Beyond Fill Rates

Adnan Contractor

1 Fill Rate

Fill rates are one of the most widely used metrics for evaluating execution quality in financial markets. Defined as the fraction of an order executed relative to its submitted size, they are often treated as a straightforward indicator of trading success. However, this apparent simplicity masks profound deficiencies. Fill rates, in isolation, fail to capture the intricate trade-offs inherent in execution. They ignore liquidity constraints, spread-induced costs, and the market impact of trading activity.

As a very simple example, consider two identical orders submitted in different markets. In a highly liquid, narrow-spread market, achieving a high fill rate might require little effort or cost. In contrast, in an illiquid, wide-spread market, achieving the same fill rate could involve significant costs, such as adverse price movements or crossing multiple levels of the order book. Without incorporating these dynamics, fill rates offer an incomplete and potentially misleading view of execution quality.

In my opinion, an examination of fill rate should include a nod to

1. Liquidity Elasticity: Captures how the availability of liquidity diminishes with order size, modeled using stochastic processes and integrals.
2. Spread Dynamics: Quantifies the direct costs associated with accessing liquidity, using proportional penalties and dynamic spread adjustments.
3. Market Impact: Models the nonlinear feedback between order size and price movement, penalizing strategies that prioritize immediate fills at excessive costs.

In this paper, I attempt to formulate a nuanced and unified metric that accounts for the complex interdependencies of market microstructure, specifically focusing on the above three areas.

2 Liquidity Elasticity

Liquidity, broadly defined as the capacity of a market to absorb trades without significant price impact, is a dynamic and stochastic resource. Unlike a static quantity, liquidity evolves continuously due to the complex interactions between market participants, external events such as macroeconomic shocks, and the ongoing process of order submissions and cancellations. Understanding liquidity as a function of order size and time is critical for accurately modeling execution efficiency, as larger orders inherently interact with deeper and less liquid portions of the order book.

To formalize liquidity, let $L(x,t)$ denote the cumulative liquidity function, which represents the total available liquidity up to an order size x at time t. This function provides a macroscopic description of the order book by aggregating the volume accessible at progressively larger order sizes. The instantaneous availability of liquidity, or liquidity density, is defined as the negative gradient of the cumulative liquidity:

$$\ell(x,t) = -\frac{\partial L(x,t)}{\partial x}.$$

This definition ensures that $\ell(x,t)$ is non-negative and decreases with x, reflecting the empirical observation that liquidity is more abundant near the top of the book and diminishes as larger orders interact with deeper levels. The negative sign captures the natural depletion of liquidity as a function of order size.

To model the inherent randomness of liquidity, I represent $L(x,t)$ as a stochastic process:

$$L(x,t) = \mathbb{E}[L(x,t)] + \sigma(x,t)W(t),$$

where $\mathbb{E}[L(x,t)]$ is the deterministic component capturing the expected liquidity profile, $\sigma(x,t)$ quantifies the volatility of liquidity as a function of size and time, and $W(t)$ is a standard Brownian motion representing the stochastic fluctuations driven by market dynamics. This formulation explicitly incorporates both the predictable structure of liquidity and its random variations.

Empirical evidence suggests that liquidity density frequently exhibits an exponential decay pattern with respect to order size, particularly in low-liquidity markets. This observation motivates the functional form:

$$\mathbb{E}[\ell(x)] = L_0 e^{-\gamma x},$$

where L_0 represents the baseline liquidity density at the top of the book, and $\gamma > 0$ is the decay parameter. The parameter γ serves as a measure of market resilience: smaller values indicate gradual liquidity depletion, characteristic of robust markets, while larger values reflect steeper decay, signaling fragility. This exponential model is consistent with observations that deeper levels of the order book are sparsely populated, particularly in volatile or illiquid markets.

From this density function, the expected cumulative liquidity available for an order of size Q is obtained by integration:

$$\mathbb{E}[L(Q)] = \int_0^Q L_0 e^{-\gamma x}\, dx.$$

Evaluating this integral yields:

$$\mathbb{E}[L(Q)] = \frac{L_0}{\gamma}\left(1 - e^{-\gamma Q}\right).$$

This result encapsulates the nonlinear relationship between order size and liquidity consumption. For small orders, where $Q \ll 1/\gamma$, the term $e^{-\gamma Q}$ approaches unity, and the cumulative liquidity scales approximately linearly:

$$\mathbb{E}[L(Q)] \approx L_0 Q.$$

As Q increases, the cumulative liquidity asymptotically approaches the finite value $\frac{L_0}{\gamma}$, reflecting the exhaustion of accessible liquidity deeper in the order book. This asymptotic behavior highlights the diminishing returns associated with larger orders in markets characterized by steep liquidity decay.

To evaluate execution efficiency in this context, I introduce the Liquidity-Adjusted Fill Efficiency (LAFE), which normalizes the fill rate by the cumulative liquidity consumed. LAFE is defined as:

$$\text{LAFE} = \frac{\mathbb{E}[L(Q_{\text{filled}})]}{\mathbb{E}[L(Q_{\text{submitted}})]}.$$

Substituting the expression for $\mathbb{E}[L(Q)]$, LAFE becomes:

$$\text{LAFE} = \frac{\frac{L_0}{\gamma}\left(1 - e^{-\gamma Q_{\text{filled}}}\right)}{\frac{L_0}{\gamma}\left(1 - e^{-\gamma Q_{\text{submitted}}}\right)} = \frac{1 - e^{-\gamma Q_{\text{filled}}}}{1 - e^{-\gamma Q_{\text{submitted}}}}.$$

This expression reveals that LAFE depends solely on the relative sizes of Q_{filled} and $Q_{\text{submitted}}$, as well as the decay parameter γ. The metric adjusts the fill rate to account for the nonlinear depletion of liquidity, offering a more nuanced evaluation of execution efficiency.

The introduction of γ in the LAFE framework is critical. It provides a quantitative measure of market resilience, with higher γ values indicating markets where liquidity depletes rapidly, increasing the cost of executing large orders. LAFE also enables cross-market comparisons by normalizing fill rates to the liquidity consumed, ensuring that traders in less resilient markets are not unfairly penalized for lower

fill rates. Finally, LAFE highlights the diminishing returns of larger orders in steeply decaying liquidity environments, where the marginal cost of accessing additional volume grows exponentially.

Consider two markets with identical baseline liquidity, L_0, but different decay parameters, $\gamma_1 = 0.2$ and $\gamma_2 = 0.5$. For an order of size $Q = 10$, the expected cumulative liquidity consumed is:

$$\mathbb{E}[L(10)]_{\gamma_1} = \frac{L_0}{0.2}\left(1 - e^{-2}\right) \approx 0.864\frac{L_0}{0.2},$$

$$\mathbb{E}[L(10)]_{\gamma_2} = \frac{L_0}{0.5}\left(1 - e^{-5}\right) \approx 0.993\frac{L_0}{0.5}.$$

The higher liquidity consumption in the steeper decay market, γ_2, underscores the increased difficulty of executing large orders in less resilient environments.

To extend the LAFE framework to dynamic settings, I redefine the liquidity density as a time-dependent function:

$$\ell(x, t) = L_0(t)e^{-\gamma(t)x},$$

where $L_0(t)$ and $\gamma(t)$ evolve with time. This dynamic model accommodates changing market conditions, such as volatility spikes or liquidity replenishment. The cumulative liquidity then becomes:

$$\mathbb{E}[L(Q, t)] = \int_0^Q L_0(t)e^{-\gamma(t)x}\,dx = \frac{L_0(t)}{\gamma(t)}\left(1 - e^{-\gamma(t)Q}\right).$$

This extension enables real-time adjustments to execution strategies, allowing traders to respond dynamically to transient liquidity shocks or evolving market resilience.

The Liquidity-Adjusted Fill Efficiency framework provides a robust tool for quantifying execution efficiency in terms of liquidity dynamics. By explicitly modeling the nonlinear depletion of liquidity, LAFE offers profound insights into the trade-offs inherent in executing large orders and navigating diverse market conditions.

3 Spread Dynamics

The bid-ask spread, denoted S, serves as a critical measure of transaction costs in financial markets. It represents the price difference between the best available ask and bid quotes and functions as a cost of immediacy for liquidity takers. When a trader submits a market order, they cross the spread to execute the trade, paying the ask price for a buy order or accepting the bid price for a sell order. This transaction imposes a direct cost that scales with the order size Q and the relative magnitude of the spread $\frac{S}{P}$, where P is the midpoint price of the security. While the Liquidity-Adjusted Fill Efficiency (LAFE) accounts for liquidity availability and consumption, it does not capture these financial costs. Incorporating spread-induced costs is essential to fully understand execution efficiency, particularly in volatile or illiquid markets where spreads can widen significantly.

To formalize this, let the total spread cost incurred for an executed order of size Q be expressed as:

$$C_{\text{spread}}(Q) = Q \cdot \frac{S}{P}.$$

This formulation assumes that liquidity within the spread is uniformly distributed. The linear dependence of $C_{\text{spread}}(Q)$ on Q and $\frac{S}{P}$ is intuitive but oversimplified, as real-world order books often exhibit non-uniform liquidity profiles. To generalize, we consider the liquidity density function $f(x)$, which represents the marginal liquidity available at a price distance x from the midpoint. The spread cost then becomes:

$$C_{\text{spread}}(Q) = \int_0^Q \frac{S}{P} f(x)\,dx,$$

where $f(x)$ must satisfy the normalization condition:

$$\int_0^S f(x)\,dx = 1.$$

Assuming uniform liquidity across the spread, $f(x) = \frac{1}{S}$ for $x \in [0, S]$, the spread cost simplifies to:

$$C_{\text{spread}}(Q) = Q \cdot \frac{S}{P}.$$

This uniform case provides a baseline for spread cost estimation but fails to account for liquidity clustering, which often occurs near the best bid and ask prices. Empirical studies suggest that $f(x)$ frequently follows an exponential decay pattern:

$$f(x) = \frac{\lambda}{1 - e^{-\lambda S}} e^{-\lambda x}, \quad x \in [0, S].$$

This distribution captures the fact that liquidity is densest at the quotes and tapers off deeper into the spread. The normalization factor $\frac{\lambda}{1-e^{-\lambda S}}$ ensures that $f(x)$ integrates to unity over the interval $[0, S]$. Substituting this form of $f(x)$, the spread cost becomes:

$$C_{\text{spread}}(Q) = \frac{S}{P} \cdot \frac{\lambda}{1 - e^{-\lambda S}} \int_0^Q e^{-\lambda x}\, dx.$$

Evaluating the integral yields:

$$C_{\text{spread}}(Q) = \frac{S}{P} \cdot \frac{\lambda}{1 - e^{-\lambda S}} \left(1 - e^{-\lambda Q}\right).$$

This result generalizes the linear case by introducing nonlinear corrections that reflect the decay of liquidity density within the spread. For small Q, the term $e^{-\lambda Q}$ approaches unity, and the spread cost scales linearly with order size:

$$C_{\text{spread}}(Q) \approx Q \cdot \frac{S}{P}, \quad \text{if } Q \ll \frac{1}{\lambda}.$$

For larger orders, the nonlinear term dominates, reflecting the compounding effect of liquidity depletion.

To incorporate these costs into the evaluation of execution efficiency, I define the Spread-Adjusted Fill Efficiency (SAFE) metric as:

$$\text{SAFE} = \frac{\text{LAFE}}{1 + \frac{S}{P}}.$$

This formulation introduces a proportional penalty, $1 + \frac{S}{P}$, that adjusts LAFE according to the relative spread cost. In narrow-spread markets, where $\frac{S}{P} \to 0$, the penalty term approaches unity, and SAFE converges to LAFE:

$$\text{SAFE} \approx \text{LAFE}, \quad \text{if } \frac{S}{P} \to 0.$$

This limiting case reflects the negligible cost of crossing the spread in such environments, leaving liquidity consumption as the primary determinant of execution efficiency. Conversely, in wide-spread markets, where $\frac{S}{P} \to \infty$, the penalty term grows without bound, leading to:

$$\text{SAFE} \to \frac{\text{LAFE}}{\frac{S}{P}}, \quad \text{as } \frac{S}{P} \to \infty.$$

This asymptotic behavior captures the intuitive inefficiency of execution under high transaction costs, emphasizing the importance of liquidity conservation in these markets.

The sensitivity of SAFE to changes in $\frac{S}{P}$ can be explored by analyzing its derivative:

$$\frac{\partial \text{SAFE}}{\partial \frac{S}{P}} = -\frac{\text{LAFE}}{\left(1 + \frac{S}{P}\right)^2}.$$

The negative sign indicates that an increase in the relative spread always reduces SAFE, with the magnitude of this reduction inversely proportional to the square of the penalty term. This convex relationship implies that wider spreads impose increasingly severe penalties, further discouraging aggressive execution strategies.

Spreads are not static but evolve dynamically with market conditions such as volatility, order flow imbalances, and liquidity shocks. Let $S(t)$ denote the spread at time t, and $P(t)$ the corresponding midpoint price. The time-dependent SAFE metric is then given by:

$$\text{SAFE}(t) = \frac{\text{LAFE}}{1 + \frac{S(t)}{P(t)}}.$$

To analyze this dynamic behavior, suppose $S(t)$ follows a mean-reverting process:

$$S(t) = S_0 + \sigma_S W(t),$$

where S_0 is the long-term mean of the spread, σ_S quantifies spread volatility, and $W(t)$ is a standard Brownian motion. Substituting this into the SAFE formula, the expected SAFE value becomes:

$$\mathbb{E}[\text{SAFE}(t)] = \frac{\text{LAFE}}{\mathbb{E}\left[1 + \frac{S_0 + \sigma_S W(t)}{P(t)}\right]}.$$

If $P(t)$ is assumed constant, the penalty term exhibits mean-reverting behavior, increasing during periods of heightened volatility and reducing execution efficiency. By anticipating changes in $S(t)$, traders can time their orders to optimize SAFE.

The integration of spread costs into the execution efficiency framework reveals profound trade-offs between liquidity consumption and transaction costs. While LAFE quantifies the availability and depletion of liquidity, SAFE contextualizes this measure within the cost framework imposed by the spread. Markets with narrow spreads and high liquidity density naturally exhibit higher SAFE values, indicating cost-efficient access to liquidity. Conversely, wide-spread markets impose dual penalties, reducing both LAFE and SAFE, and emphasizing the need for timing and precision in execution strategies.

These results underscore the importance of considering both structural and dynamic factors when evaluating execution efficiency. The SAFE metric provides a powerful tool for navigating the complexities of modern markets, offering insights that extend beyond raw fill rates to encompass the full spectrum of transaction costs.

4 Market Impact

Market impact refers to the price movement induced by executing a trade. Unlike the bid-ask spread, which reflects a static transaction cost, market impact is inherently dynamic, evolving as orders interact with the liquidity landscape of the order book. This phenomenon arises from the depletion of liquidity and the displacement of quotes, creating a feedback loop between the size of a trade and its resulting cost. Understanding market impact is crucial for quantifying execution efficiency, as it often dominates transaction costs for large or aggressive orders.

To formalize market impact, let $\Delta P(Q)$ denote the price displacement caused by executing an order of size Q. Empirical and theoretical evidence suggests that market impact follows a nonlinear relationship with order size, typically characterized by a power-law function:

$$\Delta P(Q) = \alpha Q^{\delta},$$

where α is a sensitivity parameter reflecting market conditions, and $\delta \in (0, 1]$ governs the degree of nonlinearity. The exponent δ captures the diminishing returns effect, where the marginal impact of additional volume decreases as the order size grows. For small orders ($Q \to 0$), the relationship is nearly linear ($\delta \approx 1$), while for larger orders, the impact grows sub-linearly ($\delta < 1$).

The parameters α and δ are not constants; they depend on factors such as market liquidity, volatility, and structure. For instance, in highly liquid markets, α is small, indicating that large trades can be executed with minimal price displacement. Conversely, in illiquid markets, α is significantly larger, reflecting the increased cost of execution. The parameter δ encapsulates market resilience: higher values correspond

to environments where price impact grows rapidly with order size, while lower values indicate markets capable of absorbing trades more efficiently.

The total cost incurred due to market impact is the product of the price displacement per share and the total number of shares executed:

$$C_{\text{impact}}(Q) = Q \cdot \Delta P(Q) = \alpha Q^{1+\delta}.$$

This expression highlights the compounding nature of market impact, where the cost scales super-linearly with order size. Large trades incur disproportionately higher costs, necessitating strategies that minimize Q or distribute the order over time to mitigate these effects.

To analyze the sensitivity of market impact to changes in order size, consider the first derivative of $\Delta P(Q)$:

$$\frac{\partial \Delta P(Q)}{\partial Q} = \alpha \delta Q^{\delta-1}.$$

This derivative reveals that the marginal price impact decreases as Q increases, with the rate of decrease governed by δ. In highly resilient markets ($\delta \to 0$), marginal impact diminishes rapidly, whereas in fragile markets ($\delta \to 1$), the marginal impact remains relatively constant, amplifying the cost of executing large trades.

The second derivative provides further insights into the curvature of the impact function:

$$\frac{\partial^2 \Delta P(Q)}{\partial Q^2} = \alpha \delta(\delta - 1) Q^{\delta-2}.$$

For $\delta < 1$, this second derivative is negative, confirming that the impact function is concave. This concavity reflects the diminishing sensitivity of price impact to additional order size, a critical feature for understanding the trade-offs in execution strategies.

To incorporate market impact into the evaluation of execution efficiency, I introduce the Impact-Adjusted Fill Efficiency (IAFE). This metric penalizes aggressive orders that induce significant price displacement, modifying the SAFE metric to account for impact. IAFE is defined as:

$$\text{IAFE} = \text{SAFE} \cdot \exp(-\beta A),$$

where β is a penalty parameter, and A is the aggression score, given by:

$$A = \frac{\Delta P(Q)}{S} = \frac{\alpha Q^{\delta}}{S}.$$

The aggression score normalizes the price impact by the bid-ask spread S, creating a dimensionless measure of the trade's disruptiveness relative to the static cost of crossing the spread. Higher values of A correspond to more aggressive trades, where the induced impact significantly exceeds the spread cost.

The exponential penalty term in the IAFE formulation ensures that the metric responds nonlinearly to aggression. For moderate trades, where $A \ll 1$, the penalty term approximates unity, and IAFE converges to SAFE:

$$\text{IAFE} \approx \text{SAFE}, \quad \text{if } A \to 0.$$

For aggressive trades, where $A \gg 1$, the penalty term reduces IAFE exponentially, emphasizing the importance of mitigating impact in execution strategies:

$$\text{IAFE} \to \frac{\text{SAFE}}{\exp(\beta A)}, \quad \text{as } A \to \infty.$$

To understand how IAFE behaves in different market conditions, consider a trader submitting an order of size $Q = 1000$ in two markets with identical spreads but different impact parameters. In the first market, $\alpha = 0.01$ and $\delta = 0.5$; in the second, $\alpha = 0.05$ and $\delta = 0.8$. The aggression scores are:

$$A_1 = \frac{0.01 \cdot 1000^{0.5}}{S}, \quad A_2 = \frac{0.05 \cdot 1000^{0.8}}{S}.$$

Substituting these values into the IAFE formula, the efficiency in the second market is significantly lower, reflecting the amplified cost of execution in the higher-impact environment.

The dynamic nature of market impact necessitates time-dependent modeling. Let $\alpha(t)$ and $\delta(t)$ represent time-varying parameters influenced by transient liquidity conditions and volatility. The time-dependent price impact is then expressed as:

$$\Delta P(Q,t) = \alpha(t) Q^{\delta(t)}.$$

The corresponding dynamic IAFE is given by:

$$\text{IAFE}(t) = \frac{\text{SAFE}}{\exp\left(\beta \frac{\alpha(t) Q^{\delta(t)}}{S(t)}\right)}.$$

This formulation enables real-time adjustments to execution strategies. For example, during periods of high liquidity replenishment, $\alpha(t)$ decreases, reducing the penalty term and improving IAFE. Conversely, during periods of heightened volatility, where both $\alpha(t)$ and $\delta(t)$ increase, IAFE declines, signaling higher execution costs and the need for conservative strategies.

By integrating market impact into the execution efficiency framework, IAFE provides a comprehensive measure that accounts for liquidity consumption, spread costs, and dynamic price displacement. This metric offers profound insights into the trade-offs inherent in execution strategies, highlighting the importance of balancing aggression with efficiency in diverse market conditions.

5 Putting it All Together

The metrics developed in the preceding sections—Liquidity-Adjusted Fill Efficiency (LAFE), Spread-Adjusted Fill Efficiency (SAFE), and Impact-Adjusted Fill Efficiency (IAFE)—capture distinct and essential dimensions of execution efficiency. LAFE quantifies the availability and consumption of liquidity, SAFE incorporates the transaction costs imposed by the bid-ask spread, and IAFE penalizes the adverse effects of market impact. While individually insightful, these metrics do not fully encapsulate the interplay between their components. To address this, I define a unified metric, denoted Φ, that integrates liquidity consumption, spread costs, and market impact into a single framework.

Recall the definitions of LAFE, SAFE, and IAFE:

$$\text{LAFE} = \frac{1 - e^{-\gamma Q_{\text{filled}}}}{1 - e^{-\gamma Q_{\text{submitted}}}}, \quad \text{SAFE} = \frac{\text{LAFE}}{1 + \frac{S}{P}}, \quad \text{IAFE} = \frac{\text{SAFE}}{\exp\left(\beta \frac{\alpha Q^{\delta}}{S}\right)}.$$

Here, γ represents the liquidity decay parameter, S is the bid-ask spread, P is the midpoint price, α and δ govern the magnitude and nonlinearity of market impact, and β is a penalty parameter scaling the impact.

To develop a composite metric which incorporates all of the above, we define a **uniform execution efficiency** Φ by

$$\Phi = \frac{\frac{\frac{1 - e^{-\gamma Q_{\text{filled}}}}{1 - e^{-\gamma Q_{\text{submitted}}}}}{1 + \frac{S}{P}}}{\exp\left(\beta \frac{\alpha Q^{\delta}}{S}\right)}.$$

Simplifying, the unified metric is expressed as:

$$\Phi = \frac{\left(1 - e^{-\gamma Q_{\text{filled}}}\right)}{\left(1 - e^{-\gamma Q_{\text{submitted}}}\right)\left(1 + \frac{S}{P}\right)\exp\left(\beta \frac{\alpha Q^{\delta}}{S}\right)}.$$

This formulation highlights the multiplicative structure of the penalties associated with liquidity depletion, spread costs, and market impact. The numerator, $1 - e^{-\gamma Q_{\text{filled}}}$, captures the efficiency of liquidity

consumption, weighted by the decay parameter γ. The first denominator term, $1 + \frac{S}{P}$, introduces a scaling factor for the transaction costs imposed by the spread. The second denominator term, $\exp\left(\beta \frac{\alpha Q^{\delta}}{S}\right)$, penalizes aggressive execution strategies that induce significant market impact. By integrating these components, Φ provides a holistic evaluation of execution efficiency.

The utility of Φ is best understood by examining its limiting behavior under different market conditions. In highly liquid markets, where γ is small, $S \ll P$, and α and δ are minimal, the penalty terms in the denominator approach unity, and Φ converges to:

$$\Phi \approx 1 - e^{-\gamma Q_{\text{filled}}}.$$

This reflects the dominance of liquidity efficiency in such environments, where spread and market impact costs are negligible. Conversely, in illiquid markets with large γ, wide spreads $S \gg P$, and significant market impact $\alpha Q^{\delta} \gg S$, the denominator terms dominate, severely reducing Φ. This behavior underscores the compounding inefficiencies associated with liquidity depletion, high transaction costs, and aggressive execution.

To illustrate the behavior of Φ, consider a market with $\gamma = 0.3$, $S/P = 0.02$, $\alpha = 0.01$, $\delta = 0.7$, and $\beta = 1.5$. For an order size $Q = 500$, the components are calculated as:

$$\text{LAFE} = \frac{1 - e^{-0.3 \cdot 500}}{1 - e^{-0.3 \cdot 1000}} \approx 0.645,$$

$$1 + \frac{S}{P} = 1 + 0.02 = 1.02,$$

$$\exp\left(\beta \frac{\alpha Q^{\delta}}{S}\right) = \exp\left(1.5 \cdot \frac{0.01 \cdot 500^{0.7}}{0.02}\right) \approx \exp(1.84).$$

Substituting these values into the Φ formula, the unified efficiency metric is:

$$\Phi \approx \frac{0.645}{1.02 \cdot \exp(1.84)} \approx 0.230.$$

This low value reflects the compounded inefficiency caused by high liquidity consumption, non-negligible spread costs, and significant market impact.

The unified metric can also accommodate dynamic market conditions. Let $\gamma(t)$, $S(t)$, $\alpha(t)$, and $\delta(t)$ represent time-dependent parameters that evolve in response to transient liquidity changes, spread fluctuations, and market volatility. The dynamic unified metric is then expressed as:

$$\Phi(t) = \frac{\left(1 - e^{-\gamma(t) Q_{\text{filled}}}\right)}{\left(1 - e^{-\gamma(t) Q_{\text{submitted}}}\right)\left(1 + \frac{S(t)}{P(t)}\right) \exp\left(\beta \frac{\alpha(t) Q^{\delta(t)}}{S(t)}\right)}.$$

This time-dependent extension enables real-time adjustments to execution strategies. For instance, during periods of high liquidity replenishment, $\gamma(t)$ decreases, reducing the numerator penalty and improving $\Phi(t)$. Conversely, widening spreads $S(t)$ amplify the denominator penalty, signaling the need to reduce order aggression.

In summary, the unified metric Φ encapsulates the interplay between liquidity consumption, transaction costs, and market impact. Its multiplicative structure ensures that the penalties associated with each component are compounded, reflecting the full complexity of execution efficiency. By integrating these dimensions into a single framework, Φ offers a powerful tool for evaluating and optimizing trading strategies across diverse market conditions.

6 Implications for Modern Markets

Traditional measures of execution performance, such as fill rates, fail to capture the nuanced trade-offs inherent in modern markets. A simple fill rate might suggest high efficiency, yet conceal the steep

costs incurred through liquidity depletion, wide spreads, or excessive market impact. Conversely, low fill rates might indicate a carefully calibrated approach that minimizes transaction costs while preserving market stability. The unified efficiency metric Φ transforms fill rates from a simplistic measure into a sophisticated evaluation framework, offering actionable insights into the interplay of liquidity, cost, and execution strategy.

Liquidity elasticity, modeled through the decay parameter γ, defines the structural constraints of the market, penalizing strategies that overreach and exhaust available liquidity. Spread dynamics, encapsulated in the term $1 + \frac{S}{P}$, provide a measure of transaction cost efficiency, highlighting the cost of immediacy in accessing liquidity. Market impact, modeled as a power-law function of order size, penalizes aggressive strategies that induce adverse price movements, emphasizing the importance of balancing speed and stability.

Traders operating in highly liquid markets with narrow spreads can achieve higher Φ values, reflecting cost-efficient access to liquidity. Conversely, traders in illiquid or volatile environments face greater penalties, encouraging conservative strategies that minimize impact.

The metric Φ informs the design of advanced trading algorithms. By quantifying the penalties associated with liquidity consumption, spread costs, and market impact, Φ serves as an objective function for optimizing order execution. In markets with rapidly widening spreads, the penalty term $1 + \frac{S}{P}$ increases sharply, signaling the need to reduce order aggression and delay execution. Similarly, during periods of high liquidity replenishment, a declining decay parameter γ indicates improved market resilience, enabling larger orders to be executed with minimal penalties.

Markets with persistently low Φ values may indicate structural inefficiencies, such as monopolistic market-making behavior or inadequate liquidity provision. Conversely, markets that consistently deliver high Φ values demonstrate cost-efficient and stable environments that promote equitable access to liquidity.

Future work should explore extensions of the framework, such as integrating additional dimensions of market microstructure, incorporating techniques for parameter estimation, and applying the metric to emerging asset classes like cryptocurrencies.