A Unified Theory for Volume, Impact, and Volatility

Youssef O. Chahdi* Johannes Muhle-Karbe † Mathieu Rosenbaum‡
Grégoire Szymanski§

January 7, 2026

Abstract

We introduce a two-layer Hawkes model for the autonomous "core" order flow and the market's endogenous reaction. This model has a natural scaling limit, which reconciles a number of salient but apparently contradictory empirical facts. To wit, the aggregate unsigned order flow increments are rough (like price volatilities), whereas the signed order flow is the sum of a martingale and a process with long-range dependence. Under no-arbitrage constraints, the order flow also pins down a corresponding price model, where volatility is rough and the price impact of each trade decays according to a power law. All the quantities are determined by a single parameter (measuring the persistence of core and reaction flow in the micro-model), but nevertheless turn out to be remarkably consistent with empirical estimates.

Keywords: Trading volume, order flow, rough volatility, market impact, long memory, market microstructure, Hawkes processes, limit theorems.

Mathematics Subject Classification (2020): 60F05, 60G22, 60G55, 62P05, 91G15, 91G80

*MICS, CentraleSupélec, youssef.ouazzani-chahdi@centralesupelec.fr
†Department of Mathematics, Imperial College London, j.muhle-karbe@imperial.ac.uk
‡Ceremade, Université Paris Dauphine-PSL, mathieu.rosenbaum@dauphine.psl.eu
§DMATH, Université du Luxembourg, gregoire.szymanski@uni.lu

Youssef Ouazzani Chahdi, Mathieu Rosenbaum and Grégoire Szymanski gratefully acknowledge support from the ILB Chair *Artificial Intelligence and Quantitative Methods for Finance* at University Paris Dauphine-PSL

1 Introduction

Prices and the corresponding traded quantities are the fundamental observables in any financial market. Yet, "although most models of asset markets have focused on the behavior of returns [...] their implications for trading volume have received far less attention" [28].

Indeed, over reasonably short time horizons, the natural baseline model for prices are martingales, for which systematic trading profits are difficult to attain. In contrast, trading volume is well known to exhibit strong persistence, but no canonical model has emerged yet. The challenge here is to consistently model a number robust empirical properties that appear almost contradictory. On the one hand, the long-range dependence of order flow [4, 26–28] naturally motivates models based on fractional Brownian motion with Hurst index $H \gg 0.5$. On the other hand, (unsigned) daily traded amounts are almost synonymous with volatility [10, 29, 32], which is better modeled by rough fractional Brownian motions with Hurst index $H \ll 0.5$ [15].

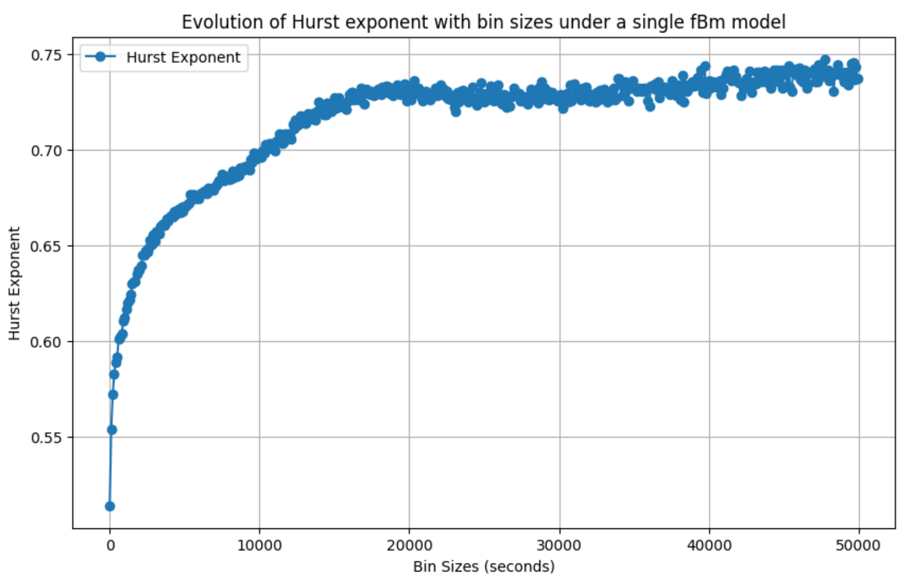

Figure 1: Average Hurst exponent estimations over the 40 stocks of the CAC40 index (2021–2024) using a single fractional Brownian motion specification.

This ambiguity is further compounded by the fact that such measurements for trading volume are not scale invariant. This is illustrated in Figure 1, which displays a standard roughness estimator computed from (signed) trading volume sampled at different frequencies. We observe that at high frequencies the estimates are close to $H = 0.5$ in line with diffusion models as in [5, 6, 17]. However, as observation frequencies become coarser, the estimates increase steadily and reach values near $H = 0.75$ at daily frequencies.

In addition to this multiscale behavior, the corresponding unsigned trading volume displays a completely different behavior. Indeed, as illustrated in Fig. 2, the time series of traded amounts is very rough, with roughness estimates well below 0.5 at all timescales just like for price volatilities.

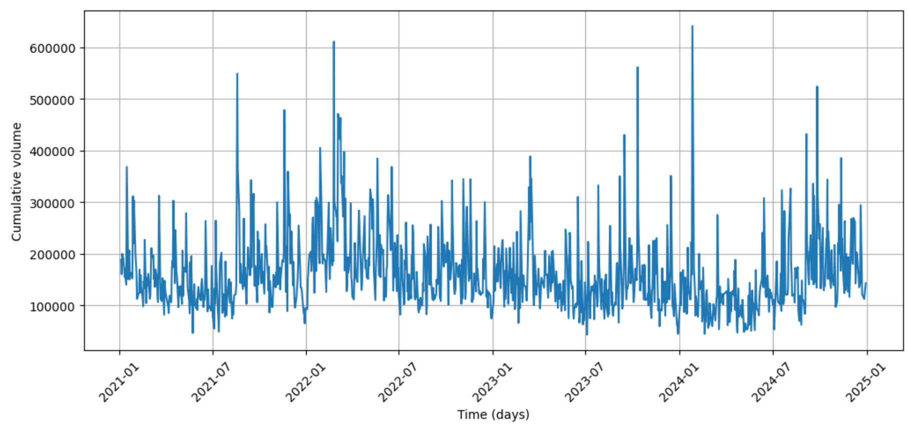

Figure 2: Daily increments of the unsigned trading volume for LVMH.

In the present study, we reconcile these apparently conflicting stylized facts. To this end, we start from a micro-model for the order flow based on two different types of Hawkes processes. The first models the "core flow", i.e., trades submitted for autonomous reasons.[1] The second describes the "reaction flow", i.e., the strong reaction of the market to core and reaction orders in the spirit of ([11, 19, 20, 23–25], where the core flow is constant). We show that this micro-model has a natural scaling limit with the following properties, determined by a *single statistic* H_0 of the micro model (a measure of the persistence of the core and reaction flow):

- The (cumulative) unsigned order flow converges to a (integral) of a rough process with Hurst exponent $H_0 - 1/2$, paralleling results for (integrated) volatility of asset prices.
- The cumulative signed order flow is the sum of a fractional component with Hurst index $H_0 \in (1/2, 1)$ and an independent martingale.

Thereby, the model reconciles the rough estimates for unsigned order flow with the strong persistence of the signed flow. What is more, such a mixed fractional model (as studied by [7] as a model for asset *prices*) also matches the multiscale behavior of signed trading volume observed in Figure 1.

But that is not all: via the future expected flow, the present model also naturally leads to a consistent price model as in [25]. To ensure prices are martingales, the price impact of trades in turn needs to decay according to a power law kernel with exponent $2 - 2H_0$. The corresponding price volatility then pinned down as another fractional process with Hurst exponent $2H_0 - 3/2$.

Whence, all of these quantities are determined by the single parameter H_0. Despite this extreme parsimony, the model's testable predictions turn out to be remarkably consistent with empirical estimates. Indeed, for $H_0 = 3/4$, the signed order flow has this Hurst index, which closely matches the empirical estimates one obtains from estimating a mixed fractional Brownian motion model. As displayed in Figure 3, these estimates are now very stable across a wide range of aggregation time scales.

[1] The key role of such "fundamental" trading practices in the overall behavior of the market is also underlined in [30, 31], who develop a detailed model of metaorders, characterizing their size, duration, and mutual correlations. In contrast, our approach is more reduced-form but its natural scaling limit turns out to have very precise implications.

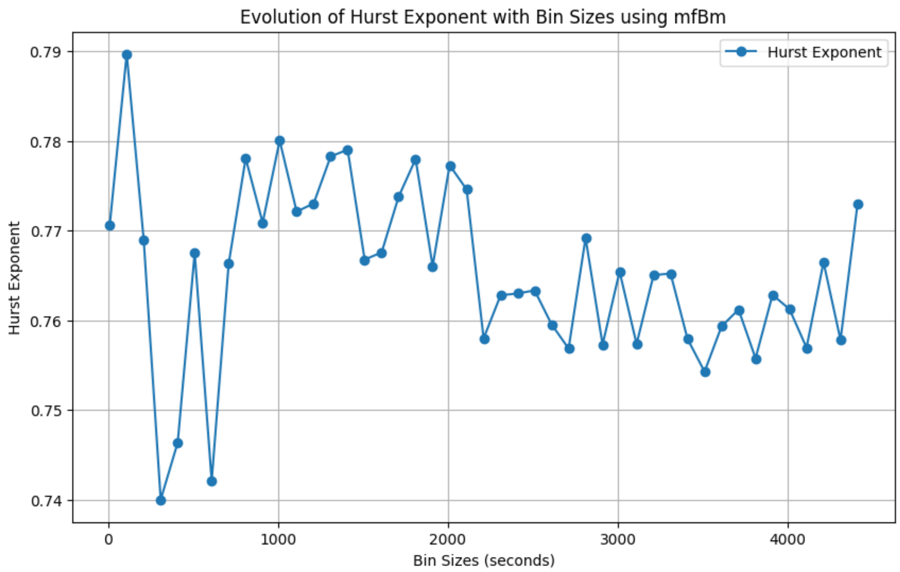

Figure 3: Average Hurst exponent estimations over 40 stocks on the period 2021-2024 using a mixed fractional Brownian motion.

The corresponding unsigned order flow in turn has Hurst index 1/4, which closely resembles the properties of time series data as displayed in Figure 2. Moreover, for $H_0 = 3/4$, we obtain a very rough volatility process with Hurst index ≈ 0 in line with, e.g., [3, 9, 13, 15]. Finally, price impact then decays with a square root kernel. For orders with constant participation rate, this leads to the celebrated "square-root law" linking the price impact of metaorders to the quantity executed. In previous studies, reconciling this feature with diffusive prices requires much more autocorrelated order flow in the data. In contrast, the interplay of core and reaction flow in our model achieves this while remaining consistent with the empirical properties of the flow.

In summary, the scaling limit of our Hawkes model suggests an extremely parsimonious limiting model that consistently matches a wide range of empirical properties for trading volume, price impact dynamics, and the corresponding price volatilities.

The remainder of the paper is organized as follows. Section 2 introduces the two-layer branching Hawkes framework. Section 3 derives the scaling limits for core, reaction, and total flows. Section 4 connects the core order flow dynamics to volatility and impact.

2 A two-layer Hawkes model for order flow

This section introduces a model that will reconcile the apparently contrasting properties of unsigned and signed order flows and establish a unified framework connecting it to rough volatility and the decay of market impact according to a power law kernel.

To this end, we decompose the aggregate order flow into two conceptually distinct building blocks:

- **Core order flow:** The "core order flow" arises from a heterogeneous mixture of trading motives and horizons. In particular, it comprises medium and low frequency strategies, often grounded in fundamental information, long term valuation views, or trend following dynamics. Such strategies explain part of the empirically observed persistence in order flow, where the signs of trades exhibit long-range dependence. This is complemented by metaorder splitting: large institutional trades are executed incrementally over time to minimize market impact.

- **Reaction orders:** unlike the core flow, "reaction orders" are not initiated for autonomous reasons but arise endogenously as a response to other trades. This applies both to the core flow (which contains both informed trades and trading opportunities) and to other reaction orders. Such reaction orders reflect the dynamic interplay among liquidity providers, high frequency market makers and quantitative strategies that continuously adjust their positions and inventories. The resulting feedback mechanisms generate additional layers of correlation within the order flow, that in turn amplify the persistence of the core order flow.

2.1 Core order flow

We model the core buy and sell orders by two independent univariate Hawkes processes, denoted by F^+ and F^-, respectively. This is a very natural modeling tool for the splitting of a metaorder or a trend following strategy, for example: once a child order is submitted, the probability of observing further orders of the same sign increases, reflecting the continuation of an execution program.

Formally, both F^+ and F^- have the same baseline intensity $\nu > 0$ and the same excitation kernel $\varphi_0 \colon \mathbb{R}_+ \to \mathbb{R}_+$ that governs the temporal dependence between trades.[2] Hence, the intensities of core buy and sell orders are given by

$$\lambda_t^+ = \nu + \int_0^{t-} \varphi_0(t-s)\, dF_s^+, \qquad \lambda_t^- = \nu + \int_0^{t-} \varphi_0(t-s)\, dF_s^-.$$

When the kernel φ_0 decays rapidly, the process approximates a memoryless sequence of orders, with limited interaction between successive trades. Conversely, a slowly decaying φ_0 implies that each trade continues to elevate the probability of subsequent trades in the same direction over an extended horizon. This allows to capture that large metaorders or trend-following strategies, once initiated, generate persistent streams of transactions.

The signed and unsigned core order flow are in turn given by

$$F_t = F_t^+ - F_t^-, \qquad V_t = F_t^+ + F_t^-,$$

which measure the directional and overall trading volumes due to core trading activity.

2.2 Reaction orders

We now turn to the market's endogenous reaction to incoming orders. As discussed above, it is difficult to discriminate autonomous baseline orders from other trades, so we model this reaction flow via Hawkes process driven by baseline and other reaction trades in the same manner. More specifically, we consider a two-dimensional Hawkes process

$$\mathbf{N}_t = (N_t^+, N_t^-),$$

[2]This symmetry is required to obtain a finite svaling limit below.

where N^+ describes reaction buys and N^- models reaction sells.

The baseline intensity of $\mathbf{N}$ is driven by the reaction to core orders through a symmetric kernel matrix

$$\boldsymbol{\phi} = \begin{pmatrix} \varphi_1 & \varphi_2 \\ \varphi_2 & \varphi_1 \end{pmatrix},$$

so that

$$\boldsymbol{\mu}_t = \int_0^t \boldsymbol{\phi}(t-s) \cdot d\mathbf{F}_s \qquad \text{where} \qquad \mathbf{F}_t = (F_t^+, F_t^-).$$

The global intensity of $\mathbf{N}$ is given by

$$\boldsymbol{\lambda}_t = \boldsymbol{\mu}_t + \int_0^t \boldsymbol{\phi}(t-s) \cdot d\mathbf{N}_s = \int_0^t \boldsymbol{\phi}(t-s) \cdot \mathrm{d}(\mathbf{F}_s + \mathbf{N}_s).$$

This structure of Hawkes process branching on Hawkes processes captures the asymmetry of reactions. More precisely:

- Following a baseline buy order at t_0, a wave of reaction buy orders (with intensity φ_1) is triggered on the ask side, reflecting for instance momentum strategies, while a smaller wave of reaction sell orders (with intensity φ_2) may appear on the bid side, reflecting inventory rebalancing or contrarian liquidity provision;
- The situation is symmetric for core sell orders, with φ_1 and φ_2 swapping roles (again, to obtain a finite scaling limit below).
- Non-core orders are digested by the market through the same mechanism. This is represented by the integral term in the intensity of $\mathbf{N}$. The market uses the very same kernel to process core and non-core orders as there is no way to distinguish between them.

2.3 Aggregate Order Flow

The aggregate order flow combines both baseline and reaction flow:

$$U_t = F_t^+ + F_t^- + N_t^+ + N_t^-, \qquad S_t = F_t^+ - F_t^- + N_t^+ - N_t^-,$$

where U_t is the unsigned aggregate order flow and S_t is its signed counterpart.

This decomposition underscores that long memory in the signed order flow S_t need not arise solely from persistence in the core component. Instead, the observed persistence of the aggregate flow is generated by the interaction between exogenous persistence in the core flow and its endogenous amplification by reaction trades. We now formalize this intuition by studying the scaling limit of the Hawkes model introduced in this section as the observation scale goes to infinity.

3 Scaling limits of the order flows

In this section, we are interested in the macroscopic behaviour of the different order flows introduced above. We first addresses the scaling limits of baseline orders and then turn to ones for reaction orders. Subsequently, we examine the scaling limits of the aggregate unsigned and signed order flow, respectively.

3.1 Scaling limit of the baseline flow

We consider the same model as in Section 2, with the additional exponent T, and assume that the processes are observed on a finite time horizon $[0, T]$ with $T > 0$. The goal of this section is to establish scaling limits for the core order flow process as T goes to infinity , thus capturing its macroscopic behavior. Following [24], we work in a nearly unstable, heavy–tailed Hawkes regime that captures both the high level of clustering of the core flow and the long memory of trading activity. We formalize this through the following assumptions.

Assumption A. *There exists a nonnegative sequence $(a_0^T)_{T\geqslant 0}$ converging to one such that $a_0^T < 1$ and*

$$\varphi_0^T = a_0^T \varphi_0,$$

for some completely monotone kernel φ_0 (see [1] for definition) such that $\|\varphi_0\|_{L^1} = 1$. Furthermore, there exists $0 < \alpha_0 < 1$ and a positive constant K_0 such that as t tends to infinity,

$$\alpha_0 t^{\alpha_0} \int_t^\infty \varphi_0(t)\, dt \to K_0.$$

From a probabilistic perspective, a Hawkes process can be viewed as a population process and the norm of the corresponding self-exciting kernel, in this case φ_0, can be interpreted as the proportion of descendants in the whole population. In the financial setting, the norm $\|\varphi_0\|_{L^1}$ can be seen as the proportion of endogenous orders in the market. In fact, markets are highly endogenous, in the sense that a large fraction of orders are not driven by exogenous economic information but are instead generated algorithmically in reaction to past order flow. In our model, this translates into the assumption that the norm of the self-exciting kernel converges to one, while remaining strictly less than one. In particular, the condition $\|\varphi_0\|_{L^1} < 1$ ensures the existence of a stationary solution for the intensity. The second part of Assumption A imposes a heavy-tailed kernel, which captures strong clustering in order arrivals induced, e.g., by the splitting of metaorders.

To obtain non-degenerate limits for our signed and unsigned core flows, the parameters a_0^T, α_0 and the baseline intensity ν^T of the baseline flow have to be scaled appropriately:

Assumption B. *There exists two constants $\lambda_0, \mu_0 > 0$ such that*

$$\lim_{T\to\infty} T^{\alpha_0}(1 - a_0^T) = \lambda_0 K_0 \frac{\Gamma(1-\alpha_0)}{\alpha_0} \qquad \textit{and} \qquad \lim_{T\to\infty} T^{1-\alpha_0}\nu^T = \mu_0 \frac{\alpha_0}{K_0\Gamma(1-\alpha_0)}$$

where Γ represents the Gamma function.

Under these assumptions, the long-term average intensity of the Hawkes process $F^{\pm,T}$ is $(1 - a_0^T)^{-1}\nu^T$. Therefore the average number of trades from $F^{\pm,T}$ on $[0, T]$ scales as $T\nu^T(1 - a_0^T)^{-1}$. Thus it is natural to scale each of the Hawkes processes by $(1 - a_0^T)^{-1}\nu^T T$ and consider the scaled processes

$$\overline{F}_t^{\pm,T} = \frac{1 - a_0^T}{T\nu^T} F_{tT}^{\pm,T}.$$

Let us define for $\alpha_0 > 0$ and $\lambda_0 > 0$ the function f^{α_0,λ_0} by

$$f^{\alpha_0,\lambda_0}(x) = \lambda_0 x^{\alpha_0 - 1} E_{\alpha_0,\alpha_0}(-\lambda_0 x_0^{\alpha}),$$

where $E_{\alpha,\beta}$ is the (α,β)-Mittag-Leffler function

$$E_{\alpha,\beta}(x) = \sum_{k=0}^{\infty} \frac{x^k}{\Gamma(\alpha k + \beta)},$$

see [18]. The following theorem is proved in Appendix B.

Theorem 3.1. *Under Assumptions A and B, the process* $(\overline{F}_t^{+,T}, \overline{F}_t^{-,T})_{t\in[0,1]}$ *is tight for the Skorohod topology. Furthermore, any limit point* (F_t^+, F_t^-) *of* $(\overline{F}_t^{+,T}, \overline{F}_t^{-,T})$ *satisfies*

$$F_t^{\pm} = \int_0^t s f^{\alpha_0,\lambda_0}(t-s)\,ds + \frac{1}{\sqrt{\mu_0\lambda_0}} \int_0^t f^{\alpha_0,\lambda_0}(t-s) Z_s^{\pm}\,ds,$$

where Z^+ *and* Z^- *are two continuous martingales with quadratic variations respectively* F^+ *and* F^-.

When $\alpha_0 > 1/2$, then the limiting processes in Theorem 3.1 are differentiable and their derivatives belong to the class of rough Heston models developed in [12, 24]. However, the empirically relevant case will turn out to be $\alpha_0 < 1/2$, where the core flow is highly persistent and the limiting processes become non-differentiable. This corresponds to the hyper-rough Heston models introduced in [25]. We give in the following proposition an accurate statement regarding the regularity of such processes. This notion is of particular interest since, in the high-frequency regime and for the class of processes in our framework, the Hölder regularity coincides with the Hurst exponent. As a consequence, characterizing the Hölder regularity provides direct information on the roughness of the sample paths and on the associated Hurst parameter.

Proposition 3.2. *For any* $\varepsilon > 0$*, the processes* F^+ *and* F^- *are almost surely Hölder continuous with exponent* $(1 \wedge 2\alpha_0) - \varepsilon$ *on* $[0,1]$*. In the case* $\alpha_0 < 1/2$*, they are moreover exactly* $2\alpha_0$*-Hölder continuous in* L^2*, that is, there exists a constant* $C > 0$ *such that for any* $t \in [0,1]$*, as* h *tends to zero:*

$$\Big(\mathbb{E}|F_{t+h} - F_t|^2\Big)^{1/2} = C h^{2\alpha_0} + o(h^{2\alpha_0}).$$

As T goes to infinity, the scaled signed and unsigned core processes satisfy

$$\overline{F}_t^{+,T} + \overline{F}_t^{-,T} \longrightarrow F_t^+ + F_t^- \qquad \text{and} \qquad \overline{F}_t^{+,T} - \overline{F}_t^{-,T} \longrightarrow F_t^+ - F_t^-.$$

From Theorem 3.1, we obtain the following proposition.

Proposition 3.3. *Let*

$$F_t = F_t^+ + F_t^- \qquad \textit{and} \qquad V_t = F_t^+ - F_t^-$$

denote the scaling limits of the unsigned and signed core flows, respectively, where F^+ *and* F^- *are given in Theorem 3.1. We have*

$$F_t = 2\int_0^t s\, f^{\alpha_0,\lambda_0}(t-s)\,ds + \frac{1}{\sqrt{\mu_0\,\lambda_0}} \int_0^t f^{\alpha_0,\lambda_0}(t-s)\, Z_s^F\,ds$$

and

$$V_t = \frac{1}{\sqrt{\mu_0\,\lambda_0}} \int_0^t f^{\alpha_0,\lambda_0}(t-s)\, Z_s^V\,ds, \tag{1}$$

where Z^F and Z^V are two continuous martingales with quadratic variation F and quadratic co-variation V such that

$$Z^F = Z^+ + Z^- \qquad \text{and} \qquad Z^V = Z^+ - Z^-,$$

where Z^+ and Z^- are given in Theorem 3.1.

When $\alpha_0 < 1/2$, F and V are exactly $2\alpha_0$-Hölder continuous in L^2. Thus, the signed and unsigned core processes have the same regularity as a fractional Brownian motion with Hurst exponent

$$H_0 = 2\alpha_0.$$

3.2 Scaling limit of the reaction orders

As for the baseline flows, we augment the notations for the reaction flow from Section 2 with the additional exponent T for the time horizon. We write

$$\mathbf{\Lambda}_t^T = \int_0^t \boldsymbol{\lambda}_s^T ds$$

for the compensator of our Hawkes process and the associated martingale is denoted by

$$\mathbf{M}_t^T = \mathbf{N}_t^T - \mathbf{\Lambda}_t^T.$$

We are again interested in the macroscopic scaling behavior of the reaction orders. Therefore, in the same spirit as in [11], we make the following assumption that again reflects the high degree of endogeneity of financial markets.

Assumption C. *There exists a nonnegative sequence $(a_1^T)_{T\geqslant 0}$ converging to one such that $a_1^T < 1$ and*

$$\boldsymbol{\phi}^T = a_1^T \boldsymbol{\phi}$$

for some matrix $\boldsymbol{\phi}$ such that its spectral radius satisfies

$$\mathcal{S}(\|\boldsymbol{\phi}\|_{L^1}) = \|\varphi_1\|_{L^1} + \|\varphi_2\|_{L^1} = 1.$$

From [25], we also know that Assumption C is necessary in order to obtain non-trivial market impact on the market. We write $k_1(t) \geqslant k_2(t)$ for the eigenvalues of $\boldsymbol{\phi}(t)$, i.e.,

$$k_1(t) = \varphi_1(t) + \varphi_2(t), \quad k_2(t) = \varphi_1(t) - \varphi_2(t)$$

and denote by v_1, v_2 their associated eigenvectors

$$v_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad v_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

The following assumption relates to the slowly decreasing behavior of the kernel matrix, that is also necessary to obtain non-trivial market impact, see [25].

Assumption D. *There exists $1/2 < \alpha_1 < 1$ and $K_1 > 0$ such that*

$$\lim_{t\to\infty} \alpha_1 x^{\alpha_1} \int_t^\infty k_1(s)\, ds \to K_1.$$

We finally need to specify an asymptotic framework similar to that in Assumption B to ensure our limiting processes are not degenerate. In [11] there is a constant baseline μ^T and two positive constants λ_1 and μ_1 such that

$$T^{\alpha_1}(1-a_1^T) \to \lambda_1, \qquad \text{and} \qquad T^{1-\alpha_1}\mu^T \to \mu_1.$$

However, in our setting the baseline intensity $\boldsymbol{\mu}^T$ is itself stochastic and time-dependent. In [11], μ^T behaves like T^{α_1-1} as $T \to \infty$, so that the expected number of baseline-driven jumps on $[0,T]$, namely $T\,\mu^T$, grows like T^{α_1}. In our case, the number of baseline event between 0 and T is $F_T^{+,T} + F_T^{-,T}$, that is of order $(1-a_0^T)^{-1}T\,\nu^T$. Therefore, it is natural to replace $T\,\mu^T$ by $T\nu^T(1-a_0^T)^{-1}$ and to make the following assumption.

Assumption E. *There exist $\lambda_1, \mu_1 > 0$ such that*

$$T^{\alpha_1}(1-a_1^T) \to \lambda_1 \qquad \textit{and} \qquad \frac{T^{1-\alpha_1}\nu^T}{1-a_0^T} \to \mu_1.$$

By Assumption B, the product $T\nu^T$ is of order T^{α_0}, which implies from Assumption E that $1-a_0^T$ must be of order $T^{\alpha_0-\alpha_1}$. However, we already have that $1-a_0^T$ scales as $T^{-\alpha_0}$. Hence, to accommodate both of these scalings, we necessarily need

$$\alpha_1 = 2\,\alpha_0.$$

Also note that from Assumption D, we have $1/4 < \alpha_0 < 1/2$. As a consequence, the existence of a nontrivial scaling limit imposes strong structural constraints on the underlying Hawkes model.

In summary, we consider the scaled processes

$$\overline{N}_t^{\pm,T} = \frac{(1-a_0^T)(1-a_1^T)}{T\nu^T} N_{tT}^{\pm,T}, \qquad \overline{\Lambda}_t^{\pm,T} = \frac{(1-a_0^T)(1-a_1^T)}{T\nu^T} \Lambda_{tT}^{\pm,T},$$
$$\overline{M}_t^{\pm,T} = \Big(\frac{(1-a_0^T)(1-a_1^T)}{T\nu^T}\Big)^{1/2} M_{tT}^{\pm,T}.$$

We are now ready to state the convergence in distribution of these processes.

Theorem 3.4. *Under Assumptions A, B, C, D and E:*

- *The process $(\overline{N}^{+,T}, \overline{N}^{-,T}, \overline{\Lambda}^{+,T}, \overline{\Lambda}^{-,T}, \overline{M}^{+,T}, \overline{M}^{-,T})$ is C-tight for the Skorokhod topology. Moreover, each of its limit points (X, X, X, X, Z^+, Z^-) satisfies the rough Heston-type dynamics*
$$X_t = \frac{1}{2}\int_0^t f^{\alpha_1,\lambda_1}(t-s)F_s\,ds + \frac{1}{2\sqrt{\lambda_1\mu_1}}\int_0^t f^{\alpha_1,\lambda_1}(t-s)Z_s\,ds,$$
where $Z = Z^+ + Z^-$, with Z^+ and Z^- two continuous martingales with quadratic variation X and zero quadratic covariation, and F is given in Proposition.3.3. Furthermore, X behaves as an integrated rough process, and its derivative has Hölder regularity of order $(H_1 - \varepsilon)$ for any $\varepsilon > 0$ on $[0,1]$, where $H_1 = \alpha_1 - 1/2 = H_0 - 1/2$,
- *The scaled signed reaction flow $\overline{N}^{+,T} - \overline{N}^{-,T}$ converges in probability to zero.*

The first part of Theorem 3.4 suggests that the roughness in the unsigned order flow originates from reaction orders. The second statement shows that under the natural rescaling for the unsigned reaction flow, the signed reaction flow actually vanishes. This means unsigned and signed flows have a different order of magnitude, which will play a crucial role in the study of the asymptotic behaviors of the global flows in the next section.

3.3 Scaling limits for the global order flow

We now turn to the scaling limits of the aggregate order flow processes

$$\begin{aligned} U_t^T &= F_t^{T,+} + F_t^{-,T} + N_t^{+,T} + N_t^{-,T}, \\ S_t^T &= F_t^{T,+} - F_t^{-,T} + N_t^{+,T} - N_t^{-,T}. \end{aligned}$$

We start with the unsigned order flow, which is the easiest case as all the terms have already been fully investigated in Sections 3.1 and 3.2. We define

$$\overline{U}_t^T = \frac{(1-a_0^T)(1-a_1^T)}{T\nu^T} U_{tT}^T.$$

We have the following theorem.

Theorem 3.5. *Under Assumptions A, B, C, D and E, the scaled unsigned order flow* $\overline{U}^T$ *is C-tight in the Skorohod topology. Furthermore if* U *is a limit point of* $\overline{U}^T$*, then* U *satisfies*

$$U_t = 2X_t = \int_0^t f^{\alpha_1,\lambda_1}(t-s)F_s\,ds + \sqrt{\frac{1}{\lambda_1\mu_1}}\int_0^t f^{\alpha_1,\lambda_1}(t-s)Z_s\,ds, \tag{2}$$

where X *is defined in Theorem 3.4.*

We see that the contribution of the baseline flow vanishes in the limit of the aggregate unsigned trading volume, which is instead fully determined by the reaction flow.[3] As a consequence, just like the unsigned reaction flow, the aggregate unsigned flow is an integrated rough process and for any $\varepsilon > 0$, its derivative has Hölder regularity of order $H_1 - \varepsilon$ with $H_1 = \alpha_1 - 1/2 = H_0 - 1/2$.

We now turn to the signed order flow. As already observed in Theorem 3.4 for the reaction flow, the same scaling as for the unsigned order flow leads to a trivial limit here:

$$\frac{(1-a_0^T)(1-a_1^T)}{T\nu^T} S_{tT}^T \to 0.$$

The intuition for this is provided by Theorem 3.4: in the reaction flow, the buy and sell order flows have the same asymptotic scaling limits, which implies a vanishing difference. Therefore, we need to adapt the scaling for the signed order flow similarly as in [11]:

$$\overline{S}_t^T = \Big(\frac{(1-a_0^T)(1-a_1^T)}{T\nu^T}\Big)^{1/2} S_{tT}^T.$$

In this regime, we then obtain the following nontrivial limiting result:

[3]The only trace of the baseline flow is through the term F_s in the first integral in U_t, coming from the baseline intensity of the reaction orders.

Theorem 3.6. *Under Assumptions A, B, C, D and E, the scaled signed order flow $\overline{S}^T$ converges in the sense of finite dimensional laws to*

$$S_t = \frac{\sqrt{\lambda_1\mu_1}(\|\varphi_1\|_1 - \|\varphi_2\|_1)}{1-(\|\varphi_1\|_1 - \|\varphi_2\|_1)} V_t + \frac{1}{1-(\|\varphi_1\|_1 - \|\varphi_2\|_1)}(Z_t^+ - Z_t^-),$$

where V is given by Proposition 3.3, and Z^+ and Z^- are given by Theorem 3.4.

Theorem 3.6 shows that the aggregate signed order flow can be decomposed into two distinct components. The first is the contribution of the core order flow. It has the same regularity as a fractional Brownian motion with Hurst exponent $H_0 = 2\alpha_0$ and therefore induces persistence in aggregate signed flow. The second component is a martingale term, which originates from the reaction orders.

As discussed in the introduction, this decomposition is crucial to resolving the apparent lack of scale invariance in empirical order flow data. To this end, we replace the complex model from Theorem 3.6 with the simplest process with the same local behavior: a mixed fractional Brownian motion, that is, the sum of a fractional Brownian motion and an independent Brownian motion. Put differently, we use $S_t = W_t + B_t^{H_0}$, where W is a standard Brownian motion, used as a proxy for the reaction driven martingale component, and B^{H_0} is a fractional Brownian motion with Hurst exponent $H_0 = 2\alpha_0 = \alpha_1$, mirroring the regularity of the fractional component in the scaling limit of the aggregate signed flow.

We can now apply standard roughness estimators under this mixed fractional Brownian motion approximation. Figure 1 shows that when the aggregate signed order flow is approximated by a single fractional Brownian motion, the estimated Hurst exponent depends strongly on the bin size. For very fine sampling the estimate is close to 0.5, reflecting the dominance of the martingale term originating from reaction orders, while at larger bin sizes the estimate increases steadily as the persistent influence of core orders becomes more pronounced. In contrast, Figure 3 demonstrates that when the flow is modeled as a mixture of a fractional Brownian motion and a Brownian motion, the estimated Hurst exponent stabilizes around 0.77 across all bin sizes. Therefore, the testable implication of Theorem 3.6 are confirmed by the data.

4 From order flow to market impact and rough volatility

By adapting the arguments of [25], we now demonstrate that the order flow model discussed so far also allows to deduce the shape of the corresponding market impact decay kernel. The latter in turn pins down the corresponding prices, and it will turn out that the single parameter H_0 not only pins down the trading volume but also impact dynamics and the roughness of the volatility process.

The starting point of [25] is to assume no statistical arbitrage opportunities exist, which is tantamount to a martingale price. If the permanent price impact is linear (to rule out profitable roundstrips, cf. [14, 21]), it is show in [25] that the price P_t needs to equal

$$P_t = P_0 + \lim_{s\to\infty} \kappa\,\mathbb{E}[Q_s^+ - Q_s^- \mid \mathcal{G}_t], \tag{3}$$

where κ is the permanent impact coefficient, Q^+ and Q^- represent the cumulative buy and sell volumes up to time t, respectively, and $(\mathcal{G}_t)_{t\geqslant 0}$ is the natural filtration generated by the order flows

(Q^+, Q^-). Price movements thus correspond to the market's anticipation of future order flow. This relationship provides a general and model-independent link between order flow dynamics and price evolution, reconciling the strong persistence of order flow with the (approximate) martingale nature of prices (over reasonably short horizons).

If Q^a and Q^b are two independent Hawkes processes, then (3) takes the explicit propagator form

$$P_t = P_0 + \kappa \int_0^t \xi(t-s)\,(dQ_s^+ - dQ_s^-),$$

where ξ is a decay kernel compensating the memory of the flow that can be computed from the Hawkes excitation kernel. In this setting, [25] show that there necessarily exists some $\beta \in (0,1)$ such that the temporal market impact $MI(t)$ of a regularly scheduled metaorder over a renormalized time interval $[0,1]$, that is the average price deviation at a given time t due to the execution of the metaorder satisfies

$$MI(t) \sim t^{1-\beta}, \text{ for } t \leqslant 1,$$
$$MI(t) \sim t^{1-\beta} - (t-1)^{1-\beta}, \text{ for } t > 1.$$

The parameter β also is linked to the tail of the kernel of the Hawkes processes driving the flow: $\phi(t) \sim t^{-(1+\beta)}$ as t tends to infinity, see [25] for details. The celebrated square-root law in time (and in turn also in quantities, for metaorders executed with a constant participation rate) corresponds to the case $\beta = 1/2$ in the above formulas.[4]

Another important implication of this framework is the fact that the scaling limit of the price is a rough volatility model. Indeed, the volatility of the price is driven by a rough fractional process with roughness exponent $\beta - 1/2$, see again [25].

For more general order flow processes as in the present study, we see from Equation (3) that the key part of the flow for the link with price dynamics is its predictable component. If the limiting signed flow is approximated by a mixed fractional Brownian motion $(Y_t)_t$ whose fractional component has Hurst exponent $H_0 = 2\alpha_0$, then we know from [7] that, provided that $H_0 > 3/4$, Y is a semimartingale in its natural filtration. Empirically, we find that H_0 is greater than 3/4, see Figure 3. We therefore assume Y to be a semimartingale and decompose it into an unpredictable (local) martingale component M and a second component with finite variation A:

$$Y_t = M_t + A_t.$$

To proceed, the key idea now is to approximate the finite variation process A with a difference of two independent Hawkes processes $\widetilde{N}^a$ and $\widetilde{N}^b$ with the same baseline intensity and similar self-exciting kernel. More precisely, if this kernel decays as $t^{-(1+\alpha)}$ with $\alpha \in (1/2, 1)$, then the scaling limit of $\widetilde{N}^a - \widetilde{N}^b$ is continuous with a derivative that has regularity of order $(\alpha - 1/2 - \varepsilon)$ for any $\varepsilon > 0$. In addition, we know from [8] that A_t is $(2H_0 - 1/2 - \varepsilon)$ Hölder continuous for any $\varepsilon > 0$. Since $H_0 > 3/4$, A_t is differentiable and its derivative has a Hölder regularity of order $(2H_0 - 3/2 - \varepsilon)$ for any $\varepsilon > 0$. Hence the natural choice for the Hawkes approximation is to take

$$\alpha = 2H_0 - 1.$$

We therefore obtain the following link between the core order flow, the market impact exponent *and* the roughness of the price volatility.

[4] Note, however, that the exact shape of the relaxation phase is less agreed upon than that of the increasing phase of the impact, see [30].

Theorem 4.1. *Under the previous approximations, we have*

$$MI(t) \sim t^{2-2H_0}, \text{ for } t \leqslant 1,$$

$$MI(t) \sim t^{2-2H_0} - (t-1)^{2-2H_0}, \text{ for } t > 1.$$

Furthermore, the volatility of the price exhibits a rough behavior with a Hurst parameter $H_{vol} = 2H_0 - 3/2$.

Theorem 4.1 provides a structural relation between core order flow memory, market impact shape, and rough volatility. Intuitively, stronger persistence in the order flow (larger H_0) implies a faster-decaying propagator kernel (to compensate the memory), a more concave impact function (smaller β), and simultaneously rougher volatility at macroscopic scales. The obtained empirical value $H_0 = 0.77$ yields $1 - \beta = 0.46$ and $H_{\text{vol}} = 0.04$. The square-root law ($1 - \beta = 0.5$) corresponds to $H_0 = 3/4$, implying $H_{\text{vol}} = 0$.

References

[1] Serge Bernstein. Sur les fonctions absolument monotones. Acta Mathematica, 52:1–66, 1929.

[2] Patrick Billingsley. Convergence of Probability Measures. Wiley-Interscience, 1968.

[3] Anine E. Bolko, Kim Christensen, Mikko S. Pakkanen, and Bezirgen Veliyev. A GMM approach to estimate the roughness of stochastic volatility. Journal of Econometrics, 235(2):745–778, 2023.

[4] Jean-Philippe Bouchaud, Doyne Farmer, and Fabrizio Lillo. How markets slowly digest changes in supply and demand. In Handbook of financial markets: dynamics and evolution, pages 57–160. Elsevier, 2009.

[5] R. Carmona and L. Leal. Optimal execution with quadratic variation inventories. SIAM Journal on Financial Mathematics, 14(3):751–776, 2023.

[6] R. Carmona and K. Webster. The self-financing equation in limit order book markets. Finance and Stochastics, 23(3):729–759, 2019.

[7] Patrick Cheridito. Mixed fractional Brownian motion. Bernoulli, 7(6):913–934, 2001.

[8] Pavel Chigansky and Marina Kleptsyna. A remark on the drift of the mixed fractional Brownian motion. Working paper, 2025.

[9] Carsten Chong and Viktor Todorov. A nonparametric test for rough volatility. Journal of the American Statistical Association, 120(552):2772–2783, 2025.

[10] Khalil Dayri and Mathieu Rosenbaum. Large tick assets: implicit spread and optimal tick size. Market Microstructure and Liquidity, 1(01):1550003, 2015.

[11] Omar El Euch, Masaaki Fukasawa, and Mathieu Rosenbaum. The microstructural foundations of leverage effect and rough volatility. Finance and Stochastics, 22(2):241–280, 2018.

[12] Omar El Euch and Mathieu Rosenbaum. The characteristic function of rough Heston models. Mathematical Finance, 29(1):3–38, 2019.

[13] Masaaki Fukasawa, Tetsuya Takabatake, and Rebecca Westphal. Consistent estimation for fractional stochastic volatility model under high-frequency asymptotics. Mathematical Finance, 32(4):1086–1132, 2022.

[14] Jim Gatheral. No-dynamic-arbitrage and market impact. Quantitative Finance, 10(7):749–759, 2010.

[15] Jim Gatheral, Thibault Jaisson, and Mathieu Rosenbaum. Volatility is rough. Quantitative Finance, 18(6):933–949, 2018.

[16] G. Gripenberg, S.-O. Londen, and O. Staffans. Volterra Integral and Functional Equations, volume 34 of Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge, 1990.

[17] Paolo Guasoni and Marko Weber. Dynamic trading volume. Mathematical Finance, 27(2):313–349, 2017.

[18] Hans Haubold, Arakaparampil Mathai, and Ram Saxena. Mittag-Leffler Functions and Their Applications. Journal of Applied Mathematics, page 298628, 2011.

[19] Ulrich Horst and Wei Xu. The microstructure of stochastic volatility models with self-exciting jump dynamics. The Annals of Applied Probability, 32(6):4568–4610, 2022.

[20] Ulrich Horst, Wei Xu, and Rouyi Zhang. Convergence of heavy-tailed Hawkes processes and the microstructure of rough volatility. arXiv preprint arXiv:2312.08784, 2023.

[21] Gur Huberman and Werner Stanzl. Price manipulation and quasi-arbitrage. Econometrica, 72(4):1247–1275, 2004.

[22] Jean Jacod and Albert N. Shiryaev. Limit Theorems for Stochastic Processes. Springer, Berlin, Heidelberg, 1987.

[23] Thibault Jaisson and Mathieu Rosenbaum. Limit theorems for nearly unstable Hawkes processes. The Annals of Applied Probability, 25(2):600–631, 2015.

[24] Thibault Jaisson and Mathieu Rosenbaum. Rough fractional diffusions as scaling limits of nearly unstable heavy tailed Hawkes processes. The Annals of Applied Probability, 26(5):2860–2882, 2016.

[25] Paul Jusselin and Mathieu Rosenbaum. No-arbitrage implies power-law market impact and rough volatility. Mathematical Finance, 30(4):1309–1336, 2020.

[26] Fabrizio Lillo, Szabolcs Mike, and Doyne Farmer. Theory for long memory in supply and demand. Physical Review E, 71(6):066122, 2005.

[27] Fabrizio Lillo, Szabolcs Mike, and J. Doyne Farmer. Theory for long memory in supply and demand. Physical Review E, 71:066122, 2005.

[28] Andrew W. Lo and Jiang Wang. Trading volume: Definitions, data analysis, and implications of portfolio theory. The Review of Financial Studies, 13(2):257–300, 2000.

[29] Ananth Madhavan, Matthew Richardson, and Mark Roomans. Why do security prices change? a transaction-level analysis of nyse stocks. The Review of Financial Studies, 10(4):1035–1064, 1997.

[30] Guillaume Maitrier and Jean-Philippe Bouchaud. The subtle interplay between square-root impact, order imbalance & volatility: A unifying framework. arXiv preprint arXiv:2506.07711, 2025.

[31] Guillaume Maitrier, Grégoire Loeper, and Jean-Philippe Bouchaud. The subtle interplay between square-root impact, order imbalance & volatility II: An artificial market generator. arXiv preprint arXiv:2509.05065, 2025.

[32] Matthieu Wyart, Jean-Philippe Bouchaud, Julien Kockelkoren, Marc Potters, and Michele Vettorazzo. Relation between bid–ask spread, impact and volatility in order-driven markets. Quantitative finance, 8(1):41–57, 2008.

A Useful results about Hawkes processes

In this section, we summarize some useful results about Hawkes processes with time-varying baseline. The proofs are omitted for conciseness. They can however be easily adapted from the constant baseline case, see for instance [12].

Definition A.1. *A Hawkes process with baseline (or background rate) $\mu : [0, \infty) \to [0, \infty)$ and self-exciting kernel $\varphi : [0, \infty) \to \mathbb{R}$ is a process N adapted to some filtration $(\mathcal{F}_t)_t$ such that the compensator Λ of N has the form $\Lambda_t = \int_0^t \lambda_s \, ds$ where*

$$\lambda_t = \mu_t + \int_0^{t-} \varphi(t-s) \, dN_s.$$

Lemma A.2. *Define $M = N - \Lambda$ and $\psi = \sum_{k \geqslant 1} \varphi^{*k}$ where φ^{*k} stands for the k-fold convolution of φ. Then for any $0 \leqslant t \leqslant T$, we have*

$$\lambda_t = \mu_t + \int_0^t \psi(t-s)\mu_s \, ds + \int_0^{t-} \psi(t-s) dM_s,$$
$$\int_0^t \lambda_s \, ds = \int_0^t \mu_s \, ds + \int_0^t \psi(t-s) \int_0^s \mu_u \, du \, ds + \int_0^t \psi(t-s) M_s \, ds.$$

Lemma A.3. *For any $0 \leqslant t \leqslant T$, we have*

$$\mathbb{E}[\lambda_t] = \mu_t + \int_0^t \psi(t-s)\mu_s \, ds.$$

B Proof of the results of Section 3

B.1 Proof of Theorem 3.1

Consider a standard Hawkes process N^T with same baseline intensity ν^T and kernel φ_0^T as $F^{\pm,T}$. We then define

$$\overline{N}_t^T = \frac{1-a_0^T}{T\nu^T} N_{tT}^T,$$
$$\overline{\Lambda}_t^T = \frac{1-a_0^T}{T\nu^T} \Lambda_{tT}^T,$$
$$\overline{M}_t^T = \Big(\frac{1-a_0^T}{T\nu^T}\Big)^{1/2} M_{tT}^T.$$

The proof is then split into five parts:

- Step 1: We show that the sequence $(\overline{\Lambda}^T)$ is C-tight.
- Step 2: We show that the sequences of martingales $(\overline{X}^T - \overline{\Lambda}^T)$ tends to zero in probability, uniformly on $[0,1]$.
- Step 3: Under Assumptions A and B, the sequence $(\overline{M}^T, \overline{X}^T)$ is tight. Furthermore, if (Z,X) is a limit point of $(\overline{M}^T, \overline{X}^T)$, then Z is a continuous martingale and $[Z,Z]=X$.
- Step 4: We conclude the convergence of the process $(\overline{N}_t^T, \overline{\Lambda}_t^T, \overline{M}_t^T)$ in distribution for the Skorohod topology towards (X,X,Z) where X and Z are given in Theorem 3.1.
- Step 5: We prove the Hölder property for X.

In this paper, we only prove that $\overline{\Lambda}^T$ is tight; the remaining steps can be found in the proof of Theorem 3.1 in [24]. Let us now prove the following lemma.

Lemma B.1. *The sequence $(\overline{\Lambda}^T)$ is C-tight.*

Proof. Let $\psi_0^T = \sum_{k\geqslant 1}(\varphi_0^T)^{*k}$. We know from Lemma A.3 and Assumption A that

$$\mathbb{E}[\lambda_t^T] = \nu^T + \int_0^t \psi_0^T(t-s)\nu^T\, ds \leqslant \nu^T(1+\|\psi_0^T\|_1) \leqslant \frac{\nu^T}{1-a_0^T}$$

and from Assumption A that $\|\psi_0^T\|_1 = (1-a_0^T)^{-1}a_0^T$. This implies

$$\frac{1-a_0^T}{\nu^T}\sup_t \mathbb{E}[\lambda_t^T] \leqslant 1$$

and therefore

$$\mathbb{E}[\overline{X}_1^T] = \mathbb{E}[\overline{\Lambda}_1^T] \leqslant 1.$$

Moreover, since

$$\langle \overline{M}_t^T, \overline{M}_t^T\rangle = \overline{\Lambda}_t^T$$

the Burkholder-Davis-Gundy inequality then ensures

$$\mathbb{E}[\sup_{t\leqslant 1} |\overline{M}_t^T|^2] \leqslant C$$

for a constant $C > 0$. We now prove the tightness of $\overline{\Lambda}^T$. We write

$$\begin{aligned}\overline{\Lambda}_t^T &= \frac{1-a_0^T}{T\nu^T}\Big(\nu^T tT + \int_0^{tT} \psi_0^T(tT-s)s\,ds\nu^T + \int_0^{tT} \psi_0^T(tT-s)M_s^T\,ds\Big)\\ &= \Big((1-a_0^T)t + T(1-a_0^T)\int_0^{tT} \psi_0^T(T(t-s))s\,ds\Big) + \frac{1-a_0^T}{T\nu^T}\int_0^t T\psi_0^T(T(t-s))M_{Ts}^T\,ds.\end{aligned}$$

The authors in [24] prove the uniform convergence of the first term towards the process

$$\int_0^t sf^{\alpha_0,\lambda_0}(t-s)ds,$$

and therefore is tight. We then focus on the second one and we set

$$\begin{aligned}\widetilde{\Lambda}_t^T &= \frac{1-a_0^T}{T\nu^T}\int_0^t T\psi_0^T(T(t-s))M_{Ts}^T\,ds\\ &= \Big(\frac{1-a_0^T}{T\nu^T}\Big)^{1/2}\int_0^t T\psi_0^T(T(t-s))\overline{M}_s^T\,ds\\ &= \Big(\frac{1}{(1-a_0^T)T\nu^T}\Big)^{1/2}\int_0^t \rho^T(t-s)\overline{M}_s^T\,ds\end{aligned}$$

with

$$\rho_0^T(t) = (1-a_0^T)T\psi_0^T(Tt).$$

To prove the tightness of $\widetilde{\Lambda}^T$, we use Theorem 7.3. in [2] which states that $\widetilde{\Lambda}^T$ is tight provided the following two conditions hold:

- For each $\eta > 0$, there exist $a > 0$ such that

$$\limsup_T \mathbb{P}(|\widetilde{\Lambda}_0^T| \geqslant a) \leqslant \eta.$$

- For each $\varepsilon > 0$, we have

$$\lim_{\delta\to 0}\limsup_T \mathbb{P}(\omega(\widetilde{\Lambda}^T;\delta) \geqslant \varepsilon) = 0$$

  where we use the notation

$$\omega(x;\delta) = \sup_{|t-s|\leqslant\delta,0\leqslant s\leqslant t\leqslant 1} |x(t)-x(s)|.$$

for $\delta > 0$. The first condition clearly holds. We prove that $\widetilde{\Lambda}^T$ verifies the second condition. We first write for $0 \leqslant s \leqslant t \leqslant s+\delta \leqslant 1$

$$\begin{aligned}|\widetilde{\Lambda}_t^T - \widetilde{\Lambda}_s^T| &= \Big|\Big(\frac{1}{(1-a_0^T)T\nu^T}\Big)^{1/2}\int_0^t \rho_0^T(t-u)\overline{M}_u^T\,du - \Big(\frac{1}{(1-a_0^T)T\nu^T}\Big)^{1/2}\int_0^s \rho_0^T(s-u)\overline{M}_u^T\,du\Big|\\ &= \Big(\frac{1}{(1-a_0^T)T\nu^T}\Big)^{1/2}\Big|\int_s^t \rho_0^T(t-u)\overline{M}_u^T\,du + \int_0^s (\rho_0^T(t-u)-\rho_0^T(s-u))\overline{M}_u^T\,du\Big|\\ &\leqslant \Big(\frac{1}{(1-a_0^T)T\nu^T}\Big)^{1/2}\Big(\int_s^t \rho_0^T(t-u)\,du + \int_0^s |\rho_0^T(t-u)-\rho_0^T(s-u)|\,du\Big)\sup_{u\leqslant 1}|\overline{M}_u^T|.\end{aligned}$$

Under Assumption A, the kernel φ_0 is completely monotone and it follows from Theorem 5.4 in [16] that ρ^T is decreasing. Since $|t-s| \leqslant \delta$, we get

$$\begin{aligned}
|\widetilde{\Lambda}_t^T - \widetilde{\Lambda}_s^T| &\leqslant \Big(\frac{1}{(1-a_0^T)T\nu^T}\Big)^{1/2}\Big(\int_0^\delta \rho_0^T(u)\,du + \int_0^s |\rho_0^T(t-s+u) - \rho_0^T(u)|\,du\Big)\sup_{u\leqslant 1}|\overline{M}_u^T| \\
&\leqslant \Big(\frac{1}{(1-a_0^T)T\nu^T}\Big)^{1/2}\Big(\int_0^\delta \rho_0^T(u)\,du + \int_0^s \rho_0^T(u)\,du - \int_{t-s}^t \rho_0^T(u)\,du\Big)\sup_{u\leqslant 1}|\overline{M}_u^T| \\
&\leqslant 2\Big(\frac{1}{(1-a_0^T)T\nu^T}\Big)^{1/2}\int_0^\delta \rho_0^T(u)\,du \sup_{u\leqslant 1}|\overline{M}_u^T|.
\end{aligned}$$

Using Markov's inequality, we deduce that

$$\mathbb{P}(\omega(\widetilde{\Lambda}^T;\delta) \geqslant \varepsilon) \leqslant 2\varepsilon^{-1}\Big(\frac{1}{(1-a_0^T)T\nu^T}\Big)^{1/2}\int_0^\delta \rho_0^T(u)\,du\,\mathbb{E}[\sup_{u\leqslant 1}|\overline{M}_u^T|] \leqslant C'\int_0^\delta \rho_0^T(u)\,du$$

for some positive constant C' and we conclude using

$$\lim_{\delta\to 0}\limsup_{T\to\infty}\int_0^\delta \rho_0^T(u)\,du = 0$$

Furthermore, since the maximum jump size of $\overline{\Lambda}^T$, that is $(1-a_T)(T\nu^T)^{-1}$, goes to zero, we conclude that $\overline{\Lambda}^T$ is C-tight using Proposition VI-3.26 in [22]. The rest of the proof can be found in [24].

B.2 Proof of Proposition 3.2

Suppose $\alpha_0 < 1/2$ with the convention $f^{\alpha_0,\lambda_0}(u) = 0$ for $u \leqslant 0$. We define the forward increment operator $\Delta_h f(t) := f(t+h) - f(t)$ for $t, h > 0$ and (X, Z) to denote either (F^+, Z^+) or (F^-, Z^-). We set

$$V(t,h) := \mathbb{E}\big[(X_{t+h} - X_t)^2\big].$$

Proving Proposition 3.2 is equivalent to proving that

$$V(t,h) = O(h^{4\alpha_0}).$$

We first decompose $X_t = g(t) + \widehat{X}_t$ where

$$g(t) := \mathbb{E}[X_t] = \int_0^t s\, f^{\alpha_0,\lambda_0}(t-s)\,\mathrm{d}s,$$

and

$$\widehat{X}_t = \int_0^t f^{\alpha_0,\lambda_0}(t-s) Z_s\,\mathrm{d}s.$$

With these notations, we obtain

$$V(t,h) = \mathbb{E}\Big[\big(\Delta_h g(t) + \Delta_h \widehat{X}_t\big)^2\Big] = \big(\Delta_h g(t)\big)^2 + \mathbb{E}\big[(\Delta_h \widehat{X}_t)^2\big]$$

since $\mathbb{E}[\Delta_h \widehat{X}_t] = \Delta_h \mathbb{E}[\widehat{X}_t] = 0$. Note that

$$\begin{aligned}\Delta_h \widehat{X}_t &= \int_0^{t+h} f^{\alpha_0,\lambda_0}(t+h-s) Z_s \, \mathrm{d}s - \int_0^t f^{\alpha_0,\lambda_0}(t-s) Z_s \, \mathrm{d}s \\ &= \int_0^{t+h} \big(f^{\alpha_0,\lambda_0}(t+h-s) - f^{\alpha_0,\lambda_0}(t-s) \big) Z_s \, \mathrm{d}s \\ &= \int_0^{t+h} \Delta_h f^{\alpha_0,\lambda_0}(t-s) \, Z_s \, \mathrm{d}s.\end{aligned}$$

Thus, we write

$$\begin{aligned}\mathbb{E}\big[(\Delta_h \widehat{X}_t)^2\big] &= \int_0^{t+h}\int_0^{t+h} \Delta_h f^{\alpha_0,\lambda_0}(t-s) \, \Delta_h f^{\alpha_0,\lambda_0}(t-v) \, \mathbb{E}[Z_s Z_v] \, \mathrm{d}s\mathrm{d}v \\ &= 2\int_0^{t+h} g(s) \, \Delta_h f^{\alpha_0,\lambda_0}(t-s) \left(\int_s^{t+h} \Delta_h f^{\alpha_0,\lambda_0}(t-v) \, \mathrm{d}v \right) \mathrm{d}s.\end{aligned}$$

We introduce

$$\varrho^{\alpha_0,\lambda_0}(x) := \int_0^x f^{\alpha_0,\lambda_0}(y) \, \mathrm{d}y, \qquad x \geqslant 0.$$

so that we have $\int_s^{t+h} \Delta_h f^{\alpha_0,\lambda_0}(t-v) \, \mathrm{d}v = \Delta_h \varrho^{\alpha_0,\lambda_0}(t-s)$, and thus

$$\mathbb{E}\big[(\Delta_h \widehat{X}_t)^2\big] = 2\int_0^{t+h} g(s) \, \Delta_h f^{\alpha_0,\lambda_0}(t-s) \, \Delta_h \varrho^{\alpha_0,\lambda_0}(t-s) \, \mathrm{d}s.$$

Hence,

$$\begin{aligned}V(t,h) &= \big(\Delta_h g(t)\big)^2 + 2\int_0^{t+h} g(s) \, \Delta_h f^{\alpha_0,\lambda_0}(t-s) \, \Delta_h \varrho^{\alpha_0,\lambda_0}(t-s) \, \mathrm{d}s \\ &= \big(\Delta_h g(t)\big)^2 + 2\int_0^{t} g(t-s) \, \Delta_h f^{\alpha_0,\lambda_0}(s) \, \Delta_h \varrho^{\alpha_0,\lambda_0}(s) \, \mathrm{d}s \\ &\quad + 2\int_t^{t+h} g(s) \, \Delta_h f^{\alpha_0,\lambda_0}(t-s) \, \Delta_h \varrho^{\alpha_0,\lambda_0}(t-s) \, \mathrm{d}s.\end{aligned} \tag{4}$$

We would like to bound g. We have for $0 \leqslant t \leqslant 1$ and $0 < h \leqslant 1-t$,

$$|g(t)| = \Big| \int_0^t s f^{\alpha_0,\lambda_0}(t-s) ds \Big| \leqslant |\varrho^{\alpha_0,\lambda_0}(t)| \leqslant 1,$$

and

$$\begin{aligned}|\Delta_h g(t)| &= \Big| \int_0^{t+h} s f^{\alpha_0,\lambda_0}(t+h-s) ds - \int_0^t s f^{\alpha_0,\lambda_0}(t-s) ds \Big| \\ &= \Big| \int_0^{t+h} (t+h-s) f^{\alpha_0,\lambda_0}(s) ds - \int_0^t (t-s) f^{\alpha_0,\lambda_0}(s) ds \Big| \\ &\leqslant h \Big| \int_0^t f^{\alpha_0,\lambda_0}(s) ds \Big| + \Big| \int_t^{t+h} (t+h-s) f^{\alpha_0,\lambda_0}(s) ds \Big|\end{aligned}$$

$$\leqslant h\varrho^{\alpha_0,\lambda_0}(t) + h|\varrho^{\alpha_0,\lambda_0}(t+h) - \varrho^{\alpha_0,\lambda_0}(t)|$$
$$\leqslant h.$$

In particular,

$$\big(\Delta_h g(t)\big)^2 = O(h^2).$$

Furthermore, for $s \in [t, t+h]$,

$$\Delta_h f^{\alpha_0,\lambda_0}(t-s) = f^{\alpha_0,\lambda_0}(t+h-s) \qquad \text{and} \qquad \Delta_h \varrho^{\alpha_0,\lambda_0}(t-s) = \varrho^{\alpha_0,\lambda_0}(t+h-s).$$

By the mean-value theorem, for each $t \leqslant s$, we can write $g(s) = g(t) + g'(\xi_t(s))(s-t)$ for some $t \leqslant \xi_t(s) \leqslant s$. Therefore, the last term of (4) becomes

$$\int_t^{t+h} g(s)\,\Delta_h f^{\alpha_0,\lambda_0}(t-s)\,\Delta_h \varrho^{\alpha_0,\lambda_0}(t-s)\,\mathrm{d}s = g(t)\int_t^{t+h} f^{\alpha_0,\lambda_0}(t+h-s)\varrho^{\alpha_0,\lambda_0}(t+h-s)\,\mathrm{d}s$$
$$+ \int_t^{t+h} g'(\xi_t(s))(s-t)\,f^{\alpha_0,\lambda_0}(t+h-s)\varrho^{\alpha_0,\lambda_0}(t+h-s)\,\mathrm{d}s.$$

A change of variables gives

$$\int_t^{t+h} f^{\alpha_0,\lambda_0}(t+h-s)\varrho^{\alpha_0,\lambda_0}(t+h-s)\,\mathrm{d}s = \int_0^h f^{\alpha_0,\lambda_0}(v)\varrho^{\alpha_0,\lambda_0}(v)\,\mathrm{d}v = \tfrac{1}{2}\varrho^{\alpha_0,\lambda_0}(h)^2.$$

Because $\Delta_h g$ is bounded, g' is also bounded on $[0,1]$. We obtain

$$\Big|\int_t^{t+h} g'(\xi_t(s))(s-t)\,f^{\alpha_0,\lambda_0}(t+h-s)\varrho^{\alpha_0,\lambda_0}(t+h-s)\,\mathrm{d}s\Big|$$
$$\leqslant C'h\int_t^{t+h} f^{\alpha_0,\lambda_0}(t+h-s)\varrho^{\alpha_0,\lambda_0}(t+h-s)\,\mathrm{d}s$$
$$= \frac{C'h}{2}\varrho^{\alpha_0,\lambda_0}(h)^2$$

for some constant $C' > 0$. Consequently,

$$\int_t^{t+h} g(s)\,\Delta_h f^{\alpha_0,\lambda_0}(t-s)\,\Delta_h \varrho^{\alpha_0,\lambda_0}(t-s)\,\mathrm{d}s = \frac{1}{2}\varrho^{\alpha_0,\lambda_0}(h)^2\,\big(1+O(h)\big).$$

We now explicit the behavior of $\varrho^{\alpha_0,\lambda_0}(h)$ as $h \to 0$

Lemma B.2. *As h goes to* 0*, we have*

$$\varrho^{\alpha_0,\lambda_0}(h) = \frac{\lambda_0}{\Gamma(\alpha_0+1)}h^{\alpha_0} + O(h^{2\alpha_0}).$$

Proof. We recall that the Mittag–Leffler density satisfies

$$f^{\alpha_0,\lambda_0}(t) = \lambda_0 t^{\alpha_0-1}E_{\alpha_0,\alpha_0}(-\lambda_0 t^{\alpha_0}), \qquad \varrho^{\alpha_0,\lambda_0}(t) = \lambda_0 t^{\alpha_0}E_{\alpha_0,\alpha_0+1}(-\lambda_0 t^{\alpha_0}).$$

Using the series expansion of the Mittag–Leffler function

$$E_{\alpha_0,\beta}(x) = \sum_{n=0}^{\infty} \frac{x^n}{\Gamma(\alpha_0 n + \beta)},$$

we obtain

$$\varrho^{\alpha_0,\lambda_0}(t) = \lambda_0 t^{\alpha_0} \sum_{n=0}^{\infty} \frac{(-\lambda_0 t^{\alpha_0})^n}{\Gamma(\alpha_0 n + \alpha_0 + 1)} = \sum_{n=1}^{\infty} \frac{(-\lambda_0 t^{\alpha_0})^n}{\Gamma(\alpha_0 n + 1)} = 1 - E_{\alpha_0,1}(-\lambda_0 t^{\alpha_0}).$$

From the power series expansion of $E_{\alpha_0,1}$ around 0,

$$E_{\alpha_0,1}(-\lambda_0 h^{\alpha_0}) = 1 - \frac{\lambda_0}{\Gamma(\alpha_0+1)} h^{\alpha_0} + \frac{\lambda_0^2}{\Gamma(2\alpha_0+1)} h^{2\alpha_0} + O(h^{3\alpha_0}),$$

we obtain

$$\varrho^{\alpha_0,\lambda_0}(h) = \frac{\lambda_0}{\Gamma(\alpha_0+1)} h^{\alpha_0} + O(h^{2\alpha_0}), \qquad h \to 0.$$

□

We are interested now in the second term of (4). Let

$$H(t) := E_{\alpha_0,1}(-\lambda_0 t^{\alpha_0}).$$

Then $H \in C^1(\mathbb{R}_+)$, H is continuous on $(0,\infty)$ and decreasing, and

$$f^{\alpha_0,\lambda_0}(t) = -H'(t), \qquad \varrho^{\alpha_0,\lambda_0}(t) = 1 - H(t).$$

Consider

$$I(t,h) := \int_0^t g(t-s)\, \Delta_h f^{\alpha_0,\lambda_0}(s)\, \Delta_h \varrho^{\alpha_0,\lambda_0}(s)\, \mathrm{d}s = \int_0^t g(t-s)\, (\Delta_h H')(s)\, (\Delta_h H)(s)\, \mathrm{d}s,$$

Integrating by parts, we have

$$\begin{aligned} I(t,h) &= \left[g(t-s)\, \frac{(\Delta_h H(s))^2}{2} \right]_{s=0}^{s=t} - \frac{1}{2} \int_0^t g'(t-s)\, (\Delta_h H(s))^2\, \mathrm{d}s \\ &= -\frac{1}{2} g(t)\, (\Delta_h H(0))^2 - \frac{1}{2} \int_0^t \varrho^{\alpha_0,\lambda_0}(t-s)\, (\Delta_h H(s))^2\, \mathrm{d}s, \end{aligned}$$

where we used $g(0) = 0$ and $g' = \varrho^{\alpha_0,\lambda_0}$. Since $(\Delta_h H(0))^2 = (\Delta_h \varrho^{\alpha_0,\lambda_0}(0))^2 = \varrho^{\alpha_0,\lambda_0}(h)^2$, the first term equals $-\frac{1}{2} g(t) \varrho^{\alpha_0,\lambda_0}(h)^2$. Moreover, for all $\delta > 0$, we have

$$\begin{aligned} \int_0^t \varrho^{\alpha_0,\lambda_0}(t-s)\, (\Delta_h H(s))^2\, \mathrm{d}s &\leqslant \varrho^{\alpha_0,\lambda_0}(t) \int_0^t (\Delta_h H(s))^2\, \mathrm{d}s \\ &\leqslant \varrho^{\alpha_0,\lambda_0}(t) \left\{ \int_0^\delta (\Delta_h H(s))^2\, \mathrm{d}s + \int_\delta^t (\Delta_h H(s))^2\, \mathrm{d}s \right\}. \end{aligned}$$

Using that H is bounded and for $u, h > 0$

$$(\Delta_h H(s))^2 = \Big(\int_s^{s+h} H'(v)dv\Big)^2 = \Big(\int_s^{s+h} f^{\alpha_0,\lambda_0}(v)dv\Big)^2 \leqslant h^2 (f^{\alpha_0,\lambda_0}(s))^2.$$

Moreoever, since $\alpha_0 < 1/2$, $E_{\alpha_0,\alpha_0}(-s) \leqslant 1$ for any positive s, and we have that

$$(f^{\alpha_0,\lambda_0}(s))^2 \leqslant \lambda_0^2 s^{2\alpha_0-2}.$$

Therefore, we write

$$\begin{aligned}
\int_0^t \varrho^{\alpha_0,\lambda_0}(t-u)\,(\Delta_h H(u))^2\,\mathrm{d}u &\leqslant \varrho^{\alpha_0,\lambda_0}(t)\left\{C\,\delta + h^2 \int_\delta^t (f^{\alpha_0,\lambda_0}(u))^2\,\mathrm{d}u\right\} \\
&\leqslant \varrho^{\alpha_0,\lambda_0}(t)\left\{C\,\delta + h^2 \int_\delta^\infty (f^{\alpha_0,\lambda_0}(u))^2\,\mathrm{d}u\right\} \\
&\leqslant \varrho^{\alpha_0,\lambda_0}(t)\left\{C\,\delta + h^2 \int_\delta^\infty \lambda_0^2 u^{2\alpha_0-2}\,\mathrm{d}u\right\} \\
&= \varrho^{\alpha_0,\lambda_0}(t)\left\{C\,\delta + h^2\,\lambda_0^2\,\frac{\delta^{2\alpha_0-1}}{1-2\alpha_0}\right\},
\end{aligned}$$

which is finite since $\alpha_0 < 1/2$. Choosing $\delta = h^{1/(1-\alpha_0)}$ balances the two terms, giving

$$\int_0^t \varrho^{\alpha_0,\lambda_0}(t-u)\,(\Delta_h H(u))^2\,\mathrm{d}u \leqslant C'\,\varrho^{\alpha_0,\lambda_0}(t)\,h^{1/(1-\alpha_0)}.$$

Using $\alpha_0(1-\alpha_0) \leqslant 1/4$ we have $\frac{1}{1-\alpha_0} \geqslant 4\alpha_0$, hence for $h \in (0, 1]$,

$$h^{1/(1-\alpha_0)} \leqslant h^{4\alpha_0}.$$

Therefore,

$$I(t,h) = \frac{1}{2} g(t) \varrho^{\alpha_0,\lambda_0}(h)^2 + O\big(\varrho^{\alpha_0,\lambda_0}(t)\,h^{4\alpha_0}\big).$$

Using Lemma (B.2), we obtain

$$I(t,h) = \frac{g(t)\,\lambda_0^2}{2\,\Gamma(\alpha_0+1)^2}\,h^{2\alpha_0} + O\big(h^{4\alpha_0 \wedge 1}\big).$$

Hence, going back to (4), we write

$$\begin{aligned}
V(t,h) &= \big(\Delta_h g(t)\big)^2 + 2I(t,h) + 2\int_t^{t+h} g(u)\,\Delta_h f^{\alpha_0,\lambda_0}(t-u)\,\Delta_h \varrho^{\alpha_0,\lambda_0}(t-u)\,\mathrm{d}u \\
&= O(h^2) + 2\cdot \frac{g(t)\,\lambda_0^2}{2\,\Gamma(\alpha_0+1)^2}\,h^{2\alpha_0} + 2\cdot\frac{1}{2}\varrho^{\alpha_0,\lambda_0}(h)^2\,\big(1+O(h)\big) + O\big(h^{4\alpha_0}\big) \\
&= \frac{2\lambda_0^2}{\Gamma(\alpha_0+1)^2}\,(1+g(t))\,h^{2\alpha_0} + O\big(h^{4\alpha_0\wedge 1}\big) + O(h^2).
\end{aligned}$$

Since $2\alpha_0 < 1$, the remainder $O(h^2)$ is negligible with respect to $h^{2\alpha_0}$ as h goes to 0.
In summary, for $\alpha_0 \in (0, \frac{1}{2})$ the following holds uniformly for $t, h > 0$, as h tends to 0,

$$V(t,h) = \mathbb{E}\big[(X_{t+h} - X_t)^2\big] = \frac{2\lambda_0^2}{\Gamma(\alpha_0+1)^2}\,(1+g(t))\,h^{2\alpha_0} + O\big(h^{4\alpha_0\wedge 1}\big).$$

We conclude that F^+ and F^- are exactly $2\alpha_0$–Hölder continuous in L^2. $\square$

B.3 Proof of Proposition 3.3

From Theorem 3.1 we can see that

$$F_t = F_t^+ + F_t^- = 2\int_0^t s\, f^{\alpha_0,\lambda_0}(t-s)\, ds + \frac{1}{\sqrt{\mu_0\,\lambda_0}}\int_0^t f^{\alpha_0,\lambda_0}(t-s)\, Z_s^F\, ds$$

where $Z^F = Z^+ + Z^-$. Notice that Z^F is a continuous martingale with quadratic variation F. On the other hand, the process V satisfies

$$V_t = F_t^+ - F_t^- = \frac{1}{\sqrt{\mu_0\,\lambda_0}}\int_0^t f^{\alpha_0,\lambda_0}(t-s)\, Z_s^V\, ds$$

where $Z^V = Z^+ - Z^-$. Note also that Z^V is a continuous martingale with quadratic variation F. We can compute the quadratic covariance of the two resulting martingales

$$< Z^V, Z^F > = < Z^+ - Z^-, Z^+ + Z^- > = F^+ - F^- = V.$$

B.4 Proof of Theorem 3.4

The proof relies on replicating the findings of [11] with the stochastic time-varying baseline $\boldsymbol{\mu}^T$. We start by providing multiple elements needed for the proof.
First, note that we have

$$\int_0^t \boldsymbol{\lambda}_s^T\, ds = \int_0^t \boldsymbol{\mu}_s^T\, ds + \int_0^t \psi^T(t-s)\cdot \int_0^s \boldsymbol{\mu}_u^T\, du\, ds + \int_0^t \psi^T(t-s)\cdot \mathbf{M}_s^T\, ds.$$

Now, $\boldsymbol{\mu}^T = \phi^T * d\mathbf{F}^T$ and since $\mathbf{F}_0^T = \mathbf{0}$ we have

$$\int_0^t \boldsymbol{\mu}_s^T\, ds = \phi^T * \mathbf{F}_t^T.$$

Using also the identity $\psi^T * \phi^T = \psi^T - \phi^T$, we obtain

$$\int_0^t \boldsymbol{\lambda}_s^T\, ds = \int_0^t \psi^T(t-s)\cdot \mathbf{F}_s^T\, ds + \int_0^t \psi^T(t-s)\cdot \mathbf{M}_s^T\, ds.$$

In this setting, it is more suitable to work with the 2-dimensional rescaled processes

$$\begin{aligned}
\overline{\mathbf{N}}_t^T &= \frac{(1-a_0^T)(1-a_1^T)}{T\nu^T}\mathbf{N}_{tT}^T,\\
\overline{\boldsymbol{\Lambda}}_t^T &= \frac{(1-a_0^T)(1-a_1^T)}{T\nu^T}\boldsymbol{\Lambda}_{tT}^T,\\
\overline{\mathbf{M}}_t^T &= \Big(\frac{(1-a_0^T)(1-a_1^T)}{T\nu^T}\Big)^{1/2}\mathbf{M}_{tT}^T.
\end{aligned}$$

The scaled unsigned reaction flow is then given by

$${}^{\mathrm{t}}v_1 \cdot \overline{\mathbf{N}}^T = \overline{N}^{+,T} + \overline{N}^{-,T},$$

and the scaled signed reaction flow by

$$ {}^{\mathrm{t}}v_2 \cdot \overline{\mathbf{N}}^T = \overline{N}^{+,T} - \overline{N}^{-,T}. $$

We can then write

$$ \begin{aligned} \overline{\mathbf{\Lambda}}_t^T &= \frac{(1-a_0^T)(1-a_1^T)}{T\nu^T} \int_0^{tT} \boldsymbol{\lambda}_s^T \, ds \\ &= \int_0^t T(1-a_1^T)\psi^T(T(t-s)) \cdot \overline{\mathbf{F}}_s^T \, ds + \frac{1-a_0^T}{T\nu^T} \int_0^t T(1-a_1^T)\psi^T(T(t-s)) \cdot \mathbf{M}_{sT}^T \, ds. \end{aligned} $$

Note that

$$ \mathbb{E}[\overline{\mathbf{\Lambda}}_t^T] = \int_0^t T(1-a_1^T)\psi^T(T(t-s)) \cdot \mathbb{E}[\overline{\mathbf{F}}_s^T] \, ds, $$

and from Section B.1, we know that $\mathbb{E}[\overline{\mathbf{F}}_s^{\pm,T}] \leqslant 1$, then

$$ {}^{\mathrm{t}}v_1 \cdot \mathbb{E}[\overline{\mathbf{\Lambda}}_t^T] \leqslant T(1-a_1^T)\, {}^{\mathrm{t}}v_1 \cdot \Big(\int_0^t \psi^T(T(t-s)) ds \Big) \cdot v_1 \leqslant (1-a_1^T)\varrho \left(\int_0^\infty \psi^T(s) ds \right) < 1. $$

Therefore, using the Burkholder-Davis-Gundy inequality from **??**, we get that

$$ \mathbb{E}\Big[\sup_{t \leqslant 1} \big\| \overline{\mathbf{M}}_t^T \big\|_2^2 \Big] \leqslant C $$

for some constant $C > 0$.

For $i = 1, 2$, v_i is the eigenvector associated with the eigenvalue k_i so we have

$$ \phi^T \cdot v_i = k_i^T v_i. $$

By induction,

$$ v_i^T \cdot (\phi^T)^{*n} = (k_i^T)^{*n} \, v_i^T, $$

and we define scalar kernels

$$ \psi_i^T(x) = \sum_{n \geqslant 1} (a_1^T)^n (k_i^T)^{*n}(x), \quad \rho_i^T(x) = T(1-a_1^T)\psi_i^T(Tx), \quad \varrho_i^T(t) = \int_0^t \rho_i^T(s) \, ds. $$

Consequently, we have ${}^{\mathrm{t}}v_i \cdot \psi^T = \psi_i^T \, {}^{\mathrm{t}}v_i$ and

$$ {}^{\mathrm{t}}v_i \cdot \overline{\mathbf{\Lambda}}_t^T = \int_0^t \rho_i^T(t-s) \, {}^{\mathrm{t}}v_i \cdot \overline{\mathbf{F}}_s^T \, ds + c^T \int_0^t \rho_i^T(t-s) \, {}^{\mathrm{t}}v_i \cdot \overline{\mathbf{M}}_s^T \, ds $$

where

$$ c^T = \sqrt{(1-a_0^T)/(T\nu^T(1-a_1^T))} \to \sqrt{\frac{1}{\lambda_1 \mu_1}}. $$

We are interested in studying the convergence of this process for $i \in \{1, 2\}$.

Convergence of ${}^{\mathrm{t}}v_i \cdot \overline{\mathbf{N}}_t^T - {}^{\mathrm{t}}v_i \cdot \overline{\mathbf{\Lambda}}_t^T$. We have

$$ {}^{\mathrm{t}}v_i \cdot \overline{\mathbf{N}}_t^T - {}^{\mathrm{t}}v_i \cdot \overline{\mathbf{\Lambda}}_t^T = \Big(\frac{(1-a_0^T)(1-a_1^T)}{T\nu^T}\Big)^{1/2} {}^{\mathrm{t}}v_i \cdot \overline{\mathbf{M}}_t^T $$

Using Doob's inequality, we get

$$ \begin{aligned} \mathbb{E}\Big[\sup_{t\leqslant 1}\Big|{}^{\mathrm{t}}v_i \cdot \overline{\mathbf{N}}_t^T - {}^{\mathrm{t}}v_i \cdot \overline{\mathbf{\Lambda}}_t^T\Big|^2\Big] &\leqslant 4\Big(\frac{(1-a_0^T)(1-a_1^T)}{T\nu^T}\Big)\mathbb{E}[|\,{}^{\mathrm{t}}v_i \cdot \overline{\mathbf{M}}_T^T|^2] \\ &\leqslant 4\Big(\frac{(1-a_0^T)(1-a_1^T)}{T\nu^T}\Big)\big\|v_i\big\|^2\mathbb{E}[\big\|\overline{\mathbf{M}}_T^T\big\|_2^2] \\ &\leqslant C'\Big(\frac{(1-a_0^T)(1-a_1^T)}{T\nu^T}\Big). \end{aligned} $$

Since $(1-a_0^T)(1-a_1^T)(T\nu^T)^{-1}$ is of the order of $T^{-2\alpha_1}$, we obtain the convergence to zero in L^2 and in probability of ${}^{\mathrm{t}}v_i \cdot \overline{\mathbf{N}}_t^T - {}^{\mathrm{t}}v_i \cdot \overline{\mathbf{\Lambda}}_t^T$.

Convergence of ${}^{\mathrm{t}}v_2 \cdot \overline{\mathbf{N}}_t^T$. We know from [11] that ϱ_2^T converges uniformly to zero and we write

$$ \mathbb{E}\Big[\Big|\int_0^t \rho_2^T(t-s)\,{}^{\mathrm{t}}v_2 \cdot \overline{\mathbf{F}}_s^T\,ds\Big|\Big] \leqslant \|v_2\| \sup_{t\leqslant 1}\mathbb{E}[\|\overline{\mathbf{F}}_t^T\|]\varrho_2(t). $$

Using the fact that $\mathbb{E}[\|\overline{\mathbf{F}}_t^T\|]$ is bounded, which has been proven in Section B.1, we conclude the first integral of ${}^{\mathrm{t}}v_2 \cdot \overline{\mathbf{N}}_t^T$ converges to zero in L^1. For the second integral, we write using an integration by part

$$ c^T\int_0^t \rho_2^T(t-s)\,{}^{\mathrm{t}}v_2 \cdot \overline{\mathbf{M}}_s^T\,ds = c^T\int_0^t \varrho_2^T(t-s)({}^{\mathrm{t}}v_2 \cdot \overline{\mathbf{M}}_s^T). $$

Thus, there exists a constant C' such that

$$ \mathbb{E}\Big[\Big(c^T\int_0^t \rho_2^T(t-s)\,{}^{\mathrm{t}}v_2 \cdot \overline{\mathbf{M}}_s^T\,ds\Big)^2\Big] \leqslant C'\int_0^t (\varrho_2^T(s))^2 ds $$

and therefore the second integral converges to 0 in L^2. Therefore we obtain that ${}^{\mathrm{t}}v_2 \cdot \overline{\mathbf{\Lambda}}_t^T$ goes to zero in L^1. Consequently, ${}^{\mathrm{t}}v_2 \cdot \overline{\mathbf{N}^T}$ converges to 0 in probability and in L^1.

Convergence of ${}^{\mathrm{t}}v_1 \cdot \overline{\mathbf{N}}_t^T$. We write

$$ {}^{\mathrm{t}}v_1 \cdot \overline{\mathbf{\Lambda}}_t^T = \int_0^t \rho_1^T(t-s)\,{}^{\mathrm{t}}v_1 \cdot \overline{\mathbf{F}}_s^T\,ds + c^T\int_0^t \rho_1^T(t-s)\,{}^{\mathrm{t}}v_1 \cdot \overline{\mathbf{M}}_s^T\,ds $$

Using the same arguments and methodology as in Section B.1, we get that ${}^{\mathrm{t}}v_1 \cdot \overline{\mathbf{\Lambda}}^T$ is C-tight and we conclude that $({}^{\mathrm{t}}v_1 \cdot \overline{\mathbf{N}}^T, {}^{\mathrm{t}}v_1 \cdot \overline{\mathbf{\Lambda}}^T, {}^{\mathrm{t}}v_1 \cdot \overline{\mathbf{M}}^T)$ is C-tight. Furthermore, if (X, X, Z) is a limit point of $({}^{\mathrm{t}}v_1 \cdot \overline{\mathbf{N}}^T, {}^{\mathrm{t}}v_1 \cdot \overline{\mathbf{\Lambda}}^T, {}^{\mathrm{t}}v_1 \cdot \overline{\mathbf{M}}^T)$, then Z is a continuous martingale with $[Z, Z] = X$.

Moreover, we know from [24] that the sequence of measures with density $\rho_1^T(x)$ converges weakly towards the measure with density $\lambda_1 x^{\alpha_1 - 1} E_{\alpha_1,\alpha_1}(-\lambda_1 x_1^{\alpha})$. In particular, over $[0, 1]$,

$$\varrho_1^T(t) = \int_0^t \rho_1^T(x)dx$$

converges uniformly towards

$$\varrho^{\alpha_1,\lambda_1}(t) = \int_0^t f^{\alpha_1,\lambda_1}(x)dx.$$

Therefore, using the same approach as in [24] yields

$$\int_0^t \rho_1^T(t-s)\, {}^{\mathrm{t}}v_1 \cdot \overline{\mathbf{F}}_s^T\, ds \to \int_0^t f^{\alpha_1,\lambda_1}(t-s)F_s\, ds$$
$$c^T \int_0^t \rho_i^T(t-s)\, {}^{\mathrm{t}}v_i \cdot \overline{\mathbf{M}}_s^T\, ds \to \sqrt{\frac{1}{\lambda_1\mu_1}} \int_0^t f^{\alpha_1,\lambda_1}(t-s)Z_s\, ds$$

where F is the scaling limit of ${}^{\mathrm{t}}v_1 \cdot \overline{\mathbf{F}}^T$ from Proposition 3.3.

Convergence of $(\overline{N}^{+,T}, \overline{N}^{-,T}, \overline{\Lambda}^{+,T}, \overline{\Lambda}^{-,T}, \overline{M}^{+,T}, \overline{M}^{-,T})$. We use the fact that the sum process $({}^{\mathrm{t}}v_1.\overline{\mathbf{N}}^T, {}^{\mathrm{t}}v_1.\overline{\mathbf{\Lambda}}^T, {}^{\mathrm{t}}v_1.\overline{\mathbf{M}}^T)$ is C-tight, which implies the C-tightness of the process $(\overline{N}^{+,T}, \overline{N}^{-,T}, \overline{\Lambda}^{+,T}, \overline{\Lambda}^{-,T}, \overline{M}^{+,T}, \overline{M}^{-,T})$. Furthermore, using the same arguments as in Section B, the previous result, and the fact that

$$\overline{N}^{+,T} = \frac{1}{2}({}^{\mathrm{t}}v_1.\overline{\mathbf{N}}^T + {}^{\mathrm{t}}v_2.\overline{\mathbf{N}}^T) \quad \text{and} \quad \overline{N}^{-,T} = \frac{1}{2}({}^{\mathrm{t}}v_1.\overline{\mathbf{N}}^T - {}^{\mathrm{t}}v_2.\overline{\mathbf{N}}^T),$$

if (X, X, X, X, Z^+, Z^-) is an accumulation point of $(\overline{N}^{+,T}, \overline{N}^{-,T}, \overline{\Lambda}^{+,T}, \overline{\Lambda}^{-,T}, \overline{M}^{+,T}, \overline{M}^{-,T})$, then

$$X_t = \frac{1}{2}\int_0^t f^{\alpha_1,\lambda_1}(t-s)F_s\, ds + \frac{1}{2\sqrt{\lambda_1\mu_1}} \int_0^t f^{\alpha_1,\lambda_1}(t-s)Z_s\, ds \qquad \text{and} \qquad Z_t = Z_t^+ + Z_t^-$$

where Z^+ and Z^- are two continuous martingales with quadratic variation X and zero quadratic covariation. Seeing that F is smoother than Z, the regularity of X is determined by that the second integral, which is $(H_1 - \varepsilon)$-Holder continuous for every $\varepsilon > 0$ on $[0, 1]$, with $H_1 = \alpha_1 - 1/2$.

B.5 Proof of Hölder regularity in Theorem 3.4

From the previous section, we know that:

- X is Lipschitz continuous,
- Z is $(1/2 - \varepsilon)$-Hölder continuous for all $\varepsilon > 0$, since its quadratic variation, which is X, is continuous,
- F is $(2\alpha_0 - \varepsilon)$-Hölder continuous for all $\varepsilon > 0$,
- Z^F is $(\alpha_0 - \varepsilon)$-Hölder continuous for all $\varepsilon > 0$.

Then for $0 < \gamma < 1$, we know from Proposition A.1 in [24] that

- X admits a fractional derivative of order γ and $D^\gamma X$ is $(1-\gamma)$-Hölder regular,
- If $\gamma < 2\alpha_0 = \alpha_1$, then F admits a fractional derivative of order γ and $D^\gamma F$ is $(2\alpha_0 - \gamma - \varepsilon)$-Hölder regular for all $\varepsilon > 0$,
- If $\gamma < 1/2$, then Z admits a fractional derivative of order γ and $D^\gamma Z$ is $(1/2 - \gamma - \varepsilon)$-Hölder regular for all $\varepsilon > 0$.

Let $1/2 < \gamma < \alpha_1$. From Proposition 3.1 and Corollary A.2 in [24], we have

$$\begin{aligned} X_t &= \frac{1}{2}\int_0^t f^{\alpha_1,\lambda_1}(t-s)F_s\,ds + \frac{1}{2\sqrt{\lambda_1\mu_1}}\int_0^t f^{\alpha_1,\lambda_1}(t-s)Z_s\,ds \\ &= \frac{1}{2}\int_0^t D^\gamma f^{\alpha_1,\lambda_1}(t-s)I^\gamma F_s\,ds + \frac{1}{2\sqrt{\lambda_1\mu_1}}\int_0^t D^\gamma f^{\alpha_1,\lambda_1}(t-s)I^\gamma Z_s\,ds \end{aligned}$$

Furthermore, F and Z are fractionally differentiable and we have

$$I^\gamma F_s = \int_0^s D^{1-\gamma}F_u du \qquad \text{and} \qquad I^\gamma Z_s = \int_0^s D^{1-\gamma}Z_u du.$$

We rewrite the expression of X as follows

$$X_t = \frac{1}{2}\int_0^t\int_0^s D^\gamma f^{\alpha_1,\lambda_1}(t-s)D^{1-\gamma}F_u du ds + \frac{1}{2\sqrt{\lambda_1\mu_1}}\int_0^t\int_0^s D^\gamma f^{\alpha_1,\lambda_1}(t-s)D^{1-\gamma}Z_u du ds \qquad (5)$$

We use Fubini's theorem and we write

$$\begin{aligned} \int_0^t\int_0^s D^\gamma f^{\alpha_1,\lambda_1}(t-s)D^{1-\gamma}Z_u du ds &= \int_0^t\int_u^t D^\gamma f^{\alpha_1,\lambda_1}(t-s)D^{1-\gamma}Z_u ds du \\ &= \int_0^t\int_u^t D^\gamma f^{\alpha_1,\lambda_1}(s-u)D^{1-\gamma}Z_u ds du \\ &= \int_0^t\int_0^s D^\gamma f^{\alpha_1,\lambda_1}(s-u)D^{1-\gamma}Z_u du ds. \end{aligned}$$

Applying the same computations to the first integral in (5), we get

$$X_t = \int_0^t Y_s ds$$

with

$$Y_s = \frac{1}{2}\int_0^s D^\gamma f^{\alpha_1,\lambda_1}(s-u)D^{1-\gamma}F_u du + \frac{1}{2\sqrt{\lambda_1\mu_1}}\int_0^s D^\gamma f^{\alpha_1,\lambda_1}(s-u)D^{1-\gamma}Z_u du.$$

Since $2\alpha_0 > 1/2$, we know that F is smoother than Z, and thus the regularity of Y is that of its second term. From Propositions 3.1 and A.3 in [24], we have that the second integral has Hölder regularity $(\alpha_1 - \gamma)$ for $1/2 < \gamma < 1$. Thus, for every $\varepsilon > 0$, the second integral, and therefore Y, has Hölder regularity $(\alpha_1 - 1/2 - \varepsilon)$.

B.6 Proof of Theorem 3.5

First, note that $\overline{F}_t^{T,+} + \overline{F}_t^{-,T}$ scales as $T\nu^T(1-a_0^T)^{-1}$. Seeing that $(1-a_1^T)$ is of the same order as $T^{-\alpha_1}$, we conclude that $(1-a_0^T)(1-a_1^T)(T\nu^T)^{-1}(F_{tT}^{T,+} + F_{tT}^{-,T})$ converges to zero. Moreover, Theorem 3.4 ensures that the process

$$\frac{(1-a_0^T)(1-a_1^T)}{T\nu^T}(N_{tT}^{+,T} + N_{tT}^{-,T}) = \overline{\mathbf{N}}^{+,T} + \overline{\mathbf{N}}^{-,T} = {}^{\mathrm{t}}v_1 \cdot \overline{\mathbf{N}}^T$$

is C-tight and it converges in distribution in the Skorohod topology. Therefore, the same applies to $\overline{U}^T$, and if U is a limit of ${}^{\mathrm{t}}v_1 \cdot \overline{\mathbf{N}}^T$, then it is also a limit of $\overline{U}^T$ and it satisfies Equation (2).

B.7 Proof of Theorem 3.6

Notice that on one hand, $(1-a_0^T)(T\nu^T)^{-1}$ is of the same order as $T^{-2\alpha_0} = T^{-\alpha_1}$. But we also know that $(1-a_1^T)$ grows like $T^{-\alpha_1}$. Therefore, we can see that

$$\Big(\frac{(1-a_0^T)(1-a_1^T)}{T\nu^T}\Big)^{1/2} \qquad \text{and} \qquad \frac{1-a_0^T}{T\nu^T}$$

are of the same order. Thus, Theorem 3.1 guarantees that

$$\Big(\frac{(1-a_0^T)(1-a_1^T)}{T\nu^T}\Big)^{1/2}(F_{tT}^{T,+} - F_{tT}^{-,T}) \to V_t$$

where V is given by (1). We just need to compute the limit of

$$\Big(\frac{(1-a_0^T)(1-a_1^T)}{T\nu^T}\Big)^{1/2}(N_{tT}^{+,T} - N_{tT}^{-,T}).$$

We write

$$N_t^{T,+} - N_t^{-,T} = M_t^{T,+} - M_t^{-,T} + \Lambda_t^{T,+} - \Lambda_t^{-,T},$$

and

$$\begin{aligned}
\Lambda_t^{T,+} - \Lambda_t^{-,T} = {}^{\mathrm{t}}e_2\mathbf{\Lambda}_t^T &= \int_0^t \psi_2^T(t-s)\,{}^{\mathrm{t}}v_2 \cdot \mathbf{F}_s^T\,ds + \int_0^t \psi_2^T(t-s)(M_s^{+,T} - M_s^{-,T})\,ds \\
&= \int_0^t \psi_2^T(t-s)\,{}^{\mathrm{t}}v_2 \cdot \mathbf{F}_s^T\,ds + \int_0^t\int_0^{t-s} \psi_2^T(u)\,du\,d(M_s^{+,T} - M_s^{-,T}) \\
&= \int_0^t \psi_2^T(t-s)\,{}^{\mathrm{t}}v_2 \cdot \mathbf{F}_s^T\,ds + \int_0^\infty \psi_2^T(u)\,du(M_t^{+,T} - M_t^{-,T}) \\
&\quad - \int_0^t\int_{t-s}^\infty \psi_2^T(u)\,du\,d(M_s^{+,T} - M_s^{-,T}).
\end{aligned}$$

Therefore we get

$$\begin{aligned}
N_t^{T,+} - N_t^{-,T} = &\int_0^t \psi_2^T(t-s)\,{}^{\mathrm{t}}v_2 \cdot \mathbf{F}_s^T\,ds + (1 + \int_0^\infty \psi_2^T(u)\,du)(M_t^{+,T} - M_t^{-,T}) \\
&- \int_0^t\int_{t-s}^\infty \psi_2^T(u)\,du\,d(M_s^{+,T} - M_s^{-,T}).
\end{aligned}$$

$$= \int_0^t \psi_2^T(t-s)\,{}^{\mathrm{t}}v_2 \cdot \mathbf{F}_s^T\,ds + \frac{1}{1-a_1^T(\|\varphi_1\|_1 - \|\varphi_2\|_1)}(M_t^{+,T} - M_t^{-,T})$$
$$- \int_0^t \int_{t-s}^\infty \psi_2^T(u)\,du\,d(M_s^{+,T} - M_s^{-,T}).$$

After rescaling, we obtain

$$\Big(\frac{(1-a_0^T)(1-a_1^T)}{T\nu^T}\Big)^{1/2}\big(N_{tT}^{+,T} - N_{tT}^{-,T}\big) = \Big(\frac{(1-a_0^T)(1-a_1^T)}{T\nu^T}\Big)^{1/2} \int_0^{tT} \psi_2^T(Tt-s)\,{}^{\mathrm{t}}v_2 \cdot \mathbf{F}_s^T\,ds + \frac{1}{1-a_1^T(\|\varphi_1\|_1 - \|\varphi_2\|_1)}(\overline{M}_t^{+,T} - \overline{M}_t^{-,T}) - R_t^T \tag{6}$$

where

$$R_t^T = \int_0^t \int_{T(t-s)}^\infty \psi_2^T(u)\,du\,d(\overline{M}_s^{+,T} - \overline{M}_s^{-,T}).$$

Following the same argument as in the proof of Theorem 3.2 in [11], we conclude the convergence of R^T to zero in the sense of finite dimensional laws.
Furthermore, from Theorem 3.4, we know that the second term in (6) converges to

$$\frac{1}{1-(\|\varphi_1\|_1 - \|\varphi_2\|_1)}(Z_t^+ - Z_t^-).$$

It remains to study the first term

$$\Big(\frac{(1-a_0^T)(1-a_1^T)}{T\nu^T}\Big)^{1/2} \int_0^{tT} \psi_2^T(Tt-s)\,{}^{\mathrm{t}}v_2 \cdot \mathbf{F}_s^T\,ds.$$

After proper rescaling, we obtain

$$\Big(\frac{(1-a_0^T)(1-a_1^T)}{T\nu^T}\Big)^{1/2} \int_0^{tT} \psi_2^T(tT-s)\,{}^{\mathrm{t}}v_2 \cdot F_s^T\,ds = c^T \int_0^t T\psi_2^T(T(t-s))\,{}^{\mathrm{t}}v_2 \cdot \overline{F}_s^T\,ds$$

where

$$c^T = \sqrt{(T\nu^T(1-a_1^T))/(1-a_0^T)} \to \sqrt{\lambda_1\mu_1}.$$

To understand its asymptotic behavior as T goes to infinity, one can compute the Fourier transform $\widehat{\psi_2^T(T.)}$ of $\psi_2^T(T.)$. We have

$$\widehat{\psi_2^T(T.)}(z) = \int_{x\in\mathbb{R}_+} \psi_2^T(Tx)e^{ixz}dx = \frac{1}{T}\sum_{n\geqslant 1}(a_1^T)^n\Big(\widehat{k}_2(z/T)\Big)^n = \frac{a_1^T\widehat{k}_2(z/T)}{T\Big(1-a_1^T\widehat{k}_2(z/T)\Big)}$$

As T goes to infinity, $\widehat{k}_j(z/T)$ tends to $\|k_2\|_1$ and recall that $\|k_2\|_1 < 1$. Therefore, we see that

$$T\widehat{\psi_2^T(T.)}(z) \to \frac{\|k_2\|_1}{1-\|k_2\|_1} = \frac{\|\varphi_1\|_1 - \|\varphi_2\|_1}{1-(\|\varphi_1\|_1 - \|\varphi_2\|_1)}$$

and consequently, if we define

$$\chi_T(dt) := T\psi_2^T(Tt)dt$$

then we have

$$\chi_T(dt) \to \frac{\|\varphi_1\|_1 - \|\varphi_2\|_1}{1-(\|\varphi_1\|_1 - \|\varphi_2\|_1)}\delta_0(dt)$$

Thus, we have shown that

$$\Big(\frac{(1-a_0^T)(1-a_1^T)}{T\nu^T}\Big)^{1/2}\int_0^{tT}\psi_2^T(Tt-s)\,{}^{\mathrm{t}}v_2\cdot\mathbf{F}_s^T\,ds \to \frac{\sqrt{\lambda_1\mu_1}(\|\varphi_1\|_1 - \|\varphi_2\|_1)}{1-(\|\varphi_1\|_1 - \|\varphi_2\|_1)}V_t.$$

Eventually, we obtain

$$\Big(\frac{(1-a_0^T)(1-a_1^T)}{T\nu^T}\Big)^{1/2}S_{tT}^T \to \frac{\sqrt{\lambda_1\mu_1}(\|\varphi_1\|_1 - \|\varphi_2\|_1)}{1-(\|\varphi_1\|_1 - \|\varphi_2\|_1)}V_t + \frac{1}{1-(\|\varphi_1\|_1 - \|\varphi_2\|_1)}(Z_t^+ - Z_t^-)$$

in the sense of finite dimensional laws.