1. Motivation for Measuring OFI at Multiple Depth Levels

Order flow imbalance (OFI) computed solely at the best bid–ask *level 0* gives an incomplete picture of supply-and-demand pressure. The paper *Cross-Impact of Order Flow Imbalance in Equity Markets* justifies extending OFI down the book as follows:

1. **Liquidity resides beyond the top quote.** Empirical depth studies show that size resting two or three ticks away can exceed the displayed size at the best price levels. Ignoring those queues understates the true inventory that market participants can immediately trade against. Consequently, deeper-level order submissions and cancellations carry information about future price moves.

2. **A "natural extension" of best-level OFI.** Section 2.1.2 formally defines level–m OFI,

$$\mathrm{OFI}_{i,t}^{(m)}(h) = \sum_{n=N(t-h)+1}^{N(t)} \left(\mathrm{OF}_{i,n}^{(m,b)} - \mathrm{OF}_{i,n}^{(m,a)}\right),$$

   noting that multi-level vectors $\left(\mathrm{ofi}^{(1)}, \ldots, \mathrm{ofi}^{(10)}\right)$ simply generalise equation (1) for $m > 0$.

3. **Higher explanatory power for returns.** Adding deeper levels raises in-sample and out-of-sample R^2 until about eight levels, demonstrating incremental predictive content over the top quote alone.

4. **Strong common structure yet non-redundant information.** A principal-component analysis of 10-level OFIs shows the first component explains $\approx 89\%$ of total variance. That common factor, termed the *integrated OFI*, delivers markedly better price-impact fits than best-level OFI, confirming that deeper levels contribute information not captured by level 0.

5. **Diminished need for cross-asset terms once depth is included.** When multi-level information is aggregated into an integrated OFI, adding cross-stock OFIs provides little extra explanatory power; by contrast, a model with only best-level OFIs *does* benefit from cross-asset terms. This result underscores that multiple depth levels already embed the bulk of the relevant order-flow signal.

Intuitive explanation. Picture a football stadium with two automated turnstiles: one for *entries* (buys) and one for *exits* (sells). If you only count *how many* people pass through both gates in a minute (total volume), you know the stadium was busy but not whether the crowd inside grew or shrank. What moves the noise level in the stands—the "price" of excitement—is the *net* difference between entries and exits: 200 fans in and 180 fans out means the crowd swells by 20, pushing the volume up; 180 in and 200 out means it falls by 20, and the atmosphere quiets.

Order Flow Imbalance is that net difference. It tells you immediately whether the crowd pressure is building or easing, whereas raw volume cannot distinguish a balanced flow from a one-sided surge. Because short-term prices react to which side has the upper hand, the signed imbalance is the sharper predictor.

2. Why the Authors Prefer LASSO to OLS in Cross–Impact Estimation

When modeling cross–asset impact, each regression window contains roughly $N \approx 100$ candidate predictors (one OFI series per peer stock) but far fewer observations (e.g. a 30-minute window at 1-minute frequency provides only 30 data points). Ordinary Least Squares (OLS) fails in such "large-p, small-n" settings and delivers unstable or undefined estimates. The paper therefore replaces OLS with the Least Absolute Shrinkage and Selection Operator (LASSO) for three specific reasons:

1. **Ill-posedness.** With more parameters than observations, the OLS normal equations do not have a unique solution; LASSO remains well-posed under the same dimensionality because the ℓ_1 penalty regularizes the problem :contentReference[oaicite:0]index=0.

2. **Severe multicollinearity among OFIs.** Cross-asset order flows are highly correlated (about 10% of pairwise correlations exceed 0.30, see Figure 3 in the paper), violating OLS assumptions and inflating variances. The LASSO penalty shrinks correlated coefficients toward zero, mitigating multicollinearity.

3. **Economic sparsity and interpretability.** Empirically, only a small subset of stocks meaningfully influence a given asset; most cross-impact coefficients are negligible. LASSO performs automatic variable selection, producing sparse coefficient matrices that are easier to interpret and consistent with the assumption that "only a few assets matter".

4. **Predictive performance.** Out-of-sample R^2 improves once LASSO selects the relevant cross-impact terms; OLS would over-fit in-sample and generalize poorly. Table 3 shows modest but consistent gains when cross-asset LASSO models are added on top of single-asset price-impact baselines.

Intuitive explanation. Imagine trying to explain a child's grade using test scores from the entire class: with dozens of nearly identical predictors and only a handful of report cards, an unregularized fit crumbles. A method that keeps only the few most informative classmates and dampens redundancy yields stabler, clearer insights—precisely what LASSO provides for cross-impact.

3. Why Order Flow Imbalance Outperforms Trade Volume for Forecasting Minute-Ahead Returns

The paper argues that *signed* order flow imbalance (OFI) is a more informative state variable than *unsigned* trade volume when modelling intraday price changes. Four main points are made:

1. **Directionality of pressure.** Volume treats buys and sells symmetrically, whereas OFI nets the two sides of the book. A million shares traded can leave the mid-price unchanged if buy and sell pressure offset; the same notional activity produces a large positive OFI if buying dominates. Cont et al. (2014) therefore conclude that "over short time intervals, *price changes are mainly driven by OFI*".

2. **Earlier information than prints.** OFI incorporates limit-order submissions, cancellations, and quote revisions *before* trades occur. As a result it anticipates price moves that volume, which is recorded only once execution happens, can at best confirm after the fact. Section 2.1 explains how every book event contributes to OFI (equations 1–2).

3. **Higher explanatory power in the data.** When the authors regress one-minute returns on multi-level OFIs (model PI[m]) the out-of-sample R^2 peaks at $\approx 83\%$ for $m = 8$ levels (Table A4). Earlier work that relied on contemporaneous or lagged volume typically achieved single-digit R^2 at the same horizon; the OFI specification therefore captures far more of the short-term variance.

4. **Economic interpretation.** Price moves whenever marginal buyers must pay up or marginal sellers must accept less. Net order imbalance is the mechanical proxy for that pressure, while sheer volume conflates informed trades with liquidity provision. Hence OFI is a closer "cause" of returns, whereas volume is only an imperfect correlate.

Intuitive picture. Imagine watching a tug-of-war: knowing *how many* people are pulling (volume) tells you little about the rope's motion unless you also know which side they stand on. OFI supplies that missing sign, so it predicts which direction the rope — here, the mid-price — will move next.