Order Flow Modelization

Wentao Wang, Yuhang Guo

Présentation du cours APM_52112_EP

20 jan 2026

Overview

- **Huang et al.** model the limit order book as a state-dependent Markov queue system linking order flows to price dynamics.
- **Shternshis et al.** Price predictability at ultra-high frequency: Entropy-based randomness test.

Motivation

- Classical LOB models assume independent Poisson order flows, fails to capture empirical dependence on order book.
- Need for model that considers queue sizes, liquidity imbalance and local microstructure.
- **Key idea:** View the LOB as a Markov queuing system.
    - Proof ergodicity and stationary distribution under periods of fixed reference price;
    - Introduce inter-level dependencies and bid-ask interations;
    - Constructs model linking microscopic queue dynamics to price formation and volatility.

Model Setup

Goal: Model LOB dynamics conditional on a fixed reference price.

$$X(t) = \big(q_{-K}(t), \ldots, q_{-1}(t), q_1(t), \ldots, q_K(t)\big) \in \mathbb{N}^{2K}$$

With $q_i(t)$ the queue size at distance $i - \frac{1}{2}$ ticks from reference price, bid-ask treated symmetrically.
We now allow only single-queue unit jumps:

$$q \ \longrightarrow \ q \pm e_i, \ \ e_i = (0, ..., position\ i \ = \ 1, ..., 0)$$

Thus the infinitesimal generator:

$$Q_{q,q+e_i} = f_i(q), \qquad Q_{q,q-e_i} = g_i(q),$$

$$Q_{q,q} = -\sum_{p \neq q} Q_{q,p}, \qquad Q_{q,p} = 0 \quad \text{otherwise.}$$

Stability and Ergodicity

Assumption 1: Negative drift:

$$q_i > C \quad \Rightarrow \quad f_i(q) - g_i(q) < -\delta.$$

Assumption 2: Bounded arrivals:

$$\sum_i f_i(q) \leq H.$$

Theorem (Ergodicity)

Under given assumptions, the process $X(t)$ admits a unique stationary distribution:

$$\lim_{t\to\infty} \mathbb{P}\big(X(t) = p \mid X(0) = q\big) = \pi(p).$$

This justifies long-run queue distributions and enables simulations and estimations.

Model 1: Independent Queues

Key Assumptions: Each price level evolves independently; LO insertions, cancellations, MO conditionally independent.
Each queue is modeled as a birth-death process, respectively for arrival/departure:

$$f_i(q) = \lambda_i^L(q_i), \qquad g_i(q) = \lambda_i^C(q_i) + \lambda_i^M(q_i).$$

To estimate these constants, we consider for the triplet:

$$\big(\Delta t_i(\omega),\ T_i(\omega),\ q_i(\omega)\big), \qquad T_i(\omega) \in \{E^+, E^-, E^t\}.$$

We apply the MLE method and finally get the explicite values:

$$\widehat{\Lambda}_i(n) = \frac{1}{\mathbb{E}[\Delta t_i \mid q_i = n]}.$$

Model 1: Independent Queues

Now we define the local balance ratio:

$$\rho_i(n) = \frac{\lambda_i^L(n)}{\lambda_i^C(n+1) + \lambda_i^M(n+1)}.$$

Since queues are independent, the stationary distribution of the queue is:

$$\pi_i(n) = \pi_i(0) \prod_{k=1}^{n} \rho_i(k-1),$$

With normalization constant:

$$\pi_i(0) = \left(1 + \sum_{n=1}^{\infty} \prod_{k=1}^{n} \rho_i(k-1)\right)^{-1}.$$

Model 2a: Same-side Dependency

Now we consider dependency to neighboring queue states and liquidity of the opposite side. In this case, f_i and g_i are no longer functions of q_i alone.
Model 2a: we now consider the behavior of the second-best queue $Q_{\pm 2}$ depends on whether the best queue $Q_{\pm 1}$ is empty (same-side dependency).

- Best queues $Q_{\pm 1}$:

$$f_{\pm 1}(q) = \lambda^L_{\pm 1}(q_{\pm 1}), \qquad g_{\pm 1}(q) = \lambda^C_{\pm 1}(q_{\pm 1}) + \lambda^M_{\text{buy/sell}}(q)$$

- Second-best queues $Q_{\pm 2}$:

$$f_{\pm 2}(q) = \lambda^L_{\pm 2}\left(q_{\pm 2}, \mathbf{1}_{\{q_{\pm 1} > 0\}}\right), \qquad g_{\pm 2}(q) = \lambda^C_{\pm 2}\left(q_{\pm 2}, \mathbf{1}_{\{q_{\pm 1} > 0\}}\right)$$

- Thus we have the market order allocation:

$$g_i(q) = \lambda^C_i(q) + \lambda^M_{\text{buy}}(q)\,\mathbf{1}_{\{i = \text{best ask}\}} + \lambda^M_{\text{sell}}(q)\,\mathbf{1}_{\{i = \text{best bid}\}}$$

The pair $(Q_{\pm 1}, Q_{\pm 2})$ forms a QBD process, allowing computation of the stationary distribution via matrix-geometric methods.

Model 2b: Bid-Ask Interaction

Model 2b: Now order placement and aggressiveness depend on liquidity on the opposite side of the book.
We define:

$$S_{m,l}(x) = \begin{cases} Q_0, & x = 0, \\ Q_-, & 0 < x \le m, \\ Q_\circ, & m < x \le l, \\ Q_+, & x > l. \end{cases}$$

Then we have the intensity dependence at best quotes for the best bid/ask:

$$\lambda_{\pm 1}^{L,C,M} = \lambda_{\pm 1}^{L,C,M}\left(q_{\pm 1},\, S_{m,l}(q_{\mp 1})\right).$$

This structure does not have closed-form stationary distribution, so Monte-Carlo Method is applied to obtain its stationary behavior.

Model 3: Queue-Reactive Model

Model 3: Couple queue dynamics with a stochastic evolution of the reference price p^{ref}, yielding a joint price–queue Markov model.
Now we define the system state variable:

$$\widetilde{X}(t) = \big(q_{-K}(t), \ldots, q_{-1}(t), q_1(t), \ldots, q_K(t),\ p^{\mathrm{ref}}(t)\big).$$

And the system follows the dynamics as below:

- Between price moves: queues evolve as in Model 1/2;
- Extreme queue events may trigger price jumps.

Model 3: Reference Price Dynamics

Let δ denote the tick size.

Price jump triggers:

- Market order consumes the last unit at $Q_{\pm 1}$;
- Cancellation empties the best queue $Q_{\pm 1}$;
- Limit order arrives inside the spread while $Q_{\mp 1} = 0$.

Price update rule:

$$p^{\mathrm{ref}} \longrightarrow p^{\mathrm{ref}} \pm \delta \quad \text{with probability } \theta$$

Queue relabeling:

- The former best queue becomes empty;
- Remaining queues are shifted by one level;
- Queue sizes are renormalized using AES ratios.

Model 3: Slippage and Exogenous Information

We define the Slippage as follow:

$$\text{Slippage} = \frac{P_{\text{benchmark}} - P_{\text{exec}}}{P_{\text{benchmark}}}, \qquad P_{\text{exec}} = \frac{1}{n_{\text{total}}} \sum_k p_k \, \Delta n_k .$$

The benchmark can be chosen from arrival price or global VWAP.
After a price jump, with probability θ_{reinit}, the LOB state is redrawn from its stationary distribution:

$$(q_{-K}, \dots, q_K) \sim \pi_{\text{stat}}.$$

With θ the endogenous (mechanical) price dynamics and θ_{reinit} the exogenous information intensity.

Conclusion

This work develops a queue-based framework for modeling order book dynamics and price formation:

- Starting from Markov queue models with a fixed reference price;
- Extending to state-dependent interactions within and across sides;
- Introducing a queue-reactive mechanism for endogenous price moves.

Experiments and simulations show that:

- Queue-based models reproduce stationary structure of OB;
- Purely endogenous dynamics yield low volatility and strong mean reversion;
- Exogenous reinitialization is needed to match market volatility.

$\Rightarrow$ The queue-reactive model provides a time-consistent joint description of order flow and price dynamics for execution and market impact analysis.

Summary

- Research question: Are tick-by-tick price directions **statistically predictable** (i.e., not i.i.d.)?
- Method: two hypothesis tests for independence
  - Hypotheses: H_0 next symbol is independent of past history;
  - (1) Entropy-bias test (B): estimate **Shannon entropy** of symbol blocks and test deviation from the maximum-entropy i.i.d. benchmark (implemented with **non-overlapping** blocks).
  - (2) NP/KL test (D): Neyman–Pearson likelihood-ratio / **KL divergence** on **overlapping** blocks;
- Why predictable? Compare predictable vs. non-predictable days
- When within the day? Localize predictability by partitioning the day and applying multiple testing (Šidák)

Method I: Entropy-bias test (non-overlapping blocks)

- **Input:** a symbolic sequence $X = \{x_t\}_{t=1}^{n}$, alphabet size s (in this paper $s = 2$ for up/down).
- **Build non-overlapping blocks** of length k:

$$\hat{x}_t = (x_{(t-1)k+1}, \ldots, x_{tk}), \quad n_b = \lfloor n/k \rfloor.$$

- **Empirical frequencies** $\hat{f}_j$ over all s^k blocks $\{a_j\} \subset A^k$ and **Shannon entropy estimator**

$$\hat{H} = -\sum_{j=0}^{s^k-1} \frac{\hat{f}_j}{n_b} \ln\Big(\frac{\hat{f}_j}{n_b}\Big).$$

- **Test statistic (entropy bias):**

$$B = 2n_b\,(k \ln s - \hat{H}).$$

- **Lemma 1 (asymptotic null):** under the **maximum-entropy i.i.d. benchmark** (equiprobable blocks),

$$B \xrightarrow{d} \chi^2_{s^k-1}.$$

- Reject H_0 (independence) if p-value $< \alpha$ (paper uses $\alpha = 0.01$); reject $\Rightarrow$ **predictable**.

Method II: NP / KL-divergence test (overlapping blocks)

- **Key idea:** use **overlapping** contexts of length $(k-1)$ and test whether the next symbol depends on the past.
- **Overlapping blocks:**

$$\bar{x}_t = (x_t, \dots, x_{t+k-2}), \quad t = 1, \dots, n-k+2.$$

- **Counts / empirical frequencies:**

$$f_{ij} = \sum_{t=1}^{n-k+1} \mathbf{1}(\bar{x}_t = a_i)\, \mathbf{1}(x_{t+k-1} = a_j), \quad f_{i.} = \sum_j f_{ij}, \quad f_{.j} = \sum_i f_{ij}.$$

- **NP statistic (scaled KL divergence):**

$$D = 2 \sum_{i,j} f_{ij} \ln\left(\frac{(n-k+1) f_{ij}}{f_{i.} f_{.j}} \right).$$

- **Lemma 2 (asymptotic null):** if H_0 (independence) holds,

$$D \xrightarrow{d} \chi^2_{(s^{k-1}-1)(s-1)}.$$

- **Advantage vs Method I:** does **not require** symbols to be equiprobable; uses more data via overlapping blocks.

Data preprocessing: transaction-time aggregation & 0-duration trades

- **Transaction time:** sample prices at each execution (tick-by-tick).

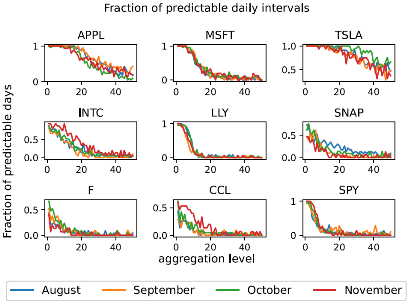
$$r_t = \ln \frac{P_t}{P_{t-1}}, \qquad s_t = \mathbf{1}(r_t > 0) \in \{0, 1\} \quad (\text{remove } r_t = 0).$$

- **Aggregation level a:** number of transactions per one time step.
  - Construct aggregated prices by taking the **last available price** every a transactions:

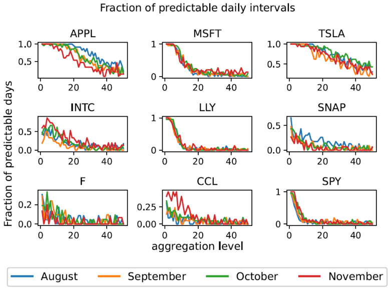
$$P_m^{(a)} = P_{ma}.$$

  - Compute $r_m^{(a)} = \ln\left(P_m^{(a)} / P_{m-1}^{(a)}\right)$, discretize to $\{0, 1\}$, remove zero returns.
- **0-duration (same nanosecond) transactions:** timestamps have 1ns precision, so multiple trades may share the same time.
  - Build an alternative dataset by **aggregating within each nanosecond**: sum volumes and keep the **final available price** of that nanosecond.
  - Then apply the same pipeline (sign encoding + optional aggregation level a).

- **Conclusion:** As aggregation level *a* increases (coarser transaction-time sampling), the fraction of predictable days **decreases** (prices "whiten" under aggregation).



(a) NP-statistics



(b) entropy bias

Impact of 0-duration trades (simultaneous transactions)

- **Conclusion:** Removing/aggregating same-nanosecond trades (*w/o 0-duration*) leads to a **lower measured predictability** than the full record (*with 0-duration*).

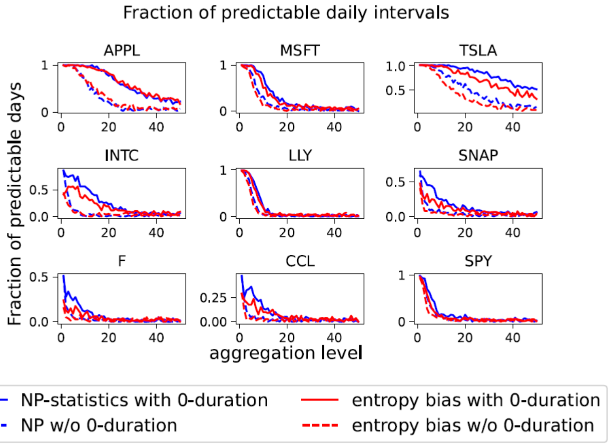


Fig. 7. Fraction of predictable days for different assets with and without simultaneous transactions.

Predictable vs non-predictable days: what differs?

Predictable days of AAPL, MSFT, INTC, LLY, F, CCL, and SPY are characterized by larger trading volumes and a larger number of non-zero price changes compared with non-predictable days.

Table 3
Statistics for predictable and not predictable days.

Parameter	AAPL		MSFT	TSLA	INTC	LLY	SNAP	F	CCL	SPY
Aggregation level	15	30	10	25	2	5	1	1	1	5
Number of predictable days	35	3	26	35	44	44	53	41	38	29
Number of non-zero returns	$>^{**}$		$>^{**}$	$>^{*}$	$>^{**}$	$>^{**}$	$>^{**}$	$>^{**}$	$>^{*}$	$>^{*}$
Fraction of 0-returns							$>^{*}$	$>^{**}$	$>^{**}$	
k		$>^{**}$				$>^{*}$		$>^{**}$	$>^{*}$	
$(\hat{p}(0\dots0)+\hat{p}(1\dots1))2^k$	$>^{**}$		$>^{**}$	$>^{**}$	$>^{**}$	$>^{**}$				$>^{**}$
Magnitude of daily log-price increment								$>^{**}$		
Mean price returns							$>^{*}$		$>^{*}$	
Magnitude of autocorrelation of non-zero returns	$>^{**}$			$>^{**}$	$>^{**}$	$>^{**}$	$>^{**}$	$>^{**}$	$>^{**}$	$>^{**}$
Magnitude of autocorrelation of absolute values	$>^{**}$	$<^{**}$					$>^{**}$	$>^{*}$	$>^{**}$	
ν of t-distribution		$<^{*}$						$>^{*}$	$>^{**}$	$<^{*}$
Scale of t-distribution		$<^{*}$				$<^{*}$	$<^{*}$			
Daily volume	$>^{*}$		$>^{**}$		$>^{*}$	$>^{**}$	$>^{**}$	$>^{**}$	$>^{**}$	$>^{*}$
Fraction of jumps		$<^{**}$							$>^{*}$	

For each stock and aggregation level in columns and price characteristics in rows, we compare means of the statistics between predictable and not predictable days. Predictability of days are defined by NP-statistics. $>$ stands for larger mean for predictable days, $<$ stands for larger mean for not predictable days. * and ** stand for the rejection of the hypothesis that characteristics in two types of days have the same mean. There is no significant difference found for parameters $|\hat{p}(1)-\hat{p}(0)|$ and magnitude of shift of t-distribution. For AAPL stock, the results for two aggregation levels are shown.

* Is for 0.05 level of significance.

** Stands for 0.01 significance.

When within the day is it predictable? (localization results)

- **Procedure:** split each trading day into S **non-overlapping** intervals (min length ≥ 1000 symbols); apply multiple testing with **Šidák correction**

$$\alpha_S = 1 - (1-\alpha)^{1/S} \quad (\alpha = 0.01),$$

and detect whether predictability appears in any sub-interval.

- **Main findings:**
  - Two predictable intervals generally do not occur consecutively. In other words, following an interval with detected predictability, the subsequent interval does not display a significant level of predictability.
  - However, considering all transactions of the SNAP stock (Snap Inc.), we are able to detect several predictable time intervals going in a row.