

On a Simple Relationship Between Order Imbalance, Skew and Width in Over-The-Counter Trading

Peter Cotton*

August 7, 2026

Abstract

We consider a market maker who can only obtain and dispose of inventory by responding to a sequence of sealed-bid enquiries, and whose customers arrive with imbalanced intent: sellers more often than buyers, or the reverse. Under the assumption that the best competing response is exponentially distributed around a commonly discerned fair price, we observe a symmetry in the steady state solution that compresses the imbalanced problem onto the perfectly balanced one. Order imbalance is absorbed, exactly, by a translation of the market maker’s skew, a widening of her quotes, and a multiplication of her effective cost of carry. The adjustment is simple even though the solution it adjusts is not, and it involves no free parameter beyond the observable market width. The exponential assumption is needed only locally, at the quotes actually made, and the width that enters is the locally observed one. Among the consequences: a market maker with zero inventory should still skew; skew responds to imbalance at first order whereas width responds only at second order; and the popular “constant width, linear skew” heuristic is recovered as the small-skew solution in the special case of balanced flow and quadratic holding cost.

Keywords: market making, order imbalance, skew, bid–ask spread, request for quote, inventory cost.

MSC codes: 91G15, 93E20.

1 Introduction

The facilitation of trade by intermediaries often takes the form of a sequence of sealed-bid auctions. A customer wishing to sell an item will typically obtain several bids from dealers and select the best price; a customer wishing to buy will perform a similar search and transact with whomever offers the lowest. The market maker takes on price risk and carrying cost, and must respond to each enquiry with that in mind. She is said to mark up her offer, or mark down her bid, relative to a fair market price.

*The work was completed around 2015, and the central symmetry was presented at NYU Tandon some time later; it was first written up in April 2022. The author thanks Pascal Tomecek, Morten Bai Andersen, Nick West and Erik Clossen for useful discussions. Joint work with Andrew Papanicolaou (*Trading Illiquid Goods: Market Making as a Sequence of Sealed-Bid Auctions, with Analytic Results*, working paper) extends the model considered here to clustered arrivals and stochastically varying imbalance; the present paper isolates the original observation, which is a symmetry. Comments welcome.

We have in mind fungible goods for which this description is apt (dealer markets in physical commodities, corporate and municipal bonds, blocks of securities, or for that matter rare baseball cards) and one empirical regularity of such markets that the standard theory does not deliver. Dealers shade their quotes with the direction of the flow. A dealer who expects the next call to be a seller quotes differently from one who expects a buyer, and does so even when her book is flat. Practitioners regard this as too obvious to remark upon. The optimal market making literature has meanwhile concentrated on the other lever: the benchmark policies descending from Avellaneda and Stoikov [4] hold width constant and skew linearly in inventory, with buy and sell arrivals assumed symmetric.

The purpose of this paper is to show that flow-shading is not a heuristic layered on top of inventory management but an exact feature of optimal behavior, and that it follows from a symmetry. In the steady state, the market making problem facing imbalanced flow compresses onto the perfectly balanced problem: the imbalance is absorbed, without approximation, by a translation of the skew, a widening of the quotes, and a multiplication of the effective cost of carry. The three adjustments are explicit constants involving nothing but the imbalance itself and the market width.

The solution being adjusted is not simple. The indifference cost that characterizes optimal behavior must in general be computed numerically, and its shape depends on the entire cost of carry function. The imbalance correction nevertheless rides on top of any such solution unchanged, so a single balanced solve covers every level of imbalance. Nor is the exponential assumption load-bearing in its global form: only the hazard at the quotes actually made enters, a point Remark 2 makes precise.

We note three consequences. A market maker with zero inventory should still skew, by half the market width times the log-odds of the arrival direction, which is approximately twice the market width times the imbalance when the imbalance is mild. The width response is second order in the imbalance while the skew response is first order, which we believe is why practitioners skew before they widen. And the constant width, linear skew heuristic reappears in the special case of balanced flow and quadratic holding cost, which locates what that benchmark omits: the flow term. The magnitudes involve no free parameter beyond the observable width and are falsifiable in principle; we make no empirical claims here.

1.1 The CWLS benchmark. Making constant width markets with skew linear in inventory.

To fix terminology, a dealer's *width* is the difference between where she is prepared to bid and where she is prepared to offer, and her *skew* is the difference between the average of the two and her estimate of fair value. The Constant Width Linear Skew heuristic maintains a fixed width and sets skew proportional to signed inventory, with a sign that leans the dealer towards a neutral position. Motivations for CWLS range from its *prima facie* reasonableness to formal justification as an approximate solution of a stochastic control problem [4, 5]. These justifications require some approximation, however, and they assume symmetric flow through-out. From a practical perspective CWLS is not a straw man, because it is simple, and because it sometimes *is* the solution. Section 4 makes the second point precise, and the symmetry of Section 3 supplies the benchmark's missing column: what to do about imbalance.

1.2 Literature

The dealer-inventory line begins with Garman [1], whose dealer faces genuinely asymmetric Poisson buy and sell rates, the oldest imbalanced-flow model, but sets a single static price pair, the content being ruin rather than policy. Amihud and Mendelson [2] are the closest classical antecedent in spirit: imbalanced arrivals move a preferred inventory position and monotone quotes, though the results are structural rather than closed form. Ho and Stoll [3] quote from value function differences, and the *slope* half of the slope-and-convexity characterization of Section 2.2 is implicit there and in everything since; the *convexity* half states that the discretionary component of a dealer’s width reveals the second difference of her inventory cost exactly, so that her inventory cost can be read off her quoting. We have not found the latter stated elsewhere.

Avellaneda and Stoikov [4] anchor the modern literature. Their constant width, linear skew policy is reached by dropping terms, and the dropped terms are precisely the ones needed to establish any rule regarding width. Constant width is an artifact of the approximation, not a conclusion of the model, and the line consequently has nothing to say about how width should respond to inventory or to flow.

The exact treatment of Guéant, Lehalle and Fernandez-Tapia [5] takes arrival intensities to be symmetric; the one directional asymmetry handled in closed form in that line is drift in the mid-price, which shifts quotes but neither widens them nor taxes carry. Guéant [6, 7] and Bergault et al. [13, 14] accommodate asymmetric intensities in generality but handle them numerically. We have found no equivalence statement, no log-odds shift, and no cost multiplier in that literature.

Cartea, Jaimungal and coauthors [8, 9, 10] treat order-flow imbalance as a predictive signal, a different object from the stationary structural imbalance considered here. Bergault and Guéant [18] observe numerically that flow-aware market makers skew their quotes even in the absence of inventory; Corollary 1 is the closed form of that observation.

The request-for-quote literature [11, 12, 15] arrived some years after this work was completed. Fermanian, Guéant and Pu [11] provide econometric support for the win-curve assumption we make, Hendershott and Madhavan [16] document the sealed-bid mechanism itself, and Butz and Oomen [17] document electronic FX dealers skewing on flow at flat inventory, the practice that Corollary 1 derives rather than assumes.

2 Model for an imbalanced dealer market

Trade opportunities arrive as a Poisson process with mean inter-arrival time τ . Each arrival is a customer, met simultaneously by quotes from several market makers, and the trade is done at the winning quote. At each arrival there is a chance $q \in (0, 1)$ that the customer is a seller (a *buy opportunity* for our market maker) and chance $1 - q$ that the customer is a buyer. All trades are of size s . There is a fair market price commonly discerned by all dealers; in the event of a buy opportunity our market maker submits a markdown $m^\downarrow > 0$ below it, and in the event of a sell opportunity a markup $m^\uparrow > 0$ above it. She wins if her response beats the best competing response, whose displacement beyond fair value has survival function F with hazard rate $h(z)$; we will shortly take the hazard constant, an assumption that Remark 2 weakens to local constancy, so that the best competitor is exponentially distributed with mean

$$w = 1/h,$$

which we call the *market width*. Every won trade concedes a fixed amount ϵ to adverse selection: when we mark up an item by m we assume we make only $m - \epsilon$. Holding inventory x (a real number; negative values accommodate short positions) costs $c(x)$ per unit time with $c(0) = 0$, and the game is played forever: we study the ergodic, steady state policy.

The model deliberately leaves two things out, and the companion paper extends it in exactly these directions: arrivals may cluster, and the imbalance q may itself fluctuate. The symmetry is exact in the stationary core, which is what we isolate here.

2.1 Indifference liquidation markup

Suppose there is a liquidation cost $\nu(x) \geq 0$ that the market maker can pay at any time to unload her inventory, and suppose she is indifferent between paying it and carrying on. The function ν can be thought of as the dealer-to-dealer “haircut” at which she would clear her book. It characterizes her policy, and two discrete derivatives of it will do all the work:

$$S(x) = \frac{\nu(x+s) - \nu(x-s)}{2s}, \quad C(x) = \frac{\nu(x+s) - 2\nu(x) + \nu(x-s)}{2s},$$

the slope and convexity of the indifference cost, per unit.

2.2 Optimal markups

Define break-even markups, which we call *strikes*: the response levels at which winning the enquiry leaves the market maker indifferent,

$$K^\downarrow(x; s) = \epsilon + \frac{\nu(x+s) - \nu(x)}{s}, \quad K^\uparrow(x; s) = \epsilon + \frac{\nu(x-s) - \nu(x)}{s},$$

for buying and selling respectively. (Note that $K^\uparrow$ may well be negative for positive inventory: a loaded dealer will pay to get out.) The best response solves $\sup_m (m - K)(1 - F(m))$, and evidently the optimal choice of markup lies precisely where the increase in profit exactly offsets the possibility of losing the trade. For say we increased the markup by Δm ; then

$$\overbrace{(m^\uparrow - K^\uparrow(x; s))}^{\text{existing benefit}} \quad \overbrace{h(m^\uparrow) \Delta m}^{\text{chance of losing it}} \quad = \quad \overbrace{\Delta m}^{\text{increase in profit}}.$$

The first-order condition is therefore independent of the distributional assumption on the inside market:

$$m^\uparrow(x; s) = \max\left(\frac{1}{h(m^\uparrow(x; s))} + K^\uparrow(x; s), 0\right), \quad (1)$$

and likewise for $m^\downarrow$ with $K^\downarrow$. Under the constant hazard assumption this reads straightforwardly:

$$\text{“markup} = \text{market width} + \text{adverse selection} + \text{marginal inventory cost”}$$

though of course the coupled quantities ν and the markups remain to be found. The first order condition itself is Lerner logic with survival-function demand and is classical; the market making content lies in the indifference-adjusted strike. Away from the region where the $\max(\cdot, 0)$ binds, the two responses arrange themselves around a displaced midpoint:

$$\text{half-width}(x) = \Delta + C(x), \quad \text{mid}(x) = \text{fair value} - S(x), \quad \Delta = w + \epsilon. \quad (2)$$

Slope of the inventory cost is the skew; convexity of the inventory cost is the discretionary width. These identities are exact, and the second runs both ways: a dealer's quoting reveals the first two derivatives of her inventory cost, whether or not she has ever written it down.

2.3 The consistency equation

The market maker can remain indifferent in two ways: liquidate now and re-enter, or carry to the next opportunity. Comparing the two yields a functional equation for ν . With exponential markups, $\sup_m(m - K)e^{-hm} = \frac{1}{h}e^{-1-hK}$, and the steady state condition takes the form: for all x ,

$$\frac{\tau c(x)}{s} = \frac{e^{-1-h\epsilon}}{h} \left[e^{-hC(x)} \left(q e^{-hS(x)} + (1-q) e^{hS(x)} \right) - \left(\text{same expression at } x = 0 \right) \right]. \quad (3)$$

In words: the direct cost of carrying x must be exactly offset by the differential option value of the next trading opportunity at x versus at zero inventory. This is why common-sense holding cost arguments, mean holding time multiplied by carrying cost and so forth, systematically overstate the cost of inventory and encourage overly defensive quoting. They ignore that a larger inventory makes the next enquiry *more* valuable, since the potential to get out helps more.

3 The symmetry

The paper rests on an elementary identity: for any S ,

$$q e^{-hS} + (1-q) e^{hS} = 2\sqrt{q(1-q)} \cosh(h(S - \delta)), \quad \delta = \frac{1}{2h} \log \frac{q}{1-q}. \quad (4)$$

An imbalanced mixture of exponentials is a balanced one in disguise: translated by δ , and rescaled by $2\sqrt{q(1-q)} \leq 1$. Writing $\gamma = \frac{1}{h} \log \frac{1}{2\sqrt{q(1-q)}} \geq 0$ for the rescaling expressed as a width, we have immediately:

Theorem 1 (The imbalanced problem compresses onto the balanced one). *The steady state market making problem with arrival imbalance q and carrying cost $c(\cdot)$ has the same solution as the perfectly balanced problem ($q = \frac{1}{2}$) with carrying cost $M(q)c(\cdot)$, where*

$$M(q) = e^{h\gamma} = \frac{1}{2\sqrt{q(1-q)}} \geq 1, \quad (5)$$

under the correspondence: skew translated by δ , and non-discretionary width Δ widened to $\Delta + \gamma$.

Proof. Substitute (4) into (3) and write $S_\delta(x) = S(x) - \delta$:

$$\frac{\tau c(x)}{s} = \frac{e^{-1-h\epsilon-h\gamma}}{h} \left[e^{-hC(x)} \cosh(hS_\delta(x)) - e^{-hC(0)} \cosh(hS_\delta(0)) \right],$$

which is the balanced consistency equation in the pair (C, S_δ) , with the cost side inflated by $e^{h\gamma} = M(q)$. $\square$

Remark 1 (The adjustment is simple; the solution is not). *The theorem neither requires nor provides a closed form. In general ν must be computed numerically, and everything about its shape depends on c . What the symmetry provides is the correction: whatever the balanced solution is, the imbalanced one is that same object translated by δ , widened by γ , and evaluated at carrying cost $M(q)c$. One balanced solve serves every imbalance.*

Remark 2 (Exponentiality is needed only locally). *The constant hazard need not hold globally. The distribution of the best competing quote enters through one object only: the value of the next enquiry as a function of its strike, $G(K) = \sup_m (m - K)(1 - F(m))$. For any smooth F the envelope theorem gives $G'(K) = -(1 - F(m^*(K)))$, and dividing by G ,*

$$-\frac{d}{dK} \log G(K) = \frac{1}{m^*(K) - K} = h(m^*(K)),$$

exactly. The proof uses only the exponential form of G , so the requirement is that the hazard be constant at the quotes actually made, across the bounded range of strikes generated by the inventories the dealer visits. If the hazard varies slowly there, every statement above holds with w the reciprocal hazard at the operating point, and the error is of the order of the hazard's relative variation across that range. We checked this numerically in `verify_local_exponentiality.py`, which accompanies the paper: the error in the zero-inventory skew is about a quarter of the hazard's relative variation across the visited strikes, and shrinks linearly with it.

Remark 3 (The algebra is old). *We arrived at the substitution in the proof independently and noticed its provenance only afterwards. It is a birth-death symmetrization, and the geometric mean of the rates it produces is familiar from the spectral theory of birth and death processes [19, 20] and from the transient analysis of the simple queue [21]. It could equally be applied to the linear systems of the limit order book literature with asymmetric intensities, though to our knowledge nobody has done so. What we claim is the economics: imbalance translates the skew, widens the quotes, and taxes the carry, with all three effects tied to a single observable.*

Corollary 1 (A flat book still skews). *By symmetry the balanced solution has $S_\delta(0) = 0$, hence*

$$S(0) = \delta = \frac{w}{2} \log \frac{q}{1-q} \approx 2w \left(q - \frac{1}{2} \right) \quad \text{for mild imbalance.} \quad (6)$$

When customers are predominantly sellers the market maker marks her midpoint down by δ before holding any inventory at all. Skew is a flow phenomenon first and an inventory phenomenon second.

Corollary 2 (Skew is first order in imbalance; width is second order). *Expanding about $q = \frac{1}{2}$,*

$$\delta \approx 2w \left(q - \frac{1}{2} \right), \quad \gamma \approx 2w \left(q - \frac{1}{2} \right)^2, \quad M(q) \approx 1 + 2 \left(q - \frac{1}{2} \right)^2.$$

A mild imbalance moves the midpoint linearly and the width only quadratically. This accords with the practitioner's instinct to skew into flow and widen only when the market becomes seriously one-sided. Both responses involve no parameter beyond the observable width w ; the skew flips sign between net-selling and net-buying regimes while the widening, being even in $q - \frac{1}{2}$, does not.

Corollary 3 (Imbalance acts as a carrying cost). *By (5), one-sided flow is indistinguishable from a multiplied cost of carry, without bound as $q \rightarrow 0$ or 1 . A market maker in a good that costs nothing to hold is nonetheless taxed on inventory whenever the flow is one-sided, and taxed at a rate she can compute from the tape.*

4 Constant width solutions, and where CWLS sits

Suppose the convexity is constant, $C(x) \equiv C_0$: a constant width market. The consistency equation then inverts to give the skew directly,

$$S_\delta(x) = \frac{1}{h} \cosh^{-1} \left(e^{hC_0} \Omega c(x) + \cosh(hS_\delta(0)) \right), \quad \Omega = \frac{\tau h}{s} e^{1+h\epsilon+h\gamma}, \quad (7)$$

an easily implementable constant width model with, in general, *nonlinear* skew. For small argument the right side is approximately proportional to $\sqrt{c(x)}$, so a linear component in c (a physical holding fee) puts a square-root kink in the skew at the origin, and funding costs or equity hurdles bend it elsewhere.

Where, then, is CWLS exactly optimal? Requiring $C(x) \equiv C_0$ together with consistency of (3) and non-negative liquidation costs forces

$$\nu(x) = \frac{C_0}{s} x^2 + \delta x + C_2, \quad (8)$$

with the imbalance appearing as the *linear term of the indifference cost*, and feeding (8) back through (3) pins the cost of carry to the shape $c(x) \propto \cosh(2hC_0x/s) - 1$. The quadratic term prices inventory risk; the linear term prices the flow; and the skew $S(x) = 2\frac{C_0}{s}x + \delta$ is linear with an intercept.

CWLS is therefore exactly correct in one corner: balanced flow and a cosh-shaped cost of carry, a shape that is quadratic up to a relative error of $(hS_\delta(x))^2/12$. For a genuinely quadratic cost, CWLS is accurate while the skew is small relative to the width, with error quadratic in that ratio. The symmetry supplies the benchmark's missing column: under imbalance, keep the linear response to inventory, add the intercept δ , and widen by γ . The moral for anyone outside the corner is procedural: solve for $\nu(x)$ rather than pretend $\nu \propto x^2$ works, since the quadratic ansatz is internally inconsistent for any other cost of carry.

5 Uses

The equivalence was stated as a fact about optimal policy, but its practical content is broader, and we note five uses.

Fill ratios. At the optimum, profitability factors cleanly. By (1) the optimal response sits the distance w beyond its strike, so the expected net gain on a won trade is sw at every level of inventory, and all variation in expected profit per enquiry is variation in the fill ratio. Under the constant hazard the fill ratio on a quote at markup m is $e^{-m/w}$, log-linear in the markup with slope $-1/w$, which is how w is estimated from a quote archive; at the optimal response to strike K it equals $e^{-1-K/w}$. In particular a dealer facing zero adverse selection and zero marginal inventory cost should, if she quotes optimally, win $e^{-1} \approx 37\%$ of her enquiries, and persistent departures from that ratio measure her effective strike.

Testing the symmetry. Request-for-quote archives record direction, quotes, and fills. Direction counts estimate q , the win curve estimates w , with Remark 2 indicating that the local estimate is the appropriate one, and the theorem then predicts the zero-inventory skew $\delta = \frac{w}{2} \log \frac{q}{1-q}$ and the widening γ with no free parameter. Cross-sectionally, instruments with more one-sided flow should show mids shifted at first order in the imbalance and quotes widened at second order, per Corollary 2.

Reading a dealer's book. The identities (2) run both ways: skew and discretionary width are the first two differences of the indifference cost, so a counterparty's quoting reveals her inventory cost function, and a shift in her zero-inventory skew reveals her view of the flow.

Increasing the statistical power of learned policies. Reinforcement learning approaches to market making [12] generalize the present model and need not assume an exponential win curve. The symmetry helps them anyway, reinterpreted as an approximate heuristic: experience gathered at one imbalance transfers to every other through the correction $(\delta, \gamma, M(q))$, so a learner can pool episodes across flow regimes instead of learning each regime separately. An invariance of this sort reduces the effective dimension of the policy space, and the sample sizes required shrink accordingly.

Assessing decisions after the fact. Illiquidity implies short time series, and short time series make back-testing unreliable; the model instead prices individual decisions. Comparing a trader's response m against the model's m^* on a single enquiry:

	model trades	model misses
trader trades	$s(m - m^*)$	$s(m - m^*) + s w$
trader misses	$-s w$	0

Every cell is computable from observables at the moment of the enquiry, and the running total is a score of quoting skill that separates flow-reading (the δ term) from inventory management (the S_δ term). Profit and loss, by contrast, mixes both with luck. The constant margin under exponential demand is itself a classical property; the diagnostic built on it is not.

References

- [1] Garman, M. B. (1976). Market microstructure. *Journal of Financial Economics* 3(3), 257–275.
- [2] Amihud, Y., Mendelson, H. (1980). Dealership market: market-making with inventory. *Journal of Financial Economics* 8(1), 31–53.
- [3] Ho, T., Stoll, H. R. (1981). Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial Economics* 9(1), 47–73.
- [4] Avellaneda, M., Stoikov, S. (2008). High-frequency trading in a limit order book. *Quantitative Finance* 8(3), 217–224.
- [5] Guéant, O., Lehalle, C.-A., Fernandez-Tapia, J. (2012). Dealing with the inventory risk: a solution to the market making problem. *Mathematics and Financial Economics* 7(4), 477–507.
- [6] Guéant, O. (2016). *The Financial Mathematics of Market Liquidity: From Optimal Execution to Market Making*. Chapman & Hall/CRC.
- [7] Guéant, O. (2017). Optimal market making. *Applied Mathematical Finance* 24(2), 112–154.

- [8] Cartea, Á., Jaimungal, S., Ricci, J. (2014). Buy low, sell high: a high frequency trading perspective. *SIAM Journal on Financial Mathematics* 5(1), 415–444.
- [9] Cartea, Á., Jaimungal, S. (2016). Incorporating order-flow into optimal execution. *Mathematics and Financial Economics* 10(3), 339–364.
- [10] Cartea, Á., Jaimungal, S., Penalva, J. (2015). *Algorithmic and High-Frequency Trading*. Cambridge University Press.
- [11] Fermanian, J.-D., Guéant, O., Pu, J. (2017). The behavior of dealers and clients on the European corporate bond market: the case of multi-dealer-to-client platforms. *Market Microstructure and Liquidity* 2(3–4), 1750004.
- [12] Guéant, O., Manziuk, I. (2019). Deep reinforcement learning for market making in corporate bonds: beating the curse of dimensionality. *Applied Mathematical Finance* 26(5), 387–452.
- [13] Bergault, P., Guéant, O. (2021). Size matters for OTC market makers: general results and dimensionality reduction techniques. *Mathematical Finance* 31(1), 279–322.
- [14] Bergault, P., Evangelista, D., Guéant, O., Vieira, D. (2021). Closed-form approximations in multi-asset market making. *Applied Mathematical Finance* 28(2), 101–142.
- [15] Barzykin, A., Bergault, P., Guéant, O. (2023). Algorithmic market making in dealer markets with hedging and market impact. *Mathematical Finance* 33(1), 41–79.
- [16] Hendershott, T., Madhavan, A. (2015). Click or call? Auction versus search in the over-the-counter market. *Journal of Finance* 70(1), 419–447.
- [17] Butz, M., Oomen, R. (2019). Internalisation by electronic FX spot dealers. *Quantitative Finance* 19(1), 35–56.
- [18] Bergault, P., Guéant, O. (2023). Liquidity dynamics in RFQ markets and impact on pricing. arXiv:2309.04216.
- [19] Ledermann, W., Reuter, G. E. H. (1954). Spectral theory for the differential equations of simple birth and death processes. *Philosophical Transactions of the Royal Society A* 246(914), 321–369.
- [20] Karlin, S., McGregor, J. (1957). The differential equations of birth-and-death processes, and the Stieltjes moment problem. *Transactions of the American Mathematical Society* 85(2), 489–546.
- [21] Bailey, N. T. J. (1954). A continuous time treatment of a simple queue using generating functions. *Journal of the Royal Statistical Society, Series B* 16(2), 288–291.