolution of Inventory and Asset Price in Simulation



(a) Example PnL

(b) PnL Distribution

Figure 3. PnL for 1000 Simulations

Conclusion

The trading environment is ever-evolving, with continuous technological progress and regulatory changes. The anticipated developments will only increase the importance of optimal execution and market making in optimal trading.

We have described and implemented foundational models in optimal execution and market making, providing intuition as to the solution methods of such problems and guidance as to the extension of the models to different market conditions. We hope the report and code constitute a comprehensive introduction to market microstructure and liquidity, and help fellow aspiring quants in their journey to high-frequency trading.

REFERENCES

[1] Almgren, R. and Chriss, N. (2000). Optimal execution of portfolio transactions. *Annalen der Physik*, 322(10):891–921.

[2] Avellanada, M. and Stoikov, S. (2000). High-frequency trading in a limit order book. *Quantitative Finance*, 8(3):217—224.

[3] Bertsimas, D. and Lo, A. W. (1998). Optimal control of execution costs. *Journal of Financial Markets*, 1(1):1–50.

[4] Dayri, K. and Rosenbaum, M. (2015). Large tick assets: Implicit spread and optimal tick size. *Market Microstructure and Liquidity*, 01(01).

[5] Foucault, T., Kadan, O., and Kandel, E. (2013). Liquidity cycles and make/take fees in electronic markets. *The Journal of Finance*, 68(1):299–341.

[6] Garman, M. (1976). Market microstructure. *Journal of Financial Economics*, 3(3):257—-275.

[7] Gatheral, J. (2010). No-dynamic-arbitrage and market impact. *Quantitative Finance*, 10(7):749–759.

[8] Gould, M. D., Porter, M. A., Williams, S., McDonald, M., Fenn, D. J., and Howison, S. D. (2013). Limit order books. *Quantitative Finance*, 13(11):1709—-1742.

[9] Guéant, O. (2016). *The Financial Mathematics of Market Liquidity - From Optimal Execution to Market Making*. Chapman Hall, New York.

[10] Guéant, O., Lehalle, C.-A., and Fernandez-Tapia, J. (2012). Dealing with the inventory risk - a solution to the market making problem. *Journal of Financial Markets*, 1(1):1–50.

[11] Harris, L., of Exchanges, W. F., and Centre for European Policy Studies (Brussels, B. (2010). *Regulated exchanges: dynamic agents of economic growth*. Oxford University Press, New York.

[12] Ho, T. and Stoll, H. (1981). Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial Economics*, 9(1):47—735.

[13] Huberman, G. and Stanzl, W. (2004). Price manipulation and quasi-arbitrage. *Econometrica*, 72(4):1247–1275.

[14] Touchette, H. (2014). Legendre-fenchel transforms in a nutshell - stellenbosch university.