MTH 9879 Market Microstructure Models

Inventory models and market-making

Jim Gatheral

Baruch College
The City University of New York

Spring 2014

Outline of lecture

- The Garman (1976) model
- Amihud and Mendelson (1980)
- The Stoll (1978) model
- Ho and Stoll (1981)
- Dynamic programming principle and the HJB equation
- Avellaneda and Stoikov (2008)
- Guéant, Lehalle and Fernandez-Tapia (2013)
- Guilbaud and Pham (2013)

Garman (1976)

A quote from Garman (1976):

> *We depart from the usual approaches of the theory of exchange by (1) making the assumption of asynchronous, temporally discrete market activities on the part of market agents and (2) adopting a viewpoint which makes the temporal microstructure,* i.e., *moment-to-moment aggregate exchange behavior, as an important descriptive aspect of such markets.*

- Dealers are needed because buyers and sellers don't arrive synchronously.
  - In our context, agents may act as market makers without necessarily being dealers as such.

Garman (1976)

- MB orders arrive at rate $\lambda_A(p)$ and MS orders at rate $\lambda_B(p)$, both functions of the price p.
  - $\lambda_A(p)$ is decreasing in p and $\lambda_B(p)$ is increasing in p.
- If the dealer were to quote a choice price, the graphs would cross at $p^\star$ given by:

$$\lambda_A(p^\star) = \lambda_B(p^\star) =: \lambda^\star$$

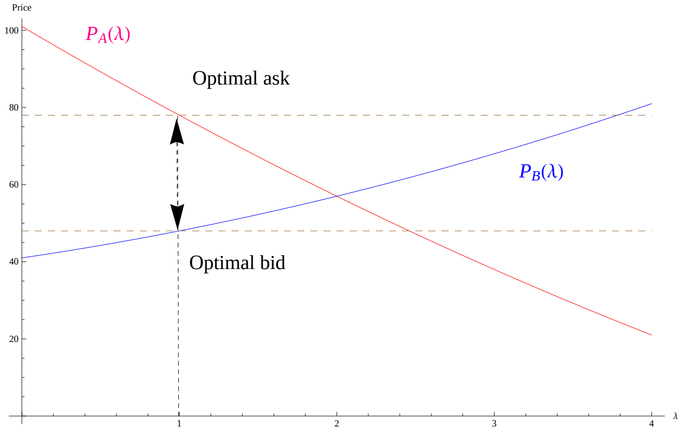
- If instead the dealer quotes a two-way price $\{B, A\}$, and on average $\lambda_A(A) = \lambda_B(B) =: \tilde{\lambda}$, the P&L per unit of time will be given by

$$\pi(B, A) = (A - B)\,\tilde{\lambda}.$$

  - The wider the spread, the greater the P&L per trade but the lower the rate of trading.
  - The dealer finds the optimal bid and offer by maximizing P&L per unit time.

Garman model schematic

Figure 1: As the price increases, the rate λ_B of MS increases, the rate λ_A of MB decreases.



Garman (1976)

- The dealer needs to hold cash and securities in order to be able to accommodate asynchronous buying and selling.
- Garman assumes that the bid B and ask A are set once and for all. This implies that
  - Cash holdings follow a zero-drift random walk.
  - Stock holdings follow a zero-drift random walk.
- The dealer is ruined with probability one.
  - With realistic parameters, he is ruined in a few days!
- The main intuition is then

> *[Garman] "The order of magnitude makes it clear that the specialists must pursue a policy of relating their prices to their inventories in order to avoid failure."*

Amihud and Mendelson (1980)

- Framework similar to Garman.
- Inventory is constrained to lie between given lower and upper bounds.
- The market maker maximizes expected profit per unit time.

Amihud and Mendelson implications

- B and A are monotone decreasing functions of inventory.
- There is a spread $s = A - B > 0$.
- There is a preferred inventory and s is an increasing function of the distance from this preferred inventory position.
- The quotes are not set symmetrically about fair value.

CARA utility

- CARA stands for *Constant Absolute Risk Aversion*.
- The utility function is

$$U(W) = -e^{-\lambda\, W}$$

- If W is normally distributed,

$$\mathbb{E}\left[U(W)\right] = \mathbb{E}\left[-e^{-\lambda\, W}\right] = -\exp\left\{-\lambda\, \left(\mathbb{E}[W] - \lambda/2\, \mathrm{Var}[W]\right)\right\}$$

  so maximizing CARA utility is equivalent to maximizing

$$\mathcal{L}(W) = \mathbb{E}[W] - \lambda/2\, \mathrm{Var}[W]$$

  which is just mean-variance optimization.
- $\mathbb{E}[W] - \lambda/2\, \mathrm{Var}[W]$ is called the *certainty equivalent* of W.

Stoll (1978)

- The dealer has exponential utility:

$$U(W) = -e^{-\alpha\,W}.$$

- Inventory is constrained to lie between given lower and upper bounds.
- The security has the random payoff $S \sim N(\mu, \sigma^2)$.

Stoll (1978) optimal bid-ask computation

- Suppose the dealer's current inventory is q.
- He posts a bid at B.
  - If his bid is not hit, his terminal wealth will be $W = q\,S$.
  - If his bid is hit, his terminal wealth will be $W = (q+1)S - B$.
- For the dealer to be indifferent, we must have

$$\mathbb{E}[U\left((q+1)S - B\right)] = \mathbb{E}[U(q\,S)]$$

so

$$(q+1)\mathbb{E}[S] - B - \frac{\alpha}{2}\,\mathrm{Var}[(q+1)\,S] = q\,\mathbb{E}[S] - \frac{\alpha}{2}\,\mathrm{Var}[q\,S] \qquad (1)$$

Then

$$B = \mu - \frac{\alpha}{2}\,\left[(q+1)^2 - q^2\right]\,\sigma^2 = \mu - \frac{\alpha}{2}\,(2\,q+1)\,\sigma^2. \qquad (2)$$

Stoll (1978) optimal inventory

- This indifference price argument works only if the dealer starts with optimal inventory.
  - What is his optimal inventory?
- Suppose that the current market price of the stock is P.
- Terminal wealth is then $q\,(S - P)$.
- Maximizing terminal wealth over q then gives optimal inventory:

$$\frac{d}{dq}\,\left\{q\,(\mu - P) - \frac{\alpha}{2}\,q^2\,\sigma^2\right\} = 0$$

when

$$q = \frac{\mu - P}{\alpha\,\sigma^2} =: q^\star$$

- Substituting back into (2) gives

$$B = P - \alpha\,\frac{\sigma^2}{2}. \tag{3}$$

Extension to n assets

- In the n asset case, we can make the same indifference price argument to derive the optimal bids and offers.
  - Assume the joint distribution of the payoffs of the n assets is multivariate normal.
- The analog of equation (1) for finding the optimal bid B_j for asset j is

$$\begin{aligned} & \sum_i^n q_i\,\mu_i + \mu_j - B_j - \frac{\alpha}{2}\,\mathrm{Var}\left[\sum_i^n q_i\,S_i + S_j\right] \\ = & \sum_i^n q_i\,\mu_i - \frac{\alpha}{2}\,\mathrm{Var}\left[\sum_i^n q_i\,S_i\right] \end{aligned} \tag{4}$$

This gives

$$
\begin{aligned}
B_j &= \mu_j - \alpha \sum_i^n q_i \,\mathrm{Cov}\left[S_i, S_j\right] - \frac{\alpha}{2}\,\mathrm{Var}[S_j] \\
&= \mu_j - \alpha \sum_i^n q_i \,\rho_{ij}\,\sigma_i\,\sigma_j - \frac{\alpha}{2}\,\sigma_j^2 \qquad (5)
\end{aligned}
$$

Extending further to a quote in size n_j, we see that

$$
B_j = \mu_j - \alpha \sum_i^n q_i \,\rho_{ij}\,\sigma_i\,\sigma_j - \frac{\alpha}{2}\,n_j\,\sigma_j^2 \qquad (6)
$$

A completely analogous argument gives the fair ask price

$$
A_j = \mu_j - \alpha \sum_i^n q_i \,\rho_{ij}\,\sigma_i\,\sigma_j + \frac{\alpha}{2}\,n_j\,\sigma_j^2
$$

From equation (6), we note that

- The optimal bid for the jth asset depends in general on existing positions in all other assets.
- Assuming the correlations ρ_{ij} are positive, the greater the initial inventory, the lower the bid is biased.
  - The lower the correlation with existing inventory, the less the quote bias.
- The spread

$$s_j = A_j - B_j = \alpha \, n_j \, \sigma_j^2$$

  is
  - Increasing in the risk aversion coefficient (price of risk) α.
  - Increasing in the asset volatility σ_j.
  - Increasing in the quote size n_j.
- The spread does not depend on inventory.

- Mean-variance optimization is equivalent to adopting variance as a risk measure.
  - The units ($\$^2$) are wrong: We don't think that double the position has four times the risk!
- Suppose we take value-at-risk (VaR) as our risk measure instead.
  - This choice of risk measure is more consistent with allocated cost of capital in reality.
- The analog of (4) for a quote of size m is

$$\begin{aligned} & \sum_i^n q_i\,\mu_i + m\,(\mu_j - B_j) - \nu\,\sqrt{\mathrm{Var}\left[\sum_i^n q_i\,S_i + m\,S_j\right]} \\ = \; & \sum_i^n q_i\,\mu_i - \nu\,\sqrt{\mathrm{Var}\left[\sum_i^n q_i\,S_i\right]}. \end{aligned}$$

Rearranging gives

$$B_j = \mu_j - \frac{\nu}{m}\left\{\sqrt{\mathrm{Var}\left[\sum_i^n q_i\,S_i + m\,S_j\right]} - \sqrt{\mathrm{Var}\left[\sum_i^n q_i\,S_i\right]}\right\} \tag{7}$$

To get intuition for this expression, consider the one-stock case where there is existing inventory q. The optimal bid is then

$$B = \mu - \frac{\nu}{m}\,\sigma\,\{|q+m| - |q|\} = \mu - \frac{\nu}{m}\,\sigma \begin{cases} m & \text{if } q \geq 0 \\ 2\,q + m & \text{if } -m \leq q < 0 \\ -m & \text{if } q < -m. \end{cases}$$

- If the potential trade is not risk reducing, the quoted spread is linear in the volatility σ and (approximately) independent of the quoted size m.
  - Spread is linear in volatility both in previous models and empirically.
  - Intuitively, quoted spread should somewhat depend on quoted size in dealer markets (but it may be independent of size over a large range of sizes).
  - Note also that ν is dimensionless.

Monetary risk measures

Let $\mathcal{X}$ denote the set of all financial positions.

The following definition is from [Fölmer and Schied]:

Definition (4.1)

A mapping $\rho : \mathcal{X} \to \mathbb{R}$ is called a *monetary risk measure* if it satisfies the following conditions for all $X, Y \in \mathcal{X}$:

- *Monotonicity*: If $X \leq Y$, then $\rho(X) \geq \rho(Y)$.
- *Cash invariance*: If $m \in \mathbb{R}$, then $\rho(X + m) = \rho(X) - m$.

- Without loss of generality, we could have $\rho(0) = 0$.
- Obviously, variance does not qualify as a monetary risk measure.

Ho and Stoll (1981)

- Ho and Stoll extend Stoll (1978) to a multiperiod framework.
  - Both order flow and stock price are stochastic.
- The log-stock price x_t is modeled as Brownian motion with drift.
- Market order arrivals are Poisson with rates λ_B and λ_A. These rates depend linearly on the bid and ask prices B and A set by the dealer.
- Inventory has a stochastic evolution in addition to its natural evolution resulting from the arrival of market orders.
- The dealer's utility function is quadratic.

- The solution to this problem involves stochastic optimal control techniques and in fact, this specific problem seems to be too hard to solve in closed-form.

Quick review of Itô's Lemma

Assume

$$dx = \mu\, dt + \sigma\, dW$$

and $V(x,t)$ some function of x and t. Then

$$\begin{aligned} dV &= \frac{\partial V}{\partial t}\, dt + \frac{\partial V}{\partial x}\, dx + \frac{1}{2}\, \frac{\partial^2 V}{\partial x^2}\, dx^2 \\ &= \frac{\partial V}{\partial t}\, dt + \frac{\partial V}{\partial x}\, (\mu\, dt + \sigma\, dW) + \frac{1}{2}\, \frac{\partial^2 V}{\partial x^2}\, \sigma^2\, dt \\ &= \left\{ \frac{\partial V}{\partial t} + \mu\, \frac{\partial V}{\partial x} + \frac{\sigma^2}{2}\, \frac{\partial^2 V}{\partial x^2} \right\} dt + \sigma\, \frac{\partial V}{\partial x}\, dW \end{aligned}$$

Diffusion equation in 1D

Then, if $J = \mathbb{E}[V(x,t)]$, J satisfies

$$\frac{dJ}{dt} = \partial_t J + \mathcal{L} J$$

where

$$\mathcal{L} = \mu \, \partial_x + \frac{\sigma^2}{2} \, \partial_{x,x}$$

is the infinitesimal generator of the Itô diffusion.

Bellman's principle of optimality

"An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision."

(See Bellman, 1957, Chap. III.3.)

Stochastic optimal control

Consider a utility functional of the form

$$J_t = \mathbb{E}\left[\int_t^T h(s, y_s, v_s)\, ds \middle| \mathcal{F}_t\right]$$

where y_s is a state vector, v_s is a vector-valued control and the evolution of the system is determined by a stochastic differential equation (SDE):

$$dy_t = \mu(t, y_t, v_t)\, dt + \sigma(t, y_t, v_t)\, dZ_t$$

The principle of dynamic programming amounts to stating that

$$\begin{aligned} J_t &= \max_{v \in \mathcal{G}} \mathbb{E}\left[h(t, y_t, v_t)\, dt + \mathbb{E}\left[\int_{t+dt}^T h(s, y_s, v_s)\, ds \middle| \mathcal{F}_{t+dt}\right]\right] \\ &= \max_{v \in \mathcal{G}} \mathbb{E}\left[h(t, y_t, v_t)\, dt + J_{t+dt}\right] \\ &= \max_{v \in \mathcal{G}} \left\{h(t, y_t, v_t)\, dt + J_t + \left\{\frac{\partial J}{\partial t} + \mathcal{L}_t^v J\right\} dt\right\}. \end{aligned}$$

The HJB equation

Canceling the J_t terms and dividing by dt gives the HJB equation:

$$\frac{\partial J}{\partial t} + \max_{v \in \mathcal{G}} \left\{ \mathcal{L}_t^v \, J + h(t, y_t, v) \right\} = 0$$

where $\mathcal{L}_t^v$ is the infinitesimal generator of the Itô diffusion:

$$\mathcal{L}_t^v = \frac{1}{2} \, \sigma(t, y, v)^2 \, \partial_{y,y} + \mu(t, y, v) \, \partial_y .$$

Ho and Stoll setup

- The dealer quotes prices

$$p_b = p - b;\, p_a = p + a$$

- Cash:

$$dF = r\,F\,dt - (p - b)\,dq_b + (p + a)\,dq_a$$

- Stock:

$$dX = r_X\,X\,dt + \sigma_X\,X\,dZ_X$$

- Inventory:

$$dI = r_I\,I\,dt + p\,dq_b - p\,dq_a + I\;\sigma_I\,dZ_I$$

  where dq_a and dq_b represent Poisson distributed jumps with intensities λ_a, λ_b.

- Base wealth:

$$dY = r_Y\,Y\,dt + Y\,dZ_Y$$

Ho and Stoll objective

The dealer's objective is to maximize $\mathbb{E}\left[U(W_T)\right]$ with

$$W_T = F_T + I_T + Y_T$$

The value function $J(\cdot)$ is a function of F, I and Y. Since there is no intermediate consumption $h(\cdot)$, the problem reduces to

$$\max_{a,b \in \mathcal{G}} dJ(t, F, I, Y) = 0$$

with $J(T, F, Y, I) = U(W_T)$.

HJB equation for Ho and Stoll value function

Ignoring interest rates and returns (which are effectively irrelevant for small times), we get

$$\frac{\partial J}{\partial t} + \mathcal{L}_t\, J + \max_{a,b \in \mathcal{G}} \{\lambda_a\, [J(t, F + p\, Q + a\, Q, I - p\, Q, Y) - J(t, F, I, Y)] \\ + \lambda_b\, [J(t, F - p\, Q + b\, Q, I + p\, Q, Y) - J(t, F, I, Y)]\} = 0$$

where $\mathcal{L}_t$ is the infinitesimal generator of the Itô diffusion:

$$\mathcal{L}_t = \frac{1}{2}\, \sigma_Y^2\, Y^2\, \partial_{Y,Y} + \frac{1}{2}\, \sigma_I^2\, I^2\, \partial_{I,I} + \sigma_Y\, \sigma_I\, Y\, I\, \partial_{Y,I}$$

- Not surprisingly, this equation is near impossible to solve, even approximately.

Conclusions of Ho and Stoll (1981)

- The spread depends on the time horizon of the dealer.
  - The shorter the time remaining, the lower the risk – this is somewhat artificial!
- The risk adjustment is increasing in both the risk aversion α and the variance σ^2.
- The spread is independent of inventory level.
  - The dealer moves his mid-price depending on inventory so as to influence the arrival rates of MB and MS orders.

- All of this is consistent with the Stoll (1978) conclusions.

- Also, MB orders cause the mid-price to increase, MS orders cause the mid-price to decrease; there is market impact.

- As in Stoll (1978), the agent's utility is exponential.
- Stock price follows arithmetic Brownian motion

$$dS = \sigma\, dW$$

- Consider first the "frozen inventory limit" where the agent holds inventory q until time T without trading.
- Wlog, take initial wealth to be x. Then the value function is

$$\begin{aligned} J(x,S,q,t) &= -e^{-\alpha\, x}\, \mathbb{E}_t\left[e^{-\alpha\, q\, S_T}\right] \\ &= -e^{-\alpha\, x}\, e^{-\alpha\, q\, S_0}\, \exp\left\{\frac{\alpha^2\, q^2\, \sigma^2\, (T-t)}{2}\right\} \\ &=: -e^{V(x,q,t)} \end{aligned}$$

where x is initial wealth and q is inventory.

Reservation bid price

The reservation bid price r_B solves

$$V(x - r_B, S, q+1, t) = V(x, S, q, t).$$

The solution, again as in Stoll, is

$$r_B(x, S, q, t) = S - \frac{\alpha}{2}\,(2\,q + 1)\,\sigma^2\,(T - t)$$

Similarly,

$$r_A(x, S, q, t) = S + \frac{\alpha}{2}\,(-2\,q + 1)\,\sigma^2\,(T - t)$$

The average of the reservation bid and ask prices

$$r^\star = (r_A + r_B)/2 = S - \alpha\,q\,\sigma^2\,(T - t) \tag{8}$$

is referred to as the *reservation price*.

Model setup

- The MB arrival rate λ^A is decreasing in $\delta^A = A - S$.
- The MS arrival rate λ^B is decreasing in $\delta^B = S - B$.
- Wealth in cash jumps every time there is a market order:

$$dx_t = A\, dN_t^A - B\, dN_t^B$$

- Stock inventory at time t is given by:

$$q_t = N_t^B - N_t^A$$

  - Inventory thus depends only on arrivals of market orders and does not have any additional random component. This considerably simplifies the analysis relative to Ho and Stoll.

The HJB equation

The objective of the dealer is to maximize

$$J(S,x,q,t) = \max_{\delta^B,\delta^A} \mathbb{E}_t\left[-e^{-\alpha\,(x_T+q_T\,S_T)}\right]$$

where the optimal controls δ^B and δ^A are in general time and state dependent.

The change in utility due to a market buy order is given by

$$J(S,x+A,q+1,t) - J(S,x,q,t)$$

and the change in utility due to a market sell order by

$$J(S,x-B,q-1,t) - J(S,x,q,t)$$

$J(\cdot)$ satisfies the HJB equation

$$
\begin{aligned}
& \partial_t J + \frac{\sigma^2}{2}\,\partial_{S,S} J \\
& + \max_{\delta^B} \lambda^B(\delta^B)\,[J(S, x - B, q + 1, t) - J(S, x, q, t)] \\
& + \max_{\delta^A} \lambda^A(\delta^A)\,[J(S, x + A, q - 1, t) - J(S, x, q, t)] \\
= \; & 0 \qquad (9)
\end{aligned}
$$

Because utility is exponential, we can simplify (9) with the ansatz

$$J(S, x, q, t) = -e^{-\alpha\, x}\, e^{-\alpha\, \theta(S,q,t)}$$

By definition of the reservation price r_A, we have

$$J(S, x + r_A, q - 1, t) = J(S, x, q, t)$$

and from the ansatz,

$$J(S, x + A, q - 1, t) = e^{-\alpha\,(A - r_A)}\, J(S, x + r_A, q - 1, t).$$

Then

$$J(S, x + A, q - 1, t) - J(S, x, q, t) = J(S, q, t)\, \left[e^{-\alpha\,(A - r_A)} - 1\right]$$

and

$$J(S, x - B, q + 1, t) - J(S, x, q, t) = J(S, q, t)\, \left[e^{-\alpha\,(-B + r_B)} - 1\right].$$

Substituting the ansatz

$$J(S,x,q,t) = -e^{-\alpha\,x}\,e^{-\alpha\,\theta(S,q,t)}$$

back into the HJB equation (9) gives the following equation for $\theta(\cdot)$:

$$\begin{aligned} & \partial_t\theta + \frac{\sigma^2}{2}\,\partial_{S,S}\theta - \alpha\,\frac{\sigma^2}{2}\,(\partial_S\theta)^2 \\ & +\frac{1}{\alpha}\,\max_{\delta^B}\,\lambda^B(\delta^B)\,\left[1 - e^{-\alpha\,(-B+r_B)}\right] \\ & +\frac{1}{\alpha}\,\max_{\delta^A}\,\lambda^A(\delta^A)\,\left[1 - e^{-\alpha\,(A-r_A)}\right] \\ = \; & 0 \end{aligned} \tag{10}$$

with boundary condition

$$\theta(S,q,T) = q\,S.$$

The first order condition for the bid price B is then

$$\partial_B \left\{ \lambda_B(B) \, \left[e^{-\alpha\,(-B+r_B)} - 1 \right] \right\} = 0$$

which gives the implicit relation

$$\lambda_B{}'(B) \, \left[e^{-\alpha\,(-B+r_B)} - 1 \right] + \alpha \, \lambda^B(B) \, e^{-\alpha\,(-B+r_B)} = 0$$

which may be further simplified to

$$B = r_B - \frac{1}{\alpha} \log \left(1 + \alpha \, \frac{\lambda_B(B)}{\lambda_B{}'(B)} \right) \tag{11}$$

Similarly, the optimal ask price is given by

$$A = r_A + \frac{1}{\alpha} \log \left(1 - \alpha \, \frac{\lambda_A(A)}{\lambda_A{}'(A)} \right) \tag{12}$$

Discussion

- Equations (11) and (12) set optimal bid and ask levels as a function of inventory q and the arrival rate function $\lambda(\cdot)$.
- Optimal bid and ask prices are obtained in two steps:
  - Find the reservation prices r_B and r_A.
  - Adjust these to the current best estimates of the market order arrival rate functions.

Estimating order arrival rates

Following Avellaneda and Stoikov again, we may estimate the $\lambda(\cdot)$ as follows:

- Assume that the density of market order size is

$$f(x) \propto \frac{1}{x^{1+\nu}}.$$

  It is a stylized fact that $\nu \approx 3/2$.

- Assume that market impact is given roughly by

$$\Delta P \propto x^{\beta}$$

  where x is order size. Another stylized fact is that $\beta \sim 1/2$.

Let δ be distance from the best quote and assume that order arrival rates are constant in aggregate. Then the rate of arrival of market orders at δ can be (handwavingly) estimated as:

$$
\begin{aligned}
\lambda(\delta) &\approx \Pr(\Delta p > \delta) \\
&\propto \Pr(x > \delta^{1/\beta}) \\
&\propto \int_{\delta^{1/\beta}}^{\infty} \frac{dx}{x^{1+\nu}} \\
&= \frac{1}{\nu}\,\delta^{-\nu/\beta}
\end{aligned}
$$

Substituting $\nu = 3/2$ and $\beta = 1/2$ gives

$$\lambda(\delta) \sim \frac{1}{\delta^3}$$

This is nothing other than the famous cubic law of [Gabaix et al.]!

Then

$$\frac{\lambda'(\delta)}{\lambda(\delta)} \approx -\frac{3}{\delta}$$

and recalling that $B = S - \delta_B$, we get

$$\begin{aligned} B &= r_B - \frac{1}{\alpha} \log\left(1 + \alpha\, \frac{\lambda_B(B)}{\lambda_B{}'(B)}\right) \\ &= r_B - \frac{1}{\alpha} \log\left(1 + \alpha\, \delta_B/3\right) \\ &\approx r_B - \delta_B/3 \end{aligned}$$

Similarly, $A \approx r_A + \delta_A/3$.

Infinite horizon objective

- The reservation price (8) depends explicitly on time because it relates to the variance of the payoff at some final time.
- For a market-making business, (we hope) there is no final time. Avellaneda and Stoikov suggest considering an objective of the form

$$
\begin{aligned}
V(S,x,q) &= \mathbb{E}\left[-\int_0^\infty e^{-\omega\, t}\, e^{-\alpha\,(x+q\,S_t)}\, dt\right] \\
&= -e^{-\alpha\,(x+q\,S)} \int_0^\infty e^{-(\omega-\alpha^2\,q^2\,\sigma^2/2)\,t}\, dt \\
&= -\frac{e^{-\alpha\,(x+q\,S)}}{\omega-\alpha^2\,q^2\,\sigma^2/2}
\end{aligned}
$$

Infinite horizon reservation price

Now we have the value function $V(\cdot)$, we can apply the definition of reservation price r_B:

$$V(S, x - r_B, q + 1) = V(S, x, q)$$

to obtain

$$r_B = S + \frac{1}{\alpha} \log \left(\frac{\tilde{\omega} - (q+1)^2}{\tilde{\omega} - q^2} \right) \tag{13}$$

where

$$\tilde{\omega} = \frac{2\,\omega}{\alpha^2\,\sigma^2} =: (q_{\max} + 1)^2$$

Similarly

$$r_A = S - \frac{1}{\alpha} \log \left(\frac{\tilde{\omega} - (q-1)^2}{\tilde{\omega} - q^2} \right) \tag{14}$$

Small inventory limit

In the limit of small inventory $|q| \ll q_{\rm max}$, (13) and (14) become

$$r_B \approx r^\star - s/2;\; r_A \approx r^\star + s/2$$

with

$$\begin{aligned} s/2 &\approx \frac{1}{\alpha\, q_{\rm max}^2} \\ r^\star &\approx S - 2\,q\,\frac{1}{\alpha\, q_{\rm max}^2} = S - q\,s \end{aligned}$$

Fair price P

- We may repeat the foregoing analysis with the current stock price S replaced with the market maker's private valuation P (which we may think of as the fair price of the stock).
  - P may be estimated by observing the state of the order book.
  - λ_A and λ_B may be estimated using the autocorrelation of order flow.

Symmetric exponential arrival rates

Following Avellaneda and Stoikov, suppose that

$$\lambda^A(\delta) = \lambda^B(\delta) = A\, e^{-k\,\delta}.$$

- This is a questionable approximation because arrival rates probably depend on inventory.
  - Adverse selection again!
- However, making this approximation gives the same expression for the indifference price as the "frozen inventory" indifference price (8) and simplifies the analysis.

Solution of Avellaneda-Stoikov HJB equation

The HJB equation (10) looks impossible to solve. However, by assuming exponential arrival of limit orders and expanding for small q, [Avellaneda and Stoikov] get

$$r^\star = S - q\,\alpha\,\sigma^2\,(T-t)$$

with

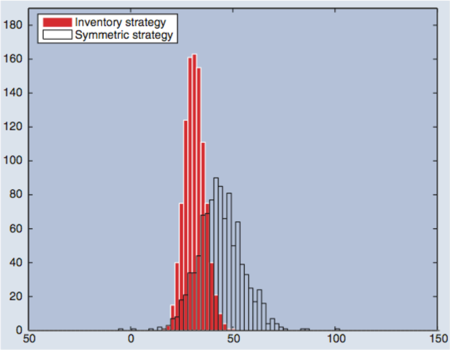
$$\delta_A + \delta_B = \alpha\,\sigma^2\,(T-t) + \frac{2}{\alpha}\,\log\left(1+\frac{\alpha}{k}\right)$$

- Set the reservation price $r^\star$.
- Make a spread around this price.

Results of simulation

Table 3. 1000 simulations with $\gamma = 1$.

Strategy	Average spread	Profit	Std (Profit)	Final q	Std (Final q)
Inventory	3.02	31.4	5.0	0.02	1.7
Symmetric	3.02	44.0	11.0	0.00	5.1



An exact solution of the HJB equation

- [Guéant, Lehalle and Fernandez-Tapia] later came up with a clever way of solving the HJB equation (10) assuming exponential market order arrival rates $\lambda^A(\delta) = \lambda^B(\delta) = A\, e^{-k\,\delta}$.
- The key trick is to impose inventory limits so that $q \in [-Q, Q]$ and discretizing q. Then, by solving an ODE for each q, they can solve the problem in closed-form.
- They obtain

$$\begin{aligned}
\delta_B &= \frac{1}{k} \log \frac{v_q(T)}{v_{q+1}(T)} + \frac{1}{\alpha} \log\left(1 + \frac{\alpha}{k}\right) \\
\delta_A &= \frac{1}{k} \log \frac{v_q(T)}{v_{q-1}(T)} + \frac{1}{\alpha} \log\left(1 + \frac{\alpha}{k}\right)
\end{aligned}$$

where the $v_q(\cdot)$ are solutions of the aforementioned ODEs.

Improved asymptotics

- Contrary to the conclusion drawn by [Avellaneda and Stoikov] from their asymptotic analysis, [Guéant, Lehalle and Fernandez-Tapia] show that the optimal quotes are almost independent of T if time remaining (T) is big enough.
- With a heuristic argument and assuming q small, [Guéant, Lehalle and Fernandez-Tapia] propose the following asymptotic approximation:

$$\begin{aligned}
\delta_B &= \frac{1}{\alpha}\log\left(1+\frac{\alpha}{k}\right)+(q+1/2)\sqrt{\frac{\sigma^2\,\alpha}{2\,k\,A}\,\left(1+\frac{\alpha}{k}\right)^{1+k/\alpha}} \\
\delta_A &= \frac{1}{\alpha}\log\left(1+\frac{\alpha}{k}\right)-(q-1/2)\sqrt{\frac{\sigma^2\,\alpha}{2\,k\,A}\,\left(1+\frac{\alpha}{k}\right)^{1+k/\alpha}}
\end{aligned}$$

which are of course independent of T.

Adding market orders

- The frameworks of both [Avellaneda and Stoikov] and [Guéant, Lehalle and Fernandez-Tapia] assume that the market maker's problem is to place limit orders optimally so as to maximize utility assuming Poisson arrivals of market orders.
  - The optimal limit price could easily be outside the current spread.
- In practice, it could obviously be optimal for the market maker to sometimes cross the spread and send market orders to reduce inventory.
- In the more realistic setup of [Guilbaud and Pham], the market maker may post limit orders at the best bid and offer prices, improve the spread, or send market orders.
  - Market order arrival rates are allowed to be time-dependent (as they are in practice).
  - Market order arrival rates are additionally (real-time estimated) functions of the spread.

Guilbaud and Pham solution

- The resulting HJB equation is discretized and solved numerically.
  - Obviously too complicated to solve analytically.
- Pseudo-code is provided to facilitate implementation of the algorithm.
- The algorithm is tested on simulated data and compared with the following three strategies:
  - WoMO: Same strategy with no market orders
  - Cst: Place constant quantities on each of the best quotes
  - Random: Place constant quantity on each side at random prices that are at the best quote or improve the spread.

Guilbaud and Pham simulation results

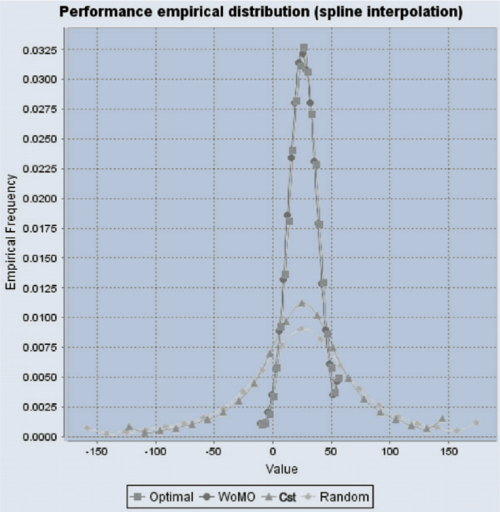


Figure 4. Empirical distribution of the terminal wealth X_T (spline interpolation).

Guilbaud and Pham optimal frontier

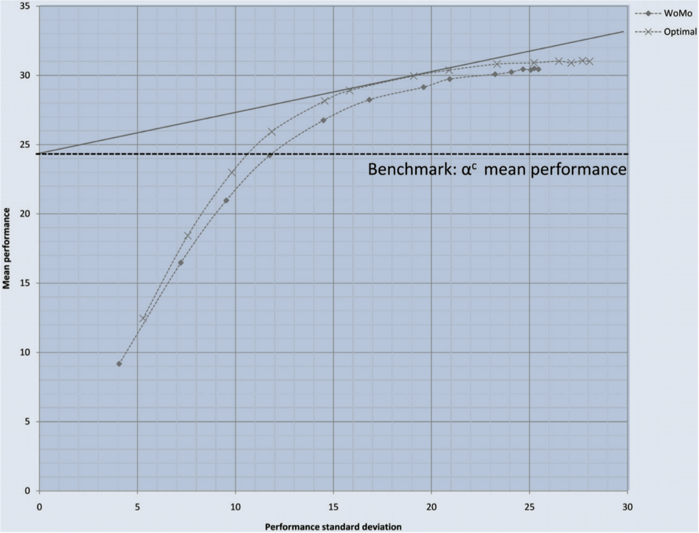


Figure 6. Efficient frontier plot.

Final remarks

- All inventory models have the following characteristics:
  - It is optimal for the market maker to keep inventory close to zero.
  - There will therefore be mean reversion in prices due to inventory effects.
  - The spread is compensation for risk.
  - Biasing prices to encourage liquidation of inventory reduces the variance of returns.
    - This also causes short-term mean reversion in prices.

- In real markets, as in [Guilbaud and Pham], the spread is given.
  - A market maker either joins or improves the best quote, or does no business.
- However, market order arrival rates are not symmetric: they depend on the book imbalance $\mathcal{I}$.
  - This work remains to be done!

References

Marco Avellaneda and Sasha Stoikov, High-frequency trading in a limit order book, *Quantitative Finance*, **8**(3) 217–224 (2008).

Hans Föllmer and Alexander Schied, *Stochastic Finance: An Introduction in Discrete Time*, De Gruyter, Berlin (2011).

Xavier Gabaix, Parameswaran Gopikrishnan, Vasiliki Plerou and H. Eugene Stanley, Understanding the cubic and half-cubic laws of financial fluctuations, *Physica A: Statistical Mechanics and its Applications*, **324**(1-2) 1–5 (2003).

Olivier Guéant, Charles-Albert Lehalle, and Joaquin Fernandez-Tapia, Dealing with the inventory risk: a solution to the market making problem, *Mathematics and Financial Economics* **7**(4) 477–507 (2013).

Fabien Guilbaud and Huyên Pham, Optimal high-frequency trading with limit and market orders, *Quantitative Finance* **13** (1) 79–94 (2013).

Joel Hasbrouck, *Empirical Market Microstructure*, Oxford University Press, Chapter 11 (2007).

Albert J Menkveld, High frequency trading and the new-market makers **SSRN** (2012).