MTH 9879 Market Microstructure Models, Spring 2019

Lecture 3: Inventory models and market-making

Jim Gatheral Department of Mathematics

Baruch COLLEGE
WEISSMAN SCHOOL OF ARTS AND SCIENCES
Master of Financial Engineering

Outline of Lecture 3

- The Garman (1976) model
- Amihud and Mendelson (1980)
- The Stoll (1978) model
- Ho and Stoll (1981)
- Dynamic programming principle and the Hamilton-Jacobi-Bellman (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

$$\pi(B,A) = (A - B)\,\tilde{\lambda} = \left(P_A(\tilde{\lambda}) - P_B(\tilde{\lambda})\right)\,\tilde{\lambda}.$$

  - 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.

- The two-way price $\{B,A\}$ is set once and for all.

Garman model schematic

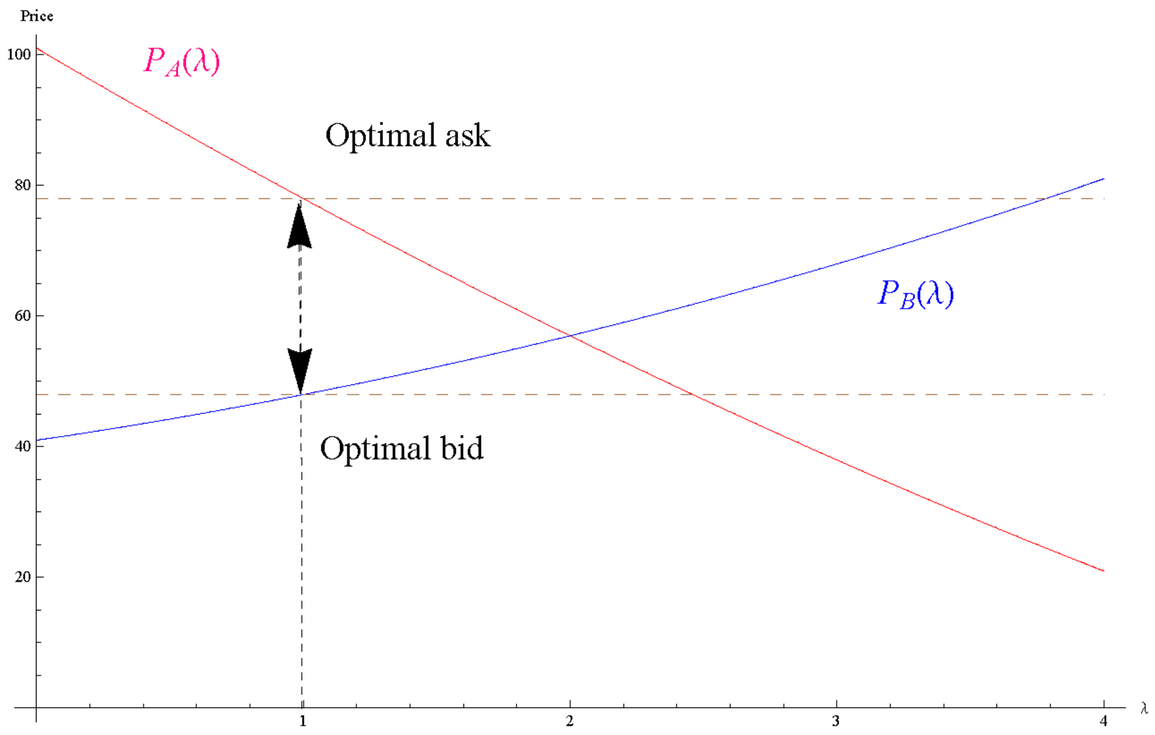

Figure 1: As the price increases, the rate λ_B of MS increases, the rate λ_A of MB decreases.

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 monopolistic market maker sets two-way price dynamically in order to maximize expected profit per unit time in equilibrium.

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 Arrow-Pratt measure of absolute risk-aversion (ARA), a.k.a., the *coefficient of absolute risk aversion*, A of a utility function U is defined as $A = -\frac{U''}{U'}$.
  - Exponential utility $U(x) = \frac{1}{\alpha}\left(1 - e^{-\alpha x}\right)$ is unique in exhibiting constant absolute risk aversion (CARA). Indeed, $A(x) = \alpha$ for exponential utility.
  - Sometimes exponential utility is quoted in the form $U(x) = -e^{-\alpha\, x}$.

- If W is normally distributed,

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

  so maximizing CARA utility is equivalent to maximizing

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

  which is just mean-variance optimization.

- $\mathbb{E}[W] - \frac{\alpha}{2}\,\mathrm{Var}[W]$ 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

  (1)

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

  Then

  (2)

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

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 expected terminal utility 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

  (3)

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

- Similarly, the ask price is given by

$$A = P + \alpha \frac{\sigma^2}{2}.$$

Note

- Bid and ask are symmetric around the current price P and is independent of μ.
- Spread is equal to $\alpha\sigma^2$. The more risk averse the market maker is or the more volatile the security is, the wider the spread.
- Spread is quadratic in volatility.

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

(4)

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

This gives

(5)

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

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

(6)

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

A completely analogous argument gives the fair ask price

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

Observations

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.

Choice of risk measure

- Mean-variance optimization is equivalent to adopting variance as a measure of risk.
  - The units ($\$^2$) are wrong: We don't think that double the position would quadraple 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

(7)
$$\begin{aligned}B_j &= \mu_j\\ &- \frac{\nu}{m}\left\{\sqrt{\mathrm{Var}\left[\sum_i^n q_i\,S_i + m\,S_j\right]} - \sqrt{\mathrm{Var}\left[\sum_i^n q_i\,S_i\right]}\right\}\end{aligned}$$

To get intuition for this expression, consider the one-stock case where there is existing inventory q. The optimal bid is then
$$\begin{aligned}B &= \mu - \frac{\nu}{m}\,\sigma\,\{|q+m| - |q|\}\\ &= \mu - \frac{\nu}{m}\,\sigma\begin{cases} m & \text{if } q \geq 0\\ 2\,q + m & \text{if } -m \leq q < 0\\ -m & \text{if } q < -m.\end{cases}\end{aligned}$$

- 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][3]:

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 $\rho(X + m) = \rho(X) - m$.

Note

- 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.

Review: Feynman-Kac formula

The solution of the PDE

$$u_t + \frac{1}{2}\sigma^2(x,t)u_{xx} + \mu(x,t)u_x + h(x,t) = 0$$

with terminal condition $u(x,T) = g(x)$ has the stochastic representation

$$u(x,t) = \mathbb{E}\left[g(X_T) + \int_t^T h(X_s,s)ds \middle| X_t = x\right],$$

where X_t is the diffusion process satisfying the SDE

$$dX_t = \mu(X_t,t)dt + \sigma(X_t,t)dW_t.$$

W_t is a Brownian motion.

Feynman-Kac formula: Quick derivation

Given the PDE

$$u_t + \frac{1}{2}\sigma^2(x,t)\,u_{xx} + \mu(x,t)u_x + h(x,t) = 0,$$

we have

$$\begin{aligned}
&\mathbb{E}\left[u(x_T,T) - u(x,t)\right|\mathcal{F}_t] \\
&= \mathbb{E}\left[g(X_T)\right|\mathcal{F}_t] - u(x,t) \\
&= \mathbb{E}\left[\int_t^T du(X_s,s)\middle|\mathcal{F}_t\right] \\
&= \mathbb{E}\left[\int_t^T \partial_s u\,ds + \partial_x u\,dX_s + \partial_{x,x}u\,dX_s^2\middle|\mathcal{F}_t\right] \quad \text{(Itô's Lemma)} \\
&= \mathbb{E}\left[\int_t^T \left[\partial_s u + \mu_s\,\partial_x u + \frac{1}{2}\sigma_s^2\,\partial_{x,x}u\right]\,ds\middle|\mathcal{F}_t\right] \\
&= -\mathbb{E}\left[\int_t^T h(X_s,s)\,ds\middle|\mathcal{F}_t\right] \quad \text{(by assumption).}
\end{aligned}$$

Stochastic control problem

A stochastic control problem is a control problem which aims to maximize certain expected rewards among all admissible controls.
Specifically, consider

$$\max_{v\in\mathcal{G}[0,T]} \mathbb{E}\left[g(X_T^{(v)}) + \int_0^T h(s,X_s^{(v)},v_s)ds\right]$$

where

- g is referred to as the *terminal reward* and h is the *running reward*
- the state variable $X_t^{(v)}$ is driven by the controlled SDE $$

  ```
  d\Xv_t = \mu( t,\Xv_t, v_t) dt + \sigma(t,\Xv_t, v_t) dW_t.
  ```

  $$
- $\mathcal{G}[0,T]$ is the collection of admissible controls in the time interval $[0,T]$

Bellman's principle of optimality

(also known as the *Dynamic Programming Principle* (DPP))

"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.)

Value function

- For a given admissible control $v \in \mathcal{G}[t,T]$, define the *performance criterion* $J^{(v)}$ as

  $$

  ```
  J^{(v)}(t,x) = \E\left[\left. g(\Xv_T) + \int_0^T h(s,\Xv_s,v_s) ds \right| \cF
  _t \right].
  ```

  $$

- The value function $J(t,x)$ for a stochastic control problem maximizes the performance criterion:

  $$

  ```
  J(t,x) = \max_{v\in\cG[t,T]} J^{(v)}(t,x),
  ```

  $$

- The value function J at (t,x) is the optimal value of the control problem conditioned on the process starting at (t,x) and applying the optimal control thereafter.
- This maximization problem is typically solved by solving the Hamilton-Jacobi-Bellman (HJB) equation.

Bellman's principle again

Bellman's principle may be recast in terms of the value function as follows. For any $0 < \epsilon < T - t$,

$$J(t,x) = \max_{v \in \mathcal{G}[t,t+\epsilon]} \mathbb{E}\left[\left. \int_t^{t+\epsilon} h(s,X_s^{(v)},v_s)ds + J(t+\epsilon, X_{t+\epsilon}^{(v)}) \right| \mathcal{F}_t \right].$$

The Hamilton-Jacobi-Bellman (HJB) equation

The value function J satisfies the terminal value problem

$$\partial_t J(t,x) + \max_{v \in \mathcal{G}[t]} \{ \mathcal{L}^{(v)} J(t,x) + h(t,x,v) \} = 0, \text{ for } t < T$$

with terminal condition

$$J(T,x) = g(x).$$

The quantity

$$\mathcal{H}(t, x, \partial_x J, \partial_{xx} J) = \max_{v \in \mathcal{G}[t]} \{ \mathcal{L}^{(v)} J(t,x) + h(t,x,v) \}$$

is known as the *Hamiltonian* of the stochastic control problem.

Note

- The HJB equation is basically an infinitesimal version of Bellman's principle.
- The optimal policy (control) is given implicitly in terms of the value function J.
- The whole argument is applicable to processes with jumps as well.

Ho and Stoll setup

- The dealer quotes prices

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

- 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:

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

Ho and Stoll objective and value function

- The dealer's objective is to maximize the expected terminal utility of wealth, i.e.,

$$\max_{a,b \in \mathcal{G}} \mathbb{E}\left[U(F_T + I_T + Y_T)\right]$$

- Since there is no intermediate running reward $h(\cdot)$, the value function $J(t,f,i,y)$ reduces to

$$J(t,f,i,y) = \max_{a,b \in \mathcal{G}[t,T]} \mathbb{E}\left[U(F_T, I_T, Y_T) \middle| F_t = f, Y_t = y, I_t = i\right].$$

HJB equation for Ho and Stoll value function

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

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

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.

Avellaneda and Stoikov (2008)

Model setup

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

$$dS_t = \sigma\,dW_t$$

- N_t^A: cumulative market buy orders up to time t
- N_t^B: cumulative market sell orders up to time t
- The MB arrival rate $\lambda_A = \lambda_A(\delta^A)$ is decreasing in $\delta^A := A - S$, the distance to midprice.
- The MS arrival rate $\lambda_B = \lambda_B(\delta^B)$ is decreasing in $\delta^B := S - B$, the distance to midprice.
- Wealth in cash jumps every time there is a market order:

$$dx_t = A\,dN_t^A - B\,dN_t^B = (S_t + \delta^A)dN_t^A - (S_t - \delta^B)dN_t^B$$

- The evolution of stock inventory is given by:

$$dq_t = dN_t^B - dN_t^A$$

Note

- 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.

Objective and value function

- The objective of the dealer is to maximize the expected exponential utility

$$\max_{\delta^B,\delta^A} \mathbb{E}\left[-e^{-\alpha\,(x_T+q_T\,S_T)}\right]$$

  over certain admissible controls.
- The optimal controls δ^B and δ^A are in general time and state dependent. In fact, it is determined in terms of the value function.

- The value function of the stochastic control problem is therefore defined by

$$J(t,x,S,q)=\max_{(\delta^B,\delta^A)\in\mathcal{G}[t,T]} \mathbb{E}\left[-e^{-\alpha\,(x_T+q_T\,S_T)}\middle|\, x_t=x, S_t=S, q_t=q\right]$$

  where $\mathcal{G}[t,T]$ denotes the admissible controls in the time interval $[t,T]$.

Itô's formula for processes with jumps

Let X_t and Y_t be processes with jump. Denote by X_t^c and Y_t^c the continuous part of X and Y respectively.

If J is a smooth function, then for $t<s$, we have the following Itô's formula

$$\begin{aligned}
&J(s,x_s,S_s,q_s)-J(t,x_t,S_t,q_t)\\
&=\int_t^s J_t d\tau+\int_t^s J_x dx_\tau^c+\int_t^s J_S dS_\tau^c+\int_t^s J_q dq_\tau^c\\
&\quad+\frac{1}{2}\int_t^s J_{xx}d[x^c]_\tau+\frac{1}{2}\int_t^s J_{SS}d[S^c]_\tau+\frac{1}{2}\int_t^s J_{qq}d[q^c]_\tau+\text{covariation terms}\\
&\quad+\sum_{t<\tau\le s} J(\tau,x_\tau,S_\tau,q_\tau)-J(\tau,x_{\tau-},S_{\tau-},q_{\tau-})
\end{aligned}$$

The HJB equation

The value function $J(\cdot)$ satisfies the HJB equation

(9)

$$\begin{aligned}
&\partial_t J+\frac{\sigma^2}{2}\,\partial_S^2 J\\
&+\max_{\delta^B}\ \lambda_B(\delta^B)\ \left[J(t,x-S+\delta^B,S,q+1)-J(t,x,S,q)\right]\\
&+\max_{\delta^A}\ \lambda_A(\delta^A)\ \left[J(t,x+S+\delta^A,S,q-1)-J(t,x,S,q)\right]\\
&=0
\end{aligned}$$

with terminal condition

$$J(T,x,S,q)=-e^{-\alpha(x+qS)}.$$

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

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

- In other words, the initial wealth does not really play a role in the solution to the utility maximization problem. This turns out to be one of the features that distinguishes exponential utility from the others in this setting.
- Note that $B = S - \delta_B$ and $A = S + \delta_A$.

Note that with the ansatz we have

$$\begin{aligned}
&\partial_t J = -\alpha\, \theta_t\, J,\\
&\partial_S^2 J = \left\{-\alpha\theta_{SS} + \alpha^2\theta_S^2\right\}\, J,\\
&J(t, x+S+\delta^A, S, q-1) = J(t,x,S,q)e^{-\alpha\left\{S+\delta^A+\theta(t,s,q-1)-\theta(t,s,q)\right\}},\\
&J(t, x+S-\delta^B, S, q+1) = J(t,x,S,q)e^{-\alpha\left\{-S+\delta^B+\theta(t,s,q+1)-\theta(t,s,q)\right\}}.
\end{aligned}$$

Substitute the above into (9), the HJB equation (9) reduces to

(9')

$$\begin{aligned}
&\partial_t\theta + \frac{\sigma^2}{2}\,\partial_S^2\theta - \alpha\,\frac{\sigma^2}{2}\,(\partial_S\theta)^2\\
&+ \frac{1}{\alpha}\,\min_{\delta^B}\,\lambda_B(\delta^B)\,\left[1 - e^{-\alpha\,\left\{-S+\delta^B+\theta(t,S,q+1)-\theta(t,S,q)\right\}}\right]\\
&+ \frac{1}{\alpha}\,\min_{\delta^A}\,\lambda_A(\delta^A)\,\left[1 - e^{-\alpha\,\left\{S+\delta^A+\theta(t,S,q-1)-\theta(t,S,q)\right\}}\right] = 0
\end{aligned}$$

with terminal condition $\theta(T,S,q) = qS$.

The first order condition for the optimal bid $B = S - \delta^B$ is then

$$\partial_{\delta^B}\,\left\{\lambda_B(\delta^B)\,\left[1 - e^{-\alpha\,\left\{-S+\delta^B+\theta(t,S,q+1)+\theta(t,S,q)\right\}}\right]\right\} = 0$$

which gives the implicit relation

$$\lambda_B'(\delta^B)\,\left[1 - e^{-\alpha\,\left\{-S+\delta^B+\theta(t,S,q+1)-\theta(t,S,q)\right\}}\right] + \alpha\,\lambda_B(\delta^B)\,e^{-\alpha\,\left\{-S+\delta^B+\theta(t,S,q+1)-\theta(t,S,q)\right\}} = 0$$

which may be further simplified to

(11)

$$B = S - \delta^B = \theta(t,S,q+1) - \theta(t,S,q) - \frac{1}{\alpha}\,\log\left(1 - \alpha\,\frac{\lambda_B(\delta^B)}{\lambda_B'(\delta^B)}\right).$$

Similarly, the optimal ask price A is given by

(12)

$$A = S + \delta^A = \theta(t,S,q) - \theta(t,S,q-1) + \frac{1}{\alpha}\,\log\left(1 - \alpha\,\frac{\lambda_A(\delta^A)}{\lambda_A'(\delta_A)}\right).$$

Reservation prices

The quantities r_B and r_B defined in the following are referred to as the *reservation prices*.

- $r_B := \theta(t, S, q+1) - \theta(t, S, q)$.
- $r_A := \theta(t, S, q) - \theta(t, S, q-1)$.

One can verify that the reservation prices r_B and r_A satisfy the relationships \begin{eqnarray} *&& J(t, x - r_B, S, q+1) = J(t, x, S, q), \ && J(t, x + r_A, S, q-1) = J(t, x, S, q). \end{eqnarray}*

In other words, the reservation price is the price that the market maker is willing to trade indifferently in utility (value function) at the instant t while the current value of the stock is S and his inventory is q.

Discussion

- Equations (11) and (12) set optimal bid and ask levels as a function of inventory q, stock price S, and the arrival rate function $\lambda(\cdot)$.
- Optimal bid and ask prices are obtained in two steps:
  - Solve equations (11) and (12) for optimal δ^B and δ^A in terms of S and the function θ.
  - Substitute the optimal δ^B and δ^A back to the HJB equation (9') then solve the resulting (nonlinear) PDE for θ, which at this point does not depend on the δ's.

Exponential arrive rates

Assuming exponential arrival rates, i.e.,

$$\lambda_A(\delta^A) = K_A e^{-\mu_A \delta^A},$$
$$\lambda_B(\delta^B) = K_B e^{-\mu_B \delta^B}.$$

Note

- This is a questionable approximation because arrival rates probably depend on inventory.
  - Adverse selection again!
- However, making this approximation simplifies the analysis a lot.

Notice that in this case since

$$\frac{\lambda_A(\delta^A)}{\lambda_A'(\delta^A)} = -\frac{1}{\mu_A}, \quad \frac{\lambda_B(\delta^B)}{\lambda_B'(\delta^B)} = -\frac{1}{\mu_B}$$

the optimal bid B and the optimal ask A are given explicitly in terms of the function θ as

$$B = S - \delta^B = \theta(t, S, q+1) - \theta(t, S, q) - \frac{1}{\alpha}\log\left(1 + \frac{\alpha}{\mu_B}\right),$$
$$A = S + \delta^A = \theta(t, S, q) - \theta(t, S, q-1) + \frac{1}{\alpha}\log\left(1 + \frac{\alpha}{\mu_A}\right).$$

In particular, the optimal controls δ^B and δ^A are given by

$$\delta^B = S - \theta(t, S, q+1) + \theta(t, S, q) + \frac{1}{\alpha}\log\left(1 + \frac{\alpha}{\mu_B}\right),$$
$$\delta^A = -S + \theta(t, S, q) - \theta(t, S, q-1) + \frac{1}{\alpha}\log\left(1 + \frac{\alpha}{\mu_A}\right).$$

Further reducing HJB equation under exponential arrival rates

By substituting the optimal controls into the HJB equation (9') we obtain the following nonlinear PDE

$$\begin{aligned}&\partial_t\theta + \frac{\sigma^2}{2}\,\partial_S^2\theta - \alpha\,\frac{\sigma^2}{2}\,(\partial_S\theta)^2\\ &+ \frac{K_B}{\mu_B}\left(1+\frac{\alpha}{\mu_B}\right)^{-\left(1+\frac{\mu_B}{\alpha}\right)} e^{-\mu_B(S-\theta(t,S,q+1)+\theta(t,S,q))}\\ &+ \frac{K_A}{\mu_A}\left(1+\frac{\alpha}{\mu_A}\right)^{-\left(1+\frac{\mu_A}{\alpha}\right)} e^{-\mu_A(-S+\theta(t,S,q)-\theta(t,S,q-1))} = 0\end{aligned}$$

Symmetric exponential arrival rate

Further assume that $\mu := \mu_A = \mu_B$ and $K := K_A = K_B$, i.e., the arrivals of MB and MS orders are symmetric. The HJB equation then reduces to

$$\theta_t + \frac{\sigma^2}{2}\,\theta_{SS} - \alpha\frac{\sigma^2}{2}\,\theta_S^2 + \frac{\eta}{\mu}\left\{e^{-\mu(S-\theta(t,S,q+1)+\theta(t,S,q))} + e^{-\mu(-S+\theta(t,S,q)-\theta(t,S,q-1))}\right\} = 0,$$

where we denoted by $\eta = K\left(1+\frac{\alpha}{\mu}\right)^{-\left(1+\frac{\mu}{\alpha}\right)}$ for simplicity.

Now use the ansatz $\theta = qS + \frac{1}{\mu}\log v(t,q)$. Since

$$\theta_t = \frac{1}{\mu}\frac{\dot v}{v}, \quad \theta_S = q, \quad \theta_{SS} = 0,$$

we end up

$$\frac{1}{\mu}\frac{\dot v}{v} - \alpha\,\frac{\sigma^2}{2}\,q^2 + \frac{\eta}{\mu}\left\{\frac{v(t,q+1)}{v(t,q)} + \frac{v(t,q-1)}{v(t,q)}\right\} = 0.$$

Finally, we obtain a system of linear ODEs

$$\dot v(t,q) = \alpha\,\frac{\sigma^2}{2}\mu\,q^2 v(t,q) - \eta\,\{v(t,q+1) + v(t,q-1)\}\,.$$

with terminal condition $v(T,q) = 1!$

By restricting $-Q \le q \le Q$, for some $Q > 0$ and further imposing the boundary conditions

$$\begin{aligned}\dot v(t,Q) &= \alpha\,\frac{\sigma^2}{2}\mu\,q^2 v(t,Q) - \eta v(t,Q-1),\\ \dot v(t,-Q) &= \alpha\,\frac{\sigma^2}{2}\mu\,q^2 v(t,-Q) - \eta v(t,-Q+1),\end{aligned}$$

[Guéant, Lehalle and Fernandez-Tapia][5] showed that $J(t,x,S,q) = -e^{-\alpha(x+qS)}v^{-\frac{\alpha}{\mu}}(t,q)$ is the value function of the control problem.

Optimal bid, ask, and spread

Thus, the optimal bid B is given by

$$\begin{aligned} B = S - \delta^B &= \theta(t,S,q+1) - \theta(t,S,q) - \frac{1}{\alpha}\log\left(1+\frac{\alpha}{\mu}\right) \\ &= S - \frac{1}{\mu}\log\frac{v(t,q)}{v(t,q+1)} - \frac{1}{\alpha}\log\left(1+\frac{\alpha}{\mu}\right). \end{aligned}$$

Similary, the optimal ask A is given by

$$A = S + \delta^A = S + \frac{1}{\mu}\log\frac{v(t,q)}{v(t,q-1)} + \frac{1}{\alpha}\log\left(1+\frac{\alpha}{\mu}\right).$$

Thus, the bid-ask spread s is given by

$$s = A - B = \frac{1}{\mu}\log\frac{v^2(t,q)}{v(t,q+1)v(t,q-1)} + \frac{2}{\alpha}\log\left(1+\frac{\alpha}{\mu}\right).$$

Large time asymptotics of optimal bid, ask, and spread

- The optimal bid and ask prices obtained depends on the terminal horizon T.
- Intuitively, the closer the dealer is to time T, the less risky his inventory in stock is, since it can be liquidated at the mid price S_T.
- For a market-making business, (we hope) there should be no such final time.

By taking the limit as $T \to \infty$ in the optimal bid B and ask A, [Guéant, Lehalle and Fernandez-Tapia][5] obtained that

$$B_\infty := \lim_{T\to\infty} B = S - \frac{1}{\mu}\log\frac{v_\infty(q)}{v_\infty(q+1)} - \frac{1}{\alpha}\log\left(1+\frac{\alpha}{\mu}\right),$$

$$A_\infty := \lim_{T\to\infty} A = S + \frac{1}{\mu}\log\frac{v_\infty(q)}{v_\infty(q-1)} + \frac{1}{\alpha}\log\left(1+\frac{\alpha}{\mu}\right),$$

where $v_\infty(\cdot)$ is given by

(13)

$$v_\infty = \underset{\|u\|_2=1}{\operatorname{argmax}}\left\{\sum_{q=-Q}^{Q}\frac{\mu}{2}\alpha\sigma^2 q^2 u(q)^2 + \eta\sum_{q=-Q}^{Q-1}\{u(q+1)-u(q)\}^2 + \eta\{u(Q)^2 + u(-Q)^2\}\right\}.$$

Hence, the spread as $T \to \infty$ reads

$$s_\infty = A_\infty - B_\infty = \frac{1}{\mu}\log\frac{v_\infty^2(q)}{v_\infty(q+1)v_\infty(q-1)} + \frac{2}{\alpha}\log\left(1+\frac{\alpha}{\mu}\right).$$

Approximate asymptotic optimal bid and ask

Finally, by furhter approximate $v_\infty(\cdot)$ by the solution $\tilde{v}$ to the following variational problem

$$\tilde{v} = \underset{\|\tilde{u}\|_2=1}{\text{argmax}}\left\{\int_{-\infty}^{\infty}\left[\frac{\mu}{2}\alpha\sigma^2x^2\tilde{u}^2(x) + \eta|\tilde{u}'(x)|^2\right]dx\right\},$$

[Guéant, Lehalle and Fernandez-Tapia][4] obtain the following approximation:

$$\delta_B \approx \frac{1}{\alpha}\log\left(1+\frac{\alpha}{\mu}\right) + \left(q+\frac{1}{2}\right)\sqrt{\frac{\sigma^2\,\alpha}{2\,\mu\,\eta}},$$

$$\delta_A \approx \frac{1}{\alpha}\log\left(1+\frac{\alpha}{\mu}\right) - \left(q-\frac{1}{2}\right)\sqrt{\frac{\sigma^2\,\alpha}{2\,k\,\eta}}.$$

More realistic order arrival rates

Avellaneda and Stoikov suggest estimate the arrival rate $\lambda(\cdot)$ 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.][4]

Then

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

and recalling that first order criterion that δ^B satisfies

$$\begin{aligned}\delta^B &= S - \theta(t,S,q+1) + \theta(t,S,q) + \frac{1}{\alpha}\log\left(1 - \alpha\,\frac{\lambda_B(\delta^B)}{\lambda'_B(\delta^B)}\right)\\ &= S - \theta(t,S,q+1) + \theta(t,S,q) + \frac{1}{\alpha}\log\left(1 + \alpha\,\frac{\delta_B}{3}\right)\\ &\approx S - \theta(t,S,q+1) + \theta(t,S,q) + \frac{\delta_B}{3}.\end{aligned}$$

Thus, the optimal δ^B is approximately

$$\delta^B \approx \frac{3}{2}\left\{S - \theta(t,S,q+1) + \theta(t,S,q)\right\}.$$

Similarly,

$$\delta^A \approx \frac{3}{2}\left\{\theta(t,S,q) - \theta(t,S,q-1) - S\right\}.$$

Next we would repeat the whole procedure as in the case of exponential arrival rate in order to find the optimal prices. However, unfortunately in this case a closed-form expression as neat as in the exponential arrival case is not accessible. One has to resort to numerical methods for solving the resulting equation.

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.

Infinite horizon objective

Alternative in the setting of infinite horizon, Avellaneda and Stoikov suggest considering an expected discounted cumulative utility of the form

$$\max_{\delta^A,\delta^B} \mathbb{E}\left[-\int_0^{\infty} e^{-\omega\, t}\, e^{-\alpha\, (x_t+q_t\, S_t)}\, dt\right].$$

- The value function U in this case is hence given by

$$U(x,S,q) = \max_{\delta^A,\delta^B} \mathbb{E}\left[-\int_0^{\infty} e^{-\omega\, t}\, e^{-\alpha\, (x_t+q_t\, S_t)}\, dt \middle| x_0 = x, S_0 = S, q_0 = q\right].$$

- The value function U satisfies the following HJB equation

$$\begin{aligned}
&-\omega U + \frac{\sigma^2}{2}\,\partial_S^2 U \\
&+ \max_{\delta^B}\ \lambda_B(\delta^B)\ \left[U(t,x-S+\delta^B,S,q+1) - U(t,x,S,q)\right] \\
&+ \max_{\delta^A}\ \lambda_A(\delta^A)\ \left[U(t,x+S+\delta^A,S,q-1) - U(t,x,S,q)\right] \\
&= 0
\end{aligned}$$

  with the *transversality condition*

$$\lim_{T\to\infty} e^{-\omega T}\,\mathbb{E}\,[|U(x_T,S_T,q_T)|] = 0$$

  for any admissible control δ^B and δ^A.
- Likewise, we may apply the ansatz $U(x,S,q) = e^{-\alpha(x+\theta(S,q))}$ to reduce the HJB equation.

Adding market orders

- The frameworks of both [Avellaneda and Stoikov][1] and [Guéant, Lehalle and Fernandez-Tapia][4] 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][6], 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
  - Place constant quantity on each side at random prices that are at the best quote or improve the spread.

Guilbaud and Pham simulation results

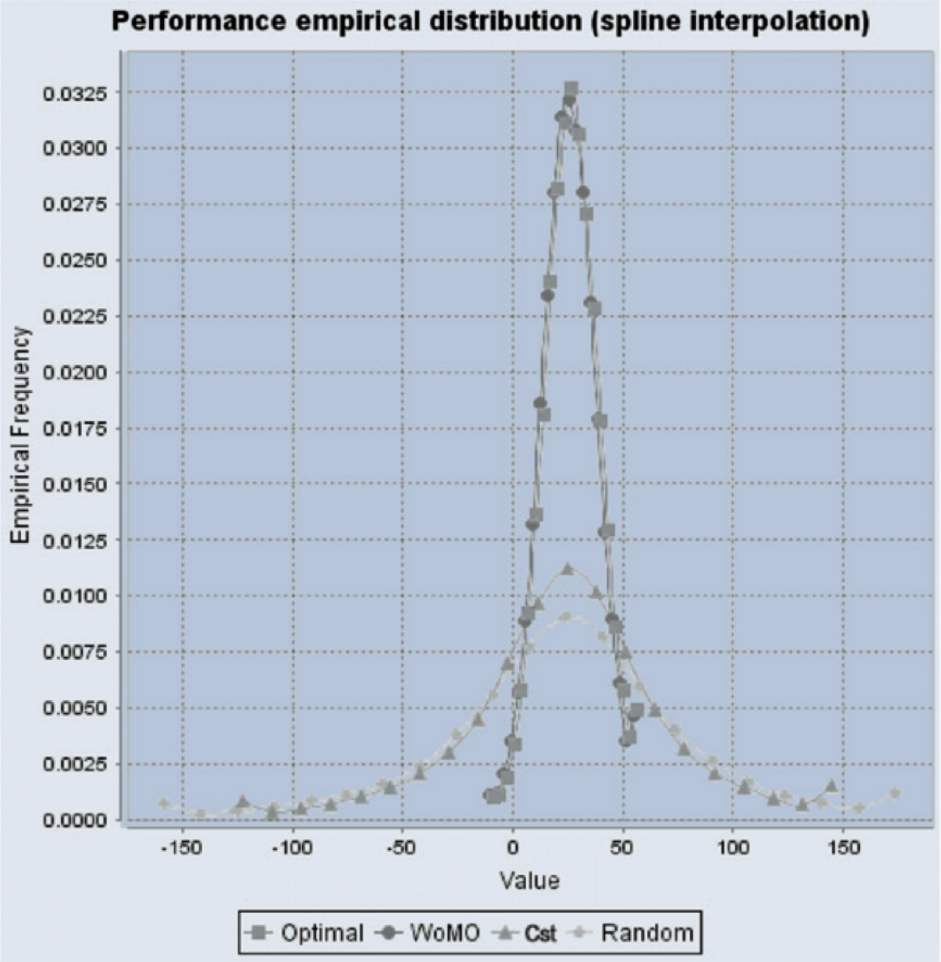

Figure 4. Empirical distribution of the terminal wealth X_T (spline interpolation).

Guilbaud and Pham optimal frontier

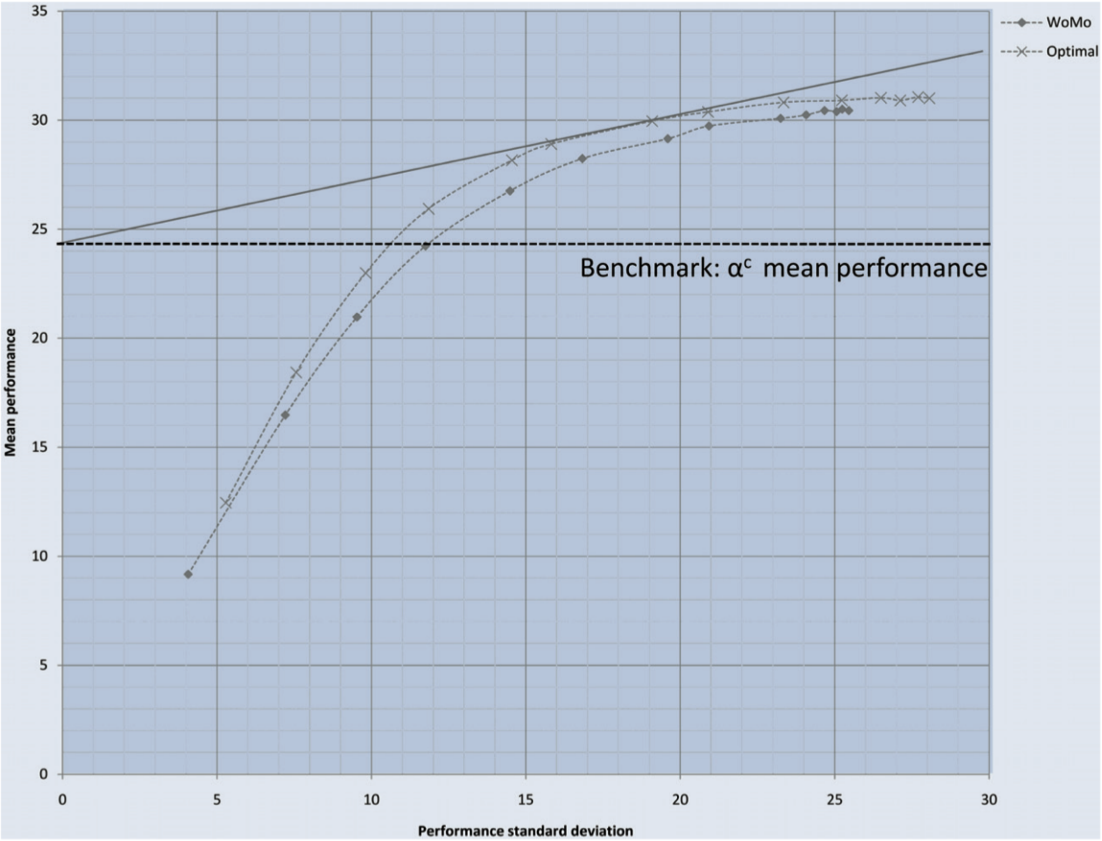

Figure 6. Efficient frontier plot.

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 market impact
    - Market sells cause the price to decrease.
    - Market buys cause the price to increase.
  - The spread is compensation for risk.
    - The spread is increasing in volatility and in the price of risk.

- In real markets, as in [Guilbaud and Pham][6], as in the case of big tick stocks, the spread is given.
  - A market maker either joins or improves the best quote, or does no business.
- Market order arrival rates are not symmetric: they depend on the book imbalance $\mathcal{I}$.
  - [Cartea, Donnelly and Jaimungal][2] solve an optimal control problem to find the optimal placement of limit orders using the book imbalance.

References

1. ^ Marco Avellaneda and Sasha Stoikov, High-frequency trading in a limit order book, *Quantitative Finance* **8**(3), 217–224, (2008).
2. ^ Álvaro Cartea, Ryan Donnelly and Sebastian Jaimungal, Enhancing trading strategies with order book signals, *SSRN* available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2668277 (2015).
3. ^ Hans Föllmer and Alexander Schied, *Stochastic Finance: An Introduction in Discrete Time*, De Gruyter, Berlin (2011).
4. ^ 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).
5. ^ 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).
6. ^ Fabien Guilbaud and Huyên Pham, Optimal high-frequency trading with limit and market orders, *Quantitative Finance* **13**(1), 79-94, (2013).
7. ^ Joel Hasbrouck, *Empirical Market Microstructure*, Oxford University Press, Chapter 11, (2007).
8. ^ Albert J Menkveld, High frequency trading and the new-market makers, *Journal of Financial Markets* **16**(4), 712-740, (2013).