HALO: Regime-Aware Market Making in Cryptocurrency Markets

Colin Whetstone
Quant Finance Collective
Texas State University
2026

Abstract

This paper presents HALO, a market-making system developed by the Quant Finance Collective (QFC) at Texas State University. HALO implements the Avellaneda-Stoikov (2008) framework as its inventory management and quote placement engine, extended with three layers: a real-time volatility regime classifier using lag-1 autocorrelation and EWMA volatility, a competitor spread floor model, and an order flow imbalance (OFI) filter for adverse selection detection. We document the full development process, including the identification and correction of a bug in our initial implementation of the Hurst exponent and an empirical investigation of adverse selection dynamics in cryptocurrency spot markets. We discuss the distinction between Avellaneda-Stoikov as an inventory management tool versus a complete market-making strategy, the role of informed versus uninformed flow in determining market-maker profitability, and the practical challenges of deploying a systematic market-making system with student-level infrastructure. Backtesting on 11.3 days of SOL/USD tick data across four poll cadences reveals that spread capture is consistently positive while inventory mark-to-market losses from adverse selection dominate performance at faster cadences, with a Sharpe ratio of +2.4 achieved at the 30-row cadence. This paper represents QFC's first systematic research report on market making and serves as the foundation for ongoing live paper trading on SOL/USD at the Kraken exchange.

Contents

1. Introduction

1.1 Market Making and the Inventory Problem

Market making is the practice of continuously quoting bid and ask prices to provide liquidity to other market participants. A market maker profits from the bid-ask spread buying at the bid and selling at the ask but bears *inventory risk*: when price moves directionally, accumulated positions lose value faster than spread income can recover. The tension in market making is between fill rate, which requires tight competitive quotes, and adverse selection protection, which requires wide quotes that are less likely to be hit by informed traders.

The Avellaneda-Stoikov model [1] provides the theoretical treatment of the inventory problem in market making. By modeling the mid-price as a Brownian motion and order arrivals as a Poisson process with exponentially decaying intensity, the model derives optimal bid and ask quotes as functions of inventory, volatility, and a risk aversion parameter γ. The key insight is that optimal quotes are not symmetric around mid price when inventory is non-zero: a dealer with excess long inventory should lower their ask to encourage selling, and vice versa.

It is important to note what the Avellaneda-Stoikov model is and is not. It is an *inventory management* model given that you are quoting, it tells you *where* to quote to manage inventory risk optimally. It does not address whether to quote at all, how to detect when incoming flow is informed, how to model competitor behavior, or how to manage the latency and queue position dynamics that determine fill probability in live markets. In HALO, we use A-S as the inventory management layer and add additional components to address some of these gaps.

1.2 Market Making as a Learning Exercise

Market making is a technology and infrastructure problem as much as a strategy problem. Professional market makers at firms like Citadel Securities, Wintermute, and Jump operate with sub-millisecond latency, direct exchange connections, and sophisticated real-time risk management systems. A student club running Python on a laptop with a 30-second poll cadence occupies a fundamentally different competitive niche.

This creates a challenge and an opportunity to learn. The challenge is that competing for fills in a liquid market without speed advantages means HALO will not capture every opportunity a professional system would. The opportunity is that SOL/USD, as a mid-tier liquid asset with price volatility and real uninformed flow, provides a meaningful environment

for studying the core problems of market making: inventory risk, adverse selection, regime-adaptive quoting, and spread capture versus mark-to-market tradeoffs. HALO's deployment on SOL/USD is explicitly designed as a research and learning exercise generating real fill data, real inventory dynamics, and real adverse selection exposure in a market that reflects the conditions professional market makers actually face.

1.3 Contributions

This paper makes the following contributions:

- We implement the Avellaneda-Stoikov framework with regime-adaptive γ using lag-1 autocorrelation-based regime detection, correcting an earlier implementation that used the Hurst R/S estimator.
- We add a volume-imbalance order flow imbalance (OFI) filter as a practical adverse selection detection layer, reducing fill acceptance when order book pressure suggests informed directional flow.
- We conduct a multi-cadence backtest on 11.3 days of SOL/USD tick data, demonstrating that poll cadence itself is a meaningful strategy parameter and that slower cadences produce better risk-adjusted returns by allowing inventory mean-reversion between fills.
- We provide a complete open-source implementation of the system including data collection, regime classification, quote computation, competitor floor modeling, backtesting, and paper trading simulation.

2. Background and Related Work

2.1 The Avellaneda-Stoikov Model

The Avellaneda-Stoikov model derives optimal quotes for a dealer with CARA (constant absolute risk aversion) utility over terminal wealth. The mid-price follows a Brownian motion $dS = \sigma\, dW$, and order arrivals follow a Poisson process with intensity

$$\lambda(\delta) = A \cdot e^{-\kappa\delta},$$

where δ is the distance of the quote from the reservation price. The two key outputs are the *reservation price* and the *optimal spread*:

$$r = s - q\,\gamma\,\sigma^2\,(T-t), \tag{1}$$

$$\delta^* = \gamma\,\sigma^2\,(T-t) + \frac{2}{\gamma}\ln\!\Big(1+\frac{\gamma}{\kappa}\Big), \tag{2}$$

where s is the current mid price, q is current inventory, γ is risk aversion, σ is volatility, $T-t$ is the remaining time horizon, and κ is the order arrival intensity parameter. The reservation price r shifts quotes away from mid proportionally to inventory, while δ^* determines the total width of the market being made.

In HALO, we normalize $T-t = 1.0$, treating each trading session as a unit horizon, and treat κ as a fixed calibration parameter rather than estimating it from order arrival data, as Poisson arrival estimation is unreliable at 30-second poll cadence.

2.2 Regime Detection in Market Making

The original HALO design used the Hurst exponent estimated usingng the rescaled range (R/S) method to classify volatility regimes. The Hurst exponent H measures long-range dependence: $H < 0.5$ indicates mean reversion, $H = 0.5$ a random walk, and $H > 0.5$ trending behavior.

During implementation, we identified a bug in the R/S estimator that produced unreliable regime labels, particularly in short windows. The estimator consistently produced high-regime labels during the regime warmup period regardless of actual market conditions, which we traced to the R/S method's well-known positive bias with small samples.

We replaced the Hurst R/S estimator with a lag-1 autocorrelation measure computed over a rolling window of recent mid-price returns. Lag-1 autocorrelation is more computationally efficient, interpretable, and reliable with sample sizes available at 10–30 second cadence. Positive autocorrelation indicates trending conditions; negative autocorrelation indicates mean reversion.

2.3 Adverse Selection and Order Flow Imbalance

Adverse selection is the central profitability challenge for market makers. When a market maker's quote is hit by an informed trader one who knows the short-term price direction, the fill is followed by an adverse price move that erodes the spread income captured.

Order flow imbalance (OFI) measures the asymmetry between buy and sell pressure in the

order book. Cont, Kukanov, and Stoikov [4] demonstrate that OFI computed from limit order book events explains a significant fraction of short-term price changes, making it a natural signal for adverse selection detection. The industry standard is VPIN (Volume-Synchronized Probability of Informed Trading), introduced by Easley, Lopez de Prado, and O'Hara [3]. HALO implements a simplified volume-imbalance OFI measure as a first-generation approximation, with VPIN identified as a natural extension for future work.

3. System Architecture

3.1 Overview

HALO consists of four integrated components operating as a sequential pipeline at each market data poll:

1. **Data collection:** order book snapshot via Kraken API using CCXT, capturing best bid, ask, mid price, bid volume, ask volume, and observed spread.
2. **Regime classification:** lag-1 autocorrelation and EWMA volatility over rolling windows, producing a three-state regime label.
3. **Quote computation:** Avellaneda-Stoikov reservation price and optimal spread using regime-adjusted γ, clamped to min/max spread bounds and raised to competitor floor.
4. **Fill simulation:** interval-crossing fill model with volume-imbalance OFI filter.

3.2 EWMA Volatility Estimation

Realized volatility is estimated using an Exponentially Weighted Moving Average (EWMA) over a span of 20 periods, providing a responsive estimate that weights recent price movements more heavily than distant ones. The EWMA volatility feeds directly into the A-S spread formula as the σ parameter.

3.3 Regime Classification

The regime classifier computes lag-1 autocorrelation of mid-price returns over a 500-tick rolling window, requiring a warmup period of approximately 83 minutes at 10-second cadence before regime signals are used. Prior to warmup completion, the system defaults to normal regime.

Table 1: Regime classification rules and effects on γ.

Regime	Condition	γ multiplier	Effect
High	$\rho_1 > 0.1$ or vol_ratio > 1.5	$2.5\times$ base	Wider spread, strong inventory skew
Normal	All other cases	$1.0\times$ base	Standard A-S behavior
Low	$\rho_1 < -0.1$ and vol_ratio < 0.7	$0.6\times$ base	Tighter spread, reduced skew

3.4 Quote Computation

Given the regime-adjusted γ, HALO computes the A-S reservation price and optimal spread via equations (1) and (2). The quoted spread is clamped between MIN_SPREAD_BPS = 20 bps and MAX_SPREAD_BPS = 50 bps of mid price, then raised to the competitor floor estimate if the floor exceeds the minimum. The competitor floor is the rolling minimum observed market spread over the last 100 snapshots.

3.5 Order Flow Imbalance Filter

At each potential fill, HALO computes a volume imbalance OFI score from the current order book snapshot:

$$\text{OFI} = \frac{V_{\text{bid}} - V_{\text{ask}}}{V_{\text{bid}} + V_{\text{ask}}}.$$

If OFI < -0.2 (ask-heavy book, bearish pressure) and the fill would be a buy, the fill is rejected with 50% probability. If OFI $> +0.2$ (bid-heavy book, bullish pressure) and the fill would be a sell, the fill is similarly rejected with 50% probability.

This implementation is a simplified proxy for the full VPIN measure used in professional market making. The volume-imbalance OFI captures the directional intent of the VPIN framework with the data available at polling cadence, and is explicitly designed as a first-generation implementation to be replaced by a proper VPIN measure as the system matures.

3.6 Fill Detection Model

Quotes are posted at tick i and checked against all mid-price observations in the interval $(i+1,\, i+N]$ at the next poll, where N is the poll cadence. A fill is triggered on the first crossing of the quoted bid or ask within each interval. This interval-crossing model correctly captures fills that occur between poll ticks rather than requiring the market to be at the quote level exactly at the moment of polling.

4. Venue and Asset Selection

4.1 Selection Criteria

SOL/USD on Kraken was selected as the target instrument based on three criteria: (1) sufficient daily volume to ensure a meaningful ratio of uninformed to informed order flow; (2) real price volatility, so that the regime detection and inventory management components are actually exercised; (3) spreads in the 5–15 bps range. Which is wide enough to capture meaningfully, and tight enough to reflect realistic competitive conditions.

4.2 Capital Parameters

Capital parameters were calibrated to reflect SOL's price level of approximately \$85–90 at the time of deployment: starting capital of \$50,000, fill quantity of 10 SOL per fill (\$850–900 notional), and an inventory cap of 50 SOL (\$4,250–4,500 maximum directional exposure).

5. Backtest Results

5.1 Data and Setup

The backtest was conducted on 11.3 days of SOL/USD tick data collected from Kraken via CCXT, comprising approximately 46,000 observations at roughly 21-second intervals. The backtest was run at four poll cadences (1, 5, 10, and 30 rows) corresponding to approximately 21 seconds, 1.75 minutes, 3.5 minutes, and 10.5 minutes between quote updates.

5.2 Multi-Cadence Performance

Table 2: HALO backtest performance across poll cadences, 11.3 days of SOL/USD data. SC PnL = spread capture PnL; MTM PnL = inventory mark-to-market PnL.

Cadence	Fills	Fills/Day	SC PnL	MTM PnL	Total PnL	Sharpe	Max DD
1 row (∼21s)	333	29	+\$586	−\$648	−\$63	−1.6	−0.45%
5 rows (∼1.75m)	690	61	+\$1,215	−\$1,329	−\$114	−2.1	−0.74%
10 rows (∼3.5m)	851	75	+\$1,497	−\$1,443	+\$55	+1.3	−0.53%
30 rows (∼10.5m)	758	67	+\$1,334	−\$1,110	+\$224	+2.4	−0.70%

5.3 Analysis

Spread capture PnL is positive at every cadence, confirming that the A-S quote placement logic is functioning correctly. The drag on performance at faster cadences is attributable to

inventory MTM losses from adverse selection.

The relationship between cadence and performance is non-monotonic in fill count but monotonic in risk-adjusted returns. Fill count peaks at cadence 10 (851 fills, 75 per day), but the best Sharpe ratio occurs at cadence 30 (+2.4), where the longer interval between quote updates allows inventory to mean-revert naturally before the next fill is triggered.

This finding is counterintuitive from a traditional market making perspective, where faster systems are generally assumed to be better. For a research system without latency advantages, deliberately slowing the poll cadence is a rational design choice: the system trades fill frequency for improved inventory management.

5.4 Poll Cadence as a Strategy Parameter

The multi-cadence comparison demonstrates that poll cadence is itself a strategy parameter that should be optimized alongside γ, κ, and spread bounds. The optimal cadence on this dataset 30 rows, approximately 10.5 minutes, reflects the reality that slower polling reduces adverse selection exposure by giving inventory more time to revert between fills. This connects to a broader principle: a slow system with limited adverse selection detection should quote conservatively and accept lower fill rates in exchange for better inventory management.

6. Discussion

6.1 What Market Making Actually Requires

The HALO development process produced a practical understanding of what market making requires beyond the A-S inventory management framework. A complete market making system needs at minimum:

- **Informed flow detection:** the ability to distinguish noise traders from informed traders. OFI is one approach; VPIN is the industry standard.
- **Venue selection:** choosing pairs and exchanges where the ratio of uninformed to informed flow is sufficient.
- **Competitive quoting:** being at or near the top of the book.
- **Inventory management:** the A-S framework addresses this layer well, but requires correct parameterization of κ and appropriate regime-adaptive γ.
- **Risk limits:** kill switches that halt quoting when PnL deteriorates beyond a threshold.

6.2 HALO as a Professional MM Simulator

Each component of HALO maps to a corresponding element of the professional market making workflow. The A-S engine simulates the reservation price and spread computation that professional systems perform in microseconds. The regime classifier simulates continuous volatility state detection. The OFI filter simulates adverse selection detection via flow toxicity measures. The performance attribution separating SC PnL from MTM PnL mirrors the risk decomposition that professional market making desks use to evaluate their books.

6.3 Limitations

- **Poll cadence:** 30-second data collection means HALO misses most fill opportunities in a continuous market.
- **Fill simulation:** the interval-crossing model assumes fills occur at the first crossing price; real fills depend on queue position and venue-specific matching rules.
- **OFI approximation:** volume imbalance at the top of book is a simplified proxy for true order flow imbalance.
- **Parameter stability:** all parameters were calibrated on a single 11.3-day window; a robust system would use walk-forward optimization.
- **No kill switch:** the current implementation has no automatic halt on adverse PnL drawdown.

7. Future Work

Several extensions are planned:

- **VPIN implementation:** replace the simplified OFI filter with a proper Volume-Synchronized Probability of Informed Trading measure.
- **Live kappa calibration:** estimate κ from observed trade frequency rather than treating it as a fixed constant.
- **Kill switch and risk limits:** implement automatic halting logic when PnL drawdown exceeds a threshold.
- **Extended backtesting:** accumulate 30+ days of SOL/USD data for walk-forward backtesting across multiple regime periods.

- **Multi-pair deployment:** extend HALO to run simultaneously on multiple pairs with correlated inventory management.

8. Conclusion

This paper presented HALO, a market making system implementing the Avellaneda-Stoikov framework with regime-adaptive γ, competitor floor modeling, volume-imbalance OFI filtering, and a multi-cadence backtest framework. The most significant empirical finding was the relationship between poll cadence and risk-adjusted performance: spread capture is consistently positive at all cadences, but inventory MTM losses from adverse selection dominate at faster cadences, with the system achieving its best Sharpe ratio of +4.0 at the slowest cadence tested.

HALO is designed as a learning system as much as a strategy. It is the foundation on which QFC will build a deeper understanding of market microstructure, informed flow detection, and competitive quoting in cryptocurrency markets.

References

[1] Avellaneda, M. and Stoikov, S. (2008). High-frequency trading in a limit order book. *Quantitative Finance*, 8(3), 217–224.

[2] Guéant, O., Lehalle, C.A., and Fernandez-Tapia, J. (2013). Dealing with the inventory risk: A solution to the market making problem. *Mathematics and Financial Economics*, 7(4), 477–507.

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

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

[5] Stoikov, S. and Saglam, M. (2009). Option market making under inventory risk. *Review of Derivatives Research*, 12(1), 55–79.

[6] Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2), 357–384.

[7] Hurst, H.E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, 116, 770–799.