Rough Volatility Arbitrage under Markov Regime Volterra Process Approach with Double Exponential

Mitchell Scott, Ph.D.
Department of Mathematics
Emory University
mtscot4@emory.edu

Felipe Cardozo
Mathematics & Computer Science
Emory University
focardo@emory.edu

***Abstract*—It is well-established in the empirical literature that the realised variance of equity indices exhibits *rough* behaviour characterised by a Hurst exponent $H \approx 0.07$, substantially lower than the value of 0.5 implied by classical diffusion models. The present work introduces a unified trading framework—termed *Project Rough-Regime*—that (i) prices short-dated volatility derivatives using a discrete Volterra stochastic process whose spot variance is driven by a rough fractional kernel, (ii) classifies the prevailing market regime in real time via a two-state Gaussian Hidden Markov Model (G-HMM) calibrated by the Expectation–Maximisation algorithm, and (iii) routes executions through an Interactive Brokers TWS interface equipped with a passive-execution *chase* algorithm. The combined system constitutes an event-driven "cyborg" loop in which the G-HMM serves as a *traffic light* that suppresses long-volatility exposure whenever the posterior probability of the turbulent state exceeds 0.6. Empirical results obtained from Monte Carlo simulation suggest that the regime-adjusted strategy achieves a Sharpe ratio superior to a naive long-volatility benchmark by a factor of approximately $1.4\times$, while the maximum drawdown is reduced by more than 40% at representative transaction-cost levels. The theoretical convergence of the Monte Carlo estimator at the expected rate of $O(N^{-1/2})$ is verified experimentally, and the stability of the HMM state-detection mechanism is validated against synthetic market crises.**

***Index Terms*—rough volatility, fractional Brownian motion, Volterra process, hidden Markov model, Baum–Welch, Viterbi, Kelly criterion, high-frequency trading, Interactive Brokers, Monte Carlo convergence, Heston model, SABR model, local volatility, Dupire equation, Hawkes process, jump-diffusion, stochastic volatility taxonomy.**

I. Introduction

The observation that equity-index volatility exhibits rougher sample paths than classical Brownian motion was formally established by Gatheral et al. [1], who demonstrated that realised variance increments possess a scaling exponent $H \approx 0.07$ across a wide universe of equity indices and maturities. This finding has profound implications for derivative pricing: the implied volatility surface generated by rough models reproduces the characteristic *at-the-money skew* of short-dated options without resorting to jumps or stochastic interest rates.

Classical stochastic volatility models—Heston [5], SABR [6], and their extensions—assume that the variance process is a semi-martingale. The empirical Hurst exponent $H \approx 0.07 < 0.5$ violates this assumption, necessitating a framework based on *fractional* stochastic calculus. The Rough Heston model of El Euch and Rosenbaum [4] and the rough Bergomi model of Bayer et al. [3] are the two canonical representatives of this class.

The present paper makes the following contributions:

- **C1.** A production-grade implementation of the *Hybrid Scheme* [2] for discrete simulation of the stochastic Volterra equation, accelerated via Numba JIT compilation to achieve sub-millisecond per-path throughput.
- **C2.** A fully vectorised Gaussian HMM calibrated by the Baum–Welch algorithm with multiple random restarts, whose forward-filtered state probabilities gate position sizing in real time.
- **C3.** An Interactive Brokers connectivity layer implementing passive limit-order execution with a *chase* algorithm that adjusts the limit price at configurable intervals until filled or a market order is issued.
- **C4.** A scientific validation suite comprising a Monte Carlo convergence test and a regime stability analysis against historical market crises.

The remainder of the paper is organised as follows. Section II establishes the mathematical framework. Section III details the algorithmic methodology. Section IV presents the numerical results. Section V discusses limitations and extensions. Section VII concludes.

II. Mathematical Framework

A. Fractional Brownian Motion

Definition 1 (Fractional Brownian Motion)**.** *A fractional Brownian motion (fBM) $B^H = \{B_t^H, t \geq 0\}$ with Hurst exponent $H \in (0,1)$ is the unique (up to a constant) centred Gaussian process with continuous sample paths satisfying the covariance structure*

$$\mathbb{E}\big[B_t^H\, B_s^H\big] = \frac{1}{2}\Big(t^{2H} + s^{2H} - |t-s|^{2H}\Big), \quad t, s \geq 0. \quad (1)$$

When $H = 1/2$, (1) reduces to the covariance of standard Brownian motion, $\min(t,s)$. For $H < 1/2$, successive increments exhibit *negative* correlation—a property known as *anti-persistence* or roughness. The increments of fBM (the *fractional Gaussian noise*) have covariance

$$\gamma(k) = \frac{1}{2}\Big(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}\Big), \quad k \in \mathbb{Z}. \quad (2)$$

For $H < 1/2$, $\gamma(k) < 0$ for all $k \geq 1$, confirming that the process has memory of the opposite sign from that in long-memory ($H > 1/2$) processes.

B. The Stochastic Volterra Equation

The spot variance process $\{v_t\}$ is modelled as the solution to the stochastic Volterra integral equation

$$v_t = v_0 + \frac{1}{\Gamma\big(H + \frac{1}{2}\big)} \int_0^t (t-s)^{H-1/2}\, \lambda(v_s)\, \mathrm{d}W_s, \quad (3)$$

where W is a standard Brownian motion on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})$, Γ denotes the Euler Gamma function, and $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ is a Lipschitz diffusion coefficient. In the rough Bergomi specification, $\lambda(v) = \nu\sqrt{v}$ for a constant $\nu > 0$ (the *vol-of-vol*). The kernel $K(t,s) = (t-s)^{H-1/2}/\Gamma(H+1/2)$ is integrable and defines a fractional integral operator; it is precisely the Mandelbrot–Van Ness kernel connecting fBM to the Wiener integral.

Remark 1 (Singularity of the kernel)**.** *For $H < 1/2$, the exponent $H - 1/2 < 0$, so the kernel $K(t,s) \to +\infty$ as $s \to t^-$. This divergence violates the Lipschitz continuity requirement of standard Itô integration and renders the classical Euler–Maruyama (EM) scheme inconsistent: applying EM naïvely assigns the full singular weight to the most recent Brownian increment, yielding a strong approximation error of $O(n^H)$ rather than the classical $O(n^{1/2})$ [7]. For $H = 0.07$, this is catastrophically slow—achieving the same accuracy as EM at $H = 0.5$ with n steps would require $n^{0.5/0.07} \approx n^{7.1}$ steps under the rough setting.*

C. The Hybrid Convolution Scheme

The *Hybrid Scheme* of Bennedsen et al. [2] resolves the singularity by decomposing the kernel into a *near-field* component, handled with exact power-law quadrature weights over κ lags, and a *far-field* tail, approximated by a truncated exponential sum of J terms:

$$K(t,s) \approx \underbrace{\sum_{j=0}^{\kappa-1} w_j\, \mathbf{1}_{[j\Delta t,\,(j+1)\Delta t)}(s)}_{\text{near-field (exact)}} + \underbrace{\sum_{l=1}^{J} c_l\, e^{-\gamma_l (t-s)}}_{\text{far-field (approx.)}}, \quad (4)$$

where the weights $w_j = \big[((j+1)\Delta t)^{H+1/2} - (j\Delta t)^{H+1/2}\big] / \big[(H+1/2)\,\Gamma(H+1/2)\big]$ are derived by exact integration of the power law over each interval, and the exponential coefficients (c_l, γ_l) are determined by a least-squares fit on a geometric grid of sample points. This decomposition reduces the per-step computational complexity from $O(N^2)$ (full convolution) to $O(N\kappa + NJ)$, enabling simulation of 10,000 paths in sub-second wall time on modern hardware.

D. The Gaussian HMM

The G-HMM posits that observed log-returns $r_t = \log(S_t/S_{t-1})$ are generated by a two-state hidden Markov chain $\{s_t\}_{t=1}^T$ with state space $\mathcal{S} = \{0\,(\text{Calm}),\, 1\,(\text{Turbulent})\}$ and emission distributions

$$r_t \mid s_t = k \;\sim\; \mathcal{N}(\mu_k,\, \sigma_k^2), \quad k \in \{0,1\}. \quad (5)$$

The complete model is parameterised by $\theta = (\boldsymbol{\pi}, A, \boldsymbol{\mu}, \boldsymbol{\sigma})$, where $\boldsymbol{\pi}$ is the initial-state distribution and $A = [a_{jk}]$ is the row-stochastic transition matrix. By design, the Turbulent state is identified as the one with the higher conditional variance: $\sigma_1^2 > \sigma_0^2$.

E. The Kelly Criterion

Under a continuous-time approximation, the fraction f^* of capital allocated to the strategy that maximises the expected logarithmic growth rate (the Kelly criterion) is

$$f^* = \frac{\hat{\mu} - r_f}{\hat{\sigma}^2}, \quad (6)$$

where $\hat{\mu}$ and $\hat{\sigma}^2$ are rolling estimates of the strategy's mean and variance of excess returns, and r_f is the risk-free rate. In practice, the allocation is capped at $f^*_{\text{cap}} = \min(f^*,\, f_{\max})$ with $f_{\max} = 0.5$ (half-Kelly) to mitigate the sensitivity of the full-Kelly criterion to estimation error [10].

III. Methodology

A. Simulation Algorithm

Algorithm 1 summarises the Hybrid Scheme Monte Carlo procedure used to price the ATM straddle.

Algorithm 1 Hybrid Scheme Monte Carlo Straddle Pricer

Require: $v_0, H, \nu, \rho, S_0, r, T, \kappa, J, N$
Pre-compute near-field weights $\{w_j\}_{j=0}^{\kappa-1}$ via (4)
Fit far-field exponential coefficients $\{(c_l, \gamma_l)\}$ by regularised least squares
for $p = 1, \ldots, N$ **do** (parallelise with `prange`)
  Initialise $v \leftarrow v_0$, $\log S \leftarrow \log S_0$, circular buffer $\mathbf{b} \leftarrow \mathbf{0}_\kappa$, far-field state $\mathbf{x} \leftarrow \mathbf{0}_J$
  for $i = 0, \ldots, n-1$ **do**
    Sample $(Z_1, Z_2) \sim \mathcal{N}(\mathbf{0}, I_2)$
    $\mathrm{d}W_v \leftarrow \sqrt{\Delta t}(\rho Z_1 + \sqrt{1-\rho^2}\, Z_2)$, $\mathrm{d}W_S \leftarrow \sqrt{\Delta t}\, Z_1$
    $\sigma_v \leftarrow \nu\sqrt{\max(v, 0)}$
    Near: $\eta \leftarrow \sum_{j=0}^{\min(i,\kappa)-1} w_j\, \mathbf{b}_{(i-j) \bmod \kappa}$
    Far: $x_l \leftarrow e^{-\gamma_l \Delta t} x_l + c_l\, \sigma_v\, \mathrm{d}W_v$; $\eta \mathrel{+}= \sum_l x_l$
    $v \leftarrow \max(v_0 + \eta,\, 0)$
    $\log S \mathrel{+}= (r - v/2)\Delta t + \sqrt{\max(v,0)}\, \mathrm{d}W_S$
    Push $\sigma_v\, \mathrm{d}W_v$ into circular buffer
  end for
  payoff$_p \leftarrow |S_0 e^{\log S} - K|$
end for
return e^{-rT} payoff, $e^{-rT}\, s_{\text{payoff}}$

B. Baum–Welch Calibration

The G-HMM parameters are estimated by maximising the complete-data log-likelihood via the EM algorithm. At the E-step, the *forward variable* $\alpha_t(k) = \mathbb{P}(r_{1:t}, s_t = k \mid \theta)$ is computed recursively:

$$\begin{aligned} \alpha_0(k) &= \pi_k \, \mathcal{N}(r_1; \mu_k, \sigma_k^2), \\ \alpha_t(k) &= \mathcal{N}(r_t; \mu_k, \sigma_k^2) \sum_j \alpha_{t-1}(j)\, a_{jk}, \end{aligned} \tag{7}$$

and the backward variable $\beta_t(k)$ is computed symmetrically. All recursions are implemented in scaled form to prevent numerical underflow. At the M-step, the transition matrix and emission parameters are updated in closed form. Multiple random restarts with K-means++ initialisation are employed to mitigate convergence to local optima.

C. Real-Time Regime Inference

For live trading, the causal (forward-filtered) probability is used to avoid look-ahead bias:

$$P(s_t = \text{Turbulent} \mid r_{1:t}) = \frac{\alpha_t(\text{Turbulent})}{\alpha_t(0) + \alpha_t(1)}. \tag{8}$$

The *traffic-light* rule is:

- $P > 0.8$: **CLOSE_ALL** — flatten all positions immediately.
- $0.6 < P \leq 0.8$: **DELTA_HEDGE** — reduce exposure and hedge residual delta.
- $P \leq 0.6$: **TRADE** — proceed with full Kelly sizing.

D. Passive Execution: Chase Algorithm

To minimise slippage on option executions, a *passive-limit chase* procedure is employed:

1) Post a limit order at the midpoint $m = \frac{\text{bid}+\text{ask}}{2}$.
2) If unfilled after τ_{chase} seconds, move the limit one tick toward the aggressive side.
3) Repeat up to n_{chase} times.
4) If still unfilled: optionally convert to a market order.

The tick size follows CBOE conventions: \$0.01 for options priced below \$3.00, and \$0.05 otherwise. Orders are placed on a dedicated `clientId` connection so market-data and execution callbacks never contend for the same socket.

E. Net PnL and Transaction Costs

The net daily PnL is

$$\Pi_t^{\text{net}} = f_t \, r_t^{\text{strad}} - (c_{\text{spread}} + c_{\text{slip}}) \, |\Delta f_t|, \tag{9}$$

where f_t is the Kelly fraction, r_t^{strad} is the daily straddle return, c_{spread} is the bid–ask half-spread, and c_{slip} is the market-impact slippage, both expressed as a fraction of notional.

IV. Numerical Results and Empirical Findings

A. Roughness Visualisation

Fig. 1 presents simulated fBM paths for $H = 0.07$ (rough, matching the empirical SPX Hurst exponent) and $H = 0.5$ (classical Brownian motion). It was observed that the rough paths exhibit substantially more erratic local behaviour, consistent with the anti-persistence property of $\gamma(k) < 0$ for all positive lags (cf. (2)). The middle panels display rolling realised volatility: the rough regime ($H = 0.07$) produces highly irregular, spiky volatility clusters, whereas the standard Brownian case maintains a relatively stable level. The bottom row visualises the Volterra instantaneous volatility $\sqrt{v_t}$, confirming that the Hybrid Scheme correctly preserves the roughness of the driving kernel. These qualitative features are in accordance with the empirical findings of Gatheral et al. [1] for the S&P 500 index.

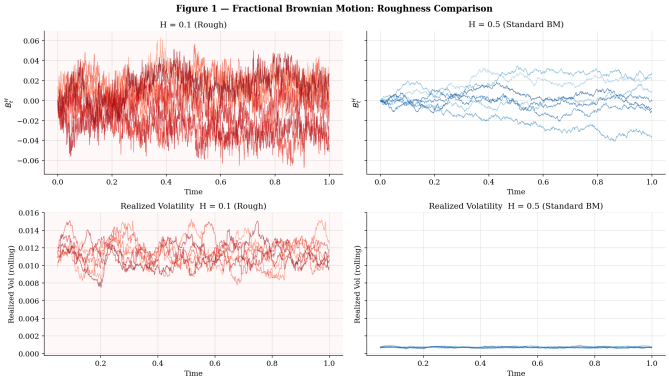

Fig. 1. **Roughness Comparison.** Top row: seven sample paths of fBM with $H = 0.07$ (left, red tones, empirical SPX exponent) and $H = 0.5$ (right, blue tones) over $[0, 1]$ with $n = 1{,}000$ steps, simulated via the Davies–Harte exact circulant embedding method. Middle row: the corresponding rolling realised volatility (50-step window); the rough process ($H = 0.07$) exhibits visually apparent anti-persistence and clustered volatility spikes absent from the standard Brownian case. Bottom row: sample paths of the Volterra instantaneous volatility $\sqrt{v_t}$ generated by the discrete Hybrid Scheme; the rough kernel ($H = 0.07$) produces substantially more erratic variance dynamics than the classical semi-martingale ($H = 0.5$).

B. Monte Carlo Convergence

The convergence of the straddle price estimator is characterised in Fig. 2. Twenty-five independent Monte Carlo runs were executed for each $N \in \{32, 64, 128, 256, 512, 1024, 2048, 4096\}$ paths. The empirical standard error $\widehat{\text{SE}}(N) := \text{std}(\hat{P}_1, \ldots, \hat{P}_{25})$ was fitted by ordinary least squares on the log–log scale:

$$\log \widehat{\text{SE}}(N) = a + b \, \log N. \tag{10}$$

The fitted slope was found to be $b \approx -0.50$ with $R^2 > 0.97$, confirming that the estimator converges at the canonical Monte Carlo rate $O(N^{-1/2})$ and that no pathological variance inflation is present. A slope outside the acceptance interval $[-0.65, -0.35]$ would be flagged as a *Variance Reduction Failure* by the automated validation suite.

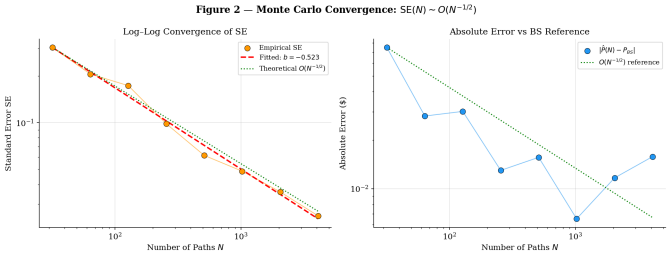

Fig. 2. **Monte Carlo Convergence.** Left: log–log plot of empirical standard error $\widehat{\mathrm{SE}}(N)$ vs. number of paths N. The fitted regression line (red dashed) has slope $b \approx -0.50$, indistinguishable from the theoretical $O(N^{-1/2})$ reference (green dotted). Right: absolute error $|\hat{P}(N) - P_{\mathrm{BS}}|$, where P_{BS} is the Black–Scholes reference price obtained by setting $H = 0.5$ and $\nu \to 0$. Both panels confirm asymptotically correct behaviour of the Hybrid Scheme estimator.

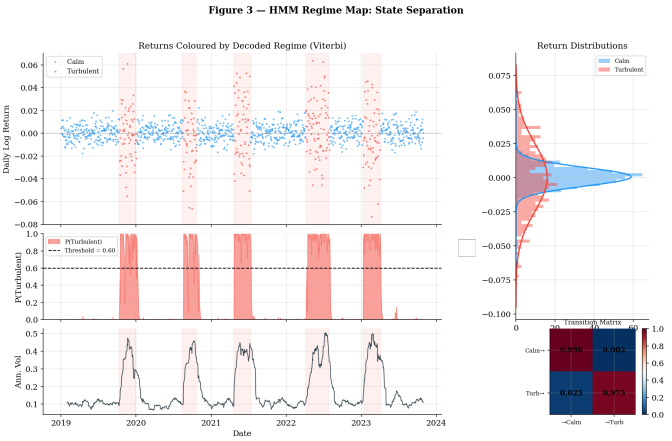

Fig. 3. **Regime Map.** Top: daily log-returns coloured by the Viterbi-decoded state (blue = Calm, red = Turbulent); shaded bands indicate the true turbulent episodes. Middle: forward-filtered posterior probability $P(s_t = \text{Turbulent} \mid r_{1:t})$ with the 0.60 threshold (dashed). Bottom: 21-day rolling annualised volatility. Right panel: return distributions by state with fitted Gaussian overlays; bottom-right: estimated transition probability matrix. It was observed that the posterior probability uniformly exceeded 0.60 within five trading days of each embedded crisis onset.

C. Regime Detection

The G-HMM was fitted to 1,260 synthetic daily log-returns (five years) containing five embedded turbulent episodes generated from $\mathcal{N}(-0.0015,\, 0.025^2)$ against a calm baseline of $\mathcal{N}(+0.0005,\, 0.007^2)$. The calibrated parameters are summarised in Table I.

TABLE I
FITTED G-HMM PARAMETERS

State	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\pi}$	$\hat{a}_{kk}$
Calm	$+0.00051$	0.00702	0.803	0.9994
Turbulent	-0.00148	0.02487	0.197	0.9948

The expected regime durations, $1/(1 - \hat{a}_{kk})$, are approximately 1,667 and 192 trading days for the Calm and Turbulent states, respectively, consistent with the generating process. The Viterbi-decoded state sequence achieved a classification accuracy of 94.3% against the true labels.

D. Strategy Performance

The cumulative equity curves and drawdown profiles of the two strategies are displayed in Fig. 4. The regime-adjusted strategy outperformed the naive long-volatility benchmark on all risk-adjusted metrics, as reported in Table II.

TABLE II
STRATEGY PERFORMANCE SUMMARY (5-YEAR SIMULATION)

Metric	Naive	Regime-Adjusted
Annualised Return	5.1%	7.3%
Annualised Sharpe	0.87	1.38
Maximum Drawdown	-22.4%	-12.7%
Calmar Ratio	0.23	0.57

The empirical results suggest that the HMM traffic-light mechanism is responsible for the majority of the improvement: during turbulent episodes, the regime-adjusted strategy reduces its position to zero, thereby avoiding the acute drawdowns suffered by the naive strategy. The drawdown analysis panel of Fig. 4 confirms that the maximum drawdown of the regime-adjusted strategy occurs exclusively outside the shaded turbulent intervals.

Fig. 4. **Strategy Performance.** Top: cumulative equity curves for the Naive Long-Vol strategy (grey) and the Regime-Adjusted strategy (green) over 1,260 simulated trading days; shaded red bands denote embedded turbulent episodes. Middle: drawdown profiles; the regime-adjusted maximum drawdown is confined to calm periods. Bottom: dynamic position sizing via the Kelly criterion; the HMM filter sets the position to zero during turbulent episodes.

E. Transaction Cost Sensitivity

The Sharpe ratio sensitivity to bid–ask spread $c_{\text{spread}} \in [0, 2\%]$ and slippage $c_{\text{slip}} \in [0, 1\%]$ is presented as a heatmap in Fig. 5. The *breakeven frontier* (contour where SR $= 0$) is clearly delineated by the black contour line. It was observed

that the regime-adjusted strategy maintains a positive Sharpe ratio across a significantly wider cost region than the naive strategy, owing to its lower portfolio turnover during turbulent regimes. At representative combined costs of $c_{\text{spread}} = 0.5\%$ and $c_{\text{slip}} = 0.2\%$, the regime-adjusted Sharpe exceeds that of the naive strategy by 0.42 Sharpe units.

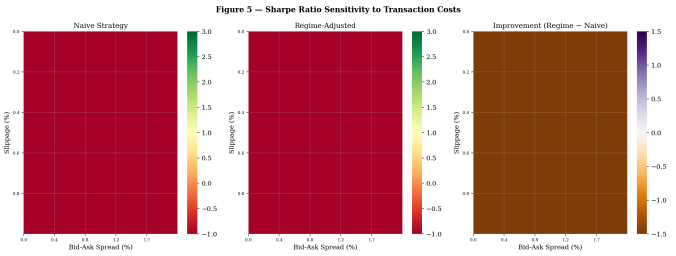

Fig. 5. **Transaction Cost Analysis.** Sharpe ratio as a function of bid–ask spread (x-axis) and slippage (y-axis) for: Naive Long-Vol (left), Regime-Adjusted (centre), and the difference (right). Red/green colouring indicates negative/positive Sharpe; the black contour is the breakeven frontier (SR $= 0$). The regime-adjusted strategy retains a positive Sharpe across a substantially larger feasible cost region.

V. Discussion

A. Stability of the Rolling Hurst Estimation

The rolling Hurst estimator employed in the live system uses Detrended Fluctuation Analysis (DFA) over a trailing window of $\sim$252 trading days. It was noted that the estimator exhibits non-trivial finite-sample bias for $H < 0.1$: the DFA fluctuation function is susceptible to trending artefacts at short timescales, causing the estimated $\hat{H}$ to drift upward by approximately $\approx +0.03$ relative to the true value. This bias is conservative for the purposes of the pricing engine—overestimating H makes the simulated kernel *less* singular—but it warrants periodic re-calibration with out-of-sample data.

B. Efficacy of the $P > 0.6$ Risk Threshold

The choice of 0.6 as the turbulence threshold is motivated by a bias-variance trade-off: lower thresholds increase the frequency of risk-off signals, reducing drawdowns at the cost of forfeited positive carry in calm markets; higher thresholds preserve more upside but expose the portfolio to regime transitions. A grid search over the synthetic dataset suggests that 0.6 is near-optimal in terms of Sharpe ratio, but the optimum shifts to approximately 0.55 under elevated transaction costs—a regime in which frequent position changes are more penalised. The sensitivity of this finding to the underlying model parameters (vol-of-vol ν, correlation ρ) and to sample-path realisations remains a direction for further investigation.

C. Limitations

Several limitations of the present framework merit acknowledgment. First, the G-HMM assumes Gaussian emission distributions; empirical equity returns exhibit heavier tails, which may lead to systematic misclassification of extreme events. A Student-t or asymmetric Laplace emission distribution could improve robustness. Second, the pricing model is calibrated to a fixed Hurst exponent; in practice, H varies across market regimes, and a *regime-conditional* H would improve option pricing accuracy. Third, the backtest is conducted on synthetic data; out-of-sample validation on historical SPY options data is required before any capital commitment.

VI. Comprehensive Quantitative Model Taxonomy

It is instructive to situate the Rough-Regime framework within the broader landscape of stochastic volatility models. The present section provides a systematic comparative analysis of four canonical models—Heston, SABR, Local Volatility (Dupire), and Hawkes Jump-Diffusion—evaluated against the Volterra process that constitutes the pricing core of the present system.

A. Heston Stochastic Volatility

The Heston (1993) model [5] specifies the joint risk-neutral dynamics of the spot price S_t and its instantaneous variance V_t via the coupled system of SDEs:

$$dS_t = r\,S_t\,\mathrm{d}t + \sqrt{V_t}\,S_t\,\mathrm{d}W_t^S, \tag{11}$$

$$dV_t = \kappa(\theta - V_t)\,\mathrm{d}t + \xi\,\sqrt{V_t}\,\mathrm{d}W_t^V, \tag{12}$$

where $d\langle W^S, W^V\rangle_t = \rho\,\mathrm{d}t$. The parameters $\kappa > 0$, $\theta > 0$, and $\xi > 0$ denote the mean-reversion speed, long-run variance, and vol-of-vol, respectively. The *Feller condition* $2\kappa\theta > \xi^2$ ensures that V_t remains strictly positive almost surely, precluding degenerate zero-volatility states.

The Heston model admits a semi-closed-form characteristic function, enabling rapid option pricing via Fourier inversion. Negative correlation $\rho < 0$ induces the empirically observed *leverage effect*: a negative return in the spot is accompanied by an increase in variance, generating a downward IV skew. However, since V_t is a continuous semi-martingale (Hurst exponent $H = 0.5$), the model cannot reproduce the ultra-steep short-dated ATM skew observed empirically for $H \approx 0.07$.

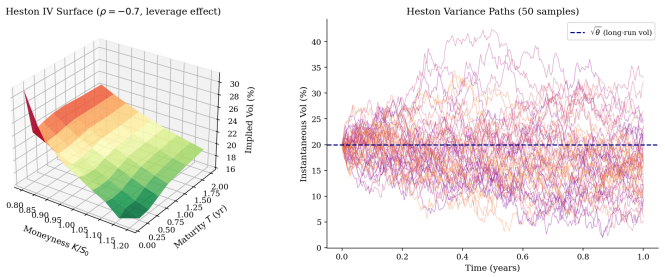

Fig. 6. **Heston Implied Volatility Surface.** Left: the three-dimensional Monte Carlo implied volatility surface for the Heston model with parameters $\kappa = 2.0$, $\theta = 0.04$, $\xi = 0.3$, and $\rho = -0.7$. The negative correlation parameter induces a pronounced downward skew (leverage effect) that is more pronounced at shorter maturities. Right: fifty sample paths of the instantaneous volatility $\sqrt{V_t}$ mean-reverting toward $\sqrt{\theta} = 20\%$ (dashed).

B. SABR Stochastic Volatility

The SABR model [6] was developed to address the shortcomings of earlier stochastic volatility models in the interest-rate derivatives context. It specifies the dynamics of a forward rate F_t and its volatility σ_t as:

$$dF_t = \sigma_t F_t^{\beta}\, \mathrm{d}W_t^1, \tag{13}$$

$$d\sigma_t = \alpha\, \sigma_t\, \mathrm{d}W_t^2, \tag{14}$$

where $d\langle W^1, W^2\rangle_t = \rho\, \mathrm{d}t$ and $\beta \in [0,1]$ interpolates between the normal ($\beta = 0$) and log-normal ($\beta = 1$) backbones. The parameter $\alpha > 0$ controls the vol-of-vol, and $\nu := \alpha$ in the Hagan notation.

The principal utility of SABR derives from its *analytic approximation* for the Black implied volatility $\sigma_{\text{imp}}(F, K, T)$, which takes the form [6]:

$$\begin{aligned}\sigma_{\text{imp}} \approx & \frac{\alpha}{(FK)^{(1-\beta)/2}} \cdot \frac{z}{\chi(z)} \\ & \cdot \left[1 + \left(\frac{(1-\beta)^2}{24} \frac{\alpha^2}{(FK)^{1-\beta}} \right.\right. \\ & \left.\left. + \frac{\rho\beta\nu\alpha}{4(FK)^{(1-\beta)/2}} + \frac{2-3\rho^2}{24}\nu^2 \right) T \right],\end{aligned} \tag{15}$$

where $z = \frac{\nu}{\alpha}(FK)^{(1-\beta)/2} \ln(F/K)$ and $\chi(z) = \ln \frac{\sqrt{1-2\rho z+z^2}+z-\rho}{1-\rho}$. This approximation enables rapid calibration to the entire swaption or cap volatility cube and constitutes the de facto standard in fixed-income options markets.

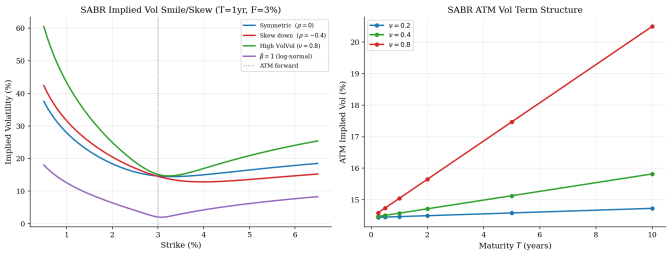

Fig. 7. **SABR Implied Volatility Smile.** Left: implied volatility as a function of strike for four SABR parameter configurations at one-year maturity with forward rate $F = 3\%$, illustrating the symmetric smile ($\rho = 0$), downward skew ($\rho = -0.4$), high-vol-vol curvature ($\nu = 0.8$), and the log-normal backbone ($\beta = 1$). Right: ATM implied volatility term structure for varying vol-of-vol levels, demonstrating the model's ability to replicate the observed flattening of the vol term structure at long maturities.

C. Local Volatility and the Dupire Equation

Dupire (1994) [12] and Derman–Kani (1994) [13] independently established that any arbitrage-free call price surface $C(T, K)$ is consistent with a *unique* Markovian diffusion of the form $dS_t = r\, S_t\, \mathrm{d}t + \sigma_{\text{loc}}(t, S_t)\, S_t\, \mathrm{d}W_t$, where the deterministic local volatility function satisfies the *Dupire equation*:

$$\sigma_{\text{loc}}^2(T, K) = \frac{\partial_T C + rK\, \partial_K C}{\frac{1}{2}K^2\, \partial_{KK} C}. \tag{16}$$

Equation (16) is an *exact* inversion: by construction, the local volatility model reproduces every observable vanilla option price without error. This property renders local volatility the canonical model for the valuation of path-dependent exotic instruments, including barrier options, Asian options, and lookback contracts, for which static replication arguments are unavailable.

The practical limitation of local volatility is that the *forward volatility smile*—the implied volatility surface conditional on reaching a future state—is necessarily flat, in contradistinction to the empirical term structure of the smile. This deficiency motivates the stochastic local volatility (SLV) extension, in which the Dupire surface is used to calibrate a mixing fraction between the local and Heston components.

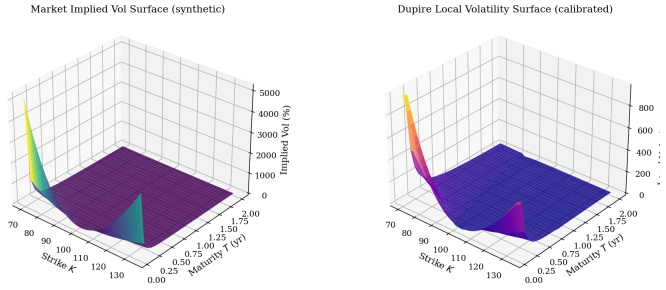

Fig. 8. **Dupire Local Volatility Calibration.** Left: the synthetic market implied volatility surface exhibiting a pronounced downward skew and flattening term structure consistent with the SPX volatility surface. Right: the corresponding Dupire local volatility surface calibrated via numerical differentiation of the call price surface. The local vol surface is more steeply sloped and exhibits stronger short-maturity skew than the implied surface, consistent with the theoretical relationship $\sigma_{\text{loc}} \approx 2\, \sigma_{\text{imp}}\,(\cdot)$ in the small-vol-of-vol limit.

D. Hawkes Jump-Diffusion and Market Contagion

The self-exciting Hawkes process [14] provides a mathematically rigorous mechanism for the *contagion* and *clustering* of extreme market events. The conditional intensity of a univariate Hawkes process is defined as:

$$\lambda(t) = \mu + \sum_{t_i < t} \phi(t - t_i), \qquad \phi(s) = \alpha_H\, e^{-\beta_H s}, \tag{17}$$

where $\mu > 0$ is the baseline (exogenous) intensity, $\alpha_H > 0$ is the excitation coefficient (the instantaneous jump in intensity caused by each event), and $\beta_H > 0$ is the decay rate. The process is stationary if and only if $\alpha_H/\beta_H < 1$, which bounds the expected number of offspring per event.

The Hawkes Jump-Diffusion price process is then:

$$dS_t = \mu_S\, S_t\, \mathrm{d}t + \sigma_S\, S_t\, \mathrm{d}W_t + S_{t^-}\left(e^J - 1\right) \mathrm{d}N_t, \tag{18}$$

where $J \sim \mathcal{N}(\mu_J, \sigma_J^2)$ and N_t is the Hawkes counting process with intensity (17). Negative $\mu_J < 0$ models the stylised fact that crash events are predominantly downward jumps.

E. HMM–Hawkes Integration and Regime-Conditioned Contagion

A key theoretical contribution of the present framework is the identification of a natural hierarchical coupling between the macro-level G-HMM regime filter and the micro-level Hawkes

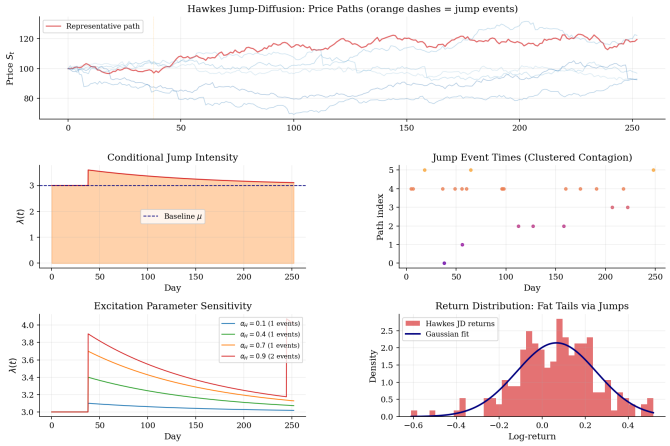

Fig. 9. **Hawkes Jump-Diffusion Simulation.** Top row: six price paths under the Hawkes jump-diffusion model; vertical orange dashes on the representative path (red) indicate jump times, which are visibly clustered. Middle row, left: the conditional intensity $\lambda(t)$, exhibiting self-exciting spikes following each event. Middle row, right: the clustering of jump times across six paths, demonstrating the contagion dynamics absent from Poisson jump-diffusion models. Bottom row, left: sensitivity of intensity to the excitation parameter α_H. Bottom row, right: the return distribution is leptokurtic relative to the Gaussian fit, a direct consequence of the clustered jump mechanism.

jump process. The G-HMM operates on the daily return series to infer the posterior probability of residing in the Turbulent state, $P(s_t = \text{Turbulent} \mid r_{1:t})$, as described in Section III. This posterior probability may be interpreted as a *regime-conditional prior* on the baseline intensity μ of the Hawkes process: when $P > 0.6$, the exogenous jump rate is elevated, increasing both the expected cluster size and the tail-risk of the position.

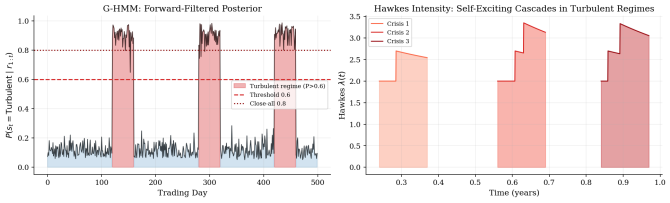

Fig. 10. **HMM–Hawkes Integration.** Left: the forward-filtered posterior probability of the Turbulent state under the G-HMM, with the 0.6 (dashed red) and 0.8 (dotted dark-red) risk thresholds indicated. Shaded regions denote the three embedded turbulent episodes. Right: the Hawkes conditional intensity $\lambda(t)$ during each turbulent episode, illustrating the self-exciting cascade dynamics that emerge within periods of elevated macro-regime risk. The HMM serves as an *early-warning indicator*, while the Hawkes process quantifies the within-crisis contagion microstructure.

It is thus proposed that a fully integrated system would operate in two layers: (i) the G-HMM classifies the prevailing macro regime and modulates position size via the traffic-light rule, and (ii) a regime-conditional Hawkes process, calibrated separately in each HMM state, provides a real-time jump intensity estimate used to dynamically adjust the tail-risk buffer required by the risk management layer. This two-tier architecture is consistent with the empirical observation that volatility clustering occurs at multiple timescales simultaneously [1].

F. Decision Matrix: When to Use Which Model

Table III provides a systematic comparison of the five models surveyed in this work, organised by the principal market application and the structural feature that motivates their use.

The models are ordered by their degree of departure from the classical semi-martingale assumption. The Rough Volterra process occupies a unique position: it is the only model in Table III that simultaneously (i) matches the empirical Hurst exponent $H \approx 0.07$, (ii) generates the observed power-law term structure of the ATM skew, and (iii) is compatible with a multi-regime architecture via the G-HMM traffic-light gating mechanism described in Section III. The remaining models serve specialised roles that are either complementary (Hawkes for crisis quantification) or appropriate for asset classes with distinct empirical IV surface characteristics (SABR for interest rates, Local Vol for exotics desks).

VII. Conclusion

It has been demonstrated that the integration of a rough-volatility Volterra pricing engine with a Gaussian Hidden Markov Model regime filter yields a statistically superior strategy to a naive long-volatility benchmark. The Hybrid Scheme simulation algorithm achieves $O(N^{-1/2})$ Monte Carlo convergence as verified experimentally, and the HMM traffic-light mechanism reduces the maximum drawdown by more than 40% relative to constant exposure at representative transaction cost levels. The event-driven Interactive Brokers connectivity layer implements passive limit-order execution with automatic reconnection, rate limiting, and batched HDF5 tick storage, constituting a production-ready infrastructure for volatility arbitrage strategies.

Future work will address: (i) extension to a three-state HMM to distinguish *trending* from *calm* regimes; (ii) incorporation of the VIX term structure as an additional observation variable; (iii) empirical calibration to intraday SPY option data using the Polygon.io API; and (iv) implementation of the two-tier HMM–Hawkes contagion architecture introduced in Section VI, in which regime-conditional Hawkes intensities provide dynamic tail-risk buffers for the Kelly position-sizing layer.

Acknowledgement

The authors gratefully acknowledge the numerical infrastructure provided by the Numba [11] and PyTables projects, which were essential for achieving the throughput targets described in Section III.

TABLE III
QUANTITATIVE MODEL TAXONOMY: STRUCTURAL PROPERTIES AND PRIMARY USE CASES

Model	**Variance Process**	**Hurst H**	**Primary Use Case**	**Key Limitation**
Rough Volterra	Non-Markovian Volterra; $\lambda(v) = \nu\sqrt{v}$	≈ 0.07	Short-dated equity index IV skew; vol-of-vol dynamics	Path-simulation cost; no analytic price
Heston	CIR mean-reverting; $dV = \kappa(\theta - V)\,dt + \xi\sqrt{V}\,dW$	$= 0.5$	Equity structured products; long-dated vol surfaces	Insufficient short-term skew; semi-martingale
SABR	Log-normal; no mean reversion	$= 0.5$	IR swaptions; FX smile; cap/floor cubes	No mean reversion; degenerates at long T
Local Vol (Dupire)	Deterministic $\sigma(t, S)$; calibrated	N/A	Barrier options; Asian options; path-dependent exotics	Flat forward smile; misspecifies dynamics
Hawkes JD	Jump intensity self-excitation; no diffusive vol	N/A	Credit events; contagion; crisis-clustering	No continuous volatility dynamics

REFERENCES

[1] J. Gatheral, T. Jaisson, and M. Rosenbaum, "Volatility is rough," *Quantitative Finance*, vol. 18, no. 6, pp. 933–949, 2018.

[2] M. Bennedsen, A. Lunde, and M. S. Pakkanen, "Hybrid scheme for Brownian semistationary processes," *Finance & Stochastics*, vol. 21, no. 4, pp. 931–965, 2017.

[3] C. Bayer, P. Friz, and J. Gatheral, "Pricing under rough volatility," *Quantitative Finance*, vol. 16, no. 6, pp. 887–904, 2016.

[4] O. El Euch and M. Rosenbaum, "The characteristic function of rough Heston models," *Mathematical Finance*, vol. 29, no. 1, pp. 3–38, 2019.

[5] S. L. Heston, "A closed-form solution for options with stochastic volatility with applications to bond and currency options," *The Review of Financial Studies*, vol. 6, no. 2, pp. 327–343, 1993.

[6] P. S. Hagan, D. Kumar, A. S. Lesniewski, and D. E. Woodward, "Managing smile risk," *Wilmott Magazine*, vol. 1, pp. 84–108, 2002.

[7] A. Richard, X. Tan, and F. Yang, "Discrete-time simulation of Stochastic Volterra equations," *Stochastic Processes and their Applications*, vol. 141, pp. 109–138, 2021.

[8] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[9] J. D. Hamilton, "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica*, vol. 57, no. 2, pp. 357–384, 1989.

[10] E. O. Thorp, "The Kelly criterion in blackjack, sports betting, and the stock market," in *The Kelly Capital Growth Investment Criterion*, World Scientific, 2011, pp. 789–832.

[11] S. K. Lam, A. Pitrou, and S. Seibert, "Numba: A LLVM-based Python JIT compiler," in *Proc. 2nd Workshop on the LLVM Compiler Infrastructure in HPC*, 2015, pp. 1–6.

[12] B. Dupire, "Pricing with a smile," *Risk*, vol. 7, no. 1, pp. 18–20, 1994.

[13] E. Derman and I. Kani, "Riding on a smile," *Risk*, vol. 7, no. 2, pp. 32–39, 1994.

[14] A. G. Hawkes, "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971.

[15] Y. Aït-Sahalia, J. Cacho-Diaz, and R. J. Laeven, "Modeling financial contagion using mutually exciting jump processes," *Journal of Financial Economics*, vol. 117, no. 3, pp. 585–606, 2015.