DEEP PPDES FOR ROUGH (LOCAL) STOCHASTIC VOLATILITY

ANTOINE JACQUIER AND MUGAD OUMGARI

Abstract. We introduce a new deep-learning based algorithm to evaluate options in a large class of rough stochastic volatility models, possibly including a local volatility component. In this setting, we show that the pricing function is the solution to a path-dependent PDE, for which we develop a numerical scheme based on Deep Learning techniques. Numerical simulations suggest that the latter is extremely efficient, and provides a good alternative to classical Monte Carlo simulations.

1. Introduction

Stochastic models in financial modelling have undergone many transformations since the Black and Scholes model [14], and its most recent revolution, pioneered by Gatheral, Jaisson and Rosenbaum [36], has introduced the concept of rough volatility. In this setting, the instantaneous volatility is the solution to a stochastic differential equation driven by a fractional Brownian motion with small (less than a half) Hurst exponent, synonym of low Hölder regularity of the paths. Not only is this feature consistent with historical time series [36], but it further allows to capture the notoriously steep at-the-money skew of Equity options, as highlighted in [3, 9, 8, 31, 32]. Since then, a lot of effort has been devoted to advocating this new class of models and to showing the full extent of their capabilities, in particular as accurate dynamics for a large class of assets [11], and for consistent pricing of volatility indices [47, 52]. Nothing comes for free though, and the flip side of this new paradigm is the computational cost. With the notable exception of the rough Heston model [1, 28, 29, 27, 26] and its affine extensions [2, 37], the absence of Markovianity of the fractional Brownian motion prevents any pricing tools other than Monte Carlo simulations; the simulation of continuous Gaussian processes, including fractional Brownian motion, is traditionally slow as soon as one steps away from the standard Brownian motion. However, the clear superiority–for estimation and calibration–of these rough volatility models has encouraged deep and fast innovations in numerical methods for pricing, in particular the now standard Hybrid scheme [10, 44] as well as Donsker-type theorems [46, 65], numerical approximations [6, 43, 38] and machine learning-based techniques [48, 72].

In fact, industry practice is often entrenched, not in pure stochastic volatility models, but in models enhanced with a local volatility component, à la Dupire [19], thereby ensuring an exact fit to the observed European option price surface. The natural next step for rough volatility models is therefore to include such a component. We naturally include this possibility in our setting, albeit only in a theoretical way. Although this could be performed by simulation, we adopt a different strategy, following the recent development by Viens and Zhang [73], who proved an analogous version of the Feynman-Kac theorem for rough volatility models. The fundamental difference, though, is that the corresponding partial differential equation is now path-dependent. This forces us to revisit classical market completeness results in the setting of rough local stochastic volatility models. Path-dependent partial differential equations (PPDEs) have been studied extensively by Touzi, Zhang and co-authors [24, 23, 22], and we shall draw existence and uniqueness from their works. Despite these advances, though, very little has been developed to solve these PPDEs numerically, with the sole exception of a path-dependent version of the Barles and Souganidis' monotone scheme [5] by Zhang and Zhuo [74] and Ren and Tan [67]. The implementation thereof is however far from obvious, and we propose instead a novel algorithm, based on the discretisation of the pricing PPDE, which we then solve

using the deep learning technique pioneered by E, Han and Jentzen [21], which does not suffer the curse of dimensionality.

We note in passing that using machine learning (or deep learning) techniques to solve high-dimensional PDEs has recently been the focus of several approaches. Neural networks have indeed been used to solve PDEs for a long time [58, 59]; more recently, Sirignano and Spiliopoulos [71] proposed an algorithm not depending on a given mesh (as opposed to the previous literature), thus allowing for an easier extension to the multi-dimensional case. During the writing-up of the present paper, a further two ideas, similar in spirit to the one we are borrowing from [21] came to light: Sabate-Vidales, Šiška and Szpruch [70], as well as and Huré, Pham and Warin [49] also used the BSDE counterpart of the PDE–albeit in different ways–to apply machine learning techniques. Since our set-up here is not solely about solving a high-dimensional PDE, we shall leave the precise comparison of these different schemes to rest for the moment.

We shall follow a natural progression in this paper: Section 2 introduces the financial modelling setup of the analysis, introducing rough local stochastic volatility models, and proving preliminary results fundamental for their application in quantitative finance. In Section 3, we show that, in this context, a financial derivative is the solution to a path-dependent PDE, for which we propose a discretisation algorithm, and adapt the Deep Learning methodology developed in [21]. We show the validity and accuracy of this technique in Section 5 for the rough Heston model. We gather in appendix some long proofs and reminders in order not to disrupt the flow of the paper.

2. Modelling framework

In order to dive right into the modelling framework and our main setup, we postpone to Appendix A a review of the functional Itô formula for stochastic Volterra systems, as developed by Viens and Zhang [73]. We introduce a rough local stochastic volatility model for the dynamics of a stock price process. Before diving into numerical considerations, we adapt the classical framework of no-arbitrage and market completeness to this setup in order to ensure that pricing and calibration make any sense at all.

2.1. **Rough (local) stochastic volatility model.** We are interested here in stochastic volatility models, where the volatility is rough, in the sense of [36]. This can be written, under the historical measure, as

$$
\begin{aligned}
S_t &= S_0 + \int_0^t \mu_r S_r \mathrm{d}r + \int_0^t l(r, S_r, V_r) S_r \mathrm{d}W_r, \\
V_t &= V_0 + \int_0^t \mathrm{K}(t-r)\Big(b(V_r)\mathrm{d}r + \xi(V_r)\mathrm{d}B_r\Big), \\
\mathrm{d}\langle W, B\rangle_t &= \rho\,\mathrm{d}t,
\end{aligned}
$$

(2.1)

where $\rho \in [-1, 0]$, S_0, V_0 are strictly positive real numbers, and W and B are two standard Brownian motions. Setting $\mathbf{X} = (S, V)$, we can rewrite the system as

$$
\mathbf{X}_t = \mathbf{X}_0 + \int_0^t \mathbf{b}(t, r, \mathbf{X}_r)\mathrm{d}r + \int_0^t \boldsymbol{\sigma}(t, r, \mathbf{X}_r) \cdot \mathrm{d}\mathbf{B}_r,
$$

(2.2)

where

$$
\mathbf{b}(t, r, \mathbf{X}_r) = \begin{pmatrix} \mu_r S_r \\ \mathrm{K}(t-r)b(V_r) \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\sigma}(t, r, \mathbf{X}_r) = \begin{pmatrix} \overline{\rho}\, l(t, S_r, V_r)S_r & \rho\, l(t, S_r, V_r)S_r \\ 0 & \mathrm{K}(t-r)\xi(V_r) \end{pmatrix},
$$

where $\overline{\rho} := \sqrt{1 - \rho^2}$ and $\mathbf{B} = (B^\perp, B)$, with $W := \rho B + \overline{\rho}B^\perp$. The SDE for $\mathbf{X}$ represents a stochastic Volterra system which is, in general, not Markovian. We could in principle allow for more generality and assume, following [73], that for any $t \geq 0$ and $r \in [0, t]$, the coefficients $\mathbf{b}$ and $\boldsymbol{\sigma}$ depend on the past trajectory $\mathbf{X}_{r\wedge\cdot}$, for example to include models with delay. However, such an extension is not needed in the application we are interested in, and we shall not pursue it.

Remark 2.1. We shall only make general assumptions on the behaviour of the function $l(\cdot)$, enough to ensure existence and uniqueness of the system. A classical example in Mathematical Finance is $l(t, S, V) = \varsigma(V)$, for some function $\varsigma(\cdot)$, in which case (2.1) corresponds to a (rough) stochastic volatility model. A more general setting is that of a local (rough) stochastic volatility model, with $l(\cdot)$ of the form $l(t, S, v) = \mathfrak{L}(t, S)\varsigma(v)$,

where $\mathfrak{L}(\cdot, \cdot)$ is usually called the leverage function. From the results by Dupire [19] and Gyöngy [42], if one wants to ensure that this model calibrates exactly to European option prices, then the equality

$$\sigma_L^2(t, s) = \mathbb{E}^{\mathbb{Q}}\left[l(t, S_t, V_t)^2 | S_t = s\right] = \mathbb{E}^{\mathbb{Q}}\left[\mathfrak{L}(t, S_t)^2 \varsigma(V_t)^2 | S_t = s\right] = \mathfrak{L}(t, s)^2 \mathbb{E}^{\mathbb{Q}}\left[\varsigma(V_t)^2 | S_t = s\right]$$

must hold for every $t, s \geq 0$, where the function σ_L is called the local volatility is obtained (at least in theory) directly from European option prices. Here $\mathbb{Q}$ denotes any given risk-neutral measure (we show later that there are in fact infinitely many of them). We will show below (Assumption 2.2 and Theorem 2.5) that such a probability measure exists, thereby making the model meaningful for option pricing. The leverage function can then be recovered directly as

$$(2.3) \qquad \mathfrak{L}(t, s) = \frac{\sigma_L(t, s)}{\sqrt{\mathbb{E}^{\mathbb{Q}}\left[\varsigma(V_t)^2 | S_t = s\right]}},$$

as long as the right-hand side makes sense, and Assumption 2.2 below ensures this is indeed the case. The term on the right-hand side is a conditional expectation with respect to the stock price, and therefore only depends on t and $\{S_t = s\}$, no matter whether the variance process is Markovian or not. This class of models has the advantage of ensuring perfect (at least theoretically) calibration to European option prices, while giving flexibility to price other options, in particular path-dependent or exotic options. A rigorous proof of the existence and uniqueness of (2.1) is outside the scope of this paper; in fact, even in the classical (non-rough case), a general answer does not exist yet, and only recent advances [55, 57] have been made in this direction. From a numerical perspective, calibration of the leverage function can be performed precisely using the particle method developed by Guyon and Henry-Labordère. Ideally, and we hope to achieve this in a near future, one should combine the latter with the simulation of the rough volatility component in order to calibrate vanilla smiles perfectly while capturing other specificities of the market.

We shall always work under the following considerations:

Assumption 2.2.
 (i) The kernel $K \in L^2_{\text{loc}}(\mathbb{R}_+ \to \mathbb{R})$ admits a resolvent of the first kind, and there exists $\gamma \in (0, 2]$ such that

$$\int_0^h \mathrm{K}(t)^2 \mathrm{d}t = \mathcal{O}(h^\gamma) \qquad \text{and} \qquad \int_0^T [\mathrm{K}(t + h) - \mathrm{K}(t)]\, \mathrm{d}t = \mathcal{O}(h^\gamma), \quad \text{for every } T \geq 0, \text{ as } h \text{ tends to zero};$$

 (ii) the functions b and ξ^2 are linear of the form $b(y) = b_0 + b_1 y$ and $\xi(y)^2 = a_0 + a_1 y$;
 (iii) for any $t \geq 0$, the map $\mathfrak{L}(t, \cdot, v)$ is bounded away from zero and bounded above by $\mathfrak{C}$ for any $v > 0$, and $\mathfrak{L}(t, s, \cdot)$ is strictly positive, bounded away from zero, uniformly Hölder continuous with $\mathfrak{L}(t, s, y) \leq C_\varsigma \left(1 + |y|^{p_\varsigma}\right)$ for some $C_\varsigma > 0$ and $p_\varsigma \in (0, 1)$.;
 (iv) the system (2.1) admits a unique (weak) solution;

The form of the kernel K ensures that the variance process has stationary increments. We recall that if the equalities $K * L = L * K = 1$ hold with $L : \mathbb{R}_+ \to \mathbb{R}$ of locally bounded variation, then L is called the resolvent of the first kind of K (here $*$ stands for the convolution operator). We borrow Condition (ii) from [2] so that, in the purely stochastic volatility case $l(t, s, v) \equiv \sqrt{v}$, we are exactly in the setting of an affine Volterra system $(\log(S), V)$. In fact, this condition alone ensures that the process V is an affine Volterra process, and by [2, Theorem 3.3], Assumption 2.2(i) ensures that the SDE for V admits a continuous weak solution. This in particular implies [2, Theorem 4.3] that, under suitable integrability conditions,

$$(2.4) \qquad \mathbb{E}\left[\mathrm{e}^{uV_T} \big| \mathcal{F}_t\right] = \exp\left\{\phi(T - t) + \psi(T - t)V_t\right\},$$

for any $0 \leq t \leq T$, where the two functions ϕ and ψ satisfy the system of Riccati equations

$$\dot{\psi}(t) = b_1 \psi(t) + \frac{a_1}{2}\psi(t)^2 \qquad \text{and} \qquad \dot{\phi}(t) = b_0 \psi(t) + \frac{a_0}{2}\psi(t)^2,$$

with boundary conditions $\psi(0) = u$, $\phi(0) = 0$. Assumption 2.2(i) is again borrowed from [2], and we refer the interested reader to this paper for examples of kernels satisfying this condition. Most examples so far in quantitative finance, such as power-law kernel $\mathrm{K}(t) \equiv t^{H-1/2}$ and Gamma kernel $\mathrm{K}(t) \equiv t^{H-1/2}\mathrm{e}^{-\lambda t}$, for $H \in (0, 1)$, $\lambda > 0$, fit into this framework. Finally, Assumption 2.2(iii) allows the representation

$$(2.5) \qquad S_t = S_0 \exp\left\{\int_0^t \left(\mu_u - \frac{l^2(u, S_u, V_u)}{2}\right) \mathrm{d}u + \int_0^t l(u, S_u, V_u)\mathrm{d}W_u\right\}$$

to be valid since the stochastic integrand is square integrable, by virtue of

$$\int_0^t \mathbb{E}\left[l^2(u, S_u, V_u)\right] \mathrm{d}u \leq C_\varsigma^2 \mathfrak{C}^2 \int_0^t \mathbb{E}\left[\left(1 + |V_u|^{2p_\varsigma}\right)\right] \mathrm{d}u.$$

Remark 2.3. The assumption on the function $l(\cdot)$ may look restrictive at first, in particular regarding the boundedness in the S-space; in the context of local volatility models (Remark 2.1), this implies boundedness for each $t \geq 0$ of the map $\mathfrak{L}(t, \cdot)$; when inferring it from market data through (2.3), it may not be bounded; but it is customary to truncate it for large values of the underlying, and we follow this convention throughout.

Remark 2.4. One could consider a slightly different setup as (2.1), where the kernel does not apply to the whole dynamics of the process V, but only to the diffusion part, in the spirit of diffusions driven by Volterra Gaussian noises [8, 36, 45]. In the simple case, following the seminar paper by Comte and Renault [16], the variance process is the unique strong solution to the Volterra stochastic equation

$$(2.6) \qquad\qquad V_t = V_0 \mathrm{e}^{bt} + a \int_0^t \mathrm{e}^{b(t-u)} \mathrm{d}B_u^H.$$

where B^H is a fractional Brownian motion with Hurst exponent $H \in (0, 1)$, admitting the representation $B_t^H = \int_0^t \mathrm{K}(t-u) \mathrm{d}B_u$ for some standard Brownian motion B with the same filtration as B^H [17]. The process V in (2.6) is a continuous Gaussian process with finite variance, and Fernique's estimate [30] yields that $\mathbb{E}\left[\exp\left(\alpha \sup_{t \in [0,T]} |V_t|^p\right)\right]$ is finite for any $\alpha, T > 0$ and $p < 2$, so that for any $\alpha > 0$,

$$\mathbb{E}\left[\exp\left\{\alpha \int_0^t l(u, S_u, V_u)^2 \mathrm{d}u\right\}\right] \leq \mathbb{E}\left[\exp\left\{\alpha C_\varsigma^2 \mathfrak{C}^2 \int_0^t \left(1 + |V_u|^{2p_\varsigma}\right) \mathrm{d}u\right\}\right]$$

$$\leq \mathbb{E}\left[\exp\left\{\alpha C_\varsigma^2 \mathfrak{C}^2 t \left(1 + \sup_{u \in [0,t]} |V_u|^{2p_\varsigma}\right)\right\}\right]$$

is finite and (2.5) holds.

2.2. **Market completeness and arbitrage freeness.** In the general case where the correlation parameter is different from -1 and 1, the presence of the two noises renders the market incomplete. In order to be able to use the system (2.1) for pricing purposes, we need to show how to complete the market, and to check for the existence of some probability measure under which the stock price is a true martingale. The latter issue was solved for the rough Heston model by El Euch and Rosenbaum [27], while the rough Bergomi case with non-positive correlation was recently proved by Gassiat [33]. We show that this still holds in our framework. We assume the existence of a money market account yielding a risk-free interest rate $(r_t)_{t \geq 0}$.

Theorem 2.5. *The market is incomplete and free of arbitrage.*

Proof. We wrote (2.1) under the historical measure $\mathbb{P}$, but for pricing purposes, we need it under the pricing measure $\mathbb{Q}$. Introduce the market prices of risk as adapted processes $(\lambda_t)_{t \geq 0}$ and $(\beta_t)_{t \geq 0}$ such that

$$(2.7) \qquad\qquad \rho \lambda_t + \overline{\rho} \beta_t = \frac{\mu_t - r_t}{l(t, S_t, V_t)},$$

which is well defined since the denominator is strictly positive by Assumption 2.2(iii). We now define the probability measure $\mathbb{Q}$ via its Radon-Nikodym derivative

$$\left.\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right|_{\mathcal{F}_t} := \exp\left\{-\frac{1}{2} \int_0^t \left(\lambda_s^2 + \beta_s^2\right) \mathrm{d}s - \int_0^t \left(\lambda_s \mathrm{d}B_s + \beta_s \mathrm{d}B_s^\perp\right)\right\},$$

so that Girsanov's Theorem [56, Chapter 3.5] implies that the processes $B^\mathbb{Q}$ and $B^{\mathbb{Q},\perp}$ defined by

$$B_t^\mathbb{Q} := B_t + \int_0^t \lambda_u \mathrm{d}u \qquad \text{and} \qquad B_t^{\mathbb{Q},\perp} := B_t^\perp + \int_0^t \beta_u \mathrm{d}u$$

are orthogonal Brownian motions under $\mathbb{Q}$. Therefore, under $\mathbb{Q}$, the stock price satisfies

$$(2.8) \qquad \begin{aligned} S_t &= S_0 + \int_0^t r_u S_u \mathrm{d}u + \int_0^t l(u, S_u, V_u) \left(\rho \, \mathrm{d}B_u^\mathbb{Q} + \overline{\rho} \, \mathrm{d}B_u^{\mathbb{Q},\perp}\right), \\ V_t &= V_0 + \int_0^t \mathrm{K}(t-u) \left(\widehat{b}(V_u) \mathrm{d}u + \xi(V_u) \mathrm{d}B_u^\mathbb{Q}\right), \end{aligned}$$

where, under $\mathbb{Q}$, the drift of the variance process is now of the form $\widehat{b}(V_t) = b(V_t) - \lambda_t \xi(V_t)$. Introducing the Brownian motion $W^{\mathbb{Q}}$ (under $\mathbb{Q}$), the first SDE then reads

$$S_t = S_0 + \int_0^t r_u S_u \mathrm{d}u + \int_0^t l(u, S_u, V_u) \mathrm{d}W^{\mathbb{Q}}.$$

The proposition then follows directly from Novikov's criterion thanks to the assumptions on the function $l(\cdot)$ in Assumption 2.2(iii). $\qquad\square$

Remark 2.6. Following [33], we could in fact partially relax the assumption on the function $l(\cdot)$. Assume for example a rough local stochastic volatility setting (Remark 2.1) where $l(t, S, v) = \mathfrak{L}(t, S)\varsigma(v)$, but with Assumption 2.2(iii) replaced by $\varsigma(v) = \exp(\eta v)$, for some $\eta > 0$. Then, recalling that $\rho \leq 0$, the proof of [33, Theorem 1], using a localisation argument with the increasing sequence of stopping times $\tau_n := \inf\{t > 0 : V_t = n\}$, remains the same, and the stock price is a true martingale. This in particular means that a local volatility version of the rough Bergomi model [8, 52, 53] falls within our current setup.

3. Pricing via Path-dependent PDEs

We shall from now on only work under the risk-neutral measure and, with a slight abuse of notations, write b instead of $\widehat{b}$ for the drift of the variance process (equivalently taking the market price of volatility risk to be null in Theorem 2.5). In the classical Itô diffusion setting, where the kernel K is constant, Feynman-Kac formula transforms the pricing problem from a probabilistic setting to a PDE formulation. The formulation of our rough local stochastic volatility model (2.1) goes beyond this scope since the system is not Markovian any longer. We adapt here the results developed by Viens and Zhang [73] to show that the pricing problem is equivalent to solving a path-dependent PDE.

We first start with the following lemma, which shows, not surprisingly, that the option price should be viewed, not as a function of the state variable at a fixed given time, but as a functional over paths. The argument is not new, and versions have already appeared for the rough Heston model in [27] and in the context of Itô's formula for stochastic Volterra equations in [73]. Consider a European option with maturity T and payoff function $g(\cdot)$ written on the stock price S. By standard no-arbitrage arguments, from Theorem 2.5, its price at time $t \in [0, T]$ reads

$$(3.1) \qquad\qquad\qquad\qquad P_t = \mathbb{E}[g(S_T)|\mathcal{F}_t].$$

The key object in our analysis below is the infinite-dimensional stochastic process $(\boldsymbol{\Theta}^t)_{t \in [0,T]}$, adapted to the filtration $(\mathcal{F}_t)_{t \in [0,T]}$, defined as

$$(3.2) \qquad\qquad \boldsymbol{\Theta}_u^t := \begin{cases} \mathbf{X}_u, & \text{if } u \in [0, t], \\ \mathbb{E}\left[\mathbf{X}_u - \int_t^u \mathbf{b}(u, r, \mathbf{X}_r)\mathrm{d}r \,\middle|\, \mathcal{F}_t\right], & \text{if } u \in [t, T]. \end{cases}$$

Lemma 3.1. *There exists a map* $\mathrm{P} : [0, T] \times [0, \infty) \times \mathcal{C}([0, T] \to \mathbb{R}^2)$, *such that, for any* $t \in [0, T]$,

$$P_t = \mathrm{P}\left(t, S_t, \left(\boldsymbol{\Theta}_u^t\right)_{t \leq u \leq T}\right).$$

Proof. Following [73], for any $0 \leq t < u \leq T$, we decompose (2.2) as

$$\mathbf{X}_u = \underbrace{\mathbf{X}_0 + \int_0^t \mathbf{b}(u, r, \mathbf{X}_r)\mathrm{d}r + \int_0^t \boldsymbol{\sigma}(u, r, \mathbf{X}_r) \cdot \mathrm{d}\mathbf{B}_r}_{\boldsymbol{\Theta}_u^t \in \mathcal{F}_t} + \underbrace{\int_t^u \mathbf{b}(u, r, \mathbf{X}_r)\mathrm{d}r + \int_t^u \boldsymbol{\sigma}(u, r, \mathbf{X}_r) \cdot \mathrm{d}\mathbf{B}_r}_{\boldsymbol{I}_u^t \notin \mathcal{F}_t}.$$

Applying Itô's lemma to (2.1), denoting $l(t, \mathbf{X}_t) = l(t, S_t, V_t)$ we can write, for any $t \in [0, T]$,

$$S_T = S_t \exp\left\{\int_t^T \left(\mu_r - \frac{l(r, \mathbf{X}_r)^2}{2}\right)\mathrm{d}r + \int_t^T l(r, \mathbf{X}_r)\mathrm{d}W_r\right\}$$

$$= S_t \exp\left\{\int_t^T \left(\mu_r - \frac{l(r, \boldsymbol{\Theta}_r^t + \boldsymbol{I}_r^t)^2}{2}\right)\mathrm{d}r + \int_t^T l(r, \boldsymbol{\Theta}_r^t + \boldsymbol{I}_r^t)\mathrm{d}W_r\right\},$$

and therefore the option price at time $t \in [0,T]$ reads

$$\mathbb{E}\left[g(S_T)|\mathcal{F}_t\right] = \mathbb{E}\left[g\left(S_t \exp\left\{\int_t^T \left(\mu_r - \frac{l(r,\boldsymbol{\Theta}_r^t + \boldsymbol{I}_r^t)^2}{2}\right)\mathrm{d}r + \int_t^T l(r,\boldsymbol{\Theta}_r^t + \boldsymbol{I}_r^t)\mathrm{d}W_r\right\}\right)\middle|\mathcal{F}_t\right],$$

and the lemma follows since $\boldsymbol{I}_r^t \notin \mathcal{F}_t$ for all $r \in [t,T]$. □

The general setup we have considered so far allows for dependence on the stock price process S as well, and Lemma 3.1 indicates that the option price at time t depends on a two-dimensional curve $\boldsymbol{\Theta}^t$. This can be simplified in the setting of a (rough) stochastic volatility model, where the process $\boldsymbol{\Theta}$ in (3.2) now reads

$$(3.3) \qquad \Theta_u^t := \begin{cases} V_u, & \text{if } u \in [0,t], \\ \mathbb{E}\left[V_u - \int_t^u \mathrm{K}(u-r)b(r,V_r)\mathrm{d}r \middle| \mathcal{F}_t\right], & \text{if } u \in [t,T]. \end{cases}$$

We have here used the usual (not bold) notation for Θ as it is only one-dimensional. For any $t \in [0,T]$, the one-dimensional curve Θ^t (or equivalently the infinite-dimensional process $(\Theta^t)_{t\in[0,T]}$) plays a fundamental role. It does not exactly represent the forward variance curve, which is used in [27], and we refer to [73] for the precise correspondence between the two.

Corollary 3.2. *Assume that the function l is of the form $l(t,S_t,V_t) = \varsigma(t,V_t)$, then there exists a function $\mathrm{P}^V : [0,T] \times [0,\infty) \times \mathcal{C}([0,T] \to \mathbb{R})$, such that, for any $t \in [0,T]$,*

$$P_t = \mathrm{P}^V\left(t, S_t, \left(\Theta_u^t\right)_{t\leq u\leq T}\right).$$

Proof. The proof of the corollary is almost identical to that of Lemma 3.1 above. Indeed, we can also write a decomposition of the form $V_u = \Theta_u^t + I_u^t$, for all $0 \leq t < u \leq T$, with $\Theta^t \in \mathcal{F}_t$ and $I^t \notin \mathcal{F}_t$, and with the obvious representation (3.3) for Θ^t. □

The following theorem is the main result here (proved in Appendix B), and shows how to extend the classical Feynman-Kac formula to the path-dependent case. From now on, we shall adopt the notation

$$(3.4) \qquad \mathrm{K}^t := \mathrm{K}(\cdot - t), \qquad \text{for any } t \geq 0,$$

to denote the curve K^t seen at time t. It will become clear in the following theorem how this becomes handy.

Theorem 3.3. *The option price (3.1) is the unique solution to the linear path-dependent PDE*

$$(3.5) \qquad \left(\partial_t + \mathcal{L}_x + \mathcal{L}_{xx} + \mathcal{L}_{x\omega} + \mathcal{L}_\omega + \mathcal{L}_{\omega\omega} - r_t\right)\mathrm{P}\left(t, S_t, \boldsymbol{\Theta}^t\right) = 0,$$

for $t \in [0,T)$, with boundary condition $P\left(T, S_T, \boldsymbol{\Theta}^T\right) = g(S_T)$, where, at the point $(t,x,\boldsymbol{\Theta}^t)$,

$$\mathcal{L}_{x\omega} := \rho l(t,x,\Theta_t^t)\xi(\Theta_t^t)x\left\langle\partial_{x,\omega},\mathrm{K}^t\right\rangle, \quad \mathcal{L}_{xx} := \frac{1}{2}l(t,x,\Theta_t^t)^2x^2\partial_x^2, \quad \mathcal{L}_x := r_t x\partial_x,$$

$$\mathcal{L}_{\omega\omega} := \frac{1}{2}\xi(\Theta_t^t)^2\left\langle\partial_\omega^2,(\mathrm{K}^t,\mathrm{K}^t)\right\rangle, \qquad \mathcal{L}_\omega := b(\Theta_t^t)\left\langle\partial_\omega,\mathrm{K}^t\right\rangle.$$

Note that the function P in the theorem depends on the two-dimensional curve $\boldsymbol{\Theta}^t$, but the coefficients only depend on the one-dimensional curve Θ^t corresponding to the variance process. The pricing PDE (3.5) has both state-dependent terms, involving derivatives with respect to x, and path-dependent ones, involving functional derivatives with respect to ω. In order to streamline the presentation, we defer to Appendix A the precise framework, borrowed from [73], to define these derivatives. Without essential loss of understanding for the rest of the analysis, the reader can view them basically as Fréchet derivatives (with some regularisation due to the singularity of the kernel along the diagonal). Path-dependent PDEs have been studied in great details, and we refer the interested reader to [24, 23, 22].

4. A NEW NUMERICAL FRAMEWORK FOR PPDES

Theorem 3.3 showed that pricing under rough volatility could be analysed through the lens of path-dependent PDEs. However, numerical schemes for such equations are scarce, and the only approaches we are aware of is the extension of Barles and Souganidis' monotone scheme [5] to the path-dependent case by Zhang and Zhuo [74], the convergence of which was proved by Ren and Tan [67]. However, the actual implementation of this scheme in the PPDE context is far from trivial, and we consider a different route here, more amenable to computations in our opinion. We first discretise the PPDE along some basis of functions, reducing the infinite-dimensional problem to a finite-, yet high-, dimensional problem. High-dimensional PDEs suffer from the so-called curse of dimensionality, and are notoriously difficult to solve. We then adopt the deep learning approach recently developed by E, Han and Jentzen [21] to solve this system of PDEs.

4.1. Discretisation of the PPDE.
For each $t \in [0, T]$, we consider a basis $(\psi_a^t)_{a=1,\ldots,p}$ of càdlàg functions, for some fixed integer p, and use it to approximate Θ^t and K^t by

$$\widehat{\Theta}^t := \sum_{a=1}^p \theta_a^t \psi_a^t \qquad \text{and} \qquad \widehat{\mathrm{K}}^t := \sum_{a=1}^p \kappa_a^t \psi_a^t,$$

for some sequence of real coefficients $\boldsymbol{\theta}^t := (\theta_a^t)_{a=1}^p$ and $\boldsymbol{\kappa}^t := (\kappa_a^t)_{a=1}^p$. Since $\mathrm{K}^t \in \mathcal{D}_t$, the space of càdlàg functions on $[t, T]$ (see Appendix A.2), then, from Definition A.4,

$$\left\langle \partial_\omega \mathrm{P}\left(t, x, \Theta^t\right), \mathrm{K}^t \right\rangle := \partial_\varepsilon \mathrm{P}\left(t, x, \Theta^t + \varepsilon \mathrm{K}^t \mathbf{1}_{[t,T]}\right)\big|_{\varepsilon=0} = \partial_\varepsilon \mathrm{P}\left(t, x, \Theta^t + \varepsilon \mathrm{K}^t\right)\big|_{\varepsilon=0},$$

and we can introduce the following approximations of the path derivatives along the direction K^t:

$$\left\langle \partial_\omega \mathrm{P}\left(t, x, \widehat{\Theta}^t\right), \widehat{\mathrm{K}}^t \right\rangle := \partial_\varepsilon \mathrm{P}\left(t, x, \widehat{\Theta}^t + \varepsilon \widehat{\mathrm{K}}^t\right)\Big|_{\varepsilon=0} = \partial_\varepsilon \mathrm{P}\left(t, x, \sum_{a=1}^p \left(\theta_a^t + \varepsilon \kappa_a^t\right)\psi_a\right)\Bigg|_{\varepsilon=0}$$

$$=: \partial_\varepsilon \widehat{\mathrm{P}}\left(t, x, \left(\theta_a^t + \varepsilon \kappa_a^t\right)_{a=1}^p\right)\big|_{\varepsilon=0} = \sum_{i=1}^n \partial_{\theta_a^t} \widehat{\mathrm{P}}\left(t, x, \boldsymbol{\theta}^t\right) \kappa_a^t = \nabla_{\boldsymbol{\theta}^t} \widehat{\mathrm{P}}\left(t, x, \boldsymbol{\theta}^t\right) \cdot \boldsymbol{\kappa}^t,$$

where the new function $\widehat{\mathrm{P}}$ now acts on $[0, T] \times [0, \infty) \times \mathbb{R}^n$. Likewise, for the second functional derivative, we have the approximation

$$\left\langle \partial_\omega^2 \mathrm{P}\left(t, x, \widehat{\Theta}^t\right), \left(\widehat{\mathrm{K}}^t, \widehat{\mathrm{K}}^t\right)\right\rangle = \sum_{a,j=1}^p \partial_{\theta_a^t \theta_j^t} \widehat{\mathrm{P}}\left(t, x, \boldsymbol{\theta}^t\right) \kappa_a^t \kappa_j^t = \left(\boldsymbol{\kappa}^t\right)^\top \cdot \Delta_{\boldsymbol{\theta}^t} \widehat{\mathrm{P}}\left(t, x, \boldsymbol{\theta}^t\right) \cdot \boldsymbol{\kappa},$$

and finally the cross derivatives can be approximated similarly as

$$\left\langle \partial_{x,\omega} \mathrm{P}\left(t, x, \widehat{\Theta}^t\right), \widehat{\mathrm{K}}^t \right\rangle := \partial_x \nabla_{\boldsymbol{\theta}^t} \widehat{\mathrm{P}}\left(t, x, \boldsymbol{\theta}^t\right) \cdot \boldsymbol{\kappa}^t.$$

The PPDE (3.5) therefore becomes

$$(4.1) \qquad \left(\partial_t + \mathcal{L}_x + \mathcal{L}_{xx} + \sum_{i=1}^n \mathcal{L}_{x\theta_a^t} + \sum_{a=1}^p \mathcal{L}_{\theta_a^t} + \sum_{a,j=1}^p \mathcal{L}_{\theta_a^t \theta_j^t} - r_t\right) \widehat{\mathrm{P}} = 0,$$

where the differential operators are defined, for each $a, j = 1, \ldots, n$, as

$$\mathcal{L}_{x\theta_a} := \rho\, l(t, x, \boldsymbol{\theta}^t)\xi(\boldsymbol{\theta}^t)x\kappa_a^t \partial_{x\theta_a^t}, \quad \mathcal{L}_{xx} := \frac{l(t, x, \boldsymbol{\theta}^t)^2}{2}x^2 \partial_x^2, \qquad \mathcal{L}_x := r_t x \partial_x,$$

$$\mathcal{L}_{\theta_a^t \theta_j^t} := \frac{\xi(\boldsymbol{\theta}^t)^2}{2}\kappa_a^t \kappa_j^t \partial_{\theta_a^t \theta_j^t}, \qquad\qquad \mathcal{L}_{\theta_a} := b(\boldsymbol{\theta}^t)\kappa_a^t \partial_{\theta_a^t}.$$

We can rewrite this system in a more concise way as

$$(4.2) \qquad \partial_t \widehat{\mathrm{P}} + \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma} \cdot \boldsymbol{\Sigma}^\top \cdot \Delta \widehat{\mathrm{P}}\right) + \boldsymbol{\mu} \cdot \nabla \widehat{\mathrm{P}} - r_t \widehat{\mathrm{P}} = 0,$$

where $\boldsymbol{\mu}\left(t, x, \boldsymbol{\theta}^t\right) := \left(x r_t, b(\boldsymbol{\theta}^t)\kappa_1^t, \ldots, b(\boldsymbol{\theta}^t)\kappa_n^t\right)^\top$ and

$$\boldsymbol{\Sigma}\left(t, x, \boldsymbol{\theta}^t\right) \cdot \boldsymbol{\Sigma}\left(t, x, \boldsymbol{\theta}^t\right)^\top := \begin{pmatrix} \frac{1}{2}l(t, x, \boldsymbol{\theta}^t)^2 x^2 & \rho l(t, x, \boldsymbol{\theta}^t)\xi(\boldsymbol{\theta}^t)x\kappa_1^t & \cdots & \rho l(t, x, \boldsymbol{\theta}^t)\xi(\boldsymbol{\theta}^t)x\kappa_n^t \\ \rho\, l(t, x, \boldsymbol{\theta}^t)\xi(\boldsymbol{\theta}^t)x\kappa_1^t & \xi(\boldsymbol{\theta}^t)^2(\kappa_1^t)^2 & \cdots & \xi(\boldsymbol{\theta}^t)^2\kappa_1^t\kappa_n^t \\ \vdots & \vdots & \ddots & \vdots \\ \rho\, l(t, x, \boldsymbol{\theta}^t)\xi(\boldsymbol{\theta}^t)x\kappa_n^t & \xi(\boldsymbol{\theta}^t)^2\kappa_1^t\kappa_n^t & \cdots & \xi(\boldsymbol{\theta}^t)^2\left(\kappa_n^t\right)^2 \end{pmatrix}$$

Remark 4.1. The simplest example is to consider piecewise constant curves $\psi_i^t = \mathbf{1}_{\delta_i^t}$, for $i = 1, \ldots, n$, where $(\delta_i^t)_{i=1}^n$ represents a mesh of the interval $[t, T]$. In fact, to simplify the notations below, we consider from now on the mesh $\delta_i^t = [t_{i-1}, t_i)$, and we write $\theta_i^t = \Theta_{t_i}^t$ and $\kappa_i^t = \mathrm{K}(t_i - t)$.

There is an interesting connection between the functional Itô formula in Theorem A.6 and backward stochastic differential equations. Following [25], consider the multidimensional BSDE

$$(4.3) \qquad \begin{cases} X_t = \xi + \displaystyle\int_0^t \overline{\mu}(r, X_r)\mathrm{d}s + \int_0^t \overline{\Sigma}(r, X_r)\mathrm{d}W_r, \\ Y_t = g(X_T) + \displaystyle\int_t^T f(r, X_r, Y_r, Z_r)\mathrm{d}r - \int_t^T Z_r^\top \cdot \mathrm{d}W_r, \end{cases}$$

on some filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$ supporting a d-dimensional Brownian motion W. The solution process (X, Y, Z) takes values, at each point in time, in $\mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d$. Consider further the PDE

$$(4.4) \quad \partial_t u(t, x) + \frac{1}{2}\mathrm{Tr}\left(\overline{\Sigma}(t, x)\overline{\Sigma}(t, x)^\top \Delta u(t, x)\right) + \overline{\mu}(t, x)\nabla u(t, x) + f\left(t, x, u(t, x), \overline{\Sigma}(t, x)^\top \nabla u(t, x)\right) = 0,$$

with terminal boundary condition $u(T, x) = g(x)$. In the regular (Assumption A.3(i)) case, as shown in [73], assuming that (4.4) has a classical solution $u(\cdot)$ in $\mathcal{C}_+^{1,2}(\Lambda)$ (see Appendix A.2), then the couple (Y, Z) defined as $Y_t := u(t, X_t)$ and $Z_t := \overline{\Sigma}(t, X_t)^\top \cdot \nabla u(t, X_t)$ is the solution to (4.3). In light of this result, and following up on the main idea in [21], if the solution to the option problem satisfies (4.2), it also solves the BSDE (4.3), namely

$$(4.5) \qquad \widehat{\mathrm{P}}\left(t, S_t, \boldsymbol{\theta}^t\right) = \widehat{\mathrm{P}}\left(T, S_T, \boldsymbol{\theta}^T\right) + \int_t^T r_u \widehat{\mathrm{P}}\left(u, S_u, \boldsymbol{\theta}^u\right)\mathrm{d}u - \int_t^T \boldsymbol{\Sigma}\left(u, S_u, \boldsymbol{\theta}^u\right)^\top \cdot \nabla\widehat{\mathrm{P}}\left(u, S_u, \boldsymbol{\theta}^u\right)\mathrm{d}W_u,$$

or, written in forward form,

$$(4.6) \qquad \widehat{\mathrm{P}}\left(t, S_t, \boldsymbol{\theta}^t\right) = \widehat{\mathrm{P}}\left(0, S_0, \boldsymbol{\theta}^0\right) - \int_0^t r_u \widehat{\mathrm{P}}\left(u, S_u, \boldsymbol{\theta}^u\right)\mathrm{d}u + \int_0^t \boldsymbol{\Sigma}\left(u, S_u, \boldsymbol{\theta}^u\right)^\top \cdot \nabla\widehat{\mathrm{P}}\left(u, S_u, \boldsymbol{\theta}^u\right)\mathrm{d}W_u.$$

4.2. **Neural network structure.** We now introduce the neural network structure that will help us solve the high-dimensional pricing problem above. We concentrate on the setting in Remark 4.1.

4.2.1. *Simulation of the network inputs.* The first step is a discretisation in time of the stochastic Volterra system (2.2). Many different possible discretisation schemes exist, and we shall not here explore them in great details. We note the fundamental feature that the kernel is singular on the diagonal, and therefore requires special care. The hybrid scheme, recently developed by Bennedsen, Lunde and Pakkanen [10] is a key tool for this. We postpone to Appendix C a detailed analysis of the discretisation for the rough Heston model, which we will use in our numerical application later on. For the sake of our argument here, all we require at the moment is a discretised process $\left(S_{t_i}, V_{t_i}, (\Theta_{t_j}^{t_i})_{0 \leq j \leq n}\right)$ along a grid $(t_i = \frac{iT}{n})_{0 \leq i \leq n}$, for some integer n.

4.2.2. *Euler discretisation scheme.* We are iteratively computing the price of the option at each time step. Both $\widehat{\mathrm{P}}_0$ and $(\nabla\widehat{\mathrm{P}}_0)$ are the initial price and gradient that will be optimised with the weights of the network. On the two-dimensional grid $(t_i, t_j)_{0 \leq i, j \leq n}$, we can discretise the forward stochastic equation (4.6) as

$$\begin{cases} \widehat{\mathrm{P}}\left(t_0, S_{t_0}, \boldsymbol{\theta}^{t_0}\right) &= \widehat{\mathrm{P}}_0, \\ \nabla\widehat{\mathrm{P}}\left(t_0, S_{t_0}, \boldsymbol{\theta}^{t_0}\right) &= \left(\nabla\widehat{\mathrm{P}}\right)_0, \\ \widehat{\mathrm{P}}\left(t_{i+1}, S_{t_{i+1}}, \boldsymbol{\theta}^{t_{i+1}}\right) &= (1 + r_{t_i}\Delta_i)\widehat{\mathrm{P}}\left(t_i, S_{t_i}, \boldsymbol{\theta}^{t_i}\right) + \Sigma\left(t_i, S_{t_i}, \boldsymbol{\theta}^{t_i}\right)^\top \nabla\widehat{\mathrm{P}}\left(t_i, S_{t_i}, \boldsymbol{\theta}^{t_i}\right)\Delta W_{t_i}, \\ \widehat{\mathrm{P}}\left(t_n, S_{t_n}, \boldsymbol{\theta}^{t_n}\right) &= g(S_T), \end{cases}$$

where $W_{t_i} = \left(B_{t_i}^{\mathbb{Q},\perp}, B_{t_i}, \ldots, B_{t_i} \right)^{\top} \in \mathbb{R}^n$. Discretising the backward SDE (4.5) would give

$$
\begin{cases}
\widehat{\mathrm{P}}\left(t_n, S_{t_n}, \boldsymbol{\theta}^{t_n}\right) &= g(S_T), \\
\widehat{\mathrm{P}}\left(t_i, S_{t_i}, \boldsymbol{\theta}^{t_i}\right) &= \left(1 - r_{t_{i+1}}\Delta_{i+1}\right)\widehat{\mathrm{P}}\left(t_{i+1}, S_{t_{i+1}}, \boldsymbol{\theta}^{t_{i+1}}\right) - \Sigma\left(t_i, S_{t_i}, \boldsymbol{\theta}^{t_i}\right)^{\top}\nabla\widehat{\mathrm{P}}\left(t_i, S_{t_i}, \boldsymbol{\theta}^{t_i}\right)\Delta W_{t_i}.
\end{cases}
$$

We shall follow here this backward approach as it is more natural for pricing exotic (path-dependent) derivatives, such as Bermudan or American options.

4.3. **Neural networks.** Based on the discretisation of the process, we introduce a coarser discretisation grid $(\tau_i)_{i=1,\ldots,m}$ such that $\tau_m = t_n$, $\tau_0 = t_0$ and $m \leq n$. In [21], E, Han and Jentzen assumed $n = m$, but we allow here for more flexibility. This also greatly improves the speed of the algorithm, as it reduces the number of networks–hence the number of network parameters–and simplifies the computation of the loss function. For each $i = 1, \ldots, m$, the price as well as the model parameters are known. The only unknown is the term $\Sigma\left(\tau_i, S_{\tau_i}, \boldsymbol{\theta}^{\tau_i}\right)^{\top}\nabla\widehat{\mathrm{P}}\left(\tau_i, S_{\tau_i}, \boldsymbol{\theta}^{\tau_i}\right)$ involving the gradient and the diffusion matrix. We therefore use a neural network to infer its value:
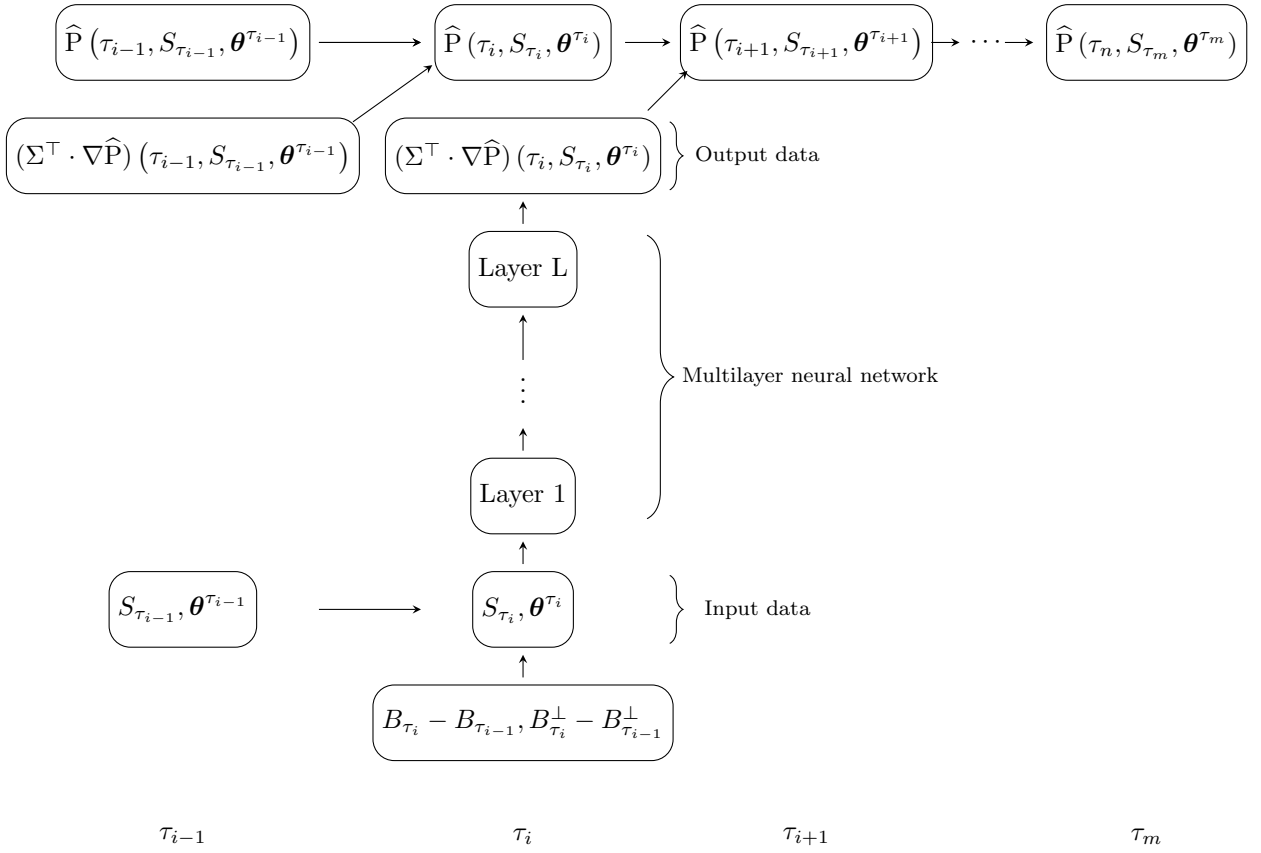
The input consists of one input layer with the value of the processes at t_i. Each hidden layer is computed by multiplying the previous layer by the weights $\mathbf{w}$ and adding a bias $\boldsymbol{\delta}$. After computing a layer, we apply a batch normalisation by computing the mean $\mathfrak{m}$ and the standard deviation s of the layer, and by applying the linear transformation $\mathrm{T}(x) := \gamma\frac{x-\mathfrak{m}}{s} + \beta$ to each element of the layer, where the scale γ and the offset β are to be calibrated. We also use the `ReLu` activation function $a(x) = x_+$ on each element of the layer.

4.4. **Optimisation of the algorithm.** In the algorithm, n_l denotes the number of layers per sub-network, n_N the number of neurons per layer, n_B the number of batches used to separate the samples, N the size of each batch and *epoch* shall denote the number of times all the samples are fed to the training algorithm. We

finally introduce the following loss function that we aim to minimise:

$$\mathrm{L}\left(\mathbf{w}, \boldsymbol{\delta}, \beta, \gamma, \widehat{\mathrm{P}}_0, (\nabla\widehat{\mathrm{P}})_0\right) := \mathbb{E}\left[\left|g\left(S_{\tau_m}\right) - \widehat{\mathrm{P}}\left(\tau_m, S_{\tau_m}, \boldsymbol{\theta}^{\tau_m}\right)\right|^2\right],$$

or, in fact, its version on the sample,

$$(4.7) \qquad \widehat{\mathrm{L}}\left(\mathbf{w}, \boldsymbol{\delta}, \beta, \gamma, \widehat{\mathrm{P}}_0, (\nabla\widehat{\mathrm{P}})_0\right) := \frac{1}{N}\sum_{k=1}^{N}\left|g\left(S_{\tau_m}^k\right) - \widehat{\mathrm{P}}\left(\tau_m, S_{\tau_m}, (\boldsymbol{\theta}^{\tau_m})^k\right)\right|^2.$$

It represents the variance of the initial price found by backward iterations for each simulated path. Since the initial price is unique and deterministic, minimising its variance is natural good way to compute the initial price. Regarding the optimisation itself, we use the Adaptive Moment Estimation method [4] for the first iterations, and switch to the stochastic gradient method, with slower but more stable convergence properties. The different parameters to calibrate are then

- The weights $\mathbf{w}$ and biases $\boldsymbol{\delta}$ for each layer of each sub-network;
- The β and γ in the batch normalisation;
- The initial price $\widehat{\mathrm{P}}_0$;
- The initial gradient $(\nabla\widehat{\mathrm{P}})_0$.

5. NUMERICS: APPLICATION TO THE ROUGH HESTON MODEL

We develop numerics for the rough Heston model, developed in [27, 26, 28], which has the following form:

$$(5.1) \qquad \begin{cases} \mathrm{d}S_t = S_t\sqrt{V_t}\mathrm{d}W_t, \\ V_t = V_0 + \displaystyle\int_0^t \mathrm{K}(t-s)\left[\kappa(\theta - V_s)\mathrm{d}s + \nu\sqrt{V_s}\mathrm{d}Z_s\right], \\ \mathrm{d}\langle W, Z\rangle_t = \rho\,\mathrm{d}t, \end{cases}$$

where the kernel is defined as $\mathrm{K}(t) := \frac{1}{\Gamma(\alpha)}t^{\alpha-1}$, for $\alpha \in (\frac{1}{2}, 1)$. By [2], the variance process V admits a non-negative weak solution if both θ and V_0 are non-negative. The model is not Markovian, and a precise Monte Carlo scheme with accurate convergence is not available yet. We adapt the hybrid scheme from [10] to the rough Heston model, detailed in Appendix C. El Euch and Rosenbaum [26] showed that, similarly to the standard Heston model ($\alpha = 1$), the characteristic function of the stock price can be computed in semi-closed form, and the next section details this, as well as a numerical scheme, that we use for comparisons.

5.1. **Pricing via fractional Riccati equations.** For $t \geq 0$, $\Phi_t : \mathbb{R} \to \mathbb{C}$ denotes the characteristic function

$$\Phi_t(u) = \mathbb{E}\left[\left(\frac{S}{S_0}\right)^{\mathtt{i}u}\right].$$

El Euch and Rosenbaum [26] showed that

$$\log\Phi_t(u) = \theta\mathfrak{I}^1 h(u, t) + V_0\mathfrak{I}^{1-\alpha}h(u, t),$$

where, for any $u \in \mathbb{R}$, $h(u, t)$ solves the fractional Riccati equation

$$(5.2) \qquad \mathfrak{D}^\alpha h(u, t) = \mathfrak{F}(u, h(u, t)),$$

with

$$\mathfrak{F}(u, x) := -\frac{u(u + \mathtt{i}u)}{2} + (\mathtt{i}u\rho\nu - 1)x + \frac{\nu^2 x^2}{2}.$$

Here, $\mathfrak{I}^r$ and $\mathfrak{D}^r$ denote the fractional integral and the fractional derivative defined by

$$\mathfrak{I}^r f(t) := \frac{1}{\Gamma(r)}\int_0^t (t-s)^{r-1}f(s)\mathrm{d}s \qquad \text{and} \qquad \mathfrak{D}^r f(t) := \frac{1}{1-r}\frac{\mathrm{d}}{\mathrm{d}t}\int_0^t (t-s)^{-r}f(s)\mathrm{d}s.$$

Contrary to the Heston model however, the fractional equation (5.2) does not admit a closed-form expression, and we recall the Adams scheme, proposed in [26] to solve it numerically. Taking the fractional integrals on both sides of (5.2) yields

$$h(u, t) = \frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}\mathfrak{F}(u, h(u, s))\,\mathrm{d}s.$$

We now define an equidistant grid $(t_k)_{0 \le k \le n}$ with mesh size Δ such that $t_k = k\Delta$. For each $k = 0, \ldots, n$, we approximate $h(u, t_k)$ by

$$\widehat{h}(u, t_k) := \frac{1}{\Gamma(\alpha)} \int_0^{t_k} (t_k - s)^{\alpha-1} \widehat{g}(u, s)) \, \mathrm{d}s,$$

where, for $t \in [t_j, t_{j+1})$,

$$\widehat{g}(u, t) := \frac{t_{j+1} - t}{t_{j+1} - t_j} \mathfrak{F}(u, h(u, t_j)) + \frac{t - t_j}{t_{j+1} - t_j} \mathfrak{F}(u, h(u, t_{j+1}))$$

is a linear interpolation function. Therefore

$$\widehat{h}(u, t_k) = \sum_{j=0}^{k-1} a_{j,k} \mathfrak{F}(u, h(u, t_j)) + a_{k,k} \mathfrak{F}(u, h(u, t_k)),$$

with

$$\begin{cases} a_{0,k} = \dfrac{\Delta^\alpha}{\Gamma(\alpha+2)} \left((k-1)^{\alpha+2} - (k - \alpha - 1)k^\alpha \right), \\[2ex] a_{j,k} = \dfrac{\Delta^\alpha}{\Gamma(\alpha+2)} \left[(k-j+1)^{\alpha+1} + (k-j-1)^{\alpha+1} - 2(k-j)^{\alpha+1} \right], 1 \le j \le k-1, \\[2ex] a_{k,k} = \dfrac{\Delta^\alpha}{\Gamma(\alpha+2)}. \end{cases}$$

Since $h(u, t_k)$ appears on both sides, it is an implicit scheme which we approximate by an explicit scheme. Consider a predictor $\widehat{h}^P(u, t_k)$ derived from the Riemann sum approximating $\widehat{h}(a, t_k)$ by the integral

$$\widehat{h}^P(u, t_k) = \frac{1}{\Gamma(\alpha)} \int_0^{t_k} (t_k - s)^{\alpha-1} \widetilde{g}(u, s) \mathrm{d}s,$$

with $\widetilde{g}(u, t) := \mathfrak{F}(u, h(u, t_j))$ for $t \in [t_j, t_{j+1})$, and therefore

$$\widehat{h}^P(u, t_k) = \sum_{j=0}^{k-1} b_{j,k} \mathfrak{F}(u, \widehat{h}(u, t_j)) \qquad \text{with} \qquad b_{j,k} = \frac{\Delta^\alpha}{\Gamma(\alpha+1)} \left[(k-j)^\alpha - (k-j-1)^\alpha \right].$$

The final scheme therefore reads

$$\widehat{h}(u, t_k) = \sum_{j=0}^{k-1} a_{j,k} \mathfrak{F}(u, h(u, t_j)) + a_{k,k} \mathfrak{F}(u, h^P(u, t_k)).$$

Lewis [60] showed that we can recover Call option prices via inverse Fourier transform as

$$C(S, T, K) = S - \frac{\sqrt{SK}}{\pi} \int_0^\infty \Re \left(\mathrm{e}^{\mathrm{i}uk} \Phi_T \left(u - \frac{\mathrm{i}}{2} \right) \right) \frac{\mathrm{d}u}{u^2 + \frac{1}{4}}.$$

5.2. **Analysis of the algorithm.** We consider the following computer and software specifications:

Intel Core i7-6600U, CPU 2.60GHz, 32GB RAM, Python 3.6.1, Anaconda 4.4.0, Tensorflow 1.5.0.

and all times below are indicated in seconds.

5.2.1. *Discussion on the number of required networks.* We consider the following parameters for the rough Heston model, without interest rate nor dividend:

$$(5.3) \qquad \kappa = 1, \qquad \nu = 0.1, \qquad \alpha = 0.6, \qquad \rho = -0.7, \qquad V_0 = 0.04, \qquad \theta = 0.06, \qquad S_0 = 1,$$

as well the following network configuration: we consider $50{,}000$ Monte Carlo paths, with 200 time steps. For the deep learning algorithm, each neural network consists of 3 layers with 5 neurons each, the learning rate is set to 0.2 and the number of iterations set to 1000. In this example, we discretise the curve Θ on a space of functions of dimension 10. We consider 20 log-moneynesses ranging from -0.4 to 0.4, for maturities in $\{0.1, 0.5, 1.6, 5.\}$ (expressed in years). The loss function we consider in fact takes into account all strikes for each maturity, and not each single strike individually, therefore increasing the speed of the algorithm. Table 2 below shows the differences between the BSDE scheme, the standard Monte Carlo and the Riccati method, for different number (m) of BSDE time steps, corresponding in fact to different numbers of neural networks. Surprisingly at first, increasing the number of BSDE time steps (m)–clearly more computationally intensive–does not improve the accuracy. This can be explained by the fact that a larger number of networks

implies more parameters to optimise over, and therefore reduces accuracy. This leads us to advocate an algorithm between the one by E, Han, Jentzen [21], where the number of networks is equal to the number of Monte Carlo time steps, and that by Chan-Wai-Nam, Mikael, Warin [64], who consider a single network. For each table, *Build time* corresponds to the time to build the network, *train time* is the training time, *DL Avg Er* is the average error (across all strikes) between the deep learning price and the Riccati price (computed with 100 discretisation steps), and *DL Max Error* is the max absolute errors across all strikes. As a comparison, the errors and computation times between the Monte Carlo and the Riccati prices are detailed in Table 1.

	MC time	Riccati time	Avg Error	Max Error
T=0.1	133	54	-5.E-6	2.2E-3
T=0.5	138	57	-1.4E-4	4.E-4
T=1.6	138	54	1.2E-3	2.7E-3
T=5	140	55	3.7E-3	5.6E-3

Table 1: Monte Carlo and Riccati prices for comparisons

	T	Build time	Train time	DL Avg Err	DL Max Error
$m=2$	0.1	31.31	24.28	3.2E-4	2.2E-3
	0.5	26.10	25.49	-5.E-6	2.0E-4
	1.6	26.78	21.55	-2.7E-3	4.4E-3
	5	25.91	20.52	-2.4E-2	2.7E-2
$m=4$	0.1	43.48	23.30	4.5E-4	2.3E-3
	0.5	43.36	23.98	9.8E-4	1.2E-3
	1.6	53.54	33.87	5.9E-3	7.4E-3
	5	47.12	26.69	3.9E-2	4.1E-2
$m=5$	0.1	46.46	25.77	3.4E-4	2.2E-3
	0.5	44.12	30.70	-7.3E-4	9.0E-4
	1.6	40.72	21.98	1.5E-3	2.9E-3
	5	44.25	21.03	6.8E-3	8.6E-3
$m=10$	0.1	64.14	21.06	2.8E-4	2.2E-3
	0.5	65.81	20.38	-1.5E-2	1.6E-2
	1.6	72.67	23.14	-2.3E-2	2.5E-2
	5	68.30	21.04	4.6E-2	4.8E-2
$m=20$	0.1	132.06	24.247	-2.3E-2	2.5E-2
	0.5	132.24	27.149	-5.2E-2	5.2E-2
	1.6	133.12	22.513	1.3E-1	1.4E-1
	5	124.92	20.067	-1.E-1	1.0E-1

Table 2: Prices computed via the deep learning algorithm with different (m) BSDE steps

5.2.2. *Shapes of the smile.* We now consider the following set of parameters (we have bumped the volatility of volatility and the correlation on purpose to capture the level of the skew on Equity markets):

$$(5.4) \qquad \kappa = 1, \qquad \nu = 0.9, \qquad \alpha = 0.6, \qquad \rho = -0.8, \qquad V_0 = 0.04, \qquad \theta = 0.06, \qquad S_0 = 1.$$

We shall consider two maturities, and, with the same hyper-parameters as above, amended with the following configurations (*Networks* is the number of BSDE steps, *Neurons* stands for the number of neurons per layer, and *Layers* is the number of layers). The implied volatility smiles look as in Figures 1 and 1. On Figure 2, we added the results computed with the Adams scheme for the Riccati equation with 400 discretisation steps. It is well known that the Riccati version is not so accurate for small maturities and very steep skews as in our example here (with the parameters in (5.4)), and one could for example use the power series algorithm proposed by Callegaro, Grasselli and Pagès [15]. The conclusion here is that one should overall be parsimonious with the number of hyper-parameters: increasing the number of layers or the number of

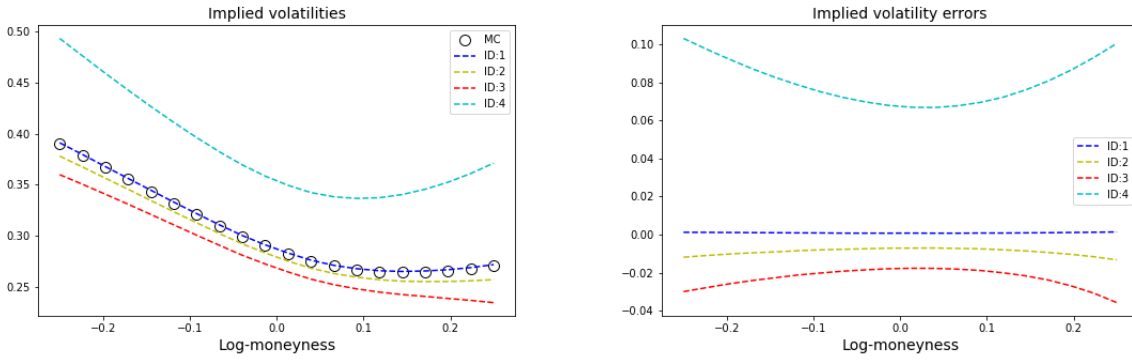|	Config ID	Networks (m)	Neurons	Layers
$T = 0.6$	1	4	6	4
	2	4	3	2
	3	2	3	4
	4	4	10	4
$T = 0.2$	1	5	6	4
	2	3	10	2
	3	3	6	6
	3	2	6	4

Table 3: List of configurations

FIGURE 1. Implied volatility smiles for different configurations as in Table 3 for $T = 0.6$ years.
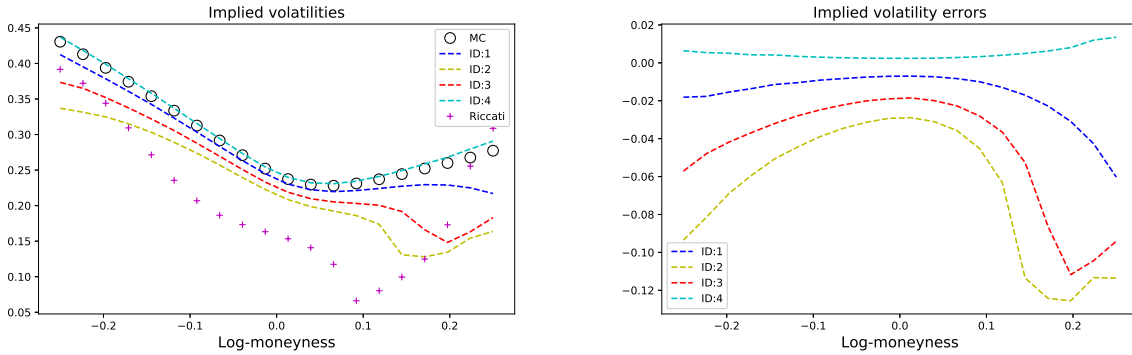

FIGURE 2. Implied volatility smiles for different configurations as in Table 3 for $T = 0.2$ years.

neurons per layer is far from optimal, and a few of them are enough for sufficient accuracy. We leave for future research and deeper numerical analysis the study and convergence of such an algorithm for path-dependent options. In this case, we believe that the flexibility of our network compared to the competitors is key, as one can match the BSDE time steps (hence number of networks) with the path-dependent constraints of the setup: early exercise features for American options, no-exercise periods for convertible bonds for example.

APPENDIX A. A REVIEW OF THE FUNCTIONAL ITÔ FORMULA FOR STOCHASTIC VOLTERRA SYSTEMS

We gather here several results from Viens and Zhang [73], which are key to our analysis.

A.1. **Stochastic Volterra systems.** We consider here a stochastic Volterra process of the form

$$(A.1) \qquad \mathbf{X}_t = \mathbf{X}_0 + \int_0^t \mathbf{b}(t, r, \mathbf{X}_.) \mathrm{d}r + \int_0^t \boldsymbol{\sigma}(t, r, \mathbf{X}_.) \cdot \mathrm{d}\mathbf{W}_r,$$

where $\mathbf{X}_0 \in \mathbb{R}^d$, $\mathbf{W}$ is a Brownian motion in $\mathbb{R}^n$ on a given filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$, and $\mathbf{b} : \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^d$ and $\boldsymbol{\sigma} : \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^{d \times n}$ satisfy the following assumptions:

Assumption A.1. The processes $\mathbf{b}$ and $\boldsymbol{\sigma}$ are adapted to $(\mathcal{F}_t)_{t \geq 0}$, and the derivatives $\partial_t \mathbf{b}$ and $\partial_t \boldsymbol{\sigma}$ exist. Furthermore, for $\boldsymbol{\varphi} \in \{\mathbf{b}, \boldsymbol{\sigma}, \partial_t \mathbf{b}, \partial_t \boldsymbol{\sigma}\}$, $|\boldsymbol{\varphi}(t, s, \omega)| \leq C \left(1 + \|\omega\|_T^a\right)$ for some $C, a > 0$.

Here, $\|\omega\|_T := \sup_{0 \leq t \leq T} |\omega_t|$ denotes the supremum norm on the interval $[0, T]$. Saying that $\boldsymbol{\varphi}$ is adapted here is equivalent to the fact that it can be written as $\boldsymbol{\varphi}(t, r, \mathbf{X}_.) = \boldsymbol{\varphi}(t, r, \mathbf{X}_{r \wedge .})$. This assumption ensures that the system (A.1) is well defined in the following sense:

Assumption A.2. The SDE (A.1) admits a weak solution and $\mathbb{E}\left[\sup_{t \in [0, T]} |\mathbf{X}_t|^p\right]$ is finite for any $p \geq 1$.

This assumption follows [73]. We do not require strong solutions, as the noise W is not observable, and only $\mathbf{X}$ is (or at least some of its components, for example S in the rough Heston model). We refer the reader to [12, 13] for precise conditions on $\mathbf{b}$ and $\boldsymbol{\sigma}$ ensuring weak existence of a solution. The moment condition is more technical and needed for the functional Itô formula in Theorem A.6. Most interesting models in the finance literature satisfy these assumptions, and we refer to [73, Appendix] for sufficient conditions ensuring Assumption A.2, in particular for the class of rough affine models [2]. The key differences between (A.1) and a classical stochastic differential equation is that both drift and diffusion depend (a) on two time variables (thus violating the flow property), and (b) on the whole path. The other classical issue is that the coefficients may blow up, as for the Riemann-Liouville fractional Brownian motion $\int_0^t (t-s)^{H-1/2} \mathrm{d}W_s$, where the power-law kernel explodes on the diagonal whenever the Hurst exponent H lies in $(0, \frac{1}{2})$. Following the terminology introduced in [73], two cases have to be distinguished:

Assumption A.3.
  (i) (Regular case) For any $s \in [0, T]$, $\partial_t \mathbf{b}(t, s, \cdot)$ and $\partial_t \boldsymbol{\sigma}(t, s, \cdot)$ exist on $[s, T]$, and for $\boldsymbol{\varphi} \in \{\mathbf{b}, \boldsymbol{\sigma}, \partial_t \mathbf{b}, \partial_t \boldsymbol{\sigma}\}$,

$$|\boldsymbol{\varphi}(t, s, \omega)| \leq C \left(1 + \|\omega\|_T^a\right), \qquad \text{for some } a, C > 0;$$

  (ii) (Singular case) Let $\boldsymbol{\varphi} \in \{\mathbf{b}, \boldsymbol{\sigma}\}$. For any $s \in [0, T]$, $\partial_t \boldsymbol{\varphi}(t, s, \cdot)$ exists on $(s, T]$, and there exists $h \in \left(0, \frac{1}{2}\right)$ such that, for some $a, C > 0$,

$$|\boldsymbol{\varphi}(t, s, \omega)| \leq C \left(1 + \|\omega\|_T^a\right) (t-s)^{h-1/2} \qquad \text{and} \qquad |\partial_t \boldsymbol{\varphi}(t, s, \omega)| \leq C \left(1 + \|\omega\|_T^a\right) (t-s)^{h-3/2}.$$

The first case mainly deals with the path dependence and the absence of the Markov property, while the second one allows us to treat the presence of two time variables in the kernel, which occurs in fractional models, and in particular in the setting of Section 2 above. For any $0 \leq t \leq u$, we can decompose (A.1) as
(A.2)

$$\mathbf{X}_u = \underbrace{\mathbf{X}_0 + \int_0^t \mathbf{b}(u, r, \mathbf{X}_{r \wedge .}) \mathrm{d}r + \int_0^t \boldsymbol{\sigma}(u, r, \mathbf{X}_{r \wedge .}) \mathrm{d}\mathbf{W}_r}_{\Theta_u^t \in \mathcal{F}_t} + \underbrace{\int_t^u \mathbf{b}(s, r, \mathbf{X}_{r \wedge .}) \mathrm{d}r + \int_t^u \boldsymbol{\sigma}(u, r, \mathbf{X}_{r \wedge .}) \mathrm{d}\mathbf{W}_r}_{I_u^t \notin \mathcal{F}_t}.$$

We further recall (from [73]) the concatenation notation of the paths $\mathbf{X}$ and Θ^t before and after time t,

$$(A.3) \qquad \left(\mathbf{X} \otimes^t \Theta^t\right)_u := \mathbf{X}_u \mathbf{1}_{\{0 < u < t\}} + \Theta_u^t \mathbf{1}_{\{t < u < T\}}, \qquad \text{for any } u, t \in [0, T].$$

A.2. **Functional Itô calculus.** For any $t \in [0, T]$, let $\mathcal{D}_t$ and $\mathcal{C}_t$ denote respectively the space of càdlàg functions on $[t, T]$ and that of continuous functions on $[t, T]$, as well as

$$\overline{\Lambda} := \left\{(t, \omega) \in [0, T] \times \mathcal{D}_0 : \omega_{[t, T]} \in \mathcal{C}_t\right\} \qquad \text{and} \qquad \Lambda := [0, T] \times \mathcal{C}\left([0, T], \mathbb{R}^d\right),$$

where $\omega_{[t,T]}$ refers to the truncation of the path ω to the interval $[t,T]$. We denote by $\mathcal{C}(\overline{\Lambda})$ the space of all functions on $\overline{\Lambda}$, continuous with respect to the distance function $\mathbf{d}((t,\omega),(t',\omega')) := |t-t'| + \|\omega - \omega'\|_T$. For a given $u \in \mathcal{C}(\overline{\Lambda})$, we define its (right) time derivative as

$$\partial_t u(t,\omega) := \lim_{\varepsilon \downarrow 0} \frac{u(t+\varepsilon,\omega) - u(t,\omega)}{\varepsilon}. \qquad \text{for all } (t,\omega) \in \overline{\Lambda}.$$

Following [73], we then define spatial derivatives of $u \in \mathcal{C}(\overline{\Lambda})$ as linear or bilinear operators on $\mathcal{C}_t$:

Definition A.4. The spatial derivatives of $u \in \mathcal{C}(\overline{\Lambda})$ are defined as Fréchet derivatives. For any $(t,\omega) \in \overline{\Lambda}$,

$$\langle \partial_\omega u(t,\omega), \eta \rangle \quad := \lim_{\varepsilon \downarrow 0} \frac{u(t,\omega + \varepsilon\eta_{[t,T]}) - u(t,\omega)}{\varepsilon}, \qquad \text{for any } \eta \in \mathcal{C}_t,$$

$$\langle \partial_\omega^2 u(t,\omega), (\eta,\zeta) \rangle \quad := \lim_{\varepsilon \downarrow 0} \frac{\langle \partial_\omega u(t,\omega + \varepsilon\eta_{[t,T]}), \zeta \rangle - \langle \partial_\omega u(t,\omega), \zeta \rangle}{\varepsilon}, \quad \text{for any } \eta, \zeta \in \mathcal{C}_t.$$

This definition of the spatial derivative in the direction $\eta \in \mathcal{C}_t$ is obviously equivalent to

$$\langle \partial_\omega u(t,\omega), \eta \rangle = \frac{\mathrm{d}}{\mathrm{d}\varepsilon} u\left(t, \omega + \varepsilon\eta_{[t,T]}\right)\big|_{\varepsilon=0}.$$

This definition is consistent with that of Dupire [20], as the perturbation acts on the time interval $[t,T]$, but not on $[0,t]$, and the distance function $\mathbf{d}(\cdot)$ is similar to Dupire's pseudo-distance (see also [68]). We shall further need the following two spaces:

$$\mathcal{C}^{1,2}(\overline{\Lambda}) := \left\{ u \in \mathcal{C}(\overline{\Lambda}) : \varphi \in \mathcal{C}(\overline{\Lambda}) \text{ for } \varphi \in \{\partial_t u, \partial_\omega u, \partial_\omega^2 u\} \right\},$$

$$\mathcal{C}_+^{1,2}(\overline{\Lambda}) := \left\{ u \in \mathcal{C}^{1,2}(\overline{\Lambda}) : \varphi \text{ has polynomial growth for } \varphi \in \{\partial_t u, \partial_\omega u, \partial_\omega^2 u\} \right.$$

$$\left. \text{and } \langle \partial_\omega^2 u, (\eta,\eta) \rangle \text{ is locally uniformly continuous in } \omega \text{ with polynomial growth} \right\}.$$

The definition of polynomial growth here is as follows:

Definition A.5 (Definition 3.3 in [73]). Let $u \in \mathcal{C}(\overline{\Lambda})$ such that $\partial_\omega u$ is well defined on $\overline{\Lambda}$. The functional $\partial_\omega u$ is said to have polynomial growth if

$$|\langle \partial_\omega u(t,\omega), \eta \rangle| \leq C\left(1 + \|\omega\|_T^\alpha\right) \|\eta_{[t,T]}\|_T, \qquad \text{for all } (t,\omega) \in \overline{\Lambda}, \eta \in \mathcal{C}_t,$$

for some $C, \alpha > 0$. It is continuous if $\overline{\Lambda} \ni (t,\omega) \mapsto \langle \partial_\omega u(t,\omega), \eta \rangle$ is continuous under $\mathbf{d}$ for every $\eta \in \mathcal{C}$.

We now recall the main result by Viens and Zhang [73, Theorem 3.10 and Theorem 3.17], extending the Itô formula to the stochastic Volterra framework, for both regular and singular cases. The issue with the singular case (Definition A.3(ii)) is that the coefficients $\mathbf{b}$ and $\boldsymbol{\sigma}$ do not belong to $\mathcal{C}_t$ any longer, so that the Fréchet derivatives in Definition A.4 do not make sense any more. In order to develop an Itô formula, those need to be amended. We refer the reader to [73, Definition 3.16] for a precise definition of the space $\mathcal{C}_{+\beta}^{1,2}(\overline{\Lambda})$, where $\beta \in (0,1)$ intuitively monitors the rate of explosion on the (time) diagonal.

Theorem A.6. For $t \in [0,T]$, define $\mathbf{Z}^t := \mathbf{X} \otimes^t \Theta^t$, and let $\varphi^{t,\omega} := \varphi(\cdot, t, \omega)$ for $\varphi \in \{\mathbf{b}, \boldsymbol{\sigma}\}$ to emphasise the time dependence of the coefficients. Under Assumptions A.1-A.2, the following Itô formula holds:

$$\mathrm{d}u\left(t, \mathbf{Z}^t\right) = \left(\partial_t u\left(t, \mathbf{Z}^t\right) + \left\langle \partial_\omega u\left(t, \mathbf{Z}^t\right), \mathbf{b}^{t,\mathbf{X}} \right\rangle + \frac{1}{2}\left\langle \partial_\omega^2 u\left(t, \mathbf{Z}^t\right), \left(\boldsymbol{\sigma}^{t,\mathbf{X}}, \boldsymbol{\sigma}^{t,\mathbf{X}}\right) \right\rangle \right) \mathrm{d}t$$

$$+ \left\langle \partial_\omega u\left(t, \mathbf{Z}^t\right), \boldsymbol{\sigma}^{t,\mathbf{X}} \right\rangle \mathrm{d}\mathbf{W}_t,$$

(1) in the regular case (Assumption A.3(i)), whenever $u \in \mathcal{C}_+^{1,2}(\Lambda)$;

(2) in the singular case (Assumption A.3(ii)) for $u \in \mathcal{C}_{+,\beta}^{1,2}(\Lambda)$ with $\beta + h - \frac{1}{2} > 0$, where the spatial derivatives should be understood in the regularised sense:

$$\langle \partial_\omega u(t,\omega), \phi \rangle := \lim_{\delta \downarrow 0} \langle \partial_\omega u(t,\omega), \phi^\delta \rangle \qquad and \qquad \langle \partial_\omega^2 u(t,\omega), (\phi, \phi) \rangle := \lim_{\delta \downarrow 0} \langle \partial_\omega^2 u(t,\omega), (\phi^\delta, \phi^\delta) \rangle,$$

with the truncated function $\phi^\delta(t,s,\omega) := \phi\left(t \vee (s+\delta), s, \omega\right)$.

In the theorem, we invoked the space $\mathcal{C}_+^{1,2}(\Lambda)$, which represents the space of functions $u \in: \Lambda \to \mathbb{R}$ such that there exists $v \in \mathcal{C}_+^{1,2}(\overline{\Lambda})$ for which $v = u$ on Λ. The derivatives are defined similarly as restrictions on Λ.

APPENDIX B. PROOF OF THEOREM 3.3

Since the path-dependent PDE in the theorem is linear, existence and uniqueness of the solution is well known, and we refer to [22]. We consider a self-financing portfolio Π consisting of the derivative P given in (3.1), some quantity Δ of stock and some other derivative Ψ, i.e. at any time $t \in [0, T]$,

$$\Pi_t = P_t - \Delta_t S_t - \gamma_t \Psi_t.$$

From Theorem A.6, we can write a functional Itô formula for the option price using Lemma 3.1 under the pricing measure $\mathbb{Q}$:

$$\mathrm{d}P_t = \mathrm{d}\mathrm{P}\left(t, S_t, \boldsymbol{\Theta}^t\right) = \mathcal{A}\mathrm{P}\mathrm{d}t + l(t, \mathbf{X}_t) S_t \partial_x \mathrm{P} \mathrm{d}W_t + \xi(V_t) \left\langle \partial_\omega \mathrm{P}, \mathrm{K}^t \right\rangle \mathrm{d}B_t,$$

where again $\mathbf{X}_t = (S_t, V_t)$, and

$$\mathcal{A}\mathrm{P} := \partial_t \mathrm{P} + r_t S_t \partial_x \mathrm{P} + \frac{l(t, \mathbf{X}_t)^2}{2} S_t^2 \partial_x^2 \mathrm{P} + \frac{\xi(V_t)^2}{2} \left\langle \partial_\omega^2 \mathrm{P}, (\mathrm{K}^t, \mathrm{K}^t) \right\rangle + b(V_t) \left\langle \partial_\omega \mathrm{P}, \mathrm{K}^t \right\rangle + l(t, \mathbf{X}_t) \rho \xi(V_t) S_t \left\langle \partial_{x,\omega} \mathrm{P}, \mathrm{K}^t \right\rangle.$$

Here, the x-derivative refers to the classical derivative with respect to the second component S_t, whereas the ω-derivative is the Fréchet-type derivative in the direction given by the one-dimensional path ω. Applying directly Theorem A.6 with the SDE (2.2) for the process $\mathbf{X}$, we should normally obtain terms of the form $\left\langle \partial_{\boldsymbol{\omega}} \mathrm{P}, \mathbf{b}^t \right\rangle$ (and similarly for the second and the cross derivatives), where $\mathbf{b}_r^t = \mathbf{b}(r, t, \mathbf{X}_r)$ for $r \in [0, T]$ is in $\mathbb{R}^2$ and $\boldsymbol{\omega} = (\omega_1, \omega_2)$ the two-dimensional path. We can then write

$$\left\langle \partial_{\boldsymbol{\omega}} \mathrm{P}\left(t, S_t, \boldsymbol{\Theta}^t\right), \mathbf{b}^t \right\rangle = \mu_t S_t \partial_{\omega_1} \mathrm{P}\left(t, S_t, \boldsymbol{\Theta}^t\right) + \left\langle \partial_{\omega_2} \mathrm{P}\left(t, S_t, \boldsymbol{\Theta}^t\right), \mathrm{K}^t b(V_t) \right\rangle$$
$$= \mu_t S_t \partial_x \mathrm{P}\left(t, S_t, \boldsymbol{\Theta}^t\right) + b(V_t) \left\langle \partial_{\omega_2} \mathrm{P}\left(t, S_t, \boldsymbol{\Theta}^t\right), \mathrm{K}^t \right\rangle.$$

This yields, similarly, for the portfolio Π, under $\mathbb{Q}$,

$$\mathrm{d}\Pi_t = \mathrm{d}\mathrm{P} - \Delta_t \mathrm{d}S_t - \gamma_t \mathrm{d}\Psi_t$$
$$= \mathcal{A}\mathrm{P}\mathrm{d}t + l(t, \mathbf{X}_t) S_t \partial_x \mathrm{P} \mathrm{d}W_t + \xi(V_t) \left\langle \partial_\omega \mathrm{P}, \mathrm{K}^t \right\rangle \mathrm{d}B_t$$
$$- \Delta_t \left(\mu_t S_t \mathrm{d}t + l(t, \mathbf{X}_t) S_t \mathrm{d}W_t \right) - \gamma_t \left(\mathcal{A}\Psi_t \mathrm{d}t + l(t, \mathbf{X}_t) S_t \partial_x \Psi_t \mathrm{d}W_t + \xi(V_t) \left\langle \partial_\omega \Psi_t, \mathrm{K}^t \right\rangle \mathrm{d}B_t \right).$$

The portfolio is risk free if $\mathrm{d}\Pi_t = r\Pi_t \mathrm{d}t$ and the random noise is cancelled, meaning that

$$(\mathrm{B}.1) \qquad \begin{cases} \left(\mathcal{A}\mathrm{P} - \gamma_t \mathcal{A}\Psi_t - \Delta_t r_t S_t \right)\mathrm{d}t & = r_t \left(\mathrm{P} - \Delta_t S_t - \gamma_t \Psi_t \right)\mathrm{d}t, \\ \partial_x \mathrm{P} - \gamma_t \partial_x \Psi_t - \Delta_t & = 0, \\ \left\langle \partial_\omega \mathrm{P}, \mathrm{K}^t \right\rangle - \gamma_t \left\langle \partial_\omega \Psi_t, \mathrm{K}^t \right\rangle & = 0, \end{cases}$$

since both functions $l(\cdot, \cdot, \cdot)$ and $\xi(\cdot)$ are nowhere null. The last two equalities yield

$$\gamma_t = \frac{\left\langle \partial_\omega \mathrm{P}, \mathrm{K}^t \right\rangle}{\left\langle \partial_\omega \Psi_t, \mathrm{K}^t \right\rangle} \qquad \text{and} \qquad \Delta_t = \partial_x \mathrm{P} - \frac{\left\langle \partial_\omega \mathrm{P}, \mathrm{K}^t \right\rangle}{\left\langle \partial_\omega \Psi_t, \mathrm{K}^t \right\rangle} \partial_x \Psi_t.$$

We can now rewrite the first equality in (B.1) as

$$\left(\mathcal{A}\mathrm{P} - \gamma_t \mathcal{A}\Psi_t - \Delta_t r_t S_t \right)\mathrm{d}t = r_t \left(\mathrm{P} - \Delta_t S_t - \gamma_t \Psi_t \right)\mathrm{d}t,$$

which is equivalent to

$$\frac{\left(\mathcal{A} - r_t \right)\mathrm{P}}{\left\langle \partial_\omega \mathrm{P}, \mathrm{K}^t \right\rangle} = \frac{\left(\mathcal{A} - r_t \right)\Psi_t}{\left\langle \partial_\omega \Psi_t, \mathrm{K}^t \right\rangle}.$$

The left-hand side is a function of P only, whereas the right-hand side only depends on Ψ. Therefore, the only way for this equality to hold is for both sides to be equal to some function $-\widehat{b}$ that depends on S_t, Θ^t and t, but not on P nor Ψ. The pricing equation for the price function is therefore

$$\left(\mathcal{A} - r_t \right)\mathrm{P} = - \left\langle \partial_\omega \mathrm{P}, \mathrm{K}^t \right\rangle \widehat{b}_t.$$

Following similar computations in classical (Markovian) stochastic volatility models, we consider $\widehat{b}_t$ of the form $\widehat{b}_t = b(V_t) - \xi(V_t)\lambda_t$, where λ_t is called the market price of risk. The final pricing PDE is therefore

$$\partial_t + r_t S_t \partial_x + \frac{l(t, \mathbf{X}_t)^2}{2} S_t^2 \partial_x^2 + \frac{\xi(V_t)^2}{2} \left\langle \partial_\omega^2, (\mathrm{K}^t, \mathrm{K}^t) \right\rangle + b(V_t) \left\langle \partial_\omega, \mathrm{K}^t \right\rangle + l(t, \mathbf{X}_t) \rho \xi(V_t) S_t \left\langle \partial_{x,\omega}, \mathrm{K}^t \right\rangle + \left\langle \partial_\omega, \mathrm{K}^t \right\rangle \widehat{b}_t = r_t.$$

With a slight abuse of notations, writing b in place of $\widehat{b}$ proves the statement.

Appendix C. Simulation of the rough Heston model

We provide here details about the simulation of the rough Heston model in (5.1). Introducing the infinite-dimensional process $(\Theta^t)_{t \geq 0}$ as above, we can write, for any $t \geq 0$ and $u \geq t$,

$$(C.1) \qquad \Theta_u^t = V_0 + \int_0^t \mathrm{K}(u-s) \Big[\kappa(\theta - V_s)\mathrm{d}s + \xi\sqrt{V_s}\mathrm{d}B_s \Big].$$

Given a fixed time horizon $T > 0$ and a given number of time steps n, we introduce an equidistant grid for the closed interval $[0, T]$ as $t_i = i/n$, for $i = 0, \ldots, n$. Discretising the rough SDE for the variance process in (5.1) along this grid, and denoting $V_i = V_{t_i}$ for simplicity, we can write $V_0 = V_0$ and, for any $i = 1, \ldots, n$,

$$V_i = V_0 + \int_0^{t_i} \mathrm{K}(t_i - s)\Big[\kappa(\theta - V_s)\mathrm{d}s + \xi\sqrt{V_s}\mathrm{d}B_s\Big]$$

$$\approx V_0 + \sum_{j=0}^{i-1} \kappa(\theta - V_j) \int_{t_j}^{t_{j+1}} \mathrm{K}(t_i - s)\mathrm{d}s + \sum_{j=0}^{i-1} \xi\sqrt{V_j} \int_{t_j}^{t_{j+1}} \mathrm{K}(t_i - s)\mathrm{d}B_s$$

$$(C.2) \qquad = V_0 + \sum_{j=0}^{i-1} \kappa(\theta - V_j)A_{j,i} + \sum_{j=0}^{i-2} \xi\sqrt{V_j} \int_{t_j}^{t_{j+1}} \mathrm{K}(t_i - s)\mathrm{d}B_s + \xi\sqrt{V_{i-1}} \int_{t_{i-1}}^{t_i} \mathrm{K}(t_i - s)\mathrm{d}B_s,$$

where we freeze the variance process on each subinterval to its left-point value, and single out the singular part of the kernel in the last integral. We also introduced the quantity

$$A_{j,i} := \int_{t_j}^{t_{j+1}} \mathrm{K}(t_i - s)\mathrm{d}s, \qquad \text{for } i = 1, \ldots, n \text{ and } j = 0, \ldots, i-1,$$

which can be pre-computed and stored. Note that, for $i = 1$, the middle sum in (C.2) does not appear. Following [10], we can write the middle term in the discretisation as

$$\sum_{j=0}^{i-2} \xi\sqrt{V_j} \int_{t_j}^{t_{j+1}} \mathrm{K}(t_i - s)\mathrm{d}B_s = \sum_{j=0}^{i-2} \xi\sqrt{V_j}\,\mathrm{K}\left(\frac{b_{i-j}^*}{n}\right) \int_{t_j}^{t_{j+1}} \mathrm{d}B_s = \sum_{k=2}^{i} \xi\sqrt{V_{i-k}}\,\mathrm{K}\left(\frac{b_k^*}{n}\right) \int_{t_{i-k}}^{t_{i-k+1}} \mathrm{d}B_s$$

$$(C.3) \qquad\qquad =: \sum_{k=2}^{i} \xi\sqrt{V_{i-k}}\,\mathrm{K}\left(\frac{b_k^*}{n}\right) \overline{B}_{i-k},$$

with $\overline{B}_i := \int_{t_i}^{t_{i+1}} \mathrm{d}B_s$ for $i = 0, \ldots, n-1$. Finally, for the last term in (C.2), where the singularity occurs, we introduce the vector $(\widetilde{B}_i)_{i=0,\ldots,n-1}$ as $\widetilde{B}_i := \int_{t_i}^{t_{i+1}} \mathrm{K}(t_{i+1} - s)\mathrm{d}B_s$. In the notations of [10], $\widetilde{B}_i$ is denoted $\overline{B}_{i,1}$, but we remove the double index here. For any $i = 0, \ldots, n-1$, the couple $(\overline{B}_i, \widetilde{B}_i)$ forms a two-dimensional Gaussian vector, with covariance matrix Σ given by

$$\Sigma_{11} = \frac{1}{n}, \qquad \Sigma_{22} = \int_{t_i}^{t_{i+1}} \mathrm{K}(t_{i+1} - s)^2\mathrm{d}s, \qquad \Sigma_{12} = \Sigma_{21} = \int_{t_i}^{t_{i+1}} \mathrm{K}(t_{i+1} - s)\mathrm{d}s.$$

Summarising, we discretise the variance process as

$$(C.4) \qquad V_i = V_0 + \sum_{j=0}^{i-1} \kappa(\theta - V_j)A_{j,i} + \sum_{k=2}^{i} \xi\sqrt{V_{i-k}}\,\mathrm{K}\left(\frac{b_k^*}{n}\right) \overline{B}_{i-k} + \xi\sqrt{V_{i-1}}\,\widetilde{B}_{i-1}, \qquad \text{for } i = 1, \ldots, n.$$

Remark C.1. For computational purposes, the steps above can be sped up bearing in mind that the matrix $(A_{j,i})_{i,j}$ is a strictly upper triangular Toeplitz matrix and that the last expression on the right-hand side of (C.3) can be computed as a discrete Fourier transform.

Remark C.2. In the power law case $\mathrm{K}(t) = t^{H-\frac{1}{2}}$ with $H \in (0,1)$, the expressions above simplify to

$$A_{j,i} = \frac{1}{H + \frac{1}{2}} \left\{ (t_i - t_j)^{H+\frac{1}{2}} - (t_i - t_{j+1})^{H+\frac{1}{2}} \right\},$$

$$(C.5) \qquad \Sigma = \begin{pmatrix} \dfrac{1}{n} & \dfrac{1}{(H+\frac{1}{2})n^{H+\frac{1}{2}}} \\ \dfrac{1}{(H+\frac{1}{2})n^{H+\frac{1}{2}}} & \dfrac{1}{2Hn^{2H}} \end{pmatrix},$$

and, as shown in [10], the coefficients (b_k^*) are explicitly computed as

$$(C.6) \qquad b_k^* := \left(\frac{k^{H+\frac{1}{2}} - (k-1)^{H+\frac{1}{2}}}{H+\frac{1}{2}} \right)^{\frac{1}{H-1/2}}.$$

In this case, with the uniform grid $t_i = i/n$, denoting $\widetilde{A}_{i-j} := A_{j,i}$, we can rewrite (C.4) as

$$V_i = V_0 + \sum_{j=0}^{i-1} \kappa(\theta - V_j)\widetilde{A}_{i-j} + \sum_{k=2}^{i} \xi\sqrt{V_{i-k}}\, \mathrm{K}\left(\frac{b_k^*}{n}\right) \overline{B}_{i-k} + \xi\sqrt{V_{i-1}}\,\widetilde{B}_{i-1}, \qquad \text{for } i = 1,\dots,n,$$

and the vector $(\widetilde{A}_k)_{k=1,\dots,n}$ reads

$$\widetilde{A}_k = \frac{k^{H+\frac{1}{2}} - (k-1)^{H+\frac{1}{2}}}{\left(H+\frac{1}{2}\right)n^{H+\frac{1}{2}}}.$$

Regarding the process Θ in (C.1), we discretise it analogously as $\Theta_k^i = V_k$ whenever $k \le i$ and, for $k > i$,

$$\Theta_k^i = V_0 + \int_0^{t_i} \mathrm{K}(t_k - s)\Big[\kappa(\theta - V_s)\mathrm{d}s + \xi\sqrt{V_s}\mathrm{d}B_s\Big] = V_0 + \sum_{j=0}^{i-1} \int_{t_j}^{t_{j+1}} \mathrm{K}(t_k - s)\Big[\kappa(\theta - V_s)\mathrm{d}s + \xi\sqrt{V_s}\mathrm{d}B_s\Big]$$

$$\approx V_0 + \sum_{j=0}^{i-1} \kappa(\theta - V_j)\int_{t_j}^{t_{j+1}} \mathrm{K}(t_k - s)\mathrm{d}s + \xi\sum_{j=0}^{i-1}\sqrt{V_j}\int_{t_j}^{t_{j+1}} \mathrm{K}(t_k - s)\mathrm{d}B_s$$

$$= V_0 + \sum_{j=0}^{i-1} \kappa(\theta - V_j)A_{j,k} + \xi\sum_{j=0}^{i-1}\sqrt{V_j}\int_{t_j}^{t_{j+1}} \mathrm{K}(t_k - s)\mathrm{d}B_s$$

$$= V_0 + \sum_{j=0}^{i-1} \kappa(\theta - V_j)A_{j,k} + \xi\sum_{j=0}^{i-1}\sqrt{V_j}\,\mathrm{K}\left(\frac{b_{k-j}^*}{n}\right)\overline{B}_j$$

$$= \left(V_0 + \sum_{j=0}^{i-2} \kappa(\theta - V_j)A_{j,k} + \xi\sum_{j=0}^{i-2}\sqrt{V_j}\,\mathrm{K}\left(\frac{b_{k-j}^*}{n}\right)\overline{B}_j\right) + \kappa(\theta - V_{i-1})A_{i-1,k} + \xi\sqrt{V_{i-1}}\,\mathrm{K}\left(\frac{b_{k-(i-1)}^*}{n}\right)\overline{B}_{i-1}$$

$$= \Theta_k^{i-1} + \kappa(\theta - V_{i-1})A_{i-1,k} + \xi\sqrt{V_{i-1}}\,\mathrm{K}\left(\frac{b_{k-(i-1)}^*}{n}\right)\overline{B}_{i-1}.$$

For the stock price, starting from $S_{t_0} = S_0$, we use the discretised explicit form, for $i = 1,\dots,n$,

$$(C.7) \qquad S_i = S_{i-1}\exp\left\{-\frac{V_{i-1}}{2n} + \sqrt{V_{i-1}}\,W_{i-1}\right\},$$

where $W_i := \int_{t_i}^{t_{i+1}} \mathrm{d}W_s$ for some standard Brownian motion W such that $\mathrm{d}\langle W, B\rangle_t = \rho\,\mathrm{d}t$.

Remark C.3. The simulation recipe is as follows:

- Pre-compute the vector (b_k^*) in (C.6) for $k = 1,\dots,n$;
- Generate three iid Gaussian samples $(\mathrm{N}_i^1, \mathrm{N}_i^2, \mathrm{N}_i^3)$ for $i = 0,\dots,n-1$;
- Recalling the covariance matrix Σ in (C.5), compute the Gaussian vector $(\overline{B}_i, \widetilde{B}_i, W_i)_{i=0,\dots,n-1}$ as

$$\begin{pmatrix} \overline{B}_i \\ \widetilde{B}_i \\ W_i \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \rho/n \\ \Sigma_{12} & \Sigma_{22} & \rho\Sigma_{12} \\ \rho/n & \rho\Sigma_{12} & 1/n \end{pmatrix}^{1/2} \begin{pmatrix} \mathrm{N}_i^1 \\ \mathrm{N}_i^2 \\ \mathrm{N}_i^3 \end{pmatrix},$$

since $\mathbb{E}[W_i\widetilde{B}_i] = \rho\int_{t_i}^{t_{i+1}} \mathrm{K}(t_{i+1} - s)\mathrm{d}s = \rho\Sigma_{12}$, and the square-root understood in the Cholesky sense;
- Compute $(V_i)_{i=0,\dots,n}$ using (C.4) and $(S_i)_{i=0,\dots,n}$ using (C.7);
- For each $i = 0,\dots,n$, generate the discretised curves $(\Theta_k^i)_{k=0,\dots,n}$.

References

[1] E. Abi Jaber. Lifting the Heston model. To appear in *Quantitative Finance*.

[2] E. Abi Jaber, M. Larsson, S. Pulido. Affine Volterra processes. To appear in *Annals of Applied Probability*.

[3] E. Alòs, J. León and J. Vives. On the short-time behavior of the implied volatility for jump-diffusion models with stochastic volatility. *Finance and Stochastics*, 11(4), 571-589, 2007.

[4] J. Ba and D. Kingma. Adam: a method for stochastic optimization. *Proceedings of the International Conference on Learning Representations*, May 2015.

[5] G. Barles and P.E. Souganidis. Convergence of approximation schemes for fully nonlinear second-order equation. *Asymptotic Analysis*, 4: 271-283, 1991.

[6] C. Bayer, C. Ben Hammouda and R. Tempone. Hierarchical adaptive sparse grids for option pricing under the rough Bergomi model. arXiv:1812.08533, 2018.

[7] C. Bayer, P. Friz, P. Gassiat, J. Martin and B. Stemper. A regularity structure for rough volatility. To appear in *Mathematical Finance*.

[8] C. Bayer, P. Friz and J. Gatheral. Pricing under rough volatility. *Quantitative Finance*, 16(6): 1-18, 2015.

[9] C. Bayer, P. Friz, A. Gulisashvili, B. Horvath and B. Stemper. Short-time near the money skew in rough fractional stochastic volatility models. To appear in *Quantitative Finance*.

[10] M. Bennedsen, A. Lunde and M.S. Pakkanen. Hybrid scheme for Brownian semistationary processes. *Finance and Stochastics*, 21(4): 931-965, 2017.

[11] M. Bennedsen, A. Lunde and M.S. Pakkanen. Decoupling the short- and long-term behavior of stochastic volatility. arXiv:1610.00332, 2017.

[12] M.A. Berger and V.J. Mizel. Volterra Equations with Itô Integrals, I. *Journal of Integral Equations*, 2(3): 187-245, 1980.

[13] M.A. Berger and V.J. Mizel. Volterra Equations with Itô Integrals, II. *Journal of Integral Equations*, 2(4): 319-337, 1980.

[14] F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Econ.*, 81(3): 637-654, 1973.

[15] G. Callegaro, M. Grasselli and G. Pagès. Fast hybrid schemes for fractional Riccati equations (rough is not so tough). To appear in *Mathematics of Operations Research*.

[16] F. Comte and E. Renault. Long memory continuous time models. *Journal of Econometrics*, 73(1): 101-149, 1996.

[17] L. Decreusefond and A. Ustünel. Stochastic analysis of fractional Brownian motion. *Potential Analysis*, 10: 177-214, 1996.

[18] M. Djehiche and M. Eddahbi. Hedging options in market models modulated by the fractional Brownian motion. *Stochastic Analysis and Applications*, 19(5): 753-770, 2001.

[19] B. Dupire. Pricing with a smile. *Risk*, 1994.

[20] B. Dupire. Functional Itô Calculus. *Quantitative Finance*, 19(5): 721-729, 2019.

[21] W. E, J. Han and A. Jentzen. Solving high-dimensional partial differential equations using deep learning. To appear in *Proceedings of the Natural Academy of Sciences*, 2018.

[22] I. Ekren, N. Touzi and J. Zhang. On viscosity solutions of path-dependent PDEs. *Annals Prob.*, 42(1): 204-236, 2014.

[23] I. Ekren, N. Touzi and J. Zhang. Viscosity solutions of fully nonlinear parabolic path-dependent PDEs: Part I. *Annals of Probability*, 44(2): 1212-1253, 2016.

[24] I. Ekren, N. Touzi and J. Zhang. Viscosity solutions of fully nonlinear parabolic path-dependent PDEs: Part II. *Annals of Probability*, 44(4): 2507-2553, 2016.

[25] N. El Karoui, S. Peng and M.C. Quenez. Backward stochastic differential equations in Finance. *Math. Fin.*, 7(1): 1-71, 1997.

[26] O. El Euch and M. Rosenbaum. The characteristic function of rough Heston models. *Math. Finance*, 29(1): 3-38, 2019.

[27] O. El Euch and M. Rosenbaum. Perfect hedging in rough Heston models. *Annals Applied Prob.*, 28(6): 3813-3856, 2018.

[28] O. El Euch, M. Fukasawa and M. Rosenbaum. The microstructural foundations of leverage effect and rough volatility. *Finance and Stochastics*, 22(2): 241-280, 2018.

[29] O. El Euch, J. Gatheral and M. Rosenbaum. Roughening Heston. *Risk*, April 2019.

[30] X. Fernique. Intégrabilité des vecteurs Gaussiens. *CRAS Paris*, 270: 1698-1699, 1970.

[31] M. Forde and H. Zhang. Asymptotics for rough stochastic volatility models. *SIAM Fin. Math.*, 8: 114-145, 2017.

[32] M. Fukasawa. Asymptotic analysis for stochastic volatility: martingale expansion. *Finance and Stoch.*, 15: 635-654, 2011.

[33] P. Gassiat. On the martingale property in the rough Bergomi model. To appear in *Electronic Comm. Prob.*, 2019.

[34] J. Gatheral. The Volatility Surface: a practitioner's guide. John Wiley & Sons, 2006.

[35] J. Gatheral and A. Jacquier. Arbitrage-free SVI volatility surfaces. *Quantitative Finance*, 14: 59-71, 2014.

[36] J. Gatheral, T. Jaisson and M. Rosenbaum. Volatility is rough. *Quantitative Finance*, 18(6): 933-949, 2018.

[37] J. Gatheral and M. Keller-Ressel. Affine forward variance models. To appear in *Finance and Stochastics*, 2019.

[38] J. Gatheral and R. Radoičić. Rational approximation of the rough Heston solution. *IJTAF*, 22(3), 2019.

[39] H. Guennoun, A. Jacquier, P. Roome and F. Shi. Asymptotic behaviour of the fractional Heston model. *SIAM Journal on Financial Mathematics*, 9(3): 1017-1045, 2018.

[40] A. Gulisashvili. Large deviation principle for Volterra type fractional stochastic volatility models. *SIAM Journal on Financial Mathematics*, 9(3): 1102-1136, 2018.

[41] J. Guyon and P. Henry-Labordère. Nonlinear option pricing, CRC Press, 2013.

[42] I. Gyöngy. Mimicking the one-dimensional marginal distributions of processes vaving an Itô differential. *Probability Theory and Related Fields*, 71(4): 501-516, 1986.

[43] P. Harms. Strong convergence rates for Markovian representations of fractional Brownian motion. arXiv:1902.01471, 2019.

[44] C. Heinrich, M. Pakkanen and A.E.D. Veraart. Hybrid simulation scheme for volatility modulated moving average fields. *Mathematics and Computers in Simulation*, 166: 224-244, 2019.

[45] B. Horvath, A. Jacquier and C. Lacombe. Asymptotic behaviour of randomised fractional volatility models. *Journal of Applied Probability*, **56**(2), 2019.

[46] B. Horvath, A. Jacquier and A. Muguruza. Functional central limit theorems for rough volatility arXiv:1711.03078, 2018.

[47] B. Horvath, A. Jacquier and P. Tankov. Volatility options in rough volatility models. To appear in *SIAM Fin. Math.*

[48] B. Horvath, A. Muguruza and M. Tomas. Deep learning volatility. SSRN:3322085, 2019.

[49] C. Huré, H. Pham and X.Warin. Some machine learning schemes for high-dimensional nonlinear PDEs. To appear in *Mathematics of Computation*.

[50] H.E. Hurst. Long-term storage capacity of reservoirs. *Trans. American Society of Civil Engineers*, **116**(1): 770-799, 1951.

[51] H.E. Hurst, R.P. Black and Y.M. Simaika. Long-term storage: an experimental study. *London: Constable,* 1965.

[52] A. Jacquier, C. Martini and A. Muguruza. On VIX futures in the rough Bergomi model. *Quant. Finance*, **18**(1): 45-61, 2018.

[53] A. Jacquier, M. Pakkanen and H. Stone. Pathwise large deviations for the rough Bergomi model. *Journal of Applied Probability*, **55**(4): 1078-1092, 2018.

[54] B. Jourdain. Loss of martingality in asset price models with lognormal stochastic volatility. Preprint *CERMICS*, 2004.

[55] B. Jourdain and A. Zhou. Existence of a calibrated regime switching local volatility model and new fake Brownian motions. To appear in *Mathematical Finance.*

[56] I. Karatzas and S. Shreve. Brownian motion and stochastic calculus. Springer-Verlag, New-York, 1988.

[57] D. Lacker, M. Shkolnikov and J. Zhang. Inverting the Markovian projection, with an application to local stochastic volatility models. To appear in *Annals of Probability*.

[58] I. Lagaris, A. Likas, and D. Fotiadis, Artificial neural networks for solving ordinary and partial differential equations. *IEEE Transactions on Neural Networks*, **9**(5): 987-1000, 1998.

[59] H. Lee. Neural algorithm for solving differential equations. *Journal of Computational Physics*, **91**: 110-131, 1990.

[60] A. Lewis. A simple option formula for general jump-diffusion and other exponential Lévy processes, Available at optioncity.net/pubs/ExpLevy.pdf, 2001.

[61] B. Mandelbrot and J. Van Ness. Fractional Brownian motions, fractional noises and applications. *SIAM Review*, **10**(4): 422-437, 1968.

[62] R. McCrickerd and M.S. Pakkanen. Turbocharging Monte Carlo pricing for the rough Bergomi model. *Quantitative Finance*, **18**(11), 1877-1886, 2018.

[63] W. McGhee. An artificial neural network representation of the SABR stochastic volatility model. SSRN:3288882, 2018.

[64] Q. Chan-Wai-Nam, J. Mikael and X. Warin. Machine Learning for semi linear PDEs *Journal of Scientific Computing*, **79**(3): 1667-1712, 2019.

[65] P. Parczewski. Donsker-type theorems for correlated geometric fractional Brownian motions and related processes. *Electronic Communications in Probability*, **22**(55): 1-13, 2017.

[66] J. Picard. Representation formulae for the fractional Brownian motion. *Séminaire de Probabilités*, **43**: 3-70, 2011.

[67] Z. Ren and X. Tan. On the convergence of monotone schemes for path-dependent PDEs. *Stochastic Processes and their Applications*, **127**(6): 1738-1762, 2017.

[68] Z. Ren, N. Touzi and J. Zhang, An overview of viscosity solutions of path-dependent PDEs. *Stochastic Analysis and Applications*. Springer Proceedings in Mathematics and Statistics, **100**: 397-454, 2014.

[69] M. Romano and N. Touzi. Contingent claims and market completeness in a stochastic volatility model. *Mathematical Finance*, **7**(4): 399-412, 1997.

[70] M. Sabate Vidales, D. Šiška and L. Szpruch. Martingale functional control variates via deep learning. arXiv:1810.05094, 2018.

[71] J. Sirignano and K. Spiliopoulos. DGM: A deep learning algorithm for solving partial differential equations. *Journal of Computational Physics*, **375**: 1339-1364, 2018

[72] H. Stone. Calibrating rough volatility models: a convolutional neural network approach. To appear in *Quantitative Finance.*

[73] F. Viens and J. Zhang. A martingale approach for fractional Brownian motions and related path-dependent PDEs. To appear in *Annals of Applied Probability.*

[74] J. Zhang and J. Zhuo. Monotone schemes for fully nonlinear parabolic path dependent PDEs. *Journal Fin. Eng.*, **1**, 2014.

DEPARTMENT OF MATHEMATICS, IMPERIAL COLLEGE LONDON, AND ALAN TURING INSTITUTE
E-mail address: `a.jacquier@imperial.ac.uk`

LLOYDS BANKING GROUP PLC, COMMERCIAL BANKING, 10 GRESHAM STREET, LONDON, EC2V 7AE, UK
E-mail address: `Mugad.Oumgari@lloydsbanking.com`