A Theoretical Jump Spike Pressure Framework for High-Frequency Prediction Calibration

Jatin Chauhan*
chauhanjatin100@gmail.com

Fred Xu
fredxu@cs.ucla.edu

Abstract

We introduce *Jump Spike Pressure Index* (JSPI), a theoretical and practical framework for calibrating fixed high-frequency predictors under sparse jump-like perturbations in limit order book inputs. We model observed windows as noisy discretizations of latent jump-diffusion dynamics, where sparse spikes may encode liquidity shocks, queue depletion, quote dislocations, or price-discovery events rather than mere nuisance noise. JSPI estimates robust local baselines and scales, detects sparse signed residuals, aggregates them into a scalar price-jump pressure signal, and applies a lightweight exponential tilt to the fixed model probabilities. We prove stability bounds for the induced probability correction in total variation and KL divergence, establish rolling-window consistency of the JSPI estimator under sparse contamination and jump-separation conditions, and derive convergence of the calibrated predictor to its oracle counterpart. On the FI-2010 benchmark, JSPI improves F1-score across DeepLOB, LOBster, MLP, and XGBoost over all evaluated prediction horizons, with the most consistent gains at longer horizons. The framework provides a transparent, model-agnostic, and computationally inexpensive calibration layer for high-frequency prediction.

1 Introduction

Modern financial prediction systems increasingly operate on high-frequency, high-dimensional market data. In limit order book (LOB) markets, the state of supply and demand is revealed through a rapidly evolving feature set of bid prices, ask prices, bid depths, ask depths, cancellations, and executions across multiple book levels. This representation is informative but also fragile: the same local movement in the book may reflect routine queue dynamics, temporary quote noise, liquidity withdrawal, aggressive order flow, or a genuine jump-like price discovery event. Distinguishing between these regimes is central to short-horizon prediction, market making, and risk control. A predictor that treats all abrupt movements as nuisance noise may discard precisely the information that precedes directional price moves. Conversely, a predictor that reacts strongly to every local irregularity may become poorly calibrated under transient dislocations, stale quotes, or microstructure artifacts.

This paper studies the problem of calibrating fixed LOB predictors in the presence of sparse, jump-like perturbations in high-frequency financial inputs. We model the observed LOB stream as a discretized noisy observation of an underlying latent order-book process with a regular component and a finite-activity jump component. Building on the assumption of sparsity in the jump component, we do not interpret it merely as exogenous corruption since in high-frequency markets, such events may encode liquidity shocks, quote dislocations, queue depletion, one-sided order-flow pressure, or jump-like price discovery. The central question is therefore not whether spikes should always be removed, but whether their directional content can be extracted in a way that improves the predictions of an already trained model.

*Work done independently.

Preprint.

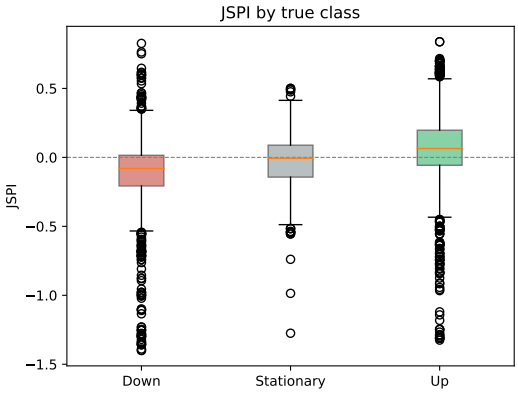


Figure 1: JSPI values across the price movement: a 3-class classification setup here on the FI-2010 (Ntakaris et al., 2018) high-frequency benchmark. WHile a noticeable overlap exists, we observe that even a fraction of the predictions when calibrated can induce a huge change in the downstream prediction as discussed in table 1 which is sufficient for financial predictions.

This fixed-predictor perspective is deliberate. In many financial applications, models are costly to retrain, tightly integrated into production infrastructure, or subject to model-risk constraints. Moreover, when a model has already been trained on a benchmark or production data stream, it is often more realistic to ask whether its outputs can be calibrated at test time than to redesign the entire architecture. We therefore treat the base predictor

$$f_\theta : \mathbb{R}^{T\times D} \to \Delta^2$$

as fixed, where the three classes correspond to downward, stationary, and upward future mid-price movement. The object of study is the base probability vector

$$p_t^{\text{base}} = f_\theta(X_t),$$

and how it should be directionally adjusted when the observed input window contains sparse price-jump evidence.

To this end, we introduce the *Jump Spike Pressure Index* (JSPI), a model-agnostic test-time calibration procedure. JSPI first estimates a local robust baseline for each time-feature coordinate using rolling medians and estimates local scale using a rolling median absolute deviation. It then constructs a sparse jump mask by thresholding standardized deviations from this baseline. The detected signed residuals are aggregated over bid-price and ask-price channels to form a scalar jump-pressure statistic.

Positive values of JSPI_t indicate upward detected price-jump pressure, negative values indicate downward detected pressure, and values near zero indicate either weak jump activity or offsetting signed residuals. This statistic is related in spirit to the order-flow imbalance literature, which links short-horizon price changes to supply-demand imbalances at the book, but it is distinct: JSPI is a thresholded signed price-residual statistic rather than a depth-normalized volume or quote-event imbalance (Cont et al., 2013).

Given JSPI_t, we calibrate the fixed base probabilities by an exponential directional tilt. This form preserves the base predictor and changes only the output probabilities. It increases the upward class probability when the detected jump pressure is positive, increases the downward class probability when the detected jump pressure is negative, and affects the stationary class only through normalization. Operationally, the procedure is lightweight: after computing p_t^{base}, the additional operations are rolling median/MAD estimation, thresholding, aggregation over selected price channels, and a three-class normalization. Figure 1 quantifies our JSPI framework and the behavior across the price movement.

The cross-asset relevance of this formulation is broader than equity LOB prediction alone. In equities, sparse spikes may represent queue depletion, aggressive order-flow bursts, liquidity replenishment, or transient quote dislocations. In cryptocurrency markets, fragmented liquidity and exchange-specific disruptions can generate local spikes, while liquidation cascades and market-wide stress can generate synchronized spikes across venues. In power and energy markets, price spikes are

a fundamental economic feature rather than a mere measurement artifact, arising from scarcity, congestion, outages, load shocks, and renewable forecast errors. In options and derivatives markets, jumps in the underlying or in implied volatility surfaces have direct consequences for hedging, tail-risk pricing, and short-dated option valuation; classical jump-diffusion and stochastic-volatility jump models were introduced precisely to address discontinuous price dynamics and volatility smiles (Merton, 1976; Bates, 1996). Thus, the sparse spike decomposition provides a common language for modeling rare but consequential financial events whose interpretation is market-dependent. We provide detailed discussion on each of the asset classes in Section A.

Our theoretical analysis formalizes three aspects of this calibration scheme. First, we bound the perturbation induced by the exponential tilt on the Total Variation and KL divergence in Theorem 1. These bounds quantify the extent to which JSPI can modify the base probability vector and show that the calibration is stable when the effective tilt is small. Second, in corollary 1 we prove a margin condition under which the calibrated argmax is unchanged; therefore, highly confident base predictions are not necessarily altered by JSPI. Third, under a high-frequency sparse-jump regime, local smoothness, local quantile regularity, sparse contamination, and a jump-separation condition, we show in theorem 2 that the rolling median baseline and MAD scale are uniformly consistent, that the jump mask is recovered with probability tending to one, and that the estimated JSPI converges to its residual oracle at rate $O_p(s_t e_n)$, where s_t is the number of oracle jump coordinates. This further yields convergence of the calibrated prediction distribution to the oracle-calibrated distribution in total variation and KL in theorem 3. Building on the classical high probability results from machine learning theory, we provide in theorem 4 the bound for the calibration hyperparameter η.

Empirically, we evaluate JSPI on the FI-2010 LOB benchmark (Ntakaris et al., 2018), using the standard mid-price movement classification task over horizons $10, 20, 30, 50,$ and 100 ticks - results discussed in Section 5. We test four base predictors: DeepLOB, LOBster, MLP, and XGBoost (Zhang et al., 2019; Jeonghwan-Cheon; Chen & Guestrin, 2016). Across all models and horizons, JSPI improves F1-score. The gains are most consistent at longer horizons, where sparse price-jump residuals appear to contain directional information not fully captured by the fixed predictor. The results support a moderate conclusion: JSPI is a simple, transparent, and computationally inexpensive calibration layer that can improve class-balanced predictive performance while preserving the fixed base model.

2 Problem Setup

2.1 Notation and Discretized Observation Model

We model the observed limit order book (LOB) stream as a discretized high-frequency observation of an underlying latent order-book process. Let

$$(\Omega, \mathcal{F}, (\mathcal{F}_u)_{u\geq 0}, \mathbb{P})$$

be a filtered probability space. For each observed LOB feature channel $j \in \{1, \dots, D\}$, let $Y^j = (Y^j_u)_{u\geq 0}$ denote the latent regular component of the corresponding order-book feature. We use the jump-diffusion representation

$$dY^j_u = b_j(u)\, du + \sigma_j(u)\, dW^j_u + dJ^j_u,$$

where $b_j(u)$ is a local drift, $\sigma_j(u)$ is a local volatility coefficient, W^j is a Brownian motion, and J^j is a finite-activity jump component. Equivalently, one may write

$$J^j_u = \int_{(0,u]\times\mathcal{E}} \gamma_j(s,z)\, N_j(ds, dz),$$

where N_j is a Poisson random measure and $\gamma_j(s,z)$ is the jump size associated with mark z.

The observed LOB feature is a noisy high-frequency observation of this latent process,

$$\widetilde{X}^j_u = Y^j_u + \zeta^j_u,$$

where ζ^j_u denotes market microstructure noise. This formulation is meant to capture the fact that high-frequency order-book observations are not clean samples from a smooth latent process; they

may contain microstructure noise, one-sided quote effects, and economically meaningful jump-like movements (Aït-Sahalia & Jacod, 2012; Bibinger et al., 2024).

The resulting discrete multivariate time series is denoted by

$$\mathbf{x}_i = (X_{i,1}, \ldots, X_{i,D}) \in \mathbb{R}^D.$$

In the LOB setting, $\mathbf{x}_i$ may contain bid prices, ask prices, bid volumes, ask volumes, and additional engineered microstructure features across multiple book levels.

Given a lookback horizon T, we construct the input window

$$\mathbf{X}_t = \begin{bmatrix} \mathbf{x}_{t-T+1}^\top \\ \mathbf{x}_{t-T+2}^\top \\ \vdots \\ \mathbf{x}_t^\top \end{bmatrix} \in \mathbb{R}^{T \times D}.$$

For a prediction horizon T', the corresponding future target sequence may be written as

$$\mathbf{Y}_t = \begin{bmatrix} \mathbf{y}_{t+1}^\top \\ \mathbf{y}_{t+2}^\top \\ \vdots \\ \mathbf{y}_{t+T'}^\top \end{bmatrix} \in \mathbb{R}^{T' \times d_y}.$$

In this work, we focus on the classification-style mid-price movement setting. The target is a discrete class label

$$y_t^{(T')} \in \{-1, 0, +1\},$$

corresponding respectively to downward, stationary, or upward future mid-price movement over horizon T'.

Sparse spike decomposition. The discretized observation window is modeled as the superposition of a latent regular component and a sparse jump or spike component:

$$\mathbf{X}_t = \widehat{\mathbf{X}}_t + \mathbf{S}_t,$$

where

$$\widehat{\mathbf{X}}_t \in \mathbb{R}^{T \times D}$$

denotes the latent regular LOB component and

$$\mathbf{S}_t \in \mathbb{R}^{T \times D}$$

denotes a sparse spike component. Entrywise,

$$S_{t,i,j} = Z_{t,i,j} A_{t,i,j}, \qquad i = 1, \ldots, T, \quad j = 1, \ldots, D,$$

where $Z_{t,i,j} \in \{0, 1\}$ is the spike-location indicator and $A_{t,i,j} \in \mathbb{R}$ is the corresponding spike amplitude. The sparsity assumption in this work is

$$\|\mathbf{S}_t\|_0 \ll TD,$$

meaning that only a small fraction of time-feature coordinates are affected by abrupt jump-like events.

Noteworthy is that in high-frequency LOB data, these sparse spikes correspond to abrupt liquidity shocks, transient quote dislocations, order-flow pressure, queue depletion, liquidity withdrawal, or jump-like price discovery. Rather than removing the spike contamination, our focus instead here is to exploit this estimation for calibration of downstream predictions (discussed in next section).

We provide detailed implicitation of the Spiking model across asset classes in appendix section A **where we consider - equities, crypto, energy and derivaties**.

For structured LOB events, the spike support may be concentrated on a feature subset

$$\mathcal{J}_{\text{spike}} \subseteq \{1, \ldots, D\},$$

such as bid prices, ask prices, bid volumes, ask volumes, or engineered imbalance features. For example, over a contiguous temporal block

$$B = \{\tau, \tau+1, \ldots, \tau+\ell-1\} \subseteq \{1, \ldots, T\},$$

a block spike can be represented as

$$S_{t,i,j} = \begin{cases} \kappa_j, & i \in B,\ j \in \mathcal{J}_{\text{spike}}, \\ 0, & \text{otherwise}, \end{cases}$$

where $\kappa_j > 0$ controls the channel-specific spike magnitude. A liquidity-withdrawal event may similarly be represented as a negative spike over selected volume channels:

$$S_{t,i,j} = \begin{cases} -\kappa_j, & i \in B,\ j \in \mathcal{J}_{\text{liq}}, \\ 0, & \text{otherwise}, \end{cases}$$

where $\mathcal{J}_{\text{liq}}$ denotes the subset of volume-related LOB features whose depletion represents a liquidity shock.

2.2 Fixed Predictor and Classification Target

We consider a fixed trained predictor

$$f_\theta : \mathbb{R}^{T\times D} \to \Delta^{C-1}, \qquad C = 3,$$

where Δ^{C-1} is the probability simplex over the three movement classes $\{-1, 0, +1\}$. Given the observed input window $\mathbf{X}_t$, the predictor outputs

$$\widehat{\mathbf{p}}_t = f_\theta(\mathbf{X}_t) = (\widehat{p}_{t,-1}, \widehat{p}_{t,0}, \widehat{p}_{t,+1}), \qquad \sum_{c\in\{-1,0,+1\}} \widehat{p}_{t,c} = 1.$$

The associated predicted class is

$$\widehat{y}_t^{(T')} = \arg\max_{c\in\{-1,0,+1\}} \widehat{p}_{t,c}.$$

Throughout this work, f_θ is treated as fixed. That is, our analysis does not concern the training procedure used to obtain θ OR the the hypothesis class f.

Instead, we **explicitly** study how the predictions of a **fixed** test-time model, which is trained with the respective objectives (see works such as Rundo et al. (2019),(Rouf et al., 2021) and their citations thereof for the evolution of machine learning techniques in quantitative finance), can be interpreted, adjusted, or calibrated in the presence of the sparse spike component $\mathbf{S}_t$.

Equivalently, one may view the base prediction as

$$\mathbf{p}_t^{\text{base}} = f_\theta(\mathbf{X}_t),$$

where the input $\mathbf{X}_t$ may contain both regular LOB dynamics and sparse spike events. The role of the JSPI construction is then to extract a spike-aware directional signal from $\mathbf{X}_t$ and use it to study or calibrate the fixed base probabilities, without retraining the underlying predictor.

3 Proposed Calibration Method: Jump Spike Pressure Index (JSPI)

We propose a test-time calibration procedure that extracts a sparse price-jump signal from the observed LOB window X_t and uses it to directionally tilt the fixed base predictor f_θ. All notation for X_t, the sparse decomposition $X_t = \widehat{X}_t + S_t$, and the three-class prediction task is as in Section 2.

Step 1: Robust local residuals. For each coordinate (i, j), define a rolling median baseline

$$B_{t,i,j} = \text{median}\{X_{t,k,j} : k \in \mathcal{N}_{r_B}(i)\}, \qquad \mathcal{N}_{r_B}(i) = \{k : |k-i| \le r_B,\ 1 \le k \le T\}.$$

The corresponding residual is

$$R_{t,i,j} = X_{t,i,j} - B_{t,i,j}.$$

We estimate the local scale by the rolling Median Absolute Deviation

$$\widehat{\sigma}_{t,i,j} = 1.4826 \cdot \text{median}\left\{|X_{t,k,j} - B_{t,k,j}| : k \in \mathcal{N}_{r_\sigma}(i)\right\}.$$

The constant $1.4826 = 1/\Phi^{-1}(0.75)$ is the standard normal-reference consistency factor for the MAD. We emphasize that the Gaussian consistency factor in the MAD is a scale convention, not a distributional assumption. Indeed the log-normal price models imply Gaussian log-returns only in an idealized setting (notice the logarithmic transform), high-frequency LOB residuals are typically noisy and heavy-tailed. Our use of the MAD is therefore the robust scale estimate on the usual standard-deviation scale under a normal reference law.

The use of local order statistics is motivated by jump-detection methods for noisy high-frequency order prices, where jumps are localized using robust local statistics under LOB microstructure noise (Bibinger et al., 2024). We built our model on similar framework, albeit under local statistics not necessarily order statistics.

Step 2: Jump mask and sparse residual. Define the standardized residual

$$A_{t,i,j} = \frac{|R_{t,i,j}|}{\widehat{\sigma}_{t,i,j} + \varepsilon},$$

where $\varepsilon > 0$ prevents division by small local scales. The estimated jump mask is

$$\widehat{Z}_{t,i,j} = \mathbf{1}\{A_{t,i,j} > \tau\},$$

with threshold $\tau > 0$. The estimated sparse signed residual is

$$\widehat{S}_t = \widehat{Z}_t \odot R_t, \qquad \widehat{S}_{t,i,j} = \widehat{Z}_{t,i,j}(X_{t,i,j} - B_{t,i,j}).$$

Thus, non-detected coordinates are set to zero and detected coordinates retain their signed deviation from the robust local baseline.

Step 3: Price-jump pressure aggregation. Let $\mathcal{J}_{\text{bidPrice}}$ and $\mathcal{J}_{\text{askPrice}}$ denote the bid-price and ask-price feature indices. We aggregate the detected signed price residuals as

$$P_t^{\text{bid}} = \sum_{i=1}^{T} \sum_{j \in \mathcal{J}_{\text{bidPrice}}} \widehat{S}_{t,i,j}, \qquad P_t^{\text{ask}} = \sum_{i=1}^{T} \sum_{j \in \mathcal{J}_{\text{askPrice}}} \widehat{S}_{t,i,j}.$$

The Jump Spike Pressure Index is

$$\text{JSPI}_t = P_t^{\text{bid}} + P_t^{\text{ask}}.$$

Positive values may indicate upward detected price-jump pressure, negative values indicate downward detected price-jump pressure, and values near zero indicate either weak detected jump activity or offsetting signed residuals across levels and sides. This differs from order-flow imbalance: JSPI is a thresholded signed price-residual statistic, whereas OFI is a quote-event imbalance statistic based on supply-demand changes at the book (Cont et al., 2013).

Step 4: Directional probability tilting. Let

$$p_t^{\text{base}} = f_\theta(X_t)$$

be the fixed model probability vector. For classes $c \in \{-1, 0, +1\}$, define

$$p_t^{\text{final}}(c) = \frac{p_t^{\text{base}}(c) \exp(\eta c \, \text{JSPI}_t)}{\sum_{c' \in \{-1,0,+1\}} p_t^{\text{base}}(c') \exp(\eta c' \, \text{JSPI}_t)},$$

where $\eta \geq 0$ controls the calibration strength. The final prediction is

$$\widehat{y}_t = \arg\max_{c \in \{-1,0,+1\}} p_t^{\text{final}}(c).$$

This exponential tilt increases the probability of the upward class when $\text{JSPI}_t > 0$, increases the probability of the downward class when $\text{JSPI}_t < 0$, and leaves the stationary class affected only through normalization.

The end to end pipeline is described in Algorithm 1.

Remark on short temporal windows. When $T = 1$, temporal rolling baselines and scales cannot be estimated from the input window alone. In that case, JSPI requires either a pre-aggregation temporal window, cross-sectional local structure, or externally estimated baseline and scale quantities. We therefore treat the $T = 1$ case as requiring additional assumptions.

Figure 3 provides an example jump detection over a 100 temporal window length block over the FI-2010 benchmark.

4 Theoretical Results

This section states the main theoretical properties of the proposed JSPI calibration framework. We focus on three questions. First, how much can the exponential JSPI tilt move the fixed base prediction? Second, under what high-frequency sparse-jump conditions is the JSPI estimator consistent? Third, how does consistency of JSPI translate into consistency of the calibrated prediction distribution? An additional independent contribution is the derivation of a selection rule for η using empirical risk bounds under negative association of the risk (loss) at per sample level. Proofs are deferred to Appendix B with the restatements of these theorems for completeness.

4.1 Prediction stability under JSPI tilting

Theorem 1 (Prediction variation under JSPI exponential tilting)**.** *For any fixed $p_t^{\text{base}} \in \Delta^2$, the JSPI-calibrated distribution satisfies*

$$d_{\text{TV}}\left(p_t^{\text{base}}, p_t^{\text{final}}\right) \leq \tanh\left(\frac{|a_t|}{2}\right).$$

Equivalently,

$$\left\|p_t^{\text{base}} - p_t^{\text{final}}\right\|_1 \leq 2\tanh\left(\frac{|a_t|}{2}\right) \leq |a_t|.$$

Moreover, both KL divergences are quadratically controlled:

$$D_{\text{KL}}\left(p_t^{\text{base}} \,\|\, p_t^{\text{final}}\right) \leq \frac{a_t^2}{2}, \qquad D_{\text{KL}}\left(p_t^{\text{final}} \,\|\, p_t^{\text{base}}\right) \leq \frac{a_t^2}{2}.$$

Consequently,

$$D_{\text{KL}}\left(p_t^{\text{base}} \,\|\, p_t^{\text{final}}\right) + D_{\text{KL}}\left(p_t^{\text{final}} \,\|\, p_t^{\text{base}}\right) \leq a_t^2.$$

The bound is independent of the architecture of f_θ, since the predictor is treated as fixed and only its output probability vector is tilted.

Corollary 1 (Argmax stability under a base margin)**.** *Let*

$$c_t^\star = \arg\max_{c \in \{-1,0,+1\}} p_t^{\text{base}}(c)$$

and define the base prediction margin

$$\gamma_t = p_t^{\text{base}}(c_t^\star) - \max_{c \neq c_t^\star} p_t^{\text{base}}(c).$$

If

$$\gamma_t > 2\tanh\left(\frac{\eta|\text{JSPI}_t|}{2}\right),$$

then

$$\arg\max_c p_t^{\text{final}}(c) = \arg\max_c p_t^{\text{base}}(c).$$

Thus, JSPI does not necessarily change highly confident predictions. Its impact is concentrated near decision boundaries, where a small directional tilt may alter the predicted class.

4.2 Consistency of the JSPI estimator

We begin by stating the assumptions under which the rolling-window JSPI estimator recovers the *sparse* residual oracle. Without loss of generality, the theory is stated for the price channels

$$\mathcal{J}_P = \mathcal{J}_{\text{bidPrice}} \cup \mathcal{J}_{\text{askPrice}},$$

since our framework is foused on price imbalances, but the core method is directly applicable to volume channels as well.

For $j \in \mathcal{J}_P$, the local observation is modeled as the discrete sampling-bin realization of a latent jump-diffusion price process:

$$X_{t,i,j} = \mu_j(u_i) + G^\star_{t,i,j} + \xi_{t,i,j}, \qquad i = 1, \dots, T_n.$$

Here $\mu_j(u_i)$ is the local regular component, $G^\star_{t,i,j}$ is the structural jump innovation, and $\xi_{t,i,j}$ collects continuous martingale fluctuations and microstructure noise after local median centering (discussed in detail in section B.2). The oracle jump mask is

$$Z^\star_{t,i,j} = \mathbf{1}\{G^\star_{t,i,j} \neq 0\}.$$

The oracle local baseline is

$$B^\star_{t,i,j} = \mu_j(u_i).$$

We work on an equispaced high-frequency grid similar to (Bibinger et al., 2024) with

$$u_i = i\Delta_n, \qquad \Delta_n \to 0.$$

Let the rolling-window radius satisfy

$$r_n \to \infty, \qquad m_n = 2r_n + 1, \qquad h_n = m_n\Delta_n \to 0.$$

For each index i, define the local neighborhood

$$\mathcal{N}_n(i) = \{k : |k - i| \leq r_n,\ 1 \leq k \leq T_n\}.$$

For simplicity, the theoretical statements are made on the interior set

$$\mathcal{I}_n = \{r_n + 1, \dots, T_n - r_n\},$$

so that $|\mathcal{N}_n(i)| = m_n$. Boundary indices can be handled by one-sided windows or padding, affecting only constants.

Assumption 1 (Local smoothness)**.** *There exist constants $L < \infty$ and $\alpha \in (0, 1]$ such that, uniformly over $j \in \mathcal{J}_P$,*

$$|\mu_j(u) - \mu_j(v)| \leq L|u - v|^\alpha.$$

Thus, for $k \in \mathcal{N}_n(i)$,

$$|\mu_j(u_k) - \mu_j(u_i)| \leq Lh_n^\alpha.$$

Assumption 2 (Local quantile regularity and concentration)**.** *For each price channel j, the local noise distribution is stationary within windows of width h_n, has median zero, and has density bounded away from zero in a neighborhood of its median. The same type of regularity holds for the distribution of $|\xi_{t,i,j}|$ at its median. In addition, the local empirical distribution functions satisfy*

$$\max_{i \in \mathcal{I}_n} \max_{j \in \mathcal{J}_P} \sup_x \left| \widehat{F}_{i,j}(x) - F_{i,j}(x) \right| = O_p\left(\sqrt{\frac{\log(T_n D_n)}{m_n}} \right).$$

For independent local observations, this type of bound follows from standard DKW–Massart concentration, together with a union bound over local windows and channels - which we derive in Lemma 2

Assumption 3 (Sparse local jump contamination)**.** *Let*

$$\pi_n = \max_{i \in \mathcal{I}_n} \max_{j \in \mathcal{J}_P} \frac{1}{m_n} \sum_{k \in \mathcal{N}_n(i)} Z^\star_{t,k,j}.$$

We assume

$$\pi_n \to 0.$$

Thus, each local window contains only an asymptotically vanishing fraction of jump-contaminated observations. ***Note on*** π_n***:*** *While we can relax the requirement of $\pi_n \to 0$ by keeping a controlled growth rate and derive the theoretical results without much adjustments, for a substantial fraction of the trading windows, we believe it is a realistic enough assumption (Aït-Sahalia & Jacod, 2012).*

Assumption 4 (Residual boundedness)**.** *There exists $C_X < \infty$ such that, with probability tending to one,*

$$\max_{i \in \mathcal{I}_n} \max_{j \in \mathcal{J}_P} |X_{t,i,j} - B^\star_{t,i,j}| \le C_X.$$

This can be weakened by allowing $C_X = C_{X,n}$ to grow slowly with n, in which case the final rates acquire the corresponding multiplicative factor.

Assumption 5 (Jump separation)**.** *Let*

$$\sigma^\star_{t,i,j} = 1.4826 \cdot \text{median}\left(|\xi_{t,i,j}|\right)$$

denote the local population MAD scale, and define the oracle standardized residual

$$A^\star_{t,i,j} = \frac{|X_{t,i,j} - B^\star_{t,i,j}|}{\sigma^\star_{t,i,j} + \varepsilon}.$$

Let

$$e_n = h_n^\alpha + \pi_n + \sqrt{\frac{\log(T_n D_n)}{m_n}}.$$

There exists a constant $M > 0$ such that, with probability tending to one,

$$\min_{(i,j): Z^\star_{t,i,j}=1} A^\star_{t,i,j} > \tau_n + M e_n,$$

and

$$\max_{(i,j): Z^\star_{t,i,j}=0} A^\star_{t,i,j} < \tau_n - M e_n.$$

Thus, true jump and non-jump coordinates are separated by a margin larger than the local estimation error.

The implemented JSPI estimator replaces the unknown local baseline $B^\star_{t,i,j} = \mu_j(u_i)$ by a rolling median and detects jumps through standardized deviations from this robust baseline.

Theorem 2 (Rolling-window JSPI consistency)**.** *Suppose Assumptions 1–5 hold. Define the rolling median baseline*

$$B_{t,i,j} = \text{median}\left\{X_{t,k,j} : k \in \mathcal{N}_n(i)\right\},$$

and the rolling MAD scale estimator

$$\widehat{\sigma}_{t,i,j} = 1.4826 \cdot \text{median}\left\{|X_{t,k,j} - B_{t,k,j}| : k \in \mathcal{N}_n(i)\right\}.$$

Let

$$\widehat{A}_{t,i,j} = \frac{|X_{t,i,j} - B_{t,i,j}|}{\widehat{\sigma}_{t,i,j} + \varepsilon}, \qquad \widehat{Z}_{t,i,j} = \mathbf{1}\{\widehat{A}_{t,i,j} > \tau_n\}.$$

Then

$$\max_{i \in \mathcal{I}_n} \max_{j \in \mathcal{J}_P} |B_{t,i,j} - B^\star_{t,i,j}| = O_p(e_n),$$

and

$$\max_{i \in \mathcal{I}_n} \max_{j \in \mathcal{J}_P} |\widehat{\sigma}_{t,i,j} - \sigma^\star_{t,i,j}| = O_p(e_n).$$

Consequently,

$$\mathbb{P}\left(\widehat{Z}_{t,i,j} = Z^\star_{t,i,j} \text{ for all } i \in \mathcal{I}_n,\ j \in \mathcal{J}_P\right) \to 1.$$

Define the estimated sparse residual and oracle sparse residual by

$$\widehat{S}_t = \widehat{Z}_t \odot (X_t - B_t), \qquad S^\star_t = Z^\star_t \odot (X_t - B^\star_t).$$

Let

$$\widehat{\text{JSPI}}_t = \sum_{i \in \mathcal{I}_n} \sum_{j \in \mathcal{J}_{\text{bidPrice}}} \widehat{S}_{t,i,j} + \sum_{i \in \mathcal{I}_n} \sum_{j \in \mathcal{J}_{\text{askPrice}}} \widehat{S}_{t,i,j},$$

and define $\text{JSPI}^\star_t$ analogously using $S^\star_t$. If

$$s_t = \left|\{(i,j) : i \in \mathcal{I}_n,\ j \in \mathcal{J}_P,\ Z^\star_{t,i,j} = 1\}\right|,$$

then

$$\left|\widehat{\mathrm{JSPI}}_t - \mathrm{JSPI}_t^\star\right| = O_p(s_t e_n).$$

In particular, if $s_t e_n \to 0$, then

$$\widehat{\mathrm{JSPI}}_t - \mathrm{JSPI}_t^\star \xrightarrow{p} 0.$$

If the number of oracle price-jump coordinates is fixed, so that $s_t = O(1)$, then

$$\left|\widehat{\mathrm{JSPI}}_t - \mathrm{JSPI}_t^\star\right| = O_p(e_n).$$

Remark 1 (Simplified rate under finite local jump support)**.** *If the number of jump-contaminated coordinates in each local window is uniformly bounded, then $\pi_n = O(m_n^{-1})$. Since this is absorbed by the empirical-quantile rate, the error rate simplifies to*

$$e_n = h_n^\alpha + \sqrt{\frac{\log(T_n D_n)}{m_n}}.$$

4.3 Prediction convergence under JSPI consistency

We next show that consistency of the JSPI statistic implies consistency of the calibrated predictive distribution.

Theorem 3 (Prediction convergence under JSPI consistency)**.** *Suppose*

$$\widehat{\mathrm{JSPI}}_t - \mathrm{JSPI}_t^\star = O_p(\kappa_n), \qquad \kappa_n \to 0.$$

Define

$$p_t^{\mathrm{final}}(c) = \frac{p_t^{\mathrm{base}}(c)\exp\left(\eta c\,\widehat{\mathrm{JSPI}}_t\right)}{\sum_{c'} p_t^{\mathrm{base}}(c')\exp\left(\eta c'\,\widehat{\mathrm{JSPI}}_t\right)},$$

and define $p_{t,\star}^{\mathrm{final}}$ analogously with $\mathrm{JSPI}_t^\star$ in place of $\widehat{\mathrm{JSPI}}_t$. Then

$$d_{\mathrm{TV}}\left(p_t^{\mathrm{final}}, p_{t,\star}^{\mathrm{final}}\right) \le \tanh\left(\frac{\eta\left|\widehat{\mathrm{JSPI}}_t - \mathrm{JSPI}_t^\star\right|}{2}\right) = O_p(\eta\kappa_n).$$

Furthermore,

$$D_{\mathrm{KL}}\left(p_t^{\mathrm{final}} \,\|\, p_{t,\star}^{\mathrm{final}}\right) = O_p(\eta^2\kappa_n^2),$$

and

$$D_{\mathrm{KL}}\left(p_{t,\star}^{\mathrm{final}} \,\|\, p_t^{\mathrm{final}}\right) = O_p(\eta^2\kappa_n^2).$$

Corollary 2 (No-jump self-neutrality)**.** *Suppose the assumptions of Theorem 2 hold and there are no oracle price-jump coordinates in $\mathcal{J}_P$, so that*

$$\mathrm{JSPI}_t^\star = 0.$$

If

$$\widehat{\mathrm{JSPI}}_t = O_p(\kappa_n), \qquad \kappa_n \to 0,$$

then

$$d_{\mathrm{TV}}\left(p_t^{\mathrm{base}}, p_t^{\mathrm{final}}\right) = O_p(\eta\kappa_n),$$

and

$$D_{\mathrm{KL}}\left(p_t^{\mathrm{base}} \,\|\, p_t^{\mathrm{final}}\right) = O_p(\eta^2\kappa_n^2).$$

Thus, in the absence of oracle price jumps, the JSPI correction is asymptotically neutral.

4.4 Hold-out selection of the calibration strength

Finally, we state a high-probability guarantee for choosing the scalar calibration strength η. This result is stated under a hold-out validation assumption. While we use the IID application of concentration inequalities here, it is trivially extended to negative association of the risk - more discussion in section B.3. For temporally dependent validation streams, the same structure suggests replacing n_{val} by an effective sample size under suitable dependence conditions.

Theorem 4 (Hold-out selection of the JSPI calibration parameter)**.** *Let*

$$\mathcal{D}_{\text{val}} = \{(X_t, Y_t)\}_{t=1}^{n_{\text{val}}}$$

be an independent validation set, drawn i.i.d. Assume that f_θ, the JSPI estimator, and the base probabilities p_t^{base} are fixed before evaluating the validation set.

For $\eta \in [0, H]$, define

$$p_{t,\eta}(c) = \frac{p_t^{\text{base}}(c) \exp\left(\eta c \, \text{JSPI}_t\right)}{\sum_{c' \in \{-1,0,+1\}} p_t^{\text{base}}(c') \exp\left(\eta c' \, \text{JSPI}_t\right)},$$

and the validation loss

$$\ell_\eta(t, Y_t) = -\log p_{t,\eta}(Y_t).$$

Let

$$R(\eta) = \mathbb{E}[\ell_\eta(X, Y)], \qquad \widehat{R}_{\text{val}}(\eta) = \frac{1}{n_{\text{val}}} \sum_{t=1}^{n_{\text{val}}} \ell_\eta(t, Y_t),$$

and define

$$\widehat{\eta} \in \arg\min_{\eta \in [0,H]} \widehat{R}_{\text{val}}(\eta).$$

Assume that $|\text{JSPI}_t| \leq B$ almost surely and that the base probabilities are clipped away from zero:

$$p_t^{\text{base}}(c) \geq \rho, \qquad c \in \{-1, 0, +1\},$$

for some $\rho \in (0, 1/3]$. Then, for all $\eta, \eta' \in [0, H]$,

$$0 \leq \ell_\eta(t, Y_t) \leq M_\rho, \qquad M_\rho = \log(1/\rho) + 2HB,$$

and

$$|\ell_\eta(t, Y_t) - \ell_{\eta'}(t, Y_t)| \leq 2B|\eta - \eta'|.$$

Consequently, with probability at least $1 - \delta$,

$$R(\widehat{\eta}) \leq \inf_{\eta \in [0,H]} R(\eta) + 2M_\rho \sqrt{\frac{\log(2N_n/\delta)}{2n_{\text{val}}}} + \frac{4M_\rho}{\sqrt{n_{\text{val}}}},$$

where

$$N_n = 1 + \left\lceil \frac{2BH\sqrt{n_{\text{val}}}}{M_\rho} \right\rceil.$$

Equivalently,

$$R(\widehat{\eta}) \leq \inf_{\eta \in [0,H]} R(\eta) + O\left(M_\rho \sqrt{\frac{\log\left(1 + \frac{BH\sqrt{n_{\text{val}}}}{M_\rho}\right) + \log(1/\delta)}{n_{\text{val}}}} \right).$$

5 Experiments

Setup: We consider the FI-2010 dataset (Ntakaris et al., 2018) due to its size mirroring practical settings along with open access. The task is mid-price forecasting for limit order book data which was originally transformed to a classification problem by the authors to predict - (i) upward movement; (ii) downward movement or (iii) no movement. It has 10 levels of LOB data that has been normalized with three different methods: Z-scoring, min-max and decimal-precision normalization. We provide experimental results for ZScore in table 1 as it is the most widely adopted framework in the literature. It also provides annotated label for 5 different time horizons: 10, 20, 30, 50, and 100 ticks, again we provide empirical results across all the settings.

Models: DeepLOB (Zhang et al., 2019), LOBster Jeonghwan-Cheon, MLP, XGBoost (Chen & Guestrin, 2016).

Table 1: Model performance, accuracy and F1-score, of the base predictions against JSPI calibration, where the columns are the prediction horizons. Statistically Significant results over 10 random seed runs are **highlighted in bold** and we avoid writing the deviations due to space shortage here.

Model	Setting	10		20		30		50		100	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
DeepLOB	Base	0.8230	0.6983	0.7356	0.6199	0.7504	0.6889	0.7512	0.7265	0.7313	0.7315
	W/ JSPI	0.8233	**0.7035**	0.7359	**0.6256**	**0.7526**	**0.6953**	**0.7559**	**0.7338**	**0.7426**	**0.7430**
LOBster	Base	0.8231	0.6939	**0.6239**	0.3292	0.7451	0.6831	0.7527	0.7277	0.7331	0.7329
	W/ JSPI	0.8226	**0.7025**	0.5797	**0.4275**	**0.7484**	**0.6902**	**0.7587**	**0.7363**	**0.7429**	**0.7428**
MLP	Base	**0.7292**	0.4851	**0.6334**	0.4366	**0.5981**	0.4793	0.5055	0.4871	0.4721	0.4503
	W/ JSPI	0.7264	**0.5020**	0.6017	**0.4704**	0.5817	**0.5072**	**0.5165**	**0.5079**	**0.4966**	**0.4691**
XGBoost	Base	0.7842	0.5973	**0.7032**	0.5487	0.7014	0.6220	0.6806	0.6591	0.6152	0.6116
	W/ JSPI	0.7833	**0.6163**	0.7015	**0.5638**	**0.7051**	**0.6340**	**0.6861**	**0.6677**	**0.6279**	**0.6225**

5.1 Results

Table 1 shows that JSPI calibration improves the F1 score for all model families and all FI-2010 prediction horizons. The gains are most consistent at the longer horizons 50 and 100, where sparse price-jump residuals appear to provide directional information not fully captured by the fixed base predictor. For example, DeepLOB improves from 0.7265 to 0.7338 at horizon 50 and from 0.7315 to 0.7430 at horizon 100, while LOBster improves from 0.7277 to 0.7363 and from 0.7329 to 0.7428 over the same horizons.

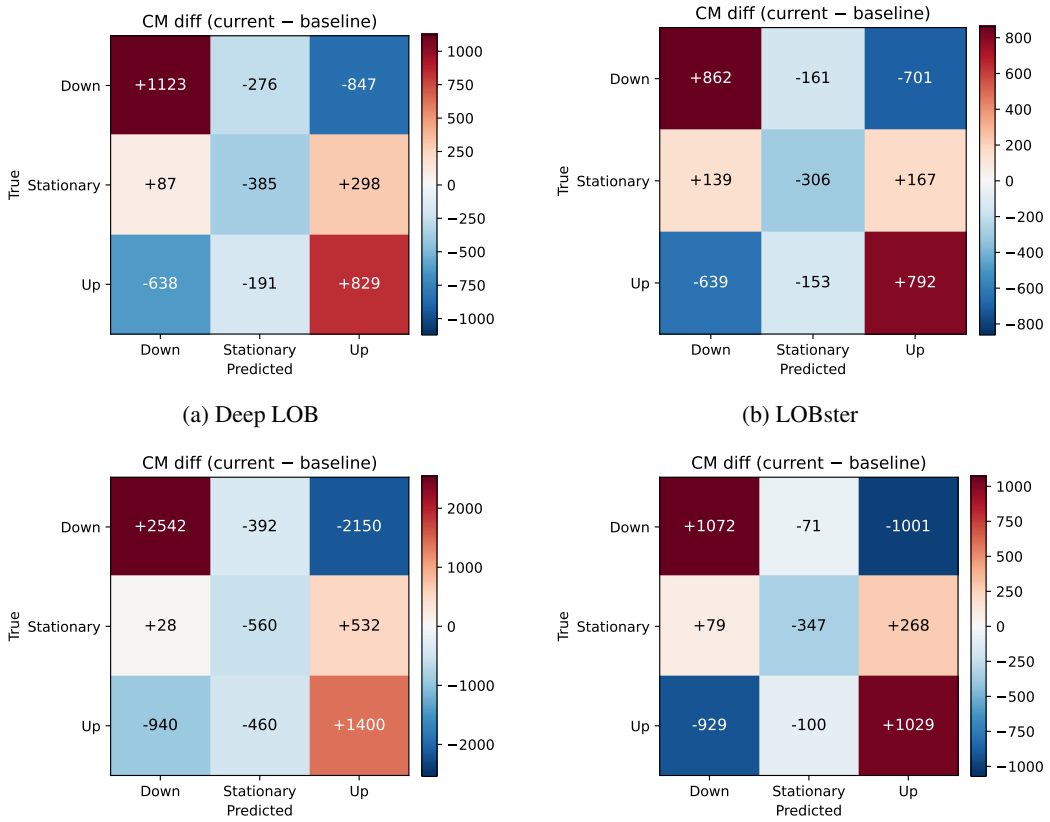
The behavior of accuracy is more mixed. In several cases JSPI improves both accuracy and F1, but there are also settings where F1 improves while accuracy decreases. This is expected for a directional post-hoc calibration: the exponential tilt can move probability mass toward harder or less frequent directional classes, improving class-balanced behavior without necessarily increasing the overall fraction of correct labels. Thus, the empirical benefit of JSPI is better reflected by F1 than by accuracy alone.

The table also suggests that JSPI should be viewed as a lightweight calibration layer rather than a replacement for a strong predictor. When the fixed base model is weak, as in several MLP settings, JSPI improves F1 but does not close the gap to stronger architectures. This is consistent with the design of the method: it does not retrain f_θ, modify the learned representation, or solve poor base-model calibration globally. It only applies a structured directional correction using the sparse price-jump signal extracted from the input window.

The practical advantage is that the calibration is model-agnostic and computationally cheap. After obtaining p_t^{base}, JSPI only requires rolling median/MAD estimation, thresholding, aggregation over price channels, and a three-class exponential normalization. This cost is negligible relative to training or evaluating the models. such as Overall, the results support a moderate conclusion: JSPI provides a simple and transparent test-time calibration mechanism that tends to improve class-balanced predictive performance, especially at longer horizons, while preserving the fixed base predictor.

6 Conclusion and Future Directions

We introduced JSPI as a lightweight, model-agnostic calibration layer for fixed high-frequency predictors under sparse jump-like perturbations. While the empirical results suggest consistent F1-score improvements, the method remains limited by its reliance on robust local thresholding, a scalar directional pressure summary, and assumptions such as sparse contamination and jump separation. Our ongoing and potential future work is focused on adaptive spike masks, multiscale JSPI variants, cross-asset extensions. Another natural direction is to combine JSPI with uncertainty quantification for non-exchangeable time series, since recent conformal and predictive-inference work explicitly studies coverage under temporal dependence and mixing-type assumptions. Finally, connecting JSPI more directly with formal high-frequency jump tests under LOB-specific microstructure noise may yield sharper finite-sample diagnostics and stronger deployment guarantees.



(a) Deep LOB

(b) LOBster

(c) MLP

(d) XGBoost

Figure 2: Difference of values on horizon 100 ticks: computed as the JSPI calibrated confusion matrix minus the base prediction confusion matrix across the 3 classes.

References

Yacine Aït-Sahalia and Jean Jacod. Testing for jumps in noisy high frequency data. *Journal of Econometrics*, 168(2):207–222, 2012. doi: 10.1016/j.jeconom.2011.12.004.

Carol Alexander, Ding Chen, and Arben Imeraj. Crypto quanto and inverse options. *Mathematical Finance*, 33(4):1005–1043, 2023.

David S. Bates. Jumps and stochastic volatility: Exchange rate processes implicit in deutsche mark options. *The Review of Financial Studies*, 9(1):69–107, 1996. doi: 10.1093/rfs/9.1.69.

Christoph Bergmeir, Rob J Hyndman, and Bonsoo Koo. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120: 70–83, 2018.

Markus Bibinger, Nikolaus Hautsch, and Alexander Ristig. Jump detection in high-frequency order prices. *arXiv preprint arXiv:2403.00819*, 2024.

Svetlana Borovkova and Ferry Jaya Permana. Modelling electricity prices by the potential jump-diffusion. In *Stochastic Finance*, pp. 239–263. Springer, 2006.

Likai Chen and Wei Biao Wu. Concentration inequalities for empirical processes of linear time series. *Journal of Machine Learning Research*, 18(231):1–46, 2018.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

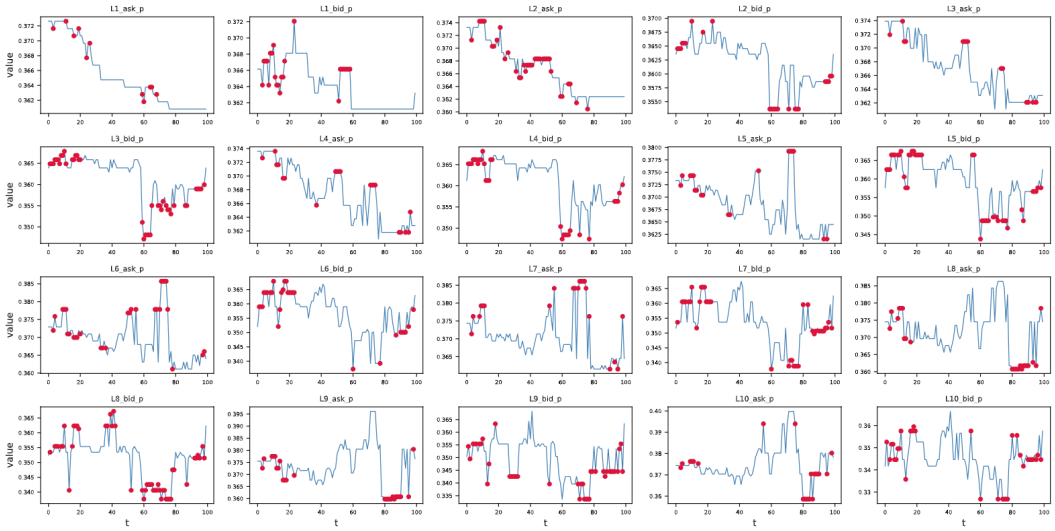


Figure 3: Jumps (in red) detected across the order levels on a data sample of temporal block length 100

R. Cont, A. Kukanov, and S. Stoikov. The price impact of order book events. *Journal of Financial Econometrics*, 12(1), 2013. ISSN 1479-8417. doi: 10.1093/jjfinec/nbt003. URL `http://dx.doi.org/10.1093/jjfinec/nbt003`.

Rama Cont, Sasha Stoikov, and Rishi Talreja. A stochastic model for order book dynamics. *Operations Research*, 58(3):549–563, 2010. doi: 10.1287/opre.1090.0780.

David Easley, Maureen O'Hara, Songshan Yang, and Zhibai Zhang. Microstructure and market dynamics in crypto markets, 2024.

Hanyuan Hang and Ingo Steinwart. A bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. 2017.

Jeonghwan-Cheon. LOBster: Limit order book deep learning method. `https://github.com/Jeonghwan-Cheon/lob-deep-learning/tree/main`. GitHub repository.

Carlo Mari and Cristiano Baldassari. Ensemble methods for jump-diffusion models of power prices. *Energies*, 14(8):2084, 2021.

Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pp. 1269–1283, 1990.

Robert C Merton. Option pricing when underlying stock returns are discontinuous. *Journal of financial economics*, 3(1-2):125–144, 1976.

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for non-iid processes. volume 20, 2007.

H Jabbari Nooghabi and HA Azarnoosh. Exponential inequality for negatively associated random variables. *Statistical Papers*, 50(2):419–428, 2009.

Adamantios Ntakaris, Martin Magris, Juho Kanniainen, Moncef Gabbouj, and Alexandros Iosifidis. Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods. *Journal of Forecasting*, 37(8):852–866, August 2018. ISSN 1099-131X. doi: 10.1002/for.2543. URL `http://dx.doi.org/10.1002/for.2543`.

Nusrat Rouf, Majid Bashir Malik, Tasleem Arif, Sparsh Sharma, Saurabh Singh, Satyabrata Aich, and Hee-Cheol Kim. Stock market prediction using machine learning techniques: a decade survey on methodologies, recent developments, and future directions. *Electronics*, 10(21):2717, 2021.

Abhishek Roy, Krishnakumar Balasubramanian, and Murat A Erdogdu. On empirical risk minimization with dependent and heavy-tailed data. volume 34, pp. 8913–8926, 2021.

Francesco Rundo, Francesca Trenta, Agatino Luigi Di Stallo, and Sebastiano Battiato. Machine learning for quantitative finance applications: A survey. *Applied Sciences*, 9(24):5574, 2019.

Vladimir Vapnik and Vlamimir Vapnik. Statistical learning theory wiley. *New York*, 1(624):2, 1998.

Zihao Zhang, Stefan Zohren, and Stephen Roberts. Deeplob: Deep convolutional neural networks for limit order books. *IEEE Transactions on Signal Processing*, 67(11):3001–3012, 2019. doi: 10.1109/TSP.2019.2907260. arXiv:1808.03668.

Algorithm 1 Jump Spike Pressure Index: Test-Time Prediction Calibration

Require: Input window $X_t \in \mathbb{R}^{T\times D}$, fixed predictor f_θ, threshold τ, correction strength $\eta \geq 0$
Ensure: Calibrated prediction $\widehat{y}_t$

1: Estimate the rolling median baseline

$$B_{t,i,j} = \text{median}\left\{X_{t,k,j} : k \in \mathcal{N}_{r_B}(i)\right\}.$$

2: Estimate the rolling MAD scale

$$\widehat{\sigma}_{t,i,j} = 1.4826 \cdot \text{median}\left\{|X_{t,k,j} - B_{t,k,j}| : k \in \mathcal{N}_{r_\sigma}(i)\right\}.$$

3: Form the jump mask

$$\widehat{Z}_{t,i,j} = \mathbf{1}\left\{\frac{|X_{t,i,j} - B_{t,i,j}|}{\widehat{\sigma}_{t,i,j} + \varepsilon} > \tau\right\}.$$

4: Estimate the sparse signed residual

$$\widehat{S}_t = \widehat{Z}_t \odot (X_t - B_t).$$

5: Compute the base probabilities

$$p_t^{\text{base}} = f_\theta(X_t).$$

6: Aggregate signed price-jump pressure

$$\text{JSPI}_t = \sum_{i=1}^{T} \sum_{j\in\mathcal{J}_{\text{bidPrice}}} \widehat{S}_{t,i,j} + \sum_{i=1}^{T} \sum_{j\in\mathcal{J}_{\text{askPrice}}} \widehat{S}_{t,i,j}.$$

7: Apply exponential directional tilting

$$\widetilde{p}_t(c) = p_t^{\text{base}}(c) \exp\left(\eta c\, \text{JSPI}_t\right), \qquad c \in \{-1, 0, +1\}.$$

8: Normalize

$$p_t^{\text{final}}(c) = \frac{\widetilde{p}_t(c)}{\sum_{c'\in\{-1,0,+1\}} \widetilde{p}_t(c')}, \qquad c \in \{-1, 0, +1\}.$$

9: **return**

$$\widehat{y}_t = \arg\max_{c\in\{-1,0,+1\}} p_t^{\text{final}}(c).$$

A Implications of the Spiking Model

The sparse spike decomposition discussed in Section 2 provides a unified way to reason about abrupt, localized, and economically meaningful perturbations in high-frequency financial inputs.

An important assumption made in this work is the sparsity of the matrix $\widehat{S}_t$, thus, only a small subset of time-feature coordinates is affected by abrupt events. This assumption is natural in high-frequency financial data, where most local observations correspond to regular order-book evolution, while rare intervals contain sharp changes induced by liquidity withdrawal, aggressive order flow, price jumps, market stress, or data-quality artifacts. Importantly, $\mathbf{S}_t$ should not be interpreted merely as exogenous corruption. In many financial markets, spike-like observations may themselves carry predictive information.

A spike-aware predictor may be written as

$$f_\theta : \mathbb{R}^{T\times D} \times \{0,1\}^{T\times D} \to \Delta^{C-1}, \qquad \widehat{\mathbf{p}}_t = f_\theta(\mathbf{X}_t, \widehat{\mathbf{Z}}_t),$$

where

$$\widehat{\mathbf{Z}}_t \approx \mathbf{1}\{\mathbf{S}_t \neq \mathbf{0}\}$$

is an estimated spike-location mask. Equivalently, one may view f_θ as learning a decomposition-sensitive representation

$$h_\theta(\mathbf{X}_t) = h_\theta(\widehat{\mathbf{X}}_t + \mathbf{S}_t),$$

whose desired behavior is not necessarily invariance to $\mathbf{S}_t$, but rather selective stability:

$$f_\theta(\widehat{\mathbf{X}}_t + \mathbf{S}_t) \approx f_\theta(\widehat{\mathbf{X}}_t) \quad \text{when } \mathbf{S}_t \text{ is nuisance noise,}$$

and

$$f_\theta(\widehat{\mathbf{X}}_t + \mathbf{S}_t) \neq f_\theta(\widehat{\mathbf{X}}_t) \quad \text{when } \mathbf{S}_t \text{ encodes economically relevant state information.}$$

A.1 Equities and Limit Order Books

In equity markets, limit order book data provide a high-dimensional view of liquidity supply and demand across price levels. Classical order-book models emphasize that the book evolves through arrivals, cancellations, and executions of limit and market orders, and that short-horizon price dynamics are closely tied to queue depletion, order-flow imbalance, and the state of available liquidity Cont et al. (2010).

For example, suppose the feature coordinates are partitioned as

$$\{1, \ldots, D\} = \mathcal{J}_{\text{price}} \cup \mathcal{J}_{\text{vol}} \cup \mathcal{J}_{\text{imb}},$$

where $\mathcal{J}_{\text{price}}$ contains bid and ask prices, $\mathcal{J}_{\text{vol}}$ contains bid and ask depths, and $\mathcal{J}_{\text{imb}}$ contains engineered imbalance or pressure features. A liquidity-withdrawal event over a short block $B \subseteq \{1, \ldots, T\}$ can be represented as

$$S_{t,i,j} = -\kappa_j \mathbf{1}\{i \in B,\ j \in \mathcal{J}_{\text{vol}}\}, \qquad \kappa_j > 0.$$

Similarly, an aggressive buy-order event that depletes best-ask liquidity and shifts the ask-side book can be represented by a structured spike over both ask-volume and ask-price coordinates:

$$S_{t,i,j} = A_{t,i,j} \mathbf{1}\{i \in B,\ j \in \mathcal{J}_{\text{ask,vol}} \cup \mathcal{J}_{\text{ask,price}}\}.$$

Such spikes may be highly localized in time but predictive of future mid-price movement.

Thus, for equity LOB prediction, studying spiking models is important for at least three reasons. First, spikes correspond to economically interpretable events such as queue depletion, liquidity replenishment, hidden-liquidity exposure, and aggressive order-flow bursts. Second, these events may be rare but disproportionately important for predicting short-horizon price moves. Third, robustness to spikes is necessary for deployment because models trained on smoothed or benchmark data may fail under real market stress, where the conditional distribution of $\mathbf{S}_t$ changes rapidly.

A.2 Energy and Power Trading

Energy and power markets provide an even more direct motivation for spike-aware modeling. Electricity prices are well known to exhibit mean reversion, seasonality, extreme volatility, and pronounced price spikes. These spikes arise because electricity is difficult to store at scale, demand and supply must balance in real time, and marginal generation costs can change sharply during scarcity, outages, congestion, or renewable forecast errors Borovkova & Permana (2006); Mari & Baldassari (2021). Therefore, in power markets, spike-like behavior is not merely a microstructure artifact; it is a core feature of the underlying economics.

Let P_t denote a spot or short-term forward power price and let $\mathbf{X}_t$ collect market features such as recent prices, load forecasts, renewable generation forecasts, fuel prices, grid congestion indicators, reserve margins, and order-book variables. A stylized spike decomposition may be written as

$$\mathbf{X}_t = \widehat{\mathbf{X}}_t + \mathbf{S}_t,$$

where $\widehat{\mathbf{X}}_t$ captures normal seasonal and mean-reverting dynamics, while $\mathbf{S}_t$ captures scarcity shocks, imbalance shocks, outage events, or extreme weather-driven deviations. In contrast to equity LOBs, where spikes may be concentrated at millisecond or second horizons, power-market spikes may persist over longer blocks:

$$S_{t,i,j} = \kappa_j \mathbf{1}\{i \in B_{\text{scarcity}},\ j \in \mathcal{J}_{\text{load}} \cup \mathcal{J}_{\text{reserve}} \cup \mathcal{J}_{\text{price}}\}.$$

Here B_{scarcity} denotes a time block associated with scarcity or system stress.

This setting highlights an important modeling distinction. In equity LOB data, spike suppression may sometimes be desirable when spikes reflect fleeting quote noise. In electricity markets, however, removing spikes may destroy the most economically important part of the signal.

From a trading perspective, the implications are substantial. Spikes affect day-ahead and intraday bidding, virtual trading, imbalance-market exposure, and the valuation of power derivatives. In such markets, predictive models that are overly smooth may underestimate tail risk and fail during scarcity episodes. By explicitly modeling S_t one can study whether the model learns a stable representation of normal conditions while still responding strongly to rare but economically dominant events.

A.3 Cryptocurrency Markets

Cryptocurrency markets are another natural domain for sparse spiking models. Compared with mature equity markets, crypto markets often exhibit fragmented liquidity, heterogeneous exchange quality, high retail participation, leverage-driven liquidations, and continuous 24/7 trading. These features can generate abrupt price moves, temporary liquidity gaps, and cross-exchange dislocations. Recent empirical work on cryptocurrency market microstructure shows that liquidity and price-discovery variables have predictive power for price dynamics, volatility estimation, dynamic hedging, and electronic market making Easley et al. (2024). Other studies emphasize the role of jumps, cojumps, and microstructure effects in crypto volatility and option pricing Alexander et al. (2023).

Let $\mathbf{X}_t^{(e)}$ denote the LOB or trade-flow input from exchange $e \in \{1, \ldots, E\}$. A multi-exchange representation can be written as

$$\mathbf{X}_t = \left(\mathbf{X}_t^{(1)}, \ldots, \mathbf{X}_t^{(E)}\right), \qquad \mathbf{X}_t^{(e)} = \widehat{\mathbf{X}}_t^{(e)} + \mathbf{S}_t^{(e)}.$$

In this setting, the sparse component may be exchange-specific:

$$\mathbf{S}_t^{(e)} \neq \mathbf{0}, \qquad \mathbf{S}_t^{(e')} = \mathbf{0} \quad \text{for many } e' \neq e,$$

which may indicate a local liquidity disruption, exchange-specific liquidation cascade, or data-quality issue. Alternatively, a systemic crypto-market event may induce synchronized spikes:

$$\mathbf{S}_t^{(1)}, \ldots, \mathbf{S}_t^{(E)} \neq \mathbf{0}$$

over approximately the same time block.

This distinction is important for prediction. A single-exchange spike may be less informative if it reflects a local anomaly, whereas a cross-exchange spike may be highly informative about market-wide order-flow pressure. A spike-aware model can therefore incorporate both local and global masks:

$$\widehat{\mathbf{Z}}_t = \left(\widehat{\mathbf{Z}}_t^{(1)}, \ldots, \widehat{\mathbf{Z}}_t^{(E)}\right),$$

and learn whether the relevant predictive signal is exchange-local or system-wide. For example, if a sudden negative depth spike occurs on one venue but not on others, the model may downweight it. If the same liquidity withdrawal appears across many venues, the model may interpret it as a stronger predictor of future volatility or directional movement.

Crypto markets also make the robustness-calibration tradeoff especially visible. Since jumps may be caused by genuine information arrival, leverage liquidations, oracle failures, stablecoin stress, or exchange outages, a model that is invariant to all spikes may become under-responsive. Conversely, a model that reacts strongly to every local spike may be vulnerable to exchange-specific noise and manipulation.

An explicit sparse decomposition, akin to our work here, provides a precise way to study this tradeoff by separating the regular component $\widehat{\mathbf{X}}_t$ and the spike component $\mathbf{S}_t$.

A.4 Options and Derivatives

Options markets provide a further implication of the sparse spike model because option prices are highly sensitive to discontinuities, tail risk, and volatility jumps in the underlying asset. Classical Black–Scholes-style diffusion models do not fully capture discontinuous price movements. This limitation motivated jump-diffusion models for option pricing, in which the underlying price process

contains both a continuous diffusion component and a jump component Merton (1976). Later stochastic-volatility jump models further showed that jumps and stochastic volatility are important for explaining option-implied distributions and volatility smiles Bates (1996).

Let S_t^{under} denote the underlying asset price and V_t denote a latent variance or volatility state. A generic jump-sensitive dynamics may be written as

$$\frac{dS_t^{\text{under}}}{S_{t^-}^{\text{under}}} = \mu_t\, dt + \sigma_t\, dW_t + J_t\, dN_t,$$

where N_t is a counting process and J_t is a jump size. In an options prediction or market-making setting, the observed feature $\mathbf{X}_t$ may include the underlying LOB, option-chain features, implied volatilities, Greeks, spreads, volumes, and open interest. The sparse spike model then becomes

$$\mathbf{X}_t = \widehat{\mathbf{X}}_t + \mathbf{S}_t,$$

where $\mathbf{S}_t$ may encode jumps in the underlying LOB, implied-volatility surface dislocations, sudden widening of option spreads, or abrupt changes in order flow.

This has direct implications for hedging. Suppose a model predicts a probability vector

$$\widehat{\mathbf{p}}_t = f_\theta(\mathbf{X}_t)$$

or a conditional return distribution

$$\widehat{F}_t = g_\theta(\mathbf{X}_t).$$

If $\mathbf{S}_t$ corresponds to a jump in the underlying or in implied volatility, then the hedging quantities derived from $\widehat{F}_t$, such as delta, gamma, or vega exposure, may change discontinuously. A model that ignores $\mathbf{S}_t$ may underestimate jump risk and produce unstable hedges. Conversely, a spike-aware model can use $\widehat{\mathbf{Z}}_t$ to distinguish between transient quote noise and economically meaningful discontinuities:

$$\widehat{F}_t = g_\theta(\mathbf{X}_t, \widehat{\mathbf{Z}}_t).$$

This formulation is particularly relevant for short-dated options, where small abrupt changes in the underlying or implied volatility can have large effects on option prices and hedging errors.

The sparse spike view also connects naturally to volatility-surface modeling. Let $\sigma_{\text{impl}}(K, \tau; t)$ denote the implied volatility at strike K, maturity τ, and time t. A spike in the option surface can be written as

$$\sigma_{\text{impl}}(K, \tau; t) = \widehat{\sigma}_{\text{impl}}(K, \tau; t) + S_\sigma(K, \tau; t),$$

Such a spike may reflect local illiquidity, stale quotes, or genuine repricing of crash risk. A spike-aware model can therefore improve both prediction and risk management by learning when implied-volatility dislocations should be smoothed away and when they should be interpreted as informative signals about tail risk.

B Detailed Theoretical Results

We start by characterizing the variation of p_t^{base} and p_t^{final} in Total Variation and KL using elementary identities, which can be extended to other divergences using the connections well highlighted here - Connection between Divergences by Frank Nielsen.

Note: We make very standard assumptions in our proofs and thus defer those to the proof statements. For assumptions made beyond the general scope of well known results, we explicitly highlight those in the theorem statements.

B.1 Prediction stability under JSPI tilting

Lemma 1 (Sharp total variation bound for a bounded likelihood-ratio tilt)**.** *Let r be a nonnegative random variable under a probability measure P, and suppose that*

$$\alpha \le r \le \beta \qquad P\text{-a.s.},$$

where $0 \le \alpha < \beta < \infty$. Assume also that $m := \mathbb{E}_P r \in (0, \infty)$, and define the r-tilted probability measure Q by

$$\frac{dQ}{dP} = \frac{r}{\mathbb{E}_P r}.$$

Then

$$d_{\mathrm{TV}}(P,Q) \le \frac{\sqrt{\beta}-\sqrt{\alpha}}{\sqrt{\beta}+\sqrt{\alpha}}.$$

If $\alpha > 0$, the bound is sharp and is attained by a two-point law supported on $\{\alpha, \beta\}$. If $\alpha = 0$, the same bound holds with right-hand side equal to 1, and the value 1 is the supremum in general, approached by two-point laws on $\{0, \beta\}$.

Proof. Let

$$m := \mathbb{E}_P r.$$

Since Q is the r-tilt of P,

$$d_{\mathrm{TV}}(P,Q) = \frac{1}{2}\mathbb{E}_P\left|1-\frac{r}{m}\right| = \frac{1}{2m}\mathbb{E}_P|r-m|.$$

Fix the value of $m \in [\alpha, \beta]$. The function

$$x \mapsto |x-m|$$

is convex on $[\alpha, \beta]$. By the Edmundson–Madansky inequality for the expectation of a convex function of a bounded random variable, among all laws supported on $[\alpha, \beta]$ with mean m, the maximum of $\mathbb{E}|r-m|$ is attained by the two-point law on $\{\alpha, \beta\}$ with the same mean. Equivalently, it suffices to consider

$$P(r=\alpha) = p, \qquad P(r=\beta) = 1-p,$$

where

$$m = p\alpha + (1-p)\beta.$$

For this two-point law,

$$\begin{aligned} d_{\mathrm{TV}}(P,Q) &= \frac{1}{2m}\left[p(m-\alpha) + (1-p)(\beta-m)\right] \\ &= \frac{p(1-p)(\beta-\alpha)}{p\alpha+(1-p)\beta}. \end{aligned}$$

It remains to maximize this expression over $p \in [0,1]$. Define

$$F(p) := \frac{p(1-p)(\beta-\alpha)}{p\alpha+(1-p)\beta}. \tag{1}$$

When $\alpha > 0$, differentiating gives that the unique maximizer is

$$p^* = \frac{\sqrt{\beta}}{\sqrt{\alpha}+\sqrt{\beta}}.$$

Substituting this value into F yields

$$\max_{0\le p\le 1} F(p) = \frac{\sqrt{\beta}-\sqrt{\alpha}}{\sqrt{\beta}+\sqrt{\alpha}}.$$

Hence

$$d_{\mathrm{TV}}(P,Q) \le \frac{\sqrt{\beta}-\sqrt{\alpha}}{\sqrt{\beta}+\sqrt{\alpha}}.$$

The preceding calculation also shows sharpness for $\alpha > 0$, since the two-point law with $p = p^*$ attains equality.

If $\alpha = 0$, the same formula gives the bound $d_{\mathrm{TV}}(P,Q) \le 1$. For the two-point law $P(r=0) = p$ and $P(r=\beta) = 1-p$, one has $m = (1-p)\beta$ and

$$d_{\mathrm{TV}}(P,Q) = p.$$

Letting $p \uparrow 1$ shows that the supremum is 1, although it is not attained by any admissible law with $\mathbb{E}_P r > 0$ as seen by eq 1. □

Theorem 5 (Prediction variation under JSPI exponential tilting)**.** *Let $p_t^{\text{base}} \in \Delta^2$ be the base prediction over $\{-1, 0, +1\}$, and define p_t^{final} by*

$$p_t^{\text{final}}(c) = \frac{p_t^{\text{base}}(c)\exp(a_t c)}{\sum_{c' \in \{-1,0,+1\}} p_t^{\text{base}}(c')\exp(a_t c')}, \qquad a_t = \eta\,\text{JSPI}_t.$$

Then

$$d_{\text{TV}}\left(p_t^{\text{base}}, p_t^{\text{final}}\right) \le \tanh\left(\frac{|a_t|}{2}\right), \text{ie}, \qquad \left\|p_t^{\text{base}} - p_t^{\text{final}}\right\|_1 \le 2\tanh\left(\frac{|a_t|}{2}\right) \le |a_t|$$

Moreover, the two KL divergences satisfy

$$D_{\text{KL}}\left(p_t^{\text{base}} \,\|\, p_t^{\text{final}}\right) \le \frac{a_t^2}{2}, \qquad D_{\text{KL}}\left(p_t^{\text{final}} \,\|\, p_t^{\text{base}}\right) \le \frac{a_t^2}{2}$$

Resulting in

$$D_{\text{KL}}\left(p_t^{\text{base}} \,\|\, p_t^{\text{final}}\right) + D_{\text{KL}}\left(p_t^{\text{final}} \,\|\, p_t^{\text{base}}\right) \le a_t^2.$$

Proof. Let the random variable of the predictions be $Y \in \{-1, 0, +1\}$ and distributed according to p_t^{base}, We write its MGF as

$$Z_t(a) = \mathbb{E}_{p_t^{\text{base}}}\left[e^{aY}\right].$$

The JSPI calibrated distribution is the exponentially tilted law

$$\frac{p_t^{\text{final}}}{p_t^{\text{base}}}(c) = \frac{e^{a_t c}}{Z_t(a_t)}.$$

For the total-variation bound, set $r(c) = e^{a_t c}$. Since $c \in \{-1, 0, +1\}$, we have

$$r(c) \in \left[e^{-|a_t|}, e^{|a_t|}\right].$$

Invoking lemma 1, where we utilize the Edmundson–Madansky upper bound for expectations of convex functions since L_1 norm is convex, we have that

Taking $\alpha = e^{-|a_t|}$ and $\beta = e^{|a_t|}$ yields the desired

$$d_{\text{TV}}\left(p_t^{\text{base}}, p_t^{\text{final}}\right) \le \tanh\left(\frac{|a_t|}{2}\right).$$

For the KL bounds: Let

$$F_t(a) := \log Z_t(a).$$

Since $Y \in \{-1, 0, +1\}$, the moment generating function $Z_t(a)$ is finite for every $a \in \mathbb{R}$. Moreover $Z_t(0) = 1$, and therefore

$$F_t(0) = 0.$$

By differentiating the log-partition function,

$$F_t'(a) = \frac{Z_t'(a)}{Z_t(a)} = \frac{\mathbb{E}_{p_t^{\text{base}}}\left[Ye^{aY}\right]}{\mathbb{E}_{p_t^{\text{base}}}\left[e^{aY}\right]} = \mathbb{E}_{p_t^{\text{final}}}[Y].$$

Differentiating once more gives

$$F_t''(a) = \text{Var}_{p_{t,a}}(Y).$$

Since $Y \in [-1, 1]$, we have

$$\text{Var}_{p_{t,a}}(Y) \le 1.$$

Thus

$$0 \le F_t''(a) \le 1 \qquad \text{for all } a \in \mathbb{R}.$$

Now,

$$\begin{aligned} D_{\mathrm{KL}}\left(p_t^{\mathrm{base}} \,\middle\|\, p_t^{\mathrm{final}}\right) &= \mathbb{E}_{p_t^{\mathrm{base}}}\left[\log \frac{p_t^{\mathrm{base}}(Y)}{p_t^{\mathrm{final}}(Y)}\right] \\ &= \mathbb{E}_{p_t^{\mathrm{base}}}\left[-a_t Y + F_t(a_t)\right] \\ &= F_t(a_t) - a_t F_t'(0). \end{aligned}$$

Since $F_t(0) = 0$, Taylor's theorem with integral remainder gives

$$F_t(a_t) - F_t(0) - a_t F_t'(0) = \int_0^{a_t} (a_t - s) F_t''(s)\, ds.$$

Equivalently,

$$F_t(a_t) - a_t F_t'(0) = \int_{\min\{0,a_t\}}^{\max\{0,a_t\}} \big(|a_t| - u\big) F_t''(\mathrm{sgn}(a_t)u)\, du,$$

where the case $a_t = 0$ is trivial. Using $F_t''(s) \le 1$, we obtain

$$D_{\mathrm{KL}}\left(p_t^{\mathrm{base}} \,\middle\|\, p_t^{\mathrm{final}}\right) \le \int_0^{|a_t|} (|a_t| - u)\, du = \frac{a_t^2}{2}.$$

We next bound the reverse KL divergence:

$$\begin{aligned} D_{\mathrm{KL}}\left(p_t^{\mathrm{final}} \,\middle\|\, p_t^{\mathrm{base}}\right) &= \mathbb{E}_{p_t^{\mathrm{final}}}\left[\log \frac{p_t^{\mathrm{final}}(Y)}{p_t^{\mathrm{base}}(Y)}\right] \\ &= \mathbb{E}_{p_{t,a_t}}\left[a_t Y - F_t(a_t)\right] \\ &= a_t F_t'(a_t) - F_t(a_t). \end{aligned}$$

Since $F_t(0) = 0$,

$$\begin{aligned} a_t F_t'(a_t) - F_t(a_t) &= a_t F_t'(a_t) - F_t(a_t) + F_t(0) \\ &= \int_0^{a_t} s F_t''(s)\, ds. \end{aligned}$$

Equivalently,

$$a_t F_t'(a_t) - F_t(a_t) = \int_0^{|a_t|} u F_t''(\mathrm{sgn}(a_t)u)\, du.$$

Using $F_t''(s) \le 1$, we get

$$D_{\mathrm{KL}}\left(p_t^{\mathrm{final}} \,\middle\|\, p_t^{\mathrm{base}}\right) \le \int_0^{|a_t|} u\, du = \frac{a_t^2}{2}.$$

Combining the two we have the desired result. □

Corollary 3 (Margin Conditions for JSPI calibration)**.** *Let* $c_t^\star = \arg\max_c p_t^{\mathrm{base}}(c)$

If the base margin satisfies

$$\gamma_t = p_t^{\mathrm{base}}(c_t^\star) - \max_{c \neq c_t^\star} p_t^{\mathrm{base}}(c) > 2\tanh\left(\frac{\eta|\mathrm{JSPI}_t|}{2}\right).$$

Then

$$\arg\max_c p_t^{\mathrm{final}}(c) = \arg\max_c p_t^{\mathrm{base}}(c).$$

implying the calibration does not affect the outputs if the model is extremely confident. This is in line with the scope of this work since the assumptions of JSPI *downstream impact are controlled by the trained model.*

On the contrary*, when the margin* γ_t *floats around the vicinity of the decision boundaries, the calibration can indeed flip the predictions.*

Note *again that the prediction that the follows the class label flip being correct then depends on the ability of* JSPI_t *to predict the price behavior.*

B.2 JSPI Consistency Results

We now provide a consistency result for the JSPI estimator. The argument is based on three ingredients: local empirical-quantile concentration, robustness of the rolling median and MAD to sparse contamination where we assume a fraction $\pi_n \to 0$ of the timesteps of the discretized SDE are jump contaminated, and a jump-separation condition ensuring consistent support recovery. Rewriting the SDE formulation and the assumptions over the discretization for theoretical treatment:

Continuous-time motivation for the sparse jump model. For the price channels $j \in \mathcal{J}_P$, we view the observed LOB feature as a noisy discretization of a latent efficient order-price process. Let

$$(\Omega, \mathcal{F}, (\mathcal{F}_u)_{u\geq 0}, \mathbb{P})$$

be a filtered probability space. For each price channel $j \in \mathcal{J}_P$, let $Y^j = (Y^j_u)_{u\geq 0}$ denote the latent regular order-price component. We model Y^j as a jump-diffusion semimartingale

$$dY^j_u = b_j(u)\,du + \sigma_j(u)\,dW^j_u + dJ^j_u,$$

where $b_j(u)$ is a local drift, $\sigma_j(u)$ is a local volatility coefficient, W^j is a Brownian motion, and J^j is a finite-activity jump process of the form

$$J^j_u = \int_{(0,u]\times\mathcal{E}} \gamma_j(s,z)\,N_j(ds,dz).$$

Here N_j is a Poisson random measure on $\mathbb{R}_+ \times \mathcal{E}$, and $\gamma_j(s,z)$ denotes the jump size associated with mark z. The observed order-book price feature is

$$\widetilde{X}^j_u = Y^j_u + \zeta^j_u,$$

where ζ^j_u denotes microstructure noise. For bid and ask quote channels, ζ^j_u may be side-dependent or one-sided; in the local analysis below, its conditional median is absorbed into the regular component so that the remaining noise has median zero.

Let

$$u_i = u_0 + i\Delta_n, \qquad i = 1,\dots,T_n,$$

be the sampling grid. The jump innovation over the i-th sampling interval is

$$G^\star_{t,i,j} = \int_{(u_{i-1},u_i]\times\mathcal{E}} \gamma_j(s,z)\,N_j(ds,dz).$$

The **discretized local observation** may therefore be written as

$$X_{t,i,j} = \mu_j(u_i) + G^\star_{t,i,j} + \xi_{t,i,j}, \qquad i = 1,\dots,T_n,$$

where

$$\mu_j(u_i) = \int_{u_{i-1}}^{u_i} b_j(s)\,ds$$

is the local regular component, while

$$\xi_{t,i,j} = \int_{u_{i-1}}^{u_i} \sigma_j(s)\,dW^j_s + \Delta_i\zeta^j$$

collects the continuous martingale fluctuation and microstructure noise, after **local median centering** (used in the consistency proof). The oracle jump mask is then

$$Z^\star_{t,i,j} = \mathbf{1}\left\{G^\star_{t,i,j} \neq 0\right\}.$$

Thus, the sparse decomposition

$$X_{t,i,j} = \mu_j(u_i) + G^\star_{t,i,j} + \xi_{t,i,j}$$

used in the theoretical analysis is the discrete sampling-bin realization of the continuous-time jump-diffusion model.

In the implemented JSPI estimator, the unknown regular component $\mu_j(u_i)$ is not estimated parametrically. Instead, it is replaced by the rolling robust baseline $B_{t,i,j}$, and jump coordinates are identified through large standardized deviations

$$\frac{|X_{t,i,j} - B_{t,i,j}|}{\widehat{\sigma}_{t,i,j} + \varepsilon}.$$

Note: To avoid a notational ambiguity, we write $G^\star_{t,i,j}$ for the latent structural jump in the observation equation, and reserve $S^\star_{t,i,j}$ for the oracle sparse residual used by the JSPI functional.

Notations and Assumptions: Let

$$\mathcal{J}_P = \mathcal{J}_{\text{bidPrice}} \cup \mathcal{J}_{\text{askPrice}}$$

denote the price dimensions, accounting for the respective levels in the order book, entering the JSPI aggregate.

It is non-trivial to provide high probability bounds for the non-asymptotic cases, thus we operate in a high-frequency sampling regime for theoretical purposes, where the asymptotics follow. Thus, we assume equi-spaced sampling

$$u_i = i\Delta_n, \qquad \Delta_n \to 0,$$

and let the rolling-window radius satisfy

$$r_n \to \infty, \qquad m_n = 2r_n + 1, \qquad h_n = m_n\Delta_n \to 0.$$

Remark 2 (Radius r_n)**.** *The assumption of $r_n \to \infty$ is indeed strong, however it proves to be sufficient for the quantile regularity and concentration used below. Relaxation (partial) of this assumption is our ongoing research effort.*

For each index i, define the local neighborhood

$$\mathcal{N}_n(i) = \{k : |k - i| \le r_n,\ 1 \le k \le T_n\}.$$

For simplicity, the proof is stated for interior indices

$$\mathcal{I}_n = \{r_n + 1, \dots, T_n - r_n\},$$

so that $|\mathcal{N}_n(i)| = m_n$. Boundary indices may be handled by one-sided windows or padding, which only changes constants.

Assumption 6 (Local smoothness of the regular component)**.** *There exist constants $L < \infty$ and $\alpha \in (0, 1]$ such that, uniformly over $j \in \mathcal{J}_P$,*

$$|\mu_j(u) - \mu_j(v)| \le L|u - v|^\alpha.$$

Thus, for $k \in \mathcal{N}_n(i)$,

$$|\mu_j(u_k) - \mu_j(u_i)| \le Lh_n^\alpha.$$

Assumption 7 (Local quantile regularity and empirical concentration)**.** *For each channel j, the noise distribution is locally stationary within windows of width h_n, has median zero, and has density bounded away from zero in a neighborhood of its median. Needless to say that this assumption is realizable since we are disambiguating the regular component from the jump's observed in the process. More precisely, there exist constants $f_{\min} > 0$, $f_{\max} < \infty$, and $\delta_0 > 0$ such that, for the local noise cumulative distribution $F_{u,j}$,*

$$F_{u,j}(0) = \frac{1}{2}, \qquad f_{u,j}(x) \ge f_{\min} \quad \textit{for all } |x| \le \delta_0.$$

The same type of regularity holds for the absolute-deviation distribution

$$|\xi_{t,i,j}|,$$

at its median. In addition, the local empirical distribution functions satisfy the uniform concentration bound

$$\max_{i \in \mathcal{I}_n} \max_{j \in \mathcal{J}_P} \sup_x \left| \widehat{F}_{i,j}(x) - F_{i,j}(x) \right| = O_p\left(\sqrt{\frac{\log(T_n D_n)}{m_n}} \right).$$

By Lemma 2, the empirical concentration condition holds for independent local observations whenever $|\mathcal{I}_n||\mathcal{J}_P| \lesssim T_n D_n$ and $\log(T_n D_n) = o(m_n)$.

Assumption 8 (Sparse local jump contamination)**.** *Let*

$$\pi_n = \max_{i \in \mathcal{I}_n} \max_{j \in \mathcal{J}_P} \frac{1}{m_n} \sum_{k \in \mathcal{N}_n(i)} Z^\star_{t,k,j}.$$

The core premise of this work is the assumption that

$$\pi_n \to 0.$$

In particular, if the number of jumps in each compact local window is It is the condition used in this work for consistency of the local median around the uncontaminated component.

Assumption 9 (Residual boundedness on the local window)**.** *There exists $C_X < \infty$ such that, with probability tending to one,*

$$\max_{i\in\mathcal{I}_n}\max_{j\in\mathcal{J}_P}|X_{t,i,j} - B^\star_{t,i,j}| \le C_X,$$

where as discussed above,

$$B^\star_{t,i,j} = \mu_j(u_i).$$

This condition may be weakened by allowing $C_X = C_{X,n} := g(n)$ to grow slowly with n, in which case the final rates in the theorem acquire the corresponding multiplicative factor.

Assumption 10 (Jump-separation condition)**.** *Let*

$$\sigma^\star_{t,i,j} = 1.4826 \cdot \text{median}\left(|\xi_{t,i,j}|\right)$$

denote the local population MAD scale, and define the ***oracle standardized residual***

$$A^\star_{t,i,j} = \frac{|X_{t,i,j} - B^\star_{t,i,j}|}{\sigma^\star_{t,i,j} + \varepsilon}.$$

Let

$$e_n = h_n^\alpha + \pi_n + \sqrt{\frac{\log(T_n D_n)}{m_n}}.$$

There exists a constant $M > 0$ such that, with probability tending to one,

$$\min_{(i,j):Z^\star_{t,i,j}=1} A^\star_{t,i,j} > \tau_n + Me_n,$$

and

$$\max_{(i,j):Z^\star_{t,i,j}=0} A^\star_{t,i,j} < \tau_n - Me_n.$$

This condition states that true jump and non-jump coordinates are separated by a margin larger than the local estimation error.

B.2.1 Supporting Lemmas

Lemma 2 (Uniform local empirical-CDF concentration)**.** *Let $\mathcal{I}_n$ denote the collection of local windows and let $\mathcal{J}_P$ denote the set of price channels. For each $(i,j) \in \mathcal{I}_n \times \mathcal{J}_P$, let*

$$Y_1^{(i,j)}, \dots, Y_{m_n}^{(i,j)}$$

be independent observations with common distribution function $F_{i,j}$, and define the corresponding empirical distribution function

$$\widehat{F}_{i,j}(x) = \frac{1}{m_n}\sum_{\ell=1}^{m_n}\mathbf{1}\{Y_\ell^{(i,j)} \le x\}.$$

Then, for every $\varepsilon > 0$,

$$\mathbb{P}\left(\max_{i\in\mathcal{I}_n}\max_{j\in\mathcal{J}_P}\sup_{x\in\mathbb{R}}\left|\widehat{F}_{i,j}(x) - F_{i,j}(x)\right| > \varepsilon\right) \le 2|\mathcal{I}_n||\mathcal{J}_P|\exp(-2m_n\varepsilon^2).$$

Consequently, if

$$\log\big(|\mathcal{I}_n||\mathcal{J}_P|\big) = o(m_n),$$

then

$$\max_{i\in\mathcal{I}_n}\max_{j\in\mathcal{J}_P}\sup_{x\in\mathbb{R}}\left|\widehat{F}_{i,j}(x) - F_{i,j}(x)\right| = O_p\left(\sqrt{\frac{\log(|\mathcal{I}_n||\mathcal{J}_P|)}{m_n}}\right).$$

In particular, if

$$|\mathcal{I}_n||\mathcal{J}_P| \lesssim T_n D_n,$$

then

$$\max_{i\in\mathcal{I}_n}\max_{j\in\mathcal{J}_P}\sup_{x\in\mathbb{R}}\left|\widehat{F}_{i,j}(x) - F_{i,j}(x)\right| = O_p\left(\sqrt{\frac{\log(T_n D_n)}{m_n}}\right).$$

Proof. Fix a pair $(i,j) \in \mathcal{I}_n \times \mathcal{J}_P$. By the Dvoretzky–Kiefer–Wolfowitz–Massart inequality Massart (1990), for every $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} \left|\widehat{F}_{i,j}(x) - F_{i,j}(x)\right| > \varepsilon\right) \le 2\exp(-2m_n\varepsilon^2).$$

Applying the union bound over all $(i,j) \in \mathcal{I}_n \times \mathcal{J}_P$ yields

$$\begin{aligned}
&\mathbb{P}\left(\max_{i\in\mathcal{I}_n}\max_{j\in\mathcal{J}_P}\sup_{x\in\mathbb{R}} \left|\widehat{F}_{i,j}(x) - F_{i,j}(x)\right| > \varepsilon\right) \\
&\quad\le \sum_{i\in\mathcal{I}_n}\sum_{j\in\mathcal{J}_P} \mathbb{P}\left(\sup_{x\in\mathbb{R}} \left|\widehat{F}_{i,j}(x) - F_{i,j}(x)\right| > \varepsilon\right) \\
&\quad\le 2|\mathcal{I}_n||\mathcal{J}_P|\exp(-2m_n\varepsilon^2).
\end{aligned}$$

Now choose

$$\varepsilon_n(M) = \sqrt{\frac{\log\bigl(2|\mathcal{I}_n||\mathcal{J}_P|\bigr) + M}{2m_n}},$$

where $M > 0$. Substituting this value into the preceding bound gives

$$\mathbb{P}\left(\max_{i\in\mathcal{I}_n}\max_{j\in\mathcal{J}_P}\sup_{x\in\mathbb{R}} \left|\widehat{F}_{i,j}(x) - F_{i,j}(x)\right| > \varepsilon_n(M)\right) \le e^{-M}.$$

Since M can be chosen arbitrarily large, this implies

$$\max_{i\in\mathcal{I}_n}\max_{j\in\mathcal{J}_P}\sup_{x\in\mathbb{R}} \left|\widehat{F}_{i,j}(x) - F_{i,j}(x)\right| = O_p\left(\sqrt{\frac{\log(|\mathcal{I}_n||\mathcal{J}_P|)}{m_n}}\right).$$

Finally, if $|\mathcal{I}_n||\mathcal{J}_P| \lesssim T_n D_n$, then

$$\log(|\mathcal{I}_n||\mathcal{J}_P|) \lesssim \log(T_n D_n),$$

up to an additive constant, and the claimed $O_p(\sqrt{\log(T_n D_n)/m_n})$ bound follows. □

Lemma 3 (Sup-norm stability of the sample median)**.** *Let $a = (a_1, \dots, a_m) \in \mathbb{R}^m$ and $b = (b_1, \dots, b_m) \in \mathbb{R}^m$. Suppose that*

$$\|a - b\|_\infty = \max_{1\le\ell\le m} |a_\ell - b_\ell| \le \delta.$$

Let $\operatorname{med}(a)$ and $\operatorname{med}(b)$ denote the sample medians of the two samples, using the same median convention for both samples. Then

$$|\operatorname{med}(a) - \operatorname{med}(b)| \le \delta.$$

Equivalently, the sample median map

$$a \mapsto \operatorname{med}(a)$$

is 1-Lipschitz as a map from $(\mathbb{R}^m, \|\cdot\|_\infty)$ to $\mathbb{R}$.

Proof. Let

$$a_{(1)} \le a_{(2)} \le \cdots \le a_{(m)} \qquad \text{and} \qquad b_{(1)} \le b_{(2)} \le \cdots \le b_{(m)}$$

denote the order statistics of a and b, respectively.

Since $\|a - b\|_\infty \le \delta$, we have, for every ℓ,

$$b_\ell - \delta \le a_\ell \le b_\ell + \delta.$$

We first show that each corresponding order statistic differs by at most δ. Fix $r \in \{1, \dots, m\}$. Because at least r coordinates of b are no larger than $b_{(r)}$, the corresponding coordinates of a are no larger than $b_{(r)} + \delta$. Hence the r-th order statistic of a satisfies

$$a_{(r)} \le b_{(r)} + \delta.$$

Similarly, because at least r coordinates of a are no larger than $a_{(r)}$, the corresponding coordinates of b are no larger than $a_{(r)} + \delta$, and therefore

$$b_{(r)} \le a_{(r)} + \delta.$$

Combining the two inequalities gives

$$|a_{(r)} - b_{(r)}| \le \delta \qquad \text{for every } r = 1, \dots, m.$$

If m is odd, say $m = 2q + 1$, then

$$\operatorname{med}(a) = a_{(q+1)}, \qquad \operatorname{med}(b) = b_{(q+1)}.$$

Therefore

$$|\operatorname{med}(a) - \operatorname{med}(b)| = |a_{(q+1)} - b_{(q+1)}| \le \delta.$$

If m is even, say $m = 2q$, and the median is defined as the average of the two central order statistics, then

$$\operatorname{med}(a) = \frac{a_{(q)} + a_{(q+1)}}{2}, \qquad \operatorname{med}(b) = \frac{b_{(q)} + b_{(q+1)}}{2}.$$

Thus

$$\begin{aligned} |\operatorname{med}(a) - \operatorname{med}(b)| &= \left| \frac{a_{(q)} - b_{(q)}}{2} + \frac{a_{(q+1)} - b_{(q+1)}}{2} \right| \\ &\le \frac{1}{2}|a_{(q)} - b_{(q)}| + \frac{1}{2}|a_{(q+1)} - b_{(q+1)}| \\ &\le \delta. \end{aligned}$$

The same argument also applies if one uses the lower or upper median convention for even m, since those are individual order statistics. Hence the sample median is 1-Lipschitz with respect to the sup norm. □

B.2.2 Main Consistency Results

Theorem 6 (Rolling-window JSPI consistency)**.** *Suppose Assumptions 6-10 hold. Let*

$$B_{t,i,j} = \operatorname{median}\{X_{t,k,j} : k \in \mathcal{N}_n(i)\}$$

be the rolling median baseline, and let

$$\widehat{\sigma}_{t,i,j} = 1.4826 \cdot \operatorname{median}\{|X_{t,k,j} - B_{t,k,j}| : k \in \mathcal{N}_n(i)\}$$

be the rolling MAD scale estimator. Define

$$\widehat{A}_{t,i,j} = \frac{|X_{t,i,j} - B_{t,i,j}|}{\widehat{\sigma}_{t,i,j} + \varepsilon},$$

and

$$\widehat{Z}_{t,i,j} = \mathbf{1}\left\{\widehat{A}_{t,i,j} > \tau_n\right\}.$$

Then

$$\max_{i \in \mathcal{I}_n} \max_{j \in \mathcal{J}_P} |B_{t,i,j} - B^\star_{t,i,j}| = O_p(e_n),$$

and

$$\max_{i \in \mathcal{I}_n} \max_{j \in \mathcal{J}_P} |\widehat{\sigma}_{t,i,j} - \sigma^\star_{t,i,j}| = O_p(e_n).$$

Consequently,

$$\mathbb{P}\left(\widehat{Z}_{t,i,j} = Z^\star_{t,i,j} \textit{ for all } i \in \mathcal{I}_n,\ j \in \mathcal{J}_P\right) \to 1.$$

Define the estimated sparse residual

$$\widehat{S}_t = \widehat{Z}_t \odot (X_t - B_t),$$

and the oracle sparse residual

$$S^\star_t = Z^\star_t \odot (X_t - B^\star_t).$$

Let

$$\widehat{\mathrm{JSPI}}_t = \sum_{i\in\mathcal{I}_n}\sum_{j\in\mathcal{J}_{\mathrm{bidPrice}}} \widehat{S}_{t,i,j} + \sum_{i\in\mathcal{I}_n}\sum_{j\in\mathcal{J}_{\mathrm{askPrice}}} \widehat{S}_{t,i,j},$$

and

$$\mathrm{JSPI}_t^\star = \sum_{i\in\mathcal{I}_n}\sum_{j\in\mathcal{J}_{\mathrm{bidPrice}}} S^\star_{t,i,j} + \sum_{i\in\mathcal{I}_n}\sum_{j\in\mathcal{J}_{\mathrm{askPrice}}} S^\star_{t,i,j}.$$

Defining

$$s_t = \left|\{(i,j) : i\in\mathcal{I}_n,\, j\in\mathcal{J}_P,\, Z^\star_{t,i,j}=1\}\right|,$$

then

$$\left|\widehat{\mathrm{JSPI}}_t - \mathrm{JSPI}_t^\star\right| = O_p(s_t e_n).$$

In particular, if $s_t e_n \to 0$, then

$$\widehat{\mathrm{JSPI}}_t - \mathrm{JSPI}_t^\star \xrightarrow{p} 0.$$

If the number of oracle price-jump coordinates is fixed, then $s_t = O(1)$, and

$$\left|\widehat{\mathrm{JSPI}}_t - \mathrm{JSPI}_t^\star\right| = O_p(e_n).$$

Proof. We prove the result in four steps.

Step 1: Uniform consistency of the rolling median baseline. Fix an interior coordinate $i\in\mathcal{I}_n$ and a price channel $j\in\mathcal{J}_P$. Consider first the hypothetical clean local sample obtained by removing the jump component:

$$X^0_{t,k,j} = \mu_j(u_k) + \xi_{t,k,j}, \qquad k\in\mathcal{N}_n(i),$$

Since $k\in\mathcal{N}_n(i)$, Assumption 6 gives

$$|\mu_j(u_k) - \mu_j(u_i)| \le Lh_n^\alpha.$$

Thus each clean local observation differs from

$$\mu_j(u_i) + \xi_{t,k,j}$$

by at most Lh_n^α.

By Lemma 3, we have that the sample median is 1-Lipschitz with respect to sup-norm perturbations of the sample, combining with Assumption 6 we have that

$$\left|\mathrm{median}\{X^0_{t,k,j} : k\in\mathcal{N}_n(i)\} - \mathrm{median}\{\mu_j(u_i)+\xi_{t,k,j} : k\in\mathcal{N}_n(i)\}\right| \le Lh_n^\alpha. \tag{2}$$

The population median of $\mu_j(u_i)+\xi_{t,k,j}$ is

$$B^\star_{t,i,j} = \mu_j(u_i),$$

since the noise has median zero. By the local empirical-process bound in Assumption 7 and quantile inversion under a density bounded below at the median,

$$\left|\mathrm{median}\{\mu_j(u_i)+\xi_{t,k,j} : k\in\mathcal{N}_n(i)\} - B^\star_{t,i,j}\right| = O_p\left(\sqrt{\frac{\log(T_nD_n)}{m_n}}\right), \tag{3}$$

uniformly over $i\in\mathcal{I}_n$ and $j\in\mathcal{J}_P$.

Now we account for **jump contamination via π_n:** In the local window $\mathcal{N}_n(i)$, at most a fraction π_n of observations are jump-contaminated. Replacing a fraction π_n of observations can shift the empirical distribution function by at most $O(\pi_n)$ in supremum norm. Since the population density around the median is bounded below by $f_{\min}$, another quantile-inversion argument implies that this contamination shifts the local median by at most $O(\pi_n)$.
Combining the - (i) stochastic quantile error (eq 3); (ii) the local Hölder bias (eq 2) and (iii) the sparse-jump contamination (above) via Union Bound yields

$$\max_{i\in\mathcal{I}_n}\max_{j\in\mathcal{J}_P} |B_{t,i,j} - B^\star_{t,i,j}| = O_p\left(h_n^\alpha + \pi_n + \sqrt{\frac{\log(T_nD_n)}{m_n}}\right) = O_p(e_n). \tag{4}$$

Step 2: Uniform consistency of the rolling MAD scale. Define the oracle local absolute deviation

$$D^\star_{t,k,j} = |X^0_{t,k,j} - B^\star_{t,k,j}| = |\xi_{t,k,j}|.$$

By Assumption 7, the distribution of $D^\star_{t,k,j}$ has density bounded away from zero at its median. Therefore, applying the same local quantile argument as in Step 1 gives

$$\max_{i\in\mathcal{I}_n}\max_{j\in\mathcal{J}_P}\left|1.4826\cdot\mathrm{median}\{D^\star_{t,k,j} : k\in\mathcal{N}_n(i)\} - \sigma^\star_{t,i,j}\right| = O_p(e_n).$$

The actual rolling MAD uses

$$|X_{t,k,j} - B_{t,k,j}|$$

instead of

$$|X^0_{t,k,j} - B^\star_{t,k,j}|.$$

For every i, j,

$$\big||X_{t,i,j} - B_{t,i,j}| - |X^0_{t,i,j} - B^\star_{t,i,j}|\big| \le |G^\star_{t,i,j}| + |B_{t,i,j} - B^\star_{t,i,j}|.$$

The sparse contamination term affects at most a fraction π_n of the local window, and hence contributes $O(\pi_n)$ to the local median by the same robustness argument as above. The baseline-estimation term contributes

$$\max_{i,j}|B_{t,i,j} - B^\star_{t,i,j}| = O_p(e_n),$$

as obtained in eq 4. Combining we obtain,

$$\max_{i\in\mathcal{I}_n}\max_{j\in\mathcal{J}_P}|\widehat{\sigma}_{t,i,j} - \sigma^\star_{t,i,j}| = O_p(e_n).$$

Step 3: Consistent jump-mask recovery. We next compare the estimated standardized residual

$$\widehat{A}_{t,i,j} = \frac{|X_{t,i,j} - B_{t,i,j}|}{\widehat{\sigma}_{t,i,j} + \varepsilon}$$

with its oracle counterpart

$$A^\star_{t,i,j} = \frac{|X_{t,i,j} - B^\star_{t,i,j}|}{\sigma^\star_{t,i,j} + \varepsilon}.$$

Let

$$\widehat{R}_{t,i,j} = |X_{t,i,j} - B_{t,i,j}|, \qquad R^\star_{t,i,j} = |X_{t,i,j} - B^\star_{t,i,j}|.$$

Then

$$\widehat{A}_{t,i,j} = \frac{\widehat{R}_{t,i,j}}{\widehat{\sigma}_{t,i,j} + \varepsilon}, \qquad A^\star_{t,i,j} = \frac{R^\star_{t,i,j}}{\sigma^\star_{t,i,j} + \varepsilon}.$$

Using the reverse triangle inequality,

$$|\widehat{R}_{t,i,j} - R^\star_{t,i,j}| = \big||X_{t,i,j} - B_{t,i,j}| - |X_{t,i,j} - B^\star_{t,i,j}|\big| \le |B_{t,i,j} - B^\star_{t,i,j}|.$$

Moreover,

$$\begin{aligned}
|\widehat{A}_{t,i,j} - A^\star_{t,i,j}| &= \left|\frac{\widehat{R}_{t,i,j}}{\widehat{\sigma}_{t,i,j} + \varepsilon} - \frac{R^\star_{t,i,j}}{\sigma^\star_{t,i,j} + \varepsilon}\right| \\
&\le \frac{|\widehat{R}_{t,i,j} - R^\star_{t,i,j}|}{\widehat{\sigma}_{t,i,j} + \varepsilon} + R^\star_{t,i,j}\left|\frac{1}{\widehat{\sigma}_{t,i,j} + \varepsilon} - \frac{1}{\sigma^\star_{t,i,j} + \varepsilon}\right| \\
&\le \frac{|B_{t,i,j} - B^\star_{t,i,j}|}{\varepsilon} + \frac{R^\star_{t,i,j}}{\varepsilon^2}|\widehat{\sigma}_{t,i,j} - \sigma^\star_{t,i,j}|.
\end{aligned}$$

Taking maxima over $i \in \mathcal{I}_n$ and $j \in \mathcal{J}_P$, and using the residual envelope

$$\max_{i,j} R^\star_{t,i,j} \le C_X$$

on the high-probability event from Assumption 9, gives

$$\max_{i,j} |\widehat{A}_{t,i,j} - A^\star_{t,i,j}| \le \frac{1}{\varepsilon} \max_{i,j} |B_{t,i,j} - B^\star_{t,i,j}| + \frac{C_X}{\varepsilon^2} \max_{i,j} |\widehat{\sigma}_{t,i,j} - \sigma^\star_{t,i,j}|.$$

By Step 1 and Step 2,

$$\max_{i,j} |B_{t,i,j} - B^\star_{t,i,j}| = O_p(e_n), \qquad \max_{i,j} |\widehat{\sigma}_{t,i,j} - \sigma^\star_{t,i,j}| = O_p(e_n).$$

Therefore,

$$\max_{i \in \mathcal{I}_n} \max_{j \in \mathcal{J}_P} |\widehat{A}_{t,i,j} - A^\star_{t,i,j}| = O_p(e_n).$$

Equivalently, there exists a constant $C_A < \infty$ such that, with probability tending to one,

$$\max_{i \in \mathcal{I}_n} \max_{j \in \mathcal{J}_P} |\widehat{A}_{t,i,j} - A^\star_{t,i,j}| \le C_A e_n.$$

If we then choose the margin constant M in Assumption 10 such that $M > C_A$. Then, on the event above,
every **true jump coordinate** satisfies

$$\widehat{A}_{t,i,j} \ge A^\star_{t,i,j} - C_A e_n > \tau_n + (M - C_A) e_n > \tau_n,$$

and every **non-jump coordinate** satisfies

$$\widehat{A}_{t,i,j} \le A^\star_{t,i,j} + C_A e_n < \tau_n - (M - C_A) e_n < \tau_n.$$

Therefore,

$$\widehat{Z}_{t,i,j} = Z^\star_{t,i,j}$$

for all $i \in \mathcal{I}_n$ and $j \in \mathcal{J}_P$, with probability tending to one. Hence

$$\mathbb{P}\left(\widehat{Z}_t = Z^\star_t \text{ on } \mathcal{I}_n \times \mathcal{J}_P\right) \to 1.$$

Step 4: JSPI consistency. On the event

$$\widehat{Z}_t = Z^\star_t \quad \text{on } \mathcal{I}_n \times \mathcal{J}_P,$$

we have

$$\widehat{S}_t = Z^\star_t \odot (X_t - B_t),$$

while the oracle sparse residual is

$$S^\star_t = Z^\star_t \odot (X_t - B^\star_t).$$

Therefore,

$$\widehat{S}_t - S^\star_t = Z^\star_t \odot (B^\star_t - B_t).$$

Taking the signed JSPI aggregate over the selected price channels gives

$$\begin{aligned} \left|\widehat{\mathrm{JSPI}}_t - \mathrm{JSPI}^\star_t\right| &\le \sum_{\substack{i \in \mathcal{I}_n \\ j \in \mathcal{J}_P}} Z^\star_{t,i,j} |B_{t,i,j} - B^\star_{t,i,j}| \\ &\le s_t \max_{i \in \mathcal{I}_n} \max_{j \in \mathcal{J}_P} |B_{t,i,j} - B^\star_{t,i,j}| \\ &= O_p(s_t e_n). \end{aligned}$$

If $s_t e_n \to 0$, this implies

$$\widehat{\mathrm{JSPI}}_t - \mathrm{JSPI}^\star_t \xrightarrow{p} 0.$$

This completes the proof. □

Remark 3 (Simplified rate under finite local jump support)**.** *If the number of jump-contaminated coordinates in each local window is uniformly bounded, then*

$$\pi_n = O(m_n^{-1}).$$

Since $m_n^{-1} \le m_n^{-1/2}$, this term is absorbed by the empirical-quantile rate, and the error rate may be written as

$$e_n = h_n^{\alpha} + \sqrt{\frac{\log(T_n D_n)}{m_n}}.$$

This recovers the simpler rate used in the main theorem statement.

Remark 4 (Residual oracle versus structural jump oracle)**.** *The consistency result above compares $\widehat{\mathrm{JSPI}}_t$ with a residual oracle rather than with a pure structural jump oracle. high-frequency order-book data.*

Under the local observation model

$$X_{t,i,j} = \mu_j(u_i) + G^{\star}_{t,i,j} + \xi_{t,i,j},$$

where $\mu_j(u_i)$ is the regular latent price component, $G^{\star}_{t,i,j}$ is the structural jump innovation, and $\xi_{t,i,j}$ is microstructure noise, the oracle local baseline is

$$B^{\star}_{t,i,j} = \mu_j(u_i).$$

Therefore,

$$X_{t,i,j} - B^{\star}_{t,i,j} = G^{\star}_{t,i,j} + \xi_{t,i,j}.$$

The implemented JSPI estimator is residual-based:

$$\widehat{S}_t = \widehat{Z}_t \odot (X_t - B_t).$$

Consequently, the natural oracle target for the implemented estimator is

$$S_t^{\mathrm{res},\star} = Z_t^{\star} \odot (X_t - B_t^{\star}),$$

and the corresponding oracle JSPI is

$$\mathrm{JSPI}_t^{\mathrm{res},\star} = \sum_{i,j \in \mathcal{J}_P} Z^{\star}_{t,i,j}(X_{t,i,j} - B^{\star}_{t,i,j}).$$

Equivalently,

$$\mathrm{JSPI}_t^{\mathrm{res},\star} = \sum_{i,j \in \mathcal{J}_P} Z^{\star}_{t,i,j} G^{\star}_{t,i,j} + \sum_{i,j \in \mathcal{J}_P} Z^{\star}_{t,i,j} \xi_{t,i,j}.$$

By contrast, a pure structural jump oracle would be

$$\mathrm{JSPI}_t^{G,\star} = \sum_{i,j \in \mathcal{J}_P} Z^{\star}_{t,i,j} G^{\star}_{t,i,j}.$$

The two oracle quantities differ by the jump-coordinate noise term

$$\mathrm{JSPI}_t^{\mathrm{res},\star} - \mathrm{JSPI}_t^{G,\star} = \sum_{i,j \in \mathcal{J}_P} Z^{\star}_{t,i,j} \xi_{t,i,j}.$$

Thus, convergence of $\widehat{\mathrm{JSPI}}_t$ to $\mathrm{JSPI}_t^{G,\star}$ does not follow from Theorem 6 alone. It requires additional assumptions or an additional denoising step that removes the jump-coordinate noise.

Potential Sufficient Conditions over ξ: *First, one may assume that the microstructure noise at detected jump coordinates is asymptotically negligible, for example*

$$\max_{(i,j): Z^{\star}_{t,i,j}=1} |\xi_{t,i,j}| = o_p(1).$$

Second, one may use a local denoising estimator of the pre- and post-jump levels and define jump size as a difference of denoised local levels rather than as a single residual. Third, one may aggregate across many independent or weakly dependent jump observations, in which case median-zero or mean-zero noise can average out under suitable dependence conditions.

In this work, we avoid imposing these assumptions and accordingly, our theoretical target is the residual oracle $\mathrm{JSPI}_t^{\mathrm{res},\star}$*, which is exactly aligned with the implemented estimator*

$$\widehat{S}_t = \widehat{Z}_t \odot (X_t - B_t).$$

This residual oracle is still meaningful: it represents the signed, locally standardized price displacement visible in the observed LOB window after subtracting the robust local baseline. In noisy high-frequency order-book data, this is the quantity directly available to us.

Theorem 7 (Prediction convergence under JSPI consistency)**.** *Suppose*

$$\widehat{\mathrm{JSPI}}_t - \mathrm{JSPI}_t^\star = O_p(\kappa_n)$$

for some $\kappa_n \to 0$*. Then*

$$d_{\mathrm{TV}}\left(p_t^{\mathrm{final}}, p_{t,\star}^{\mathrm{final}}\right) \le \tanh\left(\frac{\eta\left|\widehat{\mathrm{JSPI}}_t - \mathrm{JSPI}_t^\star\right|}{2}\right) = O_p(\eta\kappa_n).$$

Furthermore,

$$D_{\mathrm{KL}}\left(p_t^{\mathrm{final}} \,\|\, p_{t,\star}^{\mathrm{final}}\right) = O_p(\eta^2\kappa_n^2),$$

and

$$D_{\mathrm{KL}}\left(p_{t,\star}^{\mathrm{final}} \,\|\, p_t^{\mathrm{final}}\right) = O_p(\eta^2\kappa_n^2).$$

Proof. **Bounding the Total Variation:** follows from similar derivation as in Lemma 1 and we omit it.

Bounding the KL terms: Let

$$a_t = \eta\widehat{\mathrm{JSPI}}_t, \qquad a_t^\star = \eta\mathrm{JSPI}_t^\star.$$

For $a \in \mathbb{R}$, define the exponentially tilted distribution

$$p_a(c) = \frac{p_t^{\mathrm{base}}(c)e^{ac}}{Z_t(a)}, \qquad Z_t(a) = \sum_{c\in\{-1,0,+1\}} p_t^{\mathrm{base}}(c)e^{ac}.$$

Then

$$p_t^{\mathrm{final}} = p_{a_t}, \qquad p_{t,\star}^{\mathrm{final}} = p_{a_t^\star}.$$

Let

$$F_t(a) = \log Z_t(a)$$

be the log-partition function. We have that

$$D_{\mathrm{KL}}(p_a\|p_b) = F_t(b) - F_t(a) - F_t'(a)(b-a).$$

By Taylor's approximation at F_t since $\kappa_n \to 0$, for some ξ between a and b,

$$D_{\mathrm{KL}}(p_a\|p_b) = \frac{1}{2}F_t''(\xi)(b-a)^2.$$

Moreover,

$$F_t''(u) = \mathrm{Var}_{p_u}(C), \qquad C \in \{-1,0,+1\}.$$

Since $C \in [-1,1]$,

$$F_t''(u) \le 1$$

for every $u \in \mathbb{R}$. Therefore,

$$D_{\mathrm{KL}}(p_a\|p_b) \le \frac{1}{2}(a-b)^2.$$

Taking

$$a = a_t, \qquad b = a_t^\star,$$

we obtain

$$D_{\mathrm{KL}}\left(p_t^{\mathrm{final}} \,\|\, p_{t,\star}^{\mathrm{final}}\right) \le \frac{\eta^2}{2}\left|\widehat{\mathrm{JSPI}}_t - \mathrm{JSPI}_t^\star\right|^2 = O_p(\eta^2 r_n^2).$$

The reverse KL is identical:

$$D_{\mathrm{KL}}(p_b \| p_a) = F_t(a) - F_t(b) - F_t'(b)(a-b) \le \frac{1}{2}(a-b)^2.$$

Thus,

$$D_{\mathrm{KL}}\left(p_{t,\star}^{\mathrm{final}} \,\|\, p_t^{\mathrm{final}}\right) = O_p(\eta^2 r_n^2).$$

□

Corollary 4 (No-jump self-neutrality)**.** *Suppose the assumptions of Theorem* 6 *hold and there are no oracle price-jump coordinates in $\mathcal{J}_P$, so that*

$$\mathrm{JSPI}_t^\star = 0.$$

If

$$\widehat{\mathrm{JSPI}}_t = O_p(\kappa_n), \qquad \kappa_n \to 0,$$

then

$$d_{\mathrm{TV}}\left(p_t^{\mathrm{base}}, p_t^{\mathrm{final}}\right) = O_p(\eta\kappa_n),$$

and

$$D_{\mathrm{KL}}\left(p_t^{\mathrm{base}} \,\|\, p_t^{\mathrm{final}}\right) = O_p(\eta^2\kappa_n^2).$$

Thereby keeping the model predictions same. Since our recipe follows a fixed test-time predictor regime, we then admit the same performance empirically and theoretically, of the base predictor.

B.3 Hold-out selection of the calibration strength

The following theorem provides a high probability bound using standard ML techniques under IID assumptions. We make this characterization explicit as the LOB data does not admit the assumptions we consider in the following theorem and **weakly applicable** to a limited range of settings where the auto-correlation vanishes at a fast rate. However, as direct extensions of this work (which we leave for future efforts), we provide a discussion in Remark 6 for theoretical results and risk minimization bounds for dependent data.

Note on application of concentration inequalities (Hoeffding) in theorem 8 below: Using extensions of the concentration inequalities provided by Nooghabi & Azarnoosh (2009), although the observations are not independent, for the setup where the sequence of the r.v's $\ell_\eta(t, Y_t)$ (loss function as defined below) is negatively associated, the concentration inequality continue to hold. However, such negative association may imply the instability of JSPI risk as well - again, a detailed theoretical study of this is left as future work.

Theorem 8 (Hold-out selection of the JSPI calibration η)**.** *Let*

$$\mathcal{D}_{\mathrm{val}} = \{(X_t, Y_t)\}_{t=1}^{n_{\mathrm{val}}}$$

be an independent validation set, drawn i.i.d. from the same distribution as the test point. Assume that the trained predictor f_θ, the JSPI estimator, and the base probabilities p_t^{base} are fixed before evaluating the validation set.

For $\eta \in [0, H]$, define the JSPI-corrected probability

$$p_{t,\eta}(c) = \frac{p_t^{\mathrm{base}}(c)\exp\left(\eta c\,\mathrm{JSPI}_t\right)}{\sum_{c'\in\{-1,0,+1\}} p_t^{\mathrm{base}}(c')\exp\left(\eta c'\,\mathrm{JSPI}_t\right)}, \qquad c \in \{-1, 0, +1\},$$

and the negative log-likelihood loss

$$\ell_\eta(t, Y_t) = -\log p_{t,\eta}(Y_t).$$

Let

$$R(\eta) = \mathbb{E}\left[\ell_\eta(X, Y)\right]$$

be the population validation risk, and let

$$\widehat{R}_{\mathrm{val}}(\eta) = \frac{1}{n_{\mathrm{val}}}\sum_{t=1}^{n_{\mathrm{val}}} \ell_\eta(t, Y_t)$$

be the empirical validation risk. Define

$$\widehat{\eta} \in \arg\min_{\eta \in [0,H]} \widehat{R}_{\mathrm{val}}(\eta).$$

Assume that

$$|\mathrm{JSPI}_t| \le B$$

almost surely. Assume also that the base probabilities are clipped away from zero: for some $\rho \in (0, 1/3]$*,*

$$p_t^{\mathrm{base}}(c) \ge \rho, \qquad c \in \{-1, 0, +1\},$$

almost surely. Then, for all $\eta \in [0, H]$*,*

$$0 \le \ell_\eta(t, Y_t) \le M_\rho, \qquad M_\rho = \log(1/\rho) + 2HB,$$

and

$$|\ell_\eta(t, Y_t) - \ell_{\eta'}(t, Y_t)| \le 2B|\eta - \eta'|.$$

Consequently, with probability at least $1 - \delta$*,*

$$R(\widehat{\eta}) \le \inf_{\eta \in [0,H]} R(\eta) + 2M_\rho \sqrt{\frac{\log\left(\frac{2N_n}{\delta}\right)}{2n_{\mathrm{val}}}} + \frac{4M_\rho}{\sqrt{n_{\mathrm{val}}}},$$

where

$$N_n = 1 + \left\lceil \frac{2BH\sqrt{n_{\mathrm{val}}}}{M_\rho} \right\rceil.$$

Equivalently,

$$R(\widehat{\eta}) \le \inf_{\eta \in [0,H]} R(\eta) + O\left(M_\rho \sqrt{\frac{\log\left(1 + \frac{BH\sqrt{n_{\mathrm{val}}}}{M_\rho}\right) + \log(1/\delta)}{n_{\mathrm{val}}}} \right).$$

Proof sketch. For some fixed t, let

$$J_t = \mathrm{JSPI}_t$$

and

$$Z_t(\eta) = \sum_{c \in \{-1,0,+1\}} p_t^{\mathrm{base}}(c) \exp(\eta c J_t).$$

Then

$$\ell_\eta(t, Y_t) = -\log p_t^{\mathrm{base}}(Y_t) - \eta Y_t J_t + \log Z_t(\eta).$$

First, we show boundedness. Since

$$|cJ_t| \le B \qquad \text{and} \qquad \eta \in [0, H],$$

we have

$$e^{-HB} \le \exp(\eta c J_t) \le e^{HB}.$$

Also, since the base probabilities sum to one,

$$Z_t(\eta) \le e^{HB}.$$

Using $p_t^{\mathrm{base}}(Y_t) \ge \rho$, we obtain

$$p_{t,\eta}(Y_t) = \frac{p_t^{\mathrm{base}}(Y_t) \exp(\eta Y_t J_t)}{Z_t(\eta)} \ge \frac{\rho e^{-HB}}{e^{HB}} = \rho e^{-2HB}.$$

Therefore

$$\ell_\eta(t, Y_t) = -\log p_{t,\eta}(Y_t) \le \log(1/\rho) + 2HB = M_\rho.$$

Next, we show Lipschitzness in η. Under the assumptions of this theorem, we have a differentiable loss function, thus the derivative

$$\frac{\partial}{\partial \eta}\ell_\eta(t, Y_t) = -Y_t J_t + \frac{\sum_c p_t^{\text{base}}(c) c J_t \exp(\eta c J_t)}{Z_t(\eta)}$$
$$= J_t \left(\sum_c c\, p_{t,\eta}(c) - Y_t \right).$$

Because

$$c \in \{-1, 0, +1\}, \qquad Y_t \in \{-1, 0, +1\},$$

we have

$$\sum_c c\, p_{t,\eta}(c) \in [-1, 1], \qquad Y_t \in [-1, 1].$$

Thus

$$\left| \frac{\partial}{\partial \eta}\ell_\eta(t, Y_t) \right| \le 2|J_t| \le 2B.$$

Hence

$$|\ell_\eta(t, Y_t) - \ell_{\eta'}(t, Y_t)| \le 2B|\eta - \eta'|.$$

Discretization (Covering Number style argument) (Vapnik & Vapnik, 1998) Now construct a grid $\mathcal{G}_n \subset [0, H]$ with mesh size

$$\Delta_n = \frac{M_\rho}{2B\sqrt{n_{\text{val}}}},$$

so that every $\eta \in [0, H]$ is within distance Δ_n of some grid point. The grid size satisfies

$$|\mathcal{G}_n| \le 1 + \left\lceil \frac{H}{\Delta_n} \right\rceil = 1 + \left\lceil \frac{2BH\sqrt{n_{\text{val}}}}{M_\rho} \right\rceil = N_n.$$

For any fixed $\eta \in \mathcal{G}_n$, the random variables

$$\ell_\eta(t, Y_t), \qquad t = 1, \ldots, n_{\text{val}},$$

are i.i.d. and bounded in $[0, M_\rho]$. Hoeffding's inequality gives

$$\mathbb{P}\left(\left| \widehat{R}_{\text{val}}(\eta) - R(\eta) \right| > u \right) \le 2 \exp\left(-\frac{2n_{\text{val}} u^2}{M_\rho^2} \right). \tag{5}$$

Note: at the beginning of this subsection, we provide a detailed discussion of the case where the sequence of the r.v's $\ell_\eta(t, Y_t)$ is negatively associated and the above concentration bound continues to hold.

Applying the union bound over $\mathcal{G}_n$ to eq 5, with probability at least $1 - \delta$,

$$\max_{\eta \in \mathcal{G}_n} \left| \widehat{R}_{\text{val}}(\eta) - R(\eta) \right| \le M_\rho \sqrt{\frac{\log(2N_n/\delta)}{2n_{\text{val}}}}.$$

Interval to Grid bound: For any $\eta \in [0, H]$, choose $\pi(\eta) \in \mathcal{G}_n$ such that

$$|\eta - \pi(\eta)| \le \Delta_n.$$

By the $2B$-Lipschitz property, both the population and empirical risks satisfy

$$|R(\eta) - R(\pi(\eta))| \le 2B\Delta_n,$$

and

$$|\widehat{R}_{\text{val}}(\eta) - \widehat{R}_{\text{val}}(\pi(\eta))| \le 2B\Delta_n.$$

Therefore,

$$\sup_{\eta \in [0,H]} \left| \widehat{R}_{\text{val}}(\eta) - R(\eta) \right| \le M_\rho \sqrt{\frac{\log(2N_n/\delta)}{2n_{\text{val}}}} + 4B\Delta_n.$$

Since

$$4B\Delta_n = \frac{2M_\rho}{\sqrt{n_{\text{val}}}},$$

we obtain

$$\sup_{\eta\in[0,H]} \left|\widehat{R}_{\text{val}}(\eta) - R(\eta)\right| \le M_\rho \sqrt{\frac{\log(2N_n/\delta)}{2n_{\text{val}}}} + \frac{2M_\rho}{\sqrt{n_{\text{val}}}}.$$

Finally, by empirical optimality of $\widehat{\eta}$ as a minimizer of empirical risk (loss),

$$\widehat{R}_{\text{val}}(\widehat{\eta}) \le \widehat{R}_{\text{val}}(\eta) \qquad \text{for all } \eta \in [0, H].$$

Furthermore, let $\eta^\star \in \arg\min_{\eta\in[0,H]} R(\eta)$.

Then at $\widehat{\eta}$ we have

$$|R(\widehat{\eta}) - \widehat{R}_{\text{val}}(\widehat{\eta})| \le \sup_\eta |R(\eta) - \widehat{R}_{\text{val}}(\eta)| \tag{6}$$

giving

$$\begin{aligned} R(\widehat{\eta}) &\le \widehat{R}_{\text{val}}(\widehat{\eta}) + \sup_\eta |R(\eta) - \widehat{R}_{\text{val}}(\eta)| \\ &\le \widehat{R}_{\text{val}}(\eta^\star) + \sup_\eta |R(\eta) - \widehat{R}_{\text{val}}(\eta)| \\ &\le R(\eta^\star) + 2\sup_\eta |R(\eta) - \widehat{R}_{\text{val}}(\eta)|. \end{aligned}$$

Substituting the uniform convergence bound gives the theorem statement. □

Remark 5 (Role of the validation set)**.** *In this theorem we treat f_θ, the* JSPI *computation, and the candidate family $\{p_{t,\eta} : \eta \in [0, H]\}$ as fixed before hold-out selection. Thus, the only learned quantity on the validation set is the single scalar η. This is analogous to most post-hoc calibration methods where a low-dimensional calibration parameter is selected on a held-out set while the underlying predictor network remains fixed.*

Remark 6 (Temporally Correlated Data Stream)**.** ***Note on our Implementation:*** *Our implementation for the selection of η uses out of sample validation set where the data used for training and selection of the key hyperparameters is obtained before threshold of first 7 days of FI-2010 dataset, in order to avoid look-ahead bias and leakage. See (Bergmeir et al., 2018) for a broader discussion of cross-validation in autoregressive time-series prediction.*

Extensions of Theorem 8 to dependent validation data however can be obtained under stationarity and mixing assumptions by using uniform convergence or concentration inequalities for empirical processes with dependent observations. In such results, the sample size n_{val} is typically replaced, up to constants and logarithmic factors, by a dependence-adjusted effective sample size, (Mohri & Rostamizadeh, 2007), (Hang & Steinwart, 2017), and (Chen & Wu, 2018) are some of the relevant literature. Related risk minimization bounds for dependent and heavy-tailed data are studied by (Roy et al., 2021).