

A Pre-Trade Algorithmic Trading Model under Given Volume Measures and Generic Price Dynamics (GVM-GPD)

Jackie (Jianhong) Shen ¹

*Industrial and Enterprise Systems Engineering (ISE)
and Master of Science in Financial Engineering (MSFE)
216D Transportation Building, University of Illinois, Urbana, IL 61801*

Dedicated to Gil Strang on the Occasion of His 80th Birthday.

Abstract

We make several improvements to the mean-variance framework for optimal pre-trade algorithmic execution, by working with volume measures and generic price dynamics. Volume measures are the continuum analogies for discrete *volume profiles* commonly implemented in the execution industry. Execution then becomes an *absolutely continuous* measure over such a measure space, and its Radon-Nikodym derivative is commonly known as the *Participation of Volume* (PoV) function. The four impact cost components are all consistently built upon the PoV function. Some novel efforts are made for these linear impact models by having market signals more properly expressed. For the opportunistic cost, we are able to go beyond the conventional Brownian-type motions. By working directly with the auto-covariances of the price dynamics, we remove the Markovian restriction associated with Brownians and thus allow potential memory effects in the price dynamics. In combination, the final execution model becomes a constrained quadratic programming problem in infinite-dimensional Hilbert spaces. Important linear constraints such as participation capping are all permissible. Uniqueness and existence of optimal solutions are established via the theory of positive compact operators in Hilbert spaces. Several typical numerical examples explain both the behavior and versatility of the model.

Keywords: Volume, Price, Impact, Risk, Compact positive operator, Hilbert space, Existence, Uniqueness, Quadratic Programming

Email address: jshen@illinois.edu (Jackie (Jianhong) Shen)

URL: <http://publish.illinois.edu/jackieshen> (Jackie (Jianhong) Shen)

¹Jackie Shen had been a Vice President for the Equities Algorithmic Trading divisions at both J.P. Morgan and Barclays, New York, before joining the University of Illinois in 2013. Readers are encouraged to first read the disclaimer statements near the end.

1. Introduction

Algorithmic trading helps institutional investors liquidate or acquire big positions without incurring much adverse impact cost or opportunistic risk. To achieve this objective, a practical model must keep screening any real-time market characteristics and adapt the execution strategy accordingly. This necessarily means that a realistic trading model has to be *dynamic*, as in the seminal work of Bertsimas and Lo [3], and many other important ones (e.g., Huberman and Stanzl [8], Almgren [1], just to name a few).

On the other hand, both internal or external clients commonly rely on computationally less expensive *static* models to get pre-trade estimations on potential costs and risks for their positions. They may also rely on such pre-trade estimations to generate alerts if the actual execution is deviating too far, often expressed via confidence intervals, and needs immediate human intervention. Furthermore, due to the mounting computational challenges associated with fully dynamic algorithms, execution houses (e.g., broker-dealers, agency services, internal execution teams in hedge funds, etc.) also typically utilize *static algorithms* as the core for building heuristic but much faster dynamic models.

It is for these reasons that the study of utility-function based *static* algorithms is still very valuable in the execution industry. The current work focuses on some major improvements of the mean-variance framework, following largely the classical works of Almgren and Chriss [2], Huberman and Stanzl [8], and other closely related works (e.g., Kissell and Malamut [9], and Obizhaeva and Wang [10]).

The core of the current work is a single continuum model independent of any interval based discrete grids. This does not mean that any in-house implementation has to avoid interval-based setting. Instead, a single governing continuum model has numerous advantages, including for example, (i) ensuring consistency among different interval choices, (ii) staying invariant when technology advancements (e.g., on data servers, optimizer servers, or Direct Market Access (DMA)) allow executing in higher and higher frequencies, and (iii) welcoming more computational methodologies, including for example, basis function based methods that are free of interval grids. Imaging Science, for instance, has witnessed a blossoming decade very much thanks to the similar advantages as one goes from pixel or graph based discrete models to continuum models, as the latter ones are independent of camera resolutions and also befriend a wealth of mathematical tools such as PDE, variational calculus, and wavelets (e.g., Chan and Shen [4]).

The two major characteristics of the proposed model are, as the title has suggested, (i) treating market volumes as Borel measures over the execution horizon, and (ii) permitting generic price dynamics (including generic Bid-Ask spread dynamics), other than being restricted to the linear or geometric Brownians prevailing in the literature.

Generic price dynamics extend beyond the Markovian nature of conventional Brownian-type price movements, and allow short-term or long-term memories. This will be explained in more details in Section 2. Two more specific examples involving mean reversion and stochastic volatilities are also presented in

Section 6 to illustrate the flexibility of the model.

In the typical reductionism approach to static modeling of algorithmic trading [2, 10], the markets have usually been approximated by the volume profile (instead of more complex signals associated with the dynamics of *limit order books* (LOB)). Volume measures generalize discrete *volume profiles* to the continuum setting. We attempt to demonstrate that measure theory and functional analysis serve a natural foundation for the continuum modeling. Section 3 lays out the general setting for volume measures.

The proposed execution model is built upon four impact cost components, which are all *linear*. This is the third characteristics of the current work. Linear cost models allow faster calibration via linear regression techniques. They also lead to *Quadratic Programming* (QP) formulation for the final execution model, which can be readily solved via robust commercial QP solvers (e.g., the IBM CPLEX). With volume measures, all the four impact cost components are consistently build upon the *Participation of Volume* (PoV) rate function, which is the Radon-Nikodym derivative of the execution measure against the market volume measure. For both the transient and permanent costs, we have introduced volume distance for the transient impact and cumulative volume denomination for the permanent impact. These efforts are novel to the best knowledge of the author, and make the cost models more realistic. Throughout the model building process, we have also particularly emphasized:

- (a) the influence of in-house *Child Order Placement* strategies on modeling parameters, and
- (b) the role of quantitative market makers on shaping the exact forms of impact models.

All these subjects will be further elaborated in Section 4.

The rest of the paper is devoted to the analysis and computation of the established model. In Section 5, we apply the theory of positive compact operators in Hilbert Spaces to establish the existence and uniqueness of solutions to the resulted QP problem with a quadratic objective and linear constraints. A computational scheme is then proposed in Section 6, and several numerical examples are also presented to reveal the effects of various factors, including degrees of risk aversion, shapes of volume measures, dominance of individual cost components, and non-Brownian price dynamics. A feature that is also somewhat novel in the proposed model is the explicit permission of volume capping, which is popularly demanded from both internal and external clients.

2. Generic Price Dynamics

Regardless of execution styles, trading is always exposed to the overall market movements. This is especially true for large institutional orders for which immediate filling via open exchanges is infeasible. In the current work we will not touch the subject of *dark pools* where it is *not* impossible to have a large order filled via internal crossing networks run by broker-dealers, agency services, or even exchanges.

This part of the trading cost is often termed the *opportunistic cost or risk*, and is directly caused by the innate price fluctuations of the target securities. Trading models have to incorporate suitable price dynamics in order to quantify the associated opportunistic cost.

Let p_0 denote the arrival price at time $t = 0$ on the normalized trading horizon $I = [0, 1]$. Practically, this could be the mid-quote of the *National Best Bid and Offer* (NBBO). The mid-quote of the NBBO has been traditionally considered as the *true* price of the security [10], an assumption we shall follow as well. Let p_t or $p(t)$ denote the mid-quote at time $t \in I$. Define the price fluctuation to be $\delta_t = p_t - p_0$, or its homogenized version to be: $\delta_t^\circ = \delta_t/p_0$. It has been commonly assumed in the literature that without impact costs, δ_t is either linearly or geometrically Brownian [2, 10], i.e.,

$$d\delta_t = \hat{\mu}dt + \hat{\sigma}dW_t, \quad \text{or} \quad d\delta_t/p_t = \mu dt + \sigma dW_t.$$

The model in the current work does not specifically address the overnight risks, and is primarily designed for intraday executions. For the intraday horizon, the geometric Brownian can be well approximated by the linear substitute:

$$d\delta_t/p_0 = d\delta_t^\circ = \mu dt + \sigma dW_t.$$

This works well for most *non-penny* common stocks whose prices are sufficiently positive (e.g., above \$5.00), and intraday price changes are no more than a few percentage points so that the dynamic denominator p_t can be well approximated by the static reference price p_0 .

The current work, however, makes no assumptions on the specific format of the price dynamics. Instead, we only rely on the auto-covariance function: for any $t, s \in I = [0, 1]$,

$$K_\delta^\circ(t, s) = \text{E}[\delta_t^\circ \cdot \delta_s^\circ], \quad \text{or} \quad K_\delta(t, s) = \text{E}[\delta_t \cdot \delta_s] \quad (= p_0^2 K_\delta^\circ(t, s)).$$

For the intraday horizon, empirical evidences show that the drifting component μ in the Brownian framework can be assumed zero. We shall also assume that δ_t (or δ_t°) is a zero-mean process.

For the *kernel* function $K_\delta(t, s)$, the following natural assumptions are to be made:

- (i) (Symmetry) $K_\delta(t, s) = K_\delta(s, t)$, for any $t, s \in I = [0, 1]$.
- (ii) (Positivity) for any *finite* subset $\Lambda \subset I$, and any real function on the set:

$$a : \Lambda \rightarrow (-\infty, +\infty), \quad t \rightarrow a_t,$$

one always has:

$$\sum_{t, s \in \Lambda} K_\delta(t, s) a_t a_s \geq 0.$$

Conventional Brownian models always satisfy these conditions with the kernel:

$$K_\delta(t, s) = p_0^2 \cdot \sigma^2 \cdot \min(t, s) = p_0^2 \cdot \sigma^2 \cdot (t \wedge s).$$

(For convenience, we denote $\min(t, s)$ by $t \wedge s$, and $\max(t, s)$ by $t \vee s$.) Furthermore, by working with the auto-covariance kernel alone, the underlying price dynamics do not need to be Markovian or memoryless, which provides much more flexibility in applications. In Section 6, we have implemented two non-Brownian examples whose price dynamics are governed by mean reversion or stochastic volatility.

3. Volume Measures and PoV

Besides the extension beyond the conventional Brownian-type price dynamics, another major methodological innovation in the current work is we assume that the market volume distribution dV over the execution horizon $t \in I = [0, 1]$ is a finite and positive Borel measure.

Let dt denote the traditional Lebesgue measure over the horizon I . If the volume measure is Lebesgue *absolutely continuous*, the market execution *speed* $v(t)$ is then well defined to be the Radon-Nikodym derivative:

$$v(t) = \frac{dV}{dt}, \quad \text{and} \quad v(t) \in L^1(I) = L^1(I, dt).$$

Generally, embedded within a Borel volume measure could also be *atomic* measures and *continuous* singular measures. In particular, the latter means that market volumes could grow *stealthily* on a set of times whose total Lebesgue measure is zero. This could be particularly useful when studying the joint effects of *lit and dark* pools, though we will not expand this topic in the present work.

Suppose a client intends to execute X_1 shares of some security, over a normalized time interval of $I = [0, 1]$. Set X_1 positive if it is a buy to open a long position or to close an existing short position, and negative if it is a sell to liquidate an existing long position or to open a new short position. Since the current model assumes symmetry between buys and sells, from now on we shall only work with a *buy* order as the default setting, unless stated otherwise.

The *pre-trade* execution is then considered to be a Borel measure dX over the horizon I . And the following basic assumptions are made throughout the work.

- (1) (Monotone) dX is a positive measure for buys (and negative for sells). This is a basic request from clients who are generally against opportunistic selling for a buy order, and vice versa. This implies that the cumulative shares X_t or $X(t)$ always change monotonically from 0 to X_1 .
- (2) (Completion) $\int_I dX = X_1$, i.e., the targeted order is entirely filled during the horizon.
- (3) (Absolute Continuity) We assume that dX is *absolutely continuous* with respect to the market volume measure dV , so that the *Participation of Volume* (PoV) rate function $h(t)$ is the Radon-Nikodym derivative:

$$h(t) = \frac{dX}{dV}, \quad \text{and} \quad h \in L^1(I, dV).$$

The monotone assumption now simply states that h never changes sign (almost every (a.e.) with respect to dV). Similarly, the completion condition becomes

$$\int_I h(t)dV(t) = X_1.$$

Recall that in measure theory [6, 7], the absolute continuity condition is equivalent to requiring that on any time subset $\Gamma \subseteq I$, $V(\Gamma) = \int_{\Gamma} dV = 0$ would imply $X(\Gamma) = \int_{\Gamma} dX = 0$. We call this the *fundamental principle of trading*. In the reductionism approach the market volume alone represents the entire market. Traders therefore would naturally avoid trading whenever there are no market activities.

In what follows, we shall consistently build all the cost and risk components upon the PoV rate function $h(t)$ or h_t , which then becomes the decision variable to optimize with.

4. Components of Linear Impact Models

The impact model in the current work includes four pieces of components: spread cost, instantaneous cost, transient cost, and permanent cost. The first two components contribute to realized impact cost, but leave no trailing imprints on market prices, while the latter two do. All components are not unfamiliar in the literature (e.g., [3, 2]), but some innovation efforts have been made:

- (i) For the spread cost, stochastic spreads are allowed, and the component coefficient is explicitly linked to the *Child Order Placement* of each execution house.
- (ii) For the transient cost, we rely on a quantitative market-maker mechanism to build a volume-distance based exponential transition model. This effort is very much inspired by the temporal exponential resilience of Obizhaeva and Wang [10], whose driving market force has however not been elucidated.
- (iii) For the permanent cost, we employ the look-back PoV as the dependent variable, instead of the cumulative trading volume commonly used in the literature. This allows to differentiate the permanent costs arising from different market volume environments, e.g., those due to trading 1000 shares when the market volume is 5,000 shares versus when the market volume is 20,000 shares.

Furthermore, in the current work, we shall also stick to *linear* models for all the components, which result in a quadratic programming problem in Hilbert Spaces for the ultimate trading model. Another advantage of linearity is that model calibration could directly rely upon linear regression (see Section 6.1), instead of nonlinear solvers which are slower or with no guarantee on uniqueness.

Spread Cost: IC-sprd

Let θ_t denote the NBBO spread at time t (in dollar amount), and $\theta_t^\circ = \theta_t/p_0$ the homogenized spread with respect to the reference arrival price. That is,

$$\theta_t = \text{NBO}_t - \text{NBB}_t,$$

the gap between the national best offer (NBO) and best bid (NBB). For instance, for the S&P-500 (large-cap) universe, the median homogenized spread is about 5.5 basis points, i.e., θ_t° roughly fluctuates around $5.5/10,000 = 0.00055$. For the S&P-400 (mid-cap) universe, it roughly doubles.

Since the mid-quote of the NBBO is considered as the *true* price, as one crosses the spread to consume market liquidity, one has to pay at least the half spread. More generally, we denote the signed spread cost (per share) by:

$$\text{IC-sprd}_t = \alpha_0 \cdot \text{sign}(h_t) \cdot \theta_t.$$

Here the sign of the PoV indicates to pay higher for buying and receive lower for selling. Typically the coefficient α_0 is somewhere between 0.0 and 0.5. Regression results from real trades in some of the execution houses where the author has worked have confirmed this general behavior.

There is practical reason why in reality α_0 is less than 0.5 and clients on average pay less than the half spread. Each trading model is typically implemented on a discrete time grid with interval length equal to, say, 1 minute or 5 minutes (often pegged to the liquidity level of the target security). In the modern environment of high-frequency technology, each interval period provides sufficient time to make smart *Child Order Placement* (COP). This typically involves a mixture of *limit* and *market* orders. In reality, COP could also access internal or external non-displayed crossing networks, or the *dark pools*. Successful limit orders eliminate the need to cross the half spread to take liquidity as market orders do, and actually gain the half spread. Most dark orders, on the other hand, are usually pegged to the mid price of the NBBO and are thus traded at the *true* price without spread gain or loss. In combination, realized spread cost is typically less than the half spread.

Therefore, the coefficient α_0 partially reflects the underlying COP strategy. Different houses may observe different α_0 due to different COP strategies that are tailored to their accessible trading venues and proprietary anti-gaming strategies.

To proceed, we further assume that θ_t has a mean process $\bar{\theta}_t$, and auto-covariance kernel K_θ :

$$K_\theta(s, t) = \text{COV}(\theta_s, \theta_t) = \text{E} [(\theta_s - \bar{\theta}_s)(\theta_t - \bar{\theta}_t)].$$

We also assume that the NBO (or NBB) is the superposition of two *independent* processes:

$$\text{NBO}_t = p_t + \theta_t/2.0 = p_0 + \delta_t + \theta_t/2.0.$$

The independence is a non-essential technical assumption, and one could otherwise work with the joint auto-covariance kernel for (δ_t, θ_t) in later sections.

Instantaneous Cost IC-inst

Instantaneous cost represents the immediacy cost when execution consumes liquidity from the limit order books. In the reductionism approach when the market volume summarizes the overall market, market volume has to be somewhat proportional to the capacity or depth of the limit order book. This heuristic reasoning implies that bigger participation of volume (PoV) would efface deeper the limit order book and hence retreat the NBB more significantly (or NBO for sells). We therefore model this part of instantaneous cost (per share) by:

$$\text{IC-inst}_t = \alpha_1 \cdot h_t, \quad \text{where } h_t = dX/dV \text{ is the PoV rate.}$$

Here the coefficient α_1 is in dollar and can be homogenized to a dimensionless one via:

$$\alpha_1 = \alpha_1^\circ \cdot p_0.$$

Similar to the discussion for the spread cost, in practice α_1 must also depend on the underlying COP strategies and different execution houses may obtain different α_1 when regressing against their own real trading data.

Transient Cost IC-tran

Transient cost represents the short-term market digestion of the most recent wave of trades. In the literature, this component is also called the *resilience* term. For example, in the important work of Obizhaeva and Wang [10], the authors attribute this term to the resilience of the limit order book.

What actually drives such a resilient force, however, has not been made very clear in the literature. Herein, we attribute it to the activities of quantitative market makers. A market maker adjusts her quoting levels based on the expected market behavior. As a common practice, the expected behavior is often based on the moving average of the observable. We assume that at any time t , a typical quantitative market maker is primarily interested in the expected PoV via the moving average of the observed PoV's:

$$\bar{h}_t = \int_0^t h_s dM(s|t),$$

where $dM(s|t)$ is a backward moving average kernel. (Borrowed from probability theory, the vertical bar $(x|y)$ denotes variable or value x given y .) As a moving average kernel, one requires $dM(s|t)$ to be a probability measure over $s \in [0, t]$:

$$\int_{s \in [0, t]} dM(s|t) = 1.0, \quad \text{and} \quad M(A|t) = \int_{s \in A} dM(s|t) \geq 0,$$

for any Borel set $A \subseteq [0, t]$. The most popular kernel used either in academia or in practical quantitative trading firms is the *exponential* kernel (e.g., [10]):

$$dM(s|t) = \frac{1}{Z_t} \exp(-\gamma(t-s)), \quad 0 \leq s \leq t.$$

(The unital condition is often not strictly enforced, so that, for example, $Z_t \equiv \gamma^{-1}$ or any suitable constant [10].)

Based on the expected PoV \bar{h}_t , the quantitative market maker then raises the quotes in an amount proportional to \bar{h}_t :

$$\text{IC-tran}_t = \alpha_2 \cdot \bar{h}_t = \alpha_2 \int_0^t h_s \cdot dM(s|t)$$

Any incoming trade on $[t, t + dt)$ has to pay this premium cost.

In this work, we assume that dM is *absolutely continuous* under the volume measure dV , and take the exponential decay kernel under volume distance as its Radon-Nikodym derivative:

$$dM(s|t) = \frac{1}{Z_t} \exp\left(-\frac{V_t - V_s}{V_*}\right) dV_s,$$

with the normalization constant Z_t given by:

$$Z_t = \int_0^t \exp\left(-\frac{V_t - V_s}{V_*}\right) dV_s = V_* \cdot (1 - \exp(-V_t/V_*)).$$

Here the scaling constant V_* reflects the market marker's volume-based *window* size for averaging short-term PoV's. For example, a market maker could take $V_* = 1\% \cdot ADV$, the *average daily volume* of the target security. (In the United States where the normal daily core session lasts 390 minutes, this is roughly the average trading volumes over several minutes.)

Since for the majority of time, $V_t \gg V_*$, Z_t is quickly saturated to the level of V_* . For the rest of the work we will make the following simplification by directly setting:

$$Z_t \equiv V_*.$$

This constancy approximation is also the default setting for most existing works involving the *temporal* exponential kernel [10]. Analytically it helps alleviate the hardship involving a nonlinear exponential denominator.

Permanent Cost: IC-perm

By definition, as time elapses and market volume grows, the transition cost imposed by a specific trade at some past time s will fade away. The permanent cost component then captures any permanent impact a trade might have left. In the current work, we assume it is proportional to the *volume-weighted* average PoV:

$$\text{IC-perm}_t = \alpha_3 \int_0^t h_s \frac{dV_s}{V_t} = \alpha_3 \frac{X_t}{V_t}.$$

Most existing linear models only assume the proportionality to the total executed volume X_t (e.g., [2, 10]). The volume normalization introduced herein captures the difference between the impact of 5000 shares, say, executed in a period of cumulative market volume $V_t = 20,000$ shares vs. that of the same

number of shares in another period of $V_t = 200,000$ shares. The permanent impacts are clearly different.

In theory the denominator V_t vanishes in the beginning of the execution, as it is the cumulative market volume from the start at $t = 0$. This could potentially introduce some singularity when later on the total execution cost is assembled. In the scenario of *flat* profiling when $V_t = v_0 t$ for some constant market rate v_0 , for instance, the singularity is in the order of $O(1/t)$.

In practice, quantitative traders resolve such singularities through at least two general approaches. The first one is to slightly delay the computation in order to get stronger signals (with a higher *signal-to-noise ratio* (SNR)). For instance, many high-frequency strategies on the buy side will not open to trade within the first 5 to 30 minutes of the market opens, unless they specifically target at the open auctions or opening moments. This “quiet” period allows to collect *stable* signals based on moving averages. The other approach is to *regularize* the target signals. In our scenario, for example, to “boost” V_t to $V_t + \varepsilon_0$ for some thresholding volume level ε_0 , so that the signal V_t only becomes effective when $V_t \gg \varepsilon_0$. In practice, as discussed for the transient cost, ε_0 could be $1\% \times ADV$, one minute average ADV , or even a few or several round lots depending on the liquidity level of the security. In the continuum setting, we use the ε symbol to convey a sense of minuteness, as common in mathematics. As such, we finalize the permanent cost term via:

$$\text{IC-perm}_t = \alpha_3 \int_0^t h_s \frac{dV_s}{V_t + \varepsilon_0} = \alpha_3 \frac{X_t}{V_t + \varepsilon_0}.$$

We point out that all the subsequent analysis also works for “hard-thresholding” when one takes:

$$\alpha_3 \int_0^t h_s \frac{dV_s}{V_t \vee \varepsilon_0} = \alpha_3 \frac{X_t}{V_t \vee \varepsilon_0} \quad \text{instead,}$$

where $V_t \vee \varepsilon_0 = \max(V_t, \varepsilon_0)$ more explicitly regularizes V_t near $t = 0$.

In actual discrete implementation, such regularization may not be necessary since V_{t_1} is generally positive at the first grid point after $t = t_1 > 0$.

Putting Together

Combining all the components, we arrive at a model for the *execution* price \hat{p}_t . Let p_t denote the *market* price, which could be considered as the mid-quote of the NBBO at any time t . Then,

$$p_t = p_0 + \delta_t + \text{IC-tran}_t + \text{IC-perm}_t; \quad (1)$$

$$\hat{p}_t = p_t + \text{IC-sprd}_t + \text{IC-inst}_t. \quad (2)$$

Here, δ_t represents the intrinsic market price movement, regardless of “our” own trading activity. As discussed earlier, we only assume that its auto-covariance kernel $K_\delta(s, t)$ is known pre-trade.

5. The Execution Model and Analysis

The *implementation shortfall* (IS) is the amount overpaid (in the default scenario of buying) compared with the initial *paper* cost when the security price is p_0 , as first introduced by Perold [11]. In dollar amount, for any deterministic execution scheme : $X_t : 0 \leq t \leq 1$, it is defined by:

$$\begin{aligned}
\text{IS}_{\mathfrak{s}} &= \int_0^1 (\hat{p}_t - p_0) \cdot dX_t \\
&= \int_0^1 p_t \cdot dX_t + \int_0^1 \text{IC-sprd}_t \cdot dX_t + \int_0^1 \text{IC-inst}_t \cdot dX_t \\
&= \int_0^1 \delta_t \cdot dX_t + \int_0^1 \text{IC-sprd}_t \cdot dX_t + \int_0^1 \text{IC-inst}_t \cdot dX_t \\
&\quad + \int_0^1 \text{IC-tran}_t \cdot dX_t + \int_0^1 \text{IC-perm}_t \cdot dX_t \\
&= \int_0^1 (\delta_t + \alpha_0 \theta_t \cdot \text{sign}(h_t)) \cdot dX_t \\
&\quad + \int_0^1 (\text{IC-inst}_t + \text{IC-tran}_t + \text{IC-perm}_t) \cdot dX_t.
\end{aligned}$$

Due to the monotone condition discussed earlier, $\text{sign}(h_t)$ is static and global since it is always 1 for a buy order and -1 for a sell order. Therefore, we will substitute it with a static variable more commonly used: *side*, which is 1 for a buy and -1 for a sell. Notice that $\text{side} \cdot dX_t = |dX_t|$. To prepare for the mean-variance formulation, we first compute the mean $\text{E}[\text{IS}_{\mathfrak{s}}]$, which is:

$$\int_0^1 \left(\alpha_0 \bar{\theta}_t \cdot \text{side} + \alpha_1 h_t + \alpha_2 \int_0^t h_s dM(s|t) + \alpha_3 \int_0^t h_s \frac{dV_s}{V_t + \varepsilon_0} \right) \cdot dX_t. \quad (3)$$

The variance is

$$\text{VAR}[\text{IS}_{\mathfrak{s}}] = \int_0^1 \int_0^1 K_{\delta, \theta}(s, t) dX_s dX_t, \quad \text{with} \quad (4)$$

$$K_{\delta, \theta}(s, t) = K_{\delta}(s, t) + \alpha_0^2 K_{\theta}(s, t). \quad (5)$$

For any given level of risk aversion expressed via a positive weighting parameter $\lambda > 0$, the mean-variance execution model is to find the optimal PoV function \hat{h}_t , such that $d\hat{X}_t = \hat{h}_t dV_t$ and the following utility functional is minimized:

$$J_{\lambda}[h_t] = \text{E}[\text{IS}_{\mathfrak{s}}] + \lambda \cdot \text{VAR}[\text{IS}_{\mathfrak{s}}]. \quad (6)$$

Under the given volume measure dV_t , the objective functional becomes:

$$\begin{aligned}
J_\lambda[h_t] = & \alpha_0 \cdot \text{side} \cdot \int_0^1 \bar{\theta}_t h_t dV_t + \alpha_1 \int_0^1 h_t^2 dV_t \\
& + \frac{\alpha_2}{2V_*} \int_0^1 \int_0^1 h_s h_t \exp\left(-\frac{|V_t - V_s|}{V_*}\right) dV_s dV_t \\
& + \frac{\alpha_3}{2} \int_0^1 \int_0^1 h_s h_t \frac{1}{V_t \vee V_s + \varepsilon_0} dV_s dV_t \\
& + \lambda \int_0^1 \int_0^1 K_{\delta, \theta}(s, t) h_s h_t dV_s dV_t,
\end{aligned} \tag{7}$$

with at least two common constraints discussed in previous sections,

$$\text{side} \cdot h_t \geq 0, \quad \text{and} \quad \int_0^1 h_t dV_t = X_1, \tag{8}$$

for monotonicity and completion. Notice also that $\text{side} = \text{sign}(X_1)$.

According to the published marketing or sales sheets, major execution houses always allow their clients to specify the PoV capping, i.e., the linear inequality constraint:

$$\text{side} \cdot h_t \leq \max\text{PoV}. \tag{9}$$

For most clients, the popular comfort zone for $\max\text{PoV}$ is somewhere between 5% and 25%. If this constraint is turned on, there is an obvious compatibility condition for the model to yield a solution:

$$\int_0^1 \max\text{PoV} \cdot dV_t \geq \text{side} \cdot X_1, \quad \text{or} \quad \max\text{PoV} \geq \frac{|X_1|}{V_1}. \tag{10}$$

Otherwise, the execution may not be complete within the specified horizon. Except for very urgent or small orders, clients usually turn on this constraint as an ultimate safeguard.

5.1. Basic Assumptions on the Volume Measure

Since both the transient and permanent cost terms explicitly involve the cumulative market volume function:

$$V_t = \int_{[0, t]} dV_s,$$

we will make the following assumptions about the regularity of the volume measure.

- (i) (Absolute Continuity) We assume that dV is *absolutely continuous* with respect to the ordinary Lebesgue measure dt . This amounts to saying that there exists a non-negative function $v_t \in L^1(I) = L^1(I, dt)$, such that $dV_t = v_t dt$. Then the cumulative market volume function

$$V_t = \int_{[0, t]} v_t \cdot dt,$$

is well defined *pointwise*, and is continuous and non-decreasing. We call v_t the *market (trading) rate function*.

- (ii) (Lipschitz) Furthermore, we assume that the market rate function is bounded from above, i.e., there exists some constant $A > 0$, such that

$$v_t \leq A, \quad \text{almost everywhere } t \in I = [0, 1] \text{ under } dt.$$

Equivalently, we say that $v_t \in L^\infty(I)$. In terms of the cumulative market, it is equivalent to requiring V_t to be Lipschitz [6, 7].

Practically these are natural assumptions to most execution houses, where volume profiles are generated by overnight processes. One common component of these processes is a smoothing kernel to curb the effect of spurious spikes arising from direct historical averaging.

5.2. Convexity and Uniqueness

Define the kernel functions:

$$K_2(s, t) = \frac{1}{V_*} \exp\left(-\frac{|V_t - V_s|}{V_*}\right), \quad \text{and} \quad K_3(s, t) = \frac{1}{V_t \vee V_s + \varepsilon_0}.$$

Then the objective functional $J_\lambda[h]$ involves three quadratic terms in the general form of

$$Q_{K,dV}[h] = \int \int_{I \times I} K(s, t) h(s) h(t) dV(s) dV(t),$$

with $K = K_2, K_3$, or $K_{\delta, \theta}$ and $I = [0, 1]$. In what follows, we assume that the kernel function is symmetric: $K(s, t) = K(t, s)$ and that

$$K : I \times I \rightarrow \mathcal{R} \quad \text{is continuous.}$$

Here one notices the regularization role of ε_0 introduced for the permanent cost component, without which K_3 would not be continuous at $(t, s) = (0, 0)$. The following norm formula is also standard in functional analysis [6, 7]

$$\|Q_{K,dV}\| := \sup_{\|h\|_{L^2(dV)} \leq 1} Q_{K,dV}[h] \leq \|K\|_{L^2(dV \otimes dV)} \leq \|K\|_{L^\infty} \cdot V_1.$$

Definition 1. $Q_{K,dV}[h]$ is said to be positive in the Hilbert space $L^2(I, dV)$ if for any $h \in L^2(I, dV)$, $Q_{K,dV}[h] \geq 0$. A symmetric and continuous kernel $K(s, t)$ on $I \times I$ is said to be positive, if for any finite sequence $(t_i | i = 1 : N)$ in I of arbitrary length N , and real scalars $(c_i | i = 1 : N)$, one has

$$\sum_{i,j=1}^N K(t_i, t_j) c_i c_j \geq 0.$$

Notice that in some literature, it is said to be *nonnegative*.

Proposition 1. *The quadratic form $Q_{K,dV}[h]$ induced by a positive kernel K must be positive.*

Proof. This is canonical in the context of *ordinary* Lebesgue measure dt , which we do refresh here first. Assume that K is a positive kernel. First notice that

$$Q_{K,dt}[g] = \int \int_{I \times I} K(s,t)g(s)g(t)dsdt \leq \|K\|_{L^\infty(dt)}\|g\|^2,$$

for any $g \in L^2(I, dt)$. Thus $Q_{K,dt}$ is positive if and only if it is positive on a dense set of $L^2(I, dt)$. It is well known in real analysis [7] that the set of continuous functions $C(I)$ is indeed dense in $L^2(I, dt)$. For any $\phi(t) \in C(I)$, the Lebesgue integral in $L(I \times I, dt \otimes ds)$

$$Q_{K,dt}[\phi] = \int \int_{I \times I} K(s,t)\phi(s)\phi(t)dsdt$$

is identical to its Riemann integral, since $K(s,t)\phi(s)\phi(t) \in C(I \times I)$. For any partition of I :

$$0 = t_0 < t_1 < t_2 < \dots < t_N = 1,$$

the following Riemann sum is positive since K is assumed positive:

$$\sum_{i,j=1}^N K(t_i, t_j)\phi(t_i)\phi(t_j)(t_{i+1} - t_i)(t_{j+1} - t_j) \geq 0.$$

Taking proper limits, one sees that $Q_{K,dt}$ is indeed positive on $C(I)$.

Now consider a general volume measure dV with $v = dV/dt \in L^\infty(dt)$. For any $h_t \in L^2(dV)$, one has

$$\int_I h^2 v^2 dt \leq \|v\|_\infty \cdot \int h^2 v dt = \|v\|_\infty \|h\|_{L^2(dV)}^2,$$

implying that $g(t) = h(t)v(t) \in L^2(dt)$. Then the proof is complete following that

$$Q_{K,dV}[h] = Q_{K,dt}[g] \geq 0.$$

□

Proposition 2. *The risk kernel $K = K_{\delta,\theta}$ is positive.*

By definition, $K_{\delta,\theta}(t, s)$ is the auto-covariance function of the price-spread mixed process:

$$W_t = \delta_t + \alpha_0(\theta_t - \bar{\theta}_t), \quad \text{and} \quad K_{\delta,\theta}(t, s) = E[W_t \cdot W_s].$$

Then for any finite sequence $(t_i | i = 1 : N)$ and real scalars $(c_i | i = 1 : N)$,

$$\sum_{i,j=1}^N K_{\delta,\theta}(t_i, t_j)c_i c_j = E \left[\left(\sum_{i=1}^N c_i W_{t_i} \right)^2 \right] \geq 0,$$

which establishes the positivity of $K_{\delta, \theta}$.

To proceed further, we need a useful lemma whose proof follows directly from the definition of kernel positivity.

Lemma 1. *Suppose $K(T, S)$ is a positive kernel on a subset $D \subseteq \mathbb{R}$, and $\phi : I \rightarrow D$ any real function with $T = \phi(t)$. Then the pullback kernel $K_\phi(t, s) = K(\phi(t), \phi(s))$ is positive on I .*

Proposition 3. *The transient kernel $K = K_2$ is positive.*

Proof. Given a real function $g(t)$ on $t \in \mathbb{R}$, suppose its Fourier transform $G(\omega) = \int_{\mathbb{R}} g(t)e^{-\sqrt{-1}t\omega} dt$ is real and non-negative for all $\omega \in \mathbb{R}$. Then the kernel defined via: $K_g(t, s) = g(t - s)$ must be positive for $t, s \in \mathbb{R}$. This is because:

$$\begin{aligned} \sum_{i,j=1}^N K_g(t_i, t_j)c_i c_j &= \sum_{i,j=1}^N g(t_i - t_j)c_i c_j \\ &= \sum_{i,j=1}^N \frac{1}{2\pi} \int_{\mathbb{R}} G(\omega)e^{-\sqrt{-1}(t_i - t_j)\omega} d\omega \cdot c_i c_j \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} G(\omega) \left| \sum_{i=1}^N c_i e^{-\sqrt{-1}t_i \omega} \right|^2 d\omega \geq 0. \end{aligned}$$

Since the Fourier transform of an exponential $g(T) = \gamma \exp(-\gamma|T|)$ with $T \in \mathbb{R}$ is $G(\omega) = \frac{2\gamma^2}{\gamma^2 + \omega^2}$, we conclude that

$$K_g(T, S) = g(T - S) = \gamma \exp(-\gamma|T - S|)$$

must be positive on $T, S \in \mathbb{R}$. The proof is then complete by taking $\gamma = 1/V_*$, and $\phi(t) = V_t$ in the preceding lemma. \square

Proposition 4. *The permanent kernel $K = K_3(t, s)$ is positive.*

Proof. First it is easy to see from the definition that, if $K(T, S)$ is positive on a subset $D \subseteq \mathbb{R}$, and $f(t) : D \rightarrow \mathbb{R}$ a real function, the new kernel

$$K_f(T, S) = K(T, S)f(T)f(S)$$

must be positive on the same domain. Now define $D = (0, \infty)$, and

$$K_B(T, S) = T \wedge S, \quad \text{and} \quad f(T) = 1/T.$$

Notice that $K_B(T, S)$ is positive since it is the auto-covariance of the canonical Brownian motion B_T :

$$K_B(T, S) = \mathbb{E}[B_T \cdot B_S], \quad T, S > 0.$$

Therefore, the kernel

$$K(T, S) = \frac{1}{T \vee S} = \frac{T \wedge S}{T \cdot S} = K_B(T, S)f(T)f(S)$$

must be positive on D . The proof is complete via the preceding lemma with $\phi(t) = \varepsilon_0 + V_t$. \square

Theorem 1 (Convexity and Uniqueness). *The objective functional $J_\lambda[h_t]$ is strictly convex in $L^2(I, dV)$. As a result, the optimal execution solution h_t^* to the constrained model must be unique.*

Proof. The preceding propositions establish that the last three quadratic forms in $J_\lambda[h]$ are all positive and thus convex in the Hilbert space $L^2(dV)$. The second quadratic term (from instantaneous cost)

$$\alpha_1 \int_I h_t^2 dV_t = \alpha_1 \|h_t\|_{L^2(dV_t)}^2$$

is the squared norm and hence strictly convex. Since the first term on the average spread cost is linear, $J_\lambda[h]$ must be *strictly* convex. \square

5.3. Compactness and Existence

For convenience, define the combined kernel

$$\begin{aligned} \mathbf{K}_\lambda(t, s) &= \frac{\alpha_2}{2} K_2(t, s) + \frac{\alpha_3}{2} K_3(t, s) + \lambda K_{\delta, \theta}(t, s) \\ &= \frac{\alpha_2}{2V_*} \exp\left(-\frac{|V_t - V_s|}{V_*}\right) + \frac{\alpha_3}{2} \frac{1}{V_t \vee V_s + \varepsilon_0} + \lambda K_{\delta, \theta}(t, s). \end{aligned} \quad (11)$$

Since $\mathbf{K}_\lambda(t, s)$ is a continuous kernel over $I \times I = [0, 1]^2$, we have

$$\mathbf{K}_\lambda(t, s) \in L^2(I \times I, dV \otimes dV), \text{ and } \|\mathbf{K}_\lambda(t, s)\|_{L^2(dV \otimes dV)} \leq \|\mathbf{K}_\lambda(t, s)\|_\infty \cdot V_1.$$

We also use the same symbol to denote the induced linear operator in $L^2(I, dV)$:

$$\mathbf{K}_\lambda h_t := \int_I \mathbf{K}_\lambda(t, s) h(s) dV_s.$$

It is well known [5] that (i) the L^2 function norm dominates the operator norm:

$$\|\mathbf{K}_\lambda\| \leq \|\mathbf{K}_\lambda\|_{L^2(dV \otimes dV)} < \infty,$$

and (ii) such a linear operator \mathbf{K}_λ must be *compact* in the Hilbert space of $L^2(I, dV)$.

Theorem 2. *There exists a unique optimal execution solution h_t^* to the mean-variance model with $J_\lambda[h]$ defined in Eqn. (7) as the objective function, and with the following constraints (Eqn. (8) and (9)): monotonicity, completion, and PoV capping, as long as the latter two are compatible as expressed in Eqn. (10).*

Proof. By Theorem 1, it suffices to further establish the existence portion. Define

$$\hat{\phi}_t = \alpha_0 \cdot \text{side} \cdot \bar{\theta}_t.$$

Then the objective $J_\lambda[\cdot]$ can be expressed more compactly by:

$$J_\lambda[h] = \langle \hat{\phi}_t, h_t \rangle + \alpha_1 \langle h_t, h_t \rangle + \langle \mathbf{K}_\lambda h_t, h_t \rangle,$$

where the inner product is in the Hilbert space of $L^2(I, dV)$. Let \mathbf{I} denote the identity operator. Then we have

$$J_\lambda[h] = \langle \hat{\phi}_t, h_t \rangle + \langle (\alpha_1 \mathbf{I} + \mathbf{K}_\lambda) h_t, h_t \rangle.$$

Since \mathbf{K}_λ is the linear combination of three positive operators (or kernels) with positive coefficients, it must be *positive*. It is well known in the *spectral theory* [5] that the spectrum set $\sigma(\mathbf{K}_\lambda)$ of such a compact and positive operator must be a subset of the positive half real axis $[0, \infty)$, and for any $\mu \notin \sigma(\mathbf{K}_\lambda)$, the inverse $\mathbf{K}_\lambda - \mu \mathbf{I}$ must exist and be bounded in the Hilbert space $L^2(I, dV)$. In particular, with $\mu = -\alpha_1 \notin \sigma(\mathbf{K}_\lambda)$, $(\mathbf{K}_\lambda + \alpha_1 \mathbf{I})^{-1}$ is a well-defined bounded linear operator, and one can define

$$\phi_t = (\alpha_1 \mathbf{I} + \mathbf{K}_\lambda)^{-1} \hat{\phi}_t \in L^2(I, dV).$$

We also further introduce a new symmetric bilinear function: for $g, h \in L^2(I, dV)$,

$$(g, h) = \langle (\alpha_1 \mathbf{I} + \mathbf{K}_\lambda) g_t, h_t \rangle$$

It is *strictly* positive since it *dominates* the ordinary inner product:

$$(h, h) \geq \alpha_1 \langle h, h \rangle.$$

Thus it introduces a new inner product, which is actually *equivalent* to the natural one since it is also bounded above:

$$(h, h) = \alpha_1 \langle h, h \rangle + \langle \mathbf{K}_\lambda h, h \rangle \leq (\alpha_1 + \|\mathbf{K}_\lambda\|_\infty \cdot V_1) \langle h, h \rangle.$$

For convenience, we denote by $L^2(I, dV | (\cdot, \cdot))$ the same function space $L^2(I, dV)$ endowed with this new equivalent inner product, which is a Hilbert space. Furthermore, the original objective function $J_\lambda[h]$ simplifies to:

$$J_\lambda[h] = (\phi_t, h_t) + (h_t, h_t). \quad (12)$$

Define $\psi_t = (\alpha_1 \mathbf{I} + \mathbf{K}_\lambda)^{-1} \mathbf{1}$. Then the completion constraint becomes: $(\psi_t, h_t) = X_1$. And the constraints on monotonicity and participation limit remain the same:

$$0 \leq \text{side} \cdot h_t \leq \text{maxPoV}.$$

The constant PoV execution strategy:

$$h_t^{\text{const}} \equiv \frac{X_1}{V_1}, \quad t \in I = [0, 1]$$

is clearly admissible under the given constraints due to the compatibility assumption in Eqn.(10), and has a finite objective value. Then there must exist a non-empty sequence of minimizing execution strategies with finite objectives J_λ 's:

$$h_t^{(1)}, h_t^{(2)}, \dots,$$

in $L^2(I, dV)$, such that

$$\lim_{n \rightarrow \infty} J_\lambda[h^{(n)}] = \inf J_\lambda[h] < \infty,$$

and each meets the constraints. The sequence must be bounded since one can easily show from Eqn. (12) that

$$\|h_t\|_{(\cdot, \cdot)} \leq \left(J_\lambda[h_t] + \frac{1}{4} \|\phi_t\|_{(\cdot, \cdot)}^2 \right)^{1/2} + \frac{1}{2} \|\phi_t\|_{(\cdot, \cdot)}.$$

Now that in Hilbert spaces, any bounded sequence must be *weakly* pre-compact, possibly replaced by one of its subsequences, $(h_t^{(n)} | n = 1, 2, \dots)$ could be assumed to *weakly* converge to some element $h_t^* \in L^2(I, dV | (\cdot, \cdot))$, so that for any $g \in L^2(I, dV | (\cdot, \cdot))$:

$$(g_t, h_t^*) = \lim_{n \rightarrow \infty} (g_t, h_t^{(n)}).$$

Since the Hilbert norm is known to be *lower semi-continuous* (l.s.c.) under weak convergence, we have

$$J_\lambda[h^*] \leq \lim_{n \rightarrow \infty} (\phi_t, h_t^{(n)}) + \liminf_{n \rightarrow \infty} (h_t^{(n)}, h_t^{(n)}) = \liminf_{n \rightarrow \infty} J_\lambda[h_t^{(n)}] = \inf J_\lambda[h].$$

Finally the admissible space defined by the three constraints is easily seen to be a *closed* and *convex set*, which must be *closed* under weak convergence [5]. This means h^* also meets all the three constraints on completion, monotonicity, and PoV capping. Then h^* must be the optimal execution strategy with $J_\lambda[h^*] = \inf J_\lambda[h]$. \square

6. Computation and Numerical Examples

6.1. Model Calibration

In this subsection, we briefly explain the major steps for model calibration. Actual implementation should depend on the trade database each execution house owns. For example, as explained in Section 4, COP strategies employed by an execution house may directly impact data bookkeeping and subsequent model calibration,

Although the proposed model involves a varieties of kernels, its main characteristics is the linearity for all its four major parameters: $\alpha_0, \alpha_1, \alpha_2$, and α_3 . Model calibration can thus be directly based upon linear regression.

We first make the following assumptions about the empirical trades already made in the past and stored in the database of the execution house.

- (i) (Universe Coverage) For targeted securities, one has enough samples of past trades. For any major broker-dealers, exchanges, or agency houses, this is typically not an issue. For example, in the United States, securities from the universes of all major indices (e.g., S&P 500) are heavily traded. (Due to the proprietary nature of trade information, this often imposes much challenge for academic researchers.)
- (ii) (Data Cleaning) Trade data have been properly *filtered*. This has been a very common practice in execution houses. To filter is to remove erroneous trades or insignificant trades from participating in the calibration. For example, trades lasting less than 5 minutes or trades whose average PoVs are below 0.1% can be considered insignificant and filtered out.
- (iii) (Recording Intervals) For each trade, the house has kept on record its execution details, which may include for example, shares traded over each 10 seconds and average execution prices associated with. The periodic recording interval should not be too long so that the calibrated model can properly capture the transient effect.
- (iv) (Profiles and Risks) Using statistical methods and consolidated exchange data, the house has already created the standard profiles for the Bid-Ask spread $\bar{\theta}_t$, and the volume measure dV_t for each security, as well as the risk metrics for the auto-covariance matrices $K_\delta(t, s)$ and $K_\theta(t, s)$ as risk metrics.

At the execution houses where the author had worked, trading models are typically calibrated daily, or weekly the longest. (Risk parameters involved could stay longer, however.)

Implementation shortfalls (IS) are typically recorded and reported as basis points (bps). One basis point is 0.0001 (of a given currency). We thus define $\text{bps} = 1/10,000$. For a given trade, IS in bps is related to IS in dollars (or any other working currency) by

$$\text{IS}_{\text{bps}} = \frac{\text{IS}_\$}{p_0 |X_1|} \cdot \frac{1}{\text{bps}}. \quad (13)$$

For example, if a buying trade of targeted initial notional $p_0 * |X_1| = \$10,000,000$ is reported to have 12 bps of IS_{bps} , it means the final net cost the client actually has to pay is $\$10,012,000$. This even does not include any trading commissions or fees.

In order to calibrate the model to the reported IS_{bps} in databases, and to have normalized magnitudes for the model coefficients, we make the following coefficient normalization based on proper dimensionality analysis (with the circular superscript \circ representing normalized coefficients)

$$\begin{aligned} \bar{\theta}_t &= \bar{\theta}_t^\circ \cdot p_0 \cdot \text{bps}, & \alpha_0 &= \alpha_0^\circ \\ \alpha_i &= \alpha_i^\circ \cdot p_0 \cdot \text{bps}, & i &= 1, 2, 3. \end{aligned}$$

Then by Eqn. (3) for $\text{IS}_{\mathbb{S}}$, and the definition for IS_{bps} in Eqn. (13),

$$\begin{aligned} \mathbb{E} [\text{IS}_{\text{bps}}[h]] &= \alpha_0^\circ \int_I \bar{\theta}_t^\circ \frac{dX_t}{X_1} + \alpha_1^\circ \int_I h_t \frac{dX_t}{|X_1|} \\ &\quad + \alpha_2^\circ \int_I \frac{dX_t}{|X_1|} \int_0^t h_s dM(s|t) + \alpha_3^\circ \int_I \frac{dX_t}{|X_1|} \int_0^t h_s \frac{dV_s}{V_t + \varepsilon_0} \\ &= \alpha_0^\circ \cdot C_0[h] + \alpha_1^\circ \cdot C_1[h] + \alpha_2^\circ \cdot C_2[h] + \alpha_3^\circ \cdot C_3[h] \end{aligned}$$

We thus apply linear regression to all validated historical trades (i.e., h 's already observed) based on the linear model:

$$\text{IS}_{\text{bps}}[h] \sim \alpha_0^\circ \cdot C_0[h] + \alpha_1^\circ \cdot C_1[h] + \alpha_2^\circ \cdot C_2[h] + \alpha_3^\circ \cdot C_3[h] + w[h].$$

Notice the *heteroskedasticity* of the model as the residual noise term $w[h]$ carries a variance that depends on the execution profile h_t , as shown in Eqn. (4). We thus apply either the weighted least-square estimator (w-L.S.E.) or the equivalent ordinary L.S.E. to the variance normalized data.

We leave some other auxiliary details to the execution houses who are interested in the current work and who can conduct actual model calibration from their proprietary trading database. The author welcomes any feedback or suggestion from the industry.

6.2. Numerical Computing via Quadratic Programming

In this section, we show how the proposed model can be efficiently computed via quadratic programming algorithms and available commercial software (e.g., MOSEK or IBM CPLEX).

Unlike dynamic trading models adapted to evolving market environments, pre-trade models are typically built upon *historical* “profiles” for spreads, volumes, correlations and volatilities. These profiles are often generated via robust statistical methods daily or weekly, and stored into data files. Typically they are discrete intraday vectors with intervals ranging from 30 seconds to 5 minutes, based on both the liquidity levels of the target securities and the COP machinery each house employs. Therefore, below we study the discrete implementation on a regular discrete time grid.

Throughout the work we have been using the normalized trading horizon $I = [0, 1]$, mainly to introduce both notional and analytical convenience and clarity. This is not necessary for actual computing and we can work directly with the physical time $I = [T_0, T_1]$, say, from 9:45am to 1:35pm. Given a time interval Δt , suppose the trading horizon is discretized to:

$$t_0 = T_0, t_1 = \Delta t, \dots, t_N = T_1.$$

Then the continuous volume measure dV is discretized to a discrete volume “profile”:

$$d_n = V([t_{n-1}, t_n]) = \int_{t_{n-1}}^{t_n} dV_t, \quad n = 1 : N,$$

which is usually made available by the analytics teams of execution houses. Let

$$\mathbf{D} = \text{diag}(d_1, \dots, d_N)$$

be the diagonal matrix. The target PoV function h_t is similarly discretized to a column vector $H = (h_1, \dots, h_N)^T$, with

$$h_n = \frac{1}{d_n} \int_{t_{n-1}}^{t_n} h_t dV_t, \quad n = 1 : N,$$

whenever $d_n > 0$ (i.e., market volume is non-zero), and $h_n = 0$ otherwise.

Following the operator expression in Eqn. (11), we recycle the same symbol of \mathbf{K}_λ to denote the matrix corresponding to the discretization of the continuous kernel $\mathbf{K}_\lambda(t, s)$. Let $B = (b_1, \dots, b_N)^T$ denote the column vector with

$$b_n = \alpha_0 \cdot \text{side} \cdot \bar{\theta}_{t_{n-1/2}}, \quad \text{with } t_{n-1/2} = \frac{t_{n-1} + t_n}{2}, \quad n = 1 : N.$$

Then the model is discretized to:

$$\text{minimize over } H: \quad B^T \cdot \mathbf{D} \cdot H + H^T \cdot (\alpha_1 \mathbf{D} + \mathbf{D} \cdot \mathbf{K}_\lambda \cdot \mathbf{D}) \cdot H$$

subject to:

$$\text{side} \cdot H \leq \text{maxPoV};$$

$$\text{diag}(\mathbf{D})^T \cdot H = X_1;$$

$$\text{side} \cdot h_n \geq 0, \quad \text{where } d_n > 0, \quad \text{and}$$

$$h_n \equiv 0, \quad \text{where } d_n = 0.$$

Here $\text{diag}(\mathbf{D})$ denotes the column vector consisting of the diagonals of \mathbf{D} (following MATLAB). The last condition can also be used to reduce the actual dimension of the problem by eliminating zero PoV's where there are no market volumes (i.e., with $d_n = 0$).

This discrete problem fits well into the framework of quadratic programming, and can be efficiently solved numerically by commercial optimizers such as MOSEK and IBM CPLEX, which are often integrated into the local Java or C++ libraries of in-house execution analytics.

6.3. Numerical Examples

In this section, we present several examples that help reveal the general behavior of the proposed model. We have relied on the built-in quadratic programming optimizer *quadprog.m* in MATLAB. The MATLAB software generating these examples are available from the author upon request.

Throughout we assume a hypothetical market that opens for 390 minutes from 9:30am to 4:00pm. The target security is assumed to have an arrival price of $p_0 = \$30.00$, and be moderately liquid with *average daily volume* (ADV) about 5,000,000 shares. We also assume that the volume and spread profiles have been established discretely in minutes, and that the volume profile bears

a typical U-shape with more volumes at the Open and Close. The primary task is to buy $X_1 = 90,000$ shares during a horizon that spans 90 minutes.

Under these general settings, we assume that the hypothetical client is allowed to customize on the following three trading factors: (1) starting time T_0 , (2) PoV capping maxPoV, and (3) level of risk aversion λ . The choice of starting time affects the “shape” of the volume measure over the trading horizon due to the daily U-shape. In most examples we set $\text{maxPoV} = 20\%$, and the risk aversion to a medium level of $\lambda = 10^{-3}$. To better compare with the existing literature, unless indicated otherwise, the price dynamics is assumed to be Brownian with constant spreads.

6.3.1. Effect of Risk Aversion

Plotted in Fig. 1 and Fig. 2 are the optimal execution solutions corresponding to three different levels of risk aversion: medium (Fig. 1), and high and low (left and right panels in Fig. 2). As common for mean-variance models, more risk aversion implies more front-loading behavior. But unlike most results illustrated in the classical literature, front-loading is allowable only up to the level of the maxPoV (which is 20% in this series of examples) typically set by the clients.

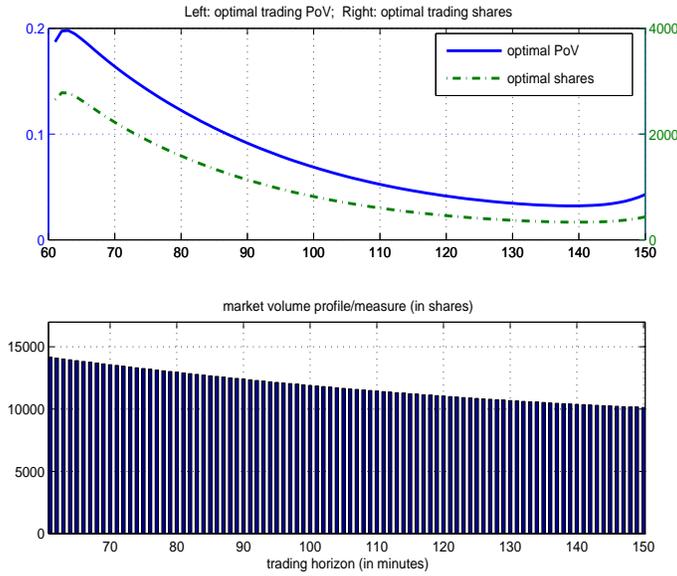


Figure 1: Medium risk aversion with $\lambda = 10^{-3}$

6.3.2. Effect of Volume Measures

Demonstrated in Fig. 3 and Fig. 4 are the effects of volume measures. Under a fixed moderate level of risk aversion ($\lambda = 10^{-3}$), trading in a high volume market environment can drive down the average PoV’s and hence trading costs

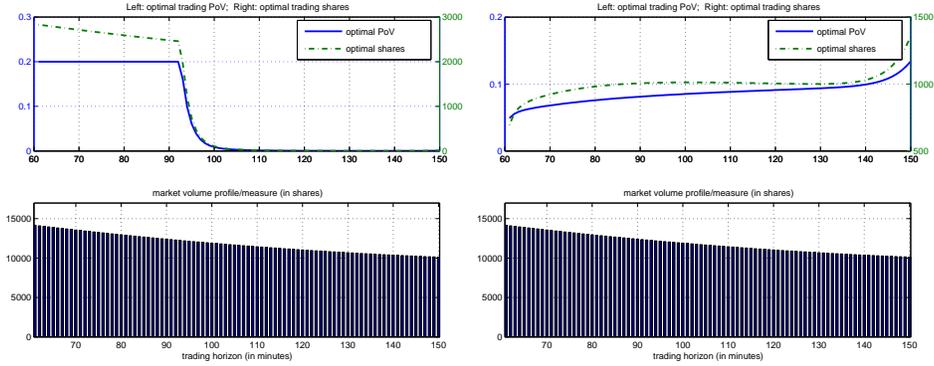


Figure 2: (Effect of Risk Aversion) Left: high risk aversion with $\lambda = 10^{-1}$; Right: low with $\lambda = 10^{-5}$. Higher risk aversion implies more front-loading behavior. Unlike many results in existence, front loading is allowable only up to the level of the maxPoV set by a client.

(e.g., the left panel of Fig. 4 which simulates typical risk-averse trading in a “morning” session).

6.3.3. Effect of Cost Components

In order to better understand the role of each cost component in the modeling, in the next three figures (Fig. 5 to 7), we plot optimal solutions corresponding to the boosting of the individual component coefficients: $\alpha_i^\circ, i = 1 : 3$. For each figure, we have boosted up one of the target coefficients by 10 times, while holding the other two at the original moderate level. The captions of the individual figures give more details and discussions.

6.3.4. Effect of Price Dynamics

In the two panels of Fig. 8, we also demonstrate the flexibility of the proposed model in dealing with price dynamics other than the classical (linear or geometric) Brownian motions. The left panel plots the optimal solution for the *mean-reversal* model, and the right panel for an *asymmetric-volatility* model.

Mean Reversion

In the mean reversal model, one assumes the homogenized price change obeys the following equation:

$$d\delta_t^\circ = -\kappa\delta_t^\circ dt + \alpha dW_t,$$

with the initial condition $\delta_{t=0}^\circ = 0.0$. Here κ stands for the strength of mean reversal, and αdW_t for the Brownian component. It can be shown that the solution bears the closed-form:

$$\delta_t^\circ = \alpha \cdot \int_0^t e^{-\kappa(t-s)} dW_s,$$

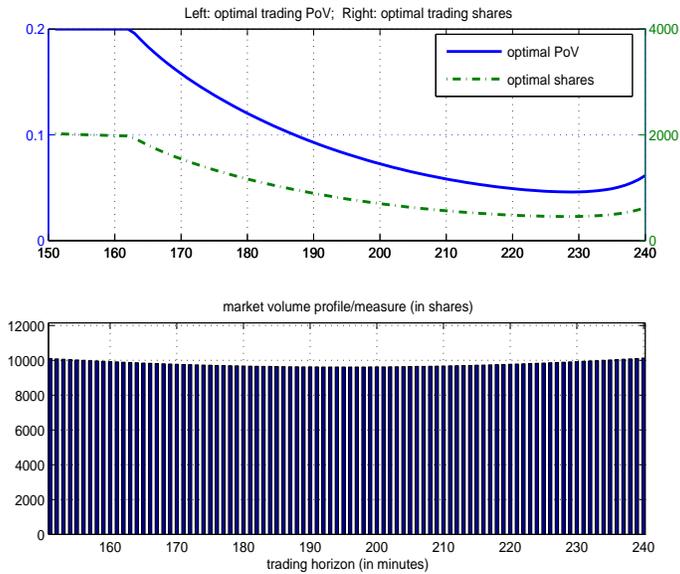


Figure 3: (Effect of Volume Measures) In a U-shaped daily volume measure, volume is relatively low and almost “flat” near the hypothetical “noon” time (12:00pm - 13:30pm). Under the given $\text{maxPoV} = 20\%$ and a moderate level of risk aversion, the combined effect of risk aversion and flat volumes encourages to trade faster in the beginning of the horizon.

and that the auto-covariance function is given by

$$K_{\delta}^{\circ}(t, s) = \frac{\alpha^2}{2\kappa} \cdot e^{-\kappa|t-s|} \cdot \left(1 - e^{-2\kappa(s \wedge t)}\right). \quad (14)$$

Notice that for bigger t, s , the covariance function behaves more like an exponential kernel.

Thus both conceptually and quantitatively, mean reversion erases long-term memory and only keeps a shift-invariant (when sufficiently away from $t_0 = 0$) short-term correlation. Unlike Brownian motions for which price uncertainties grow in the order of $\sqrt{t - t_0}$ as time elapses, mean reversion maintains almost a constant level of uncertainty at $\alpha/\sqrt{2\kappa}$. This implies that faster front-loading trading does not necessarily help reduce risks as risks are time invariant. This is indeed confirmed from the left panel of Fig. 8. Furthermore, as the kernel asymptotically behaves like the exponential function, we observe the “wall” effect at the two boundaries as well documented in the classical literature [10].

Asymmetric Stochastic Volatility

In the next example, we consider a simple form of stochastic volatility:

$$d\delta_t^{\circ} = \sigma_0 \cdot \left(\exp(-\beta\delta_t^{\circ}) \cdot 1_{\delta_t^{\circ} \leq 0}\right) \wedge 2.0 \cdot dW_t, \quad (15)$$

where σ_0 and β are constant. Under this model, instantaneous volatility increases when the price delta is negative, but has been capped under $2\sigma_0$. In

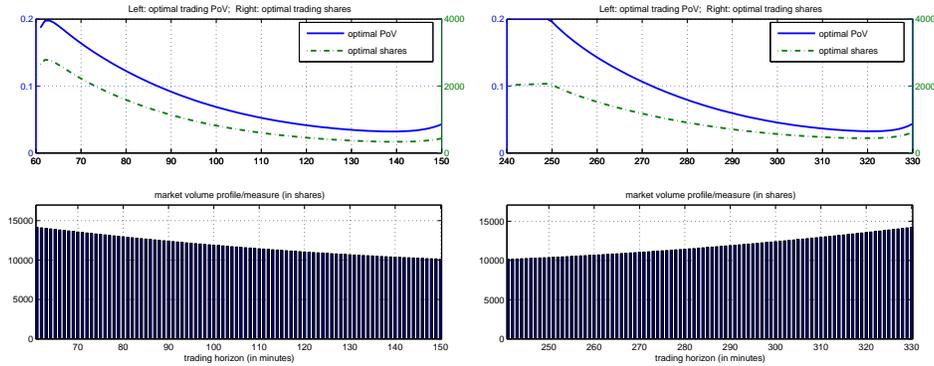


Figure 4: (Effect of Volume Measures) Model and trading parameters are being held the same for Fig. 3 except for the starting time, which affects the volume shape during the trading horizon. Left: trading in the “morning” session with more volumes skewed towards the Open; Right: trading in the “afternoon” session with volumes more concentrated towards the Close. For the “afternoon” session, risk aversion hinders taking full advantage of the richer volumes near the end of the trading horizon, and encourages instead to participate more in the beginning even the volume is comparatively lower.

the positive regime when $\delta_t^\circ > 0$, the instantaneous volatility remains at the constant level of σ_0 . This behavioral transition has been made possible by the indicator function $1_{\delta_t^\circ \leq 0}$. The volatility is thus asymmetric with respect to the directions of price movements.

As no closed form exists for the auto-covariance function, we turn to the Monte-Carlo estimation method using thousands of simulated paths (40,000 in this example). The estimated kernel function is then applied in the proposed execution model.

The resulted optimal solution has been plotted in the right panel of Fig. 8, in which one could observe the non-smoothness associated with the Monte-Carlo kernel estimation. Since the asymmetry increases the effective volatility, risk aversion enhances the front loading behavior compared with the symmetric case with a constant volatility σ_0 , which is evident from the plotting.

7. Conclusion

While realistic trading models have to be *dynamic*, static pre-trade models are still important and widely implemented in the execution industry for a number of reasons and applications explained in the Introduction section.

The primary objective of the current work has been to build a continuum model that

- (a) automatically accommodates the broadest varieties of price dynamics,
- (b) more faithfully engages the role of market volumes in the general reductionism approach of static modeling,
- (c) employs impact cost components that bear low complexities but with more market signals embedded and represented,

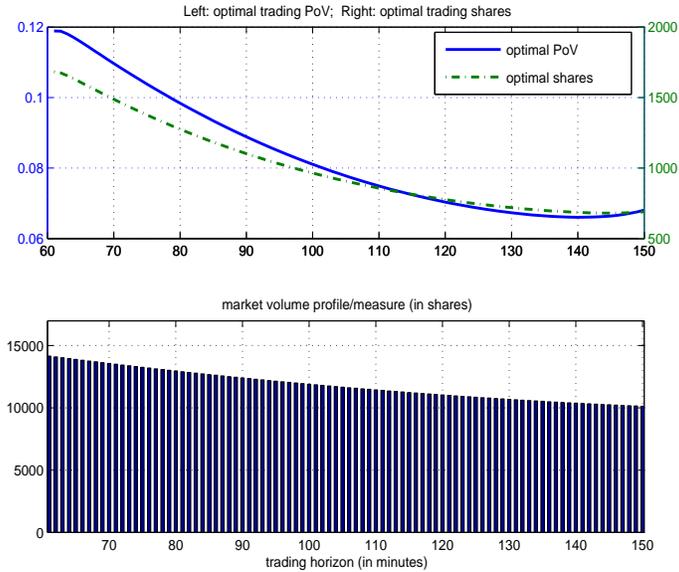


Figure 5: (Effect of Instantaneous Cost) The instantaneous cost coefficient α_1^o is boosted up by 10 times, with α_2^o and α_3^o fixed at the original moderate level. By definition, the instantaneous cost component is highly localized and trading now vs. later bears no extra penalties. The optimal execution is thus mostly shaped by risk aversion and is typically front-loading.

- (d) is versatile enough to allow most popular constraints from clients, and yet
- (e) is still analytically tractable and computationally feasible.

By working directly with the auto-covariance functions, the model virtually has allowed any price dynamics, in particular, Markovians like Brownians or non-Markovians with memories. Building upon the foundation of measure theory, we have also treated market volumes as Borel measures over the execution horizons. Pre-trade executions are then considered to be absolutely continuous measures over such measure spaces, which naturally results in the target decision variable to optimize with – the PoV rate function. All the four impact cost components have been consistently built upon the PoV function. They are all kept linear but with more market signals integrated in, including volume distances for the transient costs and cumulative volume normalization for the permanent costs. We have also considered heuristically the influence of in-house *Child Order Placement* strategies and *quantitative market makers* on impact cost building.

In combination, the proposed pre-trade model has led to a constrained quadratic programming problem in infinite-dimensional Hilbert spaces, which accommodates most linear constraints frequently requested by internal or external clients. We have in particular worked with the following three primary constraints: (i) monotonicity, (ii) completion, and probably the most important, (iii) PoV capping or volume limits, which is frequently requested from clients.

We have applied the theories of positive quadratic operators and compact

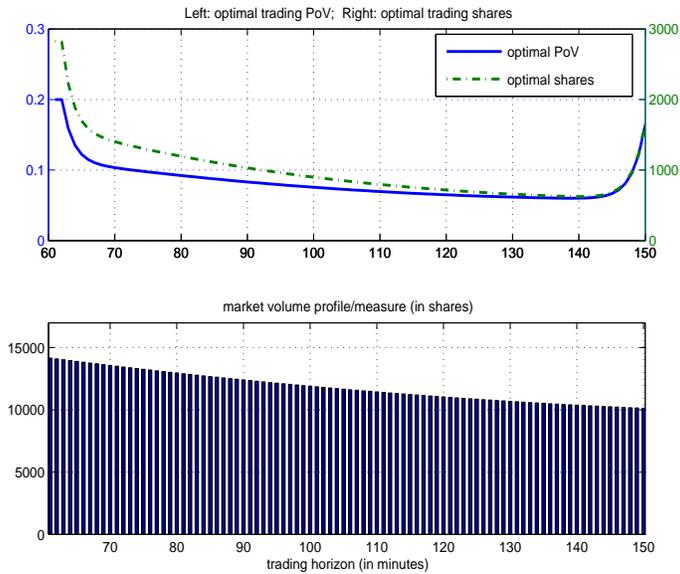


Figure 6: (Effect of Transient Cost) The transient cost coefficient α_2^o is boosted up by 10 times, with α_3^o and α_1^o held at the original moderate level. The notable boundary “wall” effect arising from the exponential transiency has been well documented in the classical literature (e.g., see Proposition 2 and 3 in Obizhaeva and Wang [10]).

operators in Hilbert spaces to establish both the positivity and compactness of the operators involved, and hence also the existence and uniqueness of optimal executions. One possible numerical scheme for projecting the continuum model onto interval-based grids has also been provided, and several computational examples have been carefully designed to address the effects of all the major factors.

Despite the versatility of the proposed model, we have to make some necessary warnings. Firstly, the model has been primarily designed to be a *pre-trade* static model and to become a major service component in the pre-trade packages offered to both internal or external clients by execution houses. Real *dynamic* trading models could be *heuristically* built upon such pre-trade models, but have to be integrated with a sound re-optimization strategy. Next, the current work has been carried out still in the conventional reductionistic approach based on market volumes. With growing analytics and understanding about the *limit order book* (LOB) dynamics, it would be naturally interesting to develop LOB based pre-trade models which are both analytically tractable and computationally feasible, and also to compare them with volume-based models (via real trading databases) to quantify the net improvements in pre-trade analytics.

We conclude the work by emphasizing that without the collective academic understanding and industrial practicing started by the many pioneers and leading practitioners, a portion of whose works have been frequently mentioned, the current work and modeling efforts would be absolutely impossible.

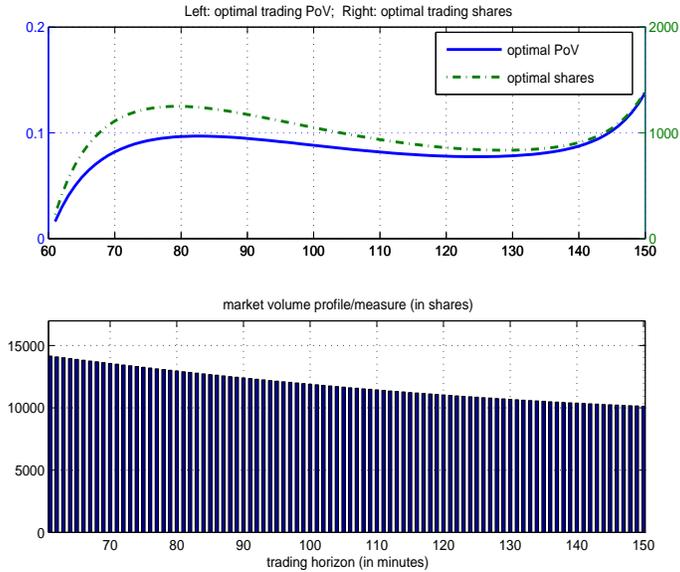


Figure 7: (Effect of Permanent Cost) The permanent cost coefficient α_3^o is boosted up by 10 times, with α_1^o and α_2^o held at the original moderate level. Permanent cost alone embraces back loading, so that shares traded at later times could pay less permanent cost built up by the earlier shares. The optimal trading profile plotted here achieves a balance under the front-loading pressure from risk aversion.

8. Disclaimers

- (1) The proposed model has not been the internal or external product of any execution houses where the author had worked. Any potential industrial conflict should be promptly directed to the attention of the author.
- (2) Due to the proprietary nature of the industry and the resulting scarcity of real trading data to the public, general expressions like “based on the experience in the execution houses where the author has worked, ...”, are purely for providing bona fide academic views based on the past working experience with real trading data and results.
- (3) Any mentioning of certain brand names, e.g., MATLAB, IBM CPLEX Optimizer, or MOSEK, etc., is not a product endorsement from the author for purchasing or investment, but an indication of some popular practices in the contemporary industry.
- (4) Execution houses who are interested in the current work and plan to implement it in their systems should be aware of any other operational risks, including for examples, the complexities at market opens or closes, trading halts, stock splitting, or any extreme market events, etc.

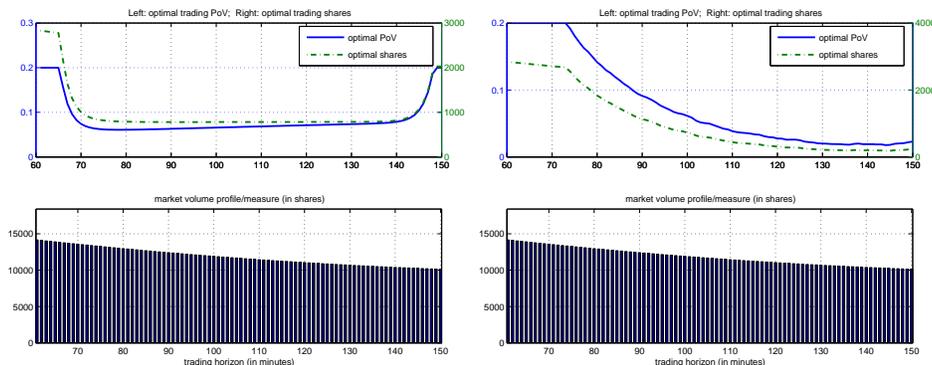


Figure 8: (Effect of Price Dynamics) Left Panel: under a mean-reversal (MR) price dynamics; Right Panel: under a price dynamics driven by asymmetric stochastic volatilities (ASV). As the MR dynamics erases long-term memories and converges to time-invariant short-term correlations, execution delay does not necessarily increase risk. Consequently, the front-loading behavior normally associated with risk aversion is weakened. The “wall” effect is the classical behavior of the exponential kernel [10], which the MR auto-covariance function converges to (see Eqn. (14)). On the right panel, the asymmetric stochastic volatility model (15) increases the effective instantaneous volatility compared with the normal Brownian dynamics and hence the execution risk associated with delaying. Therefore, it enhances the front-loading behavior associated with risk aversion.

9. Acknowledgment

The author wishes to thank the following colleagues on Wall Street: Robert Almgren for explaining to the author, then a freshman on Wall Street in 2007, how Wall Street operates under the introduction of my dear friend Andrea Bertozzi; Adlar Kim, Max Hardy, Daniel Nehren, Xu Fan, Chang Lin, Jesus Ruiz-Mata, Kathryn Zhao, Calvin Kim, Harry Rana, Arun Rajasekhar, et al., for all the delightful days at J. P. Morgan working so hard together to build solid Delta One and portfolio hedging, risk, and trading products; Anlong Li, Iaryna Gryniv, Paul Radovanovich, Federico de Francisco, Mark Skinner, Merrell Hora, Rishi Dhingra, Lada Kyj, Nazed Mannan, Allison Greene, Alan Chen, Ajit Kumar, Yihu Fang, Li Xu, Sunmbal Raza, Peter Ciaccio, Peter Norr, Hasan Ahmed, Bing Song, Ming Yang, Huaguang Feng, and George Liu, et al., for working so closely at the Barclays Capital from day to day as a grand team to deliver high-quality quantitative and technological solutions to modern algorithmic trading and portfolio analytics, and for the favorite donuts and coffees from the Dunkin Donuts, and all mini or max cupcakes from Melissa or Magnolia!

The author is particularly grateful to these colleagues and close friends, who had not only shaped his wall street career but also deeply and warmly touched his everyday life: Adlar Kim, Max Hardy, Daniel Nehren, Paul Radovanovich, Bing Song, and Huaguang Feng.

References

- [1] R. Almgren. Optimal trading with stochastic liquidity and volatility. *SIAM J. Financial Math.*, 3:163–181, 2012.
- [2] R. Almgren and N. Chriss. Optimal execution of portfolio transactions. *J. Risk*, 3:5–39, 2000.
- [3] D. Bertsimas and A. W. Lo. Optimal control of execution costs. *J. Financial Markets*, 1(1):1–50, 1998.
- [4] T. F. Chan and J. Shen. *Image Processing and Analysis: variational, PDE, wavelet, and stochastic methods*. SIAM Publisher, Philadelphia, 2005.
- [5] Y. Eidelman, V. Milman, and A. Tsolomitis. *Function Analysis - An Introduction*. Amer. Math. Soc., 2004.
- [6] L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, Inc., 1992.
- [7] G. B. Folland. *Real Analysis - Modern Techniques and Their Applications*. John Wiley & Sons, Inc., second edition, 1999.
- [8] G. Huberman and W. Stanzl. Optimal liquidity trading. *Yale School of Management Working Papers*, YSM 165, 2001.
- [9] R. Kissell and R. Malamut. Algorithmic decision making framework. *J. Trading*, 1(1):12–21, 2006.
- [10] A. Obizhaeva and J. Wang. Optimal trading strategy and supply/demand dynamics. *J. Financial Markets*, 16(1):1–32, 2013.
- [11] A. F. Perold. The implementation shortfall: Paper versus reality. *J. Portfolio Management*, 14:4–9, 1988.