# Mathematical Market Microstructure – An Optimization Approach

## Lecture II – Practical aspects of trading algorithm optimization and an overview of market microstructure theory

*FINM 37601 – Fall, 2023*

By Hongsong Chou, Ph.D.

# AGENDA

- Practical aspects of trading algorithm optimization – using VWAP as an example

- A review of “classic” market microstructure theories

# AGENDA

- Practical aspects of trading algorithm optimization – using VWAP as an example

- A review of “classic” market microstructure theories

# WHY DO TRADERS USE ALGOS? (I)

- Meet or beat certain benchmarks:
  - VWAP/TWAP;
  - Implementation Shortfall;
  - TargetClose or MOC;
- Capture liquidity when it appears;
  - WithVolume (or POV);
  - WorkAndPounce;
  - RelativeStep against certain price levels;
  - ActiveIS that uses VWAP slippage as a gauge;
- Automation of certain strategies;
  - Execution algorithms for pair strategies;
  - SmartOrderRouter based on order size (%ADV);
  - Etc.

# WHY DO TRADERS USE ALGOS? (II)

- Hide their hands:
  - Instead of sending a large order, slice it to different pieces;
  - Randomization of order submission time and size;
- Minimizing market impact:
  - Small-size orders generate smaller impact than large-size orders;
  - Market impact function is concave;
- Accessing different venues:
  - Execution algorithms for pair strategies;
  - SmartOrderRouter based on order size (%ADV);
  - Etc.

# WHAT EXEC ALGOS ARE OUT THERE?

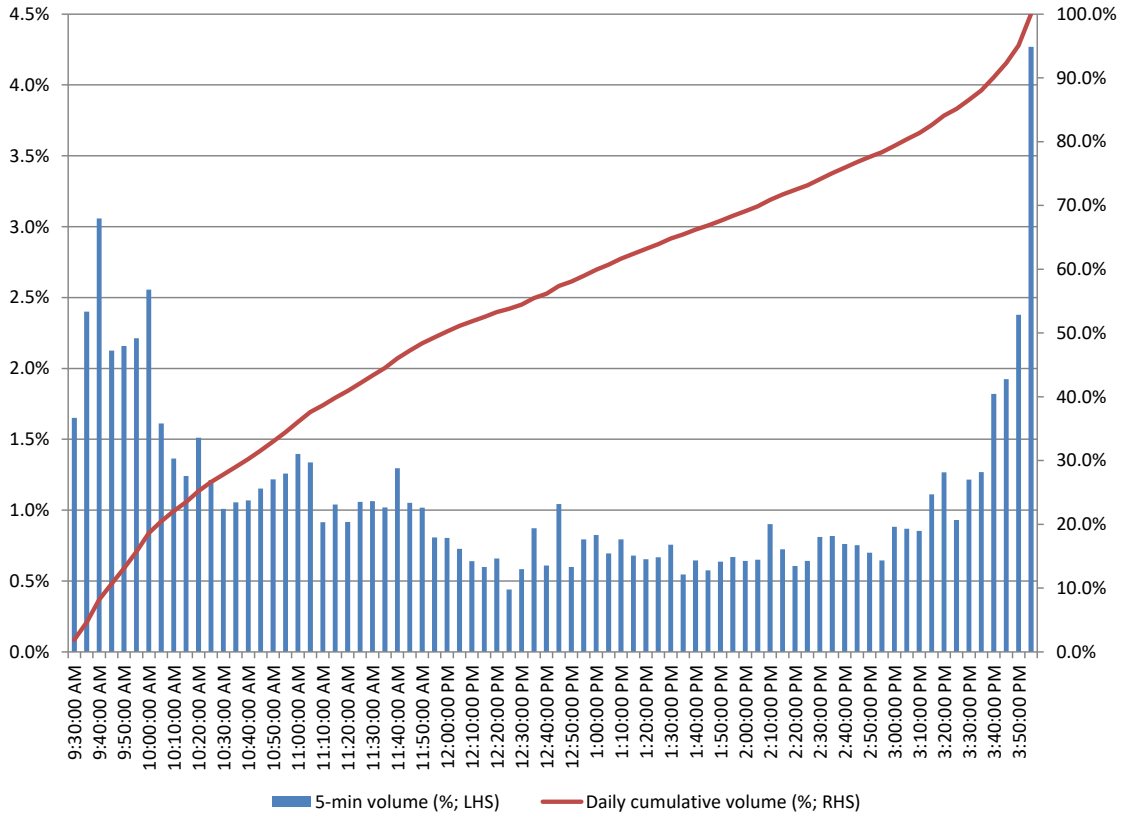
- Schedule-based but relatively static:
  - VWAP, TWAP;
  - WithVolume, IS;
  - TargetOpen and TargetClose;
- Schedule-based but relatively adaptive:
  - Adaptive IS;
  - Conditional algos that switch between static algos based on price/volume/etc. signals;
- Liquidity-driven:
  - Relative to price and/or volume (WorkAndPounce; WaitAndPounce, etc.);
  - Liquidity aggregators among various venues;
- Specific-purpose algos:
  - Pairs execution algos;
  - Portfolio IS;
  - SOR;
  - Etc.

# HOW TO BUILD A VWAP ALGO?

- VWAP algo is still a fairly sophisticated algo that can incorporate all basic elements in algo design and development;
- In theory, VWAP is the algo that minimizes market impact, not execution risk;
- There are five levels of considerations that an algo designer has to take into account when developing/enhancing a VWAP algo:
  - **Dynamic tranche time** based on both intraday seasonality and real-time market activities;
  - **Medium- and long-term alpha signals** that help gauge general trend of mid-quote movements;
  - **Volume skew** based on medium- and long-term alpha signals to capture favorable price points within an execution window;
  - **Intraday volume and price forecast and control** that help avoid chasing the wrong volume/price movements;
  - Order placement that utilizes **market making tactics** in bid-ask spread capturing.

# THE VWAP SCHEDULE (I)

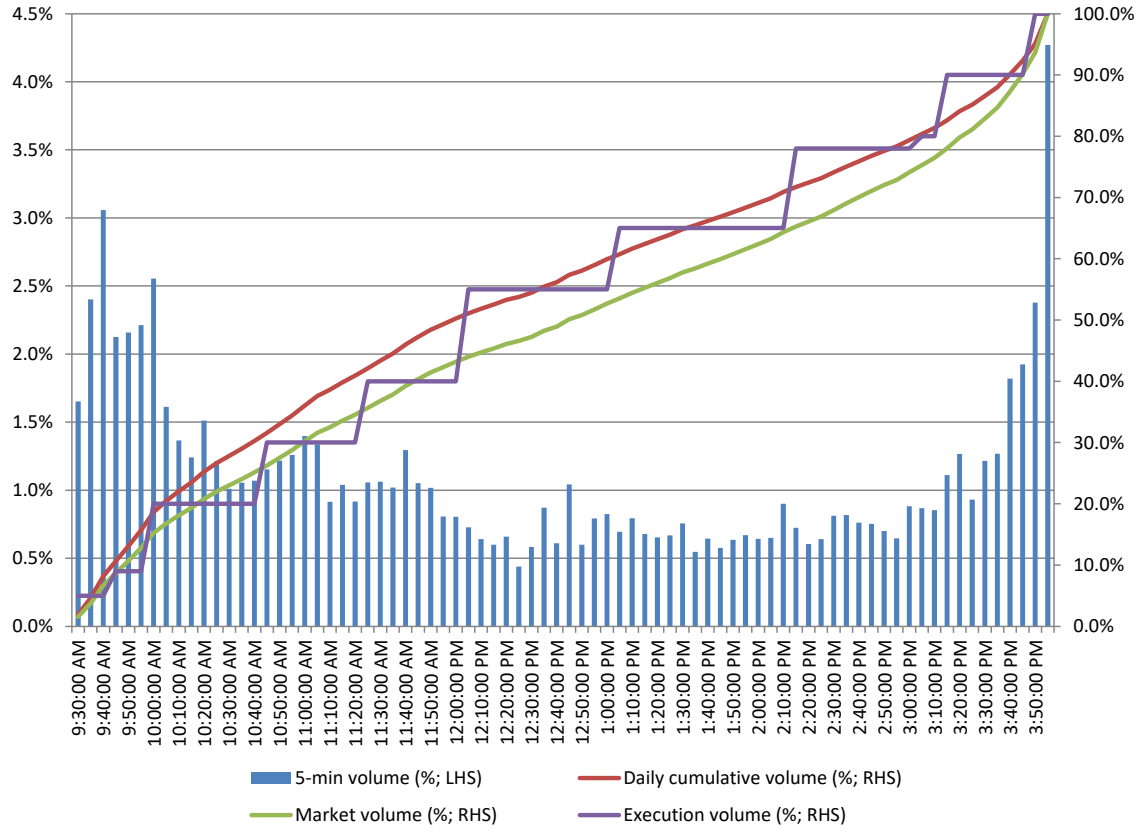
- To minimize market impact (not execution risk), a VWAP needs to participate according to the intraday volume profile:




(IBM UN intraday volume as of April 1, 2013; no open and close auction prints considered)

# THE VWAP SCHEDULE (II)

- In reality, for reasons such as capturing real-time volume burst as well as the discrete nature of executions, execution profile often zigzags the "model profile", which can deviates from the realized "market profile":




(IBM UN intraday volume as of April 1, 2013; no open and close auction prints considered)

# THE IMPORTANCE OF VOLUME PROFILE (I)

- The performance of a VWAP algo is often measured as a difference between the execution price of the algo's trade and a market VWAP:

  VWAP slippage = (VWAP of the same stock in the market – VWAP of the algo's executions) * signOfTrade; (1)

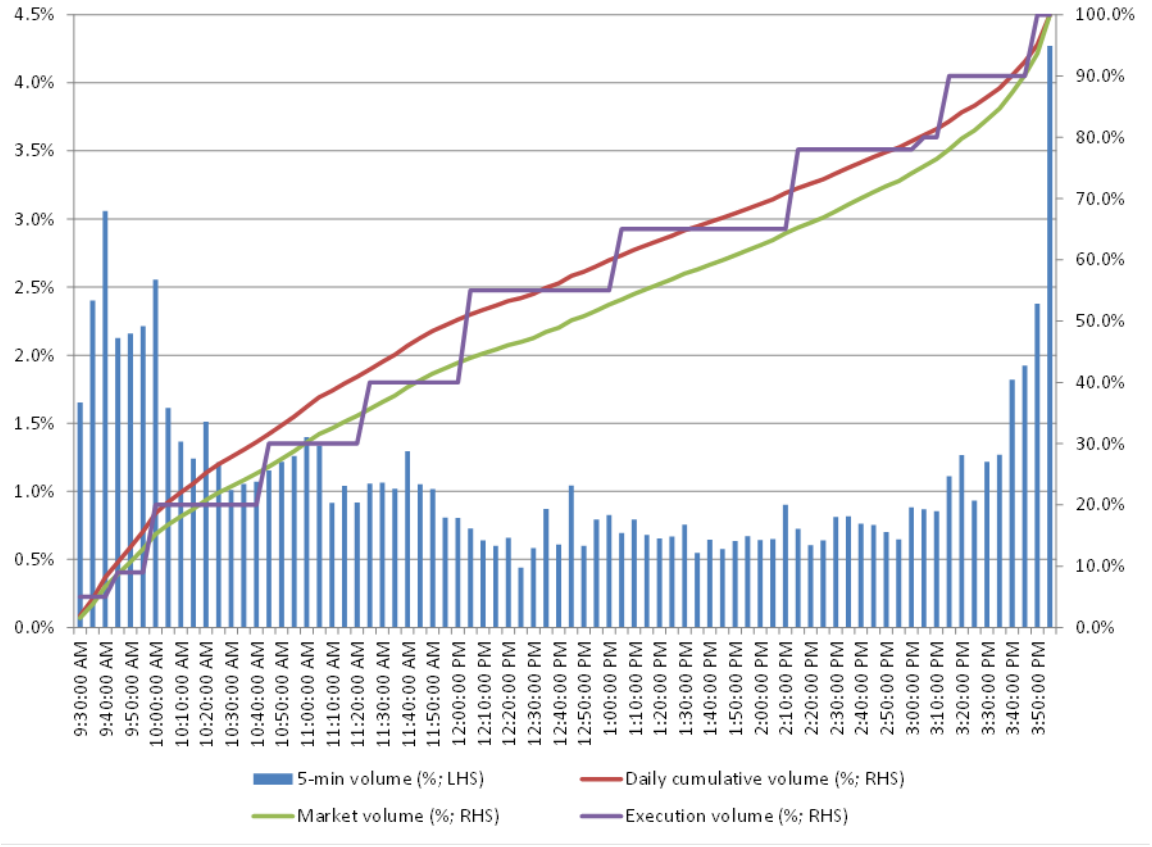
- It is often expressed in basis points by dividing the above slippage by the market VWAP, then multiplied by 10000;
- If we use $i$ to index individual "slices" of a VWAP trade ($i$ = 1, 2, …), $P_i$ to index market's prices "bucketed around" the $i$th slice of the algo's execution, $p_i$ the trade price of the $i$th execution of the algo, $v_i$ the percentage of the market's trade volume "bucketed around" the $i$th slice of the algo's execution, $w_i$ the percentage of the trade size of the $i$th execution of the algo among the total execution volume of the algo order, then we can re-write definition (1) above as:

$$VWAP\ Slippage\ (for\ Buy) = \sum_i P_i v_i - \sum_i p_i w_i$$

- Note that $v_i$ and $w_i$ are also called "realized market volume profile" and "realized execution profile".

# THE IMPORTANCE OF VOLUME PROFILE (II)

- Example profiles:




(IBM UN intraday volume as of April 1, 2013; no open and close auction prints considered)

# THE IMPORTANCE OF VOLUME PROFILE (III)

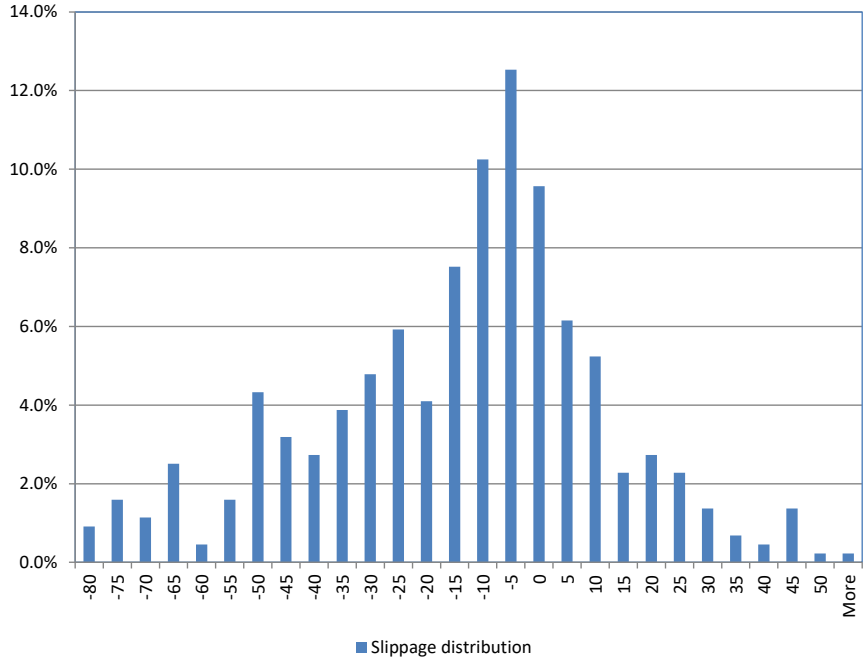
- The performance of a VWAP algo depends on three profiles: model profile, market realized profile and execution profile:

$$VWAP\ Slippage\ (for\ Buy) = \sum_i P_i v_i - \sum_i p_i w_i$$
$$= \sum_i P_i (v_i - m_i) + \sum_i P_i (m_i - w_i) + \sum_i (P_i - p_i) w_i;$$

- In the above equation, the **first term** indicates the slippage component that is due to the difference between realized market volume profile, $v_i$, and the theoretical model profile, $m_i$; the **second term** indicates the slippage component due to the difference between theoretical model profile and realized execution profile; the **third term** indicates the slippage component due to the difference between market trade price and the execution price of this individual VWAP order (at child order level).

# THE IMPORTANCE OF VOLUME PROFILE (IV)

- Another key aspect of VWAP algo performance, the  is also impacted by volume profile:

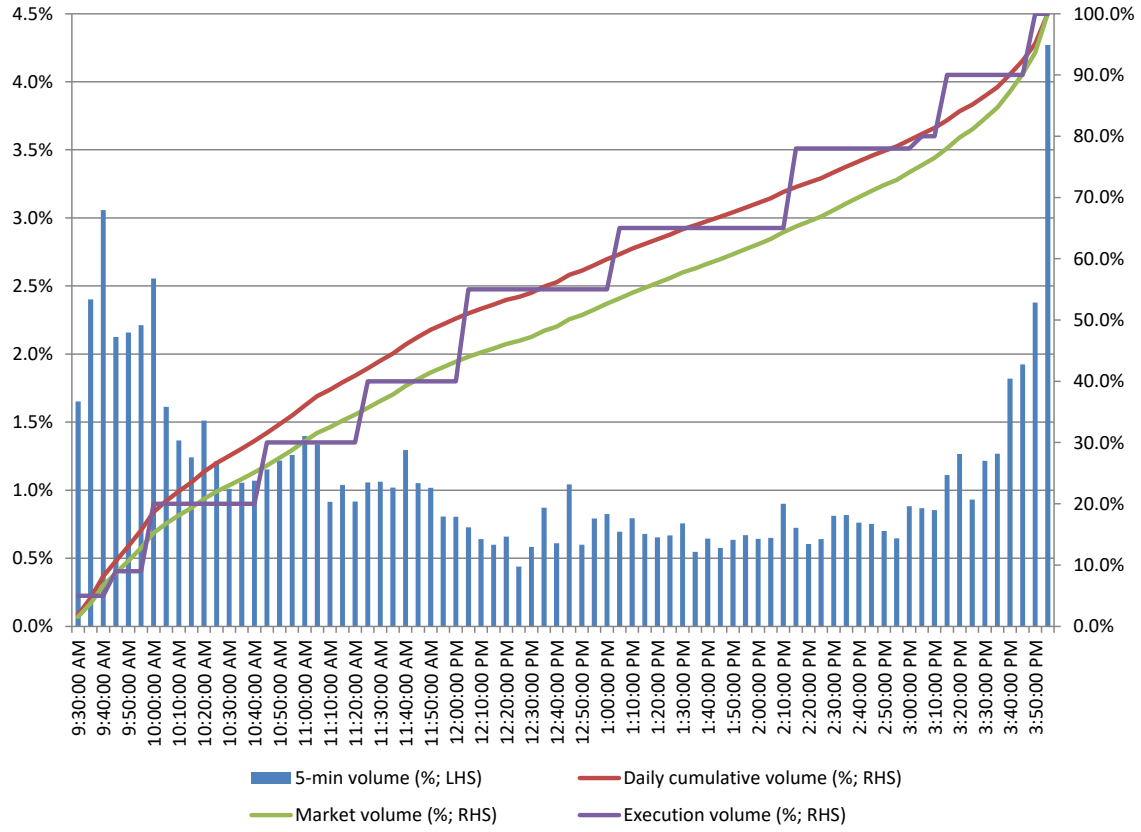



- Note that:

$$Var(Slippage) \approx \sigma^2 \int_0^T Var[v(t) - w(t)]dt$$

# THE IMPORTANCE OF VOLUME PROFILE (V)

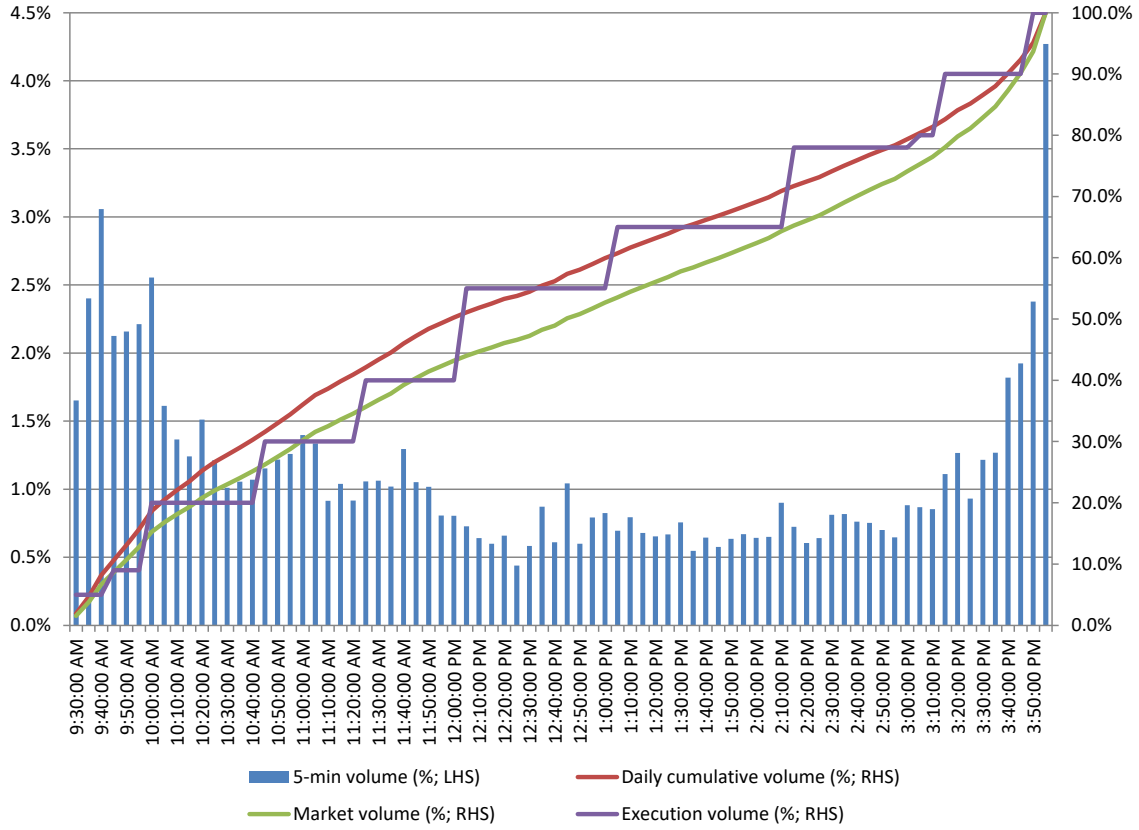
- In reality, for reasons such as capturing real-time volume burst as well as the discrete nature of executions, execution profile often zigzags the "model profile", which can deviates from the realized "market profile":




(IBM UN intraday volume as of April 1, 2013; no open and close auction prints considered)

# DYNAMIC TRANCHE TIME

- A “tranche time” is a time window within which the placement of a child order is contemplated, placed and executed;
- It often starts from the fill time of the previous child order and runs until the current child order is filled;




Note: the “static” tranche time for the case shown on the left is 15 minutes; however, the strategy adjusted this “static” tranche time based on the “bursting” of liquidity in a dynamic fashion.

# MID- TO LONG-TERM ALPHA SIGNALS (I)

- The mid- to long-term alpha signals have a “half-life” time of about 10-20 minutes to half or a whole day;
- Usually, such alpha signals can be generated based on overnight global market movements or news and past several weeks’ or months’ individual stock and/or sector movements;
- The mid- to long-term alpha signals can be used by algos like VWAP to “front-load” or “back-load” against the model profile for today; algos that leverage such “front-load” or “back-load” features will generally have good performance as measured by “profile slippage”.

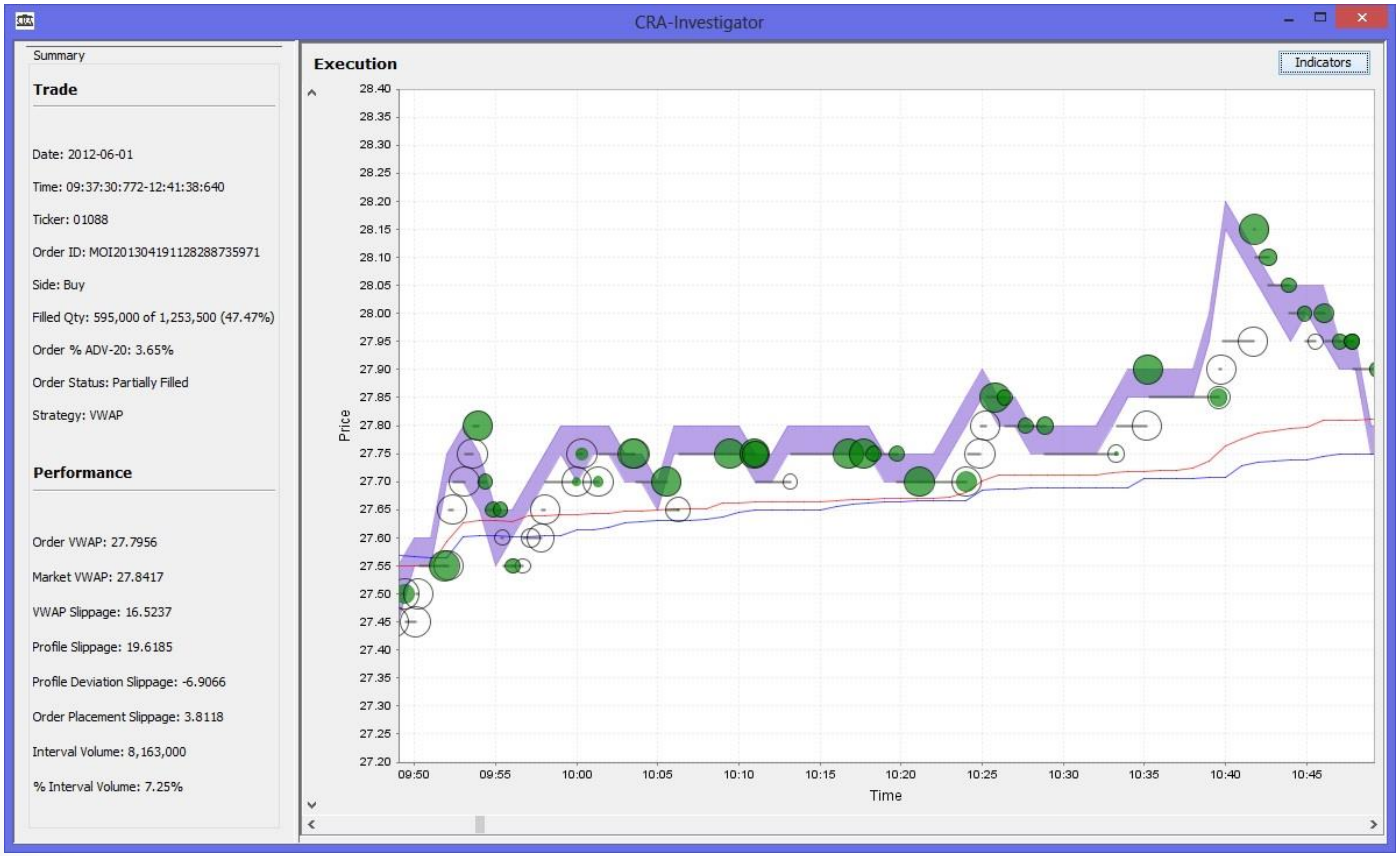
# MID- TO LONG-TERM ALPHA SIGNALS (II)

- Let's look at three trading days in which broad market exhibited different behaviors:


Source: Bloomberg

# MID- TO LONG-TERM ALPHA SIGNALS (III)

- Let's look at a few trades that tried to leverage the upward trend at the beginning of the trading day on June 1, 2012:



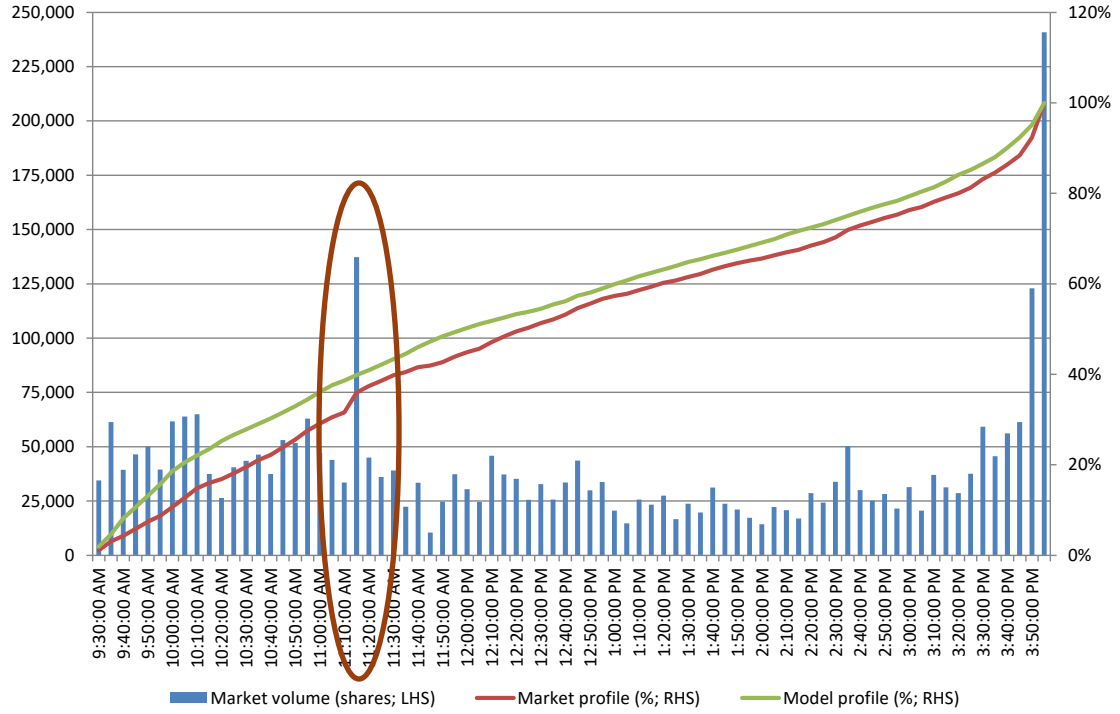

Source: Charles River Advisors Limited

# VOLUME SKEW

- "Volume Skew" is a technique that is strongly related to different behaviors of Buys vs. Sells;
- In general, market participants tend to buy slowly and sell fast, assuming the same absolute amount of price movement (in percentage, say);
- Therefore, within each tranche time, a VWAP algo can adjust the speed of execution based on whether it tries to capture up trend or down trend as well as the "side" of the order.

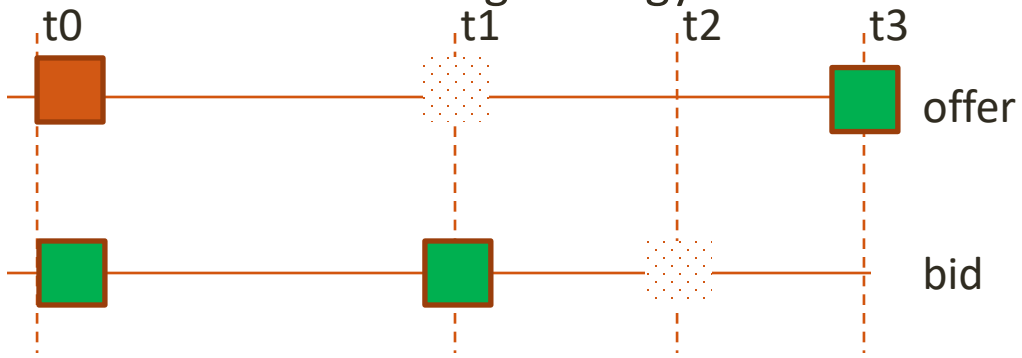
# PRICE/VOLUME SPIKES

- In general, the “price/volume spike signal” is for risk control for not chasing the wrong signal; chasing price/volume trend can lead to significant risk when the trend reverses; therefore, a good algo should be able to avoid these transient movements and stick to the model profile; note that this does not contradict the next level, market making, as they look at different time scales.



# MARKET MAKING IN VWAP ALGOS (I)

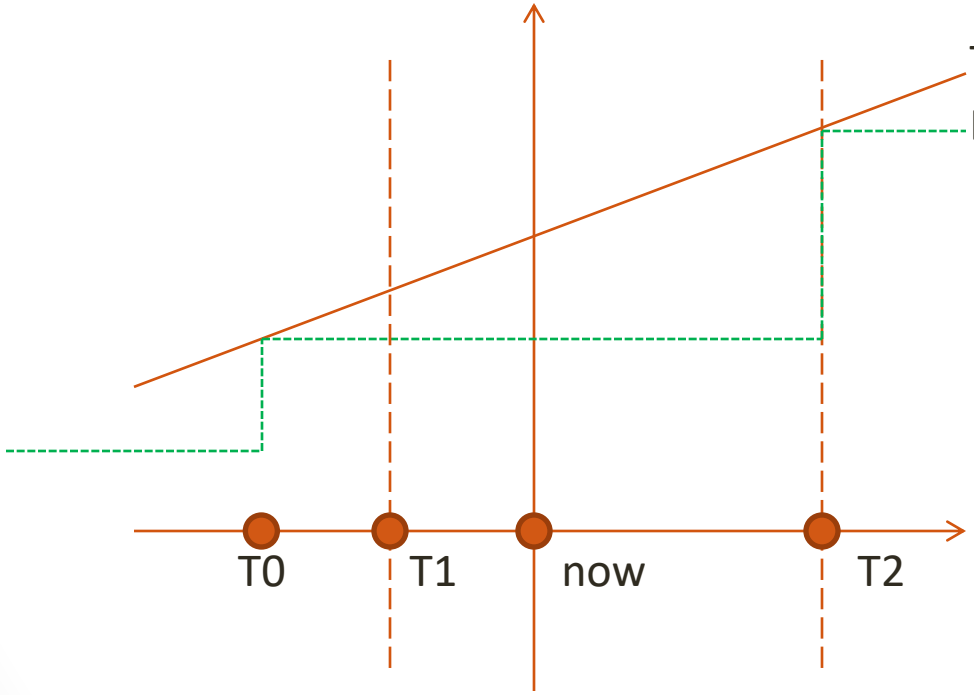
- Basic aspects of a market making strategy:




- Three key things that a market maker needs to consider:
  - P&L in which bid-ask spread plays key role;
  - Adverse selection risk: don't get "picked off" by the other side;
  - Market risk: this is essentially volatility of price.

# MARKET MAKING IN VWAP ALGOS (II)

- A “market making agent” can be deployed at the core of a VWAP algo to leverage short-term alpha signal; what the market maker will do is adjust execution aggressiveness so as to realize short-term price movement trend:




*Note: even though there are pre-calculated T1 and T2, the market maker will adjust T1 and T2 based on short-term alpha signals during real-time executions.*

# AGENDA

- Practical aspects of trading algorithm optimization – using VWAP as an example
- A review of “classic” market microstructure theories

# PRICE FORECASTING FOR MARKET MAKERS (I)

- As discussed in lecture 1, the most important role of an exchange (or any other trading venue) is to determine the price of an instrument; before a price is "realized" via trading on an exchange, it is anybody's "guess";
- "Information trader" is a group of market participants that, at least according to market microstructure theory, has more accurate prior knowledge about the true price of an instrument than anybody else in the market; the one and only goal of an "information trader" is to make money based on her proprietary knowledge of the true price of the instrument; however, whether or not she can make money depends on many factors, such as the accuracy of her guess of the true price, the strategic maneuver of other market participants (i.e., trading-related liquidity fluctuations, etc.);
- "Liquidity (or, noise) trader", on the other hand, participate in trading activities not solely on price information; in fact, they may have just random guess on the true price of an instrument;
- "Market maker" sits in between the information trader and the liquidity trader, who usually does not assume the correct knowledge of the true price of an instrument, but always tries to adjust her quoted bid and offer prices in order to facilitate the price discovery process of an exchange; as a result, it has a special role for the well being of the market; in theory, market maker does not assume the goal of "making money", but she does not want to lose money either.

# PRICE FORECASTING FOR MARKET MAKERS (II)

- The existence of a market maker is often required by an exchange; if the market is composed of only information traders, there can be times that no trade can be formed and no information can be revealed; if the market is composed of only noise traders, no economically beneficial capital allocation can be done;
- A market can achieve an equilibrium when there are only information traders and noise traders (i.e., no market makers); however, such equilibrium can be rather delicate and instable; this is particularly true for instruments that are not liquid;
- A market also can achieve an equilibrium when there are only information traders and market makers (i.e., no noise traders); however, if the information traders always systematically "adverse select" the market makers, such market may eventually breakdown as the market makers will constantly lose money thus the economic incentive to be market makers will disappear completely;
- Therefore, a well-functioning market requires the existence of all three types of market participants: information traders, noise traders and market makers; for market makers, it is of vital importance to maintain a good forecast of price movement so as to protect themselves at the same time of facilitating the price discovery process of an exchange.

# PRICE FORECASTING FOR MARKET MAKERS (III)

- The Roll (1984) model:
  - The market maker will observe past trade prices and quote prices;
  - An estimate of the "true price" of the instrument by the market maker is the mid-quote price;
  - The market maker assumes that the mid-quote of the stock will follow a process that is driven by market-related and stock-specific "public" events:

    $$m_t = m_{t-1} + u_t; \quad (2.1)$$

  - And the forecast of the price by the market maker will be written as:

    $$p_t = m_t + q_t c; \quad (2.2a)$$

    or the price increment process of

    $$\Delta p_t = u_t + (q_t - q_{t-1})c; \quad (2.2b)$$

  - Here, $q_t$ is the trade direction process that is assumed to be either 1 or -1 in the Roll model; $c$ is used here to indicate "transaction cost" of which the best approximation is the half bid-ask spread; in many cases, the market maker has to take into account the inventory that she is carrying at any forecasting moment, which will make $c$ bigger than the half bid-ask spread;
- A few comments:

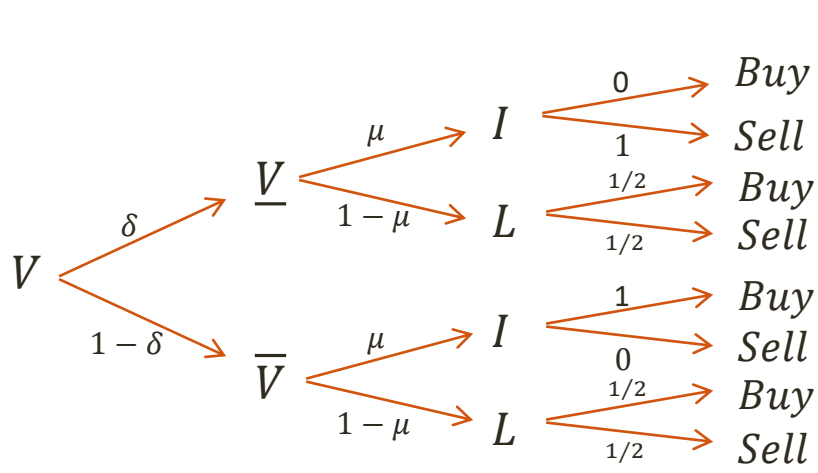
  - The Roll model is called a structural model, as it incorporate trading dynamics and market price discovery mechanism in it;

# PRICE FORECASTING FOR MARKET MAKERS (IV)

- A few comments (continued from previous page):
  - To fully understand the statistical properties of the model, one has to make assumptions about the processes of $u_t$ and $q_t$;
  - What the market maker will care about next is the mean and the standard deviation of the forecast return process:
    - $E[\Delta p_t] = E[u_t + (q_t - q_{t-1})c] = \overline{\Delta p}$; (2.3)
    - $Var[\Delta p_t] = 2c^2 + {\sigma_u}^2$ ; (2.4)
    - $Cov[\Delta p_t, \Delta p_{t-1}] = -c^2$; (2.5)
  - Here we assumed that: $E[u_t] = 0, Cov[q_t q_{t-k}] = 0$ for all $k \neq 0$ and $Cov[q_t u_{t-k}] = 0$ for all $k$;
  - Note that, if we assume that $c$ and $\sigma_u$ are both constant and $\overline{\Delta p} \neq 0$, then we call that $\Delta p_t$ follows a non-zero-mean covariance stationary process;
  - Wold Theorem (1954): Any non-zero-mean covariance stationary process $\{x_t\}$ can be represented in the form of: $x_t = \sum_{j=0}^{\infty} \theta_j \epsilon_{t-j} + \eta_t$, where $\{\epsilon_t\}$ is a zero-mean white noise process and $\{\eta_t\}$ is a linearly deterministic process by the history of $\{x_t\}$; to normalize the MA process for $\{x_t\}$ , one often makes $\theta_0 = 1$;
  - An extension of the Wold Theorem states that a covariance stationary process has zero auto-covariances at all orders higher than $k$ possesses a MA representation of order $k$;
  - In empirical market microstructure studies, one often constructs a structural model based on practical or economical considerations, then uses the Wold Theorem to swap between this structural model and its MA representations to extract statistical properties and conducts forecasting.

# EXTENSIONS OF THE ROLL MODEL (I)

- Limitations of the Roll model (1984)
  - There is no distinction between "information traders" and "noise traders" in its structure;
  - For a market maker who wants to use such a model, there is no way to determine the bid and the offer;
  - It is too simplified to incorporate meaningful (reading: practically real) order flow statistics;
- Sequential Trade Price Models (Glosten and Milgrom (1985); see also Hasbrouck (2007))
  - Consider the following event tree:




| Probability |
|---|
| $0$ |
| $\delta\mu$ |
| $\frac{1}{2}\delta(1-\mu)$ |
| $\frac{1}{2}\delta(1-\mu)$ |
| $(1-\delta)\mu$ |
| $0$ |
| $\frac{1}{2}(1-\delta)(1-\mu)$ |
| $\frac{1}{2}(1-\delta)(1-\mu)$ |

  - Here, $V$ is the observed trade price of the instrument; $\underline{V}$ and $\overline{V}$ are the lower realized value and upper realized value of the true price of the instrument, respectively; $\delta$ is the probability of realizing $\underline{V}$; $\mu$ is the probability of an information trader traded the instrument; therefore, the best way of understanding the above event tree is to view it from the "leaves" to the "root".

# EXTENSIONS OF THE ROLL MODEL (II)

- Sequential Trade Price Models (continued from previous page)
  - The non-arbitrage "offer" price can be derived using the following formula:

$$A = E[V|Buy] + c = \Pr(\underline{V}|Buy)\,\underline{V} + \Pr(\overline{V}|Buy)\overline{V} + c = \frac{\underline{V}(1-\mu)\delta + \overline{V}(1-\delta)(1+\mu)}{1+\mu(1-2\delta)} + c;$$

  - Similarly, we can derive the non-arbitrage "bid" price as:

$$B = E[V|Sell] - c = \Pr(\underline{V}|Sell)\,\underline{V} + \Pr(\overline{V}|Sell)\overline{V} - c = \frac{\underline{V}(1+\mu)\delta + \overline{V}(1-\delta)(1-\mu)}{1-\mu(1-2\delta)} - c;$$

  - The mid-quote will be:

$$\frac{A+B}{2} = \frac{2\underline{V}\delta[1+\mu^2(1-2\delta)] + 2\overline{V}(1-\delta)[1-\mu^2(1-2\delta)]}{1-\mu^2(1-2\delta)^2};$$

  - The bid-ask spread will be:

$$A - B = \frac{4(1-\delta)\delta\mu(\overline{V}-\underline{V})}{1-\mu^2(1-2\delta)^2} + 2c.$$

# EXTENSIONS OF THE ROLL MODEL (III)

- Models that are based on the stochastic nature of events and order arrival processes (Easley, Kiefer and O'Hara 1997)
  - Consider that, for any observable trade in the market, it can be driven by an information event, or not;
  - Therefore, we have to assume a process for the arrivals of information events;
  - Assume that:
    - Unconditional on any information events, the arrival of **noise** trades (buy or sell) follows a Poisson process with parameter $\varepsilon$; that is, the probability of having *n* trades (again, buy or sell) in a time interval of $\Delta t$ is $e^{-n}(\varepsilon\Delta t)^n/n!$;
    - The arrival of **information** trades is always dependent on the arrival of an information event, which is a Poisson process with parameter $\mu$; further assume that, for each trade event, the probability of it associating with an information event is $\alpha$; therefore, an arbitrary trade without information is 1-$\alpha$;
    - Assuming that the probability that market maker over estimates the true value of an asset is $\delta$ (see page 42); then, for a market maker who is observing the market, the probability of a trade that happens (buy or sell) can be written as:

      $$\Pr(trade) = (1-\alpha)\,[\Pr(b;\varepsilon) + \Pr(s;\varepsilon)] + \alpha[\delta \Pr(b;\varepsilon)\Pr(s;\varepsilon+\mu) + (1-\delta)\Pr(s;\varepsilon)\Pr(b;\varepsilon+\mu)].$$
  - A key concept in the Easley, Kiefer and O'Hara's model is the **Probability of Informed Trading (PIN)**, which is defined as the unconditional probability that a randomly chosen trade on a randomly chosen day is informed:

    $$PIN = \alpha\mu/(\alpha\mu + 2\varepsilon).$$

# KYLE MODEL (1985) (I)

- Strategic trade models
  - Kyle (1985) assumes that the interaction between the information trader and the market maker is essentially a game that determines the optimal prices for both parties;
  - The total demand from the market maker in terms of liquidity is $y = x + u$ where $x$ is the flow from information trader and $u \sim N(0, {\sigma_u}^2)$ from noise traders;
  - The information trader has an estimate of the true price of the instrument to be $v \sim N(p_0, \Sigma_0)$;
  - The market maker determines the price based on the following formula: $p = \lambda y + \mu$; if $\lambda$ is large, it indicates that the market maker may charge some premium due to elevated demand;
  - The information trader, on the other hand, expects a profit from the trade of $\pi = x[v - p]$; when plugging $p = \lambda y + \mu$, we obtain $\pi = x[v - \lambda(x + u) - \mu]$; thus, $E[\pi] = x(v - \lambda x - \mu)$ is the expected profit; by optimizing it, the information trader get an optimal demand of $\overline{x} = (v - \mu)/2\lambda$, which depends on the market maker's views on both a "baseline price ($\mu$) and the liquidity elasticity ($\lambda$) set by the market maker;
  - In this game, the market maker has to figure out two things before determining $p$, the price: one is the total demand by the information trader, and the other is the information trader's estimate on *v*, the true price of the instrument; the Kyle model finds out that these two tasks can be done at the same time as long as we assume that the market maker guesses that demand from the information trader can be written as $\overline{x} = \alpha + \beta v$; this gives the results of: $\alpha = -\mu/2\lambda$ and $\beta = 1/2\lambda$.

# KYLE MODEL (1985) (II)

- Strategic trade models (continued from previous page)
  - As the market maker observes *y* from the market with the repeated guess of $\overline{x} = \alpha + \beta v$, she will try to estimate $E[v|y]$, which can be written as (using projection theorem):

$E[v|y] = E[v] + \frac{cov(v,y)}{cov(y,y)}(y - E[y]) = p_0 + \frac{\beta\Sigma_0(y-\alpha-\beta p_0)}{{\sigma_u}^2+\beta^2\Sigma_0} = \frac{\beta\Sigma_0}{{\sigma_u}^2+\beta^2\Sigma_0}y + (p_0 - \frac{\beta\Sigma_0(\alpha+\beta p_0)}{{\sigma_u}^2+\beta^2\Sigma_0})$, which should be the price that the market maker sets up as $p = \lambda y + \mu$;

  - Therefore, $\lambda = \frac{\beta\Sigma_0}{{\sigma_u}^2+\beta^2\Sigma_0}$ and $\mu = p_0 - \frac{\beta\Sigma_0(\alpha+\beta p_0)}{{\sigma_u}^2+\beta^2\Sigma_0}$; note that λ and μ are parameters that a market maker needs to decide on the price that she needs to quote, while α and β are the parameters that the information trader needs to determine in order to finalize the amount she wants to trade on (note that she should already have an estimate on the true price of the instrument);
  - Now there are four parameters: two for the market maker, λ and μ, and two for the information trader, α and β, under the played-out game assumption that the demand from information trader is linearly dependent on the true value of the instrument via α and β; what is more, α and β are related to λ and μ via the assumption that the information trader optimizes her demand by knowing the price that the market maker will quote;
  - From previous page, we already have: $\alpha = -\mu/2\lambda$ and $\beta = 1/2\lambda$;
  - Therefore, the four parameters are determined via three market variables: $p_0$, $\Sigma_0$ and ${\sigma_u}^2$:

$$\beta = \sqrt{\frac{{\sigma_u}^2}{\Sigma_0}},\ \lambda = \frac{1}{2}\sqrt{\frac{\Sigma_0}{{\sigma_u}^2}},\ \alpha = p_0\sqrt{\frac{{\sigma_u}^2}{\Sigma_0}},\ \text{and}\ \mu = p_0.$$

# KYLE MODEL (1985) (III)

- Strategic trade models (continued from previous page)
  - An interesting conclusion from the Kyle model is that:

$$E[\pi] = \frac{(v-p_0)^2}{2}\sqrt{\frac{{\sigma_u}^2}{\Sigma_0}},$$

which indicates that the higher the noise level, the higher the profit for the information trader.

# NON-INFORMATION-BASED PRICE MODEL

- This group of market microstructure models are largely originated from the Econo-physics community, with Bouchaud, Farmer and Lillo (2009) as a good summary of discussions in this area;
- The basic idea behind these models is that the market is in ensemble-based equilibrium with fluctuations of key variables at microscopic level; as a result, the price determination process should be governed by basic physical laws, and all other phenomena can be derived from these basic laws;
- A few basic laws that are worth mentioning:
  - Order sign is a homogeneous and long memory process;
  - The market impact of each individual trade can be formulated as a "kernel" function modified by instrument-specific parameters;
  - Order flow (especially signed flow) has a profound effect on the impact of the trade that is about to happen;
- A key conclusion from this group of models is that the market maker can quote the mid-quote using the formula:

$$p_t = \int_{-\infty}^{t} G(t-\tau) f(\tau) d\tau + \varepsilon_t,$$

where *G* is a kernel (or, Green) function, *f* a function of key liquidity parameters (such as order size, average daily trading volume, etc.) and the order sign of individual trades in history; $\varepsilon_t$ is a innovation term that is determined outside of the convolution operator.

# THAT'S ALL FOR THIS LECTURE.

# THANK YOU!