Calibration of Non-Semimartingale Models - an Adjoint Approach

Christian Bender[1] and Matthias Thiel[1]

June 5, 2024

Abstract

We design and analyze a Monte-Carlo algorithm for calibrating a financial model, in which some quantities (e.g., volatility) are represented in terms of a stochastic differential equation driven by a continuous p-variation process for $p \in (1, 2)$ (e.g., a fractional Brownian motion with Hurst parameter bigger than a half). The p-variation process can be correlated to the Brownian motion, which drives the stock prices, in order to capture the so-called leverage effect. The key tool is an adjoint gradient representation via a new type of anticipating backward stochastic differential equation, which is formulated in terms of the Russo-Vallois forward integral. We provide rates of convergence for an Euler approximation of this adjoint equation. Finally, the results are illustrated by a case study, calibrating a fractional Heston model to market data.

1 Introduction

Stochastic volatility models enjoy great popularity in financial engineering, since they are able to capture several features observed in market data such as the implied volatility smile and the leverage effect (i.e., negative correlation between asset prices and volatility), see [12] for a discussion of stylized facts in stock returns. In classical volatility models of the 20th century [25, 27, 5, 47] assets price and volatility are governed by stochastic differential equations driven by correlated Brownian motions. Motivated by the phenomenon of volatility persistence (large absolute changes in stock returns tend to be followed by larges absolute changes), continuous-time models with volatility processes driven by a fractional Brownian motion with Hurst parameter bigger than a half have been suggested [8, 7, 33, 3, 36] – exploiting the long memory effect of fractional Brownian motion for this range of the Hurst parameter. More recently, the smile expansions in [18] led the

[1]Saarland University, Department of Mathematics, Postfach 151150, D-66041 Saarbrücken, Germany, bender@math.uni-sb.de, thiel@math.uni-sb.de.

authors of [20] to introduce rough volatility models, which correspond to fractional Brownian motion with Hurst parameter smaller than a half, see also [2, 19, 15].

In this paper, we extend the Monte Carlo algorithm of [29] for calibrating stochastic volatility models driven by Brownian motions to a wide class of models including fractional volatility models with Hurst parameter $H > 1/2$ (hence, in the long memory regime). More generally, we consider models consisting of two systems of stochastic differential equations (SDEs). The first one is driven by a continuous p-variation process for some $p \in (1, 2)$, e.g., fractional Brownian motion with Hurst parameter bigger than a half. This SDE is interpreted in the sense of Young integration [50] and may be used to model non-tradable quantities such as the volatilities of stocks or a short rate process governing the term structure of interest rates. The second SDE system is driven by a Brownian motion and is interpreted in the classical Itô sense. Its coefficients depend on the solution of the first system of SDEs and we may think of the solution of the second system as the prices of tradable assets in the market. We assume that the model is not fully specified in the sense that the two systems of SDEs depend on a parameter vector. The objective is to minimize the quadratic error between the model prices and observed market prices for some liquidly traded options over the parameter vector. Borrowing ideas from [29], we design a gradient-based adjoint Monte Carlo algorithm for tackling this problem. However, in contrast to [29], who first discretize the optimization problem in the sense sample average approximation [46], we follow the optimize-then-discretize approach and study the optimization problem in continuous time.

The paper is organized as follows: In Section 2, we discuss the main results. After setting the model dynamics and the optimization problem in continuous time in Subsection 2.1, the gradient of the cost functional with respect to the parameter vector is studied in Subsection 2.2. The main result (Theorem 2.6) is a new adjoint representation of the gradient in terms of an anticipating backward stochastic differential equation which is jointly driven by the p-variation process and the Brownian motion. In order to formulate this equation in a proper way, we make use of the forward integral of Russo and Vallois [43] which encompasses the Itô integral and the Young integral as special cases and, at the same time, extends the Itô integral to non-adapted integrands. The key advantage of the adjoint equation is that its dimension is independent of the number of parameters, while the Fréchet derivative of the original SDE system with respect to the parameter vector solves a linear SDE (sometimes called sensitivity equation, see [29]) whose dimension increases linearly in the number of parameters. Hence, algorithms based on the adjoint equation are expected to be more efficient than those basing on the sensitivity equation, if the dimension of the parameter vector is large. This general advantage of adjoint techniques has been found to be useful in various applications, see, e.g., [22, 21, 39].

Subsection 2.3 is devoted to the Euler discretization of the original SDE system, its sensitivity equation and the adjoint anticipating backward SDE. We provide rates of convergence for the error measured in the supremum norm in time and the L^p-norm in the sample paths (Theorem 2.8). The key technical difficulty is to control the growth of the Euler scheme for the pathwise Young differential equations. Compared to the literature on Young differential equations (e.g., [32]), we have to keep track of the dependence of the constants on the sample paths. To this end we adapt the greedy sequence technique [4, 9] to Euler partitions in a suitable way to come up with a variant of Gronwall's lemma (Lemma 4.5), which is tailor-made for our purposes. The results on the Euler schemes are then applied to estimate the error of approximations to the cost functional and to the adjoint gradient representation.

Section 3 provides some background information on Young integration and Russo-Vallois forward integration, which is required for the proofs of the main results in Section 4. We sketch the proofs of all main results emphasizing the key ideas rather than providing all technical detail. At times, we focus on simplified versions of the equations which already contain the key difficulties. Detailed proofs of all the results in full generality can be found in the second author's PhD thesis [48].

Finally, in Section 5, we present a Monte Carlo algorithm for model calibration based on the discretized adjoint gradient representation, replacing all expectations by empirical means over simulated sample paths. The algorithm is then applied to calibrate a fractional version of Heston's model to call option price data on the EUROSTOXX 50. Our case study suggests that a Hurst parameter of about $H = 0.65$ yields the best fit to the data. Additional numerical experiments illustrate the rates of convergence established in Subsection 2.3.

2 Discussion of the main results

2.1 The model dynamics and the cost functional

In this subsection, we introduce the general setting consisting of two parameter-dependent systems of stochastic differential equations (SDEs). The first SDE is driven by a multivariate stochastic process $(w_t)_{t\in[0,T]}$ which has continuous paths of finite p-variation for some $p \in (1,2)$, which includes, e.g., a fractional Brownian motion with Hurst parameter $H \in (1/2, 1)$. We here recall that a *fractional Brownian motion* $(B_t^H)_{t\in[0,T]}$ with Hurst parameter $H \in (0,1)$ is a centered Gaussian process with covariance structure

$$\mathrm{E}[B_t^H B_s^H] = \frac{1}{2}\left(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right)$$

In the financial application, this first SDE may model several factors,

which are not directly tradable and storeable, such as the volatility of primary assets or the short rate of a money market account. The second SDE is driven by a multidimensional Brownian motion and may be thought to represent, e.g., the price processes of the primary assets traded in the market.

We fix a time horizon $T > 0$ and positive integers $n_1, m_1, n_2, m_2, d \in \mathbb{N} = \{1, 2, \dots\}$. Let $(\Omega, \mathcal{F}, \mathbb{F}, P)$ be a filtered probability space (satisfying the usual conditions) carrying an m_1-dimensional stochastic process $(w_t)_{t\in[0,T]}$, whose paths are almost surely continuous and have finite p-variation for $p \in (1, 2)$, and an m_2-dimensional standard Brownian motion $(B_t)_{t\in[0,T]}$, both adapted to the filtration $\mathbb{F} = (\mathcal{F}_t)_{t\in[0,T]}$, but possibly dependent. The dependence between the two driving processes is crucial to capture, e.g., the leverage effect between stock prices and volatility [16].

Furthermore let $\mathcal{U}$ be an open, convex and bounded subset of $\mathbb{R}^d$, which represents the parameter set. We consider the parameter dependent systems of stochastic differential equations

$$\xi_t^u = \xi_0(u) + \int_0^t b(r, \xi_r^u, u)\, dr + \sum_{j=1}^{m_1} \int_0^t \sigma^j(r, \xi_r^u, u)\, dw_r^j, \tag{1}$$

$$x_t^u = x_0(u) + \int_0^t \hat{b}(r, \xi_r^u, x_r^u, u)\, dr + \sum_{j=1}^{m_2} \int_0^t \hat{\sigma}^j(r, \xi_r^u, x_r^u, u)\, dB_r^j, \tag{2}$$

where $\xi_0 : \mathcal{U} \to \mathbb{R}^{n_1}$, $b : [0,T] \times \mathbb{R}^{n_1} \times \mathcal{U} \to \mathbb{R}^{n_1}$, $\sigma = (\sigma^1, \dots, \sigma^{m_1}) : [0,T] \times \mathbb{R}^{n_1} \times \mathcal{U} \to \mathbb{R}^{n_1 \times m_1}$ and $x_0 : \mathcal{U} \to \mathbb{R}^{n_2}$, $\hat{b} : [0,T] \times \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \mathcal{U} \to \mathbb{R}^{n_2}$, $\hat{\sigma} = (\hat{\sigma}^1, \dots, \hat{\sigma}^{m_2}) : [0,T] \times \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \mathcal{U} \to \mathbb{R}^{n_2 \times m_2}$. Conditions on the coefficient functions which guarantee wellposedness of (1)–(2) for each parameter choice $u \in \mathcal{U}$ will be imposed at the end of this subsection.

The stochastic integral with respect to w in (1) can be understood in the sense of Young integration [50, 13, 14], while the stochastic integral in (2) can be interpreted as a classical Itô integral (e.g., [30]). It is, however, more convenient to work with a unifying notion of stochastic integration, which generalizes the Young integral and the Itô integral, namely with the *forward integral* of Russo and Vallois [43, 44], which is defined as follows: Let the integrator $(X_t)_{t\in[0,T]}$ be a continuous process and the integrand $(Y_t)_{t\in[0,T]}$ be integrable in the time variable, i.e., $\int_0^T |Y_s|\, ds < \infty$, P-almost surely. For every $\varepsilon > 0$, the *ε-forward integral*

$$I^-(\varepsilon, Y, dX)(t) = \int_0^t Y_s \frac{X_{(s+\varepsilon)\wedge T} - X_s}{\varepsilon}\, ds$$

is then well-defined and may be considered as a regularized version of a forward Riemann sum (i.e., a Riemann sum with tag point at the left interval boundary point of the subintervals of the partition) of Y with respect to X.

The *forward integral* of Y with respect to X is said to exist and is denoted by $(\int_0^t Y_s \, d^- X_s)_{t\in[0,T]}$, if

$$\lim_{\varepsilon\to 0} \sup_{t\in[0,T]} \left| \int_0^t Y_s \, d^- X_s - I^-(\varepsilon, Y, dX)(t) \right| = 0 \quad \text{in probability,} \tag{3}$$

i.e., it is the uniform limit in probability of the ε-forward integral. More background information on the Russo-Vallois forward integral and on the Young integral will be provided in Section 3.

With this notation at hand, we may rewrite the system (1)–(2) in more compact form as

$$\begin{aligned} \mathcal{X}_t^u = &\begin{pmatrix} \xi_0(u) \\ x_0(u) \end{pmatrix} + \int_0^t \begin{pmatrix} b\left(r, \mathcal{X}_r^{u,1:n_1}, u\right) \\ \hat{b}\left(r, \mathcal{X}_r^u, u\right) \end{pmatrix} dr \\ &+ \sum_{j=1}^{m_1} \int_0^t \begin{pmatrix} \sigma^j\left(r, \mathcal{X}_r^{u,1:n_1}, u\right) \\ 0 \end{pmatrix} d^- w_r^j + \sum_{j=1}^{m_2} \begin{pmatrix} 0 \\ \hat{\sigma}^j\left(r, \mathcal{X}_r^u, u\right) \end{pmatrix} d^- B_r^j \end{aligned} \tag{4}$$

where $\mathcal{X}_t^u = (\mathcal{X}_t^{u,1}, \ldots, \mathcal{X}_t^{u,n_1+n_2})^\top$, $\mathcal{X}_t^{u,1:n_1} := (\mathcal{X}_t^{u,1}, \ldots, \mathcal{X}_t^{u,n_1})^\top = \xi_t^u$, and $\mathcal{X}_t^{u,n_1+1:n_2} := (\mathcal{X}_t^{u,n_1+1}, \ldots, \mathcal{X}_t^{u,n_2})^\top = x_t^u$.

We now introduce cost the functional

$$J(u) = \frac{1}{2} \sum_{\mu=1}^{M} \mathrm{E}\left[g_\mu(\mathcal{X}_{T_\mu}^u) \right]^2, \quad u \in \mathcal{U} \tag{5}$$

where $0 < T_1 \leq \cdots < T_M = T$ is a finite sequence of time points. Each of the functions $g_\mu : \mathbb{R}^{n_1+n_2} \to \mathbb{R}$ may represent the difference between the payoff function of an option with maturity T_μ and its observed market price, when calibrating a financial model.

We will aim at minimizing the cost functional J by a gradient descent, for which we derive two gradient representations in the next subsection. Before doing so, we collect the assumptions required for our results:

For the p-variation process w, we assume the following exponential moment bound:

(W) There exists $K > 0$ such that

$$E\left[e^{K\|w\|_{p,0,T}^2} \right] < \infty. \tag{6}$$

Here, $\|x\|_{p,s,t}$ denotes the p-variation norm of a path x over the interval $[s,t]$.

Remark 2.1. Let us recall some standard notation concerning p-variation functions for $p \geq 1$. We define $\mathcal{P}([s,t])$ as the set of all finite partitions of

the interval $[s,t]$. For a partition $\Pi_k = (t_i)_{i=0,\dots,k}$ of $[s,t]$ into k subintervals (i.e., $s = t_0 < t_1 < \cdots < t_k = t$) we call $|\Pi_k| = \max_{i=0,\dots,k-1}\{|t_{i+1} - t_i|\}$ the mesh of the partition and, for $i = 0, \dots, k-1$, we call $[t_i, t_{i+1}]$ a subinterval of the partition. If the number of subintervals of a partition does not need to be specified, we will omit the index k. For $1 \leq p < \infty$, the p-variation semi-norm of a function $x : [s,t] \to \mathbb{R}^{n\times m}$ is then given by

$$|x|_{p,s,t} := \sup_{k\in\mathbb{N},\Pi_k\in\mathcal{P}([s,t])} \left(\sum_{i=0}^{k-1} |x_{t_{i+1}} - x_{t_i}|^p\right)^{\frac{1}{p}},$$

where $|\cdot|$ denotes the Frobenius norm of a matrix. x is said to be *of finite p-variation* over the interval $[s,t]$, if $|x|_{p,s,t} < \infty$. We write $W^p([s,t],\mathbb{R}^{n\times m})$ for the space of finite p-variation functions over $[s,t]$, which endowed with the p-variation norm $\|x\|_{p,s,t} := |x_s| + |x|_{p,s,t}$, becomes a Banach space. The subspace of continuous functions in $W^p([s,t],\mathbb{R}^{n\times m})$ will be denoted by $C^p([s,t],\mathbb{R}^{n\times m})$.

Remark 2.2. If $(w_t)_{t\in[0,T]}$ is a Gaussian process with continuous paths, which are of bounded p-variation for some $p \in (1,2)$, then the exponential moment bound in condition (W) is satisfied by Theorem 2.3 in [28].

Concerning the coefficients of the SDE (1), we suppose:

(H_1) Let $\xi_0 : \mathcal{U} \to \mathbb{R}^{n_1}$ be continuously differentiable, such that ξ_0 and its Jacobian $D\xi_0$ are bounded.

(H_2) Let $b : [0,T]\times\mathbb{R}^{n_1}\times\mathcal{U} \to \mathbb{R}^{n_1}$ be a continuous function which satisfies:

- $b(t,\xi,u)$ is bounded and twice continuously differentiable with respect to ξ and u with bounded partial derivatives.

(H_3) Let $\sigma := (\sigma^1,\dots,\sigma^{m_1}) : [0,T]\times\mathbb{R}^{n_1}\times\mathcal{U} \to \mathbb{R}^{n_1\times m_1}$ be a continuous function which satisfies:

- $\sigma(t,\xi,u)$ is bounded and three times continuously differentiable with respect to ξ and u with bounded partial derivatives.
- $\sigma(t,\xi,u)$ and all its partial derivatives with respect to ξ and u up to order 2 are Hölder continuous in t with Hölder exponent $\beta \in [\frac{1}{2}, 1]$.

For the coefficients of SDE (2), we assume:

(B_1) Let $x_0 : \mathcal{U} \to \mathbb{R}^{n_2}$ be a continuously differentiable deterministic function, such that x_0 and its Jacobian Dx_0 are bounded.

(B_2) Let $\hat{b} : [0,T]\times\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}\times\mathcal{U} \to \mathbb{R}^{n_2}$ be a continuous function which satisfies:

- $b(t, \xi, x, u)$ is continuously differentiable with respect to x, ξ and u with bounded partial derivatives.

(B_3) Let $\hat{\sigma} = (\hat{\sigma}^1, \dots, \hat{\sigma}^{m_2}) : [0,T] \times \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \mathcal{U} \to \mathbb{R}^{n_2 \times m_2}$ be a continuous function which satisfies:

- $\sigma(t, \xi, x, u)$ is continuously differentiable with respect to x, ξ and u with bounded partial derivatives.

Finally, on the cost functional we impose the following condition:

(G) For every $\mu = 1, \dots, M$, the function $g_\mu : \mathbb{R}^{n_1+n_2} \to \mathbb{R}$ is bounded and continuously differentiable with bounded and Lipschitz continuous first derivative.

We will refer to (W), (H_1)–(H_3), (B_1)–(B_3), and (G) as the *standing assumptions*. They are supposed to be in force for the rest of this paper.

2.2 Gradient representations

In this subsection, we derive two representations for the gradient of the cost functional J introduced in (5). The first one is a simple consequence of the Fréchet differentiability of the state dynamics (4) with respect to the parameter vector u and the chain rule. The second one is an adjoint gradient representation in terms of an anticipating backward stochastic differential equation in the sense of forward integration and constitutes one of the main results of this paper.

Formally differentiating the state dynamics $\mathcal{X}^u$ in (4) with respect to u (cp. Subsection 4.1 below) suggests that its Fréchet derivative $D\mathcal{X}^u$ is an $\mathbb{R}^{(n_1+n_2)\times d}$-valued stochastic process, which solves the linear matrix SDE

$$\begin{aligned}
\mathcal{Y}_t^u = &\begin{pmatrix} D\xi_0(u) \\ Dx_0(u) \end{pmatrix} \\
&+ \int_0^t \begin{pmatrix} b_\xi\left(r, \mathcal{X}_r^{u,1:n_1}, u\right) & 0 \\ \hat{b}_\xi\left(r, \mathcal{X}_r^u, u\right) & \hat{b}_x\left(r, \mathcal{X}_r^u, u\right) \end{pmatrix} \mathcal{Y}_r^u + \begin{pmatrix} b_u\left(r, \mathcal{X}_r^{u,1:n_1}, u\right) \\ \hat{b}_u\left(r, \mathcal{X}_r^u, u\right) \end{pmatrix} dr \\
&+ \sum_{j=1}^{m_1} \int_0^t \begin{pmatrix} \sigma_\xi^j\left(r, \mathcal{X}_r^{u,1:n_1}, u\right) & 0 \\ 0 & 0 \end{pmatrix} \mathcal{Y}_r^u + \begin{pmatrix} \sigma_u^j\left(r, \mathcal{X}_r^{u,1:n_1}, u\right) \\ 0 \end{pmatrix} d^- w_r^j \\
&+ \sum_{j=1}^{m_2} \int_0^t \begin{pmatrix} 0 & 0 \\ \hat{\sigma}_\xi^j\left(r, \mathcal{X}_r^u, u\right) & \hat{\sigma}_x^j\left(r, \mathcal{X}_r^u, u\right) \end{pmatrix} \mathcal{Y}_r^u + \begin{pmatrix} 0 \\ \hat{\sigma}_u^j\left(r, \mathcal{X}_r^u, u\right) \end{pmatrix} d^- B_r^j
\end{aligned} \tag{7}$$

We will refer to (7) as the *sensitivity equation*, compare, e.g., [29].

The following theorem provides existence and uniqueness of the SDEs (4) and (7) under the standing assumptions. We will make use of the

space $L^l_{\mathbb{F}}(\Omega, C^{p,0}[0,T], \mathbb{R}^{(n_1+k)\times m})$ of $\mathbb{F}$-adapted processes $(x_t)_{t\in[0,T]}$ satisfying $\mathrm{E}[\|x\|^l_{\infty,0,T}] < \infty$ and taking values in $\mathbb{R}^{(n_1+k)\times m}$ such that P-almost every path is continuous and the component paths in the first n_1 lines of x are of finite p-variation. Here, $\|x\|_{\infty,s,t} := \sup_{r\in[s,t]} |x_r|$ denotes the supremum norm over the interval $[s,t]$.

Theorem 2.3. *For every $u \in \mathcal{U}$, the SDEs (4) and (7) have unique solutions $\mathcal{X}^u \in L^l_{\mathbb{F}}(\Omega, C^{p,0}[0,T], \mathbb{R}^{n_1+2})$ and $\mathcal{Y}^u \in L^l_{\mathbb{F}}(\Omega, C^{p,0}[0,T], \mathbb{R}^{(n_1+2)\times d})$, respectively, for every $l \geq 1$. Moreover, there is a constant $C_{\mathcal{X},\mathcal{Y},l}$ independent of $u \in \mathcal{U}$ such that*

$$\mathrm{E}\left[\|\mathcal{X}^u\|^l_{\infty,0,T}\right] + \mathrm{E}\left[\|\mathcal{Y}^u\|^l_{\infty,0,T}\right] \leq C_{\mathcal{X},\mathcal{Y},l}, \quad (l \geq 1).$$

Finally, for every $l \geq 1$, the map

$$\mathcal{U} \to L^l_{\mathbb{F}}(\Omega, C^{p,0}[0,T], \mathbb{R}^{n_1+2}), \quad u \mapsto \mathcal{X}^u$$

is Fréchet differentiable with Fréchet derivative $D\mathcal{X}^u = \mathcal{Y}^u$.

The details of the technical proof can be found in [48]. We will comment in Subsection 4.1 below on the key steps of the proof and on related results in the literature.

Applying the previous theorem in conjunction with assumption (G) and the chain rule for Fréchet derivatives (see Proposition 1.1.4 in [1]), we obtain the following representation for the gradient of the cost functional

$$\nabla J(u) = \sum_{\mu=1}^{M} \mathrm{E}[g_\mu(\mathcal{X}^u_{T_\mu})]\mathrm{E}[g'_\mu(\mathcal{X}^u_{T_\mu})\mathcal{Y}^u_{T_\mu}] \tag{8}$$

under the standing assumptions. In order to obtain the adjoint gradient representation, we first derive a variation-of-constants formula for the linear matrix SDE (7) in Theorem 2.5 below.

Notation 2.4. *To simplify the notation for the rest of this paper, we define for $r \in [0,T]$ and $u \in \mathcal{U}$: $a^u(r) := a(r, \mathcal{X}^{u,1:n_1}_r, \mathcal{X}^{u,n_1+1:n_2}_r, u)$, for some generic function a mapping from $[0,T] \times \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \mathcal{U}$ to some Euclidean space.*

With this notation at hand, we consider the following homogeneous matrix-valued SDEs with initial condition equal to the unit matrix $I_{n_1+n_2}$ in $\mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$ at time $s_0 \in [0,T)$:

$$\begin{aligned}\Phi^{s_0}_t = I_{n_1+n_2} &+ \int_{s_0}^t \begin{pmatrix} b^u_\xi(r) & 0 \\ \hat{b}^u_\xi(r) & \hat{b}^u_x(r) \end{pmatrix} \Phi^{s_0}_r \, dr + \sum_{j=1}^{m_1} \int_{s_0}^t \begin{pmatrix} \sigma^{u,j}_\xi(r) & 0 \\ 0 & 0 \end{pmatrix} \Phi^{s_0}_r \, d^- w^j_r \\ &+ \sum_{j=1}^{m_2} \int_{s_0}^t \begin{pmatrix} 0 & 0 \\ \hat{\sigma}^{u,j}_\xi(r) & \hat{\sigma}^{u,j}_x(r) \end{pmatrix} \Phi^{s_0}_r \, d^- B^j_r, \end{aligned} \tag{9}$$

and

$$\Psi_t^{s_0} = I_{n_1+n_2} - \int_{s_0}^t \Psi_r^{s_0} \left[\begin{pmatrix} b_\xi^u(r) & 0 \\ \hat{b}_\xi^u(r) & \hat{b}_x^u(r) \end{pmatrix} - \sum_{j=1}^{m_2} \begin{pmatrix} 0 & 0 \\ \hat{\sigma}_\xi^{u,j}(r) & \hat{\sigma}_x^{u,j}(r) \end{pmatrix}^2 \right] dr$$
$$- \sum_{j=1}^{m_1} \int_{s_0}^t \Psi_r^{s_0} \begin{pmatrix} \sigma_\xi^{u,j}(r) & 0 \\ 0 & 0 \end{pmatrix} d^- w_r^j - \sum_{j=1}^{m_2} \int_{s_0}^t \Psi_r^{s_0} \begin{pmatrix} 0 & 0 \\ \hat{\sigma}_\xi^{u,j}(r) & \hat{\sigma}_x^{u,j}(r) \end{pmatrix} d^- B_r^j, \tag{10}$$

for $t \in [s_0, T]$, suppressing the dependence on u by abbreviating $\Phi_t^{s_0} = \Phi_t^{s_0,u}$ and $\Psi_t^{s_0} = \Psi_t^{s_0,u}$.

Theorem 2.5. *For every $u \in \mathcal{U}$ and $s_0 \in [0,T]$ the matrix-valued SDEs (9) and (10) have a unique solution $\Phi^{s_0,u}$, respectively $\Psi^{s_0,u}$ in the space $L^l_{\mathbb{F}}(\Omega, C^{p,0}[s_0,T], \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)})$ for every $l \geq 1$, such that*

$$\mathrm{E}\left[\|\Phi^{s_0,u}\|^l_{\infty,s_0,T} \right] + \mathrm{E}\left[\|\Psi^{s_0,u}\|^l_{\infty,s_0,T} \right] \leq C_{\Phi,\Psi,l},$$

where the positive constant $C_{\Phi,\Psi,l}$ is independent of u and s_0. Moreover, $\Psi_t^{s_0,u} = (\Phi_t^{s_0,u})^{-1}$ for $t \in [s_0,T]$, P-almost surely. Furthermore the solution $\mathcal{Y}_t^u$ to (7) is, for every $t \in [0,T]$, given by the following variation-of-constants formula (here we set the initial time of the homogeneous equations to $s_0 = 0$ and skip the superscripts s_0 and u)

$$\mathcal{Y}_t^u = \Phi_t \begin{pmatrix} D\xi_0(u) \\ Dx_0(u) \end{pmatrix} + \Phi_t \int_0^t \Phi_r^{-1} \left[\begin{pmatrix} b_u^u(r) \\ \hat{b}_u^u(r) \end{pmatrix} - \sum_{j=1}^{m_2} \begin{pmatrix} 0 \\ \hat{\sigma}_x^{u,j}(r) \hat{\sigma}_u^{u,j}(r) \end{pmatrix} \right] dr$$
$$+ \sum_{j=1}^{m_1} \Phi_t \int_0^t \Phi_r^{-1} \begin{pmatrix} \sigma_u^{u,j}(r) \\ 0 \end{pmatrix} d^- w_r^j + \sum_{j=1}^{m_2} \Phi_t \int_0^t \Phi_r^{-1} \begin{pmatrix} 0 \\ \hat{\sigma}_u^{u,j}(r) \end{pmatrix} d^- B_r^j. \tag{11}$$

A sketch of the proof will be provided in Subsection 4.2 below. Inserting the variation-of-constants formula (11) into the gradient representation (8), we obtain, by manipulating the forward integrals as detailed in Lemma 4.3 of Subsection 4.3 below:

$$\nabla J(u) = \mathrm{E}\Bigg[\Lambda_0 \begin{pmatrix} D\xi_0(u) \\ Dx_0(u) \end{pmatrix} + \int_0^T \Lambda_r \left[\begin{pmatrix} b_u^u(r) \\ \hat{b}_u^u(r) \end{pmatrix} - \sum_{j=1}^{m_2} \begin{pmatrix} 0 \\ \hat{\sigma}_x^{u,j}(r) \hat{\sigma}_u^{u,j}(r) \end{pmatrix} \right] dr$$
$$+ \sum_{j=1}^{m_1} \int_0^T \Lambda_r \begin{pmatrix} \sigma_u^{u,j}(r) \\ 0 \end{pmatrix} d^- w_r^j + \sum_{j=1}^{m_2} \int_0^T \Lambda_r \begin{pmatrix} 0 \\ \hat{\sigma}_u^{u,j}(r) \end{pmatrix} d^- B_r^j \Bigg]. \tag{12}$$

Here, $(\Lambda_t)_{t\in[0,T]} = (\Lambda_t^u)_{t\in[0,T]}$ is the $\mathbb{R}^{1\times(n_1+n_2)}$-valued (i.e., row-vector valued) process defined via

$$\Lambda_t = \sum_{\mu;\, T_\mu \geq t} E[g_\mu(\mathcal{X}_{T_\mu}^u)] g'_\mu(\mathcal{X}_{T_\mu}^u)\Phi_{T_\mu}\Phi_t^{-1}, \tag{13}$$

(where again $\Phi = \Phi^{0,u}$).

We emphasize that the process Λ is not $\mathbb{F}$-adapted, but anticipates future information through the factors $g'_\mu(\mathcal{X}_{T_\mu}^u)\Phi_{T_\mu}$. In particular, the integrals with respect to the Brownian motions in (12) cannot be interpreted as Itô integrals, but are 'true' forward integrals. These anticipating forward integrals, in general, do not have zero expectation, and, hence, contribute to the *adjoint gradient representation* (12).

Theorem 2.6. *The process $\Lambda = (\Lambda^u)$ satisfies $\mathrm{E}[\|\Lambda\|_{\infty,0,T}^l] < \infty$ for every $l \geq 1$ and solves the anticipating backward SDE ($t \in [0,T]$)*

$$\begin{aligned}\Lambda_t = &\sum_{T_\mu \geq t} E[g_\mu(\mathcal{X}_{T_\mu}^u)] g'_\mu(\mathcal{X}_{T_\mu}^u) \\ &+ \int_t^T \Lambda_r \left[\begin{pmatrix} b_\xi^u(r) & 0 \\ \hat{b}_\xi^u(r) & \hat{b}_x^u(r) \end{pmatrix} - \sum_{j=1}^{m_2} \begin{pmatrix} 0 & 0 \\ \hat{\sigma}_\xi^{u,j}(r) & \hat{\sigma}_x^{u,j}(r) \end{pmatrix}^2 \right] dr \\ &+ \sum_{j=1}^{m_1} \int_t^T \Lambda_r \begin{pmatrix} \sigma_\xi^{u,j}(r) & 0 \\ 0 & 0 \end{pmatrix} d^- w_r^j + \sum_{j=1}^{m_2} \int_t^T \Lambda_r \begin{pmatrix} 0 & 0 \\ \hat{\sigma}_\xi^{u,j}(r) & \hat{\sigma}_x^{u,j}(r) \end{pmatrix} d^- B_r^j. \end{aligned} \tag{14}$$

The proof will be given in Subsection 4.3.

We call (14) the *adjoint equation*, as it resembles the adjoint equation in the Pontryagin maximum principle for optimal control problems:

1. *Deterministic case*: $\sigma \equiv 0$, $b \equiv 0$, $\xi_0 \equiv 0$, $\hat{\sigma} \equiv 0$, and $T_\mu = T$ for every $\mu = 1, \ldots, M$:
   Then (14) reduces to the terminal value problem for the ordinary differential equation (ignoring the first n_1 components of Λ)

   $$\dot{\Lambda}_t = -\Lambda_t \hat{b}_x^u(t), \quad \Lambda_T = \sum_{\mu=1}^M E[g_\mu(\mathcal{X}_{T_\mu}^u)] g'_\mu(\mathcal{X}_{T_\mu}^u)$$

   which corresponds to the adjoint equation in the Pontryagin maximum principle for the control of ordinary differential equations, compare, e.g., Chapter 3.2 in [49]. Note that the local optimality condition in terms of the Hamiltonian in the optimal control situation, e.g. Eq. (3.2.7) in [49], turns into the global condition

   $$0 = \nabla J(u) = Dx_0(u) + \int_0^T \Lambda_r \hat{b}_u^u(r) dr$$

in our case, see (12), because we are minimizing over constant parameters (in contrast to optimizing dynamically over functions).

2. *Brownian motion case*: $\sigma \equiv 0$, $b \equiv 0$, $\xi_0 \equiv 0$, and $T_\mu = T$ for every $\mu = 1, \dots, M$:
In order to avoid to work with the anticipating process Λ, one can project Λ (ignoring the first n_1 components of Λ, again) on the available information leading to

$$p_t := E[\Lambda_t | \mathcal{F}_t].$$

By Theorem 7.2.2 in [49] and its proof, there is a matrix-valued process $q = (q^1, \dots, q^{m_2})$ such that the pair (p, q) solves the following nonanticipating backward stochastic differential equation (BSDE) in terms of Itô integration

$$\begin{aligned} dp_t &= -\left(p_t \hat{b}_x^u(t) + \sum_{j=1}^{m_2} q_t^j \hat{\sigma}_x^{u,j}(t) \right) + \sum_{j=1}^{m_2} q_t^j dB_t, \\ p_T &= \sum_{\mu=1}^{M} E[g_\mu(\mathcal{X}_{T_\mu}^u)] g_\mu'(\mathcal{X}_{T_\mu}^u) \end{aligned}$$

The process q can be constructed via the martingale representation theorem. This BSDE corresponds to the first adjoint equation in the maximum principle for controlled stochastic differential equations, see Eq. (3.3.8) in [49].

Note that the techniques related to moving from Λ to p, by conditioning in the Brownian motion case, heavily rely on the martingale property of a Brownian motion. As these martingale techniques are not at our disposal in the presence of the second driving process $(w_t)_{t\in[0,T]}$, we decided to work directly with Λ and to derive the adjoint equation in terms of a new type of an anticipating backward SDE in Theorem 2.6.

2.3 Euler discretization of the gradient representations

Any numerical resolution of the gradient representations (8) or (12) requires a time-discretization scheme either for the pair $(\mathcal{X}^u, \mathcal{Y}^u)$ or for the pair $(\mathcal{X}^u, \Lambda^u)$. From a computational point of view, it is beneficial to approximate the pair $(\mathcal{X}^u, \Lambda^u)$, because both processes are $\mathbb{R}^{n_1+n_2}$-dimensional, while $\mathcal{Y}^u$ is a matrix-valued, precisely $\mathbb{R}^{(n_1+n_2)\times d}$-dimensional process, where d equals the number of parameters. However, Λ^u solves, by Theorem 2.6, a new type of anticipating backward SDE in terms of the forward integral. Devising an Euler scheme for Λ^u and analyzing its rate of convergence constitute the main results of this subsection.

For the discretization of the state dynamics (4) and the sensitivity equation (7), we apply a continuously interpolated Euler scheme along a partition $\Pi^{\mathrm{E}} = (t_i)_{i=0,\dots,n}$ of $[0,T]$, which is not necessarily equidistant. We will always assume that the time points T_μ, connected to the cost functional (5), are elements of Π^{E}.

Precisely, we consider, for $t \in (t_i, t_{i+1}]$,

$$\begin{aligned}\mathcal{X}_t^{n,u} = \begin{pmatrix} \xi_t^{n,u} \\ x_t^{n,u} \end{pmatrix} &= \begin{pmatrix} \xi_{t_i}^{n,u} \\ x_{t_i}^{n,u} \end{pmatrix} + \begin{pmatrix} b\left(t_i, \xi_{t_i}^{n,u}, u\right) \\ \hat{b}\left(r, \xi_{t_i}^{n,u}, x_{t_i}^{n,u}, u\right) \end{pmatrix} (t - t_i) \\ &+ \sum_{j=1}^{m_1} \begin{pmatrix} \sigma^j\left(t_i, \xi_{t_i}^{n,u}, u\right) \\ 0 \end{pmatrix} \left(w_t^j - w_{t_i}^j\right) \\ &+ \sum_{j=1}^{m_2} \begin{pmatrix} 0 \\ \hat{\sigma}^j\left(t_i, \xi_{t_i}^{n,u}, x_{t_i}^{n,u}, u\right) \end{pmatrix} \left(B_t^j - B_{t_i}^j\right) \end{aligned} \tag{15}$$

initialized at $\mathcal{X}_0^{n,u}$ via $\xi_0^{n,u} = \xi_0(u)$ and $x_0^{n,u} = x_0(u)$, as well as

$$\begin{aligned}\mathcal{Y}_t^{n,u} =& \mathcal{Y}_{t_i}^{n,u} + \eta_{t_i,t}^{n,u}\, \mathcal{Y}_{t_i}^{n,u} + \begin{pmatrix} b_u\left(t_i, \xi_{t_i}^{n,u}, u\right) \\ \hat{b}_u\left(t_i, \xi_{t_i}^{n,u}, x_{t_i}^{n,u}, u\right) \end{pmatrix} (t - t_i) \\ &+ \sum_{j=1}^{m_1} \begin{pmatrix} \sigma_u^j(t_i, \xi_{t_i}^{n,u}, u) \\ 0 \end{pmatrix} \left(w_t^j - w_{t_i}^j\right) \\ &+ \sum_{j=1}^{m_2} \begin{pmatrix} 0 \\ \hat{\sigma}_u^j\left(t_i, \xi_{t_i}^{n,u}, x_{t_i}^{n,u}, u\right) \end{pmatrix} \left(B_t^j - B_{t_i}^j\right) \end{aligned} \tag{16}$$

$$\mathcal{Y}_0^{n,u} = \begin{pmatrix} D\xi_0(u) \\ Dx_0(u) \end{pmatrix}, \tag{17}$$

where

$$\begin{aligned}\eta_{t_i,t}^{n,u} = & \begin{pmatrix} b_\xi(t_i, \xi_{t_i}^{n,u}, u) & 0 \\ \hat{b}_\xi(t_i, \xi_{t_i}^{n,u}, x_{t_i}^{n,u}, u) & \hat{b}_x(t_i, \xi_{t_i}^{n,u}, x_{t_i}^{n,u}, u) \end{pmatrix} (t - t_i) \\ &+ \sum_{j=1}^{m_1} \begin{pmatrix} \sigma_\xi^j(t_i, \xi_{t_i}^{n,u}, u) & 0 \\ 0 & 0 \end{pmatrix} (w_t^j - w_{t_i}^j) \\ &+ \sum_{j=1}^{m_2} \begin{pmatrix} 0 & 0 \\ \hat{\sigma}_\xi^j(t_i, \xi_{t_i}^{n,u}, x_{t_i}^{n,u}, , u) & \hat{\sigma}_x^j(t_i, \xi_{t_i}^{n,u}, x_{t_i}^{n,u}, u) \end{pmatrix} (B_t^j - B_{t_i}^j). \end{aligned} \tag{18}$$

The discretization of the adjoint equation (12) is initialized at terminal time $t_n = T$ via

$$\Lambda_{t_n}^{n,u} = \sum_{\mu;\, T_\mu = T} \mathrm{E}[g_\mu(\mathcal{X}_T^{n,u})] g_\mu'(\mathcal{X}_T^{n,u}) \tag{19}$$

and then follows a backward recursion along the grid points

$$\Lambda_{t_i}^{n,u} = \Lambda_{t_{i+1}}^{n,u} + \Lambda_{t_{i+1}}^{n,u}\, \eta_{t_i,t_{i+1}}^{n,u} + \sum_{\mu;\, T_\mu = t_i} E[g_\mu(\mathcal{X}_{t_i}^{n,u})] g_\mu'(\mathcal{X}_{t_i}^{n,u}). \tag{20}$$

We apply a piecewise constant interpolation on the interval $[0,T]$, i.e.,

$$\Lambda_t^{n,u} = \Lambda_{t_{i+1}}^{n,u}$$

for $t \in (t_i, t_{i+1})$.

Remark 2.7. In order to motivate the discretization for the adjoint equation, let us look at a 1-dimensional equation of the form

$$\lambda_t = \lambda_T + \int_t^T \lambda_r (b_r - \hat{\sigma}_r^2)\, dt + \int_t^T \lambda_r \sigma_r \, d^- w_r + \int_t^T \lambda_r \hat{\sigma}_r \, d^- B_r, \qquad (21)$$

where b, σ and $\hat{\sigma}$ are $\mathbb{F}$-adapted processes. As forward integrals are based on Riemann sums with the tag point at the left interval boundary point, we get

$$\lambda_{t_i} \approx \lambda_{t_{i+1}} + \lambda_{t_i} \left[(b_{t_i} - \hat{\sigma}_{t_i}^2)\Delta_i + \sigma_{t_i} \Delta w_i + \hat{\sigma}_{t_i} \Delta B_i \right] =: \lambda_{t_{i+1}} + \lambda_{t_i} \widetilde{\eta}_{t_i},$$

for $\Delta_i := t_{i+1} - t_i$, $\Delta w_i =: w_{t_{i+1}} - w_{t_i}$ and $\Delta B_i := B_{t_{i+1}} - B_{t_i}$. Rearranging terms and performing a second order Taylor expansion (dropping higher order terms), we arrive at

$$\begin{aligned}
\lambda_{t_i} \approx & \lambda_{t_{i+1}} (1 - \widetilde{\eta}_{t_i})^{-1} \approx \lambda_{t_{i+1}} (1 + \widetilde{\eta}_{t_i} + \widetilde{\eta}_{t_i}^2) \\
\approx & \lambda_{t_{i+1}} (1 + (b_{t_i} - \hat{\sigma}_{t_i}^2)\Delta_i + \sigma_{t_i} \Delta w_i + \hat{\sigma}_{t_i} \Delta B_i + \hat{\sigma}_{t_i}^2 \Delta B_i^2) \\
\approx & \lambda_{t_{i+1}} (1 + b_{t_i} \Delta_i + \sigma_{t_i} \Delta w_i + \hat{\sigma}_{t_i} \Delta B_i).
\end{aligned}$$

In order to measure the error of the Euler schemes under the assumed p-variation regularity, we consider

$$\delta(\omega; \Pi^E) := \max_{i=0,\dots,n-1} |t_{i+1} - t_i| + |w(\omega)|_{p,t_i,t_{i+1}},$$

which depends on the p-variation seminorm of the realized path of w over the subintervals of the Euler-partition (cp. Remark 2.1), as well as the 'averaged error measure' in the lth mean

$$\delta_l(\Pi^E) := \mathrm{E}\left[\delta(\Pi^E)^l \right]^{\frac{1}{l}}, \quad l \geq 1 \qquad (22)$$

which is finite, because w satisfies the exponential moment condition (6).

For the next theorem, we require two extra conditions:

(E_1): The Hölder exponent β from condition (H_3) is an element of the interval $[\frac{1}{p}, 1]$. Moreover, the function b from condition (H_2) and its partial derivatives b_ξ and b_u are Hölder continuous in t with Hölder exponent β.

(E_2): Let $\hat{b}$ and $\hat{\sigma}$ be the coefficient functions from condition (B_1), respectively (B_2). There exists a constant $L > 0$, such that for all $x \in \mathbb{R}^{n_2}$, $\xi \in \mathbb{R}^{n_1}$, $u \in \mathcal{U}$ and $s \leq t \in [0,T]$,

$$\begin{aligned}&|\hat{b}(t,\xi,x,u) - \hat{b}(s,\xi,x,u)| + |\hat{\sigma}(t,\xi,x,u) - \hat{\sigma}(s,\xi,x,u)| \\ &\leq L(1+|x|+|\xi|)(t-s)^{\frac{1}{2}}.\end{aligned}$$

Theorem 2.8. *Suppose that, next to the standing assumptions, (E_1)–(E_2) are in force. Then, there is a constant $C_{E,l}$ depending on $l \geq 2$ (but independent of u and n) such that for every $u \in \mathcal{U}$ and $l \geq 2$,*

$$\begin{aligned}&\mathrm{E}\left[\|\mathcal{X}^u - \mathcal{X}^{n,u}\|^l_{\infty,0,T}\right]^{\frac{1}{l}} + \mathrm{E}\left[\|\mathcal{Y}^u - \mathcal{Y}^{n,u}\|^l_{\infty,0,T}\right]^{\frac{1}{l}} + \sup_{t\in[0,T]} \mathrm{E}\left[|\Lambda^u_t - \Lambda^{n,u}_t|^l\right]^{\frac{1}{l}} \\ &\leq C_{E,l}(\delta_{4l}(\Pi^E))^{(2-p)\wedge\frac{1}{2}}.\end{aligned}$$

The key steps of the proof will be discussed in Subsection 4.4 below.

Remark 2.9. If we assume (instead of the p-variation regularity), that w has Hölder continuous paths with Hölder index $H \in (1/2, 1)$ and replace the p-variation norm by the Hölder norm in the exponential moment bound (6), then the error of the Euler approximations can be bounded in terms of the mesh size $|\Pi^E|$ of the partition. Precisely, the upper bound in Theorem 2.8 can then be replaced by $C'_{E,l}|\Pi^E|^{(2H-1)\wedge\frac{1}{2}}$ for some (possibly different) constant $C'_{E,l}$. Note that the p-variation norm $|w|_{p,t_i,t_{i+1}}$ is $\mathcal{O}(|t_{i+1} - t_i|^H)$, if w is H-Hölder continuous and $p = 1/H$, so that the rates in the p-variation case and in the Hölder case fit to each other.

Remark 2.10. The convergence analysis of Euler schemes is a classical topic in the literature on numerical SDEs. We mention some results, which are closely related to Theorem 2.8. Lejay [32] considers differential equations driven by a (deterministic) p-variation function w and derives a convergence rate of the order $(\max_{i=0,\dots,n-1} |w|_{p,t_i,t_{i+1}})^{2-p}$ under Lipschitz conditions. In order to achieve estimates in the lth mean, as for the convergence to $\mathcal{X}^{u,1:n_1}$ in Theorem 2.8, we additionally need to control the dependence of the constants on the realization of w, for which we apply the greedy sequence technique [9]. As in [26] for the fractional Brownian motion case, the boundedness of the coefficients significantly helps to carry out the analysis. In the case of Hölder paths, Euler schemes for SDEs driven by a fractional Brownian motion with Hurst parameter $H > 1/2$ are well-studied. Under Lipschitz conditions a rate of convergence of the order $2H - 1$ in the sense of almost sure convergence is known to hold and to be sharp [35, 37, 40, 38]. The equation solved by $\mathcal{X}^{u,n_1+1:n_2}$ is driven by a Brownian motion. Although the coefficients depend on $\mathcal{X}^{u,1:n_1}$, the error estimate follows classical lines. Recall that strong convergence of the Euler scheme for SDEs driven by a Brownian of the order $1/2$ in the mesh size is well-known under Lipschitz

conditions, see, e.g. [31]. Hence, the results, mentioned above, indicate that the convergence rate of the Euler scheme for $\mathcal{X}^u$ derived in Theorem 2.8 and Remark 2.9 are the best ones that one could expect. The key difficulty in analyzing the Euler scheme for the linear matrix SDE $\mathcal{Y}^u$ is to control the growth of the Euler scheme driven by the p-variation process $(w_t)_{t\in[0,T]}$ in the presence of time-dependent coefficients. We are not aware of a related convergence result in the lth mean in the p-variation context, but refer to [6] for a study of linear equations driven by a fractional Brownian motion in the Hölder space setting. For the analysis of the Euler scheme for the adjoint equation (14), the proof will be based on explicit variations-of-constants formulas for the continuous-time solution Λ^u and its Euler discretization $\Lambda^{n,u}$.

Having the convergence results in Theorem 2.8 at hand, we can apply them for the approximation of the cost functional (5) and the two representations (8) and (12) of its gradient.

Let $\Pi^{\mathrm{E}} = (t_i)_{i=0,\dots,n}$ be a partition of the interval $[0,T]$ such that all the time points T_μ, $\mu = 1,\dots,M$ are included in the partition. Then, the discretized cost function and the discretization of the gradient representation (8) are defined via

$$\begin{aligned} J^n(u) &:= \frac{1}{2}\sum_{\mu=1}^{M} \mathrm{E}\left[g_\mu(\mathcal{X}^{n,u}_{T_\mu})\right]^2 \\ (\nabla J)^n(u) &:= \sum_{\mu=1}^{M} \mathrm{E}\left[g_\mu(\mathcal{X}^{n,u}_{T_\mu})\right] \mathrm{E}\left[g'_\mu(\mathcal{X}^{n,u}_{T_\mu})\mathcal{Y}^{n,u}_{T_\mu}\right] \end{aligned} \tag{23}$$

Given assumption (G) on the functions g_μ, it is easy to check that, for every $u \in \mathcal{U}$,

$$\begin{aligned} &|J(u) - J^n(u)| + |(\nabla J)(u) - (\nabla J)^n(u)| \\ \leq& C_J \left(\mathrm{E}\left[\|\mathcal{X}^u - \mathcal{X}^{n,u}\|^2_{\infty,0,T}\right]^{\frac{1}{2}} + \mathrm{E}\left[\|\mathcal{Y}^u - \mathcal{Y}^{n,u}\|^2_{\infty,0,T}\right]^{\frac{1}{2}} \right) \end{aligned} \tag{24}$$

for some constant C_J (independent of u) and, thus, Theorem 2.8 and Remark 2.9 provide rates of convergence under the different regularity assumptions on $(w_t)_{t\in[0,T]}$.

The next theorem establishes an alternative representation of the discretized gradient $(\nabla J)^n(u)$ in terms of the Euler scheme of the adjoint equation. It can be considered as the natural discretization of the adjoint gradient representation (12).

Theorem 2.11. *For every $u \in \mathcal{U}$, the discretized gradient $(\nabla J)^n(u)$ (see (23)) can be represented by*

$$(\nabla J)^n(u) = \mathrm{E}\left[\Lambda^{n,u}_0 D\mathcal{X}^u_0 + \sum_{i=0}^{n-1} \Lambda^{n,u}_{t_{i+1}} \hat{\eta}^{n,u}_{t_i,t_{i+1}}\right],$$

where (suppressing the dependence of ξ^n and x^n on u)

$$\hat{\eta}^{n,u}_{t_i,t_{i+1}} := \begin{pmatrix} b_u(t_i,\xi^n_{t_i},u) \\ \hat{b}_u(t_i,\xi^n_{t_i},x^n_{t_i},,u) \end{pmatrix}(t_{i+1}-t_i) + \sum_{j=1}^{m_1}\begin{pmatrix} \sigma^j_u(t_i,\xi^n_{t_i},u) \\ 0 \end{pmatrix}(w^j_{t_{i+1}}-w^j_{t_i}) + \sum_{j=1}^{m_2}\begin{pmatrix} 0 \\ \hat{\sigma}^j_u(t_i,\xi^n_{t_i},x^n_{t_i},u) \end{pmatrix}(B^j_{t_{i+1}}-B^j_{t_i})$$

for all $i = 0,\dots,n-1$.

Proof. Recall that, by (16),

$$\mathcal{Y}^{n,u}_{t_{i+1}} = \mathcal{Y}^{n,u}_{t_i} + \eta^{n,u}_{t_i,t_{i+1}}\mathcal{Y}^n_{t_i} + \hat{\eta}^{n,u}_{t_i,t_{i+1}}.$$

and, by (20),

$$\Lambda^{n,u}_{t_i} = \Lambda^{n,u}_{t_{i+1}} + \Lambda^{n,u}_{t_{i+1}}\,\eta^{n,u}_{t_i,t_{i+1}} + \sum_{\mu;\,T_\mu=t_i} E[g_\mu(\mathcal{X}^{n,u}_{T_\mu})]g'_\mu(\mathcal{X}^{n,u}_{T_\mu}).$$

Hence,

$$\begin{aligned}\Lambda^{n,u}_{t_{i+1}}\mathcal{Y}^{n,u}_{t_{i+1}} =& \Lambda^{n,u}_{t_{i+1}}\mathcal{Y}^{n,u}_{t_i} + \Lambda^{n,u}_{t_{i+1}}\eta^{n,u}_{t_i,t_{i+1}}\mathcal{Y}^n_{t_i} + \Lambda^{n,u}_{t_{i+1}}\hat{\eta}^{n,u}_{t_i,t_{i+1}} \\ =& \Lambda^{n,u}_{t_i}\mathcal{Y}^{n,u}_{t_i} + \Lambda^{n,u}_{t_{i+1}}\hat{\eta}^{n,u}_{t_i,t_{i+1}} - \sum_{\mu;\,T_\mu=t_i} E[g_\mu(\mathcal{X}^{n,u}_{T_\mu})]g'_\mu(\mathcal{X}^{n,u}_{T_\mu})\mathcal{Y}^{n,u}_{t_i}.\end{aligned}$$

Therefore,

$$\Lambda^{n,u}_{t_n}\mathcal{Y}^{n,u}_{t_n} - \Lambda^{n,u}_{t_0}\mathcal{Y}^{n,u}_{t_0} = \sum_{i=0}^{n-1}\Lambda^{n,u}_{t_{i+1}}\hat{\eta}^{n,u}_{t_i,t_{i+1}} - \sum_{\mu;\,T_\mu<t_n} E[g_\mu(\mathcal{X}^{n,u}_{T_\mu})]g'_\mu(\mathcal{X}^{n,u}_{T_\mu})\mathcal{Y}^{n,u}_{T_\mu}.$$

Inserting the terminal condition (19) for $\Lambda^{n,u}$ and the initial condition (17) for $\mathcal{Y}^{n,u}$, we obtain

$$\sum_{\mu=1}^{M} E[g_\mu(\mathcal{X}^{n,u}_{T_\mu})]g'_\mu(\mathcal{X}^{n,u}_{T_\mu})\mathcal{Y}^{n,u}_{T_\mu} = \Lambda^{n,u}_0 D\mathcal{X}^u_0 + \sum_{i=0}^{n-1}\Lambda^{n,u}_{t_{i+1}}\hat{\eta}^{n,u}_{t_i,t_{i+1}}$$

Recalling the definition (23) of $(\nabla J)^n(u)$, the proof is completed by taking expectation. □

3 On the Young integral and the Russo-Vallois forward integral

3.1 Background on Young integration

The Young integral [50] can be considered as a Riemann-Stieltjes integral in the context of p-variation functions. Suppose $[s,t]$ is a compact interval,

$x : [s,t] \to \mathbb{R}^{n\times m}$ and $w : [s,t] \to \mathbb{Y}$, where either $\mathbb{Y} = \mathbb{R}$ or $\mathbb{Y} = \mathbb{R}^{m\times d}$. Given a partition $\Pi_k = (t_i)_{i=0,\dots,k}$ of $[s,t]$ and a finite sequence of tag points $\Theta_k = (\theta_i)_{i=0,\dots,k-1}$, where $t_i \le \theta_i \le t_{i+1}$, the pair (Π_k, Θ_k) is said to be a *tagged partition*. The *Riemann-Stieltjes sum* of x with respect to w on the tagged partition (Π_k, Θ_k) is defined to be

$$RS(x, dw, (\Pi_k, \Theta_k)) = \sum_{i=0}^{k-1} x_{\theta_i}(w_{t_{i+1}} - w_{t_i}).$$

The *Riemann-Stieltjes integral* is said to exist and is, then, denoted by $\int_s^t x_r dw_r$, if for every $\epsilon > 0$, there is a $\delta > 0$ such that

$$\left| \int_s^t x_r dw_r - RS(x, dw, (\Pi, \Theta)) \right| < \epsilon$$

for every tagged partition (Π, Θ) with mesh-size $|\Pi| < \delta$. In the context of p-variation functions, the Riemann-Stieltjes integral exists, e.g., under the following conditions.

Theorem 3.1 (Young-Integral)**.** *For $1 \le p,\ 1 \le q$ such that $\alpha = \frac{1}{p} + \frac{1}{q} > 1$, let $x \in W^q([s,t], \mathbb{R}^{n\times m})$ and $w \in C^p([s,t], \mathbb{Y})$. Then, the Riemann-Stieltjes Integral $\int_s^t x_r\, dw_r$ exists and the inequality*

$$\left| \int_s^t x_r\, dw_r - x_\theta(w_t - w_s) \right| \le C_{p,q} |x|_{q,s,t} |w|_{p,s,t} \tag{25}$$

holds for every $\theta \in [s,t]$, where $C_{p,q} = \zeta(\alpha)$ for $\zeta(y) = \sum_{i=1}^{\infty} \left(\frac{1}{i}\right)^y$ $(y > 1)$. Moreover, we have

$$\left| \int_s^t x_r\, dw_r \right| \le C_{p,q} \|x\|_{q,s,t} |w|_{p,s,t}. \tag{26}$$

In this situation we will speak of the Young integral *and call inequality* (26) *the* Love-Young *estimate.*

Proof. As the integrator w is continuous and taking Theorem 2.42 in [14] into account, this result is a special case of Corollary 3.91 in [14]. □

As limit of Riemann-Stieljes sums, the Young integral inherits, e.g., bilinearity as operator in integrand and integrator.

Control functions are a well-known to be a uesful tool for estimating p-variation (semi-)norms, see, e.g., [17]:

Definition 3.2. A continuous map φ taking values in the nonnegative real numbers, defined on the simplex $\Delta([s,t]) = \{(u,v) \in \mathbb{R}^2 \,|\, 0 \le u \le v \le t\}$ is called a *control function* on $[s,t]$, if it satisfies the following conditions:

1. For all $r \in [s,t]$: $\varphi(r,r) = 0$.

2. For all $u \le r \le v$ in $[s,t]$: $\varphi(u,r) + \varphi(r,v) \le \varphi(u,v)$.

The following lemma is a variant of Proposition 5.10 in [17].

Lemma 3.3. *Let $\varphi_1, \dots, \varphi_m$ be superadditive functions on $[s,t]$ (i.e., they satisfy property (2) in Definition 3.2), $p \ge 1$, $C_1, \dots, C_k$ positive constants and $x : [s,t] \to \mathbb{R}^{n\times m}$ a function on $[s,t]$. The pointwise estimate*

$$|x_v - x_u| \le \sum_{j=1}^{m} C_j \varphi_j(u,v)^{\frac{1}{p}} \text{ for all } u \le v \text{ in } [s,t]$$

implies the p-variation estimate

$$|x|_{p,u,v} \le \sum_{j=1}^{m} C_j \varphi_j(u,v)^{\frac{1}{p}} \text{ for all } u \le v \text{ in } [s,t].$$

If φ_j is a control function on $[s,t]$ for all $j = 1, \dots, m$, then x is continuous on $[s,t]$.

Note that p-variation estimates lead to estimates in the sup-norm via the relation

$$\|x\|_{\infty,s,t} \le |x|_s + |x|_{p,s,t} = \|x\|_{p,s,t}. \tag{27}$$

The following proposition states that the pth power of the p-variation semimnorm constitutes a control. For a proof, we refer to [17], Proposition 5.8.

Proposition 3.4. *Let $p \ge 1$ and $x : [s,t] \to \mathbb{R}^{n\times m}$ be a continuous function of finite p-variation, then*

$$\varphi(u,v) = |x|^p_{p,u,v}$$

defines a control function on $[s,t]$.

We state two elementary lemmas (without proof), which are useful for estimating p-variation (semi-)norms.

Lemma 3.5. *Let $p \ge 1$, $B \in W^p([s,t], \mathbb{R}^{n\times n})$, $x \in W^p([s,t], \mathbb{R}^{n\times m})$ and assume that $f : \mathbb{R}^{n\times m} \to \mathbb{R}^k$ is Lipschitz continuous with constant L. Then we have*

$$\|Bx\|_{p,s,t} \le |B_s x_s| + \|B\|_{\infty,s,t}|x|_{p,s,t} + \|x\|_{\infty,s,t}|B|_{p,s,t} \le 2\|B\|_{p,s,t}\|x\|_{p,s,t}$$

and

$$|f(x)|_{p,s,t} \le L|x|_{p,s,t}.$$

Lemma 3.6. *Let $x \in W^p([s,t],\mathbb{R}^{n\times m})$, $p \geq 1$. If $s = t_0 < t_1 < \cdots < t_k = t$, then*

$$\sum_{i=0}^{k-1} |x|^p_{p,t_i,t_{i+1}} \leq |x|^p_{p,s,t} \leq k^{p-1} \sum_{i=0}^{k-1} |x|^p_{p,t_i,t_{i+1}}.$$

The following lemma, see, e.g., Theorem 3.92 in [14], is devoted to the indefinite integral

$$I_Y(x,w)(u) = \int_s^u x_r \, dw_r \ \forall u \in [s,t]$$

Lemma 3.7. *Let $1 \leq p,\ 1 \leq q$ such that $\alpha = \frac{1}{p} + \frac{1}{q} > 1$, $x \in W^q([s,t],\mathbb{R}^{n\times m})$ and $w \in C^p([s,t],\mathbb{Y})$. The indefinite integral $I_Y(x,w)$ exists and is an element of $C^p([s,t],\mathbb{X})$, where $\mathbb{X} = \mathbb{R}^{n\times d}$ or $\mathbb{X} = \mathbb{R}^{n\times m}$ depending on the choice of $\mathbb{Y}$. Furthermore, we have*

$$\|I_Y(x,w)\|_{p,s,t} = |I_Y(x,w)|_{p,s,t} = \left| \int_s^{\cdot} x_r \, dw_r \right|_{p,s,t} \leq C_{p,q} \|x\|_{q,s,t} |w|_{p,s,t}.$$

We provide a proof in order to illustrate the control function technique.

Proof. For every $r \in [s,t]$ the indefinite Integral $I_Y(x,w)(r)$ exists by Theorem 3.1. Let $u < v \in [s,t]$, then we have by additivity of the Young integral and the Young-Love estimate

$$\begin{aligned} |I_Y(x,w)(v) - I_Y(x,w)(u)| = \left| \int_u^v x_r \, dw_r \right| &\leq C_{p,q} \|x\|_{q,u,v} |w|_{p,u,v} \\ &\leq C_{p,q} \|x\|_{q,s,t} |w|_{p,u,v}. \end{aligned}$$

Since $\varphi(u,v) = |w|^p_{p,u,v}$ is a control function on $[s,t]$, we conclude the proof by applying Lemma 3.3 and by noting that $I_Y(x,w)(s) = 0$. □

A crucial tool for the study of Young differential equations is the following variant of Gronwall's lemma:

Lemma 3.8. *Let $1 \leq p \leq q$ satisfy $\frac{1}{p} + \frac{1}{q} > 1$ and fix $T > 0$. Assume that $y \in W^q([0,T],\mathbb{R}^{n\times m})$ and $w \in C^p([s,t],\mathbb{Y})$ satisfy the following condition: There exist constants $K_1, K_2 > 0$ such that for all $[s,t] \subset [0,T]$, which satisfy $|t-s| + |w|_{p,s,t} \leq K_2$, we have*

$$|y|_{q,s,t} \leq K_1 + |y_s|. \tag{28}$$

Then,

$$|y|_{q,0,T} \leq (K_1 + |y_0|) e^{2^p K_2^{-p} (T^p + |w|^p_{p,0,T})} \tag{29}$$

and

$$\|y\|_{\infty,0,T} \leq \|y\|_{q,0,T} \leq (K_1 + 2|y_0|) e^{2^p K_2^{-p} (T^p + |w|^p_{p,0,T})}.$$

If the right hand side of (28) *only consists of the constant* K_1*, then the estimates simplify to*

$$|y|_{q,0,T} \leq K_1 2^{p-1} K_2^{-p} (T^p + |w|_{p,0,T}^p) \tag{30}$$

and

$$\|y\|_{\infty,0,T} \leq \|y\|_{q,0,T} \leq |y_0| + K_1 2^{p-1} K_2^{-p} (T^p + |w|_{p,0,T}^p). \tag{31}$$

This lemma is a matrix-valued variant of results in [9] (see their Lemma 3.3, Remark 3.4, and Corollary 3.5). We include the proof in order to illustrate the greedy sequence technique of [4] and [9]. A *greedy sequence* is an increasing sequence of time points $(\tau_i)_{i=0,\dots,N}$ of the interval $[0,T]$ with $\tau_N = T$ satisfying

$$\begin{aligned} |\tau_{i+1} - \tau_i| + |w|_{p,\tau_i,\tau_{i+1}} &= \mu \text{ for } i = 0, \dots, N-2 \\ |\tau_N - \tau_{N-1}| + |w|_{p,\tau_{N-1},\tau_N} &\leq \mu \end{aligned} \tag{32}$$

for given $\mu > 0$, $p \geq 1$. For the construction of such a sequence, one can first define $\tau_0 = 0$. Notice that $\kappa(t) = t + |w|_{p,0,t}$ is continuous and strictly increasing with respect to t, with $\kappa(0) = 0$ and $\kappa(T) = T + |w|_{p,0,T}$. The intermediate value theorem ensures, that there exists a unique $t > 0$ such that $t + |w|_{p,0,t} = \mu$, if $\mu < T + |w|_{p,0,T}$. In this case we let $\tau_1 = \sup\{0 \leq t \leq T |\ t + |w|_{p,0,t} \leq \mu\}$. Otherwise let $\tau_1 = T$. This construction can be continued inductively. For $T > 0$ and $0 \leq s < t \leq T$ denote

$$\overline{N}(t) = \sup_{k \in \mathbb{N}_0} \{\tau_k \leq t\},\ \underline{N}(t) = \inf_{k \in \mathbb{N}_0} \{\tau_k \geq t\} \text{ and } N(s,t) = \overline{N}(t) - \underline{N}(s).$$

It has been shown in [9], Lemma 2.6, that the number $N(s,t)$ of subintervals defined by the greedy sequence in an interval $[s,t] \subset [0,T]$ is bounded by

$$N(s,t) \leq \frac{2^{p-1}}{\mu^p} \left((t-s)^p + |w|_{p,s,t}^p\right). \tag{33}$$

In particular, one obtains a finite partition of the interval $[0,T]$ using this construction.

Proof of Lemma 3.8. We denote by $0 = \tau_0 < \cdots < \tau_N = T$ the greedy sequence of times with $\mu = K_2$. Hence,

$$(\tau_{i+1} - \tau_i) + |w|_{p,\tau_i,\tau_{i+1}} \leq K_2$$

for $i = 0, \dots, N-1$, where $N = N(0,T)$ satisfies (33). Then, by (28), we have

$$|y|_{q,s,t} \leq K_1 + |y_s| \tag{34}$$

for all $s, t \in [\tau_i, \tau_{i+1}]$, $s \leq t$. In view of(27), this yields

$$|y_{\tau_{i+1}}| \leq \|y\|_{\infty,\tau_i,\tau_{i+1}} \leq K_1 + 2|y_{\tau_i}|$$

for all $i = 0, \dots, N-1$. If $N = 1$, then (29) trivially holds. Now let $N \geq 2$ and fix $i \in \{0, \dots, N-1\}$ such that $\tau_i < t \leq \tau_{i+1}$. Inductively we get

$$K_1 + |y_{\tau_i}| \leq K_1 + K_1 + 2|y_{\tau_{i-1}}| \leq 2(K_1 + |y_{\tau_{i-1}}|) \leq \cdots \leq 2^i(K_1 + |y_0|).$$

Hence,

$$|y|_{q,\tau_i,\tau_{i+1}} \leq K_1 + |y_{\tau_i}| \leq 2^i(K_1 + |y_0|).$$

By Lemma 3.6, we obtain

$$\begin{aligned} |y|_{q,0,T} \leq N^{\frac{q-1}{q}} \left(\sum_{i=0}^{N-1} |y|^q_{q,\tau_i,\tau_{i+1}} \right)^{\frac{1}{q}} &\leq N^{\frac{q-1}{q}} (K_1 + |y_0|) \left(\sum_{i=0}^{N-1} 2^{iq} \right)^{\frac{1}{q}} \\ &\leq (K_1 + |y_0|) e^{2N}. \end{aligned} \tag{35}$$

Taking (33) into account, we observe that

$$|y|_{q,0,T} \leq (K_1 + |y_0|) e^{2^p K_2^{-p} (|T|^p + |w|^p_{p,0,T})}.$$

In view of the inequality (27), we conclude

$$\|y\|_{\infty,0,T} \leq \|y\|_{q,0,T} \leq (K_1 + 2|y_0|) e^{2^p K_2^{-p} (T^p + |w|^p_{p,0,T})}.$$

Now suppose (34) simplifies to

$$|y|_{q,s,t} \leq K_1.$$

Then we can directly apply the first inequality in (35) to get

$$|y|_{q,0,T} \leq N K_1.$$

By (33), the assertions in (30) and (31) follow. □

3.2 Background on Russo-Vallois forward integration

We now turn to the Russo-Vallois forward integral [43, 44], which is defined as in (3) above. The key tool from the theory of forward integration in our context is the integration-by-parts formula. It involves the following notion of a generalized covariation. Suppose that $(X_t)_{t\in[0,T]}$ and $(Y_t)_{t\in[0,T]}$ are continuous stochastic processes (extended by constant extrapolation to $t > T$, if necessary). For every $\varepsilon > 0$, the *ε-covariation* is defined as

$$C(\varepsilon, Y, X)(t) = \int_0^t \frac{(X_{s+\varepsilon} - X_s)(Y_{s+\varepsilon} - Y_s)}{\varepsilon} \, ds.$$

The *generalized covariation* is then defined to be the limit in the sense of uniform convergence in probability, as ε goes to zero, of $C(\varepsilon, Y, X)$. In the

case of existence, it is denoted by $[X,Y]_t$. We write $[X]_t = [X,X]_t$ for the *generalized quadratic variation* and note that, (provided all terms exist),

$$|[X,Y]_t| \leq ([X]_t[Y]_t)^{1/2} \tag{36}$$

The *integration-by-parts formula* (Proposition 1 in [44]) now states that

$$X_tY_t = X_0Y_0 + \int_0^t X_s d^-Y_s + \int_0^t Y_s d^-X_s + [X,Y]_s, \tag{37}$$

(provided all terms on the right-hand side exist).

The following theorem relates forward integration to Itô integration and to Young integration.

Theorem 3.9. *(1) Suppose $(B_t)_{t\in[0,T]}$ is an $\mathbb{F}$-adapted Brownian motion and $(H_t)_{t\in[0,T]}$ is an $\mathbb{F}$-adapted process satisfying $\int_0^T |H_s|^2 ds < \infty$, P-a.s. Then, the forward integral $(\int_0^t H_s d^-B_s)_{t\in[0,T]}$ exists and coincides with the Itô integral $(\int_0^t H_s dB_s)_{t\in[0,T]}$.*
(2) Suppose $(X_t)_{t\in[0,T]}$ is a stochastic process with paths in $C^p([0,T],\mathbb{R})$ and $(H_t)_{t\in[0,T]}$ is a stochastic process with paths in $W^q([0,T],\mathbb{R})$ for $p,q \geq 1$ such that $\frac{1}{p}+\frac{1}{q} > 1$. Then, the forward integral $(\int_0^t H_s d^-X_s)_{t\in[0,T]}$ exists and coincides with the (pathwise) Young integral $(\int_0^t H_s dX_s)_{t\in[0,T]}$. Moreover, $[\int_0^\cdot H_s d^-X_s] = 0$ on $[0,T]$.

Proof. Part (1) is Theorem 2 in [44]. Part (2) is proved in Proposition 3 of [44] under additional Hölder assumptions. We next provide a proof for the p-variation case.

Step 1: We show that the forward integral exists and coincides with the Young integral.

For $\varepsilon > 0$, let

$$X_t^{\varepsilon-} = \frac{1}{\varepsilon}\int_0^t X_{r+\varepsilon} - X_r \, dr.$$

Then, $X^{\varepsilon-}$ is continuously differentiable with derivative $\dot{X}_t^{\varepsilon-} = \varepsilon^{-1}(X_{t+\varepsilon} - X_t)$ and

$$I^-(\varepsilon, H, dX)(t) = \int_0^t H_s \dot{X}_s^{\varepsilon-} ds = \int_0^t H_s dX_s^{\varepsilon-},$$

where the integral on the right-hand side is the Riemann-Stieltjes integral with respect to the smooth integrator $X^{\varepsilon-}$ and, thus, equals the Young integral of H with respect to $X^{\varepsilon-}$. Fix some $p' > p$ such that $1/p' + 1/q > 1$. Then, by (27) and by the Love-Young inequality in the form of Lemma 3.7,

$$\begin{aligned} &\left\| I^-(\varepsilon, H, dX) - \int_0^\cdot H_s dX_s \right\|_{\infty,0,T} \leq \left\| \int_0^\cdot H_s d(X^{\varepsilon-} - X)_s \right\|_{p',0,T} \\ \leq\ & C_{p',q} \|H\|_{q,0,T} |X^{\varepsilon-} - X|_{p',0,T}. \end{aligned}$$

Hence, the forward integral $(\int_0^t H_s d^- X_s)_{t\in[0,T]}$ exists and coincides with the Young integral, if

$$\lim_{\varepsilon\to 0} |X^{\varepsilon-} - X|_{p',0,T} = 0, \quad P\text{-a.s.} \tag{38}$$

In order to establish (38),we define $Z_t^\varepsilon = X_t^{\varepsilon-} - X_t$ for $t \in [0,T]$. Since

$$X_t^{\varepsilon-} = \frac{1}{\varepsilon}\int_0^t X_{r+\varepsilon} - X_r\, dr = \frac{1}{\varepsilon}\int_t^{t+\varepsilon} X_r\, dr - \frac{1}{\varepsilon}\int_0^\varepsilon X_r\, dr$$

we obtain, for $0 \le s \le t \le T$,

$$\begin{aligned} Z_t^\varepsilon - Z_s^\varepsilon &= \frac{1}{\varepsilon}\int_t^{t+\varepsilon} X_r - X_t\, dr - \frac{1}{\varepsilon}\int_s^{s+\varepsilon} X_r - X_s\, dr, \\ &= \frac{1}{\varepsilon}\int_0^\varepsilon (X_{t+r} - X_t) - (X_{s+r} - X_s)\, dr. \end{aligned}$$

Thus, by Jensen's inequality,

$$\begin{aligned} & |Z_t^\varepsilon - Z_s^\varepsilon| \\ \le\ & \frac{1}{\varepsilon}\int_0^\varepsilon |X_{t+r} - X_t - (X_{s+r} - X_s)|\, dr \\ \le\ & \left(\frac{1}{\varepsilon}\int_0^\varepsilon |X_{t+r} - X_t - (X_{s+r} - X_s)|^{p'-p} |X_{t+r} - X_t - (X_{s+r} - X_s)|^p\, dr\right)^{\frac{1}{p'}} \\ \le\ & 2^{1-\frac{p}{p'}} \sup_{\substack{r,u\in[0,T]\\ |u-r|\le\varepsilon}} |X_r - X_u|^{1-\frac{p}{p'}} \left(2^{p-1}\frac{1}{\varepsilon}\int_0^\varepsilon |X_{t+r} - X_{s+r}|^p + |X_t - X_s|^p\, dr\right)^{\frac{1}{p'}} \\ \le\ & 2^{1-\frac{1}{p'}} \sup_{\substack{r,u\in[0,T]\\ |u-r|\le\varepsilon}} |X_r - X_u|^{1-\frac{p}{p'}} \left(\frac{1}{\varepsilon}\int_0^\varepsilon |X_{(\cdot+r)}|_{p,s,t}^p\, dr + |X|_{p,s,t}^p\right)^{\frac{1}{p'}}. \end{aligned}$$

In view of Proposition 3.4, it is easy to check that

$$\varphi(s,t) = \frac{1}{\varepsilon}\int_0^\varepsilon |X_{(\cdot+r)}|_{p,s,t}^p\, dr + |X|_{p,s,t}^p$$

is superadditive on $\Delta([0,T])$. Hence, Lemma 3.3 implies

$$\begin{aligned} |Z^\varepsilon|_{p',0,T} &\le 2^{1-\frac{1}{p'}} \sup_{\substack{r,u\in[0,T]\\ |u-r|\le\varepsilon}} |X_r - X_u|^{1-\frac{p}{p'}} \left(\frac{1}{\varepsilon}\int_0^\varepsilon |X_{(\cdot+r)}|_{p,0,T}^p\, dr + |X|_{p,0,T}^p\right)^{\frac{1}{p'}} \\ &\le 2^{1-\frac{1}{p'}} \left(2|X|_{p,0,T}^p\right)^{\frac{1}{p'}} \sup_{\substack{r,u\in[0,T]\\ |u-r|\le\varepsilon}} |X_r - X_u|^{1-\frac{p}{p'}}. \end{aligned}$$

As the paths of X are uniformly continuous on $[0,T]$, we obtain (38).

Step 2: We show that $[X]_t = 0$ for every process X with paths in $C^p([0,T])$.

Choosing $p' \in (p,2)$, the same Young-Love inequality argument as above with $X_{s+\varepsilon} - X_s$ in place of H_s and $X^{\varepsilon-}$ in place of $X^{\varepsilon-} - X$ shows

$$\|C(\varepsilon, X, X)\|_{\infty,0,T} \leq C_{p',p'} \|X_{\cdot+\varepsilon} - X\|_{p',0,T} |X^{\varepsilon-}|_{p',0,T}.$$

By (38), the term $|X^{\varepsilon-}|_{p',0,T}$ stays P-a.s. bounded as $\varepsilon \to 0$. Moreover, the proof of (38) can be modified in a straightforward way to show $\|X_{\cdot+\varepsilon} - X\|_{p',0,T} \to 0$ P-a.s. as $\varepsilon \to 0$. Hence, $[X] = 0$ on $[0,T]$.

Step 3: We show that $[\int_0^\cdot H_s d^- X_s] = 0$ on $[0,T]$.

By Step 1 and Lemma 3.7, the process $Z_t = \int_0^t H_s d^- X_s$ has paths in $C^p([0,T])$. Thus, Step 2 applies to Z. □

In the proof of Theorem 2.5, we will make use of the integration-by-parts formula in the following form:

Theorem 3.10. *Suppose $(B_t)_{t\in[0,T]}$ is an m_2-dimensional Brownian motion and $(w_t)_{t\in[0,T]}$ is an $\mathbb{F}$-adapted process with paths in $C^p([0,T],\mathbb{R}^{m_1})$ as in the general setting of Subsection 2.1. Suppose A, $\hat{A}$, C^j, $\hat{C}^j$, D^i, $\hat{D}^i$ ($j = 1,\dots,m_1$, $i = 1,\dots,m_2$) are $\mathbb{F}$-adapted processes taking values in $\mathbb{R}^{m\times n}$ (without hat), resp. in $\mathbb{R}^{n\times k}$ (with hat). Assume that*

$$\int_0^T \Big(|A_s| + |\hat{A}_s| + \sum_{i=1}^{m_2} \Big(|D_s^i|^2 + |\hat{D}_s^i|^2 \Big) \Big) ds < \infty, \quad P\text{-a.s.},$$

and that the processes C^j, $\hat{C}^j$ have paths of bounded q-variation for some $q \geq 1$ satisfying $\frac{1}{p} + \frac{1}{q} > 1$. Let

$$X_t = X_0 + \int_0^t A_s ds + \sum_{j=1}^{m_1} \int_0^t C_s^j d^- w_s^j + \sum_{j=1}^{m_2} \int_0^t D_s^j d^- B_s^j$$
$$Y_t = Y_0 + \int_0^t \hat{A}_s ds + \sum_{j=1}^{m_1} \int_0^t \hat{C}_s^j d^- w_s^j + \sum_{j=1}^{m_2} \int_0^t \hat{D}_s^j d^- B_s^j.$$

Then, for every $t \in [0,T]$,

$$\begin{aligned} X_t Y_t =& X_0 Y_0 + \int_0^t \Big(X_s \hat{A}_s + A_s Y_s + \sum_{j=1}^{m_2} D_s^j \hat{D}_s^j \Big) ds \\ &+ \sum_{j=1}^{m_1} (X_s \hat{C}_s^j + C_s^j Y_s) d^- w_s^j + \sum_{j=1}^{m_2} (X_s \hat{D}_s^j + D_s^j Y_s) d^- B_s^j \end{aligned}$$

Sketch of the proof. We consider the scalar-valued case $m = n = k = 1$ only, but note that the extension to the matrix-valued case is straightforward. By Theorem 3.9, all forward integrals exist. By Corollary 2 in [44] and

by polarization, the generalized covariation of two continuous local martingales coincides with the usual cross-variation of local martingales (see, e.g., Chapter 1.1.5 in [30]). Then, by bilinearity of the generalized covariation in conjunction with the zero quadratic variation property of the Young integrals (Theorem 3.9, (2)) and (36),

$$[X,Y]_t = \sum_{j=1}^{m_2}\sum_{i=1}^{m_2}[\int_0^\cdot D_s^i dB_s^i, \int_0^\cdot \hat{D}_s^j dB_s^j]_t = \sum_{j=1}^{m_2}\int_0^t D_s^j \hat{D}_s^j ds.$$

Thus, (37) applies. □

Remark 3.11. Suppose that the forward integral $\int_0^t H_s d^- X_s$ exists and that Z is a random variable. Then, by the definition of the forward integral, it easily follows that

$$\int_0^t Z H_s d^- X_s = Z \int_0^t H_s d^- X_s.$$

In particular, under the assumptions of Theorem 3.9, (1),

$$Z\int_0^t H_s dB_s = \int_0^t H_s d^- B_s = \int_0^t Z H_s d^- B_s.$$

The integral on the right-hand side cannot be interpreted as Itô integral, because the integrand $(ZH_s)_{s\in[0,T]}$ is not $\mathbb{F}$-adapted, unless Z is $\mathcal{F}_0$-measurable.

4 Proofs

4.1 On the proof of Theorem 2.3

In this subsection, we briefly explain some of the key arguments leading to Theorem 2.3. We first consider the SDE system for $\mathcal{X}^u$ and recall that it can be decomposed into one subsystem driven by the p-variation process w, and another one driven by the Brownian motion B. We will mainly concentrate on the first subsystem, which reads,

$$\xi_t^u = \xi_0(u) + \int_0^t b(r,\xi_r^u,u)\,dr + \sum_{j=1}^{m_1}\int_0^t \sigma^j(r,\xi_r^u,u)\,d^- w_r^j.$$

By Theorem 3.9, the integral with respect to w is a Young integral. Then, existence and uniqueness under (H_1)–(H_3) are a direct consequence of Theorem 3.6 in [9]. Note that [9] is concerned with Young differential equations driven by a deterministic p-variation function. This is the typical framework for Young differential equations (cp. also [32]). We, thus, apply their results pathwise, i.e., for a fixed realization $w(\omega)$ of the p-variation process w. As

the cost functional J in (5) averages over the realizations by taking an expectation, we need to control the growth of the solutions in dependence of w. This is the reason to impose the boundedness assumptions on b and σ, which are, in fact, not required for existence and uniqueness.

Lemma 4.1. *Under (H1)–(H3), there is a constant C_1 independent of u and (the realization of) w such that for every $0 \le s \le t \le T$:*

$$|\xi^u|_{p,s,t} \le \frac{1}{2C_1}(1 + |\xi^u|_{p,s,t})((t-s) + |w|_{p,s,t}).$$

A routine proof, which relies on the Young-Love inequality (26) and standard Lipschitz estimates, can be found in [48], Lemma 2.27. As a consequence of the previous lemma, we observe that

$$(t-s) + |w|_{p,s,t} \le C_1 \quad \Rightarrow \quad |\xi^u|_{p,s,t} \le 1. \tag{39}$$

Combining (39) with (31) yields

$$\|\xi^u\|_{p,0,T} \le L + 2^{p-1} C_1^{-p} \left(T^p + |w|_{p,0,T}^p \right), \tag{40}$$

where L is any upper bound for $u \mapsto |\xi_0(u)|$. Then, by (6),

$$\sup_{u \in \mathcal{U}} \mathrm{E}[\|\xi^u\|_{\infty,0,T}^l] < \infty$$

for every $l \ge 1$. Summarizing, the boundedness assumption on b and σ ensures that the p-variation norm of ξ^u grows linearly in $|w|_{p,0,T}^p$, while without the boundedness assumption the solutions can grow exponentially in $|w|_{p,0,T}^p$, cp. Proposition 1 in [32]. The p-variation estimate for ξ^u obtained in (40) will turn out to be crucial for controlling the growth of the Fréchet derivative of ξ to which we turn now.

We first provide a heuristic derivation of the SDE for the Fréchet derivative of ξ in the parameter. To this end, fix $u \in \mathcal{U}$, a vector $\bar{u} \in \mathbb{R}^d$ of length 1, and choose ϵ sufficiently small such that $u_\epsilon := u + \epsilon \bar{u} \in \mathcal{U}$. Then, the difference quotient for the directional derivative reads

$$\begin{aligned}
&\frac{\xi_t^{u_\epsilon} - \xi_t^u}{\epsilon} := \frac{\xi_0(u_\epsilon) - \xi_0(u)}{\epsilon} \\
+ \quad &\int_0^t \frac{b(r, \xi_r^{u_\epsilon}, u_\epsilon) - b(r, \xi_r^u, u_\epsilon)}{\epsilon} + \frac{b(r, \xi_r^u, u_\epsilon) - b(r, \xi_r^u, u)}{\epsilon} dr \\
+ \quad &\sum_{j=1}^{m_1} \int_0^t \frac{\sigma^j(r, \xi_r^{u_\epsilon}, u_\epsilon) - \sigma^j(r, \xi_r^u, u_\epsilon)}{\epsilon} + \frac{\sigma^j(r, \xi_r^u, u_\epsilon) - \sigma^j(r, \xi_r^u, u)}{\epsilon} \, d^- w_r^j.
\end{aligned}$$

Passing formally to the limit $\epsilon \to 0$ suggests that the directional derivative of ξ in direction $\bar{u}$ at u is given by the solution $y^{u,\bar{u}}$ of the linear SDE

$$\begin{aligned} y_t^{u,\bar{u}} &= D\xi_0(u)\bar{u} + \int_0^t \left(b_\xi(r,\xi_r^u,u)y_r^{u,\bar{u}} + b_u(r,\xi_r^u,u)\bar{u}\right)\, dr \\ &\quad + \sum_{j=1}^{m_1} \int_0^t \left(\sigma_\xi^j(r,\xi_r^u,u)y_r^{u,\bar{u}} + \sigma_u^j(r,\xi_r^u,u)\bar{u}\right)\, d^-w_r^j. \end{aligned}$$

Assuming, for the moment, that the Fréchet derivative y^u of ξ at u exists, we obtain $y^{u,\bar{u}} = y^u \cdot \bar{u}$. Thus, y^u solves

$$\begin{aligned} y_t^u &= D\xi_0(u) + \int_0^t \left(b_\xi(r,\xi_r^u,u)y_r^u + b_u(r,\xi_r^u,u)\right)\, dr \\ &\quad + \sum_{j=1}^{m_1} \int_0^t \left(\sigma_\xi^j(r,\xi_r^u,u)y_r^u + \sigma_u^j(r,\xi_r^u,u)\right)\, d^-w_r^j, \end{aligned} \tag{41}$$

corresponding to the first n_1 lines of the matrix valued SDE (7) for $\mathcal{Y}^u$. This heuristic argument can be made rigorous in a similar way as, e.g., in [24] (in the framework of controlled SDEs driven by a fractional Brownian motion, where Hölder norms are applied) or in Proposition 8 of [32] (where parameter dependence in the initial condition in a p-variation setting is considered), see Section 2.1.3 in [48] for the details. As before, all arguments leading to the differentiability of the Young SDE ξ^u are applied pathwise. Hence, in order to interchange differentiation and expectation when deriving the gradient representation (8), uniform integrability of the difference quotients is required. In view of the mean-value theorem and the de la Vallée-Poussin criterion for uniform integrability, this problem can be reduced to bounding the $L^l(\Omega,P)$-norm of $\|y^u\|_{\infty,0,T}$ uniformly in $u \in \mathcal{U}$. The following lemma explains how to derive suitable bounds for a simplified equation (to avoid unnecessary technicalities).

Lemma 4.2. *Suppose $m_1 = 1$, $z_0 : \mathcal{U} \to \mathbb{R}^{n_1}$ is bounded, $f : \mathbb{R}^{n_1} \to \mathbb{R}^{n_1 \times n_1}$ is bounded and Lipschitz continuous and that the $\mathbb{R}^{n_1}$-valued process z^u solves*

$$z_t^u = z_0(u) + \int_0^t f(\xi_s^u) z_s^u dw_s, \quad 0 \le t \le T,$$

(in the sense of Young integration). Then, there is a constant C independent of u and (the realization of) w such that

$$\|z^u\|_{p,0,T} \le 2|z_0(u)|e^{C(T^p + |w|_{p,0,T}^p)}.$$

Proof. Fix $L \ge 0$ sufficiently large such that f is bounded by L and L is a Lipschitz constant for f. By the Young-Love inequality in the form of

Lemma 3.7, there is a universal constant C_p such that for every $0 \le s \le t \le T$,

$$|z^u|_{p,s,t} \le C_p \|f(\xi^u_\cdot) z^u_\cdot\|_{p,s,t} |w|_{p,s,t}.$$

In view of Lemma 3.5,

$$\|f(\xi^u_\cdot) z^u_\cdot\|_{p,s,t} \le 2\|f(\xi^u_\cdot)\|_{p,s,t} \|z^u_\cdot\|_{p,s,t} \le 2L(1 + |\xi^u|_{p,s,t})(|z^u_s| + |z^u|_{p,s,t}).$$

Hence,

$$|z^u|_{p,s,t} \le 2LC_p(1 + |\xi^u|_{p,s,t})(|z^u_s| + |z^u|_{p,s,t})|w|_{p,s,t}. \tag{42}$$

By (39), there is a constant C_1 independent of u and w such that $|\xi^u|_{p,s,t} \le 1$, if $(t-s) + |w|_{p,s,t} \le C_1$. Let $C_2 := \min\{C_1, (8LC_p)^{-1}\}$. Then, if $(t-s) + |w|_{p,s,t} \le C_2$,

$$|z^u|_{p,s,t} \quad \le \quad 4LC_p(|z^u_s| + |z^u|_{p,s,t})|w|_{p,s,t} \le \frac{1}{2}(|z^u_s| + |z^u|_{p,s,t}),$$

i.e., $|z^u|_{p,s,t} \le |z^u_s|$. Hence, Gronwall's inequality (Lemma 3.8) yields

$$\|z^u\|_{p,0,T} \le 2|z_0(u)| e^{C(T^p + |w|^p_{p,0,T})}$$

for $C = 2^p C_2^{-p}$. □

An important observation of the proof is, that p-variation estimates for z^u depend on the p-variation regularity of ξ^u via the coefficient $f(\xi^u)$. The boundedness assumptions on b and σ allow to control $|\xi^u|_{p,s,t}$ via Lemma 4.1 and lead to an exponential bound for the p-variation norm of z^u in terms of $|w|_{p,s,t}$. In view of the exponential moment bound (6), and taking the boundedness of z_0 as a function in u into account, we conclude that for every $l \ge 1$

$$\sup_{u \in \mathcal{U}} \mathrm{E}[\|z^u\|^l_{\infty,0,T}] < \infty.$$

With a little extra effort (but essentially the same argument), the same type of estimate can be obtained for y^u in place z^u.

Having the results on the differentiability of ξ^u in the parameter at hand, one can proceed to study the second subsystem of (4) given by

$$x^u_t = x_0(u) + \int_0^t \hat{b}(r, \xi^u_r, x^u_r, u)\, dr + \sum_{j=1}^{m_2} \int_0^t \hat{\sigma}^j(r, \xi^u_r, x^u_r, u)\, d^- B^j_r,$$

where the forward integrals coincide with Itô integrals by Theorem 3.9. Existence and uniqueness are standard results under the Lipschitz conditions implied by (B_1)–(B_3), see, e.g., Chapter 1.6 in [49]. Differentiability of SDEs with respect to a parameter is also classical in the semimartingale case, see, e.g., Theorem 39 in [42]. For the particular SDE satisfied by x^u some technicalities related the coupling of ξ^u into the equation must be taken into account, but the proofs follow routine argument. Of course, in contrast to Young integration, Itô's stochastic calculus is tailor-made for obtaining the required $L^l(\Omega, P)$-bounds via the Burkholder-Davis-Gundy inequality.

4.2 On the proof of Theorem 2.5

In this section, we sketch the proof of Theorem 2.5. We first discuss existence and uniqueness of the matrix-valued homogeneous equations (9)–(10). Given the specific form of these equations, it is straightforward to check that solution processes need to be of the form

$$\Phi_t^{s_0} = \begin{pmatrix} \phi_t^{s_0} & 0 \\ \tilde{\phi}_t^{s_0} & \hat{\phi}_t^{s_0} \end{pmatrix}, \qquad \Psi_t^{s_0} = \begin{pmatrix} \psi_t^{s_0} & 0 \\ \tilde{\psi}_t^{s_0} & \hat{\psi}_t^{s_0} \end{pmatrix},$$

for every $t \in [s_0, T]$. Here, the 'component' processes solve the lower-dimensional matrix-valued SDEs

$$\begin{aligned}
\phi_t^{s_0} &= I_{n_1} + \int_{s_0}^t b_\xi^u(r)\phi_r^{s_0}\, dr + \sum_{j=1}^m \int_{s_0}^t \sigma_\xi^{u,j}(r)\phi_r^{s_0}\, d^- w_r^j \\
\hat{\phi}_t^{s_0} &= I_{n_2} + \int_{s_0}^t \hat{b}_x^u(r)\hat{\phi}_r^{s_0}\, dr + \sum_{j=1}^{m_2} \int_{s_0}^t \hat{\sigma}_x^{u,j}(r)\hat{\phi}_r^u\, d^- B_r^j \\
\tilde{\phi}_t^{s_0} &= \int_{s_0}^t \hat{b}_x^u(r)\tilde{\phi}_r^{s_0} + \hat{b}_\xi^u(r)\phi_r^{s_0}\, dr + \sum_{j=1}^{m_2} \int_{s_0}^t \hat{\sigma}_x^{u,j}(r)\tilde{\phi}_r^{s_0} + \hat{\sigma}_\xi^{u,j}(r)\phi_r^{s_0} d^- B_r^j
\end{aligned}$$

and,

$$\begin{aligned}
\psi_t^{s_0} &= I_{n_1} - \int_{s_0}^t \psi_r^{s_0} b_\xi^u(r)\, dr - \sum_{j=1}^m \int_{s_0}^t \psi_r^{s_0}\sigma_\xi^{u,j}(r)\, d^- w_r^j \\
\hat{\psi}_t^{s_0} &= I_{n_2} - \int_{s_0}^t \hat{\psi}_r^{s_0} \left(\hat{b}_x^u(r) - \sum_{j=1}^{m_2} \hat{\sigma}_x^{u,j}(r)^2 \right) dr - \sum_{j=1}^{m_2} \int_{s_0}^t \hat{\psi}_r^{s_0}\hat{\sigma}_x^{u,j}(r)\, d^- B_r^j, \\
\tilde{\psi}_t^{s_0} &= -\int_{s_0}^t \tilde{\psi}_r^{s_0} b_\xi^u(r) + \hat{\psi}_r^{s_0} \left[\hat{b}_\xi^u(r) - \sum_{j=1}^{m_2} \hat{\sigma}_x^{u,j}(r)\hat{\sigma}_\xi^{u,j}(r) \right] dr \\
&\quad - \sum_{j=1}^{m_1} \int_{s_0}^t \tilde{\psi}_r^{s_0}\sigma_\xi^{u,j}(r)\, d^- w_r^j - \sum_{j=1}^{m_2} \int_{s_0}^t \hat{\psi}_r^{s_0}\hat{\sigma}_\xi^{u,j}(r) d^- B_r^j
\end{aligned}$$

respectively. By Theorem 3.9, the first equation of each of the systems is pathwise a Young differential equation driven by the p-variation process w, for which existence and uniqueness can be reduced to Proposition 2.2 in [10]. A bound of the form $C\exp\{C|w|_{p,0,T}^p\}$ for the p-variation norm of the solutions can be derived by the techniques explained in Lemma 4.2, which, in view of the exponential moment bound (6), implies that ϕ^{s_0}, $\psi^{s_0} \in L_{\mathbb{F}}^l(\Omega, C^{p,0}[s_0, T], \mathbb{R}^{n_1 \times n_1})$. The second and the third equation in the system for Φ^{s_0} and the second equation in the system vor Ψ^{s_0} are linear matrix-valued SDEs driven by a Brownian motion in the sense of Itô integration

(applying Theorem 3.9, again). Existence, uniqueness and L^l-integrability are classical for these equations, see, e.g., Chapter 1.6 in [49]. Thus, the most interesting equation is the one for $\tilde{\psi}^{s_0}$, which features a linear term in $\tilde{\psi}^{s_0}$ inside the forward integral w.r.t. to the p-variation process w and an inhomogeneity in terms of the forward integral w.r.t to the Brownian motion B. A formal application of the variation-of-constants formula provides the candidate solution

$$\begin{aligned}\tilde{\psi}_t^{s_0} :=& \Bigg[- \int_{s_0}^t \hat{\psi}_r^{s_0} \left[\hat{b}_\xi^u(r) - \sum_{j=1}^{m_2} \hat{\sigma}_x^{u,j}(r) \hat{\sigma}_\xi^{u,j}(r) \right] (\psi_r^{s_0})^{-1} \, dr \\ & - \sum_{j=1}^{m_2} \int_{s_0}^t \hat{\psi}_r^{s_0} \hat{\sigma}_\xi^{u,j}(r) (\psi_r^{s_0})^{-1} \, dB_r^j \Bigg] \psi_t^{s_0} =: X_t \psi_t^{s_0}.\end{aligned}$$

Now, the integration-by-parts formula in Theorem 3.10 yields

$$\begin{aligned}X_t \psi_t^{s_0} \quad = \quad & - \int_{s_0}^t \left(\hat{\psi}_r^{s_0} \left[\hat{b}_\xi^u(r) - \sum_{j=1}^{m_2} \hat{\sigma}_x^{u,j}(r) \hat{\sigma}_\xi^{u,j}(r) \right] + (X_r \psi_r^{s_0}) b_\xi^u(r) \right) dr \\ & - \sum_{j=1}^{m} \int_{s_0}^t (X_r \psi_r^{s_0}) \sigma_\xi^{u,j}(r) \, d^- w_r^j - \sum_{j=1}^{m_2} \int_{s_0}^t \hat{\psi}_r^{s_0} \hat{\sigma}_\xi^{u,j}(r) \, dB_r^j,\end{aligned}$$

i.e, $\tilde{\psi}^{s_0} = X\psi^{s_0}$ is a solution to the last SDE in the Ψ^{s_0}-system. Note that ψ^{s_0} is indeed invertible and $\phi^{s_0} = (\psi^{s_0})^{-1}$, which can again be verified by integration-by-parts. The L^l-integrability of $\tilde{\psi}^{s_0}$ is a simple consequence of Hölder's inequality, the Burkholder-Davis-Gundy inequality and the already established integrability properties of $\hat{\psi}^{s_0}$, ψ^{s_0}, and ϕ^{s_0}. Uniqueness can be derived by computing $\tilde{\psi}^{s_0}\phi^{s_0}$ in the same way (where $\tilde{\psi}^{s_0}$ is an arbitrary solution to the last SDE in the Ψ^{s_0}-system) and using once more that $\phi^{s_0} = (\psi^{s_0})^{-1}$.

With existence, uniqueness, and the required integrability properties of Φ^{s_0} and Ψ^{s_0} at hand, a direct computation, using Theorem 3.10 once more, shows $\Psi^{s_0}\Phi^{s_0} = I_{n_1+n_2}$, i.e., Φ^{s_0} and Ψ^{s_0} are the inverses to each other. Write $\Phi = \Phi^0$ and define

$$\begin{aligned}Y_t \quad := \quad & \begin{pmatrix} D\xi_0(u) \\ Dx_0(u) \end{pmatrix} + \int_0^t (\Phi_r)^{-1} \left[\begin{pmatrix} b_u^u(r) \\ \hat{b}_u^u(r) \end{pmatrix} - \sum_{j=1}^{m_2} \begin{pmatrix} 0 \\ \hat{\sigma}_x^{u,j}(r) \hat{\sigma}_u^{u,j}(r) \end{pmatrix} \right] dr \\ & + \sum_{j=1}^{m_1} \int_0^t (\Phi_r)^{-1} \begin{pmatrix} \sigma_u^{u,j}(r) \\ 0 \end{pmatrix} d^- w_r^j + \sum_{j=1}^{m_2} \int_0^t (\Phi_r)^{-1} \begin{pmatrix} 0 \\ \hat{\sigma}_u^{u,j}(r) \end{pmatrix} d^- B_r^j.\end{aligned}$$

Then, a final application of the integration-by-parts formula verifies that $\mathcal{Y}_t^u = \Phi_t Y_t$ solves (7).

Note that this argument also implies existence of a solution for (7). Uniqueness can be derived in the same way, by computing $\Psi_t^0 \mathcal{Y}_t^u$ for some arbitrary solution $\mathcal{Y}_t^u$ to (7).

4.3 On the proof of Theorem 2.6

Before we derive the adjoint equation for $(\Lambda_t)_{t\in[0,T]}$, we first prove the gradient representation (12).

Lemma 4.3. *The gradient of the cost functional J admits the representation* (12).

Proof. Inserting (11) into (8), and taking Remark 3.11 into account, we get

$$\begin{aligned}\nabla J(u) = \mathrm{E}\Bigg[&\sum_{\mu=1}^{M} E[g_\mu(\mathcal{X}^u_{T_\mu})] g'_\mu(\mathcal{X}^u_{T_\mu}) \Phi_{T_\mu} \begin{pmatrix} D\xi_0(u) \\ Dx_0(u) \end{pmatrix} \\ &+ \sum_{\mu=1}^{M} \int_0^{T_\mu} E[g_\mu(\mathcal{X}^u_{T_\mu})] g'_\mu(\mathcal{X}^u_{T_\mu}) \Phi_{T_\mu} \Phi_r^{-1} \\ &\cdot \left[\begin{pmatrix} b^u_u(r) \\ \hat{b}^u_u(r) \end{pmatrix} - \sum_{j=1}^{m_2} \begin{pmatrix} 0 \\ \hat{\sigma}^{u,j}_x(r) \hat{\sigma}^{u,j}_u(r) \end{pmatrix} \right] dr \\ &+ \sum_{\mu=1}^{M} \sum_{j=1}^{m_1} \int_0^{T_\mu} E[g_\mu(\mathcal{X}^u_{T_\mu})] g'_\mu(\mathcal{X}^u_{T_\mu}) \Phi_{T_\mu} \Phi_r^{-1} \begin{pmatrix} \sigma^{u,j}_u(r) \\ 0 \end{pmatrix} d^- w^j_r \\ &+ \sum_{\mu=1}^{M} \sum_{j=1}^{m_2} \int_0^{T_\mu} E[g_\mu(\mathcal{X}^u_{T_\mu})] g'_\mu(\mathcal{X}^u_{T_\mu}) \Phi_{T_\mu} \Phi_r^{-1} \begin{pmatrix} 0 \\ \hat{\sigma}^{u,j}_u(r) \end{pmatrix} d^- B^j_r \Bigg]. \end{aligned}$$

Interchanging summation and integration we, then, obtain

$$\begin{aligned}\nabla J(u) = \mathrm{E}\Bigg[&\sum_{\mu=1}^{M} E[g_\mu(\mathcal{X}^u_{T_\mu})] g'_\mu(\mathcal{X}^u_{T_\mu}) \Phi_{T_\mu} \begin{pmatrix} D\xi_0(u) \\ Dx_0(u) \end{pmatrix} \\ &+ \int_0^{T} \sum_{\mu;\, T_\mu \geq r} E[g_\mu(\mathcal{X}^u_{T_\mu})] g'_\mu(\mathcal{X}^u_{T_\mu}) \Phi_{T_\mu} \Phi_r^{-1} \\ &\cdot \left[\begin{pmatrix} b^u_u(r) \\ \hat{b}^u_u(r) \end{pmatrix} - \sum_{j=1}^{m_2} \begin{pmatrix} 0 \\ \hat{\sigma}^{u,j}_x(r) \hat{\sigma}^{u,j}_u(r) \end{pmatrix} \right] dr \\ &+ \sum_{j=1}^{m_1} \int_0^{T} \sum_{\mu;\, T_\mu \geq r} E[g_\mu(\mathcal{X}^u_{T_\mu})] g'_\mu(\mathcal{X}^u_{T_\mu}) \Phi_{T_\mu} \Phi_r^{-1} \begin{pmatrix} \sigma^{u,j}_u(r) \\ 0 \end{pmatrix} d^- w^j_r \\ &+ \sum_{j=1}^{m_2} \int_0^{T} \sum_{\mu;\, T_\mu \geq r} E[g_\mu(\mathcal{X}^u_{T_\mu})]^\top g'_\mu(\mathcal{X}^u_{T_\mu}) \Phi_{T_\mu} \Phi_r^{-1} \begin{pmatrix} 0 \\ \hat{\sigma}^{u,j}_u(r) \end{pmatrix} d^- B^j_r \Bigg]. \end{aligned}$$

Substituting the definition (13) of Λ into this expression, finally yields (12). □

Proof of Theorem 2.6. Integrability of Λ is inherited from Φ and Ψ. Recall that $\Phi_t^{-1} = \Psi_t$ by Theorem 2.5. Inserting the expression (10) for $\Psi_t - \Psi_{T_\mu}$

into the definition of Λ and interchanging summation and integration again, we obtain, thanks to Remark 3.11,

$$\begin{aligned}
\Lambda_t &= \sum_{T_\mu \geq t} E[g_\mu(\mathcal{X}^u_{T_\mu})] g'_\mu(\mathcal{X}^u_{T_\mu}) \Phi_{T_\mu} \Phi_t^{-1} \\
&= \sum_{T_\mu \geq t} E[g_\mu(\mathcal{X}^u_{T_\mu})] g'_\mu(\mathcal{X}^u_{T_\mu}) \Phi_{T_\mu} (\Phi_{T_\mu}^{-1} + \Psi_t - \Psi_{T_\mu}) \\
&= \sum_{T_\mu \geq t} E[g_\mu(\mathcal{X}^u_{T_\mu})] g'_\mu(\mathcal{X}^u_{T_\mu}) + \int_t^T \sum_{T_\mu \geq r} E[g_\mu(\mathcal{X}^u_{T_\mu})] g'_\mu(\mathcal{X}^u_{T_\mu}) \Phi_{T_\mu} \Phi_r^{-1} \\
&\quad \cdot \left[\begin{pmatrix} b^u_\xi(r) & 0 \\ \hat{b}^u_\xi(r) & \hat{b}^u_x(r) \end{pmatrix} - \sum_{j=1}^{m_2} \begin{pmatrix} 0 & 0 \\ \hat{\sigma}^u_\xi(r) & \hat{\sigma}^{u,j}_x(r) \end{pmatrix}^2 \right] dr \\
&\quad + \sum_{j=1}^{m_1} \int_t^T \sum_{T_\mu \geq r} E[g_\mu(\mathcal{X}^u_{T_\mu})] g'_\mu(\mathcal{X}^u_{T_\mu}) \Phi_{T_\mu} \Phi_r^{-1} \begin{pmatrix} \sigma^{u,j}_\xi(r) & 0 \\ 0 & 0 \end{pmatrix} d^- w^j_r \\
&\quad + \sum_{j=1}^{m_2} \int_t^T \sum_{T_\mu \geq r} E[g_\mu(\mathcal{X}^u_{T_\mu})] g'_\mu(\mathcal{X}^u_{T_\mu}) \Phi_{T_\mu} \Phi_r^{-1} \begin{pmatrix} 0 & 0 \\ \hat{\sigma}^{u,j}_\xi(r) & \hat{\sigma}^{u,j}_x(r) \end{pmatrix} d^- B^j_r.
\end{aligned}$$

Recalling the definition of Λ_r, the proof is finished. □

4.4 On the proof of Theorem 2.8

As in Subsection 4.1, we will put emphasis on the techniques of proof for the Young SDEs. The key difficulty is to control the dependence of the constants on the driving path w in order to come up with $L^l(\Omega, P)$-estimates. For the sake of illustration, we consider the following simplified variant of the linear equation for y^u in (41):

$$z^u_t = z_0(u) + \int_0^t f(\xi^u_s) z^u_s dw_s, \quad 0 \leq t \leq T, \tag{43}$$

under the assumptions of Lemma 4.2. Given a partition $\Pi^{\mathrm{E}} = (t_i)_{i=0,\dots,n}$ of $[0,T]$, we consider the Euler scheme

$$z^{n,u}_t = z^{n,u}_{t_i} + f(\xi^u_{t_i}) z^{n,u}_{t_i} (w_t - w_{t_i}),\ t \in (t_i, t_{i+1}], \quad z^{n,u}_0 = z_0(u). \tag{44}$$

Note that the Euler scheme actually should depend on the Euler approximation $\xi^{n,u}$ for ξ^u via the coefficient $f(\xi^{n,u}_{t_i})$, leading to an extra error term, but the scheme in (44) already contains all essential difficulties, including the dependence of the coefficient on the path of ξ^u in (43).

Theorem 4.4. *Under the standing assumptions and the assumptions specified in Lemma 4.2, there is a constant C independent of $\Pi^E = (t_i)_{i=0,\dots,n}$, u, and (the realization of) w such that*

$$\|z^u - z^{n,u}\|_{p,0,T} \leq \Big(\max_{i=0,\dots,n-1} |t_{i+1} - t_i| + |w|_{p,t_i,t_{i+1}} \Big)^{2-p} C e^{C|w|^p_{p,0,T}}.$$

In view of the exponential moment bound (6), Theorem 4.4 and Hölder's inequality imply existence of a constant $C_{z,l}$ such that

$$\mathrm{E}[\|z^u - z^{n,u}\|^l_{\infty,0,T}]^{1/l} \le C_{z,l}(\delta_{2l}(\Pi^E))^{2-p}$$

for every $l \ge 1$ (cp. Theorem 2.8), where δ_l is defined in (22).

The exponential bound on $|w|^p_{p,0,T}$ in Theorem 4.4 suggests an application of Gronwall's lemma. However, Lemma 3.8 is not well-suited for the Euler approximation. Indeed, for $t_i < s < t < t_{i+1}$ any estimate for $|z^u - z^{n,u}|_{p,s,t}$ will depend on $z^{n,u}_{t_i}$ (where t_i is outside the interval $[s,t]$), while Lemma 3.8 requires an estimate in terms of $|z^u_s - z^{n,u}_s|$. The following variant of Gronwall's lemma is tailor-made to deal with such a situation and will be applied in the proof of Theorem 4.4.

Lemma 4.5 (Gronwall type lemma on the Euler partition)**.**
Let $\Pi^E = (t_i)_{i=0,\dots,n}$ be a partition of $[0,T]$ and let $x \in W^p([0,T],\mathbb{R}^{n\times d})$, where $p \in (1,2)$. Furthermore let $w : [0,T] \to \mathbb{R}^m$ ($m = d$ or $m = 1$) be a continuous function of finite p-variation, $K_1, a > 0$ be constants. If for every $t_i \in \Pi^E$, $i \in \{0,\dots,n-1\}$, we have

$$|x|_{p,t_i,t_{i+1}} \le a(K_1 + |x_{t_i}|)(|t_{i+1} - t_i| + |w|_{p,t_i,t_{i+1}}), \tag{45}$$

and if there exists a constant $K_2 \le \frac{1}{a}$ such that for every $t_l, t_k \in \Pi^E$ with $0 \le t_l < t_{l+1} < t_k \le T$

$$|t_k - t_l| + |w|_{p,t_l,t_k} \le K_2 \quad \Rightarrow |x|_{p,t_l,t_k} \le K_1 + |x_{t_l}|, \tag{46}$$

then

$$\begin{aligned}|x|_{p,0,T} \le \frac{1}{2}(K_1 + |x_0|)\left(2^p K_2^{-p}\left(T^p + |w|^p_{p,0,T}\right) + 1\right)\\ \cdot \exp\left(2^p 3 K_2^{-p}\left(T^p + |w|^p_{p,0,T}\right) + 2\right).\end{aligned}$$

Compared to Lemma 3.8, the estimate (46) only needs to hold for intervals whose boundary points are from the Euler partition Π^{E}, while the corresponding estimate (28) in Lemma 3.8 must be verified for all subintervals $[s,t] \subset [0,T]$. The price to pay is the extra condition (45) on the p-variation of x on the small subintervals $[t_i, t_{i+1}]$.

For the proof, we first need to introduce some notation. Write $\Pi^{\mathrm{g}} = (\tau_i)_{i=0,\dots,N}$ for the greedy sequence defined via (32) with $\mu = K_2$. As the points in the greedy sequence, in general, are not included in the Euler partition Π^{E}, we will approximate them by neighboring points in the Euler partition. This leads to the subpartition Π^{c} of Π^{E} consisting of the time points $t \in \Pi^{\mathrm{E}}$ satisfying

$$\exists \tau \in \Pi^{\mathrm{g}} \text{ such that } t = t_{\underline{n}(\tau)} \text{ or } t = t_{\overline{n}(\tau)},$$

where

$$\underline{n} : [0,T] \to \mathbb{N}, s \mapsto \min\{i \in \{0,\dots,n\} |\, t_i \in \Pi^{\mathrm{E}} \text{ and } t_i \geq s\}$$
$$\overline{n} : [0,T] \to \mathbb{N}, s \mapsto \max\{i \in \{0,\dots,n\} |\, t_i \in \Pi^{\mathrm{E}} \text{ and } t_i \leq s\}.$$

Generic points in Π^{c} will be denoted by θ_j (with the convention $\theta_j < \theta_{j+1}$). The construction of Π^{c} is illustrated in Figure 1.

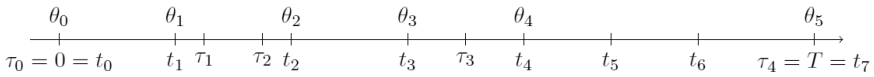

Figure 1: Graphical illustration of the construction of the partition Π^{c}.

We mention the following properties of Π^{c}:

i) If $\tau = t$ for a $\tau \in \Pi^{\mathrm{g}}$ and $t \in \Pi^{\mathrm{E}}$, then there exists $\theta \in \Pi^{\mathrm{c}}$ such that $\theta = t = t_{\underline{n}(\tau)} = t_{\overline{n}(\tau)}$.

ii) There can be multiple partition points $\tau \in \Pi^{\mathrm{E}}$ such that $\theta_j = t_{\overline{n}(\tau)}$ and $\theta_{j+1} = t_{\underline{n}(\tau)}$, e.g. τ_1, τ_2 in Figure 1.

In the situation of ii), let $\tau_{j-1} < \theta_i \leq \tau_j < \cdots < \tau_{j+m} \leq \theta_{i+1} < \tau_{j+m+1}$ Then, $m = N(\theta_i, \theta_{i+1})$, where $N(s,t)$ has been defined right before (33). Moreover, by Lemma 3.6 and the defining property (32) of the greedy sequence,

$$\begin{aligned}
&|\theta_{i+1} - \theta_i| + |w|_{p,\theta_i,\theta_{i+1}} \\
&\leq |\tau_{j+m+1} - \tau_{j-1}| + |w|_{p,\tau_{j-1},\tau_{j+m+1}} \\
&\leq \sum_{i=0}^{m+1} |\tau_{j+i} - \tau_{j-1+i}| + \left((m+2)^{p-1} \sum_{i=0}^{m+1} |w|^p_{p,\tau_{j-1+i},\tau_{j+i}} \right)^{\frac{1}{p}} \\
&\leq \sum_{i=0}^{m+1} |\tau_{j+i} - \tau_{j-1+i}| + (m+2)^{1-\frac{1}{p}} \sum_{i=0}^{m+1} |w|_{p,\tau_{j-1+i},\tau_{j+i}} \\
&\leq (N(\theta_i,\theta_{i+1})+2)^{1-\frac{1}{p}} \sum_{i=0}^{N(\theta_i,\theta_{i+1})+1} (|\tau_{j+i} - \tau_{j-1+i}| + |w|_{p,\tau_{j-1+i},\tau_{j+i}}) \\
&\leq (N(\theta_i,\theta_{i+1})+2)^{2-\frac{1}{p}} K_2. \qquad (47)
\end{aligned}$$

Concerning Π^{c}, we also introduce the notation

$$\underline{\mathcal{N}} : [0,T] \to \mathbb{N}, s \mapsto \min\{i \in \mathbb{N}_0 |\, \theta_i \in \Pi^{\mathrm{c}} \text{ and } \theta_i \geq s\}$$
$$\overline{\mathcal{N}} : [0,T] \to \mathbb{N}, s \mapsto \max\{i \in \mathbb{N}_0 |\, \theta_i \in \Pi^{\mathrm{c}} \text{ and } \theta_i \leq s\}$$

We also define $\mathcal{N}(s,t) := \overline{\mathcal{N}}(t) - \underline{\mathcal{N}}(s)$. Then, by construction $\mathcal{N}(s,t) \leq 2N(s,t) + 1$ for all $(s,t) \in \Delta([0,T])$.

We note that an alternative way to transfer the greedy sequence technique to partitions has been recently suggested by [11], which can be used to bound the discrete-time p-variation norm (i.e., for functions restricted to the grid only) of solutions to stochastic difference equations.

Proof of Lemma 4.5. Recall that the greedy sequence is constructed for the constant $\mu = K_2$. The number of subintervals defined by the partitions Π^{g} and Π^{c} is denoted by $N = N(0,T)$ and $\mathcal{N} = \mathcal{N}(0,T)$ respectively.

We consider the p-variation of x on the subintervals $[\theta_i, \theta_{i+1}]$ of the partition Π^{c} for $i \in \{0, \dots, \mathcal{N}-1\}$, and distinguish the following cases.

Case 1: There exists $\tau_l \in \Pi^{\mathrm{g}}$ and $i \in \{0, \dots, \mathcal{N}-1\}$ such that $\theta_i = t_{\overline{n}(\tau_l)}$ and $\theta_{i+1} = t_{\underline{n}(\tau_l)}$ (e.g. $[\theta_1, \theta_2]$, $[\theta_3, \theta_4]$ in figure 1). By construction it follows that there exists $j \in \{0, \dots, n\}$ such that $\theta_i = t_j$ and $\theta_{i+1} = t_{j+1}$. We estimate using (45), (47), and $aK_2 \leq 1$,

$$\begin{aligned} |x|_{p,\theta_i,\theta_{i+1}} &\leq a(K_1 + |x_{\theta_i}|)(|\theta_{i+1} - \theta_i| + |w|_{p,\theta_i,\theta_{i+1}}) \\ &\leq (K_1 + |x_{\theta_i}|)(N(\theta_i, \theta_{i+1}) + 2)^{2-\frac{1}{p}}. \end{aligned} \tag{48}$$

Case 2: There exists $\tau_j, \tau_{j+1} \in \Pi^{\mathrm{g}}$ and $i \in \{0, \dots, \mathcal{N}-1\}$ such that $\theta_i = t_{\underline{n}(\tau_j)}$ and $\theta_{i+1} = t_{\overline{n}(\tau_{j+1})}$ (e.g. $[\theta_i, \theta_{i+1}]$ for $i \in \{0, 2, 4\}$ in Figure 1). Then there exists a finite number $k - l = m \geq 1$ of subintervals of Π^{E} in the interval $[\theta_i, \theta_{i+1}]$. Let $\theta_i = t_l < t_{l+1} < \cdots < t_{l+m} = t_k = \theta_{i+1}$, if $m = 1$ we have by (45)

$$|x|_{p,\theta_i,\theta_{i+1}} = |x|_{p,t_l,t_{l+1}} \leq a(K_1 + |x_{t_l}|)(|t_{l+1} - t_l| + |w|_{p,t_l,t_{l+1}}).$$

By assumption on the form of $[\theta_i, \theta_{i+1}]$, we have

$$|t_{l+1} - t_l| + |w|_{p,t_l,t_{l+1}} \leq |\tau_{j+1} - \tau_j| + |w|_{p,\tau_j,\tau_{j+1}} \leq K_2 \leq \frac{1}{a},$$

which yields

$$|x|_{p,\theta_i,\theta_{i+1}} = |x|_{p,t_l,t_{l+1}} \leq K_1 + |x_{\theta_i}|. \tag{49}$$

Now let $m \geq 2$, since

$$|t_k - t_l| + |w|_{p,t_l,t_k} \leq K_2,$$

we have by (46) that

$$|x|_{p,\theta_i,\theta_{i+1}} = |x|_{p,t_l,t_k} \leq K_1 + |x_{t_l}| = K_1 + |x_{\theta_i}|. \tag{50}$$

Summarizing, by taking (48), (49) and (50) into account, we have

$$|x|_{p,\theta_i,\theta_{i+1}} \leq (N(\theta_i, \theta_{i+1}) + 2)^{2-\frac{1}{p}}(K_1 + |x_{\theta_i}|) \tag{51}$$

for every $i \in \{0, \dots, \mathcal{N}-1\}$. We show inductively that

$$|x_{\theta_i}| + K_1 \leq e^{2(N(0,\theta_i) + \mathcal{N}(0,\theta_i))}(K_1 + |x_0|) \tag{52}$$

for every $i \in \{0, \dots, \mathcal{N}\}$, noting that the base case $i = 0$ is trivial. Now assume (52) holds for some $i \in \{0, \dots, \mathcal{N} - 1\}$, then, by (51),

$$\begin{aligned} |x_{\theta_{i+1}}| + K_1 &\leq |x_{\theta_i}| + |x|_{p,\theta_i,\theta_{i+1}} + K_1 \\ &\leq (|x_{\theta_i}| + K_1)((N(\theta_i, \theta_{i+1}) + 2)^{2-\frac{1}{p}} + 1) \end{aligned}$$

Noting that

$$(x+2)^{2-\frac{1}{p}} \leq (x+2)^{\frac{3}{2}} \leq \frac{1}{2} e^{2(x+1)}$$

for every $x \geq 0$, we obtain

$$|x_{\theta_{i+1}}| + K_1 \leq (|x_{\theta_i}| + K_1) e^{2(N(\theta_i, \theta_{i+1})+1)}.$$

Since $\mathcal{N}(\theta_i, \theta_{i+1}) = 1$, the induction hypothesis yields

$$\begin{aligned} |x_{\theta_{i+1}}| + K_1 &\leq (|x_0| + K_1) e^{2(N(0,\theta_i)+\mathcal{N}(0,\theta_i)+N(\theta_i,\theta_{i+1})+\mathcal{N}(\theta_i,\theta_{i+1}))} \\ &\leq (|x_0| + K_1) e^{2(N(0,\theta_{i+1})+\mathcal{N}(0,\theta_{i+1}))}, \end{aligned}$$

which completes the proof of (52).

Combining (51) and (52), we have, for $0 \leq i \leq \mathcal{N} - 1$,

$$\begin{aligned} |x|_{p,\theta_i,\theta_{i+1}} &\leq (N(\theta_i, \theta_{i+1}) + 2)^{2-\frac{1}{p}} (K_1 + |x_{\theta_i}|) \\ &\leq \frac{1}{2} e^{2(N(\theta_i,\theta_{i+1})+1)} e^{2(N(0,\theta_i)+\mathcal{N}(0,\theta_i))} (K_1 + |x_0|) \\ &\leq \frac{1}{2} e^{2(N(0,\theta_{i+1})+\mathcal{N}(0,\theta_{i+1}))} (K_1 + |x_0|). \end{aligned}$$

These considerations enable us to finish the proof. We have

$$\begin{aligned} |x|_{p,0,T} &\leq \left(\mathcal{N}(0,T)^{p-1} \sum_{i=0}^{\mathcal{N}(0,T)-1} |x|^p_{p,\theta_i,\theta_{i+1}} \right)^{\frac{1}{p}} \\ &\leq \mathcal{N}(0,T)^{1-\frac{1}{p}} \frac{1}{2} (K_1 + |x_0|) \left(\sum_{i=0}^{\mathcal{N}(0,T)-1} e^{2p(N(0,\theta_{i+1})+\mathcal{N}(0,\theta_{i+1}))} \right)^{\frac{1}{p}} \\ &\leq \mathcal{N}(0,T) \frac{1}{2} (K_1 + |x_0|) e^{2(N(0,T)+\mathcal{N}(0,T))}. \end{aligned}$$

Now keep in mind that $\mathcal{N}(0,T) \leq 2N(0,T) + 1$ by construction of Π^{c}. This implies

$$|x|_{p,0,T} \leq (2N(0,T) + 1) \frac{1}{2} (K_1 + |x_0|) e^{6N(0,T)+2}.$$

Taking (33) into account, we know that

$$N(0,T) \leq 2^{p-1} K_2^{-p} \left(T^p + |w|^p_{p,0,T} \right),$$

and we conclude

$$|x|_{p,0,T} \leq \frac{1}{2}(K_1 + |x_0|)\left(2^p K_2^{-p}\left(T^p + |w|_{p,0,T}^p\right) + 1\right) \cdot \exp\left(2^p 3 K_2^{-p}\left(T^p + |w|_{p,0,T}^p\right) + 2\right).$$

□

We are now in the position to present the proof of Theorem 4.4.

Proof of Theorem 4.4. Step 1: Preliminary estimates and some notation. We fix $u \in \mathcal{U}$ and the partition $\Pi^{\mathrm{E}} = (t_i)_{i=0,\dots,n}$ and choose $L \geq 1$ sufficiently large such that L is a Lipschitz constant for f and an upper bound for $|f|$ and $|\xi_0|$. Write $A := f(\xi^u)$, $z^n := z^{n,u}$, and $z := z^u$ and let $\delta_i := |t_{i+1} - t_i| + |w|_{p,t_i,t_{i+1}}$ and $\delta := \max_{i=0,\dots,n-1} \delta_i$.

We first derive bounds for the p-variation seminorm of A and Az. By Lemma 3.5, $|A|_{p,s,t} \leq L|\xi|_{p,s,t}$ for every $0 \leq s \leq t \leq T$. Hence, by (39), there is a constant $C_1 > 0$ (independent of u, Π^{E}, and w) such that

$$(t-s) + |w|_{p,s,t} \leq C_1 \quad \Rightarrow \quad |A|_{p,s,t} \leq L. \tag{53}$$

Moreover, by Lemma 4.1 and (40),

$$|A|_{p,t_i,t_{i+1}} \leq \frac{L}{2C_1}(1 + L + 2^{p-1} C_1^{-p}(T^p + |w|_{p,0,T}^p)\delta_i$$

Similarly, by Lemma 4.2, (40), and (42), there is a constant C_2 (independent of u, Π^{E}, and w) such that

$$|z|_{p,t_i,t_{i+1}} \leq C_2(1 + |w|_{p,0,T}^p) e^{C_2(T^p + |w|_{p,0,T}^p)} \delta_i$$

By Lemma 3.5,

$$|Az|_{p,t_i,t_{i+1}} \leq \|A\|_{\infty,t_i,t_{i+1}} |z|_{p,t_i,t_{i+1}} + \|z\|_{\infty,t_i,t_{i+1}} |A|_{p,t_i,t_{i+1}}.$$

Combining the previous estimates and bounding $\|z\|_{\infty,t_i,t_{i+1}}$ by Lemma 4.2, we find a constant C' (independent of u, Π^{E}, and w) such that

$$|Az|_{p,t_i,t_{i+1}} \leq C' e^{C'|w|_{p,0,T}^p} \delta_i \tag{54}$$

for every $i = 0, \dots, n-1$. We define

$$\begin{aligned} C_1' &:= C_1'(w) := C' e^{C'|w|_{p,0,T}^p} \\ K_1 &:= K_1(w) := (C_p + 1) C_1'(w)(T + |w|_{p,0,T})^{p-1} \delta^{2-p} \\ K_2 &= \left(C_1^{-1} + 4L + 8C_p L\right)^{-1}, \end{aligned} \tag{55}$$

where $C_p \geq 1$ is the constant from the Young-Love inequality for $q = p$.

Step 2: Verification of (45) *for* $z - z^n$.
Define $\Delta_t := z_t - z_t^n$. We fix a grid point t_i and let $t_i \le s \le t \le t_{i+1}$. Then,

$$\begin{aligned}\Delta_t - \Delta_s &= \int_s^t A_r z_r - A_s z_s dw_r + (A_s z_s - A_{t_i} z_{t_i})(w_t - w_s) \\ &\quad + A_{t_i}(z_{t_i} - z_{t_i}^n)(w_t - w_s).\end{aligned}$$

Hence, by the Young-Love inequality (25), (54), and (55),

$$\begin{aligned}|\Delta_t - \Delta_s| &\le C_p |Az|_{p,s,t}|w|_{p,s,t} + |Az|_{p,t_i,s}|w|_{p,s,t} + L|\Delta_{t_i}||w|_{p,s,t} \\ &\le L\left(|\Delta_{t_i}| + (C_p+1)|Az|_{p,t_i,t_{i+1}}\right)|w|_{p,s,t} \\ &\le L\left(|\Delta_{t_i}| + (C_p+1)C_1'\delta^{2-p}\delta^{p-1}\right)(|t-s| + |w|_{p,s,t}) \\ &\le L\left(|\Delta_{t_i}| + K_1\right)(|t-s| + |w|_{p,s,t}) \qquad (56)\end{aligned}$$

noting that $L \ge 1$ and $\delta \le (T + |w|_{p,0,T})$. Then, by Lemma 3.3,

$$|\Delta|_{p,t_i,t_{i+1}} \le L\left(|\Delta_{t_i}| + K_1\right)(|t_{i+1} - t_i| + |w|_{p,t_i,t_{i+1}}). \qquad (57)$$

Step 3: Estimates for $z - z^n$ *on the grid.*
We now fix $t_\lambda, t_\kappa \in \Pi^{\mathrm{E}}$ such that $t_\lambda \le t_\kappa$. Then,

$$\Delta_{t_\kappa} - \Delta_{t_\lambda} = \sum_{i=\lambda}^{\kappa-1} \int_{t_i}^{t_{i+1}} (A_r z_r - A_{t_i} z_{t_i}) dw_r + A_{t_i}(z_{t_i} - z_{t_i}^n)(w_{t_{i+1}} - w_{t_i}).$$

The first term can be estimated (summand by summand) by the Young-Love inequality (25), while for the second term the variant of the Love-Young estimate (26) for Riemann sums (see [14], Corollary 3.87) applies. We, thus, obtain,

$$\begin{aligned}|\Delta_{t_\kappa} - \Delta_{t_\lambda}| &\le C_p \sum_{i=\lambda}^{\kappa-1} |Az|_{p,t_i,t_{i+1}}|w|_{p,t_i,t_{i+1}} + C_p \|A(z - z^n)\|_{p,t_\lambda,t_\kappa}|w|_{p,t_\lambda,t_\kappa} \\ &=: (I) + (II).\end{aligned}$$

For the first term, we note that, by (54) and Lemma 3.6,

$$\begin{aligned}(I) &\le C_p C_1' \sum_{i=\lambda}^{\kappa-1} (|t_{i+1} - t_i| + |w|_{p,t_i,t_{i+1}})|w|_{p,t_i,t_{i+1}} \\ &\le C_p C_1'(t_\kappa - t_\lambda)\delta + C_p C_1' \delta^{2-p} \sum_{i=\lambda}^{\kappa-1} |w|_{p,t_i,t_{i+1}}^p \\ &\le C_p C_1'(t_\kappa - t_\lambda)\delta^{2-p}\left(T + |w|_{p,0,T}\right)^{p-1} + C_p C_1' \delta^{2-p}|w|_{p,t_\lambda,t_\kappa}|w|_{p,0,T}^{p-1} \\ &\le K_1(|t_\kappa - t_\lambda| + |w|_{p,t_\lambda,t_\kappa}).\end{aligned}$$

For the second term, Lemma 3.5 yields

$$\begin{aligned}(II) &\leq 2C_p\|A\|_{p,t_\lambda,t_\kappa}\|\Delta\|_{p,t_\lambda,t_\kappa}|w|_{p,t_\lambda,t_\kappa}\\ &\leq 2C_p(L+|A|_{p,t_\lambda,t_\kappa})\|\Delta\|_{p,t_\lambda,t_\kappa}(|t_\kappa-t_\lambda|+|w|_{p,t_\lambda,t_\kappa}).\end{aligned}$$

Gathering terms, we obtain,

$$\begin{aligned}&|\Delta_{t_\kappa}-\Delta_{t_\lambda}|\\ \leq\ &\left(K_1+2C_p(L+|A|_{p,t_\lambda,t_\kappa})\|\Delta\|_{p,t_\lambda,t_\kappa}\right)(|t_\kappa-t_\lambda|+|w|_{p,t_\lambda,t_\kappa}).\end{aligned}\tag{58}$$

Step 4: Verification of (46) *for* $z-z^n$.
Fix $t_l,t_k\in\Pi^{\mathrm{E}}$ such that $t_l<t_{l+1}<t_k$ and $|t_k-t_l|+|w|_{p,t_l,t_k}\leq K_2$ for the constant K_2 defined in (55). Let $t_l\leq s\leq t\leq t_k$. We distinguish two cases:

a) $s,t\in[t_i,t_{i+1}]$ for some t_i, i.e., s and t are in the same subinterval.

b) $s\in[t_{\lambda-1},t_\lambda)$ and $t\in(t_\kappa,t_{\kappa+1}]$ for $t_\lambda\leq t_\kappa$, i.e., s and t are in different subintervals.

In case a), we obtain, thanks to (56) and recalling (27),

$$|\Delta_t-\Delta_s|\leq L\left(|\Delta_{t_l}|+|\Delta|_{p,t_l,t_k}+K_1\right)(|t-s|+|w|_{p,s,t})\tag{59}$$

In case b), we decompose

$$|\Delta_t-\Delta_s|\leq|\Delta_t-\Delta_{t_\kappa}|+|\Delta_{t_\kappa}-\Delta_{t_\lambda}|+|\Delta_{t_\lambda}-\Delta_s|$$

Applying (59) to the first and third term and using (58) for the second one, we get, in view of (53) and since $s\leq t_\lambda\leq t_\kappa\leq t$,

$$\begin{aligned}|\Delta_t-\Delta_s| &\leq 2L\left(|\Delta_{t_l}|+|\Delta|_{p,t_l,t_k}+K_1\right)(|t-s|+|w|_{p,s,t})\\ &\quad+\left(K_1+4C_pL\|\Delta\|_{p,t_l,t_k}\right)(|t-s|+|w|_{p,s,t}).\end{aligned}$$

In view of (59), this estimate is valid for every $t_l\leq s\leq t\leq t_k$ (and not just in case b)). Hence, by Lemma 3.3 and the definition K_2,

$$\begin{aligned}|\Delta|_{p,t_l,t_k} &\leq \left((2L+4C_pL)(|\Delta_{t_l}|+|\Delta|_{p,t_l,t_k})+3LK_1\right)(|t_l-t_k|+|w|_{p,t_l,t_k})\\ &\leq \frac{1}{2}(|\Delta|_{p,t_l,t_k}+|\Delta_{t_l}|+K_1).\end{aligned}$$

Thus,

$$|\Delta|_{p,t_l,t_k}\leq|\Delta_{t_l}|+K_1.\tag{60}$$

Step 5: Application of Gronwall's lemma for Euler partitions.
By (57) and (60) we may apply Gronwall's inequality in the form of Lemma 4.5. Taking into account that $\Delta_0=z_0-z_0^n=0$, we obtain,

$$\begin{aligned}&\|z-z^n\|_{p,0,T}=|z-z^n|_{p,0,T}\\ \leq\ &\frac{K_1}{2}\left(2^pK_2^{-p}\left(T^p+|w|_{p,0,T}^p\right)+1\right)\exp\left(2^p3K_2^{-p}\left(T^p+|w|_{p,0,T}^p\right)+2\right).\end{aligned}$$

Inserting the definition of K_1 and K_2, we observe that the right-hand side can be bounded by $\delta^{2-p} C e^{C|w|_{p,0,T}^p}$ for some constant C independent of Π^{E}, u, and w. □

The Euler schemes for the Young SDEs in the first n_1 lines of $\mathcal{X}^u$ and $\mathcal{Y}^u$ can be analyzed by the same techniques, which we illustrated in the proof of Theorem 4.4. Due to the boundedness of the coefficient functions, the Young SDEs in $\mathcal{X}^u$ are actually much simpler and do not require the Gronwall lemma on Euler partitions in the form of Lemma 4.5.

The Euler schemes for the component processes in the last n_2 lines of $\mathcal{X}^u$ and $\mathcal{Y}^u$ are driven by a Brownian motion. They can be analyzed by standard methods (see, e.g., [31]). There is a little extra work, because the coefficients in the Euler scheme of these equations depend on the Euler approximations of the Young SDEs. In particular, the error estimates for the Young SDEs enter the error analysis of the SDEs driven by the Brownian motion. For this reason, the strong convergence rates deteriorate from $1/2$ to $\min(2-p, 1/2)$.

We finally turn to the Euler approximation (20) to the adjoint equation. For fixed $t_k \in \Pi^{\mathrm{E}}$ and recalling (18), we define

$$\Phi_{t_{i+1}}^{t_k,u,n} = \Phi_{t_i}^{t_k,u,n} + \eta_{t_i,t_{i+1}}^{n,u} \Phi_{t_i}^{t_k,u,n},\ i > k, \quad \Phi_{t_k}^{t_k,u,n} = I_{n_1+n_2}.$$

A direct computation shows

$$\Lambda_{t_i}^{n,u} = \sum_{\mu;\, T_\mu \geq t_i} E[g_\mu(\mathcal{X}_{T_\mu}^{n,u})] g_\mu'(\mathcal{X}_{T_\mu}^{n,u}) \Phi_{T_\mu}^{t_i,u,n}, \tag{61}$$

while, by the definition of Λ^u in (13),

$$\Lambda_t^u = \sum_{\mu;\, T_\mu \geq t} E[g_\mu(\mathcal{X}_{T_\mu}^{u})] g_\mu'(\mathcal{X}_{T_\mu}^{u}) \Phi_{T_\mu}^{t,u}. \tag{62}$$

Now, fix $s \in [0,T]$ and let t_k be the smallest grid point bigger or equal to s. Then, $\Phi^{t_k,u,n}$ can be considered as Euler scheme for $\Phi^{s,u}$. The same convergence rates as for the Euler approximation to $\mathcal{Y}^u$ can be derived by the same techniques with the constants being independent of s. In view of (61)–(62), these rates carry over to the approximation of Λ^u by $\Lambda^{n,u}$.

5 Case study and numerical experiments

5.1 Monte-Carlo implementation

As demonstrated in Subsection 2.3, we can approximate the cost function and its gradient with respect to the parameter via Euler schemes. Since for the calculation of the discretized cost function and the discretized gradient we need to evaluate expected values, we apply the Monte-Carlo method

to come up with an implementable scheme. A comprehensive introduction to Monte-Carlo methods is given by [23]. Using A independent copies $(\mathcal{X}^{n,u,a}, \mathcal{Y}^{n,u,a})_{a=1,\dots,A}$ of the Euler schemes in (15), (16) restricted to the discrete-time grid Π^E, we can approximate the discretized cost function and the discretized gradient by the Monte-Carlo estimators

$$\begin{aligned} J^{n,A}(u) &= \frac{1}{2}\sum_{\mu=1}^{M}\left(\frac{1}{A}\sum_{a=1}^{A} g_\mu(\mathcal{X}^{n,u,a}_{T_\mu})\right)^2 \\ (\nabla J)^{n,A}(u) &= \sum_{\mu=1}^{M}\left(\frac{1}{A}\sum_{a=1}^{A} g_\mu(\mathcal{X}^{n,u,a}_{T_\mu})\right)\left(\frac{1}{A}\sum_{a=1}^{A} g'_\mu(\mathcal{X}^{n,u,a}_{T_\mu})\mathcal{Y}^{n,u,a}_{T_\mu}\right). \end{aligned} \tag{63}$$

Using the central limit theorem, it is well established that the corresponding approximation error behaves asymptotically like $\mathcal{O}(A^{-\frac{1}{2}})$, see [23].

Accordingly, we will replace the expectations by the sample mean in the Euler scheme for the adjoint equation, leading to the backward recursion

$$\Lambda^{n,u,a}_{t_i} = \Lambda^{n,u,a}_{t_{i+1}} + \Lambda^{n,u,a}_{t_{i+1}}\eta^{n,u,a}_{t_i,t_{i+1}} + \sum_{\mu;\, T_\mu = t_i}\left(\frac{1}{A}\sum_{\alpha=1}^{A} g_\mu(\mathcal{X}^{n,u,\alpha}_{t_i})\right) g'_\mu(\mathcal{X}^{n,u,a}_{t_i}),$$

initialized at

$$\Lambda^{n,u,a}_{t_n} = \sum_{\mu;\, T_\mu = T}\left(\frac{1}{A}\sum_{\alpha=1}^{A} g_\mu(\mathcal{X}^{n,u,\alpha}_{T})\right) g'_\mu(\mathcal{X}^{n,u,a}_{T}),$$

where

$$\begin{aligned} \eta^{n,u,a}_{t_i,t_{i+1}} &= \begin{pmatrix} b_\xi(t_i,\xi^{n,u,a}_{t_i},u) & 0 \\ \hat{b}_\xi(t_i,\xi^{n,u,a}_{t_i},x^{n,u,a}_{t_i},u) & \hat{b}_x(t_i,\xi^{n,u,a}_{t_i},x^{n,u,a}_{t_i},u)\end{pmatrix}(t_{i+1}-t_i) \\ &\quad + \sum_{j=1}^{m_1}\begin{pmatrix} \sigma^j_\xi(t_i,\xi^{n,u,a}_{t_i},u) & 0 \\ 0 & 0\end{pmatrix}(w^{j,a}_{t_{i+1}} - w^{j,a}_{t_i}) \\ &\quad + \sum_{j=1}^{m_2}\begin{pmatrix} 0 & 0 \\ \hat{\sigma}^j_\xi(t_i,\xi^{n,u,a}_{t_i},x^{n,u,a}_{t_i},u) & \hat{\sigma}^j_x(t_i,\xi^{n,u,a}_{t_i},x^{n,u,a}_{t_i},u)\end{pmatrix}(B^{j,a}_{t_{i+1}} - B^{j,a}_{t_i}). \end{aligned}$$

Note that the realizations $(\Lambda^{n,u,a})_{a=1,\dots,A}$ are not independent due to the presence of the sample means.

The following proposition is the analogue of Theorem 2.11 in the Monte-Carlo setup. Its proof remains unchanged except replacing the mean by the sample mean in the last step of the proof.

Proposition 5.1. *For every $u \in \mathcal{U}$, we have*

$$(\nabla J)^{n,A}(u) = \frac{1}{A}\sum_{a=1}^{A}\left(\Lambda^{n,u,a}_{t_0} D\mathcal{X}^u_0 + \sum_{i=0}^{n-1}\Lambda^{n,u,a}_{t_{i+1}}\eta^{n,u,a}_{t_i,t_{i+1}}\right). \tag{64}$$

This proposition states, that we get exactly the same result when realizing the Monte-Carlo paths and calculating the gradient via the discrete sensitivity equation (63) or the adjoint method (64). Since we are now able to approximate the value of the cost function, as well as its gradient with respect to the parameter (using two different methods), we can apply smooth gradient-based optimization algorithms to find the minimum of the cost function. As already mentioned, the adjoint method has the advantage that instead of $(n_1 + n_2) \cdot d$ forward solves of the recursion for $\mathcal{Y}^{n,u}$, we only have to perform $n_1 + n_2$ backward solves. Hence, the computational cost of a gradient evaluation does not depend on the number of parameters in the adjoint approach. In particular, in the case of time-dependent parameters this reduces the numerical effort substantially in comparison to the sensitivity method.

5.2 Case study: Calibrating a fractional Heston-type model

In this subsection, we illustrate how the results from the previous sections can be applied to calibrate a financial model with a volatility process driven by process of finite p-variation for $p \in (1, 2)$. There are several models which incorporate the long memory phenomenon of volatility, by using a fractional Brownian motion with Hurst parameter $H \in (0.5, 1)$ as driving process for the volatility, see, e.g., [8, 7, 3, 36, 33]. We choose a fractional version of the Cox-Ingersoll-Ross (CIR) process given by

$$v_t = v_0 + \int_0^t \kappa(\theta - v_r)\, dr + \int_0^t \zeta \sqrt{v_r}\, dB_r^H,$$

where B^H is a fractional Brownian motion with Hurst parameter $H \in (0.5, 1)$. It has been shown in [33] and [36] that this equation has a unique positive solution, when the integral $\int_0^t \zeta \sqrt{v_r}\, dB_r^H$ is interpreted as a pathwise Young integral. Furthermore in [33], the authors show that the process v_t is mean reverting to the parameter θ, hence the parameters can be interpreted similarly to the standard CIR model. Another feature, which we want to incorporate, is the correlation between the volatility process and the asset price process. To this end, fix $T > 0$ and let $(\Omega, \mathcal{F}, P)$ be a probability space carrying a two-sided Brownian motion $(B_t^1)_{t\in\mathbb{R}}$ and a Brownian motion $(B_t^2)_{t\in[0,T]}$ independent of B^1. Then, a fractional Brownian motion B_t^H with Hurst parameter $H \in (0.5, 1)$, can be constructed via the following integral transformation of B^1, see [34]:

$$B_t^H = C_H \left(\int_0^t (t-u)^{H-\frac{1}{2}}\, dB_u^1 + \int_{-\infty}^0 (t-u)^{H-\frac{1}{2}} - (-u)^{H-\frac{1}{2}}\, dB_u^1 \right),$$

where

$$C_H = \sqrt{\frac{2H\Gamma(\frac{3}{2} - H)}{\Gamma(H + \frac{1}{2})\Gamma(2 - 2H)}}.$$

By defining $B_t = \rho B_t^1 + \sqrt{1-\rho^2} B_t^2$, we obtain a standard Brownian motion B_t, which is correlated with B^1 via $Corr(B_t, B_t^1) = \rho$ for all $t \in [0,T]$. We write $\mathbb{F}$ for the augmentation of the filtration generated by $(B_t^H, B_t)_{t\in[0,T]}$. Note that ρ is not the correlation between B^H and the Brownian motion driving the asset price process B, but between B and the Brownian motion B_1 from which B^H has been constructed. This way we generate the desired correlation between the volatility process v and the asset price S in the following model, in a similar way as in [36]:

$$\begin{aligned} v_t &= v_0 + \int_0^t \kappa(\theta - v_s)\, ds + \int_0^t \zeta\sqrt{v_s}\, dB_s^H \\ S_t &= S_0 + \int_0^t (r-d) S_s\, ds + \int_0^t \sqrt{v_s} S_s\, d(\rho B_s^1 + \sqrt{1-\rho^2} B_s^2). \end{aligned} \tag{65}$$

Here the spot price S_0, the riskless rate r and the dividend yield d are given. We assume the market we are trading in, only consist of the asset S and a riskless bond e^{-rt} for $t \in [0,T]$.

We aim at calibrating the model with respect to the parameters $u = (v_0, \kappa, \theta, \zeta, \rho)$ to a set of market observed European call option prices. In order to apply the results derived in the theoretical part, we have to smoothen the coefficient functions of the SDE system (65). We basically follow the approach in [29], when using a piecewise polynomial error function π_1 to smoothen out the positive part $\pi(x) := (x)_+ := \max(x, 0)$ at 0 and a similar construction for the function π_2 which smoothens out the square root function at 0. These functions are given by

$$\pi_1(x) = \begin{cases} 0, & x < -\varepsilon_1 \\ -\frac{1}{16(\varepsilon_1)^3} x^4 + \frac{3}{8\varepsilon_1} x^2 + \frac{1}{2} x + \frac{3\varepsilon_1}{16}, & -\varepsilon_1 \le x \le \varepsilon_1 \\ x, & x > \varepsilon_1 \end{cases}$$

for $x \in \mathbb{R}$ and an error parameter $\varepsilon_1 > 0$ which we choose to be 0.01 for all calculations. The second function is given by

$$\pi_2(x) = \begin{cases} 0, & x < -\varepsilon_2 \\ -\frac{1}{256\varepsilon_2^{6,5}}(-15x^7 + 7\varepsilon_2 x^6 + 65\varepsilon_2^2 x^5 - 33\varepsilon_2^3 x^4 & \\ -117\varepsilon_2^4 x^3 + 77\varepsilon_2^5 x^2 + 195\varepsilon_2^6 x + 77\varepsilon_2^7), & -\varepsilon_2 \le x \le \varepsilon_2 \\ \sqrt{x}, & x > \varepsilon_2 \end{cases}$$

for $x \in \mathbb{R}$ and an error parameter $\varepsilon_2 > 0$ which we choose to be 0.001 for all calculations. To achieve boundedness of these two functions, we theoretically compose them with a smooth truncation function. Choosing the truncation level sufficiently large, this truncation can be ignored in practice, since v_t is mean reverting and for a moderate time horizon we do not expect the log-price of the asset in our model to explode along typical realizations.

This reasoning is justified by our numerical findings. The dynamics of the adjusted fractional Heston-type model are, then, given by

$$\begin{aligned} v_t &= v_0 + \int_0^t \kappa(\theta - \pi_1(v_s))\, ds + \int_0^t \zeta \pi_2(v_s)\, dB_s^H \\ S_t &= S_0 + \int_0^t (r-d) S_s\, ds + \int_0^t \pi_2(v_s) S_s\, d(\rho B_s^1 + \sqrt{1-\rho^2} B_s^2) \end{aligned} \tag{66}$$

and after a log-transformation $\hat{S}_t = \log(S_t)$ in the asset equation, this yields

$$\begin{aligned} v_t &= v_0 + \int_0^t \kappa(\theta - \pi_1(v_s))\, ds + \int_0^t \zeta \pi_2(v_s)\, dB_s^H \\ \hat{S}_t &= \hat{S}_0 + \int_0^t (r-d) - \frac{1}{2}\pi_2(v_s)^2\, ds + \int_0^t \pi_2(v_s)\, d(\rho B_s^1 + \sqrt{1-\rho^2} B_s^2). \end{aligned}$$

Note that under these adjustments (the truncated version of) $\pi_2(v_r)$ is bounded and hence the SDE (66) has the explicit solution

$$S_t = S_0 e^{\left((r-d)t - \frac{1}{2}\int_0^t \pi_2(v_s)^2\, ds + \int_0^t \pi_2(v_s)\, dB_s\right)}.$$

The dividend-adjusted discounted price process $e^{-(r-d)t} S_t$ is then a martingale with respect to P and the price for a call option with maturity T_μ and strike K_μ at time 0 in this model (taking P as pricing measure) is given by

$$e^{-rT_\mu} \mathrm{E}\left[\left(e^{\hat{S}^u_{T_\mu}} - K_\mu\right)_+\right]$$

by the risk-neutral pricing formula. We approximate this value by

$$C_\mu^{mod}(u) = e^{-rT_\mu} \mathrm{E}\left[\pi_1\left(e^{\hat{S}^u_{T_\mu}} - K_\mu\right)\right]$$

and the cost function, thus, translates to

$$J(u) = \frac{1}{2}\sum_{\mu=1}^M \mathrm{E}\left[g_\mu\begin{pmatrix} v^u_{T_\mu} \\ \hat{S}^u_{T_\mu}\end{pmatrix}\right]^2 = \frac{1}{2}\sum_{\mu=1}^M \left(C_\mu^{mod}(u) - C_\mu^{obs}\right)^2,$$

where $g_\mu(x_1, x_2) = e^{-rT_\mu}\pi_1(e^{x_2} - K_\mu) - C_\mu^{obs}$ and C_μ^{obs} is the observed market price for a call option with maturity T_μ struck at K_μ. As parameter set for the calibration problem, we choose

$$\begin{aligned} \mathcal{U} := \{(v_0, \kappa, \theta, \zeta, \rho) \in \mathbb{R}^5 |\, & v_0 \in (0.0001, 1),\ \kappa \in (0.0001, 2),\ \theta \in (0.0001, 2), \\ & \zeta \in (0.0001, 4)\, \rho \in (-0.99, 0.99)\} \end{aligned}$$

Note that a fractional Brownian motion with Hurst parameter $H > \frac{1}{2}$ has Hölder continuous paths with Hölder index H' for every $1/2 < H' < H$,

H	v_0		κ		θ		ζ		ρ	
	μ	Sd	μ	Sd	μ	Sd	μ	Sd	μ	Sd
0.5	0.072	0.0030	0.809	0.0990	0.071	0.0022	0.438	0.0428	-0.657	0.0208
0.55	0.070	0.0016	0.838	0.0662	0.054	0.0008	0.413	0.0365	-0.657	0.0115
0.6	0.070	0.0018	0.971	0.1197	0.045	0.0013	0.424	0.0400	-0.667	0.0156
0.65	0.070	0.0017	1.030	0.0733	0.042	0.0016	0.409	0.0287	-0.690	0.0155
0.7	0.069	0.0018	1.055	0.0669	0.043	0.0018	0.383	0.0221	-0.724	0.0189
0.75	0.069	0.0016	1.085	0.0609	0.043	0.0012	0.362	0.0182	-0.759	0.0152
0.8	0.067	0.0011	1.163	0.0920	0.045	0.0012	0.360	0.0164	-0.822	0.0235
0.85	0.067	0.0014	1.327	0.1149	0.047	0.0010	0.369	0.0208	-0.925	0.0233
0.9	0.065	0.0012	1.199	0.0651	0.046	0.0004	0.354	0.0141	-0.990	0.0001
0.95	0.066	0.0011	1.076	0.0537	0.039	0.0005	0.368	0.0168	-0.990	0.0000

Table 1: Calibrated parameters for different values of $H \in (\frac{1}{2}, 1)$.

see, e.g., [41], p.274. Hence, in view of Remark 2.2, it satisfies assumption (W) for every $p \in (1/H, 2)$. Moreover, the Hölder assumption in Remark 2.9 can be verified for every Hölder index $H' < H$. Finally, the assumptions $(H_1), (H_2), (H_3), (B_1), (B_2), (B_3), (E_1), (E_2), (G)$ are also fulfilled (taking the choice of $\mathcal{U}$ into account).

We calibrate the model to market prices for call options on the EUROSTOXX 50 as of October 7th, 2003. The data set is reported in [45] and consists of the prices for 144 call options (in total) with six different maturities 0.0361, 0.2000, 1.1944, 2.1916, 4.2056, 5.1639 (in years). We exclude the call option data for the strikes 2499.76 and 4990.91 in order to remove static arbitrage opportunities from the data set. After this modification of the data set, it still consists of 136 call option prices. Following [45], we set $S_0 = 2461.44$, $r = 0.03$, and $d = 0$.

For the numerical calibration, we minimize the Monte-Carlo estimate $J^{n,A}$ of the discretized cost functional using the Matlab fmincon function with the trust region reflective algorithm and a function tolerance of 10^{-6} feeding in the gradient approximation $(\nabla J)^{n,A}$ which is computed via the adjoint representation in Proposition 5.1. We initialize the parameter values at

$$v_0 = 0.1,\ \kappa = 1,\ \theta = 0.05\ , \zeta = 0.3,\ \rho = -0.7$$

and first run the calibration with $10,000$ Monte Carlo samples and a time grid which divides the time between each of the neighboring maturities into 40 subintervals (leading to $n = 240$ and a mesh size of 0.0504). The resulting parameters are stored and taken as input for a second optimization stage with $100,000$ Monte Carlo samples and a time-grid consisting of 480 subintervals, halving each of the 240 intervals of the first stage. The empirical mean and the empirical standard deviation of the parameters found in the second stage over 25 independent repetitions of the algorithm are reported in Table 5.2 for various choices of the Hurst parameter $H \in \{0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95\}$.

H	AvgErr (sample mean)	AvgErr (emp. standard dev.)	Avg runtime in sec
0.5	$8.788 \cdot 10^{-4}$	$7.229 \cdot 10^{-6}$	973.30
0.55	$6.913 \cdot 10^{-4}$	$7.014 \cdot 10^{-6}$	654.87
0.6	$6.890 \cdot 10^{-4}$	$7.628 \cdot 10^{-6}$	553.73
0.65	$6.577 \cdot 10^{-4}$	$7.598 \cdot 10^{-6}$	453.74
0.7	$6.987 \cdot 10^{-4}$	$7.501 \cdot 10^{-6}$	390.58
0.75	$8.059 \cdot 10^{-4}$	$7.659 \cdot 10^{-6}$	374.23
0.8	$8.821 \cdot 10^{-4}$	$6.467 \cdot 10^{-6}$	422.38
0.85	$9.869 \cdot 10^{-4}$	$6.725 \cdot 10^{-6}$	538.93
0.9	$2.156 \cdot 10^{-3}$	$9.565 \cdot 10^{-6}$	541.30
0.95	$3.276 \cdot 10^{-3}$	$7.675 \cdot 10^{-6}$	977.54

Table 2: Summary statistics for the calibration results.

In order to evaluate the fit of the calibration procedure, we simulate, for each choice of the Hurst parameter H, a new independent Monte Carlo sample of size 100.000 and compute the Monte Carlo estimator $\hat{C}_\mu^{mod}(u^*)$ for the model price $C_\mu^{mod}(u^*)$ along the finer time grid, where u^* is the optimal parameter vector found in the calibration routine (see Table 5.2). Table 5.2 contains some summary statistics for the estimated average error $\text{avgErr} = \frac{1}{136 S_0} \sum_{\mu=1}^{136} |\hat{C}_\mu^{mod}(u^*) - C_\mu^{obs}|$ over the 136 option prices. Precisely, we report the sample mean and the empirical standard deviation of avgErr over 100 independent repetitions as well as the run time for the calibration step (in dependence of the Hurst parameter). Our results suggest that the best fit to the data can be achieved by adding a moderate long-range dependence into the stochastic volatility process corresponding to a Hurst parameter of about $H = 0.65$.

The option price function of the calibrated model with Hurst parameter $H = 0.65$ is plotted in Figure 2 for the six maturities, for which price data is available (marked by 'x'). The figure illustrates the excellent fit of the calibrated model across all maturities and strikes.

5.3 Additional numerical experiments

We finally perform some numerical experiments in order to illustrate the rates of convergence derived in Theorem 2.8 and Remark 2.9 and the computational benefit from simulating the gradient of the cost functional via the adjoint equation Λ^u as compared to the sensitivity equation $\mathcal{Y}^u$.

Note that, in the fractional Brownian motion case, by (23) and Remark 2.9, for every $H' < H$

$$\mathrm{E}[|(\nabla J)^{n,A}(u) - (\nabla J)(u)|] = \mathcal{O}\left(A^{-\frac{1}{2}} + |\Pi^E|^{(2H'-1)\wedge\frac{1}{2}}\right).$$

In the experiment below, we fix the sample size as $A = 100,000$, but refine the time partition by decomposing the time between two maturities into 2^i subintervals for $i = 4, \ldots, 9$, leading to a partition into $n_i = 6 \cdot 2^i$ subintervals

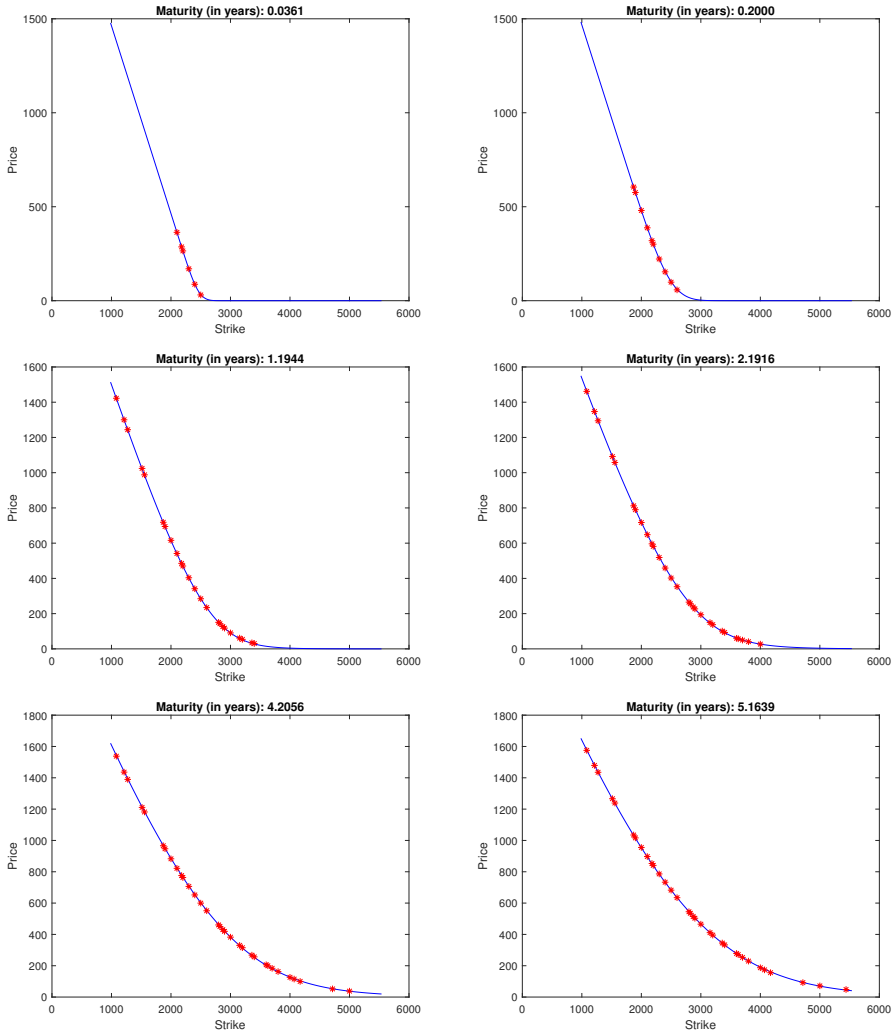

Figure 2: Call price function of the calibrated model ($H = 0.65$) and observed option prices (marked by 'x').

in total. This corresponds to a mesh size of $0.126 \cdot 2^{-(i-4)}$. Note that, by the triangle inequality,

$$\mathrm{E}\left[\left|(\nabla J)^{n_i,A}(u) - (\nabla J)^{n_{i-1},A}(u)\right|\right] = \mathcal{O}\left(A^{-\frac{1}{2}} + 2^{-i((2H'-1)\wedge\frac{1}{2})}\right).$$

We choose $H = 0.8$ leading to $(2H'-1)\wedge\frac{1}{2} = \frac{1}{2}$ for sufficiently large $H' < H$ and fix

$$u = (v_0, \kappa, \theta, \zeta, \rho)^\top = (0.016, 1, 0.02, 0.3, -0.7)^\top.$$

In this setting, we sample 20 independent copies

$$((\nabla J)^{n_4,A,j}(u), \cdots (\nabla J)^{n_9,A,j}(u))_{j=1,\ldots,20}$$

of $((\nabla J)^{n_4,A}(u), \cdots (\nabla J)^{n_9,A}(u))$ and consider the error

$$Err_i = \frac{1}{20}\sum_{j=1}^{20}\left|(\nabla J)^{n_i,A,j}(u) - (\nabla J)^{n_{i-1},A,j}(u)\right|$$

for $i = 5, \ldots, 9$. The theoretical considerations above suggest that Err_i decays as $2^{-i/2}$ provided the time-discretization error dominates the Monte Carlo error. This is confirmed by the log-log plot of the mesh size $0.126 \cdot 2^{-(i-5)}$ of the $(i-1)$th partition versus Err_i in Figure 3, where the dashed reference line exhibits a slope of 0.5.

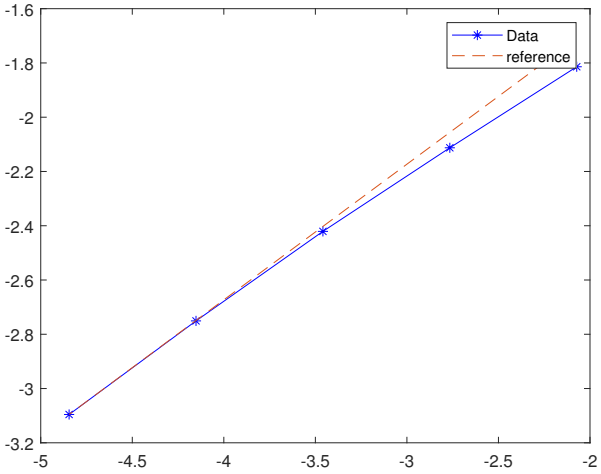

Figure 3: Log-log plot of the mesh size $0.126 \cdot 2^{-(i-5)}$ against Err_i for $i = 5, \ldots, 9$ and $H = 0.8$.

We finally compare the run time for computing $(\nabla J)^{n,A}$ based on the representation (63), for which we employ the Euler approximation of the

Number of parameters	5	9	13	17	21	25
RT adjoint	19.0	18.0	18.0	18.0	18.0	18.0
RT sensitivity	22.0	26.0	30.0	35.0	39.0	43.0

Table 3: Runtime (RT; in sec) for the computation of the gradient of the cost function with the two different methods.

sensitivity equation, and based on the discretizaton of the adjoint equation, see (64). To this end, we replace the constants κ, θ, ζ, and ρ by piecewise constant functions in time, where each of these functions can take I values. Hence, the total number of parameters, then, becomes $4I+1$. The run times reported in Table 3 correspond to a single evaluation of the gradient $(\nabla J)^{n,A}$ based on a Monte Carlo sample of size $A = 100,000$ and a time-discretization into $n = 480$ subintervals. In line with the theoretical considerations, the run time for computing the gradient via the adjoint method does not depend on the number of model parameters, while the computational time for simulating the sensitivity equation linearly increases with the number of parameters. The reported run times, thus, demonstrate the computational benefits for applying the adjoint method based on the new type of anticipating backward SDE (14) for model calibration, in particular when the parameter vector is high-dimensional.

References

[1] A. Ambrosetti and G. Prodi. *A primer of nonlinear analysis*, volume 34 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1995.

[2] C. Bayer, P. Friz, and J. Gatheral. Pricing under rough volatility. *Quant. Finance*, 16(6):887–904, 2016.

[3] V. Bezborodov, L. Di Persio, and Y. Mishura. Option pricing with fractional stochastic volatility and discontinuous payoff function of polynomial growth. *Methodol. Comput. Appl. Probab.*, 21(1):331–366, 2019.

[4] T. Cass, C. Litterer, and T. Lyons. Integrability and tail estimates for Gaussian rough differential equations. *Ann. Probab.*, 41(4):3026–3050, 2013.

[5] M. Chesney and L. Scott. Pricing European currency options: A comparison of the modified Black-Scholes model and a random variance model. *The Journal of Financial and Quantitative Analysis*, 24(3):267–284, 1989.

[6] A. Chronopoulou and S. Tindel. On inference for fractional differential equations. *Stat. Inference Stoch. Process.*, 16(1):29–61, 2013.

[7] A. Chronopoulou and F. G. Viens. Estimation and pricing under long-memory stochastic volatility. *Ann. Finance*, 8(2-3):379–403, 2012.

[8] F. Comte and E. Renault. Long memory in continuous-time stochastic volatility models. *Math. Finance*, 8(4):291–323, 1998.

[9] N. D. Cong, L. H. Duc, and P. T. Hong. Nonautonomous Young differential equations revisited. *J. Dynam. Differential Equations*, 30(4):1921–1943, 2018.

[10] N. D. Cong, L. H. Duc, and P. T. Hong. Lyapunov spectrum of nonautonomous linear Young differential equations. *J. Dynam. Differential Equations*, 32(4):1749–1777, 2020.

[11] N. D. Cong, L. H. Duc, and P. T. Hong. Numerical attractors via discrete rough paths. *J. Dynam. Differential Equations*, online first, 2023.

[12] R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quant. Finance*, 1:223–236, 2001.

[13] R. M. Dudley and R. Norvaiša. *Differentiability of six operators on nonsmooth functions and p-variation*, volume 1703 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1999.

[14] R. M. Dudley and R. Norvaiša. *Concrete functional calculus*. Springer Monographs in Mathematics. Springer, New York, 2011.

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

[16] K. French, G. Schwert, and R. Stambaugh. Expected stock returns and volatility. *Journal of Financial Economics*, 19(1):3–29, 1987.

[17] P. K. Friz and N. B. Victoir. *Multidimensional stochastic processes as rough paths*, volume 120 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2010.

[18] M. Fukasawa. Asymptotic analysis for stochastic volatility: martingale expansion. *Finance Stoch.*, 15(4):635–654, 2011.

[19] M. Fukasawa. Short-time at-the-money skew and rough fractional volatility. *Quant. Finance*, 17(2):189–198, 2017.

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

[21] M. Giles and P. Glasserman. Smoking adjoints: fast evaluation of Greeks in Monte Carlo calculations. *Risk*, 19:88–92, January 2006.

[22] M. B. Giles and N. A. Pierce. An introduction to the adjoint approach to design. *Flow, Turbulence and Combustion*, 65:393–415, 2000.

[23] P. Glasserman. *Monte Carlo methods in financial engineering*, volume 53 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 2004.

[24] Y. Han, Y. Hu, and J. Song. Maximum principle for general controlled systems driven by fractional Brownian motions. *Appl. Math. Optim.*, 67(2):279–322, 2013.

[25] S. Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6:327–343, 1993.

[26] Y. Hu, Y. Liu, and D. Nualart. Rate of convergence and asymptotic error distribution of Euler approximation schemes for fractional diffusions. *Ann. Appl. Probab.*, 26(2):1147–1207, 2016.

[27] J. Hull and A. White. The pricing of options on assets with stochastic volatilities. *The Journal of Finance*, 42(2):281–300, 1987.

[28] N. C. Jain and D. Monrad. Gaussian measures in B_p. *Ann. Probab.*, 11(1):46–57, 1983.

[29] C. Kaebe, J. H. Maruhn, and E. W. Sachs. Adjoint-based Monte Carlo calibration of financial market models. *Finance Stoch.*, 13(3):351–379, 2009.

[30] I. Karatzas and S. E. Shreve. *Brownian motion and stochastic calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1991.

[31] P. E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*, volume 23 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1992.

[32] A. Lejay. Controlled differential equations as Young integrals: a simple approach. *J. Differential Equations*, 249(8):1777–1798, 2010.

[33] E. Lépinette and F. Mehrdoust. A fractional version of the Heston model with Hurst parameter $H \in (1/2, 1)$. *available at: SSRN*, 2016.

[34] B. B. Mandelbrot and J. W. Van Ness. Fractional Brownian motions, fractional noises and applications. *SIAM Rev.*, 10:422–437, 1968.

[35] Y. Mishura and G. Shevchenko. The rate of convergence for Euler approximations of solutions of stochastic differential equations driven by fractional Brownian motion. *Stochastics*, 80(5):489–511, 2008.

[36] Y. Mishura and A. Yurchenko-Tytarenko. Approximating expected value of an option with non-Lipschitz payoff in fractional Heston-type model. *Int. J. Theor. Appl. Finance*, 23(5):2050031, 36, 2020.

[37] A. Neuenkirch. Optimal approximation of SDE's with additive fractional noise. *J. Complexity*, 22(4):459–474, 2006.

[38] A. Neuenkirch and I. Nourdin. Exact rate of convergence of some approximation schemes associated to SDEs driven by a fractional Brownian motion. *J. Theoret. Probab.*, 20(4):871–899, 2007.

[39] N. Nikolova, R. Safian, E. Soliman, M. Bakr, and J. Bandler. Accelerated gradient based optimization using adjoint sensitivities. *IEEE Transactions on Antennas and Propagation*, 52:2147–2157, January 2004.

[40] I. Nourdin. Schémas d'approximation associés à une équation différentielle dirigée par une fonction höldérienne; cas du mouvement brownien fractionnaire. *C. R. Math. Acad. Sci. Paris*, 340(8):611–614, 2005.

[41] D. Nualart. *The Malliavin calculus and related topics*. Probability and its Applications (New York). Springer-Verlag, Berlin, second edition, 2006.

[42] P. E. Protter. *Stochastic integration and differential equations*, volume 21 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, second edition, 2004.

[43] F. Russo and P. Vallois. Forward, backward and symmetric stochastic integration. *Probab. Theory Related Fields*, 97(3):403–421, 1993.

[44] F. Russo and P. Vallois. Elements of stochastic calculus via regularization. In *Séminaire de Probabilités XL*, volume 1899 of *Lecture Notes in Math.*, pages 147–185. Springer, Berlin, 2007.

[45] W. Schoutens, E. Simons, and J. Tistaert. A perfect calibration! now what? *Wilmott Magazine*, pages 66–78, March 2004.

[46] A. Shapiro. Monte Carlo sampling methods. In *Stochastic programming*, volume 10 of *Handbooks Oper. Res. Management Sci.*, pages 353–425. Elsevier Sci. B. V., Amsterdam, 2003.

[47] E. M. Stein and J. C. Stein. Stock price distributions with stochastic volatility: An analytic approach. *The Review of Financial Studies*, 4(4):727–752, 2015.

[48] M. Thiel. *Calibration of non-semimartingale models - an adjoint approach*. PhD Thesis. Saarland University, 2023.

[49] J. Yong and X. Y. Zhou. *Stochastic controls: Hamiltonian systems and HJB equations*, volume 43 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1999.

[50] L. C. Young. An inequality of the Hölder type, connected with Stieltjes integration. *Acta Math.*, 67(1):251–282, 1936.