Algorithmic And High-Frequency Trading

June 9, 2025

Contents

1 Electronic Markets and the Limit Order Book

1.1 Electronic markets and how they function

Shares are claims of ownership on corporations. These claims are used by corporations to raise money. In the US, for these shares to be traded in an electronic exchange they have to be 'listed' by an exchange, and this implies fulfilling certain requirements in terms of the number of shareholders, price, etc. The listing process is usually tied to the first issuance of the public shares (initial public offering, or IPO). The fundamental value of these

shares is derived from the nature of the contract it represents. In its simplest form, it is a claim of ownership on the company that gives the owner the right to receive an equal share of the corporation's profits (hence the name, 'share') and to intervene in the corporate decision process via the right to vote in the corporation's annual general (shareholders') meetings. Such shares are called **ordinary shares** (or **common stock**) and are the most common type of shares.

The other primary instrument used by large corporations to raise capital is **bonds**. Bonds are contracts by which the corporation commits to paying the holder a regular income (interest) but gives them no decision rights. The differences between stocks and bonds are quite clear: shareholders have no guarantees on the magnitude and frequency of dividends but have voting rights, bondholders have guarantees of regular, pre-determined payments and no voting rights.

There are other instruments with characteristics from both these contracts, the most familiar of which is **preferred stock**. Preferred stock represents a hybrid of stocks and bonds: they are like bonds in that holders have no voting rights and receive a pre-arranged income, but the income they receive has fewer guarantees: its legal treatment is that of equity, rather than debt. This difference is especially relevant when the corporation is in financial distress, as debt is senior to all equity, so that in case of liquidation, debt holders' claims have priority over the corporation's assets -they get paid first. Equity holders, if they get paid, are paid only after all debtholders' claims are settled.

A **mutual fund** is an investment product that acts as a delegated investment manager. That is, when an investor buys a mutual fund, the investor gives her cash to a financial management company that will use the cash to build a portfolio of assets according to the fund's investment objective. This objective includes the fund's assets and investment strategy, and, of course, its management fees. The fund's assets can belong to a large number of possible asset classes, including all those described above: equities, bonds, cash, FX, real estate, etc. The fund's investment strategy refers to the style of investment, primarily whether the fund is actively managed or passively tracks an index.

An investor who puts money in a fund participates in both the appreciation and depreciation of the assets as allocated by the fund manager. In order to redeem her investment, i.e. to convert her investment into cash, the investor's options depend on the type of fund she purchased. There are two main types of mutual funds: **open-end** and **closed-end** funds. Closed-end funds are mutual funds that are not redeemable: the fund issues a fixed

number of shares usually only once, at inception, and investors cannot sell the shares back to the fund. The fund sells the shares initially through an IPO and these shares are listed on an exchange where investors buy and sell these shares to each other.

Open-end funds are funds with a varying number of shares. Shares can be created to meet the demand of new investors, or destroyed (bought back by the fund) as investors seek to redeem theirs. This process takes place once a day, as the value of the fund's (net) assets (its Net Asset Value, NAV) is determined after the market close. Thus, closed-end funds, that do not have to adjust their holdings in response to investor demand, have different liquidity requirements than open-end funds and thus may trade at prices different from their NAV.

A very popular type of fund that, like closed-end funds, are traded in electronic exchanges, are ETFs. Like mutual funds, ETFs act as delegated investment managers, but they differ in two key respects. First, ETFs tend to have very specific investment strategies, usually geared towards generating the same return as a particular market index (e.g., the S&P500). Second, they are not obligated to purchase investors' shares back. Rather, if an investor wants to return their share to the fund, the fund can transfer to the investor a basket of securities that mirrors that of the ETF. This is possible because the ETF sells shares in very large units (Creation Units) which are then broken up and resold as individual shares in the exchange. A Creation Unit can be as large as 50,000 shares. Overall, the general perception one gets is that investors who are looking to reduce their trading costs and find diversified investments prefer ETFs, while investors who are looking for managers with stock-picking or similar unusual skills and who aim to beat the market will prefer mutual funds.

Some investment firms feel that the regulation that is imposed on mutual fund managers to ensure they fulfill their fiduciary duties to investors are too constraining. In response to this they have created **hedge-funds**, funds that pursue more aggressive trading strategies and have fewer regulatory and transparency requirements. Because of the softer regulatory oversight, access to these investment vehicles is largely limited to accredited investors, who are expected to be better informed and able to deal with the fund's managers. Although these funds are not traded on exchanges, their managers are active participants in those markets.

There are also other securities traded in electronic exchanges; in particular, there is a great deal of electronic trading in derivative markets, especially futures, swaps and options, and these contracts are written on a wide variety of assets (bonds, FX, commodities, equities, indices). The concepts

and techniques we develop in this book apply to the trading of any of these assets, although we primarily focus our examples and applications on equities. However, when designing algorithms and strategies one must always take into account the specific issues associated with the types of assets one is trading in, as well as the specifics of the particular electronic exchange(s) and the trading objectives of other investors one is likely to meet there.

1.2 Classifying Market Participants

Three primary classes of traders (or trading strategies) below.

1. **Fundamental** (or **noise** or liquidity) **traders**: those who are driven by economic fundamentals outside the exchange.

2. **Informed traders**: traders who profit from leveraging information not reflected in market prices by trading assets in anticipation of their appreciation or depreciation.

3. **Market makers**: professional traders who profit from facilitating exchange in a particular asset and exploit their skills in executing trades.

1.3 Trading in Electronic Markets

1.3.1 Orders and the Exchange

In the basic setup, an electronic market has two types of orders: **Market Orders** (MOs), and **Limit Orders** (LOs).

- MOs are usually considered aggressive orders that seek to execute a trade immediately. By sending an MO, a trader indicates that she wants to buy or sell a certain quantity of shares at the best available price, and this will (usually) result in an immediate trade (execution).

- On the other hand, LOs are considered passive orders, as a trader sending in an LO indicates her desire to buy or sell at a given price up to a certain, maximum, quantity of shares. As the price offered in the LO is usually worse than the current market price (higher than the best buy price for sell LOs, and lower than the best sell price for buy LOs), it will not result in an immediate trade, and will thus have to wait until either it is matched with a new order that wants to trade at the offered price (and executed) or it is withdrawn (cancelled).

Orders are managed by a matching engine and a limit order book (LOB). The LOB keeps track of incoming and outgoing orders. The matching engine uses a well-defined algorithm that establishes when a possible trade can occur, and if so, which criterion is going to be used to select the orders that will be executed. Most markets prioritise MOs over LOs and then use a price-time priority whereby, if an MO to buy comes in, the buy order will be matched with the standing LOs to sell in the following way:

1. the incoming order will be matched with the LOs that offer the best price (for buy orders, the sell LOs with the lowest price)

2. if the quantity demanded is less than what is on offer at the best price, the matching algorithm selects the oldest LOs, the ones that were posted earliest, and executes them in order until the quantity of the MO is executed completely.

3. If the MO demands more quantity than that offered at the best price, after executing all standing LOs at the best price, the matching algorithm will proceed by executing against the LOs at the second-best price, then the third-best and so on until the whole order is executed.

LOs that have increasingly worse prices are referred to as LOs that are deeper in the LOB, and the process whereby an entering market order executes against standing LOs deeper in the LOB is called 'walking the book'.

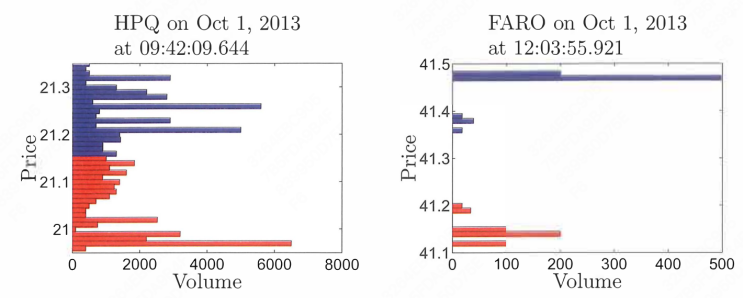

Figure 1.1 Snapshots of the NASDAQ LOB after the 10,000th event of the day. Blue bars represent the available sell LOs, red bars represent the available buy LOs.

Figure 1.3.1 shows a snapshot of the limit order book (LOB) on NASDAQ after the 10,000th event of the day for two stocks, FARO and HPQ, on Oct 1, 2013. The two are quite different. The one in the left panel corresponds to HPQ, a frequently traded and liquid asset. HPQ's LOB has LOs posted at every tick out to (at least) 20 ticks away from the midprice. In the

right panel, we have FARO's LOB. FARO is a seldom traded, illiquid asset. This asset has thinly posted bids and offers and irregular gaps in the LOB.

1.3.2 Alternate Exchange Structures

The above approach is not the only possible way to organise an exchange. For example, one could use an alternative matching algorithm, such as the **prorata rules** used in some money markets.

Beyond the legal definitions, we generically distinguish lit (open order book) from dark markets based on whether limit book informa tion is publicly available or not.

1.3.3 Colocation

Exchanges also control the amount and degree of granularity of the information you receive (e.g., you can use the consolidated/public feed at a low cost or pay a relatively much larger cost for direct/proprietary feeds from the exchanges). They also monetise the need for speed by renting out computer/server space next to their matching engines, a process called **colocation**. Through colocation, exchanges can provide uniform service to trading clients at competitive rates. Having the traders' trading engines at a common location owned by the exchange simplifies the exchange's ability to provide uniform service as it can control the hardware connecting each client to the trading engine, the cable (so all have the same cable of the same length), and the network. This ensures that all traders in colocation have the same fast access, and are not disadvantaged (at least in terms of exchange-provided hardware).

1.3.4 Extended Order Types

- **Day Orders**: orders for trading during regular trading with options to extend to pre- or post-market sessions;
- **Non-routable**: there are a number of orders that by choice or design avoid the default re-routing to other exchanges, such as 'book only', 'post only', 'midpoint peg', ...;
- **Pegged, Hide-not-Slide**: orders that move with the midpoint or the national best price;
- **Hidden**: orders that do not display their quantity;

- **Iceberg**: orders that partially display their quantity (some have options so that the visible portion will automatically be replenished when it is depleted by less than one round lot);
- **Immediate-or-Cancel**: orders that execute as much as possible at the best price and the rest are cancelled (such orders are not re-routed to another exchange nor do they walk the book);
- **Fill-or-Kill**: orders sent to be executed at the best price in their entirety or not at all;
- **Good-Till-Time**: orders with a fixed lifetime built into them so that they will be cancelled if not executed by its expiration time;
- **Discretionary**: orders display one price (the limit price) but may be executed at more aggressive (hidden) prices;

1.3.5 Exchange Fees

1.4 The Limit Order Book

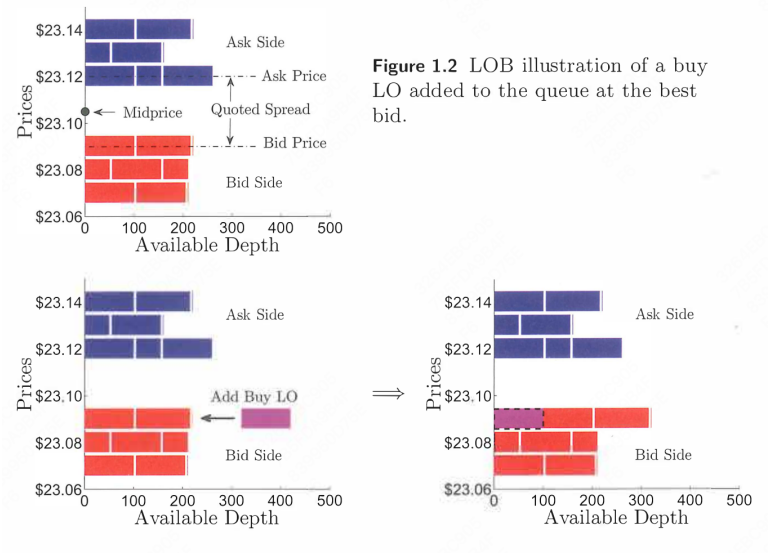

Figure 1.2 LOB illustration of a buy LO added to the queue at the best bid.

Addition of LO to LOB. In Figure 1.4, LOs are displayed as blocks of length equal to their quantities. LOs are ordered in terms of time priority from right to left, so that when a new buy LO comes in at \$23.09 (the purple block) it will be added to the line of blocks already at that price. This new

LO joins the queue at the point closest to the y-axis, becoming the third LO waiting to be executed at $23.09.

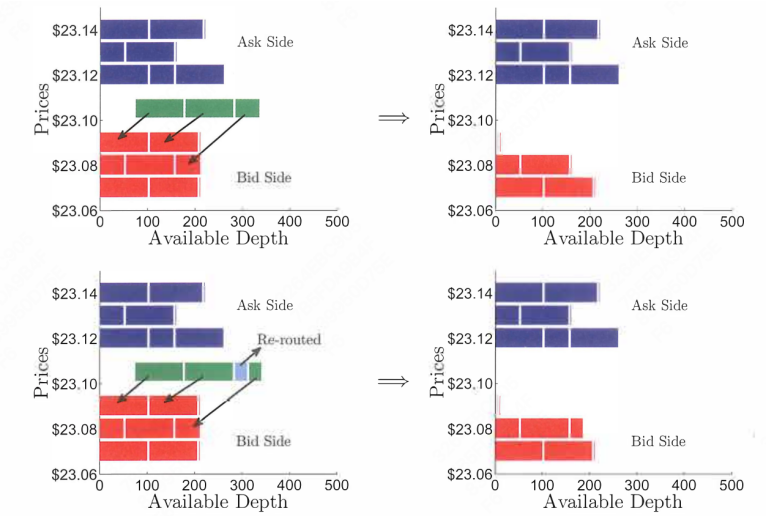

Figure 1.3 LOB illustration of a sell MO walking the LOB with and without re-routing.

MO walks the LOB or is re-routed. Suppose we are looking at the venue with the LOB depicted at the top of Figure 1.4. Assume that this venue's best bid is the best buy quote that the market, across all venues, currently displays. A new MO (to sell) 250 shares enters this market as depicted by the sum of the green blocks in the top panel of Figure 1.4. The matching engine goes through the LOB, matching existing (posted) LOs (to buy on the bid side) with the entering MO following the rules in the matching algorithm. In the LOB there are two LOs at the best bid $23.09, represented by the two red blocks, both for 100 units, totalling 200 units. These 200 units are executed at the best bid.

What happens to the final 50 units depends on the order type and the market it is operating in. In a standard market, the remaining 50 units will be executed against the LOs standing at $23.08 ordered in terms of time-priority (the MO will 'walk the book'). This is captured by the top panels in Figure 1.4: the left panel shows that the MO coming in is split into three blocks, the first two are matched with LOs at $23.09 and the last with the LOs at $23.08. After the MO is fully executed the remaining LOB is shown in the top right panel of Figure 1.4.

In the US, there are order protection rules to ensure MOs get the best possible execution, and which (depending on the order type) may require

the exchange to re-route the remaining 50 units to another exchange that is also displaying a best bid price of $23.09. In this case, as shown in the bottom left panel of Figure 1.4, part of the remaining 50 units (the light blue block) is re-routed to another venue(s) with liquidity posted at $23.09. Only once all liquidity at $23.09 in all exchanges is exhausted, can the remaining shares of the MO return and be executed in this venue against any LO resting at (the worse price of) $23.08. In this example, 25 units were re-routed to alternate exchanges, and 25 units returned to this venue and walked the book.

The MO could in principle be an Immediate-or-Cancel (IOC) order, which specifies that the remaining 50 shares that cannot be executed at the best bid should be cancelled entirely.

Because of these order protection rules (trade-through rules - there is no such rule in European markets), you will very seldom observe in the US an MO walking the book straight away. Rather, you may see a large MO being chopped up and executed sequentially in several markets in a very short span of time. This also implies that as depth disappears (as during the Flash Crash of May 6th, 2010) an MO at the end of a sequence of other orders may be executed against very poor prices, and, in the worst circumstances it may be matched with **stub quotes** - LOs at prices so ridiculous that clearly indicate they are not expected to be executed (such trades were observed during the Flash Crash in the following assets: JKE, RSP, Excelon, Accenture, amongst others). Thus, the LOB serves to keep track of LOs and apply the algorithm that matches incoming orders to existing LOs.

The LOB is defined on a fixed discrete grid of prices (the price levels). The size of the step (the difference between one price level and the next) is called the **tick**, and in the US the minimum tick size is 1 cent for all stocks with a price above one dollar. In other markets several different tick sizes coexist.

Figure 1.3.1 shows a sample plot of the limit order book (LOB) on NASDAQ after the 10,000th event of the day for two stocks, FARO and HP Q, on Oct 1, 2013. In blue you find the sell LOs -traders willing to wait to be able to sell at a high price. The best sell price, the ask, is $21.16, while the best buy price, the bid, is $21.15. The difference between the ask and the bid price, the **quoted spread** is

$$\text{Quoted Spread}_t = P_t^a - P_t^b$$

where P_t^b and P_t^a are the best bid and ask prices, which in this case, is one cent - the minimum quoted spread. However, some times the bid is equal

to the ask and the spread is zero. In that case, the market becomes **locked**, but if this happens, it tends not to last long - although for some very liquid assets it is becoming an increasingly more frequent event. Another common object used when describing the LOB is the **midprice**. The midprice is the arithmetic average of the bid and the ask:

$$\text{Midprice}_t = \frac{1}{2}(P_t^a + P_t^b)$$

As pointed out earlier, the two LOBs shown in Figure 1.3.1 are quite different. The one in the left panel corresponds to HPQ, a frequently traded and liquid asset. HPQ's LOB has LOs posted at every tick out to (at least) 20 ticks away from the midprice and the spread is the minimum spread of 1 tick. In the right panel, we have FARO's LOB. FARO is a seldom traded, illiquid asset. This asset has thinly posted bids and offers and irregular gaps in the LOB. The spread is 20 ticks (20 cents) on a (approximately) $41 priced asset. The difference in liquidity between these assets is also noticeable from the time at which the 10,000th event of the day takes place for these assets. For HPQ, the 10,000th event corresponds to a timestamp of about 9:42 a.m. (less than 15 minutes after the market opened), while for FARO the 10,000th event did not occur until about 12:04 p.m. (more than two and a half hours after market open). Also note that there are less than 100 units posted if we sum together the depth at the best two price levels on the bid and ask for FARO, while for HPQ there are more than 1,000 shares offered in those first two levels of the LOB - HPQ thus has much greater depth. If one takes into account that FARO trades at a price which is twice as high as that of HPQ, the depth in terms of dollar value of shares posted at those prices is also much greater for HPQ.

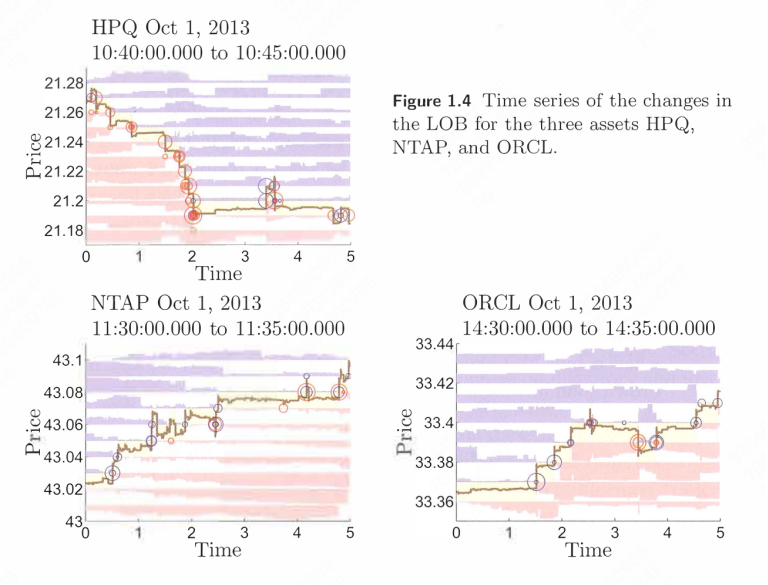

Figure 1.4 Time series of the changes in the LOB for the three assets HPQ, NTAP, and ORCL.

In Figure 1.4, we show how the LOB evolves through time for different stocks HPQ, NTAP and ORCL. On the x-axis is time in minutes, and on the y-axis are prices in dollars. The static picture we saw in Figure 1.3.1 is captured by the shaded blue and red regions - the blue regions on top represent the ask side of the LOB, the posted sell volume, while the bid side is below in red, showing the posted buy volume. The best prices, the bid and ask are identified by the edges of the intermediate light shaded beige region, which identifies the bid-ask spread. Volume at each price level, which was captured in Figure 1.3.1 by horizontal bars, is now illustrated by the size of the shaded region just above/below each price level, although the height of these regions is no longer linear, but a monotonic non-linear transformation that is visually more illustrative.

In addition, Figure 1.4 identifies when incoming orders were executed. The red/blue circles indicate the time, price and size (indicated by the size of the circle) of an aggressive MO which is executed against the LOs sitting in the LOB. When a sell MO executes against a buy LO, it is said to **hit the bid**; analogously, when a buy MO executes against a sell LO, it is said the **lift the offer**. The brown solid line depicts a variation of the asset known as the **microprice** defined as

$$\text{Microprice}_t = \frac{V_t^b}{V_t^b + V_t^a} P_t^a + \frac{V_t^a}{V_t^b + V_t^a} P_t^b$$

where V_t^b and V_t^a are the volumes posted at the best bid and ask, and P_t^b

and $|(P_t^a)$ are the bid and ask prices.

The microprice is used as a more subtle proxy for the asset's transaction cost-free price, as it measures the tendency that the price has to move either towards the bid or ask side as captured by number of shares posted, and hence indicates the buy (sell) pressure in the market. If there are a lot of buyers (sellers), then the microprice is pushed toward the best ask/bid price to reflect the likelihood that prices are going to increase (decrease).

2 A Primer on the Microstructure of Financial Markets

2.1 Market Making

2.1.1 Grossman-Miller Market Making Model

The first issue faced by an MM when providing liquidity is that by accepting one side of a trade (say buying from someone who wants to sell), the MM will hold an asset for an uncertain period of time, the time it takes for another person to come to the market with a matching demand for liquidity (wanting to buy the asset the MM bought in the previous trade). During that time, the MM is exposed to the risk that the price moves against her (in our example, as she bought the asset, she is exposed to a price decline and hence having to sell the asset at a loss in the next trade).

Grossman & Miller (1988) provide a model that captures this problem and describes how MMs obtain a liquidity premium from liquidity traders that exactly compensates MMs for the price risk of holding an inventory of the asset until they can unload it later to another liquidity trader.

Let us consider a simplified version of their model, with a finite number, n, of identical MMs for some given asset and three dates $t \in \{1, 2, 3\}$. To simplify the situation, there is no uncertainty about the arrival of matching orders: if at date $t = 1$ a liquidity trader, denoted by LT1, comes to the market to sell i units of the asset, there will be (for sure) another liquidity trader (LT2) who will arrive at the market to purchase i units (or more generally, to trade $-i$ units, so that LT1's trade (of i units) could be negative or positive (LT1 could be buying or selling). However, LT2 does not arrive to the market until $t = 2$. Let all agents start with an initial cash amount equal to W_0 , MMs hold no assets, LT1 holds i units and LT2 $-i$ units.

There are no trading costs or direct costs for holding inventory. The focus is on price changes: the asset will have a cash value at $t = 3$ of $S_3 = \mu + \epsilon_2 + \epsilon_3$, where μ is constant, ϵ_2 and ϵ_3 are independent, normally distributed random variables with mean zero and variance σ^2 . These will be publicly

announced between dates $t-1$ and t, that is ϵ_3 is announced between $t=2$ and $t=3$, and $\epsilon_2$is announced between $t=1$ and $t=2$. Hence, the realised cash value of the asset can increase or decrease (ignore the fact that there are realisations of ϵ_2 and ϵ_3 that could make the asset value negative - the model serves to illustrate a point). Because the shocks to the value of the asset are on average zero a risk-neutral trader has no cost at all from holding the asset. The model becomes interesting if we assume that all traders, MMs and liquidity traders, are risk-averse. To be more specific, suppose they have the following expected utility for the future random cash value of the asset (X_3): $\mathbb{E}[U(X_3)]$ where $U(X) = -\exp(-\gamma X)$, and where $\gamma > 0$ is a parameter capturing the utility penalty for taking risks

Solving the model backwards we obtain a description of trading behaviour and prices. At $t=3$ the cash value of the asset is realised, $S_3 = \mu + \epsilon_2 + \epsilon_3$. At $t=2$, the n MMs and LT1 come into the period with asset holdings q_1^{MM} and q_1^{LT1} respectively. LT2 comes in with $-i$ and they all exit with asset holdings q_2^j, where $j \in \{MM, LT1, LT2\}$. Note that if, for example, $q_t^j = 2$ this denotes the that agent j is holding 2 units when exiting date t, so that the agent will be long (that is, has an iventory of) two units.

Given the problem as described so far, at $t=2$ agent j chooses q_2^j to maximise his expected utility knowing the realisation of ϵ_2 that was made public before $t=2$:

$$\max_{q_2^j} \mathbb{E}\left[U(X_3^j)|\epsilon_2\right]$$

subject to

$$X_3^j = X_2^j + q_2^j S_3$$
$$X_2^j + q_2^j S_2 = X_1^j + q_1^j S_2$$

These two contraints capture:

1. the fact that the cash value of the agent's assets at $t=3$, X_3, is equal to the agent's cash holdings at $t=3$, which is equal to X_2 plus the cash value of the agent's asset inventory q_2^j
2. the fact that the cash value of the agent's assets when exiting date $t=2$ was equal to the cash value of the agent's assets when entering date $t=2$

Given the normality assumption and the properties of the expected utility function it is straightforward to show that

$$\mathbb{E}\left[U(X_3^j)|\epsilon_2\right] = -\exp\left\{-\gamma\left(X_2^j + q_2^j\mathbb{E}[S_3|\epsilon_2]\right) + \frac{1}{2}\gamma^2\left(q_2^j\right)^2\sigma^2\right\}$$

Since ϵ_2 and ϵ_3 are independent, $\epsilon_3|\epsilon_2 \sim N(0,\sigma^2)$ Therefore $S|\epsilon_2 = y = \mu + y + \epsilon_3|\epsilon_2 = y \sim N(\mu+y,\sigma^2)$ Then $S|\epsilon_2 \sim N(\mu+\epsilon_3,\sigma^2)$ Let $\mu = \mathbb{E}[S_3|\epsilon_2]$. Then $S|\epsilon \sim N(\mu,\sigma^2)$.

Assume $X \sim N(\mu,\sigma^2)$. The **Moment generating function** (**MGF**) is

$$M_X(t) = \mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} dx$$

As

$$\begin{aligned} tx - \frac{(x-\mu)^2}{2\sigma^2} &= \frac{2\sigma^2 tx - (x^2 - 2\mu x + \mu^2)}{2\sigma^2} \\ &= -\frac{x^2 - 2\mu x - 2\sigma^2 tx + \mu^2}{2\sigma^2} \\ &= -\frac{x^2 - 2x(\mu+\sigma^2 t) + \mu^2}{2\sigma^2} \\ &= -\frac{(x-(\mu+\sigma^2 t))^2 - 2\mu\sigma^2 t - \sigma^4 t^2}{2\sigma^2} \end{aligned}$$

We have

$$M_x(t) = e^{ut + \frac{\sigma^2 t^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x-(\mu+\sigma^2 t)^2}{2\sigma^2}}$$

The integral is the PDF (probability density function) of a $N(\mu+\sigma^2 t, \sigma^2)$ distribution, which integrates to 1. Thus

$$M_X(t) = \exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}$$

Now

$$\begin{aligned} \mathbb{E}[U(X_3^j) \mid \epsilon_2] &= \mathbb{E}[-\exp\left\{-\gamma_(X_2^j + q_2^j S_3)\right\} \mid \epsilon_2] \\ &= -\exp\{-\gamma X_2^j\} \cdot \mathbb{E}[\exp\{-\gamma q_2^j S_3\} \mid \epsilon_2] \\ &= -\exp\{-\gamma X_2^j\} \cdot \exp\left\{-\gamma q_2^j \mu + \frac{1}{2}\gamma^2 (q_2^j)^2 \sigma^2\right\} \end{aligned}$$

where $\mu = \mathbb{E}[S_3 \mid \epsilon_2]$

Thus, the problem is concave and the solution is characterized by

$$q_2^{j,*} = \frac{\mathbb{E}[S_3 \mid \epsilon_2] - S_2}{\gamma\sigma^2}$$

for all agents: the n MMs, LT1, and LT2

As at date $t = 2$ demand and supply for the asset have to be equal to each other, we can solve for the equilibrium price S_2:

$$nq_1^{MM} + q_1^{LT_1} + q_1^{LT2} = nq_2^{MM} + q_2^{LT1} + q_2^{LT2} \tag{1}$$

As we have established above, all q_2^j are equal, so that the right-hand side of the above equation is equal to

$$nq_2^{MM} + q_2^{LT1} + q_2^{LT2} = (n+2)\frac{\mathbb{E}[S_3 \mid \epsilon_2] - S_2}{\gamma\sigma^2} \tag{2}$$

We also know that at date 1 the total quantity of the asset available was equal to the quantity of assets LT1 brought to the market, so that LHS of (1) is

$$nq_1^{MM} + q_1^{LT1} + q_1^{LT2} = i - i = 0$$

Hence, we obtain that in equilibrium, at date $t = 2$, $S_2 = \mathbb{E}[S_3] = \mu + \epsilon_2 + \mathbb{E}[\epsilon_3] = \mu + \epsilon_2$, and therefore $q_2^j = 0$.

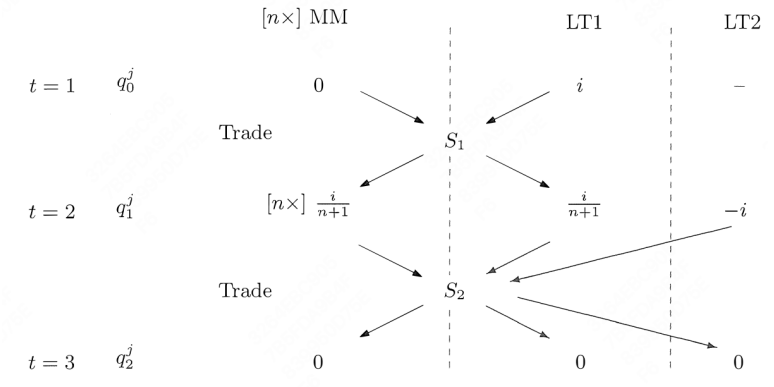

Figure 1: Trading and price setting in the Grossman-Miller model

This makes sense, as at $t = 2$ there are no asset imbalances, the price of the asset reflects its 'fundamental value' and no one will want to hold a non-zero amount of the risky asset. This analysis is captured in the bottom half of Figure 1, where we see the asset holdings of the three types of participants as they enter $t = 2$, q_1^j, $j \in \{MM, LT1, LT2\}$, and how after trading at a price equal to S_2 they end up with holdings, q_2^j, equal to zero

Consider now what happens at date $t = 1$. Participating agents (the n MMs and LT1) anticipate that whatever they do, the future market price

will be efficient and they will end up exiting date $t = 2$ with no inventories, so that $X_2 = X_3$. Thus, their portfolio decision is given by

$$\max_q \mathbb{E}\left[U(X_2^j)\right]$$

subject to

$$X_2^j = X_1^j + q_1^j S_1$$
$$X_1^j + q_1^j S_1 = X_0^j + q_0^j S_1$$

Repeating the analysis, we obtain that the optimal portfolio solution is

$$q_1^{j,*} = \frac{\mathbb{E}[S_2] - S_1}{\gamma\sigma^2}$$

for all agents, j, that are present: the n MMs and LT1. Also at date $t = 1$ demand and supply for the asset have to be equal to each other, so that

$$nq_0^{MM} + q_0^{LT1} = nq_1^{MM} + q_1^{LT1}$$

where $q_0^{LT1} = i$. Therefore

$$i = (n+1)\frac{\mu - S_1}{\gamma\sigma^2} \iff S_1 = \mu - \gamma\sigma^2\frac{i}{n+1}$$

With this expression we can interpret how the market reaches a solution for LT1's liquidity needs: LT1, a trader who wants to sell a total of $i > 0$ units at $t = 1$,finds that there is no one currently in the market with a balancing liquidity need. There are traders in the market, but they will not accept trading at the efficient price of μ because if they do, they will be taking on risky shares (they are exposed to the price risk from the realisation of ϵ_2) without compensation.

But, if they receive adequate compensation (which we call a liquidity discount, as for $i > 0$,$S_1 < \mathbb{E}[S_2] = \mu$),the n MMs will accept the LT1's shares. However, LT1 is price-sensitive, so if he has to accept a discount on the shares,he will not sell all the i shares at once. In equilibrium, both the n MMs and LT1 end up holding $q_1^{j,*} = \frac{i}{n+1}$ units of the asset each,that is LT1 sells $\frac{n}{n+1}i$ units and holds on to $\frac{i}{n+1}$ units to be sold later. Trading occurs at a price below the efficient price, $S_1 = \mu - \gamma\sigma^2\frac{i}{n+1}$. The difference between the trading price and the efficient price,namely $|S_1 - \mu| = \left|\gamma\sigma^2\frac{i}{n+1}\right|$,represents the (liquidity) discount the MMs receive in order to hold LT1's shares. This

size of the discount is influenced by the variables in the model: the size of the liquidity demand ($|i|$), the amount of competition amongst MMs (captured by n), the market's risk aversion (γ), and the risk/volatility of the underlying asset (σ^2). These variables all affect the discount in an intuitive way: the size of the liquidity shock, risk aversion, and volatility all increase the discount,while competition reduces it. This occurs when LT1 wants to sell,i.e. $i > 0$. If LT1 wanted to buy, $i < 0$,then the solution would be the same except that instead of a discount, the MMs would receive a premium equal to $|S_1 - \mu|$ per share when selling to LT1.

From this analysis we can also see that as competition n increases, the liquidity premium goes to zero, the price converges to the efficient level, $S_1 = \mu$ and LT1's optimal initial net trade, $q_1^{LT1,*} - q_0^{LT1}$, converges to his liquidity need i.

2.1.2 Trading Costs

We have seen how the Grossman & Miller framework helps to understand how the cost of holding assets (in this case, via the uncertainty it generates to the risk-averse MMs) affects liquidity via the cost of trading ($|S_1 - \mu|$) and the demand for immediacy (as at $t = 1$ LT1 only executes $\frac{n}{n+1}i$ rather than her desired i). Also,competition between MMs is crucial in determining these trading costs. But what drives n?

Grossman & Miller link competition, n, with participation costs. They do this by introducing an earlier stage to the model in which potential MMs decide whether they want to actively participate in the market and provide liquidity or prefer to do something else. The decision is determined as a function of a participation cost parameter c which proxies for the time and investments needed to keep a constant, active and competitive presence in the market, as well as the opportunity cost the MM gives up by being in the market and not doing something else. The conclusion, which can be obtained without going into the details of the analysis, is that the level of competition decreases monotonically with supplier's participation costs. Thus, participation costs, proxied by the cost parameter c, increase the size of the liquidity premium (via its effect on competition, n).

The parameter c captures the fixed costs of participating in the market, but we could also consider introducing into the model a cost of trading that depends on the level of activity in the market. In particular, we introduce trading costs that depend on the quantity traded, like actual exchange trading fees. Exchange trading fees are usually proportional to dollar-volume but here, for simplicity, we use fees proportional to number of shares traded

parameterised by η. Given that fees are known, these fees act like a participation cost for liquidity traders.

The first effect of having $\eta > 0$ is that liquidity traders with a desired trade $|i|$ that is small relative to trading fees, will find trading too expensive and refrain from trading

For sufficiently large desired trades (so that trading is preferred to not trading by all participants) the model gives us the following solution. Suppose every trader pays η per share regardless of whether they are buying or selling the asset. To simplify, assume that any remaining inventories after $t = 2$ are liquidated at $t = 3$. Also, assume LT1 wants to sell $|i|$ units ($i > 0$), and LT2 wants to buy the same amount.

Now

$$X_3^j = X_2^j + q_2(S_3 - \eta)$$

At $t = 2$, since the MMs and LT1 enter the period with a positive inventory (and will be wanting to sell now or at $t = 3$) their optimal final period holdings are

$$q_2^j = \frac{\mathbb{E}[S_3 - \eta \mid \epsilon_2] - (S_2 - \eta)}{\gamma\sigma^2}, \quad j \in \{MM, LT1\}$$

while the demand for shares by LT2 is

$$q_2^{LT2} = \frac{\mathbb{E}[S_3 + \eta \mid \epsilon_2] - (S_2 + \eta)}{\gamma\sigma^2}$$

As everyone anticipates that their trading positions need to be liquidated anyway, the trading fees do not affect the price at $t = 2$, and we obtain $S_2 = \mathbb{E}[S_3 \mid \epsilon_2] = \mu + \epsilon_2$