Optimal execution of trading orders

Simon Kogan[$*] and Richard Knight[$]

April 2018

Abstract. We develop and evaluate a trading utility function driven by Implementation Shortfall framework. The function balances between execution risk, market impact, and return expectations. A novel, conforming to empirical results, definition is provided for the market impact component. The formulas are parameterized by volume-time argument facilitating optimization of VWAP based trading strategies.

Keywords: trading utility; utility optimization; market impact; volume-time

Trading utility function

"Consider the trading process at its most basic: You have cash, and you want to buy one stock. You think the stock will go up. You want to buy soon, before the stock rises. But to avoid market impact, you are willing to be patient and assume some risk of missing the stock rise. What is your optimal trading strategy?" (Grinold & Kahn, 1999)

The trading utility function – a measure of trading efficiency, is defined as follows:

$Utility = \alpha - \lambda \cdot \psi^2 - MI$, where α – cumulative return on an asset during a trade; ψ^2 – cumulative risk; MI – cumulative market impact. All three components depend on the shape of the asset accumulation during the trade life. Our objective is choosing the accumulation shape which maximizes the utility function.

The underlying assumption here is that the order size (the total quantity to be traded) is sufficiently large for continuous trade approximation. In reality, the parent order is split into many small child orders which are submitted to an exchange for execution.

The asset under discussion and in the examples below is a single stock; nevertheless, the formulas and the utility optimization reasoning in this paper are applicable to portfolio trading as well.

The utility functional form resembles the one in (Grinold & Kahn, 1999). We also use similar notation as far as convenient. The main differences are in the risk and market impact formulas.

The asset accumulation $h(\tau)$ is given as a function of volume-time argument τ. Volume-time represents the fraction of the average daily volume traded from the execution start t_0 till clock-time t (similarly to Almgren et al., 2005). τ is scaled so that $\tau = 0$ at t_0 and $\tau = T$ at the end of the execution with T as the duration of trading in volume-days (can be either a whole number or a fraction). Volume-time facilitates usage of trading strategies dependent on volume profiles. E.g., the natural minimal market impact strategy has linearly increasing asset accumulation (in τ) by trading along the asset's volume profile – VWAP execution.

[$]Haitong International Securities Group Limited, Sydney. [*]Email: simonkog@gmail.com

Notwithstanding the above, the utility components' formulas and the optimization procedure, suggested in this paper, are readily convertible from volume-time to clock-time domain applications. One just needs to replace volume-time τ with clock-time t in the formulas and use clock-time representation of the asset's normalized cumulative volume profile as the optimization starting point (see the optimization section for the details).

In keeping with Implementation Shortfall framework, we compare various asset accumulation shapes to the idealized execution which has: $h(\tau) \equiv 1$, a=0, ψ^2=0, MI=0 – the whole quantity executes at order arrival price with zero risk and zero market impact.

In the subsequent sections we provide formulas for the trading utility function components, calibration of constants in the formulas, numerical optimization details and application examples.

Throughout the paper the meaning of the below terms is as follows:

Rate – the first derivative of a function with respect to clock-time or volume-time.

Volume profile – the average market volume of the asset on intraday intervals (similarly to Almgren et al., 2005) – e.g., 20-day average volume on 15-minute intervals.

Cumulative volume profile – the average market volume of the asset from the execution start t_0 till t in clock-time or from 0 to τ in volume-time.

Asset accumulation function – the fraction of order size, executed from t_0 to t or from 0 to τ. The function is monotonically rising and normalized to [0,1] interval, thus $h(0) = 0$ and $h(T) = 1$. In the limit scenario, when the whole quantity is executed on order arrival: $h(\tau) \equiv 1$. In the other limit, when the whole quantity is executed at the last possible moment: $h(\tau) = \begin{cases} 0, & \tau < T \\ 1 & \tau = T \end{cases}$. Sell orders are evaluated with reflected asset accumulation $h(\tau) = 1 - h(\tau)$ allowing use of the same expressions for both buy and sell orders.

Cumulative return

This component represents our viewpoint on the asset's price action during the trading period. The general form of cumulative return: $a = \int_0^T [d\tau \cdot r'(\tau) \cdot [h(\tau) - 1]]$, where T – duration of trading in volume-days; $r'(\tau)$ – return rate, which is the first derivative of return function $r(\tau)$; $h(\tau)$ – asset accumulation function. We evaluate returns relative to the benchmark which is fully invested. Thus $[h(\tau) - 1]$ term in the above. Sell orders are evaluated with an inverted return $r(\tau) = -r(\tau)$ allowing use of the same expression for both buy and sell orders.

The cumulative return formulation holds for an arbitrary return shape. Yet we usually assume linear return function $r(\tau) = \frac{f \cdot \tau}{T}$ where f is forecast on the asset's return for the duration T. This corresponds to constant return rate $r'(\tau) = \frac{f}{T}$ – the asset price rising or falling gradually and evenly. Then $a = \frac{1}{T}\int_0^T \left[d\tau \cdot f \cdot [h(\tau) - 1]\right]$. Constant return rate in volume-time matches variable return rate in clock-time – price moves faster with larger volumes. The $\frac{1}{T}$ coefficient, in the formula, allows $a = -f$ for buy orders (and $a = f$ for sell orders) when the whole quantity is executed at the last possible moment.

Cumulative risk

This component represents our risk expectations derived from the asset's historical volatility. The general form of cumulative risk: $\psi^2 = \int_0^T \left[d\tau \cdot \sigma_d^2 \cdot [1 - h(\tau)]\right]$, where σ_d – asset's daily volatility; T – duration of trading in volume-days; $h(\tau)$ – asset accumulation function.

$\psi^2 = 0$ when the whole quantity is executed on order arrival. ψ^2 is maximal when the whole quantity is executed at the last possible moment.

The risk formula is a counterpart of cumulative active risk in (Grinold & Kahn, 1999). The difference is that $[1 - h(\tau)]$ term is not squared. Thus, the risk over a specific time frame is linearly related to $h(\tau)$, assuming Implementation Shortfall (or order arrival price) benchmark. E.g., when half of the traded quantity is executed on order arrival and the other half at the last possible moment, the correspondent risk is half of the risk when the whole quantity is executed at the last possible moment.

Cumulative market impact

We define cumulative market impact of trading order as follows:

$$MI = \int_0^T \left[d\tau \cdot c \cdot \frac{Q}{ADV} \cdot [h'(\tau)]^{1.5}\right]$$

Where Q – order size (the total quantity to be executed); ADV – average daily volume; $h'(\tau)$ – asset accumulation rate, which is the first derivative of the asset accumulation $h(\tau)$; T – duration of trading in volume-days;

The proposed model integrates market impacts over each $d\tau$ which is obviously a simplification of the real process. It does not account explicitly for memory effects and does not distinguish between temporary and permanent impacts. The formula is a counterpart of cumulative active market impact in (Grinold & Kahn, 1999). The main difference is that the asset accumulation rate is incorporated in the power of 1.5 instead of squaring; which makes the model conforming to empirical rules for the three trading regimes described in (Skachkov, 2014):

- Isochronic (constant trading duration – various order size and trading rate) market impact is linearly proportional to trading rate.
- Isochoric (constant order size – various trading duration and rate) market impact is proportional to the square root of trading rate.
- Isotachic or isokinetic (constant trading rate – various trading duration and order size) market impact is proportional to the square root of order size.

In each of these regimes, we keep constant one parameter of trading (e.g., trading duration) and work out the change in market impact comparatively to the change in other parameters (e.g., order size).

The model is seamlessly adaptable to the clock-time domain by replacing volume-time τ with clock-time *t*. It conforms to the trading regimes' rules in both volume-time and clock-time (the proofs are in Appendix A).

The constant *c* in the formula can be calibrated by utilizing inventory risk model or "a trading rule of thumb that it costs roughly one day's volatility to trade one day's volume" (Grinold & Kahn, 1999). Both imply market impact is linearly proportional to volatility, without specifying the proportionality factor.

Applying the rule of thumb literally, market impact is equal to daily volatility σ_d when quantity equal to the average daily volume is traded over a day. Using the minimal market impact asset accumulation $h(\tau) = \frac{\tau}{T}$ (i.e., $h'(\tau) = \frac{1}{T}$), we obtain $\sigma_d = MI = \int_0^T \left[d\tau \cdot c \cdot \frac{Q}{ADV} \cdot \left[\frac{1}{T}\right]^{1.5} \right] = c \cdot \frac{Q}{ADV \cdot \sqrt{T}}$. Then setting *Q=ADV* and *T=1* we arrive to $c = \sigma_d$.

This simplified calibration allows us to proceed with the trading utility optimization examples below. In practical applications, however, the procedure has to be improved with rigorous statistical evaluation and tweaking.

Risk weight calibration

The λ coefficient in the utility function can be resolved by considering trades balancing historical returns and risks when market impact is negligible. Trading in a very liquid asset, such as S&P500 index ETF, the average return presumably compensates the perceived risk: $\bar{a} = \lambda \cdot \psi^2$, when buying the ETF shares on session open and selling on session close for a long period. Utilizing S&P500 historical daily prices from 1962 till 2018, we obtain $\lambda \sim 2.6$. This is a generic estimation though. E.g., buy and hold investors may require higher risk penalty for trade execution than speculators making frequent trades.

The units of λ are $\frac{1}{price}$ or $\frac{1}{relative\ price}$ (similarly to $\$^{-1}$ units in Almgren & Chriss, 2000). Thus, we are willing to accept one extra square basis point of variance if it increases our expected return by λ basis points.

Note: calendar days and volume-days are interchangeable when they are whole numbers. Thus we can use λ calculated from the daily price data in volume-time based formulas.

Trading utility optimization

The optimal trading strategy, in the context of this work, is the one having the maximal utility value when order size *Q* and trade duration *T* are fixed. The strategy thus corresponds to the shape of asset accumulation. Under the supposition of solution space smoothness, we assume the choice is between: trading heavier early and slowing down later; or trading lighter early and speeding up later. We want to trade heavier early when we are buying and expect market move up or when we are selling and expect market move down. We want to trade lighter early, and thus heavier later, when we are buying and expect market move down or when we are selling and expect market move up. We want to trade in-line with market volume to minimize market impact. We want to trade heavier early to lower the risk.

Developing the analytical solution for the above problem is beyond the scope of this work. Instead, we choose an iterative numerical approach. The logical starting point for a VWAP based trading strategy optimization is the asset accumulation shape having the minimal market impact. Such a shape replicates normalized cumulative volume profile of the asset. This initial shape is subsequently skewed, in the iterative solution search, toward trading heavier early or later.

To facilitate the optimization procedure, we need to replace $h(\tau)$, in the utility components, with a 'function of function' transform of the input asset accumulation. The transform should be the monotonically rising function of input $h(\tau)$, preserving [0, 1] range on [0, *T*] argument interval, and be controlled by one or more parameters responsible for its shape. For this, we chose the double reflection power-law transform: $hpow(h(\tau), p) = \begin{cases} p < 1, & 1 - \left(1 - h(\tau)\right)^{\frac{1}{p}} \\ p \geq 1, & \left(h(\tau)\right)^{p} \end{cases}$. Where parameter p<1 leads to concave transformation resulting in heavier early trading; p>1 leads to convex transformation resulting in heavier later trading. Both transformations can be realistically used in trade execution. In contrast, the concave transformation, produced by the simple power-law transform $\left(h(\tau)\right)^{p}$ with p<1, rises sharply at the start, which makes it unusable in practice.

The optimization starting point (minimal market impact asset accumulation) is linear in volume-time: $h(\tau) = \frac{\tau}{T}$. Therefore, we replace $h(\tau)$ in the utility components' formulas with $\mathrm{hpow}\left(\frac{\tau}{\mathrm{T}},\mathrm{p}\right) = \begin{cases} \mathrm{p}<1, & 1\text{-}\left(1\text{-}\frac{\tau}{\mathrm{T}}\right)^{\frac{1}{\mathrm{p}}} \\ \mathrm{p}\geq 1, & \left(\frac{\tau}{\mathrm{T}}\right)^{\mathrm{p}} \end{cases}$; then search for p^* – the value of p maximizing the utility function.

Various numerical optimization methods can be deployed to find p^*. However, having a function of one parameter only, we conveniently span the representative range of powers $\left[p_s, \frac{1}{p_s}\right]$, where $p_s < 1$, with small steps and calculate the utility function on each step. First, $[p_s, 1]$ range is evenly covered with $p_1 = p_s, p_2 = p_s + \Delta p, \dots, p_{n-1} = 1 - \Delta p, p_n = 1$. Then the powers are reciprocated as $\frac{1}{p_{n-1}}, \dots, \frac{1}{p_2}, \frac{1}{p_1}$. Due to the nature of the power-law transform, it results in symmetrical coverage (i.e., symmetrical around p=1 on logarithmic p scale) with function responses denser toward p=1 and sparser toward the range ends.

The numerical optimization procedure does not rely on the initial asset accumulation being an exclusively linear function, thus can be performed entirely in the time domain. To proceed there, we replace volume-time τ argument with clock-time t in the trading utility components' formulas. The initial $h(t)$ is the clock-time representation of the asset's normalized cumulative volume profile, which frequently resembles a concave followed by a convex when plotted over a trading session. The double reflection power-law transform of the input asset accumulation: $hpow(h(t), p)$. Both volume-time and clock-time domain optimizations produce the same p^* for correspondent inputs. Thus, we can perform the calculation in one domain and apply the result to the other.

Examples

Below are examples of the trading utility function optimization for a buy order when the asset return forecast is positive, neutral, and negative. With both positive and negative forecasts, we assume constant return rates in volume-time. The asset is Toyota Motor Corp (TSE 7203) stock on 2016-11-17. Order size is set to 10% of the daily volume averaged over the previous 20 days (i.e., $\frac{Q}{ADV}$=0.1). $h(\tau) = \frac{\tau}{T}$ is used as the optimization starting point. The return forecasts are expressed as fractions of the stock's daily volatility σ_d – standard deviation of daily logarithmic returns over the previous 20 days.

The optimization results (see Fig 1.) confirm our expectations for the asset accumulation shape transformation. The buy order, with the positive forecast on the asset's return, has the maximal utility value with p^* less than one (concave transformation) – executes heavier early. In the neutral scenario, p^* is still less than one due to the balance between the market impact and risk. The buy order, with the negative forecast on the asset's return, has the maximal utility value with p^* greater than one (convex transformation) – executes heavier later.

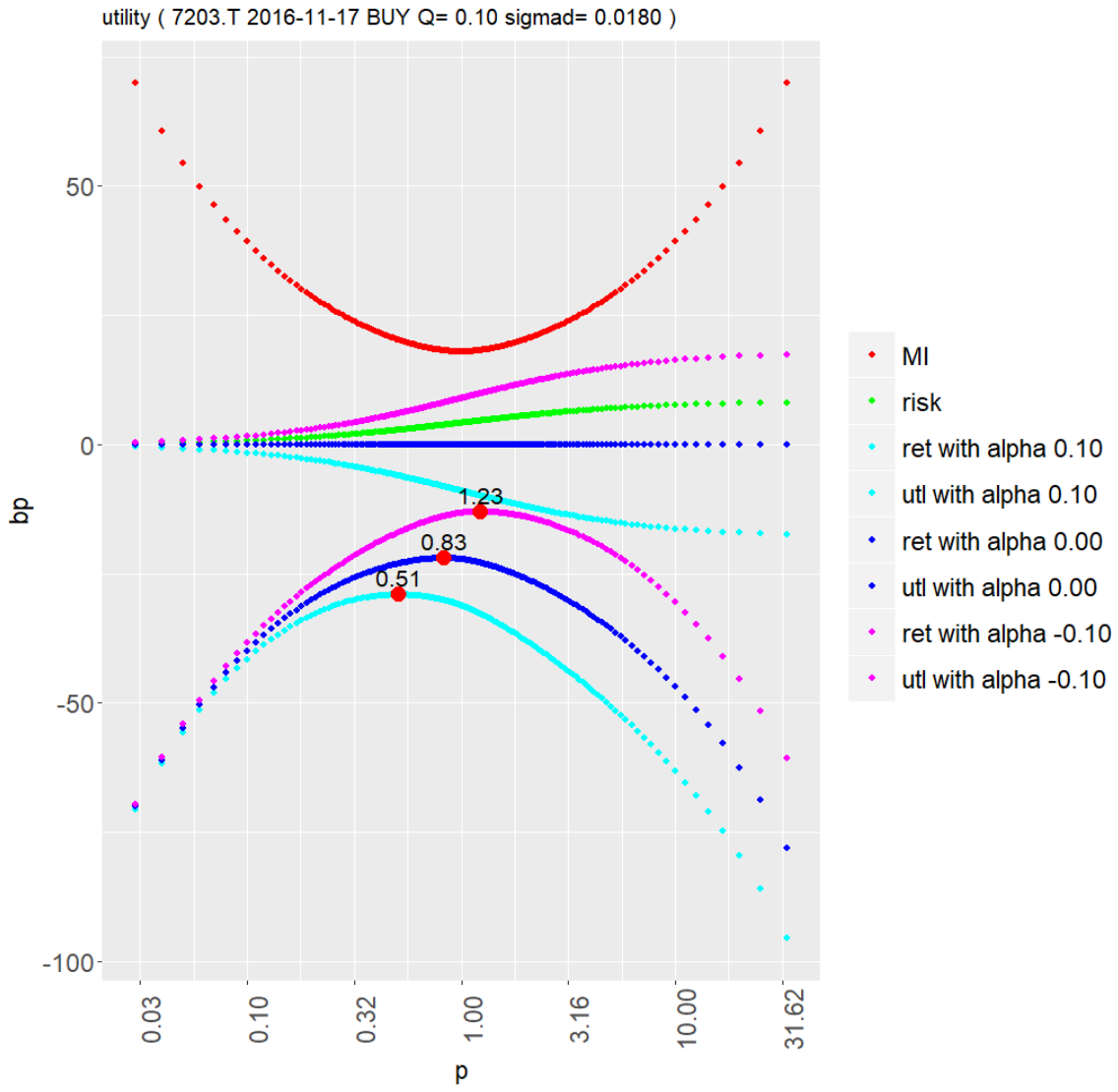

Fig 1. Buy order trading utility optimization. The asset: TSE 7203 on 2016-11-17. The colored curves – the utility and utility components as functions of the parameter p – the controlling parameter of the double reflection power-law transform. The functions' values are in basis points. Red curve – cumulative market impact. Green curve – normalized cumulative risk (i.e., $\lambda\psi^2$). Other curves are cumulative returns and the utility values, when the return forecast is positive, neutral, and negative. The forecasts' alphas are given as fractions of the asset's daily volatility: $0.1\sigma_d$, 0, and $-0.1\sigma_d$ respectively. Red dots on the utility curves, labeled with p values, mark powers p^* maximizing the utilities. The logarithmic scale is used for p axis. The utility values at p=1 correspond to the initial asset accumulation.

The optimal asset accumulation shapes for the three return forecasts are illustrated in Fig 2. Both volume-time and clock-time representations are shown. The curves are obtained by applying the double reflection power-law transform to the initial asset accumulations $h(\tau) = \frac{\tau}{T}$ and $h(t)$ with the parameters p^* calculated above. The initial $h(t)$ – the asset's normalized cumulative volume profile in clock-time. The profile is obtained by averaging intraday volume data over the previous 20 days.

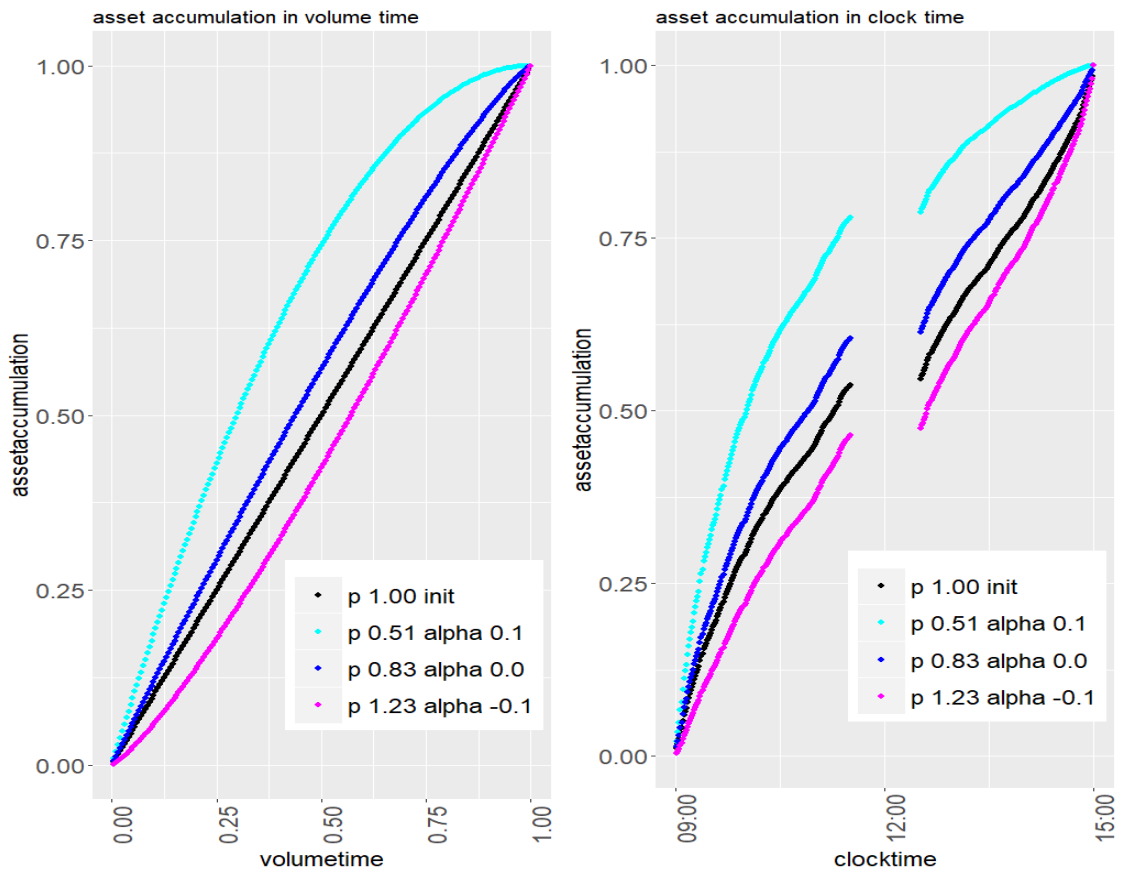

Fig 2. The colored curves – the optimal asset accumulation shapes for the three return forecasts from the above. The volume-time and clock-time dependencies are in the left and right panes respectively. The gaps in the middle of clock-time curves correspond to session break at Tokyo Stock Exchange. The initial asset accumulations (p=1) are illustrated by black curves in both panes.

Conclusion

This work proposes a new solution to the problem of maximizing trading utility. While preserving the overall utility form of (Grinold & Kahn, 1999), we use different formulas for the utility components. The new formulation of the market impact component conforms to widely observed empirical results.

The utility components are parameterized by the asset accumulation function. In the optimization context, the accumulation function is replaced with the double reflection power-law transform of the input asset accumulation controlled by one "shape" parameter; thus making the optimization amenable to the fast and simple numerical approach.

The formulas are developed in volume-time domain facilitating volume profile execution strategies. Nevertheless, the formulations are readily applicable to clock-time domain applications by simple replacing volume-time argument with clock-time. The utility optimization outcome – the optimal value of the asset accumulation transform parameter is seamlessly transferable between volume-time and clock-time.

Appendix A

Below are the proofs of conformity between the market impact formula, proposed in this paper, and the empirical rules for the trading regimes determined by the three parameters of order execution: trading duration – *T*, order size – *Q*, and trading rate – *q* (Skachkov, 2014). The parameters are not independent – each one of them is a function of the other two. The proofs hold for any valid asset accumulation $h(\tau)$; and also for any valid $h(t)$ – by replacing volume-time τ with clock-time *t* in the formulas.

In each of the three trading regimes, we keep constant one parameter (e.g., trading duration: $T_n{=}T_o{=}T$) and work out the change in market impact comparatively to the change in other parameters (e.g., order size: $Q_n = x \cdot Q_o$). Here and below: "*n*" subscript denotes new values after the change, "*o*" subscript denotes old values before the change, "*x*" – the change factor. The trading rate is defined as $q(\tau) = Q \cdot h'(\tau)$, where $h'(\tau)$ – asset accumulation rate, which is the first derivative of the asset accumulation $h(\tau)$.

Constant trading duration – various order size and trading rate

With $T_n = T_o = T$ and the new order size *x* times of the old size: $Q_n = x \cdot Q_o$. The asset accumulation, being a normalized function inside [0,1] interval, is independent of the order size and trading rate: $h_n(\tau) = h_o(\tau) = h(\tau)$.

The new trading rate $q_n(\tau) = Q_n \cdot h'(\tau) = x \cdot Q_o \cdot h'(\tau) = x \cdot q_o(\tau)$

The new market impact $MI_n = \int_0^T \left[d\tau \cdot c \cdot \frac{Q_n}{ADV} \cdot [h'(\tau)]^{1.5} \right] = \int_0^T \left[d\tau \cdot c \cdot \frac{x \cdot Q_o}{ADV} \cdot [h'(\tau)]^{1.5} \right] = x \cdot MI_o$

Market impact is linearly proportional to both trading rate and order size.

Constant order size – various trading duration and rate

With $Q_n = Q_o = Q$ and the new trading duration *x* times of the old duration: $T_n = x \cdot T_o$. The new asset accumulation is the argument scaled variant of the old one: $h_n(\tau) = h_o(\tau/x)$; the first derivative: $h'_n(\tau) = \frac{h'_o(\tau/x)}{x}$

The new trading rate $q_n(\tau) = Q \cdot h'_n(\tau) = Q \cdot \frac{h'_o(\tau/x)}{x} = \frac{q_o(\tau/x)}{x}$

The new market impact $MI_n = \int_0^{T_n} \left[d\tau \cdot c \cdot \frac{Q}{ADV} \cdot [h'_n(\tau)]^{1.5} \right] = \int_0^{x \cdot T_o} \left[d\tau \cdot c \cdot \frac{Q}{ADV} \cdot \frac{[h'_o(\tau/x)]^{1.5}}{x \cdot \sqrt{x}} \right]$

Renaming $\tau/x \to \tau$, $MI_n = \int_0^{x \cdot T_o} \left[d(x \cdot \tau) \cdot c \cdot \frac{Q}{ADV} \cdot \frac{[h'_o(\tau)]^{1.5}}{x \cdot \sqrt{x}} \right]$

Then considering $\int_0^{x \cdot T_o} d(x \cdot \tau) \cdot \ldots = x \cdot \int_0^{T_o} d\tau \cdot \ldots$

$$MI_n = \frac{x}{x \cdot \sqrt{x}} \cdot \int_0^{T_o} \left[d\tau \cdot c \cdot \frac{Q}{ADV} \cdot [h'_o(\tau)]^{1.5} \right] = \frac{MI_o}{\sqrt{x}}$$

Market impact is proportional to the square root of trading rate and inversely proportional to the square root of trade duration.

Constant trading rate – various trading duration and order size

With $T_n = x \cdot T_o$ and $Q_n = x \cdot Q_o$. The asset accumulation and its first derivative are as above: $h_n(\tau) = h_o(\tau/x)$ and $h'_n(\tau) = \frac{h'_o(\tau/x)}{x}$

The new trading rate is an argument scaled variant of the old one:

$$q_n(\tau) = Q_n \cdot h'_n(\tau) = x \cdot Q_o \cdot \frac{h'_o(\tau/x)}{x} = q_o(\tau/x)$$

The strictly constant rate $q_n = q_o = q$ is a subcase of the above. It results in the asset accumulation being a rising straight line. The calculation below holds for this subcase and the general case as well.

The new market impact $MI_n = \int_0^{T_n} \left[d\tau \cdot c \cdot \frac{Q_n}{ADV} \cdot [h'_n(\tau)]^{1.5} \right] = \int_0^{x \cdot T_o} \left[d\tau \cdot c \cdot \frac{x \cdot Q_o}{ADV} \cdot \frac{[h'_o(\tau/x)]^{1.5}}{x \cdot \sqrt{x}} \right]$

Renaming the argument and simplifying the integral as above

$$MI_n = \frac{x \cdot x}{x \cdot \sqrt{x}} \cdot \int_0^{T_o} \left[d\tau \cdot c \cdot \frac{Q_o}{ADV} \cdot [h'_o(\tau)]^{1.5} \right] = \sqrt{x} \cdot MI_o$$

Market impact is proportional to the square root of order size.

References

Almgren & Chriss, Optimal Execution of Portfolio Transactions, 2000

Almgren et al., Direct Estimation of Equity Market Impact, 2005

Grinold & Kahn, Active Portfolio Management, SECOND EDITION, McGraw-Hill, 1999

Skachkov, Market Impact Paradoxes, CFEM Seminar, 2014