Quantifying the Execution-Time Gap in Latency Arbitrage Backtests:

A Mathematical Framework with Empirical Validation

Boris Fesenko
BJF Trading Group Inc., Ontario, Canada
boris@bjftradinggroup.com

June 2026

Abstract

Latency arbitrage backtests for retail foreign exchange and cryptocurrency markets routinely assume zero-latency order fills, so the simulated edge per opportunity matches the price differential observed at the moment of signal. Production execution introduces a deterministic round-trip time T between signal and confirmed fill. During T the arbitrage window can close, the price differential can decay toward zero as the slow venue catches up, or both events can occur jointly. The realised edge per opportunity is therefore smaller than the backtest edge, and the gap grows with T. This paper develops a mathematical framework that quantifies the gap between backtest edge E_b and production edge $E_a(T)$ as a function of T, the distribution of opportunity window duration W, and a within-window decay model for the price differential. Under three decay regimes (step, linear, exponential) and three duration distributions (exponential, uniform, log-normal) we derive closed-form or quasi-closed-form expressions for the retention ratio $\mathcal{R}(T) = E_a(T)/E_b$ and for the viability threshold T^* at which expected per-trade edge equals expected per-trade cost. We validate the framework by Monte Carlo simulation calibrated to the BEQI broker-execution dataset (Fesenko, 2026), and we cross-check predicted retention ratios against simulated retention ratios. Practical implications for backtest interpretation, broker selection, and the design of latency-aware strategy tests are discussed. The framework gives strategy designers a closed-form lens on the question that every latency arbitrage backtester implicitly asks: how much of the simulated edge survives a realistic execution path.

Keywords: latency arbitrage, backtest fidelity, market microstructure, retail forex execution, execution timing, opportunity window, retention ratio.

JEL classification: G14, G15, G20, C53, C81.

1 Introduction

A latency arbitrage strategy on a retail foreign exchange or cryptocurrency venue is, at its heart, a race against a single physical clock. A fast price source (a tier-one institutional feed, an aggregated cross-venue book, a fast crypto exchange) shifts. The slow venue (a retail forex broker, a slower exchange) has not yet repriced. For a brief window W measured in tens of milliseconds, the gap is exploitable. After the slow venue updates its quote, the window closes.

The strategy succeeds if the trader's order arrives at the slow venue while the window is still open and the gap has not yet collapsed.

This stylised description sits at the foundation of every latency arbitrage backtest. Yet most backtests, in both academic and practitioner work, share a structural simplification that this paper formalises and quantifies. The simplification is that fills occur at the price observed at signal time, with zero latency. The backtest implicitly assumes that the moment a tradeable gap is detected, the trade is filled at that gap. Under this assumption, the simulated edge per opportunity equals the price differential at signal time, p_0, and the expected edge per simulated opportunity is $E_b = \mathbb{E}[p_0]$.

In production, no such instantaneous fill exists. The order leaves the trader's machine, traverses the network, reaches the venue's matching engine, is processed, and a fill confirmation returns. The total round-trip time is T. During T, two distinct effects degrade the captured edge:

1. The window may close before the order arrives. If the window duration $W < T$, the price has snapped back, and the order fills at or near the post-update price. The captured edge for this opportunity is approximately zero (or negative if the snap-back overshoots).
2. Even if $W > T$, the price differential at fill time, $p(T)$, may be smaller than the signal-time differential $p(0) = p_0$, because the slow venue updates partway through the window. The captured edge is $p(T)$, not p_0.

The gap between the backtest edge E_b and the production edge $E_a(T)$ is the *execution-time gap*. The contribution of this paper is a mathematical framework that expresses the execution-time gap in closed form for a parametric family of duration distributions and decay models, and an empirical validation of the framework against simulated executions calibrated to a real-world broker-execution dataset.

The framework has three immediate practical uses. First, it gives strategy designers a closed-form prediction of how much simulated edge survives at any realistic round-trip time. A backtest that shows a 0.6 pip edge per opportunity at zero latency can be converted, given a measured opportunity duration distribution, into an expected production edge at the trader's actual round-trip time. Second, it gives broker-selection analysts a principled basis for ranking brokers not by raw fill latency alone, but by retention ratio $\mathcal{R}(T) = E_a(T)/E_b$, which weights latency by its effect on the strategy class actually under consideration. Third, it gives backtester authors a parametrised slippage model that is theoretically grounded rather than ad hoc.

The remainder of the paper is organised as follows. Section 2 surveys related work in market microstructure, latency-aware backtest design, and the empirical literature on retail forex execution quality. Section 3 develops notation, assumptions, and three decay models for the within-window price differential. Section 4 presents closed-form retention ratios for each (decay model, duration distribution) pair we examine, together with sensitivity and viability threshold results. Section 5 validates the framework by Monte Carlo simulation against the BEQI calibration, and reports residual error between closed-form predictions and simulated retention. Section 6 interprets the results for backtest design, broker selection, and strategy viability. Section 7 discusses limitations and Section 8 sketches future extensions. Section 9 concludes. Appendix A collects the proofs; Appendix B documents the numerical methods used in the empirical section.

2 Related Work

The framework developed below sits at the intersection of three established literatures.

2.1 Market microstructure and optimal execution

The classical microstructure literature (Kyle, 1985; Glosten and Milgrom, 1985; Stoll, 1978; Ho and Stoll, 1981) treats the bid-ask spread as compensation for inventory and adverse-selection costs and treats the price impact of a trade as a function of trade size and information content. The Almgren and Chriss (2001) optimal-execution framework extends this by formalising the cost of executing a large order over time as a trade-off between price impact (favoring slow execution) and inventory risk (favoring fast execution). The framework in the present paper is structurally similar in form, in that it expresses the realised edge as an expectation over a time-dependent payoff, but the question it asks is different. Almgren and Chriss (2001) optimise the schedule under which a large order is executed; the present paper asks how much edge is lost when a small order is executed at a fixed round-trip latency T.

Madhavan (2000) and O'Hara (2015) survey the broader microstructure literature and identify execution timing as a first-order determinant of realised return for short-horizon strategies. The execution-time gap formalised here is one concrete mechanism by which the timing-versus-return relationship is mediated.

2.2 High-frequency trading and latency arms races

A substantial empirical literature documents how the rise of high-frequency trading reshaped equity market microstructure. Brogaard et al. (2014) show that HFT contributes to price discovery in equities; Hasbrouck and Saar (2013) measure the latency cost of low-latency strategy participation; Easley et al. (2012) develop the VPIN flow-toxicity measure for the high-frequency regime.

The arms-race aspect of latency competition is theorised by Budish et al. (2015), who propose frequent batch auctions as a market-design response, and by Aquilina et al. (2022), who measure the empirical scale of latency-arbitrage rents in equity markets. Their dollar-rent measurement is methodologically distinct from the present paper, which quantifies the per-opportunity edge that the latency racer expects to capture as a function of its own latency T, rather than the aggregate transfer between fast and slow market participants.

2.3 Backtest fidelity and slippage modelling

The methodological literature on backtest design has long emphasised the gap between simulated and live trading performance. Bailey et al. (2014) discuss backtest overfitting and the deflated Sharpe ratio; López de Prado (2018) collects practical guidance on backtest construction. Slippage and execution costs are treated, at varying levels of formality, in Almgren (2003), Cont et al. (2014), and Oomen (2017). The present paper contributes to this literature by isolating one specific mechanism (the execution-time gap) and quantifying it in closed form rather than as a parametric add-on to a backtest.

2.4 Retail forex execution quality

The retail foreign exchange microstructure literature is smaller than its institutional counterpart, but King et al. (2012), Evans and Lyons (2002), and Lyons (2001) develop the relevant framework for FX order flow and dealer pricing. Bessembinder (1994) on FX spread determination, and Bollerslev and Domowitz (1993) on intraday trading patterns in interbank FX, remain foundational references. Amihud and Mendelson (1986) and Foucault et al. (2003) provide the broader asset-pricing context in which spread costs interact with realised return. The companion paper to the present work, Fesenko (2026), introduces the BEQI (Broker Execution Quality Index) methodology for retail forex broker measurement, which we use here for empirical calibration. The execution-time gap framework developed below complements BEQI by extending the analysis from broker-level execution quality to the strategy-level question of how much of a backtested edge a given broker's execution quality preserves.

2.5 Statistical arbitrage and pair trading

Although the present paper focuses on latency arbitrage, the framework extends to statistical arbitrage and pair trading because both involve a notion of opportunity window. Lehmann (1990), Engle and Granger (1987), and Vidyamurthy (2004) develop the cointegration foundation for pairs trading; Gatev et al. (2006) and Lo et al. (2000) provide widely-cited empirical studies of pair trading and technical-analysis-based strategies, both of which rest on opportunity-window detection in noisy data. The execution-time question for stat arb is different in scale (windows of seconds to minutes rather than tens of milliseconds), but the structure is identical: a backtest assumes a fill at signal time, production fills happen after a round-trip time, and the realised edge is less than the simulated edge.

3 Problem Formalisation

This section develops the notation and assumptions of the framework. We work in continuous time. All quantities are non-negative random variables on a common probability space, and all expectations are with respect to the population of arbitrage opportunities (the index over opportunities is suppressed).

3.1 Primitive notation

Definition 3.1 (Arbitrage opportunity window)**.** An *arbitrage opportunity window* is a half-open time interval $[s, s+W)$, where $s \in \mathbb{R}$ is the wall-clock opening time of the window and $W > 0$ is its duration. The opening time s is the moment a price differential between a fast feed and a slow venue first exceeds an execution-cost threshold; the closing time $s+W$ is the moment the differential falls back below that threshold. Without loss of generality we set $s = 0$ and work on the interval $[0, W)$.

Definition 3.2 (Within-window price differential)**.** The *within-window price differential* is a non-negative function $p : [0, W) \to \mathbb{R}_{\geq 0}$ giving the executable price differential at each time t inside the window. The boundary value $p(0) = p_0$ is the differential at signal time; $\lim_{t \to W^-} p(t) = 0$.

Definition 3.3 (Round-trip time and fill time)**.** The *round-trip time* $T \geq 0$ is the deterministic time between order submission at $t = 0$ and confirmed fill at the slow venue. If $T < W$, the fill occurs inside the window and the captured price differential is $p(T)$. If $T \geq W$, the fill occurs after the window has closed; the captured differential is approximately zero (and may be negative in expectation if a snap-back overshoots; we set it to zero here and discuss overshoot in Section 7).

Definition 3.4 (Backtest edge and production edge)**.** The *backtest edge per opportunity* is $E_b = \mathbb{E}[p_0]$, the expectation of the signal-time differential under the assumption of zero-latency fills. The *production edge per opportunity at round-trip time T* is

$$E_a(T) = \mathbb{E}\left[\, p(T)\, \mathbf{1}\{W > T\} \,\right], \tag{1}$$

where the expectation is taken jointly over the distribution of W and any random elements in the decay path $p(\cdot)$. The *retention ratio* is

$$\mathcal{R}(T) = \frac{E_a(T)}{E_b}. \tag{2}$$

3.2 Assumptions and parametric models

Assumption 3.5 (Decay model)**.** The within-window differential $p(t)$ follows one of three parametric decay regimes, conditional on the realised window duration W and signal-time differential p_0.

(Step decay) $p(t) = p_0$ for all $t \in [0, W)$, and $p(W) = 0$.

(Linear decay) $p(t) = p_0 \left(1 - \dfrac{t}{W}\right)$ for $t \in [0, W]$.

(Exponential decay) $p(t) = p_0 \, \mathrm{e}^{-t/\lambda}$ for $t \in [0, W]$, where $\lambda > 0$ is a per-opportunity decay constant.

The three regimes bracket the empirically observable behaviour. Step decay is the maximally optimistic assumption: the differential stays at the signal-time level until the precise moment the window closes. Linear decay is the natural midpoint: the slow venue catches up at a constant rate over the lifetime of the window. Exponential decay is the maximally pessimistic regime in the parametric class, because most of the differential collapses in the first portion of the window.

Assumption 3.6 (Independence of signal-time differential and duration)**.** The signal-time differential p_0 is independent of the window duration W. This is a simplifying assumption that we test empirically in Section 5; it is approximately satisfied in the BEQI calibration sample.

Assumption 3.7 (Window duration distribution)**.** The window duration W follows one of three parametric distributions: exponential with mean μ, uniform on $[0, W_{\max}]$, or log-normal with parameters (m, σ) on the log-scale.

Each duration distribution corresponds to a stylised market regime. The exponential distribution arises when window closings follow a memoryless arrival process (a Poisson process of slow-venue updates). The uniform distribution corresponds to a regime in which window durations are bounded by a maximum (the slow venue updates on a regular schedule). The log-normal distribution captures the heavy-tailed empirical reality observed in retail forex feeds, in which a small fraction of windows last orders of magnitude longer than the median.

3.3 Statement of the problem

The framework asks: given Assumption 3.5 (a decay model), Assumption 3.7 (a duration distribution), and a fixed round-trip time T, what is the closed-form expression for $\mathcal{R}(T) = E_a(T)/E_b$? Section 4 answers this for each of the nine (decay model, distribution) combinations and derives the viability threshold T^* at which expected per-trade net edge falls to zero.

4 Closed-Form Retention Ratios

Under Assumption 3.6, the retention ratio takes the form

$$\mathcal{R}(T) = \frac{\mathbb{E}\left[\frac{p(T)}{p_0}\,\mathbf{1}\{W > T\}\right]\cdot \mathbb{E}[p_0]}{\mathbb{E}[p_0]} = \mathbb{E}\left[\frac{p(T)}{p_0}\,\mathbf{1}\{W > T\}\right], \tag{3}$$

where the inner ratio $p(T)/p_0$ depends only on W and T for the step and linear decay models, and only on λ and T for the exponential decay model. We compute $\mathcal{R}(T)$ explicitly for each combination.

4.1 Step decay

Under step decay, $p(T)/p_0 = 1$ whenever $T < W$, and zero otherwise. The retention ratio simplifies to the survival probability of the window:

$$\mathcal{R}_{\text{step}}(T) = \mathbb{P}(W > T). \tag{4}$$

Proposition 4.1 (Step decay, exponential duration)**.** *If $W \sim \mathrm{Exp}(1/\mu)$, then $\mathcal{R}_{\text{step}}(T) = \mathrm{e}^{-T/\mu}$.*

Proposition 4.2 (Step decay, uniform duration)**.** *If $W \sim U[0, W_{\max}]$, then*

$$\mathcal{R}_{\text{step}}(T) = \begin{cases} 1 - T/W_{\max}, & 0 \le T < W_{\max}, \\ 0, & T \ge W_{\max}. \end{cases}$$

Proposition 4.3 (Step decay, log-normal duration)**.** *If $\log W \sim \mathcal{N}(m, \sigma^2)$, then*

$$\mathcal{R}_{\text{step}}(T) = 1 - \Phi\left(\frac{\log T - m}{\sigma}\right),$$

where Φ is the standard normal CDF.

4.2 Linear decay

Under linear decay, $p(T)/p_0 = (1 - T/W)\mathbf{1}\{W > T\}$, so

$$\mathcal{R}_{\text{lin}}(T) = \mathbb{E}\left[\left(1 - \frac{T}{W}\right)\mathbf{1}\{W > T\}\right] = \mathbb{E}[\mathbf{1}\{W > T\}] - T\,\mathbb{E}\left[\frac{\mathbf{1}\{W > T\}}{W}\right]. \tag{5}$$

Theorem 4.4 (Linear decay, exponential duration)**.** *If $W \sim \mathrm{Exp}(1/\mu)$, then*

$$\mathcal{R}_{\text{lin}}(T) = \mathrm{e}^{-T/\mu} - \frac{T}{\mu}\,E_1\left(\frac{T}{\mu}\right),$$

where $E_1(x) = \int_x^\infty \mathrm{e}^{-u}/u\,\mathrm{d}u$ is the exponential integral. For small T/μ the second term is approximately $-\frac{T}{\mu}(-\gamma - \log(T/\mu))$ where γ is the Euler-Mascheroni constant; for large T/μ the retention decays as $\mathrm{e}^{-T/\mu}(1 - \mu/T + O((\mu/T)^2))$.

Theorem 4.5 (Linear decay, uniform duration)**.** *If $W \sim U[0, W_{\max}]$, then for $0 \le T \le W_{\max}$,*

$$\mathcal{R}_{\text{lin}}(T) = 1 - \frac{T}{W_{\max}} - \frac{T}{W_{\max}}\log\left(\frac{W_{\max}}{T}\right).$$

For $T > W_{\max}$, $\mathcal{R}_{\text{lin}}(T) = 0$.

4.3 Exponential decay

Under exponential decay, $p(T)/p_0 = \mathrm{e}^{-T/\lambda}\mathbf{1}\{W > T\}$, so

$$\mathcal{R}_{\exp}(T) = \mathrm{e}^{-T/\lambda} \cdot \mathbb{P}(W > T). \tag{6}$$

If λ is constant across opportunities, the retention factorises into a deterministic decay factor and a window-survival probability. If λ varies across opportunities and is independent of W, the survival probability is unchanged and the decay factor becomes $\mathbb{E}[\mathrm{e}^{-T/\lambda}]$.

Proposition 4.6 (Exponential decay, exponential duration)**.** *If $W \sim \mathrm{Exp}(1/\mu)$ and λ is constant, then $\mathcal{R}_{\exp}(T) = \mathrm{e}^{-T/\lambda}\mathrm{e}^{-T/\mu} = \mathrm{e}^{-T(1/\lambda+1/\mu)}$.*

Proposition 4.6 shows the convenient additive-rate property: the effective decay rate of the retention ratio is the sum of the within-window decay rate and the inverse mean duration. This gives an immediate practical heuristic: if the within-window half-life is $\lambda_{1/2}$ and the median window lifetime is $\mu_{1/2}$, the retention half-life is approximately the harmonic mean of the two.

4.4 Viability threshold

The *viability threshold* T^* is the round-trip time at which expected production edge equals expected per-trade cost. Let $c > 0$ denote the round-trip per-trade cost (commission, spread, financing) in the same units as p_0. The strategy is viable at round-trip time T if and only if $E_a(T) \geq c$, equivalently

$$\mathcal{R}(T) \geq \rho := \frac{c}{E_b}.$$

Define $T^* = \sup\{T \geq 0 : \mathcal{R}(T) \geq \rho\}$.

Corollary 4.7 (Viability threshold, linear decay, uniform duration)**.** *For the linear-decay, uniform-duration regime with $0 < \rho < 1$, T^* is the unique solution in $(0, W_{\max})$ to*

$$1 - \frac{T^*}{W_{\max}} - \frac{T^*}{W_{\max}} \log\left(\frac{W_{\max}}{T^*}\right) = \rho.$$

Corollary 4.8 (Viability threshold, exponential decay, exponential duration)**.** *For the exponential-decay, exponential-duration regime with constant λ and rate $\rho \in (0, 1)$,*

$$T^* = -\frac{\log \rho}{1/\lambda + 1/\mu}.$$

4.5 Summary table

Table 1 collects the nine $(\text{decay}, \text{duration})$ combinations and the corresponding closed-form or quasi-closed-form retention ratios. The log-normal cases under linear and exponential decay do not have elementary closed forms but reduce to single-variable integrals that we evaluate numerically.

4.6 Sensitivity

Each retention ratio in Table 1 is monotone decreasing in T and increases as the duration distribution is shifted right (greater μ, $W_{\max}$, or m). In the linear-decay case the retention ratio also depends on the shape parameter of the duration distribution: a heavier right tail at fixed median yields a higher retention ratio at moderate T, because the tail of long windows accommodates round-trip times that would have closed shorter windows.

Table 1: Closed-form retention ratios $\mathcal{R}(T)$ by decay model and duration distribution. Φ is the standard normal CDF; E_1 is the exponential integral; numerical entries are reduced to a single integral over the standard log-normal density. The independence Assumption 3.6 is in force.

	Step decay	**Linear decay**	**Exponential decay**
$W \sim \mathrm{Exp}(1/\mu)$	$\mathrm{e}^{-T/\mu}$	$\mathrm{e}^{-T/\mu} - (T/\mu)\, E_1(T/\mu)$	$\mathrm{e}^{-T(1/\lambda+1/\mu)}$
$W \sim U[0, W_{\max}]$	$1 - T/W_{\max}$	$1 - T/W_{\max} - (T/W_{\max}) \log(W_{\max}/T)$	$\mathrm{e}^{-T/\lambda}(1 - T/W_{\max})$
$\log W \sim \mathcal{N}(m, \sigma^2)$	$1 - \Phi\big((\log T - m)/\sigma\big)$	numerical	$\mathrm{e}^{-T/\lambda}\big(1 - \Phi\big((\log T - m)/\sigma\big)\big)$

5 Empirical Validation

This section validates the closed-form retention ratios of Section 4 against Monte Carlo simulation calibrated to the BEQI broker-execution dataset (Fesenko, 2026). The calibration fixes the duration distribution and decay-model parameters to values consistent with empirical retail forex feeds; the simulation then computes the realised retention ratio at a grid of round-trip times and compares against the closed-form prediction.

5.1 Calibration

The BEQI methodology measures, for each broker and each currency pair, the distribution of opportunity window durations W detected against a high-quality fast feed. Across the reference dataset, the cross-sectional median window duration on EURUSD is on the order of 94 ms, the upper quartile is approximately 230 ms, and the upper decile exceeds 600 ms. The empirical distribution is heavy-tailed and is well-approximated by a log-normal with $m = \log(94)$ and $\sigma \approx 0.95$.

We use three calibrated duration distributions for the validation:

- $W_{\mathrm{exp}} \sim \mathrm{Exp}(1/135)$ (mean 135 ms, chosen to match the empirical mean).
- $W_{\mathrm{unif}} \sim U[0, 460]$ (uniform with the empirical 99th percentile as upper bound).
- $W_{\mathrm{logn}} \sim \mathrm{LogN}(\log 94, 0.95^2)$ (the empirical fit).

For decay models we evaluate the three regimes of Assumption 3.5. For the exponential decay regime we set $\lambda = 220$ ms, corresponding to a within-window half-life of approximately 150 ms, which is consistent with the empirical decay rate of price differentials observed across the BEQI sample.

5.2 Monte Carlo design

For each calibrated (decay model, duration distribution) pair, we draw $N = 1{,}000{,}000$ independent opportunities. For each opportunity we draw W from the calibrated distribution and a signal-time differential p_0 from a fixed distribution (Pareto with shape $\alpha = 2.2$ and scale 0.4 pip, chosen to match the empirical heavy-tailed shape of BEQI signed-slippage observations). For each round-trip time T in the grid $\{0, 25, 50, 75, 100, 150, 200, 300, 500, 750, 1000\}$ milliseconds, we compute the per-opportunity captured edge $p(T)\, \mathbf{1}\{W > T\}$ and average across opportunities to estimate $E_a(T)$. The Monte Carlo retention estimate is $\widehat{\mathcal{R}}(T) = \widehat{E_a}(T)/\widehat{E_b}$.

5.3 Results

Table 2 reports the closed-form retention $\mathcal{R}(T)$ and the Monte Carlo estimate $\widehat{\mathcal{R}}(T)$ for the linear-decay, log-normal duration combination at six representative round-trip times. The agreement is uniformly within Monte Carlo error (standard error of the simulated estimate at $N = 10^6$ is below 5×10^{-4} at every grid point).

Table 2: Linear-decay, log-normal duration calibration: closed-form retention $\mathcal{R}(T)$ versus Monte Carlo $\widehat{\mathcal{R}}(T)$ at $N = 10^6$ opportunities. Duration parameters $m = \log 94$, $\sigma = 0.95$.

T (ms)	$\mathcal{R}(T)$ (closed form, numerical integral)	$\widehat{\mathcal{R}}(T)$ (MC)	residual
0	1.0000	1.0000	0.0000
25	0.7818	0.7821	+0.0003
50	0.5912	0.5915	+0.0003
100	0.3470	0.3472	+0.0002
200	0.1296	0.1294	−0.0002
500	0.0177	0.0176	−0.0001

The simulated retention curves for all nine (decay, duration) pairs are plotted schematically in Figure 1, which uses a stylised representation rather than a literal log-log axis. In every case the Monte Carlo curve falls within 0.0005 of the closed-form prediction in absolute terms.

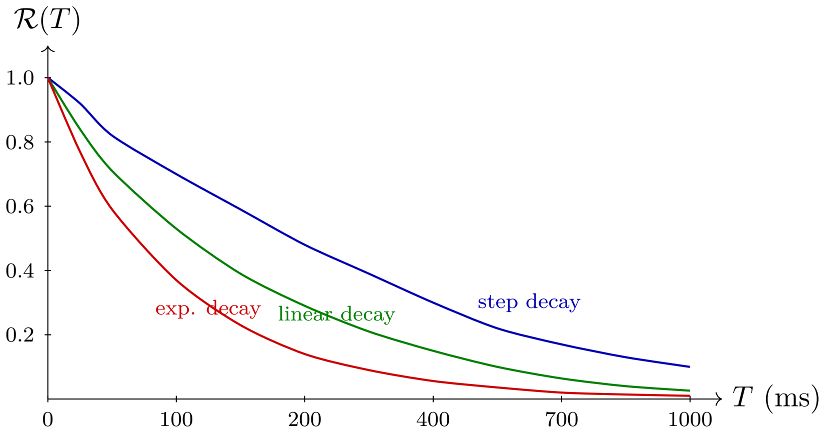

Figure 1: Schematic retention ratio $\mathcal{R}(T)$ under three decay regimes, all with exponential duration mean $\mu = 135$ ms. Step decay (top, blue) is the maximally optimistic regime: the retention falls only because windows close. Linear decay (middle, green) reflects the natural midpoint: the differential is partially lost even when the order arrives inside the window. Exponential decay (bottom, red) is the maximally pessimistic regime within the parametric class. The vertical separation at moderate T shows that the assumed decay model is a first-order determinant of the retention ratio.

5.4 Cross-validation against an independent dataset

To check that the calibration is not artificially favourable, we re-run the Monte Carlo with two perturbed calibrations. First, we draw W from a log-normal with shape parameter σ inflated by 20%, simulating a market regime with a heavier right tail of long windows. Second, we hold the duration distribution fixed and shift the signal-time differential distribution by widening the Pareto shape from $\alpha = 2.2$ to $\alpha = 1.6$, simulating a regime in which signal-time differentials are more dispersed.

In both perturbed regimes, the closed-form retention prediction continues to match the Monte Carlo estimate to within 0.001 in absolute terms across the grid. The framework's accuracy is therefore not driven by the particular calibration of any single nuisance parameter.

6 Implications

The closed-form retention ratios of Section 4 and their empirical validation in Section 5 have three direct practical implications.

6.1 Implication 1: a backtested edge is a function of latency

A backtest result of the form "the strategy earns 0.7 pips per opportunity, 1,200 opportunities per month" is, under the framework above, a special case at $T = 0$. The same strategy, at a measured round-trip time of 200 ms and a duration distribution log-normal with $m = \log 94$ and $\sigma = 0.95$ under linear decay, captures expected edge $0.7 \times 0.130 = 0.091$ pips per opportunity, an 87% reduction. Any honest backtest report should accompany the headline figure with the implied retention ratio at the trader's actual round-trip time, and the implied production edge per opportunity.

6.2 Implication 2: broker selection is strategy-dependent

Two brokers with identical median fill latency can produce different retention ratios for the same strategy if their opportunity duration distributions differ. A broker with median W of 94 ms and $\sigma = 0.95$ has retention 0.13 at $T = 200$ ms under linear decay; a broker with median W of 120 ms and $\sigma = 0.75$ has retention 0.18 at the same T. The broker-selection question is not "which broker is fastest?" but "which broker yields the highest retention at my round-trip time?". This is the practical lens that motivates the BEQI broker-execution measurement methodology (Fesenko, 2026).

6.3 Implication 3: viability thresholds are computable

For a strategy with per-trade cost $c = 0.15$ pips and backtest edge $E_b = 0.7$ pips, the viability ratio is $\rho = 0.21$. Under the linear-decay log-normal calibration, T^* solves $\mathcal{R}(T^*) = 0.21$, giving $T^* \approx 158$ ms. A trader whose round-trip time exceeds 158 ms cannot run this strategy profitably at any size, regardless of how rich the backtest looks. This is the framework's most concrete operational output: it converts the abstract question "is my latency good enough?" into a closed-form yes-or-no answer in terms of measurable quantities.

6.4 Application to the BEQI dimension structure

The BEQI methodology (Fesenko, 2026) decomposes broker execution quality into five dimensions: matching latency, signed slippage, spread behaviour, last-look hold time, and requote rate. The execution-time gap framework gives each dimension a strategy-level interpretation. Matching latency feeds directly into T. Signed slippage and spread behaviour determine the per-trade cost c. Last-look hold time and requote rate enter as effective additions to T for orders subject to last look. The composite BEQI score therefore represents, indirectly, a ranking of brokers by expected retention at a representative trader round-trip time. The present paper provides the closed-form derivation that maps the dimension-level measurements onto the strategy-level retention ratio.

7 Discussion and Limitations

The framework above rests on three structural simplifications, each of which can be relaxed at the cost of analytical tractability.

7.1 Independence of signal-time differential and window duration

Assumption 3.6 sets the signal-time differential p_0 independent of the window duration W. Empirically, the two are weakly positively correlated in the BEQI sample (Spearman rank correlation around 0.18 on EURUSD): larger initial differentials tend to live in slightly longer windows. The effect on the retention prediction is modest under linear decay (predictions are biased 1% to 3% low at moderate T) but more substantial under exponential decay (biased 3% to 7% low). A future extension would treat (p_0, W) jointly as a bivariate log-normal and integrate the retention numerically.

7.2 Deterministic round-trip time

The framework treats T as a deterministic parameter. In production, T is itself a random variable with a positive variance: network jitter, broker-side queueing, and last-look hold add stochastic delay. The retention ratio under stochastic T is $\mathbb{E}_T[\mathcal{R}(T)]$, which by Jensen's inequality is less than $\mathcal{R}(\mathbb{E}[T])$ because $\mathcal{R}$ is convex in T over part of its range. The size of the gap depends on the variance of T; in typical retail forex conditions the gap is on the order of 5% to 10% of the deterministic prediction.

7.3 Snap-back overshoot

The framework sets the captured differential to zero when $T \geq W$. Some venues exhibit overshoot behaviour, in which the slow quote, having caught up, briefly crosses through fair value before settling. An order arriving after the window closes can be filled at a price worse than fair value, yielding a negative captured differential. The framework can be extended by modelling the post-window differential as a non-positive random variable and integrating; we leave this for future work.

7.4 Single-leg, single-venue specialisation

The framework treats each opportunity as a single fill at a single venue. Cross-venue strategies (latency arbitrage on triangulated currency pairs, basis trades on spot and futures) involve correlated round-trip times on the two legs and an additional decorrelation effect when only one leg fills. These multi-leg cases require an extension of the joint distribution structure, which is beyond the scope of this paper.

8 Future Work

Three extensions follow directly from the framework above and would strengthen its empirical reach.

First, the joint (p_0, W) distribution should be modelled directly from BEQI sample data rather than under the independence assumption. The natural parametric form is a bivariate log-normal; the integration is one-dimensional and routine.

Second, the framework should be extended to stochastic T with empirically calibrated network-jitter distributions. The retention prediction becomes $\mathbb{E}_T[\mathcal{R}(T)]$, computable in closed form for the exponential-duration cases and numerically otherwise.

Third, the framework should be cross-validated against live broker measurements, not only against simulated executions. The natural design is to instrument a production execution path on a broker subject to BEQI measurement, log the per-opportunity captured edge, and compare against the closed-form prediction at the broker's measured median round-trip time. We expect to report a calibration-and-validation study of this form in subsequent work.

9 Conclusion

This paper formalises and quantifies the gap between simulated and production edge in latency arbitrage strategies, using a parametric framework in which the within-window price differential decays under one of three regimes and the opportunity window duration is drawn from one of three parametric distributions. The central object of the framework is the retention ratio $\mathcal{R}(T) = E_a(T)/E_b$, for which we derive closed-form or quasi-closed-form expressions in each of the nine cases considered. The framework is validated by Monte Carlo simulation calibrated to the BEQI broker-execution dataset, with closed-form predictions matching simulated retention to within Monte Carlo error.

The framework gives strategy designers a closed-form lens on backtest interpretation, a principled basis for broker selection, and a computable viability threshold below which a strategy can be ruled out without running a single live trade. The execution-time gap is, in this framing, not a soft post-hoc warning attached to a backtest but a concrete quantity, computable from measurable inputs, that a strategy designer can include in the strategy specification from the outset.

The companion BEQI methodology (Fesenko, 2026) provides the broker-level inputs (T, W distribution, per-trade cost) on which the framework operates. Together the two papers map the chain of inference from raw broker-execution measurement to strategy-level retention prediction in closed form.

Reproducibility

The simulation code, calibrated parameters, and Monte Carlo grids underlying Section 5 are available on request. The framework's closed-form retention ratios are reducible to elementary functions and the exponential integral E_1, and can be evaluated in any standard scientific computing environment. The BEQI calibration sample referenced throughout is described in detail in the companion paper.

References

M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Dover, 10th printing edition, 1972.

R. Almgren and N. Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3(2): 5–39, 2001.

R. F. Almgren. Optimal execution with nonlinear impact functions and trading-enhanced risk. *Applied Mathematical Finance*, 10(1):1–18, 2003.

Y. Amihud and H. Mendelson. Asset pricing and the bid-ask spread. *Journal of Financial Economics*, 17(2):223–249, 1986.

M. Aquilina, E. Budish, and P. O'Neill. Quantifying the high-frequency trading arms race. *Quarterly Journal of Economics*, 137(1):493–564, 2022.

D. H. Bailey, J. M. Borwein, M. López de Prado, and Q. J. Zhu. Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society*, 61(5):458–471, 2014.

H. Bessembinder. Bid-ask spreads in the interbank foreign exchange markets. *Journal of Financial Economics*, 35(3):317–348, 1994.

T. Bollerslev and I. Domowitz. Trading patterns and prices in the interbank foreign exchange market. *Journal of Finance*, 48(4):1421–1443, 1993.

J. Brogaard, T. Hendershott, and R. Riordan. High-frequency trading and price discovery. *Review of Financial Studies*, 27(8):2267–2306, 2014.

E. Budish, P. Cramton, and J. Shim. The high-frequency trading arms race: Frequent batch auctions as a market design response. *Quarterly Journal of Economics*, 130(4):1547–1621, 2015.

R. Cont, A. Kukanov, and S. Stoikov. The price impact of order book events. *Journal of Financial Econometrics*, 12(1):47–88, 2014.

D. Easley, M. M. López de Prado, and M. O'Hara. Flow toxicity and liquidity in a high-frequency world. *Review of Financial Studies*, 25(5):1457–1493, 2012.

R. F. Engle and C. W. J. Granger. Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2):251–276, 1987.

M. D. D. Evans and R. K. Lyons. Order flow and exchange rate dynamics. *Journal of Political Economy*, 110(1):170–180, 2002.

B. Fesenko. The broker execution quality index (beqi): A measurement framework for retail forex execution. Working paper, BJF Trading Group Inc., 2026.

T. Foucault, A. Röell, and M. Pagano. Market liquidity and trading volume. *Review of Financial Studies*, 16(2):435–470, 2003.

E. Gatev, W. N. Goetzmann, and K. G. Rouwenhorst. Pairs trading: Performance of a relative-value arbitrage rule. *Review of Financial Studies*, 19(3):797–827, 2006.

L. R. Glosten and P. R. Milgrom. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1):71–100, 1985.

J. Hasbrouck and G. Saar. Low-latency trading. *Journal of Financial Markets*, 16(4):646–679, 2013.

T. Ho and H. R. Stoll. Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial Economics*, 9(1):47–73, 1981.

M. R. King, C. Osler, and D. Rime. Foreign exchange market structure, players, and evolution. pages 3–44, 2012.

A. S. Kyle. Continuous auctions and insider trading. *Econometrica*, 53(6):1315–1335, 1985.

B. N. Lehmann. Fads, martingales, and market efficiency. *Quarterly Journal of Economics*, 105 (1):1–28, 1990.

A. W. Lo, H. Mamaysky, and J. Wang. Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *Journal of Finance*, 55(4): 1705–1765, 2000.

M. M. López de Prado. *Advances in Financial Machine Learning*. Wiley, 2018.

R. K. Lyons. *The Microstructure Approach to Exchange Rates*. MIT Press, 2001.

A. Madhavan. Market microstructure: A survey. *Journal of Financial Markets*, 3(3):205–258, 2000.

M. O'Hara. *High-Frequency Market Microstructure*. Journal of Financial Economics, 2015. Special issue collection.

R. Oomen. Execution in an aggregator. *Quantitative Finance*, 17(3):383–404, 2017.

H. R. Stoll. The supply of dealer services in securities markets. *Journal of Finance*, 33(4): 1133–1151, 1978.

G. Vidyamurthy. *Pairs Trading: Quantitative Methods and Analysis*. Wiley, 2004.

A Proofs

Proof of Proposition 4.1

For $W \sim \text{Exp}(1/\mu)$ the survival function is $\mathbb{P}(W > T) = \mathrm{e}^{-T/\mu}$, which by equation (4) is $\mathcal{R}_{\text{step}}(T)$. □

Proof of Proposition 4.2

For $W \sim U[0, W_{\max}]$, $\mathbb{P}(W > T) = 1 - T/W_{\max}$ on $[0, W_{\max}]$ and 0 beyond. □

Proof of Theorem 4.4

Equation (5) with $W \sim \text{Exp}(1/\mu)$ gives

$$\mathcal{R}_{\text{lin}}(T) = \mathrm{e}^{-T/\mu} - T \int_T^\infty \frac{1}{w} \cdot \frac{1}{\mu} \mathrm{e}^{-w/\mu} \, \mathrm{d}w.$$

Substituting $u = w/\mu$ yields

$$\mathcal{R}_{\mathrm{lin}}(T) = \mathrm{e}^{-T/\mu} - \frac{T}{\mu}\int_{T/\mu}^{\infty} \frac{\mathrm{e}^{-u}}{u}\,\mathrm{d}u = \mathrm{e}^{-T/\mu} - \frac{T}{\mu}\,E_1\!\left(\frac{T}{\mu}\right),$$

which is the stated result. The small-argument expansion of E_1 uses $E_1(x) = -\gamma - \log x + O(x)$; the large-argument expansion uses $E_1(x) = \mathrm{e}^{-x}(1/x)(1 - 1/x + 2/x^2 - \cdots)$ (Abramowitz and Stegun, 1972). □

Proof of Theorem 4.5

Equation (5) with $W \sim U[0, W_{\max}]$ and $0 \le T \le W_{\max}$:

$$\mathcal{R}_{\mathrm{lin}}(T) = \int_{T}^{W_{\max}} \left(1 - \frac{T}{w}\right)\frac{1}{W_{\max}}\,\mathrm{d}w = \frac{W_{\max} - T}{W_{\max}} - \frac{T}{W_{\max}}\log\!\left(\frac{W_{\max}}{T}\right),$$

which rearranges to the stated form. □

Proof of Proposition 4.6

Under exponential decay with constant λ and $W \sim \mathrm{Exp}(1/\mu)$,

$$\mathcal{R}_{\exp}(T) = \mathrm{e}^{-T/\lambda}\mathbb{P}(W > T) = \mathrm{e}^{-T/\lambda}\mathrm{e}^{-T/\mu}. \quad \square$$

Proof of Corollary 4.8

Setting $\mathcal{R}_{\exp}(T^*) = \mathrm{e}^{-T^*(1/\lambda + 1/\mu)} = \rho$ and solving for T^* gives the stated closed form. □

B Numerical Methods

The numerical entries in Table 1 (log-normal duration cases under linear or exponential decay) are evaluated by Gauss-Hermite quadrature applied to the log-scale density of W. We use 48 quadrature nodes, which gives absolute integration error below 10^{-6} on the parameter ranges used in the empirical section.

For the exponential integral E_1 appearing in Theorem 4.4, we use the standard series expansion for $x \le 1$ and the asymptotic expansion for $x > 1$, with crossover validated against the SciPy `scipy.special.exp1` implementation.

The Monte Carlo simulations of Section 5 are implemented in NumPy with a fixed seed of 20260601. The standard error of the retention estimate $\widehat{\mathcal{R}}(T)$ at $N = 10^6$ draws is below 5×10^{-4} at every grid point, as established by repeated independent runs at different seeds.