Rough volatility models: Monte Carlo, Asymptotics and Deep Calibration

vorgelegt von

M. Sc.

Benjamin Marco Stemper

von der Fakultät II – Mathematik und Naturwissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
- Dr. rer. nat. -

genehmigte Dissertation

Promotionsausschuss:
Vorsitzende: Prof. Dr. Barbara Zwicknagl (TU Berlin)
Gutachter: Prof. Dr. Peter K. Friz (TU & WIAS Berlin)
Gutachter: Ass.-Prof. Dr. Stefano de Marco (École Polytechnique)

Tag der wissenschaftlichen Aussprache: 14. März 2019

Berlin 2019

Abstract

So-called *rough* stochastic volatility models constitute the latest advancement in option price modeling. In contrast to popular bivariate diffusion models such as Heston, here the driving noise of volatility is modeled by a fractional Brownian motion (fBM) with scaling in the *rough* regime of Hurst parameter $H < 0.5$. A major appeal of such models lies in their ability to parsimoniously recover key stylized facts of market IV surfaces such as the exploding power-law behaviour of the ATM volatility skew near zero, a crucial feature Markovian models fail to reproduce. On the flipside, as a consequence of fBM being neither a semimartingale nor a Markov process for $H \neq \frac{1}{2}$, most currently prevalent numerical pricing and calibration routines do not (easily) carry over to the rough setting. This thesis addresses this problem and contributes to the existing literature as follows.

In chapter 2, we sharpen the large deviations results of Forde-Zhang (2017) in a way that allows us to zoom-in around the money while maintaining full analytical tractability. More precisely, this amounts to proving higher order moderate deviations (MD) estimates, only recently introduced in the option pricing context. In particular, we derive small-time asymptotic formulae for log call prices and Black-Scholes implied volatility. This in turn allows us to push the applicability range of known ATM skew approximation formulae from CLT type log-moneyness deviations of order $t^{\frac{1}{2}}$ to the wider MD regime.

In chapter 3, we present a novel Monte Carlo (MC) pricing scheme for rough volatility models based on a Karhunen-Loève-style approximation of White Noise. This complements theoretical results by Bayer et al. (2017). Our numerical experiments confirm a theoretical strong rate of H for a central object of interest and indicate a weak rate of $2H$ for the option price.

In chapter 4, we introduce a novel model calibration routine for (rough) stochastic volatility models dubbed *deep calibration*. Standard model calibration routines rely on the repetitive evaluation of the map from model parameters to Black-Scholes implied volatility, rendering calibration of many (rough) stochastic volatility models prohibitively expensive since often the map can only be approximated by costly MC simulations. As a remedy, we propose to combine the popular Levenberg-Marquardt optimization algorithm with neural network (NN) regression, replacing expensive MC simulations with cheap forward runs of a NN trained to approximate the implied volatility map. Numerical experiments confirm the high accuracy and speed of our approach.

Zusammenfassung

So genannte *raue* stochastische Volatilitätsmodelle (rSV) stellen die jüngste Weiterentwicklung der Optionspreismodellierung dar. Im Gegensatz zu bivariaten Diffusionsmodellen wie Heston wird das treibende Rauschen der Volatilität hier durch eine *fraktionelle* Brownsche Bewegung (fBB) modelliert, welche im *rauen* Regime mit Hurst index $H < 0.5$ skaliert. Mit nur wenigen Parametern können nun Kerneigenschaften von empirischen impliziten Volatilitätsoberflächen, wie z.B. das explodierende Potenzgesetzverhalten der *ATM volatility skew* für $t \to 0$, abgebildet werden. Da die fBB für $H \neq \frac{1}{2}$ weder ein Semimartingal noch Markov ist, lassen sich gängige numerische Preis- und Kalibrierungsverfahren nicht (einfach) übertragen. Diese Arbeit setzt hier an und trägt auf folgende Weise zur Fachliteratur bei.

In Kapitel 2 entwickeln wir die *Large Deviations* Ergebnisse von Forde und Zhang (2017) dahingehend weiter, dass wir am Geld nah heranzoomen, gleichzeitig aber die analytische Berechenbarkeit beibehalten. Genauer gesagt beweisen wir Abschätzungen höherer Ordnung im Kontext der *Moderate Deviations (MD)*, welche erst kürzlich in die Optionsbewertung eingeführt wurden. Konkret leiten wir für kurze Maturitäten asymptotische Formeln für Log-Call-Preise und Implizite Volatilität her. Dies ermöglicht uns, den Anwendungsbereich bekannter ATM-Skew-Approximationsformeln von log-moneyness Abweichungen des CLT-Typs der Ordnung $t^{1/2}$ auf das breitere *MD* Regime zu erweitern.

In Kapitel 3 stellen wir ein neuartiges Monte Carlo Verfahren zur Optionspreisbewertung bei rSV Modellen vor, das auf einer Karhunen-Loève-ähnlichen Näherung des Weißen Rauschens basiert. Dies ergänzt die theoretischen Ergebnisse von Bayer et al. (2017). Unsere numerischen Experimente bestätigen eine theoretische, starke Rate von H für ein zentrales Objekt von Interesse und zeigen eine schwache Rate von 2H für den Optionspreis.

In Kapitel 4 entwickeln wir ein schnelles Kalibrierungsverfahren für rSV Modelle basierend auf Neuralen Netzwerken (NN). Die Funktion, welche Input Parametern eine Implizite Volatilität zuweist, lässt sich im Kontext von rSV Modellen meist nur durch teure MC Simulationen approximieren. Wir lassen ein NN diese Funktion erlernen, sodass deren Auswertung einem schnellen und günstigen Vorwärtslauf des NN entspricht. Durch eine Kombination mit dem beliebten Levenberg-Marquardt Algorithmus, der die wiederholte Evaluation dieser Funktion bedingt, lassen sich nun beliebige rSV Modelle schnell und genau kalibrieren, wie verschiedene numerische Experimente belegen.

Acknowledgment

First and foremost, I would like to to express my sincere gratitude to my supervisors Prof. Dr. Peter K. Friz and Dr. Christian Bayer. Throughout my PhD, they have been a continuous and reliable source of support and without their exceptional knowledge, experience and guidance the thesis in this form would clearly not have been possible. In particular, I wish to express my appreciation for them exposing me to cutting-edge research topics and for allowing me to work on them together with leading experts in the field, culminating in three joint papers included in this thesis. In particular, I would like to thank the co-authors of said papers Paul Gassiat, Archil Gulisashvili, Blanka Horvath and Jörg Martin for the excellent collaboration. Exchanging ideas and constantly learning from each other was an intellectually stimulating and enriching experience.

Second, I want to gratefully acknowledge financial support from Deutsche Forschungsgemeinschaft via DFG Research Grants BA5484/1 and FR2943/2 without which this work could not have been undertaken. Apart from financing my salary, this generous funding made it possible to attend many national and international conferences and workshops where I could present my research, receive feedback from recognized experts and hear first-hand about the latest developments in my field.

Third, I would like to express my gratitude to Prof. Dr. Stéfano de Marco who readily accepted to be a co-examiner of my thesis and Prof. Dr. Barbara Zwicknagl who thankfully agreed to chair my PhD committee.

Moreover, I thank my colleagues and friends at TU Berlin and WIAS Berlin for the pleasant atmosphere in the respective working groups. Special thanks go to David Beßlich, Khalil Chouk, Tom Klose and Moritz Voß for their companionship that made my stay at TU so enjoyable.

Last but not least, I am indebted to my family and Mira for the support and encouragement I received from them since the very beginning of my university studies. They were always there for me when I needed them and provided me with the sustaining power to overcome challenging times during my studies.

Contents

1 Introduction **1**

1.1 Short-time near-the-money skew in rough fractional volatility models (Chapter 2) 5

1.2 Monte Carlo pricing under rough stochastic volatility (Chapter 3) 6

1.3 Deep calibration of rough volatility models (Chapter 4) 8

2 Short-term near-the-money skew in rough fractional volatility models **11**

2.1 Exposition and assumptions 11

2.2 Main results 16

2.3 Simulation results 22

2.4 Proof of the energy expansion 24

2.4.1 Smoothness of the energy 27

2.4.1.1 The uncorrelated case 29

2.4.1.2 The general case 32

2.4.2 Energy expansion 34

2.4.2.1 Expansion of the minimizing configuration . . 35

2.4.2.2 Energy expansion in the general case 39

2.4.2.3 Energy expansion for the Riemann-Liouville kernel 40

2.5 Proof of the pricing formula 41

2.6 Proof of the moderate deviation expansions 44

2.7 Proof of the implied volatility expansion 50

2.8 Auxiliary lemmas 52

3 Monte Carlo pricing under rough stochastic volatility **55**

3.1 Exposition and assumptions 55

3.2 Monte Carlo pricing 57

3.3 Numerical results 60

3.3.1 Strong convergence of the renormalized integral 63

3.3.2 Weak convergence 64

4 Deep calibration of rough volatility models **67**

4.1 Background 67

4.1.1 Construction of a model IV surface 67

4.1.2 Regression with neural networks 70

4.2 Calibration of option pricing models 72

4.2.1 Deep calibration 74

4.3 Neural network training 76

4.3.1 Generation of synthetic labeled data 76

4.3.2 Backpropagation and hyper parameter optimization . . 78

4.3.2.1 Tricks of the trade 81

4.4 Numerical experiments 83

4.4.1 The Heston model 84

4.4.2 The rough Bergomi model 86

4.4.2.1 Bayesian parameter inference 88

Bibliography **93**

List of Figures

1.1 **SPX Market Implied Volatility surface on 15th February 2018.** Implied volatilities have been inverted from SPX Weekly European plain vanilla Call Mid prices and the interpolation is a (non-arbitrage-free) Delaunay triangulation. Axes denote log-moneyness $m = \log(K/S_0)$ for strike K and spot S_0, time to maturity T in years and market implied volatility $\sigma_{\text{iv}}(m, T)$. 2

2.1 **Illustration of the IV term structure of the Modified Rough Bergomi model in the Moderate deviations regime with time-varying log-strike $k_t = 0.4t^\beta$.** Depicted are the asymptotic formula (Eq. (2.2.9), dashed line) and an estimate based on $N = 10^8$ samples of a MC Cholesky Option Pricer (solid line) with 500 time steps. Model parameters are given by spot vol $\sigma_0 \approx 0.2557$, vvol $\eta = 0.2928$ and correlation parameter $\rho = -0.7571$. 25

3.1 **Numerical strong rates of convergence of the renormalized integral approximation $\widetilde{\mathscr{I}}^\varepsilon$.** Depicted are strong errors on a log-log-scale as $\varepsilon \to \varepsilon' = 2^{-8}$, obtained through $M = 10^5$ Monte Carlo samples with trapezoidal rule delta of $\Delta = 2^{-17}$. Solid lines visualize empirical rates of convergence obtained by least squares regression, dashed lines provide visual reference rates. Shaded colour bands show interpolated 95% confidence levels based on normality of MC estimator. . . 62

3.2 **Numerical weak rates of convergence of the approximative option price.** Depicted are weak errors on a log-log-scale as $\varepsilon \to \varepsilon' = 2^{-8}$, obtained through $M = 10^5$ MC samples with spot $S_0 = 1$, strike $K = 1$, correlation $\rho = -0.8$, spot vol $\sigma_0 = 0.2$, vvol $\eta = 2$ and trapezoidal rule delta $\Delta = 2^{-17}$. Dashed lines represent LS estimates for rate estimation, shaded colour bands show confidence levels based on normality of the MC estimator. 65

4.1 **Schematic of a fully-connected neural network (FCNN).** Depicted FCNN has a single *hidden layer* consisting of three neurons and may learn to represent a subset of general functions $f : \mathbb{R}^2 \to \mathbb{R}^3$. In the directed acyclic graph above, vertices denote nodes and directed edges describe the flow of information from one node to the next. If number of hidden layers higher than one (typically dozens or hundreds of layers), a neural network is considered *deep*. 70

4.2 **Schematic of IV ReLU FCNN**. Depiction of 4-layer ReLU FCNN used to learn IV maps. It consists of $2^{12} = 4096$ nodes at each hidden layer. Note the output layer is a linear layer with no activation function. Rectangles denote tensors and circles denote operations, *MatMul* is matrix multiplication. The number of parallel IV calculations is given by $n \in \mathbb{N}$. 81

4.3 **Accuracy of the IV map φ_{NN} as learned by the Heston ReLU FCNN.** 85

4.4 **Accuracy of the IV map φ_{NN} as learned by the rough Bergomi ReLU FCNN.** 87

4.5 **1d- and 2d-projections of 4d Bayesian posterior over rBergomi model parameters.** On diagonal, univariate histograms of model parameters with titles stating median ± the delta to 2.5% and 97.5% quantiles. Dashed vertical lines indicate those quantiles. Off-diagonal, 2d histograms of 2d projections of MCMC samples together with isocontours from 2d Gaussian KDE. 89

1 Introduction

Almost half a century after its publication, the option pricing model by Black and Scholes (1973) remains one of the most popular analytical frameworks for pricing and hedging European options in financial markets. A part of its success stems from the availability of explicit and hence instantaneously computable closed formulas for both theoretical option prices and option price sensitivities to input parameters (*Greeks*). This comes at the expense of assuming that *volatility* – the standard deviation of log returns of the underlying asset price – is deterministic and constant. Still, in financial practice, the Black-Scholes model is often considered a sophisticated transform between option prices and Black-Scholes (BS) *implied volatility (IV)* σ_{iv} where the latter is defined as the constant volatility input needed in the BS formula to match a given (market) price. It is a well-known fact that in empirical IV surfaces obtained by transforming market prices of European options to IVs, the implied volatilities vary across strikes and maturities, exhibiting well-known smiles and at-the-money (ATM) skews and thereby contradicting the flat surface predicted by Black-Scholes (Figure 1.1). In particular, Bayer, Friz, and Gatheral (2016) report empirical at-the-money volatility skews of the form

$$\left|\frac{\partial}{\partial m}\sigma_{\text{iv}}(m,T)\right| \sim T^{-0.4}, \quad T \to 0 \tag{1.0.1}$$

for log moneyness m and time to maturity T.

While plain vanilla European Call and Put options often show enough liquidity to be marked-to-market, pricing and hedging path-dependent options (so-called *Exotics*) necessitates an option pricing model that prices European options *consistently* with respect to observed market IVs across strikes and maturities. In other words, it should parsimoniously capture stylized facts of empirical IV surfaces. To address the shortcomings of Black-Scholes and incorporate the stochastic nature of volatility itself, popular bivariate diffusion models such as SABR (Hagan, Kumar, Lesniewski, & Woodward, 2002),

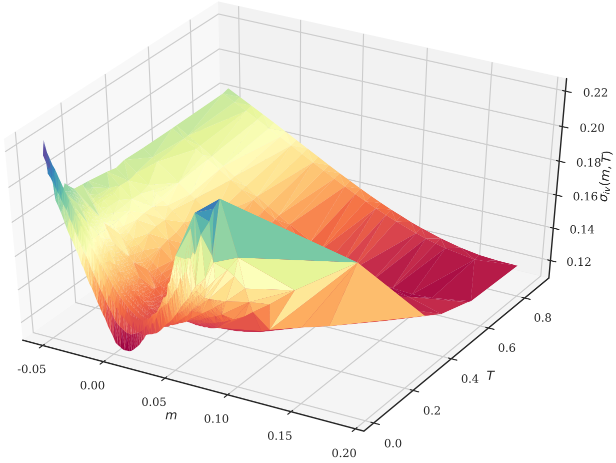


Figure 1.1: **SPX Market Implied Volatility surface on 15th February 2018.** Implied volatilities have been inverted from SPX Weekly European plain vanilla Call Mid prices and the interpolation is a (non-arbitrage-free) Delaunay triangulation. Axes denote log-moneyness $m = \log(K/S_0)$ for strike K and spot S_0, time to maturity T in years and market implied volatility $\sigma_{\text{iv}}(m, T)$.

Heston (1993) or Hull and White (1993) have been developed to capture *some* important stylized facts. However, according to Gatheral (2011), diffusive stochastic volatility models in general fail to recover the exploding power-law nature (1.0.1) of the volatility skew as time to maturity $T \to 0$ and instead predict a constant behaviour.

Since the seminal work of Gatheral, Jaisson, and Rosenbaum (2018) (a preprint had been available since 2014), the past four years have brought about a gradual shift in volatility modeling, leading away from classical diffusive stochastic volatility models towards so-called *rough* stochastic volatility models. Coined by Gatheral et al. (2018) and Bayer et al. (2016), the term essentially describes a family of continuous-path stochastic volatility models where the driving noise of volatility is modeled by a fractional Brownian motion[1] (fBM) (Mandelbrot & Van Ness, 1968) with scaling in the regime of Hurst index $H \in \left(0, \frac{1}{2}\right)$. The terminology *rough* here stems from the fact that the driving noise of the volatility has Hölder regularity H^-, hence smaller than that of Brownian motion. This modeling choice is empirically based on

[1] Instantaneous volatility is not a traded asset, so loss of semimartingality (when $H \neq 1/2$) does not imply arbitrage.

time series analysis conducted by Gatheral et al. (2018) who find that for all indices in the *Oxford-Man Institute's Realized Library*[2], the distribution of increments of log realized volatility is approximately Gaussian and log realized volatility exhibits Hölder regularity in the order of $H = 0.1$. Further analyses by Bennedsen, Lunde, and Pakkanen (2016) confirm the seminal discovery of *rough volatility* for over 5000 individual equities worldwide, suggesting that *rough volatility* is indeed a ubiquitous phenomenon. From an option-pricing point of view, a major appeal of such rough volatility models is that they allow to recover the characteristic exploding power law behaviour (1.0.1) of the ATM volatility skew for short maturities. A notable example is the *rough Bergomi* model introduced by Bayer et al. (2016) which can numerically be shown to exhibit said characteristic behaviour of the ATM volatility skew near zero. Once its only three model parameters have been calibrated to market data, it is able to approximate the empirical ATM skew quite closely (Bayer et al., 2016). At this point, we would also like to mention pioneering works by Alòs, León, and Vives (2007) and Fukasawa (2011). Quite some time before there existed any empirical evidence to consider *rough* volatility, they analytically derive a skew formula in rather restrictive rough volatility setting as an application example of a general framework.

The technical ability of rough volatility models to price European options consistently with respect to observed market IVs across moneyness and maturities makes them conceptually interesting for practitioners. On the other hand, for this new class of models to be of any practical use, efficient numerical pricing and model calibration schemes need to be developed. In fact, with some notable exceptions such as the Black and Scholes (1973) model, even for most diffusive stochastic volatility models no closed formulas for option prices are known. Indeed, drawing ideas and mathematical machinery from many areas of (applied) mathematics, a multitude of numerical methods have been developed over the years to approximate prices. These include but are not limited to:

(I) PDEs (Finite difference schemes etc.)

(II) Transforms (Fourier pricing etc.)

[2] For more information, check https://realized.oxford-man.ox.ac.uk.

(III) Large/Moderate Deviations (Small noise/time asymptotics etc.)

(IV) Stochastic simulation and related approaches (Monte Carlo (MC), Multi-Level Monte Carlo (MLMC), Quasi Monte Carlo (QMC) combined with variance reduction etc.)

Since fractional Brownian motion for $H \neq 1/2$ is neither a semimartingale nor a Markov process, currently prevalent option pricing techniques do not easily carry over to the rough setting. In particular, the non-Markovianity of fractional Brownian motion rules out any PDE related methods (I). Moreover, (off-the-shelf) Freidlin-Wentzell large deviation estimates are also not immediately applicable (II). Regarding transform methods (III), recall that the classical Heston (1993) model is amenable to Fast Fourier Pricing because the characteristic function of the log price is explicitly known. In a *rough* analogue of the Heston model dubbed *rough Heston* (Euch & Rosenbaum, 2018), it turns out that while the characteristic function of the log price is no longer explicitly known, it depends on the solution of a fractional analogue to the Riccati equations arising in the standard Heston case. These can be solved numerically, so Fourier Pricing is also applicable here. In summary, for a large class of rough stochastic volatility models, only pricing approaches in the direction of categories (III) & (IV) seem within reach at the moment and in fact, the majority of research conducted prior and also in parallel to this dissertation belongs to one of these two categories.

This thesis is dedicated to the development of novel pricing and calibration techniques in the context of rough volatility. For each chapter, we shall now provide an overview of existing works along the relevant research direction, followed by a quick summary of our approach and the results achieved in each piece of work.

1.1 Short-time near-the-money skew in rough fractional volatility models (Chapter 2)

This chapter is based on joint work with Peter K. Friz, Christian Bayer, Archil Gulisashvili and Blanka Horvath. It is a slightly revised version of an accepted manuscript of the article (Bayer, Friz, Gulisashvili, Horvath, & Stemper, 2018) published by Taylor & Francis in *Quantitative Finance* on 13 Nov 2018, available online: https://www.tandfonline.com/doi/full/10.1080/14697688.2018.1529420.

In this chapter, we rely on small-maturity approximations of option prices. This is a well-studied topic for which we mention (with no claim to completeness) a number of works, either based on large deviations or central limit type scaling regime, that inspired this work: (Alòs et al., 2007; Fukasawa, 2011; Deuschel, Friz, Jacquier, & Violante, 2014b, 2014a; Fukasawa, 2017), (Hagan et al., 2002; Berestycki, Busca, & Florent, 2004), also (Medvedev & Scaillet, 2003, 2007; Osajima, 2007; Guennoun, Jacquier, Roome, & Shi, 2014; Osajima, 2015; Mijatović & Tankov, 2016) and especially (Forde & Zhang, 2017). Rather recently, Friz, Gerhold, and Pinter (2018) introduced another regime called moderately-out-of-the-money (MOTM), which, in a sense, effectively navigates between the two regimes mentioned above, by rescaling the strike with respect to the time to maturity. This approach has various advantages. On the one hand, it reflects the market reality that as time to maturity approaches zero, strikes with acceptable bid-ask spreads tend to move closer to the money (see the original paper by Friz et al. (2018) for more details). On the other hand, it allows us to zoom in on the term structure of implied volatility around the money at a high resolution scale. To be more specific, our paper adds to the existing literature in two ways. First, we obtain a generalization of the Osajima (2015) energy expansion to a non-Markovian case, and using the new expansion, we extend the analysis of Friz et al. (2018) to the case where the volatility is driven by a rough ($H < 1/2$) fractional Brownian motion. Indeed, Laplace approximation methods on Wiener space in the spirit of Azencott (1982, 1985), Ben Arous (1988) and Bismut (1984) can be adapted to the present context, so that our analysis builds upon this framework in a fractional setting. Unlike many other works in this field, we do not rely on density expansions. We derive a small-time asymptotic expression for log call prices in the moderate deviations regime, and using a framework of Gao and Lee (2014), transform it into a term structure for Black-Scholes implied volatility. This in turn allows us to push the applicability range of known ATM skew approximation formulae from CLT type log-moneyness deviations of order $t^{1/2}$ to the wider moderate deviations regime. Finally, using a version of the rough Bergomi model by Bayer et al. (2016) with constant forward variance curve, we demonstrate numerically that our implied volatility asymptotics capture very well the geometry of the term structure of implied volatility over a wide array of maturities, extending up to a year.

1.2 Monte Carlo pricing under rough stochastic volatility (Chapter 3)

This chapter is based on joint work with Peter K. Friz, Christian Bayer, Paul Gassiat and Jörg Martin. Its contents have been partially reproduced from (Bayer, Friz, Gassiat, Martin, & Stemper, 2017). A preprint of the article has been made available at https://arxiv.org/abs/1710.07481.

With the problems pertaining to categories (I), (II) and also (III) in mind, it does not come as a surprise that a large part of research prior and in parallel to this thesis is concentrated around stochastic simulation schemes. For a large class of rough stochastic volatility models – this in particular includes the *rough Bergomi* model (Bayer et al., 2016) – the stochastic structure is such that Monte Carlo pricing requires an efficient joint simulation scheme for the bivariate Gaussian process of a Brownian motion and a (rough) Volterra process constructed from it. Numerical estimates for the stock price at maturity (and therefore also for the fair price of a European Call) then follow easily by for example an Euler discretization of the SDE describing the stock price dynamics. For the numerical simulation of the discussed bivariate process on some equidistant grid with n steps, a range of methods have been proposed. The original scheme discussed by Bayer et al. (2016) is the Cholesky method (Glasserman, 2003) which has a complexity of $\mathcal{O}(n^3)$ for the one-time Cholesky decomposition of the joint covariance matrix but then produces *ex-*

act samples at a cost of $\mathcal{O}(n^2)$. Using FFT, the Hybrid scheme of Bennedsen, Lunde, and Pakkanen (2017) pushes the complexity of obtaining a single bivariate path down to $\mathcal{O}(n \log n)$ at the cost of the scheme no longer being exact. Recently, McCrickerd and Pakkanen (2018) have "turbocharged" the Hybrid scheme by employing a mix of variance reduction methods such as control variates and antithetic variates. Finally, Horvath, Jacquier, and Muguruza (2017) provide a rough Donsker type approximation of the joint process.

In our work, we consider a general class of rough stochastic volatility models. We assume that the instantaneous volatility is given explicitly in terms of a fBM $\widehat{W}$ which is constantly correlated with the Brownian driver of the SDE describing the stock price dynamics. For such a model, a straightforward application of the conditioning formula of Romano and Touzi (1997); Willard (1997) reduces the Call price functional to the expectation of the Black-Scholes formula for some stochastic input parameters. Numerically, efficient simulation of the latter proves to be challenging task. In particular, this involves the efficient simulation of an object $\mathscr{I}$ which is given by the integration of White Noise $\dot{W}$ against a functional of the Volterra fBM $\widehat{W}$ constructed from it.[3] By integrating a Karhunen-Loève-style approximation of $\dot{W}$ against the fractional Volterra Kernel of $\widehat{W}$, we first retrieve a joint approximation $\left(\dot{W}^{\varepsilon}, \widehat{W}^{\varepsilon}\right)$ of White Noise and its corresponding Volterra process. It turns out that an approximation of $\mathscr{I}$ using $\left(\dot{W}^{\varepsilon}, \widehat{W}^{\varepsilon}\right)$ leads to the approximate integral not converging to the Itô integral $\mathscr{I}$ as the approximation becomes finer. In fact, even for the Brownian case $H = 1/2$, a famous result by Wong and Zakai (1965) shows that the limit of the approximate integral is the Stratonovich version of $\mathscr{I}$ which is given by the Ito integral plus an Itô-Stratonovich correction term. For $H < 1/2$, this correction term does not exist. As is argued in (Bayer et al., 2017), the theory of *regularity structures* introduced by Hairer (2014) provides an appropriate framework to address this issue and to derive correction terms that *renormalize* the approximative integral such that it converges to the desired object of interest.[4]

[3] The methods discussed above come to help here but we pursue a different approach.

[4] In this thesis, we provide the numerical counterpart to the theoretical results obtained in (Bayer et al., 2017). A reader more interested in the derivation of the correction terms is advised to study the original paper.

Several numerical experiments then confirm that the renormalization works as planned. In particular, with H being the Hurst index of the considered fBM, our numerical rates are consistent with a theoretical strong rate of convergence of almost H for the approximate integral across the full range of $0 < H < \frac{1}{2}$. Moreover, some weak approximation results point towards a weak rate of $2H$ for the option price approximation across the full range of $0 < H < \frac{1}{2}$.

1.3 Deep calibration of rough volatility models (Chapter 4)

This chapter is based on joint work with Christian Bayer. Its contents have been reproduced (almost) verbatim from (Bayer & Stemper, 2018). A preprint of this work can be accessed at https://arxiv.org/abs/1810.03399.

In this chapter, we introduce a novel model calibration scheme Model calibration describes the optimization procedure of finding model parameters such that the IV surface induced by the model best approximates a given market IV surface in an appropriate metric. In the absence of an analytical solution, it is standard practice to solve the arising weighted non-linear least squares problem using iterative optimizers such as Levenberg-Marquardt (LM) (Levenberg, 1944; Marquardt, 1963). However, these optimizers rely on the repetitive evaluation of the function φ from the space of model & option parameters (and external market information) to model BS implied volatility. If each such evaluation involves a time– and/or memory–intensive operation such as a Monte Carlo simulation in the case of *rough Bergomi* (Bayer et al., 2016) or other (rough) stochastic volatility models, this makes efficient calibration prohibitively expensive.

Made possible by theoretical advancements as well as the widespread availability of cheap, high performance computing hardware, *Machine Learning* has seen a tremendous rise in popularity among academics and practitioners in recent years. Breakthroughs such as (super-) human level performance in image classification (Krizhevsky, Sutskever, & Hinton, 2012; Simonyan & Zisserman, 2014; Szegedy et al., 2015) or playing the ancient Chinese board game

Go (Silver et al., 2017) may all be attributed to the advent of *Deep Learning* (Goodfellow, Bengio, & Courville, 2016). Fundamentally, its success stems from the capability of multi-layered artificial neural networks to closely approximate functions f only implicitly available through input-output pairs $\{(x_i, f(x_i))\}_{i=1}^{N}$, so-called *labeled data*.

The fundamental idea of this paper is to leverage this capability by training a fully-connected neural network on specifically tailored, synthetically generated training data to learn a map φ_{NN} approximating the true implied volatility map φ.

Remark 1.3.1. In a related but different approach, Hernandez (2017) proposes to use a neural network to learn the complete calibration routine – denoted Ψ in our notation in (4.2.1) – taking market data as inputs and returning calibrated model parameters directly. He demonstrates numerically the prowess of his approach by calibrating the popular short rate model of Hull and White (1990) to market data.

Both generating the synthetic data set as well as the actual neural network training are expensive in time and computing resource requirements, yet they only have to be performed a single time. Trained networks may then be quickly and efficiently saved, moved and deployed. The benefit of this novel approach is twofold: First, evaluations of φ_{NN} amount to cheap and almost instantaneous forward runs of a pre-trained network. Second, automatic differentiation of φ_{NN} with respect to the model parameters returns fast and accurate approximations of the Jacobians needed for the LM calibration routine. Used together, they allow for the efficient calibration of *any* (rough) stochastic volatility model including *rough Bergomi*.

To demonstrate the practical benefits of our approach numerically, we apply our machinery to Heston (1993) and *rough Bergomi* (Bayer et al., 2016) as representatives of classical and (rough) stochastic volatility models respectively. Speed-wise, no *systematic* comparison is made between the proposed neural network based approach and existing methods, yet with about 40ms per evaluation, our approach is at least competitive with existing Heston pricing methods and beats state-of-the-art rough Bergomi pricing schemes by magnitudes. Also, in both experiments, φ_{NN} exhibits small relative errors across the highly-liquid parts of the IV surface, recovering characteristic IV smiles and ATM IV skews. To quantify the uncertainty about model parameter estimates obtained by calibrating with φ_{NN}, we infer model parameters in a Bayesian spirit from (i) a synthetically generated IV surface and (ii) SPX market IV data. In both experiments, a simple (weighted) Bayesian nonlinear regression returns a (joint) posterior distribution over model parameters that (1) correctly identifies sensible model parameter regions and (2) places its peak at or close to the true (in the case of the synthetic IV) or previously reported (Bayer et al., 2016) (in the case of the SPX surface) model parameter values. Both experiments thus confirm the idea that φ_{NN} is sufficiently accurate for calibration.

2 Short-term near-the-money skew in rough fractional volatility models

The chapter is organized as follows: In Section 2.1 we set the scene, describing the class of models included in our framework ((2.1.1) and (2.1.2)) and recalling some known results ((2.1.4) and (2.1.7)), which are the starting point of our analysis. Most importantly, we argue that for small-time considerations it would suffice to restrict our attention to a class of stochastic volatility models of the form (2.1.3) with a volatility process driven by a Gaussian Volterra process such as in (2.1.2). We formulate general assumptions on the Volterra kernel (Assumptions 2.1.1 and 2.1.5) and on the function σ in (2.1.3) (Assumption 2.1.4) under which our results are valid. In Section 2.2 we gather our main results, concerning a higher order expansion of the energy (Theorem 2.2.1), and a general expansion formula for the corresponding call prices. We derive the classical Black-Scholes expansion for the call price, using the latter result mentioned above. In addition, in Section 2.2 we formulate moderate deviation expansions, which allow us to derive the corresponding asymptotic formulae for implied volatilities and implied volatility skews. Finally, Section 2.3 displays our simulation results. Sections 2.4, 2.5 and 2.6 are devoted to proofs of the energy expansion, the price expansion and the moderate deviations expansion, respectively. In the appendix, we have collected some auxiliary lemmas, which are used in different sections.

2.1 Exposition and assumptions

We consider a rough stochastic volatility model, normalized to $r = 0$ and $S_0 = 1$, of the form suggested by Forde and Zhang (2017)

$$\frac{dS_t}{S_t} = \sigma(\widehat{B}_t)d\left(\overline{\rho}W_t + \rho B_t\right). \tag{2.1.1}$$

Here (W, B) are two independent standard Brownian motions, $\rho \in (-1, 1)$ a correlation parameter, and $\overline{\rho}^2 = 1 - \rho^2$. Then $\overline{\rho}W + \rho B$ is another standard Brownian motion which has constant correlation ρ with the factor B, which drives the stochastic volatility

$$\sigma_{\text{stoch}}(t, \omega) := \sigma(\widehat{B}_t(\omega)) \equiv \sigma(\widehat{B}).$$

Here $\sigma(.)$ is some real-valued function, typically smooth but not bounded, and we will denote by $\sigma_0 := \sigma(0)$ the spot volatility, with $\widehat{B}$ a Gaussian (Volterra) process of the form

$$\widehat{B}_t = \int_0^t K(t, s)\, dB_s, \quad t \geq 0, \tag{2.1.2}$$

for some kernel K, which shall be further specified in Assumptions 2.1.1 and 2.1.5 below. The log-price $X_t = \log(S_t)$ satisfies

$$dX_t = -\frac{1}{2}\sigma^2(\widehat{B}_t)dt + \sigma(\widehat{B}_t)d\left(\overline{\rho}W + \rho B\right), \quad X_0 = 0. \tag{2.1.3}$$

Recall that by Brownian scaling, for fixed $t > 0$,

$$(B_{ts}, W_{ts})_{s\geq 0} \overset{law}{=} \varepsilon(B_s, W_s)_{s\geq 0}, \quad \text{where} \quad \varepsilon \equiv \varepsilon(t) \equiv t^{1/2}.$$

As a direct consequence, classical short-time SDE problems can be analyzed as small-noise problems on a unit time horizon. For our analysis, it will also be crucial to impose such a scaling property on the Gaussian process $\widehat{B}$ (more precisely, on the kernel K in (2.1.2)) driving the volatility process in our model:

Assumption 2.1.1 (Small time self-similarity)**.** There exists a number t_0 with $0 < t_0 \leq 1$ and a function $t \mapsto \widehat{\varepsilon} = \widehat{\varepsilon}(t)$, $0 \leq t \leq t_0$, such that

$$(\widehat{B}_{ts} : 0 \leq s \leq t_0) \overset{law}{=} (\widehat{\varepsilon}\widehat{B}_s : 0 \leq s \leq t_0).$$

In fact, we will always have

$$\widehat{\varepsilon} \equiv \widehat{\varepsilon}(t) \equiv t^H = \varepsilon^{2H},$$

which covers the examples of interest, in particular standard fractional Brownian motion $\widehat{B} = B^H$ or Riemann-Liouville fBM with explicit kernel $K(t,s) = \sqrt{2H}|t-s|^{H-1/2}$. (This is very natural, even from a general perspective of self-similar processes, see (Lamperti, 1962).)

We insist that no (global) self-similarity of $\widehat{B}$ is required, as only $\widehat{B}|_{[0,t]}$ for arbitrarily small t matters.

Remark 2.1.2. It should be possible to replace the fractional Brownian motion by a certain fractional Ornstein-Uhlenbeck process in the results obtained in this chapter. Intuitively, this replacement creates a negligible perturbation (for $t \ll 1$) of the fBm environment. A similar situation was in fact encountered in (Cass & Friz, 2010), where fractional scaling at times near zero was important. To quantify the perturbation, Cass and Friz (2010) introduced an easy to verify coupling condition (see Corollary 2 in (Cass & Friz, 2010)). It should be possible to employ a version of this condition in the present chapter to justify the replacement mentioned above. We will however not pursue this point further here.

Remark 2.1.3. Throughout this article, one can consider a classical (Markovian, diffusion) stochastic volatility setting by taking $K \equiv 1$, or equivalently $H \equiv 1/2$, by simply ignoring all hats ($\widehat{\cdot}$) in the sequel. In particular then, $\frac{\widehat{\varepsilon}}{\varepsilon} \equiv 1$ in all subsequent formulae.

General facts on large deviations of Gaussian measures on Banach spaces (Deuschel & Stroock, 1989) such as the path space $C([0,1],\mathbb{R}^3)$ imply that a large deviation principle holds for the triple $\{\widehat{\varepsilon}(W,B,\widehat{B}) : \widehat{\varepsilon} > 0\}$, with speed $\widehat{\varepsilon}^2$ and rate function

$$\begin{cases} \frac{1}{2}\|h\|^2_{H_0^1} + \frac{1}{2}\|f\|^2_{H_0^1}, & f,h \in H_0^1 \text{ and } \widehat{f} = K\dot{f}, \\ +\infty, & \text{otherwise}, \end{cases} \tag{2.1.4}$$

where

$$K\dot{f}(t) := \int_0^t K(t,s)\,\dot{f}(s)ds$$

for $f \in H_0^1$, the space of absolutely continuous paths with L^2 derivative

$$H_0^1 := \left\{ f : [0,1] \to \mathbb{R} \text{ continuous } \middle| \|f\|^2_{H_0^1} := \int_0^1 |\dot{f}(s)|^2 ds < \infty,\ f(0) = 0 \right\}. \tag{2.1.5}$$

This enables us to derive a large deviations principle for X in (2.1.3): the (local) small-time self-similarity property of $\widehat{B}$ (Assumption 2.1.1) implies that $X_t \overset{law}{=} X_1^\varepsilon$ where

$$dX_t^\varepsilon = \sigma(\widehat{\varepsilon}\widehat{B}_t)\varepsilon d\left(\overline{\rho}W_t + \rho B_t\right) - \frac{1}{2}\varepsilon^2\sigma^2(\widehat{\varepsilon}\widehat{B}_t)dt, \quad X_0^\varepsilon = 0.$$

For what follows, it will be convenient to consider a rescaled version of (2.1.3)

$$d\widehat{X}_t^\varepsilon \equiv d\left(\frac{\widehat{\varepsilon}}{\varepsilon}X_t^\varepsilon\right) = \sigma(\widehat{\varepsilon}\widehat{B}_t)\widehat{\varepsilon}d\left(\overline{\rho}W_t + \rho B_t\right) - \frac{1}{2}\varepsilon\widehat{\varepsilon}\sigma^2(\widehat{\varepsilon}\widehat{B}_t)dt, \quad \widehat{X}_0^\varepsilon = 0.$$

Under a linear growth condition on the function σ, Forde and Zhang (2017) use the extended contraction principle to establish a large deviations principle for $(\widehat{X}_1^\varepsilon)$ with speed $\widehat{\varepsilon}^2$. More precisely, with

$$\varphi_1\left(h,f\right) := \Phi_1(h,f,\widehat{f}) = \int_0^1 \sigma(\widehat{f})d\left(\overline{\rho}h + \rho f\right), \tag{2.1.6}$$

the rate function is given by

$$\begin{aligned} I\left(x\right) &= \inf_{h,f\in H_0^1}\left\{\frac{1}{2}\int_0^1 \dot{h}^2dt + \frac{1}{2}\int_0^1 \dot{f}^2dt : \varphi_1\left(h,f\right) = x\right\} \\ &= \inf_{f\in H_0^1}\left\{\frac{1}{2}\frac{\left(x - \rho\left\langle\sigma(\widehat{f}),\dot{f}\right\rangle\right)^2}{\overline{\rho}^2\left\langle\sigma^2(\widehat{f}),1\right\rangle} + \frac{1}{2}\int_0^1 \dot{f}^2dt\right\}, \end{aligned} \tag{2.1.7}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product on $L^2\left([0,1],dt\right)$. Several other proofs (under varying assumptions on σ) have appeared since (Jacquier, Pakkanen, & Stone, 2017; Bayer et al., 2017; Gulisashvili, 2017).

As a matter of fact, this chapter relies on moderate - rather than large - deviations, as emphasized in (iiic) below. To this end, let us make

Assumption 2.1.4.

(i) (Positive spot vol) Assume $\sigma : \mathbb{R} \to \mathbb{R}$ is smooth with $\sigma_0 := \sigma(0) > 0$.

(ii) (Roughness) The Hurst parameter H satisfies $H \in (0, 1/2]$.

(iiia) (Martingality) The price process $S = \exp X$ is a martingale.

(iiib) (Short-time moments) $\forall m < \infty \ \exists t > 0$: $E(S_t^m) < \infty$.

While condition (iiia) hardly needs justification, we emphasize that conditions (iiia-b) are only used to the extent that they imply condition (iiic) given below (which thus may replace (iiia-b) as an alternative, if more technical, assumption). The reason we point this out explicitly is that all the conditions (iiia-c) are implicit (growth) conditions on the function $\sigma(.)$. For instance, (iiia-b) was seen to hold under a linear growth assumption (Forde & Zhang, 2017; Gulisashvili, 2017), whereas the log-normal volatility case (think of $\sigma(x) = e^x$) is complicated. Martingality, for instance, requires $\rho \leq 0$ and there is a critical moment $m^* = m^*(\rho)$, even when $\rho < 0$. See (Sin, 1998; Jourdain, 2004; Lions & Musiela, 2007) for the case $H = 1/2$ and the forthcoming work (Friz & Gassiat, 2018) for the general rough case $H \in (0, 1]$. We view (iiic) simply as a more flexible condition that can hold in situations where (iiib) fails.

(iiic) (Call price upper moderate deviation bound) For every $\beta \in (0, H)$, and every fixed $x > 0$, and $\widehat{x}_\varepsilon := x\varepsilon^{1-2H+2\beta}$,

$$E[(e^{X_1^\varepsilon} - e^{\widehat{x}_\varepsilon})^+] \leq \exp\left(-\frac{x^2 + o(1)}{2\sigma_0^2 \varepsilon^{4H-4\beta}}\right).$$

This condition is reminiscent of the "upper part" of the large deviation estimate obtained in (Forde & Zhang, 2017)

$$E[(e^{X_1^\varepsilon} - e^{x\varepsilon^{1-2H}})^+] = \exp\left(-\frac{I(x) + o(1)}{\varepsilon^{4H}}\right) \ . \tag{2.1.8}$$

If fact, if one *formally* applies this with x replaced by $x\varepsilon^{2\beta}$, followed by Taylor expanding the rate function,

$$I(x\varepsilon^{2\beta}) \sim \tfrac{1}{2} I''(0) x^2 \varepsilon^{4\beta} = \tfrac{1}{2\sigma_0^2} x^2 \varepsilon^{4\beta} \ ,$$

one readily arrives at the estimate (iiic). Unfortunately, $o(1) = o_x(1)$ in (2.1.8), which is a serious obstacle in making this argument rigorous. Instead, we will give a direct argument (Lemma 2.6.1) to see how (iiia-b) implies (iiic).

In the sequel, we will use another mild assumption on the kernel.

Assumption 2.1.5. The kernel K has the following properties

(i) $\widehat{B}_t = \int_0^t K(t,s)dB_s$ has a continuous (in t) version on $[0,1]$.

(ii) $\forall t \in [0,1]: \int_0^t K(t,s)^2 ds < \infty$.

Note that the Riemann-Liouville kernel $K(t,s) = \sqrt{2H}(t-s)^\gamma$, $\gamma = H-1/2$ satisfies Assumption 2.1.5.

Remark 2.1.6. Assumption 2.1.5 implies that the Cameron-Martin space $\mathcal{H}$ of $\widehat{B}$ is given by the image of H_0^1 under K, i.e.,

$$\mathcal{H} = \{K\dot{f} \mid f \in H_0^1\}.$$

See Lemma 2.4.3 and Remark 2.4.4 for more details. A reference and also a sufficient condition for Assumption 2.1.5 (i) can be found e.g. in (Decreusefond, 2005, Section 3).

2.2 Main results

The following result can be seen as a non-Markovian extension of work by Osajima (2015). The statement here is a combination of Theorem 2.4.10 and Proposition (2.4.14) below. Recall that $\sigma_0 = \sigma(0)$ represents spot-volatility. We also set $\sigma_0' \equiv \sigma'(0)$.

Theorem 2.2.1 (Energy expansion). *The rate function (or energy) I is smooth in a neighbourhood of $x = 0$ (at-the-money) and it is of the form*

$$I(x) = \frac{1}{\sigma_0^2}\frac{x^2}{2} - \left(6\rho\frac{\sigma_0'}{\sigma_0^4}\int_0^1\int_0^t K(t,s)dsdt\right)\frac{x^3}{3!} + \mathcal{O}(x^4).$$

The next result is an exact representation of call prices, valid in a non-Markovian generality, and amenable to moderate- and large-deviation analysis (Theorem 2.2.4 below).

Theorem 2.2.2 (Pricing formula)**.** *For a fixed log-strike $x \geq 0$ and time to maturity $t > 0$, set $\widehat{x} := \frac{\varepsilon}{\widehat{\varepsilon}} x$, where $\varepsilon = t^{1/2}$ and $\widehat{\varepsilon} = t^H = \varepsilon^{2H}$, as before. Then we have*

$$\begin{aligned} c(\widehat{x}, t) &= E\left[(\exp(X_t) - \exp \widehat{x})^+\right] \\ &= e^{-\frac{I(x)}{\widehat{\varepsilon}^2}} e^{\frac{\varepsilon}{\widehat{\varepsilon}} x} J(\varepsilon, x), \end{aligned} \tag{2.2.1}$$

where

$$J(\varepsilon, x) := E\left[e^{-\frac{I'(x)}{\widehat{\varepsilon}^2}\widehat{U}^\varepsilon} \left(\exp\left(\tfrac{\varepsilon}{\widehat{\varepsilon}}\widehat{U}^\varepsilon\right) - 1\right) e^{I'(x) R_2^\varepsilon} \mathbf{1}_{\widehat{U}^\varepsilon \geq 0}\right]$$

and $\widehat{U}^\varepsilon$ is a random variable of the form

$$\widehat{U}^\varepsilon = \widehat{\varepsilon} g_1 + \widehat{\varepsilon}^2 R_2^\varepsilon \tag{2.2.2}$$

with g_1 a centred Gaussian random variable, explicitly given in equation (2.5.3) below, and R_2^ε is a (random) remainder term, in the sense of a stochastic Taylor expansion in $\widehat{\varepsilon}$, see Lemma 2.5.2 for more details.

Example 2.2.3 (Black-Scholes model)**.** We fix volatility $\sigma(\cdot) \equiv \sigma > 0$, and $H = 1/2$ so that $\widehat{\varepsilon} = \varepsilon$ and all $\widehat{\cdot}$ can be omitted. Energy is given by $I(x) = \frac{x^2}{2\sigma^2}$ and

$$U^\varepsilon = \varepsilon g_1 + \varepsilon^2 R_2^\varepsilon \equiv \varepsilon \sigma W_1 - \varepsilon^2 \sigma^2 / 2$$

with $R_2^\varepsilon = R_2 \equiv -\sigma^2/2$ independent of ε. Moreover,

$$\begin{aligned} J(\varepsilon, x) &= E\left[e^{-\frac{I'(x)}{\varepsilon^2} U^\varepsilon} \left(e^{U^\varepsilon} - 1\right) e^{I'(x) R_2} \mathbf{1}_{U^\varepsilon \geq 0}\right] \\ &= E\left[e^{-\frac{I'(x)}{\varepsilon} g_1} \left(e^{\varepsilon g_1 - \varepsilon^2 \frac{\sigma^2}{2}} - 1\right) \mathbf{1}_{\{g_1 \geq \frac{\varepsilon \sigma^2}{2}\}}\right] \\ &= E\left[e^{-\alpha W_1} \left(e^{\varepsilon \sigma W_1 - \frac{(\varepsilon\sigma)^2}{2}} - 1\right) \mathbf{1}_{\{W_1 \geq \frac{\varepsilon\sigma}{2}\}}\right] \\ &= e^{-\frac{(\varepsilon\sigma)^2}{2}} M(-\alpha + \varepsilon\sigma) - M(-\alpha) \end{aligned} \tag{2.2.3}$$

with $\alpha := \frac{I'(x)\sigma}{\varepsilon} = \frac{1}{\sigma}(x/\varepsilon)$, and, in terms of the standard Gaussian cdf Φ,

$$M(\beta) := E\left[e^{\beta W_1} \mathbf{1}_{\{W_1 \geq \frac{\varepsilon\sigma}{2}\}}\right] = e^{\beta^2/2} \Phi\left(\beta - \tfrac{\varepsilon\sigma}{2}\right) .$$

Using the expansion $\Phi(-y) = \frac{1}{y\sqrt{2\pi}} e^{-y^2/2}(1-y^{-2}+...)$, as $y \to \infty$ one deduces, for fixed $x > 0$, the asymptotic relation, as $\varepsilon \to 0$,

$$J(\varepsilon, x) \sim \frac{e^{-x/2}}{\sqrt{2\pi}} \frac{\varepsilon^3 \sigma^3}{x^2}. \tag{2.2.4}$$

We will be interested (cf. Theorem 2.2.4) in replacing x by $\tilde{x} = x\varepsilon^{2\beta} \to 0$ for $\beta > 0$. This gives $\tilde{\alpha} = \frac{1}{\sigma}(x/\varepsilon^{1-2\beta})$ and the above analysis, now based on $\tilde{\alpha} \to \infty$, remains valid[1] for β in the "moderate" regime $\beta \in [0, 1/2)$ and we obtain

$$\forall x > 0, \beta \in [0, 1/2) : J\left(\varepsilon, x\varepsilon^{2\beta}\right) \sim \frac{1}{\sqrt{2\pi}} \frac{\varepsilon^{3-4\beta}\sigma^3}{x^2}. \tag{2.2.5}$$

Let us point out, for the sake of completeness, that a similar expansion is *not* valid for $\beta > 1/2$. To see this, first note that (2.2.1) implies that $J(\varepsilon, x)|_{x=0}$ is precisely the ATM call price with time $t = \varepsilon^2$ from expiration. Well-known ATM asymptotics then imply that $J(\varepsilon, x)|_{x=0} \sim \frac{1}{\sqrt{2\pi}}\varepsilon\sigma$ as $\varepsilon \to 0$. These asymptotics are unchanged in case of $o(t^{1/2}) = o(\varepsilon)$ out-of-moneyness ("almost-at-the-money" in the terminology of Friz et al. (2018)), which readily implies

$$\forall x > 0, \beta > 1/2 : J\left(\varepsilon, x\varepsilon^{2\beta}\right) \sim \frac{1}{\sqrt{2\pi}}\varepsilon\sigma = \text{const} \times \varepsilon$$

At last, we have the borderline case $\beta = 1/2$, or $\tilde{x} = x\varepsilon$. From e.g. (Muhle-Karbe & Nutz, 2011, Thm 3.1), we see that $c(x\varepsilon, \varepsilon^2) \sim a(x;\sigma)\varepsilon$ with positive constant $a(x;\sigma)$. A look at (2.2.1) then reveals

$$\forall x > 0 : J(\varepsilon, x\varepsilon) \sim a(x;\sigma)\varepsilon e^{\frac{x^2}{2\sigma^2}} = \text{const} \times \varepsilon\ .$$

For the call price expansion in the large / moderate deviations regime, $\beta \in [0, 1/2)$, the polynomial in ε-behaviour of (2.2.5) implies that the J-term in the pricing formula will be negligible on the moderate / large deviation scale, in the sense for any $\theta > 0$, we have $\varepsilon^\theta \log J(\varepsilon, x\varepsilon^{2\beta}) \to 0$ as $\varepsilon \to 0$. Consequently, with $k_t = kt^\beta$, for $t = \varepsilon^2$, $k > 0$, $\beta \in [0, 1/2)$, we get the "moderate" Black-Scholes call price expansion,

$$-\log c_{BS}(k_t, t) = \frac{1}{t^{1-2\beta}} \frac{k^2}{2\sigma^2} (1 + o(1)) \text{ as } t \downarrow 0.$$

[1] More terms in the expansion of Φ are needed.

While the above can be confirmed by elementary analysis of the Black–Scholes formula, the following theorem exhibits it as an instance of a general principle. See (Friz et al., 2018) for a general diffusion statement.

Theorem 2.2.4 (Moderate Deviations)**.** *In the rough volatility regime $H \in (0, 1/2]$, consider log-strikes of the form*

$$k_t = kt^{\frac{1}{2}-H+\beta} \quad \textit{for a constant} \quad k \geq 0.$$

(i) For $\beta \in (0, H)$, and every $\theta > 0$, we have

$$-\log c(k_t, t) = \frac{I''(0)}{t^{2H-2\beta}} \frac{k^2}{2} + O(t^{3\beta-2H}) + O(t^{-\theta}) \qquad \textit{as } t \downarrow 0.$$

(ii) For $\beta \in (0, \frac{2}{3}H)$, and every $\theta > 0$, we have

$$-\log c(k_t, t) = \frac{I''(0)}{t^{2H-2\beta}} \frac{k^2}{2} + \frac{I'''(0)}{t^{2H-3\beta}} \frac{k^3}{6} + O(t^{4\beta-2H}) + O(t^{-\theta}) \qquad \textit{as } t \downarrow 0.$$

Moreover,

$$\begin{aligned} I''(0) &= \frac{1}{\sigma_0^2}, \\ I'''(0) &= -6\rho \frac{\sigma_0'}{\sigma_0^4} \int_0^1 \int_0^t K(t,s) ds dt = -6\rho \frac{\sigma_0'}{\sigma_0^4} \langle K1, 1 \rangle, \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ is the inner product in $L^2([0,1])$.

Remark 2.2.5. In principle, further terms (of order $t^{i\beta-2H}$, $i = 4, 5, \dots$) can be added to this expansion of log call prices, given that the energy has sufficient regularity, see Theorem 2.2.6. We also note that, for small enough β, the error term $O(t^{-\theta})$ can be omitted. In any case, one can replace the additive error bounds by (cruder) ones, where the right-most term in the expansion is multiplied with $(1 + o(1))$, as was done in (Friz et al., 2018).

Proof of Theorem 2.2.4. We apply Theorem 2.2.2 with $\widehat{x} = k_t = kt^{1/2-H+\beta}$, i.e., with $x = kt^\beta = k\varepsilon^{2\beta}$. In particular, we so get, with $\widehat{\varepsilon} = t^H$ and $\varepsilon = t^{1/2}$,

$$c(k_t, t) = e^{-\frac{I(x)}{\widehat{\varepsilon^2}}} e^{\frac{\varepsilon}{\widehat{\varepsilon}} x} J\left(\varepsilon, k\varepsilon^{2\beta}\right).$$

The technical Proposition 2.6.3 asserts that, for fixed $k > 0$, the factor J is negligible in the sense that, for every $\theta > 0$,

$$\varepsilon^{\theta} \log J(\varepsilon, k\varepsilon^{2\beta}) \to 0 \qquad \text{as } \varepsilon \to 0 \ .$$

The theorem now follows immediately from the Taylor expansion of $I(x)$ around $x = 0$ (see Theorem 2.2.1), plugging in $x = kt^{\beta}$. Indeed, replacing $I(x)$ by the Taylor-jet seen in (i),(ii), leads exactly to an error term $O(t^{3\beta-2H})$, resp. $O(t^{4\beta-2H})$. □

Fix real numbers $k > 0$, $0 < H < \frac{1}{2}$, $0 < \beta < H$, and an integer $n \geq 2$. For every $t > 0$, set

$$k_t = kt^{\frac{1}{2}-H+\beta},$$

and denote

$$\phi_{n,H,\beta,\theta}(t) = \max\left\{t^{2H-2\beta-\theta}, t^{(n-1)\beta}\right\}.$$

Here, $\theta > 0$ can be arbitrarily small. It is clear that for all small t and θ small enough,

$$\phi_{n,H,\beta,\theta}(t) = t^{2H-2\beta-\theta} \Leftrightarrow 2H - 2\beta \leq (n-1)\beta \Leftrightarrow \frac{2H}{n+1} \leq \beta,$$

while

$$\phi_{n,H,\beta,\theta}(t) = t^{(n-1)\beta} \Leftrightarrow 2H - 2\beta > (n-1)\beta \Leftrightarrow \beta < \frac{2H}{n+1}.$$

The following statement provides an asymptotic formula for the implied variance.

Theorem 2.2.6. *Suppose $0 < \beta < \frac{2H}{n}$ and $\theta > 0$ small enough. Then as $t \to 0$ (and for $k > 0$),*

$$\sigma_{\ impl}(k_t, t)^2 = \sum_{j=0}^{n-2} \frac{(-1)^j 2^j}{I''(0)^{j+1}} \left(\sum_{i=3}^{n} \frac{I^{(i)}(0)}{i!} k^{i-2} t^{(i-2)\beta} \right)^j + \mathcal{O}\left(\phi_{n,H,\beta,\theta}(t)\right). \tag{2.2.6}$$

The $\mathcal{O}$-estimate in (2.2.6) depends on n, H, β, θ, and k. It is uniform on compact subsets of $[0, \infty)$ with respect to the variable k.

Remark 2.2.7. Using the multinomial formula, we can represent the expression on the left-hand side of (2.2.6) in terms of certain powers of t. However, the coefficients become rather complicated.

Remark 2.2.8. Let an integer $n \geq 2$ be fixed, and suppose we would like to use only the derivatives $I^{(i)}(0)$ for $2 \leq i \leq n$ in formula (2.2.6) to approximate $\sigma_{\text{impl}}(k_t,t)^2$. Then, the optimal range for β is the following: $\frac{2H}{n+1} \leq \beta < \frac{2H}{n}$. On the other hand, if β is outside of the interval $[\frac{2H}{n+1}, \frac{2H}{n})$, more derivatives of the energy function at zero may be needed to get a good approximation of the implied variance in formula (2.2.6).

We will next derive from Theorem 2.2.6 several asymptotic formulas for the implied volatility. In the next corollary, we take $n = 2$.

Corollary 2.2.9. *As* $t \to 0$,

$$\sigma_{impl}(k_t,t) = \sigma_0 + \mathcal{O}(\phi_{2,H,\beta,\theta}(t)). \tag{2.2.7}$$

Corollary 2.2.9 follows from Theorem 2.2.6 with $n = 2$, the equality

$$I''(0) = \sigma_0^{-2} \tag{2.2.8}$$

given in Theorem 2.2.4, and the Taylor expansion $\sqrt{1+h} = 1 + \mathcal{O}(h)$ as $h \to 0$.

In the next corollary, we consider the case where $n = 3$.

Corollary 2.2.10. *Suppose* $\beta < \frac{2H}{3}$. *Then, as* $t \to 0$,

$$\sigma_{impl}(k_t,t) = \sigma_0 + \rho\frac{\sigma_0'}{\sigma_0}\langle K1,1\rangle k t^{\beta} + \mathcal{O}(\phi_{3,H,\beta,\theta}(t)). \tag{2.2.9}$$

Corollary 2.2.10 follows from Theorem 2.2.6 with $n = 3$, formula (2.2.8), the equality

$$I'''(0) = -6\rho\frac{\sigma_0'}{\sigma_0^4}\langle K1,1\rangle \tag{2.2.10}$$

(see Theorem 2.2.4), and the expansion $\sqrt{1+h} = 1 + \frac{1}{2}h + \mathcal{O}(h^2)$ as $h \to 0$.

Using Corollary 2.2.10, we establish the following implied volatility skew formula in the moderate deviation regime.

Corollary 2.2.11. *Let $0 < H < \frac{1}{2}$, $0 < \beta < \frac{2}{3}H$, and fix $y, z > 0$ with $y \neq z$. Then as $t \to 0$,*

$$\frac{\sigma_{impl}(yt^{\frac{1}{2}-H+\beta}, t) - \sigma_{impl}(zt^{\frac{1}{2}-H+\beta}, t)}{(y-z)t^{\frac{1}{2}-H+\beta}} \sim \rho \frac{\sigma_0'}{\sigma_0} \langle K1, 1 \rangle t^{H-\frac{1}{2}}. \tag{2.2.11}$$

Remark 2.2.12. Corollary 2.2.11 complements earlier works of Alòs et al. (2007) and Fukasawa (2011, 2017). For instance, the following formula can be found in (Fukasawa, 2017, p. 6), see also (Fukasawa, 2011, p. 14):

$$\frac{\sigma_{\text{impl}}(yt^{\frac{1}{2}}, t) - \sigma_{\text{impl}}(zt^{\frac{1}{2}}, t)}{(y-z)t^{\frac{1}{2}}} \sim \rho C(H) \frac{\sigma_0'}{\sigma_0} t^{H-\frac{1}{2}}. \tag{2.2.12}$$

In formula (2.2.12), we employ the notation used in the present chapter. Our analysis shows that the applicability range of skew approximation formulas is by no means restricted to the Central Limit Theorem type log-moneyness deviations of order $t^{1/2}$. It also includes the moderate deviations regime of order $t^{1/2-H+\beta}$. The previous rate is clearly $\gg t^{1/2}$ as $t \to 0$.

Remark 2.2.13 (Symmetry). Write $\Phi_1(W, B, \widehat{B}; \rho; \sigma)$ for the "Itô-type map"

$$\Phi_1(W, B, \widehat{B}) := \int_0^1 \sigma(\widehat{B}) d\left(\overline{\rho} W + \rho B\right).$$

It equals, in law, $\Phi_1(W, -B, -\widehat{B}; -\rho; \sigma(-\cdot))$, and indeed all our formulae are invariant under this transformation. In particular, the skew remains unchanged when the pair (ρ, σ_0') is replaced by $(-\rho, -\sigma_0')$.

2.3 Simulation results

We verify our theoretical results numerically with a variant of the *rough Bergomi model* (Bayer et al., 2016) which fits nicely into the general rough volatility framework considered in this chapter. As before, the model has been normalized such that $S_0 = 1$ and $r = 0$. We let (W, B) be two independent Brownian motions and $\rho \in (-1, 1)$ with $\overline{\rho}^2 = 1 - \rho^2$ such that $Z = \overline{\rho} W + \rho B$ is another Brownian motion having constant correlation ρ with B. For some spot volatility σ_0 and volatility of volatility parameter η, we then assume the following dynamics for some asset S:

$$\frac{dS_t}{S_t} = \sigma(\widehat{B}_t)dZ_t \tag{2.3.1}$$

$$\sigma(x) = \sigma_0 \exp\left(\frac{1}{2}\eta x\right) \tag{2.3.2}$$

where $\widehat{B}$ is a Riemann-Liouville fBM given by

$$\widehat{B}_t = \sqrt{2H}\int_0^t |t-s|^{H-1/2} dB_s.$$

The approach taken for the Monte Carlo simulations of the quantities we are interested in is the one initially explored in the original *rough Bergomi* pricing paper (Bayer et al., 2016). That is, exploiting their joint Gaussianity, we use the well-known Cholesky method to simulate the joint paths of $(Z, \widehat{B})$ on some discretization grid $\mathcal{D}$. With (2.3.2) being an explicit function in terms of the rough driver, an Euler discretisation of the Ito SDE (2.3.1) on $\mathcal{D}$ then yields estimates for the price paths.

The Cholesky algorithm critically hinges on the availability and explicit computability of the joint covariance matrix of $(Z, \widehat{B})$ whose terms we readily compute below.[2]

Lemma 2.3.1. *For convenience, define constants $\gamma = \frac{1}{2} - H \in [0, \frac{1}{2})$ and $D_H = \frac{\sqrt{2H}}{H+\frac{1}{2}}$ and define an auxiliary function $G : [1, \infty) \to \mathbb{R}$ by*

$$G(x) = 2H\left(\frac{1}{1-\gamma}x^{-\gamma} + \frac{\gamma}{1-\gamma}x^{-(1+\gamma)}\frac{1}{2-\gamma}{}_2F_1(1, 1+\gamma, 3-\gamma, x^{-1})\right) \tag{2.3.3}$$

where ${}_2F_1$ denotes the Gaussian hypergeometric function (Olver, 2010). Then the joint process $(Z, \widehat{B})$ has zero mean and covariance structure governed by

[2] Note that expressions for the exact same scenario have have been computed before in the original pricing paper (Bayer et al., 2016), yet in that version the expression for the autocorrelation of the fBM $\widehat{B}$ was incorrect. We compute and state here all the relevant terms for the sake of completeness.

$$\begin{cases} \mathrm{Var}[\widehat{B}_t^2] = t^{2H}, & \textit{for } t \geq 0, \\ \mathrm{Cov}[\widehat{B}_s \widehat{B}_t] = t^{2H} G\left(s/t\right), & \textit{for } s > t \geq 0, \\ \mathrm{Cov}[\widehat{B}_s Z_t] = \rho D_H \left(s^{H+\frac{1}{2}} - (s - \min(t,s))^{H+\frac{1}{2}}\right), & \textit{for } t, s \geq 0, \\ \mathrm{Cov}[Z_t Z_s] = \min(t,s), & \textit{for } t, s \geq 0. \end{cases}$$

Numerical simulations[3] confirm the theoretical results obtained in the last section. In particular - as can be seen in Figure 2.1 – the asymptotic formula for the implied volatility (2.2.9) captures very well the geometry of the term structure of implied volatility, with particularly good results for higher H and worsening results as $H \downarrow 0$. Quite surprisingly, despite being an asymptotic formula, it seems to be fairly accurate over a wide array of maturities extending up to a single year.

2.4 Proof of the energy expansion

Consider

$$\begin{aligned} dX &= -\frac{1}{2}\sigma^2(Y)dt + \sigma\left(Y\right) d\left(\overline{\rho} dW + \rho dB\right), \quad X_0 = 0 \\ dY &= d\widehat{B}, \; Y_0 = 0 \end{aligned}$$

where $\widehat{B}_t = \int_0^t K\left(t,s\right) dB_s$ for a fixed Volterra kernel (recall (2.1.3) in the previous section). We study the small noise problem $(X^\varepsilon, Y^\varepsilon)$ where $\left(W, B, \widehat{B}\right)$ is replaced by $\left(\varepsilon W, \varepsilon B, \widehat{\varepsilon}\widehat{B}\right)$. The following proposition roughly says that

$$\mathbb{P}\left(X_1^\varepsilon \approx \frac{\varepsilon}{\widehat{\varepsilon}} x\right) \approx \exp\left(-\frac{I\left(x\right)}{\widehat{\varepsilon}^2}\right).$$

Proposition 2.4.1 (Forde and Zhang (2017)). *Under suitable assumptions (cf. Section 2.1), the rescaled process* $\left(\frac{\widehat{\varepsilon}}{\varepsilon} X_1^\varepsilon : \varepsilon \geq 0\right)$ *satisfies an LDP (with speed* $\widehat{\varepsilon}^2$*) and rate function*

$$I\left(x\right) = \inf_{f \in H_0^1} \left[\frac{\left(x - \rho G\left(f\right)\right)^2}{2\overline{\rho}^2 F\left(\widehat{f}\right)} + \frac{1}{2} E\left(f\right)\right] \equiv \inf_{f \in H_0^1} \mathcal{I}_x\left(f\right) \tag{2.4.1}$$

[3] The Python 3 code used to run the simulations can be found at github.com/RoughStochVol.

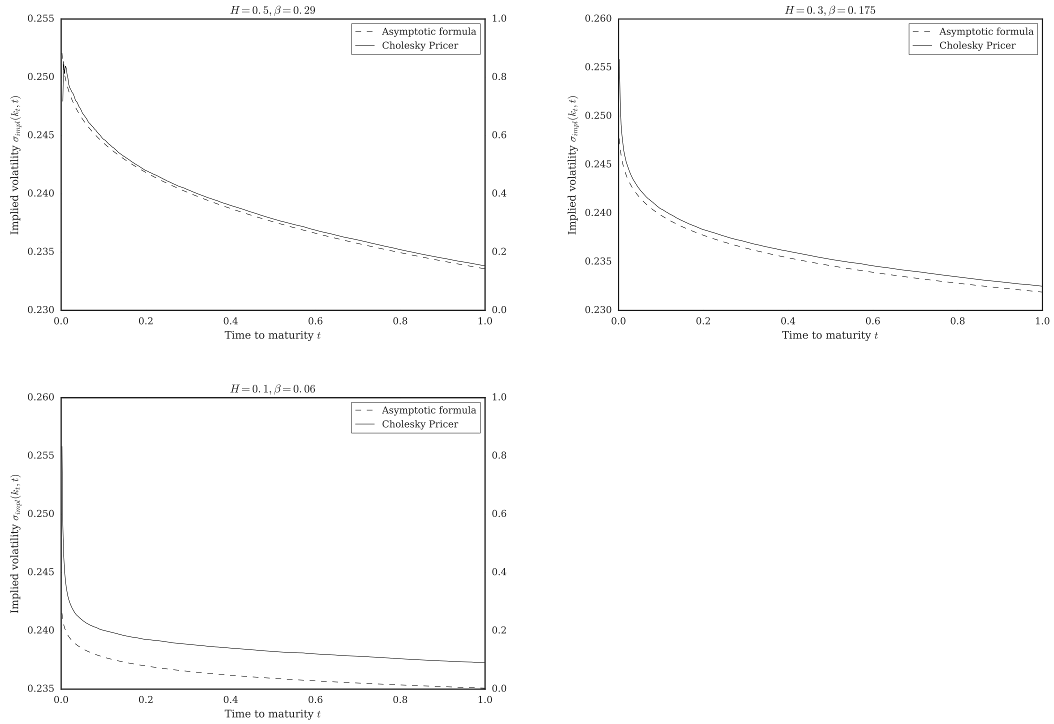


Figure 2.1: **Illustration of the IV term structure of the Modified Rough Bergomi model in the Moderate deviations regime with time-varying log-strike** $k_t = 0.4t^{\beta}$. Depicted are the asymptotic formula (Eq. (2.2.9), dashed line) and an estimate based on $N = 10^8$ samples of a MC Cholesky Option Pricer (solid line) with 500 time steps. Model parameters are given by spot vol $\sigma_0 \approx 0.2557$, vvol $\eta = 0.2928$ and correlation parameter $\rho = -0.7571$.

where

$$\begin{aligned}G(f) &= \int_0^1 \sigma\left(\left(K\dot{f}\right)(s)\right)\dot{f}_s ds \equiv \left\langle \sigma\left(K\dot{f}\right), \dot{f}\right\rangle \equiv \left\langle \sigma(\widehat{f}), \dot{f}\right\rangle \\ F(f) &= \int_0^1 \sigma\left(\left(K\dot{f}\right)(s)\right)^2 ds \equiv \left\langle \sigma^2\left(K\dot{f}\right), 1\right\rangle \equiv \left\langle \sigma^2(\widehat{f}), 1\right\rangle \\ E(f) &= \int_0^1 \left|\dot{f}(s)\right|^2 ds \equiv \left\langle \dot{f}, \dot{f}\right\rangle\end{aligned}$$

The rest of this section is devoted to analysis of the function I as defined in (2.4.1). First, we derive the first order optimality condition for the above minimization problem.

Proposition 2.4.2 (First order optimality condition)**.** *For any $x \in \mathbb{R}$ we have at any local minimizer $f = f^x$ of the functional $\mathcal{I}_x$ in* (2.4.1) *that*

$$f_t^x = \frac{\rho\left(x-\rho G\left(f^x\right)\right)\left\{\left\langle \sigma\left(K\dot{f}^x\right), 1_{[0,t]}\right\rangle + \left\langle \sigma'\left(K\dot{f}^x\right)\dot{f}^x, K1_{[0,t]}\right\rangle\right\}}{\overline{\rho}^2 F\left(f^x\right)} + \frac{\left(x-\rho G\left(f^x\right)\right)^2}{\overline{\rho}^2 F^2\left(f^x\right)}\left\langle (\sigma\sigma')\left(K\dot{f}^x\right), K1_{[0,t]}\right\rangle, \quad (2.4.2)$$

for all $t \in [0,1]$.

Proof. We denote $a \approx b$ whenever $a = b + o(\delta)$ for a small parameter δ. We expand

$$\begin{aligned}E(f+\delta g) &\approx E(f) + 2\delta\left\langle \dot{f}, \dot{g}\right\rangle \\ F(f+\delta g) &\approx F(f) + \delta\left\langle \left(\sigma^2\right)'\left(K\dot{f}\right), K\dot{g}\right\rangle \\ G(f+\delta g) &\approx G(f) + \delta\left\{\left\langle \sigma\left(K\dot{f}\right), \dot{g}\right\rangle + \left\langle \sigma'\left(K\dot{f}\right)\dot{f}, K\dot{g}\right\rangle\right\}\end{aligned}$$

If $f = f^x$ is a minimizer then $\delta \mapsto \mathcal{I}_x(f+\delta g)$ has a minimum at $\delta = 0$ for all

g. We expand

$$
\begin{aligned}
\mathcal{I}_x\left(f+\delta g\right) &= \frac{\left(x-\rho G\left(f+\delta g\right)\right)^2}{2\overline{\rho}^2 F\left(f+\delta g\right)} + \frac{1}{2}E(f+\delta g) \\
&\approx \frac{\left(x-\rho G\left(f\right)-\delta\rho\left\{\left\langle\sigma\left(K\dot{f}\right),\dot{g}\right\rangle+\left\langle\sigma'\left(K\dot{f}\right)\dot{f},K\dot{g}\right\rangle\right\}\right)^2}{2\overline{\rho}^2\left[F\left(f\right)+\delta\left\langle\left(\sigma^2\right)'\left(K\dot{f}\right),K\dot{g}\right\rangle\right]} \\
&\quad+\frac{1}{2}E(f)+\delta\left\langle\dot{f},\dot{g}\right\rangle \\
&\approx \frac{\left(x-\rho G\left(f\right)\right)^2-\delta 2\rho\left(x-\rho G\left(f\right)\right)\left\{\left\langle\sigma\left(K\dot{f}\right),\dot{g}\right\rangle+\left\langle\sigma'\left(K\dot{f}\right)\dot{f},K\dot{g}\right\rangle\right\}}{2\overline{\rho}^2 F\left(f\right)\left[1+\frac{\delta}{F(f)}\left\langle\left(\sigma^2\right)'\left(K\dot{f}\right),K\dot{g}\right\rangle\right]} \\
&\quad+\frac{1}{2}E\left(f\right)+\delta\left\langle\dot{f},\dot{g}\right\rangle \\
&\approx \frac{\left(x-\rho G\left(f\right)\right)^2-\delta 2\rho\left(x-\rho G\left(f\right)\right)\left\{\left\langle\sigma\left(K\dot{f}\right),\dot{g}\right\rangle+\left\langle\sigma'\left(K\dot{f}\right)\dot{f},K\dot{g}\right\rangle\right\}}{2\overline{\rho}^2 F\left(f\right)} \\
&\quad-\frac{\left(x-\rho G\left(f\right)\right)^2}{2\overline{\rho}^2 F\left(f\right)}\frac{\delta}{F\left(f\right)}\left\langle\left(\sigma^2\right)'\left(K\dot{f}\right),K\dot{g}\right\rangle+\frac{1}{2}E\left(f\right)+\delta\left\langle\dot{f},\dot{g}\right\rangle.
\end{aligned}
$$

As a consequence, we must have, for $f=f^x$ and every $\dot{g}\in L^2\left[0,1\right]$

$$
\begin{aligned}
0 &= \frac{d}{d\delta}\left\{\mathcal{I}_x\left(f+\delta g\right)\right\}_{\delta=0} = -\frac{\rho\left(x-\rho G\left(f\right)\right)\left\{\left\langle\sigma\left(K\dot{f}\right),\dot{g}\right\rangle+\left\langle\sigma'\left(K\dot{f}\right)\dot{f},K\dot{g}\right\rangle\right\}}{\overline{\rho}^2 F\left(f\right)} \\
&\quad-\frac{\left(x-\rho G\left(f\right)\right)^2}{\overline{\rho}^2 F^2\left(f\right)}\left\langle\left(\sigma\sigma'\right)\left(K\dot{f}\right),K\dot{g}\right\rangle+\left\langle\dot{f},\dot{g}\right\rangle.
\end{aligned}
$$

Recall $f_0^x=0$, any x. We now test with $\dot{g}=1_{[0,t]}$ for a fixed $t\in[0,1]$ and obtain

$$
\begin{aligned}
f_t^x &= \frac{\rho\left(x-\rho G\left(f^x\right)\right)\left\{\left\langle\sigma\left(K\dot{f}^x\right),1_{[0,t]}\right\rangle+\left\langle\sigma'\left(K\dot{f}^x\right)\dot{f}^x,K\mathbf{1}_{[0,t]}\right\rangle\right\}}{\overline{\rho}^2 F\left(f^x\right)} \\
&\quad+\frac{\left(x-\rho G\left(f^x\right)\right)^2}{\overline{\rho}^2 F^2\left(f^x\right)}\left\langle\left(\sigma\sigma'\right)\left(K\dot{f}^x\right),K\mathbf{1}_{[0,t]}\right\rangle.
\end{aligned}
$$

□

2.4.1 Smoothness of the energy

Having formally identified the first order condition for minimality in (2.4.1), we will now show that the energy $x\mapsto I(x)$ is a smooth function. More precisely, we will use the implicit function theorem to show that the minimizing configuration f^x is a smooth function in x (locally at $x=0$). As $\mathcal{I}_x$ is a smooth function, too, this will imply smoothness of $x \mapsto \mathcal{I}_x(f^x) = I(x)$, at least in a neighborhood of 0.

As the Cameron-Martin space $\mathcal{H}$ of the process $\widehat{B}$ continuously embeds into $C([0,1])$, K maps H_0^1 continuously into $C([0,1])$, i.e., there is a constant $C > 0$ such that for any $f \in H_0^1$ we have

$$\|K\dot{f}\|_\infty \leq C\|f\|_{H_0^1}. \tag{2.4.3}$$

This result will follow from

Lemma 2.4.3. *Let $(V_t : 0 \leq t \leq 1)$ be a continuous, centred Gaussian process and $\mathcal{H}$ its Cameron-Martin space. Then we have the continuous embedding $\mathcal{H} \hookrightarrow C[0,1]$. That is, for some constant C,*

$$\|h\|_\infty \leq C\|h\|_{\mathcal{H}}.$$

Proof. By a fundamental result of Fernique, applied to the law of V as Gaussian measure on the Banach space $(C[0,1], \|\cdot\|_\infty)$, the random variable $\|V\|_\infty$ has Gaussian integrability. In particular,

$$\sigma^2 := \mathbb{E}(\|V\|_\infty^2) < \infty,$$

On the other hand, a generic element $h \in \mathcal{H}$ can be written as $h_t = E[V_t Z]$ where Z is a centred Gaussian random variable with variance $\|h\|_{\mathcal{H}}^2$, see, e.g., (Friz & Hairer, 2014, page 150). By Cauchy–Schwarz,

$$|h_t| \leq E[|V_t|]^{1/2}\|h\|_{\mathcal{H}} \leq \sigma\|h\|_{\mathcal{H}}$$

and conclude by taking the sup over on the l.h.s. over $t \in [0,1]$. □

Remark 2.4.4. Assume V is of Volterra form, i.e. $V_t = \int_0^t K(t,s)\,dB_s$. Then it can be shown (see (Decreusefond, 2005, Section 3)) that $\mathcal{H}$ is the image of L^2 under the map

$$K : \dot{f} \mapsto \widehat{f} := \left(t \mapsto \int_0^t K(t,s)\,\dot{f}_s ds\right)$$

and $\left\|K\dot{f}\right\|_{\mathcal{H}} = \left\|\dot{f}\right\|_{L^2}$. In particular then, applying the above with $h = K\dot{f} \in$

$\mathcal{H}$, gives

$$\left\|K\dot{f}\right\|_{\infty} \leq C\left\|K\dot{f}\right\|_{\mathcal{H}} = C\left\|\dot{f}\right\|_{L^2} = C\|f\|_{H_0^1}.$$

2.4.1.1 The uncorrelated case

We start with the case $\rho = 0$ as the formulas are much simpler in this case.

By Proposition 2.4.2, any local optimizer $f = f^x$ of the functional $\mathcal{I}_x : H_0^1 \to \mathbb{R}$ in the uncorrelated case $\rho = 0$ satisfies for any $t \in [0,1]$

$$f_t = \frac{x^2}{F^2(f)} \left\langle (\sigma\sigma')\left(K\dot{f}\right), K\mathbf{1}_{[0,t]} \right\rangle.$$

We define a map $H : H_0^1 \times \mathbb{R} \to H_0^1$ by

$$H(f,x)(t) \coloneqq f_t - \frac{x^2}{F^2(f)} \left\langle (\sigma\sigma')\left(K\dot{f}\right), K\mathbf{1}_{[0,t]} \right\rangle. \tag{2.4.4}$$

Hence, for given $x \in \mathbb{R}$, any local optimizer f must solve $H(f,x) = 0$. As one particular solution is given by the pair $(0,0)$, we are in the realm of the implicit function theorem. We need to prove that

- $(f,x) \mapsto H(f,x)$ is locally smooth (in the sense of Fréchet);
- $DH(f,x) \coloneqq \frac{\partial}{\partial f} H(f,x)$ is invertible in $(0,0)$.

Note that invertibility should hold for x small enough, as $DH(f,x) = \mathrm{id}_{H_0^1} - x^2 R$ for some R, which is invertible as long as R has a bounded norm for sufficiently small x.

Remark 2.4.5. The method of proof in this section is purely local in H_0^1. Hence, we only really need smoothness of σ locally around 0. Note, however, that stochastic Taylor expansions used in Section 2.5 will actually require global smoothness of σ.

Lemma 2.4.6. *The functions $F : H_0^1 \to \mathbb{R}$ and $R_1 : H_0^1 \to C([0,1])$ defined by*

$$R_1(f)(t) \coloneqq \left\langle (\sigma\sigma')\left(K\dot{f}\right), K\mathbf{1}_{[0,t]} \right\rangle, \quad t \in [0,1],$$

are smooth in the sense of Fréchet.

Proof. For $N \geq 1$ we note that the Gateaux derivative of F satisfies

$$D^N F(f) \cdot (g_1, \ldots, g_N) = \int_0^1 \frac{d^N}{dx^N} \sigma^2(K\dot{f}) K\dot{g}_1 \cdots K\dot{g}_N ds.$$

By Lemma 2.4.3, we can bound

$$\begin{aligned} |D^N F(f) \cdot (g_1, \ldots, g_N)| &\leq \text{const} \int_0^1 |K\dot{g}_1(s)| \cdots |K\dot{g_N}(s)| ds \\ &\leq \text{const } \|K\dot{g}_1\|_\infty \cdots \|K\dot{g}_N\|_\infty \\ &\leq \text{const } C^N \|g_1\|_{H_0^1} \cdots \|g_N\|_{H_0^1}, \end{aligned}$$

for const $= \|\frac{d^n}{dx^n}\sigma^2\|_\infty$.[4] Thus, $D^N F(f)$ is a multi-linear form on H_0^1 with operator norm $\|D^N F(f)\| \leq \|\frac{d^n}{dx^n}\sigma^2\|_\infty C^N$ independent of f. As $f \mapsto D^N F(f)$ is continuous, we conclude that $D^N F(f)$ as given above is, in fact, a Fréchet derivative.

Let us next consider the functional R_1. Note that

$$\left(D^N R_1(f) \cdot (g_1, \ldots, g_N)\right)(t) = \left\langle \mathfrak{s}_N(K\dot{f}) K\dot{g}_1 \cdots K\dot{g}_N, K\mathbf{1}_{[0,t]} \right\rangle$$

for $\mathfrak{s}_N(x) \coloneqq \frac{d^N}{dx^N}\sigma(x)\sigma'(x)$. Hence, Assumption 2.1.5 implies that

$$\begin{aligned} \|D^N R_1(f) \cdot (g_1, \ldots, g_N)\|_{H_0^1}^2 &= \int_0^1 \left(\int_t^1 \mathfrak{s}_N\left((K\dot{f})(s)\right) \prod_{i=1}^N (K\dot{g}_i)(s) K(s,t) ds \right)^2 dt \\ &\leq \|\mathfrak{s}_N\|_\infty^2 \prod_{i=1}^N \|K\dot{g}_i\|_\infty^2 \int_0^1 \int_t^1 K(s,t)^2 ds dt \\ &\leq \|\mathfrak{s}_N\|_\infty^2 \; C^{2N} \prod_{i=1}^N \|g_i\|_{H_0^1}^2 \int_0^1 \int_0^s K(s,t)^2 dt ds \\ &\leq \|\mathfrak{s}_N\|_\infty^2 \; C^{2N} \int_0^1 \int_0^s K(s,t)^2 dt ds \prod_{i=1}^N \|g_i\|_{H_0^1}^2. \end{aligned}$$

We see that the multi-linear map $D^N R_1(f)$ has operator norm bounded by

$$\|D^N R_1(f)\| \leq \|\mathfrak{s}_N\|_\infty \; C^N \sqrt{\int_0^1 \int_0^s K(s,t)^2 dt ds},$$

[4] More precisely, since neither σ nor its derivatives need to be bounded, we need to actually work with a local version of the above estimate, for instance by replacing the max with a sup over a compact set containing $\{(K\dot{f})(t) : 0 \leq t \leq 1\}$.

independent of f. From continuity of $f \mapsto D^N R_1(f)$, it follows that $D^N R_1(f)$ is the N'th Fréchet derivative. □

Theorem 2.4.7 (Zero correlation)**.** *Assuming $\rho = 0$, the energy $I(x)$ (as defined in* (2.4.1)*) is smooth in a neighborhood of $x = 0$.*

Proof. By construction, we have

$$DH(f, x) = \mathrm{id}_{H_0^1} - x^2 A(f)$$

for $A : H_0^1 \to \mathcal{L}(H_0^1, H_0^1)$ defined by

$$A(f) := R_1(f) \otimes DF^{-2}(f) + F^{-2}(f) DR_1(f).$$

Here,

$$\Big(R_1(f) \otimes DF^{-2}(f)\Big) \cdot g = \underbrace{\big(DF^{-2}(f) \cdot g\big)}_{\in \mathbb{R}} \underbrace{R_1(f)}_{\in H_0^1}.$$

As verified above, H is smooth in the sense of Fréchet. Trivially, $DH(0,0) = \mathrm{id}_{H_0^1}$ is invertible and $H(0,0) = 0$. Therefore, the implicit function theorem implies that there are open neighborhoods U and V of $0 \in H_0^1$ and $0 \in \mathbb{R}$, respectively, and a smooth map $x \mapsto f^x$ from V to U such that $H(f^x, x) \equiv 0$ and f^x is unique in U with this property.

For the energy, we prove that $I(x) = \mathcal{I}_x(f^x)$ in a neighborhood of $x = 0$. First of all, we show that a minimizer exists. If not, there is a function $g \in H_0^1$ with $\mathcal{I}_x(g) < \mathcal{I}_x(f^x)$. For small enough x such a g must be inside a ball with radius ε around $0 \in H_0^1$, as $\mathcal{I}_x(g) \geq \frac{1}{2}\|g\|_{H_0^1}^2$ and $\lim_{x\to 0} \mathcal{I}_x(f^x) = 0$. Then note that for any $g \in H_0^1$

$$D^2\mathcal{I}_0(0) \cdot (g, g) = \|g\|_{H_0^1}^2 > 0,$$

where $D^2\mathcal{I}_x(f)$ denotes the second derivative of $f \mapsto \mathcal{I}_x(f)$. By continuity, $D^2\mathcal{I}_x(f)$ stays positive definite for (x, f) in a neighborhood of $(0,0)$. As noted, for x small enough, both g and f^x (and the line connecting them) lie in this neighborhood. For $h := g - f^x$, this implies

$$\mathcal{I}_x(g) - \mathcal{I}_x(f^x) = D\mathcal{I}_x(f_x) \cdot h + \int_0^1 D^2\mathcal{I}_x(f^x + th) \cdot (h, h)\, dt > 0,$$

since $D\mathcal{I}_x(f_x) \cdot h = 0$ and $D^2\mathcal{I}_x(f^x + tsh) \cdot (h,h) > 0$. This contradicts the assumption that $\mathcal{I}_x(g) < \mathcal{I}_x(f^x)$, and we conclude that f^x is, indeed, a minimizer of $\mathcal{I}_x$, implying that $I(x) = \mathcal{I}_x(f^x)$ locally.

Finally, as $x \mapsto f^x$ is smooth and $(f,x) \mapsto \mathcal{I}_x(f) = \frac{x^2}{2F(f)} + \frac{1}{2}\|f\|^2_{H_0^1}$ is smooth, we see that $x \mapsto I(x) = \mathcal{I}_x(f^x)$ is smooth in a neighborhood of 0. (Note that this arguments relies on $\sigma(0) \neq 0$, implying that $F(f) \neq 0$ for f in a neighborhood to 0.) □

Remark 2.4.8. Classical counter-examples in the context of the *direct method* of calculus of variations show that the step of verifying the existence of a minimizer should not be taken too lightly. For instance, the functional

$$J(u) \coloneqq \int_0^1 \left[(u'(s)^2 - 1)^2 + u(s)^2 \right] ds$$

does not have a minimizer in H_0^1, but J can be made arbitrarily close to 0 by choosing piecewise-linear functions u with slope $|u'| = 1$ oscillating around 0. We refer to any text book on calculus of variations. In the situation above, local "convexity" in the sense of a positive definite second derivative prevents this phenomenon. An alternative method of proof for the existence of a minimizer is to show that J is (lower semi-) continuous in the weak sense.

2.4.1.2 The general case

In the general case (cf. Proposition 2.4.2), we define the function $H : H_0^1 \times \mathbb{R} \to H_0^1$ by

$$\begin{aligned} H(f,x)(t) &\coloneqq f_t - \frac{\rho\left(x - \rho G\left(f\right)\right)\left\{\left\langle \sigma\left(K\dot{f}\right), \mathbf{1}_{[0,t]}\right\rangle + \left\langle \sigma'\left(K\dot{f}\right)\dot{f}, K\mathbf{1}_{[0,t]}\right\rangle\right\}}{\overline{\rho}^2 F\left(f\right)} \\ &\quad + \frac{\left(x - \rho G\left(f\right)\right)^2}{\overline{\rho}^2 F^2\left(f\right)} \left\langle (\sigma\sigma')\left(K\dot{f}\right), K\mathbf{1}_{[0,t]}\right\rangle \\ &= f_t - \frac{\rho\left(x - \rho G(f)\right)}{\overline{\rho}^2 F(f)} \left(R_2(f)(t) + R_3(f)(t)\right) + \frac{\left(x - \rho G(f)\right)^2}{\overline{\rho}^2 F(f)^2} R_1(f)(t), \end{aligned} \tag{2.4.5}$$

where $R_2, R_3 : H^1_0 \to H^1_0$ are defined by

$$R_2(f)(t) := \left\langle \sigma(K\dot{f}), \mathbf{1}_{[0,t]} \right\rangle, \tag{2.4.6}$$

$$R_3(f)(t) := \left\langle \sigma'(K\dot{f})\dot{f}, K\mathbf{1}_{[0,t]} \right\rangle, \tag{2.4.7}$$

$t \in [0, 1]$.

One easily checks that G, R_2, R_3 are smooth in the Fréchet sense.

Lemma 2.4.9. *The functions $G : H^1_0 \to \mathbb{R}$, $R_2 : H^1_0 \to H^1_0$ and $R_3 : H^1_0 \to H^1_0$ are smooth in Fréchet sense.*

Proof. The proof of smoothness is clear. We report the actual derivatives. For G we get

$$D^N G(f) \cdot (g_1, \dots, g_N) = \left\langle \sigma^{(N)}\left(K\dot{f}\right)\dot{f}, \prod_{i=1}^N K\dot{g}_i \right\rangle + \\ + \sum_{k=1}^N \left\langle \sigma^{(N-1)}\left(K\dot{f}\right), \dot{g}_k \prod_{i \neq k} K\dot{g}_i \right\rangle.$$

For R_2 and, respectively, R_3, we obtain

$$\left(D^N R_2(f) \cdot (g_1, \dots, g_N)\right)(t) = \int_0^t \sigma^{(N)}\left((K\dot{f})(s)\right) \prod_{i=1}^N (K\dot{g}_i)(s) ds,$$

and

$$\left(D^N R_3(f) \cdot (g_1, \dots, g_N)\right)(t) = \left\langle \sigma^{(N+1)}\left(K\dot{f}\right)\dot{f}K\mathbf{1}_{[0,t]}, \prod_{i=1}^N K\dot{g}_i \right\rangle + \\ + \sum_{k=1}^N \left\langle \sigma^{(N)}\left(K\dot{f}\right)K\mathbf{1}_{[0,t]}, \dot{g}_k \prod_{i \neq k} K\dot{g}_i \right\rangle. \quad \square$$

Theorem 2.4.10. *Let σ be smooth with $\sigma(0) \neq 0$. Then the energy $I(x)$ as defined in (2.4.1) is smooth in a neighborhood of $x = 0$.*

Proof. The proof is similar to the proof of Theorem 2.4.7. In fact, the only difference is in establishing invertibility of $DH(0, 0)$ and the existence of a minimizer.

Note that (2.4.5) contains three terms. The derivative of the first term

$(f \mapsto f)$ is always equal to $\mathrm{id}_{H_0^1}$. For the second term, we note that

$$(x - \rho G(f))\big|_{x=0,\, f=0} = 0.$$

Hence, the only non-vanishing contribution to the derivative of the second term evaluated in direction $g \in H_0^1$ at $x = 0$, $f = 0$ and $t \in [0,1]$ is

$$\frac{\rho^2 DG(0) \cdot g}{\overline{\rho}^2 F(0)} (R_2(0) + R_3(0)) = \frac{\rho^2 \sigma_0 g(1)}{\overline{\rho}^2 \sigma_0^2} (\sigma_0 t + 0) = \frac{\rho^2}{\overline{\rho}^2} g(1) t.$$

For the same reason, the derivative of the third term at $(f, x) = (0, 0)$ vanishes entirely. Hence,

$$(DH(0,0) \cdot g)(t) = g(t) + \frac{\rho^2}{\overline{\rho}^2} g(1) t.$$

It is easy to see that $g \mapsto DH(0,0) \cdot g$ is invertible. Indeed, let us construct the pre-image $g = DH(0,0)^{-1} \cdot h$ of some $h \in H_0^1$. At $t = 1$ we have

$$\frac{\overline{\rho}^2 + \rho^2}{\overline{\rho}^2} g(1) = h(1),$$

implying $g(1) = \overline{\rho}^2 h(1)$. For $0 \le t < 1$, we then get

$$g(t) + \frac{\rho^2}{\overline{\rho}^2} g(1) t = g(t) + \frac{\rho^2}{\overline{\rho}^2} \overline{\rho}^2 h(1) t = g(t) + \rho^2 h(1) t = h(t),$$

or $g(t) = h(t) - \rho^2 h(1) t$.

For existence of the minimizer, note that

$$D^2 J_0(0) \cdot (g, g) = \frac{\rho^2}{\overline{\rho}^2} g(1)^2 + \|g\|_{H_0^1}^2,$$

which is again positive definite. □

Remark 2.4.11. Though only formulated in terms of "smoothness", it is easy to show that $\sigma \in C^k$ implies that $I \in C^{k-1}$ (locally at 0).

2.4.2 Energy expansion

Having established smoothness of the energy I as well as of the minimizing configuration $x \mapsto f^x$ locally around $x = 0$, we can proceed with computing the Taylor expansion of f^x around $x = 0$. We will once more rely on the first order optimality condition given in Proposition 2.4.2. Plugging the Taylor expansion of f^x into $\mathcal{I}_x$ will then give us the local Taylor expansion of $I(x)$.

2.4.2.1 Expansion of the minimizing configuration

Theorem 2.4.12. *We have*

$$\begin{aligned} f_t^x &= \alpha_t x + \beta_t \frac{x^2}{2} + \mathcal{O}\left(x^3\right), \\ \alpha_t &= \frac{\rho}{\sigma_0} t, \\ \beta_t &= 2\frac{\sigma_0'}{\sigma_0^3}\left[\rho^2 \left\langle K1, \mathbf{1}_{[0,t]}\right\rangle + \left\langle K\mathbf{1}_{[0,t]}, 1\right\rangle - 3\rho^2 t \left\langle K1, 1\right\rangle\right]. \end{aligned}$$

Remark 2.4.13 (Non-Markovian transversality). In the RL-fBM case, $K(t,s) = \sqrt{2H}\,|t-s|^\gamma$ with $\gamma = H - 1/2$ one computes

$$\left\langle 1, K1_{[0,t]}\right\rangle = \frac{1}{(1+\gamma)(2+\gamma)}\left\{1 - (1-t)^{2+\gamma}\right\} \in C^1[0,1].$$

Interestingly, the transversality condition known from the Markovian setting ($q_1 = 0$, which readily translates to $\dot{f}_1^x = 0$ there) remains valid here (for $\rho = 0$), at least to order x^2, in the sense that

$$\dot{f}_t^x \approx \beta_t \frac{x^2}{2} = (\text{const})\,(1-t)^{1+\gamma}\,|_{t=1} = 0$$

Proof of Theorem 2.4.12. **First order expansion:**

Up to the order needed in order to get the first order term, we have

$$\begin{aligned}
f_t^x &= \alpha_t x + \mathcal{O}(x^2),\\
\dot{f}_t^x &= \dot{\alpha}_t x + \mathcal{O}(x^2),\\
\sigma(K\dot{f}^x) &= \sigma_0 + \sigma_0' K\dot{\alpha}\ x + \mathcal{O}(x^2),\\
\sigma'(K\dot{f}^x) &= \sigma_0' + \sigma_0'' K\dot{\alpha}\ x + \mathcal{O}(x^2),\\
F(f^x) &= \langle \sigma^2(K\dot{f}^x), 1\rangle\\
&= \sigma_0^2 + \mathcal{O}(x),\\
G(f^x) &= \langle \sigma(K\dot{f}^x), \dot{f}^x\rangle\\
&= \langle \sigma_0, \dot{\alpha}\rangle\, x + \mathcal{O}(x^2).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\langle \sigma(K\dot{f}^x), \mathbf{1}_{[0,t]}\rangle &= \sigma_0 t + \mathcal{O}(x),\\
\langle \sigma'(K\dot{f}^x)\dot{f}^x, K\mathbf{1}_{[0,t]}\rangle &= \mathcal{O}(x),\\
\langle \sigma\sigma'(K\dot{f}^x), K\mathbf{1}_{[0,t]}\rangle &= \mathcal{O}(1),\\
x - \rho G(f^x) &= (1 - \rho\sigma_0\alpha_1)x + \mathcal{O}(x^2),\\
(x - \rho G(f^x))^2 &= \mathcal{O}(x^2).
\end{aligned}$$

This yields for the first order term in (2.4.2)

$$\alpha_t = \frac{\rho(1 - \rho\sigma_0\alpha_1)}{\overline{\rho}^2\sigma_0} t.$$

Setting $t = 1$, we get

$$\alpha_1 = \frac{\rho}{\overline{\rho}^2\sigma_0} - \frac{\rho^2}{\overline{\rho}^2}\alpha_1,$$

which is solved by $\alpha_1 = \frac{\rho}{\sigma_0}$. Inserting this term back into the equation for α_t, we get

$$\alpha_t = \frac{\rho}{\sigma_0} t. \tag{2.4.8}$$

Second order expansion:

Using (2.4.8) and the ansatz $f_t^x = \alpha_t x + \frac{1}{2}\beta_t x^2 + \mathcal{O}(x^3)$, we re-compute the relevant terms appearing in the (2.4.2). We have

$$\sigma(K\dot{f}^x(s)) = \sigma_0 + \sigma_0' \frac{\rho}{\sigma_0}(K1)(s)x + \mathcal{O}(x^2)$$

and analogously for σ replaced by σ', $\sigma\sigma'$. This implies

$$\begin{aligned}\left\langle \sigma(K\dot{f}^x), \mathbf{1}_{[0,t]} \right\rangle &= \sigma_0 t + \sigma_0' \frac{\rho}{\sigma_0} \left\langle K1, \mathbf{1}_{[0,t]} \right\rangle x + \mathcal{O}(x^2),\\ \left\langle \sigma'(K\dot{f}^x)\dot{f}^x, K\mathbf{1}_{[0,t]} \right\rangle &= \rho \frac{\sigma'}{\sigma_0} \left\langle K\mathbf{1}_{[0,t]}, 1 \right\rangle x + \mathcal{O}(x^2),\\ \left\langle \sigma\sigma'(K\dot{f}^x), K\mathbf{1}_{[0,t]} \right\rangle &= \sigma_0\sigma_0' \left\langle K\mathbf{1}_{[0,t]}, 1 \right\rangle + \mathcal{O}(x).\end{aligned}$$

Using the notation introduced earlier, we have

$$\begin{aligned}F(f^x) &= \sigma_0^2 + 2\sigma_0'\rho \langle K1, 1 \rangle x + \mathcal{O}(x^2),\\ G(f^x) &= \rho x + \left(\frac{1}{2}\sigma_0\beta_1 + \rho^2 \frac{\sigma_0'}{\sigma_0^2} \langle K1, 1 \rangle \right) x^2 + \mathcal{O}(x^3).\end{aligned}$$

This directly implies

$$\begin{aligned}x - \rho G(f^x) &= \overline{\rho}^2 x - \rho \left(\frac{1}{2}\sigma_0\beta_1 + \rho^2 \frac{\sigma_0'}{\sigma_0^2} \langle K1, 1 \rangle \right) x^2 + \mathcal{O}(x^3),\\ (x - \rho G(f^x))^2 &= \overline{\rho}^4 x^2 - 2\overline{\rho}^2 \rho \left(\frac{1}{2}\sigma_0\beta_1 + \rho^2 \frac{\sigma_0'}{\sigma_0^2} \langle K1, 1 \rangle \right) x^3 + \mathcal{O}(x^4).\end{aligned}$$

We next compute some auxiliary terms appearing in (2.4.2).

$$\begin{aligned}N_1 &:= \rho(x - \rho G(f^x)) \left(\left\langle \sigma(K\dot{f}^x), \mathbf{1}_{[0,t]} \right\rangle + \left\langle \sigma'(K\dot{f}^x)\dot{f}^x, K\mathbf{1}_{[0,t]} \right\rangle \right)\\ &= \rho\overline{\rho}^2\sigma_0 t x + \Bigg[\rho^2\overline{\rho}^2 \frac{\sigma_0'}{\sigma_0} \left(\left\langle K1, \mathbf{1}_{[0,t]} \right\rangle + \left\langle K\mathbf{1}_{[0,t]}, 1 \right\rangle \right)\\ &\quad - \rho^4 \frac{\sigma_0'}{\sigma_0} t \langle K1, 1 \rangle - \frac{1}{2}\rho^2\sigma_0^2 t\beta_1 \Bigg] x^2 + \mathcal{O}(x^3)\end{aligned}$$

The corresponding denominator is $\overline{\rho}^2 F(f^x)$. Using the formula

$$\frac{a_1 x + a_2 x^2 + \mathcal{O}(x^3)}{b_0 + b_1 x + \mathcal{O}(x^2)} = \frac{a_1}{b_0} x + \frac{a_2 b_0 - a_1 b_1}{b_0^2} x^2 + \mathcal{O}(x^3),$$

we obtain

$$\frac{N_1}{\overline{\rho}^2 F(f^x)} = \frac{\rho}{\sigma_0} t x + \left[\rho^2 \frac{\sigma_0'}{\sigma_0^3} \left(\left\langle K1, \mathbf{1}_{[0,t]} \right\rangle + \left\langle K\mathbf{1}_{[0,t]}, 1 \right\rangle \right) \right.$$
$$\left. - \left(\frac{\rho^4}{\overline{\rho}^2} + 2\rho^2 \right) \frac{\sigma_0'}{\sigma_0^3} t \left\langle K1, 1 \right\rangle - \frac{1}{2} \frac{\rho^2}{\overline{\rho}^2} \beta_1 t \right] x^2 + \mathcal{O}(x^3) \quad (2.4.9)$$

For the second term in (2.4.2), let

$$N_2 \coloneqq (x - \rho G(f^x))^2 \left\langle (\sigma\sigma')(K\dot{f}^x), K\mathbf{1}_{[0,t]} \right\rangle = \overline{\rho}^4 \sigma_0 \sigma_0' \left\langle K\mathbf{1}_{[0,t]}, 1 \right\rangle x^2 + \mathcal{O}(x^3).$$

The corresponding denominator is $\overline{\rho}^2 F(f^x)^2 = \overline{\rho}^2 \sigma_0^4 + \mathcal{O}(x)$. Hence,

$$\frac{N_2}{\overline{\rho}^2 F(f^x)^2} = \overline{\rho}^2 \frac{\sigma_0'}{\sigma_0^3} \left\langle K\mathbf{1}_{[0,t]}, 1 \right\rangle x^2 + \mathcal{O}(x^3). \quad (2.4.10)$$

Combining (2.4.9) and (2.4.10), we get

$$f_t^x = \frac{\rho}{\sigma_0} t x + \left[\rho^2 \frac{\sigma_0'}{\sigma_0^3} \left(\left\langle K1, \mathbf{1}_{[0,t]} \right\rangle + \left\langle K\mathbf{1}_{[0,t]}, 1 \right\rangle \right) - \frac{\rho^4}{\overline{\rho}^2} \frac{\sigma_0'}{\sigma_0^3} t \left\langle K1, 1 \right\rangle \right.$$
$$\left. - \frac{1}{2} \frac{\rho^2}{\overline{\rho}^2} \beta_1 t - 2\rho^2 \frac{\sigma_0'}{\sigma_0^3} t \left\langle K1, 1 \right\rangle + \overline{\rho}^2 \frac{\sigma_0'}{\sigma_0^3} \left\langle K\mathbf{1}_{[0,t]}, 1 \right\rangle \right] x^2 + \mathcal{O}(x^3)$$

We shall next compute β_1. Taking the second order terms on both sides and letting $t = 1$, we obtain

$$\frac{1}{2}\beta_1 = \rho^2 \frac{\sigma_0'}{\sigma_0^3} 2 \left\langle K1, 1 \right\rangle - \frac{\rho^4}{\overline{\rho}^2} \frac{\sigma_0'}{\sigma_0^3} \left\langle K1, 1 \right\rangle$$
$$- \frac{1}{2} \frac{\rho^2}{\overline{\rho}^2} \beta_1 - 2\rho^2 \frac{\sigma_0'}{\sigma_0^3} \left\langle K1, 1 \right\rangle + \overline{\rho}^2 \frac{\sigma_0'}{\sigma_0^3} \left\langle K1, 1 \right\rangle .$$

Moving β_1 to the other side with $1 + \frac{\rho^2}{\overline{\rho}^2} = \frac{1}{\overline{\rho}^2}$ and collecting terms on the right hand side, we arrive at

$$\frac{1}{2} \frac{1}{\overline{\rho}^2} \beta_1 = \frac{\sigma_0'}{\sigma_0^3} \left\langle K1, 1 \right\rangle \left(2\rho^2 - \frac{\rho^4}{\overline{\rho}^2} - 2\rho^2 + \overline{\rho}^2 \right) = \frac{1 - 2\rho^2}{\overline{\rho}^2} \frac{\sigma_0'}{\sigma_0^3} \left\langle K1, 1 \right\rangle$$

We conclude that

$$\beta_1 = 2(1 - 2\rho^2) \frac{\sigma_0'}{\sigma_0^3} \left\langle K1, 1 \right\rangle$$

Hence, we obtain

$$\beta_t = 2\frac{\sigma_0'}{\sigma_0^3}\left[\rho^2\left\langle K1, \mathbf{1}_{[0,t]}\right\rangle + \left\langle K\mathbf{1}_{[0,t]}, 1\right\rangle - 3\rho^2 t\left\langle K1, 1\right\rangle\right]. \qquad \square$$

2.4.2.2 Energy expansion in the general case

Now we compute the Taylor expansion of $I(x)$ as defined in Proposition 2.4.1. We start with the second term. Plugging in the optimal path $f_t^x = \alpha_t x + \frac{1}{2}\beta_t x^2 + \mathcal{O}(x^3)$ (and using $\left\langle \dot\beta, 1\right\rangle = \beta_1$ as $\beta_0 = 0$) we obtain

$$\frac{1}{2}\left\langle \dot f^x, \dot f^x\right\rangle = \frac{1}{2}\frac{\rho^2}{\sigma_0^2}x^2 + \frac{1}{2}\frac{\rho}{\sigma_0}\beta_1 x^3 + \mathcal{O}(x^4).$$

Inserting $\beta_1 = 2(1-2\rho^2)\frac{\sigma_0'}{\sigma_0^3}\left\langle K1, 1\right\rangle$ into the above formula for $(x - \rho G(f^x))^2$, we get

$$(x - \rho G(f^x))^2 = \overline{\rho}^4 x^2 - 2\overline{\rho}^4\rho\frac{\sigma_0'}{\sigma_0^2}\left\langle K1, 1\right\rangle x^3 + \mathcal{O}(x^4).$$

Recall the denominator

$$2\overline{\rho}^2 F(f^x) = 2\overline{\rho}^2\sigma_0^2 + 4\overline{\rho}^2\sigma_0'\rho\left\langle K1, 1\right\rangle x + \mathcal{O}(x^2).$$

Using the expansion of a fraction

$$\frac{a_2x^2 + a_3x^3 + \mathcal{O}(x^4)}{b_0 + b_1 x + \mathcal{O}(x^2)} = \frac{a_2}{b_0}x^2 + \frac{a_3b_0 - a_2b_1}{b_0^2}x^3 + \mathcal{O}(x^4),$$

we obtain from

$$\begin{aligned}\frac{(x - \rho G(f^x))^2}{2\overline{\rho}^2 F(f^x)} &= \frac{\overline{\rho}^4}{2\overline{\rho}^2\sigma_0^2}x^2 + \\ &\quad + \frac{\left(-2\overline{\rho}^4\rho\frac{\sigma_0'}{\sigma_0^2}\left\langle K1, 1\right\rangle\right)2\overline{\rho}^2\sigma_0^2 - \overline{\rho}^4\left(4\overline{\rho}^2\sigma_0'\rho\left\langle K1, 1\right\rangle\right)}{4\overline{\rho}^4\sigma_0^4}x^3 + \mathcal{O}(x^4)\\ &= \frac{\overline{\rho}^2}{2\sigma_0^2}x^2 - 2\overline{\rho}^2\rho\frac{\sigma_0'}{\sigma_0^4}\left\langle K1, 1\right\rangle x^3 + \mathcal{O}(x^4).\end{aligned}$$

We note that

$$\frac{1}{2}\frac{\rho}{\sigma_0}\beta_1 - 2\overline{\rho}^2\rho\frac{\sigma_0'}{\sigma_0^4}\left\langle K1, 1\right\rangle = \left((1-2\rho^2) - 2(1-\rho^2)\right)\rho\frac{\sigma_0'}{\sigma_0^4}\left\langle K1, 1\right\rangle = -\rho\frac{\sigma_0'}{\sigma_0^4}\left\langle K1, 1\right\rangle.$$

Adding both terms, we arrive at the

Proposition 2.4.14. *The energy expansion to third order gives*

$$I(x) = \frac{1}{2\sigma_0^2}x^2 - \rho\frac{\sigma_0'}{\sigma_0^4}\left\langle K1, 1\right\rangle x^3 + \mathcal{O}(x^4).$$

2.4.2.3 Energy expansion for the Riemann-Liouville kernel

Let us specialize the energy expansion given in Proposition 2.4.14 for the Riemann-Liouville fBm. Choose $\gamma = H - \frac{1}{2}$ and recall that the kernel K takes the form $K(t,s) = (t-s)^\gamma$. We get

$$(K1)(t) = \int_0^t K(t,s)ds = \int_0^t (t-s)^\gamma ds = \frac{t^{1+\gamma}}{1+\gamma}.$$

The key term $\langle K1, 1\rangle$ appearing in the energy expansion now gives

$$\langle K1, 1\rangle = \int_0^1 (K1)(t)dt = \int_0^1 \frac{t^{1+\gamma}}{1+\gamma}dt = \frac{1}{(1+\gamma)(2+\gamma)} = \frac{1}{(H+1/2)(H+3/2)}.$$

Plugging formula (2.4.2.3) into the energy expansion, we obtain the energy expansion for the Riemann-Liouville fractional Browian motion

$$I(x) = \frac{1}{2\sigma_0^2}x^2 - \frac{\rho}{(H+1/2)(H+3/2)}\frac{\sigma_0'}{\sigma_0^4}x^3 + \mathcal{O}(x^4).$$

For completeness, let us also fully describe the time-dependence of the second order term β_t in the expansion of the optimal trajectory f_t^x. Unlike the first order time, here we do not have a linear movement any more. Indeed

$$\left\langle K1, \mathbf{1}_{[0,t]}\right\rangle = \int_0^t (K1)(s)ds = \int_0^t \frac{s^{1+\gamma}}{1+\gamma}ds = \frac{t^{2+\gamma}}{(1+\gamma)(2+\gamma)}, \tag{2.4.11}$$

$$\left\langle K\mathbf{1}_{[0,t]}, 1\right\rangle = \frac{1}{(1+\gamma)(2+\gamma)}\left(1-(1-t)^{2+\gamma}\right). \tag{2.4.12}$$

2.5 Proof of the pricing formula

Fix $x \geq 0$ and $\widehat{x} = \frac{\varepsilon}{\widehat{\varepsilon}} x$ where $\varepsilon = t^{1/2}$ and $\widehat{\varepsilon} = t^H = \varepsilon^{2H}$. We have

$$\begin{aligned} c(\widehat{x}, t) &= E\left(\exp\left(X_t\right) - \exp \widehat{x}\right)^+ \\ &= E\left(\exp\left(X_1^\varepsilon\right) - \exp \widehat{x}\right)^+ \\ &= E\left(\exp\left(\frac{\varepsilon}{\widehat{\varepsilon}} \widehat{X}_1^\varepsilon\right) - \exp\left(\frac{\varepsilon}{\widehat{\varepsilon}} x\right)\right)^+ \end{aligned}$$

where we recall

$$\widehat{X}_1^\varepsilon \equiv \frac{\widehat{\varepsilon}}{\varepsilon} X_1^\varepsilon = \int_0^1 \sigma(\widehat{\varepsilon}\widehat{B})\widehat{\varepsilon} d\left(\overline{\rho} W + \rho B\right) - \frac{1}{2}\varepsilon\widehat{\varepsilon} \int_0^1 \sigma\left(\widehat{\varepsilon}\widehat{B}_t\right)^2 dt.$$

Consider a Cameron-Martin perturbation of $\widehat{X}_1^\varepsilon$. That is, for a Cameron-Martin path h $= (h, f) \in H_0^1 \times H_0^1$ consider a measure change corresponding to a transformation $\widehat{\varepsilon}\,(W, B) \rightsquigarrow \widehat{\varepsilon}\,(W, B) + (h, f)$ (transforming the Brownian motions to Brownian motions with drift), we obtain the Girsanov density

$$G_\varepsilon = \exp\left(-\frac{1}{\widehat{\varepsilon}} \int_0^1 \dot{h}_s dW_s - \frac{1}{\widehat{\varepsilon}} \int_0^1 \dot{f}_s dB_s - \frac{1}{2\widehat{\varepsilon}^2} \int_0^1 \left(\dot{h}_s^2 + \dot{f}_s^2\right) ds\right). \tag{2.5.1}$$

Under the new measure, $\widehat{X}_1^\varepsilon$ becomes $\widehat{Z}_1^\varepsilon$, where

$$\widehat{Z}_1^\varepsilon = \int_0^1 \sigma(\widehat{\varepsilon}\widehat{B}_t + \widehat{f}_t)\left[\widehat{\varepsilon} d\left(\overline{\rho} W_t + \rho B_t\right) + d\left(\overline{\rho} h_t + \rho f_t\right)\right] - \frac{1}{2}\varepsilon\widehat{\varepsilon} \int_0^1 \sigma(\widehat{\varepsilon}\widehat{B}_t + \widehat{f}_t)^2 dt.$$

Definition 2.5.1. For fixed $x \geq 0$, write $(h, f) \in \mathcal{K}^x$ if $\Phi_1\left(h, f, \widehat{f}\right) = x$. Call such (h, f) admissible for arrival at log-strike x. Call (h^x, f^x) the cheapest admissible control, which attains

$$I\left(x\right) = \inf_{h, f \in H_0^1} \left\{\frac{1}{2} \int_0^1 \dot{h}^2 dt + \frac{1}{2} \int_0^1 \dot{f}^2 dt : \Phi_1\left(h, f, \widehat{f}\right) = x\right\},$$

where we recall that $\widehat{f} = K\dot{f}$ and

$$\Phi_1(h, f, \widehat{f}) = \int_0^1 \sigma(\widehat{f}) d\left(\overline{\rho} h + \rho f\right).$$

For any Cameron-Martin path (h, f), the perturbed random variable $\widehat{Z}_1^\varepsilon$ admits a stochastic Taylor expansion with respect to $\widehat{\varepsilon}$.

Lemma 2.5.2. *Fix $(h, f) \in \mathcal{K}^x$ and define $\widehat{Z}_1^\varepsilon$ accordingly. Then*

$$\widehat{Z}_1^\varepsilon = x + \widehat{\varepsilon} g_1 + \widehat{\varepsilon}^2 R_2^\varepsilon, \tag{2.5.2}$$

where g_1 is a Gaussian random variable, given explicitly by

$$g_1 = \int_0^1 \{\sigma(\widehat{f}_t) d\left(\overline{\rho} W_t + \rho B_t\right) + \sigma'(\widehat{f}_t)\widehat{B}_t d\left(\overline{\rho} h_t + \rho f_t\right)\}, \tag{2.5.3}$$

and

$$\begin{aligned} R_2^\varepsilon = & \int_0^1 \sigma'\left(\widehat{f}_t\right) \widehat{B}_t d\left(\overline{\rho} W_t + \rho B_t\right) - \frac{1}{2}\frac{\varepsilon}{\widehat{\varepsilon}} \int_0^1 \sigma(\widehat{\varepsilon}\widehat{B}_t + \widehat{f}_t)^2 dt \\ & + \frac{1}{2\widehat{\varepsilon}^2} \int_0^{\widehat{\varepsilon}} \int_0^1 \sigma''\left(\zeta \widehat{B}_t + \widehat{f}_t\right) \widehat{B}_t^2 \left[\widehat{\varepsilon} d\left(\overline{\rho} W_t + \rho B_t\right) + d\left(\overline{\rho} h_t + \rho f_t\right)\right] \left(\widehat{\varepsilon} - \zeta\right) d\zeta. \end{aligned} \tag{2.5.4}$$

Proof. By a stochastic Taylor expansion for the controlled process $\widehat{Z}_t^\varepsilon$ with control $(h, f) \in \mathcal{K}^x$ as in Definition 2.5.1 and thanks to $\sigma \in C^2$, we have at $t = 1$

$$\begin{aligned} \widehat{Z}_1^\varepsilon = & \int_0^1 \sigma(\widehat{\varepsilon}\widehat{B} + \widehat{f}) \left[\widehat{\varepsilon} d\left(\overline{\rho} W + \rho B\right) + d\left(\overline{\rho} h + \rho f\right)\right] - \frac{1}{2}\varepsilon\widehat{\varepsilon} \int_0^1 \sigma(\widehat{\varepsilon}\widehat{B}_t + \widehat{f}_t)^2 dt \\ = & \int_0^1 \sigma(\widehat{f}) d\left(\overline{\rho} h + \rho f\right) + \widehat{\varepsilon} \int_0^1 \{\sigma(\widehat{f}) d\left(\overline{\rho} W + \rho B\right) + \sigma'(\widehat{f})\widehat{B} d\left(\overline{\rho} h + \rho f\right)\} + \\ & + \widehat{\varepsilon}^2 \int_0^1 \sigma'\left(\widehat{f}_t\right) \widehat{B}_t d\left(\overline{\rho} W_t + \rho B_t\right) - \frac{1}{2}\varepsilon\widehat{\varepsilon} \int_0^1 \sigma(\widehat{\varepsilon}\widehat{B}_t + \widehat{f}_t)^2 dt \\ & + \frac{1}{2} \int_0^{\widehat{\varepsilon}} \int_0^1 \sigma''\left(\zeta \widehat{B}_t + \widehat{f}_t\right) \widehat{B}_t^2 \left[\widehat{\varepsilon} d\left(\overline{\rho} W_t + \rho B_t\right) + d\left(\overline{\rho} h_t + \rho f_t\right)\right] \left(\widehat{\varepsilon} - \zeta\right) d\zeta. \end{aligned}$$

Collecting terms in powers of $\widehat{\varepsilon}$ and with the random variable g_1 as in (2.5.3) (recalling that $\widehat{\varepsilon}\varepsilon \in \mathcal{O}(\widehat{\varepsilon}^2)$), we have

$$\widehat{Z}_1^\varepsilon = \int_0^1 \sigma(\widehat{f}) d\left(\overline{\rho} h + \rho f\right) + \widehat{\varepsilon} g_1 + \mathcal{O}(\widehat{\varepsilon}^2),$$

furthermore, since $(h, f) \in \mathcal{K}^x$, by the definition of Φ_1, it holds that

$$\int_0^1 \sigma(\widehat{f}) d\left(\overline{\rho} h + \rho f\right) = x.$$

This proves the statement (2.5.2) and the statement that g_1 is Gaussian is immediate from the form (2.5.3). □

Finally, we determine an explicit form of the Girsanov density G_ε for the choice where (h^x, f^x) in (2.5.1) are chosen the cheapest admissible control (cf. Definition 2.5.1. Similarly to classical works of Azencott, Ben Arous (1988) and others,we show that the stochastic integrals in the exponent of G_ε are proportional to the first order term g_1 (with factor $I'(x)$) *when evaluated at the minimizing configuration* (h^x, f^x).

Lemma 2.5.3. *We have*

$$\int_0^1 \dot{h}_t^x dW_t + \int_0^1 \dot{f}_t^x dB_t = I'(x)\, g_1.$$

Proof. See Lemma 2.8.2. □

With these preparations in place, we are now ready to prove the pricing formula from Section 2.2.

Proof of Theorem 2.2.2. With a Girsanov factor (all integrals on $[0,1]$)

$$G_\varepsilon = e^{-\frac{1}{\widehat{\varepsilon}}\int \dot{h} dW - \frac{1}{\widehat{\varepsilon}}\int \dot{f} dB - \frac{1}{2\widehat{\varepsilon}^2}\int \left(\dot{h}^2 + \dot{f}^2\right) dt}$$

and (evaluated at the minimizer)

$$G_\varepsilon|_* = e^{-\frac{I(x)}{\widehat{\varepsilon}^2}} e^{-\frac{I'(x) g_1(\omega)}{\widehat{\varepsilon}}},$$

we have, setting $\widehat{U}^\varepsilon := \widehat{Z}_1^\varepsilon - x = \widehat{\varepsilon} g_1 + \widehat{\varepsilon}^2 R_2^\varepsilon$

$$\begin{aligned}
c(\widehat{x}, t) &= E\left[\left(\exp\left(\frac{\varepsilon}{\widehat{\varepsilon}}\widehat{Z}_1^\varepsilon\right) - \exp\left(\frac{\varepsilon}{\widehat{\varepsilon}} x\right)\right)^+ G_\varepsilon|_*\right] \\
&= e^{\frac{\varepsilon}{\widehat{\varepsilon}} x} E\left[\left(\exp\left(\frac{\varepsilon}{\widehat{\varepsilon}}\widehat{U}^\varepsilon\right) - 1\right)^+ G_\varepsilon|_*\right] \\
&= e^{-\frac{I(x)}{\widehat{\varepsilon}^2}} e^{\frac{\varepsilon}{\widehat{\varepsilon}} x} E\left[\left(\exp\left(\frac{\varepsilon}{\widehat{\varepsilon}}\widehat{U}^\varepsilon\right) - 1\right)^+ e^{-\frac{I'(x) g_1}{\widehat{\varepsilon}}}\right] \\
&= e^{-\frac{I(x)}{\widehat{\varepsilon}^2}} e^{\frac{\varepsilon}{\widehat{\varepsilon}} x} E\left[\left(\exp\left(\frac{\varepsilon}{\widehat{\varepsilon}}\widehat{U}^\varepsilon\right) - 1\right) e^{-\frac{I'(x)}{\widehat{\varepsilon}^2}\widehat{U}^\varepsilon} e^{I'(x) R_2^\varepsilon} \mathbf{1}_{\widehat{U}^\varepsilon \geq 0}\right]. \\
&= e^{-\frac{I(x)}{\widehat{\varepsilon}^2}} e^{\frac{\varepsilon}{\widehat{\varepsilon}} x} J(\varepsilon, x).
\end{aligned}$$

□

2.6 Proof of the moderate deviation expansions

In Section 2, we pointed out that (iiic) is exactly what one gets from (call price) large deviations (2.1.8), if heuristically applied to $x\varepsilon^{2\beta}$. We now give a proper derivation based on moderate deviations.

Lemma 2.6.1. *Assume (iiia-b) from Assumption 2.1.4. Then an upper moderate deviation estimate holds both for calls and digital calls. That is, we have*

(iiic) For every $\beta \in (0,H)$, and every fixed $x>0$, and $\widehat{x}_\varepsilon := x\varepsilon^{1-2H+2\beta}$,

$$E[(e^{X_1^\varepsilon}-e^{\widehat{x}_\varepsilon})^+] \leq \exp\left(-\frac{x^2+o(1)}{2\sigma_0^2\varepsilon^{4H-4\beta}}\right)$$

and also

$$P[X_1^\varepsilon > \widehat{x}_\varepsilon] \leq \exp\left(-\frac{x^2+o(1)}{2\sigma_0^2\varepsilon^{4H-4\beta}}\right). \tag{2.6.1}$$

Proof. Recall $\sigma(.)$ smooth but unbounded and recall $\widehat{x}_\varepsilon := x\varepsilon^{1-2H+2\beta}$. In case of $\beta=0$ and $H=1/2$ a large deviation principle (LDP) for $(X_1^\varepsilon\widehat{\varepsilon}/\varepsilon)$ is readily reduced, via exponential equivalence, to a LDP for the family of stochastic Itô integrals given by $\int \sigma(\widehat{\varepsilon}\widehat{B})\widehat{\varepsilon}dZ$ for some Brownian Z, ρ-correlated with B. There are then many ways to establish a LDP for this family. A particularly convenient one, that requires no growth restriction on σ, uses continuity of stochastic integration with respect to the rough path $(B,Z,\int BdZ)=(B,Z,\int \widehat{B}dZ)$ in suitable metrics, for which a LDP is known (Friz & Hairer, 2014, Ch 9.3). It was pointed out in (Bayer et al., 2017) that a similar reasoning is possible when $H<1/2$, the rough path is then replaced by a "richer enhancement" of (B,Z), the precise size of which depends on H, for which again one has a LDP. A moderate deviation priniple (MDP) for $(X_1^\varepsilon\widehat{\varepsilon}/\varepsilon)$ is a LDP for $(\varepsilon^{-2\beta}X_1^\varepsilon\widehat{\varepsilon}/\varepsilon)$ for $\beta\in(0,H)$. This can be reduced to a LDP, with $\overline{\varepsilon} := \varepsilon^{-2\beta}\widehat{\varepsilon} = \varepsilon^{2H-2\beta}$, for

$$\varepsilon^{-2\beta}\int_0^1 \sigma(\widehat{\varepsilon}\widehat{B})\widehat{\varepsilon}dZ = \int_0^1 \sigma(\widehat{\varepsilon}\widehat{B})\overline{\varepsilon}dZ \equiv \int_0^1 \sigma_\varepsilon(\overline{\varepsilon}\widehat{B})\overline{\varepsilon}dZ$$

with speed $\overline{\varepsilon}^2$. Since $\sigma_\varepsilon(\cdot)\equiv\sigma(\varepsilon^{2\beta}\cdot)$ converges (with all derivatives) locally uniformly to the constant function σ_0, and one checks that the above is exponentially equivalent to the (Gaussian) family given by $\sigma_0\overline{\varepsilon}Z_1$, with law

$\mathcal{N}(0,\sigma_0^2\bar{\varepsilon}^2) = \mathcal{N}(0,\sigma_0^2\varepsilon^{4H-4\beta})$ which gives (2.6.1), even with equality. (By localization, exponential equivalence can again be done for σ without growth restrictions.)

We have not yet used either assumption (iiia-b). These become important in order to extend estimate (2.6.1) to the case of genuine call payoffs. We can follow here a well-known argument (e.g. ((Forde & Jacquier, 2009; Pham, 2010; Forde & Zhang, 2017)) with the "moderate" caveat to carry along a factor $\varepsilon^{2\beta}$. In fact, this follows precisely the argument of (Forde & Zhang, 2017) where the authors carry along a factor $\widehat{\varepsilon}/\varepsilon = \varepsilon^{2H-1}$. (This provides a *unified view on rough and moderate deviations.*) The remaining details then follow essentially "Appendix C. Proof of Corollary 4.13., part (ii) upper bound" of (Forde & Zhang, 2017), noting perhaps that the authors use their assumptions to show validity of what we simply assumed as condition (iiib), and also that one works with the quadratic rate function $I''(0)x^2 = \frac{x^2}{2\sigma_0^2}$ throughout. □

Remark 2.6.2. By an easy argument similar to "Appendix C. Proof of Corollary 4.13., part (i) lower bound" of (Forde & Zhang, 2017) one sees that validity of the call price upper bound (iiic) implies the corresponding digital call price upper bound (2.6.1.) For this reason, we only emphasized (iiic) but not (2.6.1) in Section 2.

In a classical work, Azencott (1982) (see also (Azencott, 1985), (Ben Arous, 1988, Théorème 2)) obtained asymptotic expansions of functionals of Laplace type on Wiener space, of the type "$E[\exp(-F(X^\varepsilon)/\varepsilon^2)]$", for small noise diffusions X^ε. This refines the large deviation (equivalently: Laplace) principle of Freidlin–Wentzell for small noise diffusions. In a nutshell, for fixed $X_0 = x$, Azencott gets expansions of the form $e^{-c/\varepsilon^2}(\alpha_0 + \alpha_1\varepsilon ...)$. His ideas (used by virtually all subsequent works in this direction) are a Girsanov transform, to make the minimizing path "typical", followed by localization around the minimizer (justified by a good large deviation principle), and finally a local (stochastic Taylor) type analysis near the minimizer. None of these ingredients rely on the Markovian structure (or, relatedly, PDE arguments). As a consequence (and motivation for this work) such expansions were also obtained in the (non-Markovian) context of rough differential equations driven by fractional Brownian motion (Inahama, 2013; Baudoin & Ouyang, 2015) with $H < 1/2$.

And yet, our situation is different in the sense that call price Wiener functionals do not fit the form studied by Azencott and others, nor can we in fact expect a similar expansion: Example 2.2.3 gives a Black-Scholes call price expansion of the form constant times $e^{-c\varepsilon^2}(\varepsilon^3 + ...)$. Azencott's ideas are nonetheless very relevant to us: we already used the Girsanov formula in Theorem 2.2.2 in order to have a tractable expression for J. It thus "only" remains to carry out the localization and do some local analysis.

Proposition 2.6.3. *Let $x > 0$ and $\beta \in (0, H)$. Then the factor J is negligible in the sense that, for every $\theta > 0$,*

$$\varepsilon^\theta \log J(\varepsilon, x\varepsilon^{2\beta}) \to 0 \qquad \text{as } \varepsilon \to 0 \text{ .}$$

Proof. Step 1. Localization Write $x_\varepsilon := x\varepsilon^{2\beta}, \widehat{x_\varepsilon} := x_\varepsilon \varepsilon^{1-2H} = x\varepsilon^{1-2H+2\beta}$. By definition,

$$E[(e^{X_1^\varepsilon} - e^{\widehat{x_\varepsilon}})^+]e^{\frac{I(x_\varepsilon)}{\widehat{\varepsilon}^2}} e^{-\widehat{x_\varepsilon}} = J\left(\varepsilon, x_\varepsilon\right) \text{ .}$$

Fix $x, \delta > 0$ and write $\delta_\varepsilon = \delta\varepsilon^{2\beta}$. We claim that (the positive quantity)

$$J\left(\varepsilon, x_\varepsilon\right) - J_{\delta_\varepsilon}\left(\varepsilon, x_\varepsilon\right) = e^{\frac{I(x_\varepsilon)}{\widehat{\varepsilon}^2}} e^{-\widehat{x_\varepsilon}} E[(e^{X_1^\varepsilon} - e^{\widehat{x_\varepsilon}})1_{\widehat{X}_1^\varepsilon > x_\varepsilon + \delta_\varepsilon}] \qquad (2.6.2)$$

is exponentially small, in the sense that, for some $c > 0$ and $\overline{\varepsilon}^2 = \varepsilon^{4H-4\beta}$,

$$J\left(\varepsilon, x_\varepsilon\right) - J_{\delta_\varepsilon}\left(\varepsilon, x_\varepsilon\right) = O\left(e^{-c/\overline{\varepsilon}^2}\right) .$$

There is a battle here between the exploding factor $e^{\frac{I(x_\varepsilon)}{\widehat{\varepsilon}^2}}$, with exponent

$$\frac{I\left(x_\varepsilon\right)}{\widehat{\varepsilon}^2} \sim \frac{I''(0)\left(x_\varepsilon\right)^2}{2\widehat{\varepsilon}^2} = \frac{I''(0)x^2}{2\varepsilon^{4H-4\beta}} \text{ ,}$$

and on the other hand

$$E[(e^{X_1^\varepsilon} - e^{\widehat{x_\varepsilon}})1_{\widehat{X}_1^\varepsilon > x_\varepsilon + \delta_\varepsilon}] \leq \exp\left(-\frac{(x+\delta)^2 + o(1)}{2\sigma_0^2\varepsilon^{4H-4\beta}}\right)$$

where the given estimate is an easy consequence of Lemma 2.6.1. Since $I''(0) = 1/\sigma_0^2$ we see that the last factor "exponentially over-compensates" the rest, so that the difference is indeed exponentially negligible.

Step 2. Upper bound. For any $x > 0$, recall that $\widehat{U}^{\varepsilon,x} = \widehat{U}^{\varepsilon}$ decomposes into a Gaussian random variable $g_1 = g_1^x$ and remainder $R_2^{\varepsilon,x} = R_2^{\varepsilon}$. In order to control this remainder *without* imposing a boundedness assumption on $\sigma(\cdot)$, we will crucially use a "localized remainder tail estimate" as given in Proposition 2.6.4 below. We have, for any $\varepsilon \in (0,1]$,

$$\begin{aligned} J_\delta\left(\varepsilon, x\right) &= E\left[e^{-\frac{I'(x)}{\widehat{\varepsilon}^2}\widehat{U}^{\varepsilon}}\left(\exp\left(\tfrac{\varepsilon}{\widehat{\varepsilon}}\widehat{U}^{\varepsilon}\right)-1\right)e^{I'(x)R_2^{\varepsilon}}\,\mathbf{1}_{\widehat{U}^{\varepsilon}\in[0,\delta_{\varepsilon}]}\right] \qquad (2.6.3)\\ &\leq (e^{\delta}-1)E[e^{-\frac{I'(x)}{\widehat{\varepsilon}}g_1^x};\widehat{U}^{\varepsilon,x}\in[0,\delta]]. \end{aligned}$$

To proceed, recall $\widehat{\varepsilon}^{-1}g_1^x = \widehat{\varepsilon}^{-2}\widehat{U}^{\varepsilon,x} - R_2^{\varepsilon,x}$ so that, for any $\kappa > 0$,

$$\begin{aligned} e^{-\frac{I'(x)}{\widehat{\varepsilon}}g_1^x} &= e^{-\frac{I'(x)}{\widehat{\varepsilon}}g_1^x}\mathbf{1}_{|\widehat{\varepsilon B}|_{\infty;[0,1]}\geq\kappa}\\ &+ e^{-\frac{I'(x)}{\widehat{\varepsilon}^2}\widehat{U}^{\varepsilon,x}}e^{I'(x)R_2^{\varepsilon,x}}\mathbf{1}_{|\widehat{\varepsilon B}|_{\infty;[0,1]}<\kappa}. \end{aligned}$$

Since $I'(x) > 0$ for small enough $x > 0$, it follows that $-\frac{I'(x)}{\widehat{\varepsilon}^2}\widehat{U}^{\varepsilon,x} < 0$ on the event $\{\widehat{U}^{\varepsilon,x} \in [0,\delta]\}$, which leads us to

$$\begin{aligned} J_\delta\left(\varepsilon, x\right) &\leq (e^{\delta}-1)E[e^{-\frac{I'(x)}{\widehat{\varepsilon}}g_1^x};|\widehat{\varepsilon B}|_{\infty;[0,1]}\geq\kappa] + (e^{\delta}-1)E[e^{I'(x)R_2^{\varepsilon,x}};|\widehat{\varepsilon B}|_{\infty;[0,1]}<\kappa]\\ &\leq (e^{\delta}-1)\sqrt{E[e^{-\frac{2I'(x)}{\widehat{\varepsilon}}g_1^x}]}\sqrt{P\left[|\widehat{\varepsilon B}|_{\infty;[0,1]}\geq\kappa\right]} + (e^{\delta}-1)C \end{aligned}$$

where, by Proposition 2.6.4, the constant $C = C(\kappa)$ is uniform in small ε and x. The square-root terms are computed resp. (Fernique) estimated by

$$\exp\left(\frac{(I'\left(x\right))^2\mathbb{V}(g_1^x)}{\widehat{\varepsilon}^2}\right)\times\exp(-c\kappa^2/\widehat{\varepsilon}^2)$$

for some $c > 0$ which depends on the law of B (hence H), but is uniform in ε and x. Hence, for x small enough, the resulting exponent $(I'\left(x\right))^2\mathbb{V}(g_1^x) - c\kappa^2$ is negative, which is more than enough to conclude the upper bound.

Step 3. Lower bound. Write $E_{\delta,\kappa}\left[\cdot\right] = E[\cdot\mathbf{1}_{\widehat{U}^{\varepsilon,x}\in[0,\delta_{\varepsilon}]}\mathbf{1}_{|\widehat{\varepsilon B}|_{\infty;[0,1]}<\kappa}]$ and estimate

$$
\begin{aligned}
& E_{\delta,\kappa}\left[e^{-\frac{I'(x)}{\widehat{\varepsilon}^2}\widehat{U}^\varepsilon/2}\left(\exp\left(\tfrac{\varepsilon}{\widehat{\varepsilon}}\widehat{U}^\varepsilon\right)-1\right)^{1/2}\right] \\
= \quad & E_{\delta,\kappa}\left[e^{-\frac{I'(x)}{\widehat{\varepsilon}^2}\widehat{U}^\varepsilon/2}\left(\exp\left(\tfrac{\varepsilon}{\widehat{\varepsilon}}\widehat{U}^\varepsilon\right)-1\right)^{1/2} e^{I'(x)R_2^\varepsilon/2}\, e^{-I'(x)R_2^\varepsilon/2}\right] \\
\leq \quad & J_\delta\left(\varepsilon,x\right)^{\frac{1}{2}}\, E_{\delta,\kappa}\left[e^{-I'(x)R_2^\varepsilon}\right]^{\frac{1}{2}}
\end{aligned}
$$

where we used Cauchy–Schwarz and discarded the event $\{|\widehat{\varepsilon}\widehat{B}|_{\infty;[0,1]} < \kappa\}$. The localized remainder estimate provides an upper bound on $E_{\delta,\kappa}\left[e^{-I'(x)R_2^\varepsilon}\right]$, uniformly over small (enough) ε and x. It then suffices to get a suitable lower bound of the left-hand side above. Indeed, for $u \in [0,\widehat{\varepsilon}^2\eta] = [0,\varepsilon^{4H}\operatorname{Var}\eta]$, with η small enough, not dependent on ε,

$$
u \mapsto (e^{\frac{\varepsilon}{\widehat{\varepsilon}}u}-1)^{\frac{1}{2}}e^{-\frac{I'(x)}{\widehat{\varepsilon}^2}u/2} \geq \gamma\left(\frac{\varepsilon}{\widehat{\varepsilon}}u\right)^{1/2} \tag{2.6.4}
$$

for a constant $\gamma > 0$ which can also be taken uniformly in small x,ε. Then estimate

$$
\begin{aligned}
& E_{\delta,\kappa}[(e^{\frac{\varepsilon}{\widehat{\varepsilon}}\widehat{U}^\varepsilon}-1)^{\frac{1}{2}}e^{-\frac{I'(x)}{2\varepsilon^2}\widehat{U}^\varepsilon}] \\
\geq \quad & \gamma\varepsilon^{1/2-H}E[|\widehat{U}^\varepsilon|^{1/2}1_{\widehat{U}^\varepsilon\in\left[0,\widehat{\varepsilon}^2\eta\right]}1_{|\varepsilon B|_{\infty;[0,1]}<\kappa}]\ .
\end{aligned}
$$

As a quick sanity check, pretend zero remainder so that $\widehat{U}^\varepsilon = \widehat{\varepsilon}g_1$: dropping further the (exponentially close to probability one) event $\{|\varepsilon B|_{\infty;[0,1]} < \kappa\}$, a Gaussian computation then shows that we are left with ($\gamma\varepsilon^{1/2-H}$ times $\widehat{\varepsilon}^{1/2}$ times)

$$
E[|g_1|^{1/2};g_1\in[0,\widehat{\varepsilon}]] \sim (const)\widehat{\varepsilon}^{3/2}\ .
$$

In general, set $V^\varepsilon = \widehat{U}^\varepsilon/\widehat{\varepsilon} = g_1 + \widehat{\varepsilon}R_2s^\varepsilon$, so that[5]

$$
E_\kappa\left[|\widehat{U}^\varepsilon|^{1/2};\widehat{U}^\varepsilon\in[0,\widehat{\varepsilon}^2\eta]\right] = \widehat{\varepsilon}^{1/2}\,E_\kappa[|V^\varepsilon|^{1/2};V^\varepsilon\in[0,\widehat{\varepsilon}\eta]]\ .
$$

At this stage, it is difficult to treat $\widehat{\varepsilon}R^\varepsilon$ as perturbation of g since, on the given event $\{V^\varepsilon \in [0,\widehat{\varepsilon}\eta]\}$, all terms are of order $\widehat{\varepsilon}$. We can solve this issue by realizing that we can replace, throughout, x by $x_\varepsilon = x\varepsilon^{2\beta}$. Since

[5] Write E_κ for the expected value restricted to the event $\{|\varepsilon B|_{\infty;[0,1]} < \kappa\}$

$I'(x_\varepsilon) \sim (const)x_\varepsilon$, with see from (2.6.4), that in the above estimate the event $\widehat{U}^\varepsilon \in [0, \widehat{\varepsilon}^2\eta] = [0, \varepsilon^{4H}\eta]$ (resp. $V^\varepsilon \in [0, \widehat{\varepsilon}\eta] = [0, \varepsilon^{2H}\eta]$) can be replaced by $\widehat{U}^\varepsilon \in [0, \varepsilon^{4H-2\beta}\eta]$ (resp. $V^\varepsilon \in [0, \varepsilon^{2H-2\beta}\eta]$), possibly with an insignificantly modified constant η. It is now straight-forward to show that the behaviour of $E_\kappa[|V^\varepsilon|^{1/2}; V^\varepsilon \in [0, \varepsilon^{2H-2\beta}\eta]]$ is of the same order as $E[g^{1/2}; g \in [0, \varepsilon^{2H-2\beta}\eta]]$, the correct behaviour (i.e. positive power of ε) is obtained by spelling out the (Gaussian) integral.

□

Proposition 2.6.4 (Localized remainder tail estimate)**.** *For every $\kappa > 0$, there exists $c_1, c_2 > 0$ such that, for all r and uniformly in small ε, x we have*

$$P\left[|R_2^\varepsilon| > r, |\widehat{\varepsilon}\widehat{B}|_{\infty;[0,1]} < \kappa\right] \leq c_1 \exp\left(-c_2 r\right)$$

Proof. We decompose $\widehat{\varepsilon}^2 R_2^\varepsilon = M^\varepsilon + N^\varepsilon$ in terms of the (local) martingale

$$M^\varepsilon := \widehat{\varepsilon}\int_0 \left[\sigma\left(\widehat{\varepsilon}\widehat{B} + \widehat{f}\right) - \sigma\left(\widehat{f}\right)\right] d[\overline{\rho}W + \rho B]$$

and the (bounded variation) process

$$N^\varepsilon := \int_0 \left[\sigma\left(\widehat{\varepsilon}\widehat{B} + \widehat{f}\right) - \sigma\left(\widehat{f}\right) - \sigma'\left(\widehat{f}\right)\widehat{\varepsilon}B\right] d[\overline{\rho}h + \rho f] - \frac{1}{2}\varepsilon\widehat{\varepsilon}\int_0 \sigma^2\left(\widehat{\varepsilon}\widehat{B} + \widehat{f}\right) dt\,.$$

Let $\tau^{\varepsilon,\kappa}$ be the stopping time when $\widehat{\varepsilon}\widehat{B}$ first leaves the uniform ball of radius κ. Then

$$M_t^{\kappa,\varepsilon} := M_{t\wedge\tau^{\varepsilon,\kappa}}^\varepsilon$$

still yields a (local) martingale. The point is that $\left\{|\widehat{\varepsilon}\widehat{B}|_{\infty;[0,1]} < \kappa\right\} = \{\tau^{\varepsilon,\kappa} > 1\}$. On this event, $M^\varepsilon|_{[0,1]} = M^{\kappa,\varepsilon}|_{[0,1]}$ and we can thus replace M^ε, in the definition of the remainder, by $M^{\kappa,\varepsilon}$. Let $K = K^{\kappa,x}$ be the κ-fattening of $\{f(t) : 0 \leq t \leq 1\}$, recall $f = f^x$, then, for $t \in [0,1]$,

$$d\left[M^{\kappa,\varepsilon}\right]_t / dt = \widehat{\varepsilon}^2(\sigma(\widehat{\varepsilon}\widehat{B}_t + \widehat{f}_t) - \sigma(f_t))^2 \leq \widehat{\varepsilon}^4 \left\|\sigma'\right\|_{\infty;K}^2 \; |\widehat{B}_t|^2\,.$$

Clearly, we can replace K by $\widetilde{K^\kappa}$ which contains all $K^{\kappa,x}$ for small x. To summarize, we have, on the event $\left\{|\widehat{\varepsilon}\widehat{B}|_{\infty;[0,1]} < \kappa\right\}$,

$$R^\varepsilon\left(\cdot\right) = \widehat{\varepsilon}^{-2}M^{\kappa,\varepsilon} + \widehat{\varepsilon}^{-2}N^\varepsilon$$

with $[\widehat{\varepsilon}^{-2}M^{\kappa,\varepsilon}] = O\left(|\widehat{B}|^2_{\infty;[0,1]}\right)$ and, as seen by a similar (but easier) reasoning, $\widehat{\varepsilon}^{-2}N^{\varepsilon} = O\left(|\widehat{B}|^2_{\infty;[0,1]}\right)$, always for fixed $\kappa > 0$, but uniformly in small ε (equivalently, $\widehat{\varepsilon}$) and small $x > 0$. This clearly shows that $\widehat{\varepsilon}^{-2}N^{\varepsilon}$ has exponential tails. The same is true for the martingale part, whose bracket is $O(\text{Gaussian}^2)$. This is exactly the situation for the "model" martingal increment $2\int_0^1 BdB = B_1^2 - 1$ which clearly has exponential tails. To make this rigorous, recall that Gaussian resp. exponential tails are characertized by $O(\sqrt{p})$ resp. $O(p)$-growth of the L^p-norms. The statement is then an easy consequence of the sharp (upper) BDG constant (Carlen & Kree, 1991), known to be $O(\sqrt{p})$. □

2.7 Proof of the implied volatility expansion

With Theorem 2.2.2 in place, we now turn to the proof of the implied volatility expansion, formulated in Theorem 2.2.6.

Proof of Theorem 2.2.6. We will use an asymptotic formula for the dimensionless implied variance

$$V_t^2 = t\sigma_{\text{impl}}(k_t, t)^2, \quad t > 0,$$

obtained in (Gao & Lee, 2014). It follows from the first formula in Remark 7.3 in (Gao & Lee, 2014) that

$$V_t^2 - \frac{k_t^2}{2L_t} = \mathcal{O}\left(\frac{k_t^2}{L_t^2}(k_t + |\log k_t| + \log L_t)\right), \quad t \to 0, \tag{2.7.1}$$

where $L_t = -\log c(k_t, t)$, $t > 0$.

We will need the following formula that was established in the proof of Theorem 2.2.4:

$$L_t = \frac{I(kt^{\beta})}{t^{2H}} + \mathcal{O}(t^{-\theta}) \tag{2.7.2}$$

as $t \to 0$, for all $x \geq 0$ and $\beta \in [0, H)$ and any $\theta > 0$. Let us first assume

$\frac{2H}{n+1} \leq \beta < \frac{2H}{n}$. Using the energy expansion, we obtain from (2.7.2) that

$$L_t = \sum_{i=2}^{n} \frac{I^{(i)}(0)}{i!} k^i t^{i\beta - 2H} + \mathcal{O}\left(t^{-\theta}\right) = \frac{I''(0)}{2} k^2 t^{2\beta - 2H} \times \left[1 + \sum_{i=3}^{n} \frac{2I^{(i)}(0)}{i! I''(0)} k^{i-2} t^{(i-2)\beta} + \mathcal{O}\left(t^{2H - 2\beta - \theta}\right)\right] \tag{2.7.3}$$

as $t \to 0$. The second term in the brackets on the right-hand side of (2.7.3) disappears if $n = 2$.

Remark 2.7.1. Suppose $n \geq 2$ and $\frac{2H}{n+1} \leq \beta < \frac{2H}{n}$. Then formula (2.7.3) is optimal. Next, suppose $n \geq 2$ and $0 < \beta < \frac{2H}{n+1}$. In this case, there exists $m \geq n+1$ such that $\frac{2H}{m+1} \leq \beta < \frac{2H}{m}$, and hence (2.7.3) holds with m instead of n. However, we can replace m by n, by making the error term worse. It is not hard to see that the following formula holds for all $n \geq 2$ and $0 < \beta < \frac{2H}{n+1}$:

$$L_t = \sum_{i=2}^{n} \frac{I^{(i)}(0)}{i!} k^i t^{i\beta - 2H} + \mathcal{O}\left(t^{(n+1)\beta - 2H}\right) = \frac{I''(0)}{2} k^2 t^{2\beta - 2H} \times \left[1 + \sum_{i=3}^{n} \frac{2I^{(i)}(0)}{i! I''(0)} k^{i-2} t^{(i-2)\beta} + \mathcal{O}\left(t^{(n-1)\beta}\right)\right] \tag{2.7.4}$$

as $t \to 0$ provided we choose θ small enough.

Let us continue the proof of Theorem 2.2.6. Since $k_t \approx t^{\frac{1}{2} - H + \beta}$ and $L_t \approx t^{2\beta - 2H}$ as $t \to 0$, (2.7.1) implies that

$$V_t^2 = \frac{k^2 t^{1 - 2H + 2\beta}}{2L_t} + \mathcal{O}\left(t^{1 + 2H - 2\beta - \theta}\right), \quad t \to 0. \tag{2.7.5}$$

Next, using the Taylor formula for the function $u \mapsto \frac{1}{1+u}$, and setting

$$u = \sum_{i=3}^{n} \frac{2I^{(i)}(0)}{i! I''(0)} k^{i-2} t^{(i-2)\beta} + \mathcal{O}(t^{2H - 2\beta - \theta}),$$

we obtain from (2.7.3) that

$$(2L_t)^{-1} = \frac{t^{2H - 2\beta}}{k^2 I''(0)} \left[\sum_{j=0}^{n-2} (-1)^j u^j + \mathcal{O}(u^{n-1})\right]$$

as $t \to 0$. It follows from $\frac{2H}{n+1} \le \beta < \frac{2H}{n}$ that $(n-1)\beta \ge 2H - 2\beta$, and hence

$$\begin{aligned}(2L_t)^{-1} &= \frac{t^{2H-2\beta}}{k^2 I''(0)} \left[\sum_{j=0}^{n-2} (-1)^j u^j \right] + \mathcal{O}(t^{4H-4\beta-\theta}) \\ &= \frac{t^{2H-2\beta}}{k^2 I''(0)} \left[\sum_{j=0}^{n-2} (-1)^j \left(\sum_{i=3}^{n} \frac{2I^{(i)}(0)}{i! I''(0)} k^{i-2} t^{(i-2)\beta} \right)^j \right] + \mathcal{O}(t^{4H-4\beta-\theta})\end{aligned}$$

as $t \to 0$. Now, (2.7.5) gives

$$\begin{aligned}V_t^2 = \frac{t}{I''(0)} &\left[\sum_{j=0}^{n-2} (-1)^j \left(\sum_{i=3}^{n} \frac{2I^{(i)}(0)}{i! I''(0)} k^{i-2} t^{(i-2)\beta} \right)^j \right] \\ &+ \mathcal{O}\left(t^{1+2H-2\beta-\theta} \right)\end{aligned}$$

as $t \to 0$. Finally, by cancelling a factor of t in the previous formula, we obtain formula (2.2.6) for $\frac{2H}{n+1} \le \beta < \frac{2H}{n}$. The proof in the case where $\beta \le \frac{2H}{n+1}$ is similar. Here we take into account Remark 2.7.1. This completes the proof of Theorem 2.2.6. □

2.8 Auxiliary lemmas

In this section we provide and prove some auxiliary lemmas, which are used in the preparations to the proof of Theorem 2.2.2. We start with a technical Lemma, that justifies the derivation.

Lemma 2.8.1. *Assume $\sigma(.) > 0$ and $|\rho| < 1$. Then $\mathcal{K}^x$ is a Hilbert manifold near any $\mathfrak{h} := (h, f) \in \mathcal{K}^x \subset \mathfrak{H} := H_0^1 \times H_0^1$.*

Proof. Similar to Bismut (1984, p. 25), we need to show that $D\varphi_1(\mathfrak{h})$ is surjective where $\varphi_1(\mathfrak{h}) : \mathfrak{H} \to \mathbb{R}$ with

$$\varphi_1(\mathfrak{h}) = \varphi_1(h, f) = \int_0^1 \sigma(\hat{f}) d(\bar{\rho} h + \rho f).$$

From

$$\begin{aligned}\varphi_1(\mathfrak{h}+\delta\mathfrak{h}') &= \int_0^1 \sigma(\widehat{f}+\delta\widehat{f}')d\left(\overline{\rho}h+\rho f+\delta(\overline{\rho}h'+\rho f')\right)\\ &= \varphi_1(\mathfrak{h})+\delta\int_0^1\sigma(\widehat{f})d(\overline{\rho}h'+\rho f')\\ &\quad +\delta\int_0^1\sigma'(\widehat{f})\widehat{f}'d\left(\overline{\rho}h+\rho f\right)+o\left(\delta\right).\end{aligned}$$

the functional derivative $D\varphi_1(\mathfrak{h})$ can be computed explicitly. In fact, even the computation

$$\left(D\varphi_1(\mathfrak{h}),(h',0)\right)=\overline{\rho}\int_0^1\sigma(\widehat{f})dh'$$

is sufficient to guarantee surjectivity of $D\varphi_1(\mathfrak{h})$. □

We now give the proof of Lemma 2.5.3, which determines the form of the Girsanov measure change (2.5.1) for the minimizing configuration.

Lemma 2.8.2. *(i) Any optimal control $\mathfrak{h}^0=(h^x,f^x)\in\mathcal{K}^x$ is a critical point of*

$$\mathfrak{h}=(h,f)\mapsto -I\left(\varphi_1^{\mathfrak{h}}\right)+\frac{1}{2}\left\Vert\mathfrak{h}\right\Vert_{\mathfrak{H}}^2;$$

(ii) it holds that

$$\int_0^1\dot{h}^xdW+\int_0^1\dot{f}^xdB=I'(x)\,g_1.$$

Proof. (Step 1) Write $\mathfrak{h}=(h,f)$ and

$$\varphi_1(\mathfrak{h})=\varphi_1(h,f)=\int_0^1\sigma(\widehat{f})d\left(\overline{\rho}h+\rho f\right).$$

Let $\mathfrak{h}^0=(h^x,f^x)\in\mathcal{K}^x$ an optimal control. Then

$$\mathfrak{Ker}D\varphi_1\left(\mathfrak{h}^0\right)=T_{\mathfrak{h}^0}\mathcal{K}^x=\left\{\mathfrak{h}\in\mathfrak{H}^1:D\varphi_1(\mathfrak{h})=0\right\}.$$

(This requires $\mathcal{K}^x$ to be a Hilbert manifold near $\mathfrak{h}^0$, as was seen in the last lemma.)

(Step 2) For fixed $\mathfrak{h}\in\mathfrak{H}$, define

$$u(t):=-I\left(\varphi_1^{\mathfrak{h}^0+t\mathfrak{h}}\right)+\frac{1}{2}\left\Vert\mathfrak{h}^0+t\mathfrak{h}\right\Vert_{\mathfrak{H}}^2\geq 0$$

with equality at $t=0$ (since $x=\varphi_1^{\mathfrak{h}^0}$ and $I(x)=\frac{1}{2}\left\Vert\mathfrak{h}^0\right\Vert_{\mathfrak{H}}^2$) and non-negativity

for all t because $\mathfrak{h}^0 + t\mathfrak{h}$ is an admissible control for reaching $\widetilde{x} = \varphi_1^{\mathfrak{h}^0 + t\mathfrak{h}}$ (so that $I(\widetilde{x}) = \inf\{...\} \leq \frac{1}{2}\|\mathfrak{h}^0 + t\mathfrak{h}\|_{\mathfrak{H}}^2$.)
(Step 3) We note that $\dot{u}(0) = 0$ is a consequence of $u \in C^1$ near 0, $u(0) = 0$ and $u \geq 0$. In other words, $\mathfrak{h}^0$ is a critical point for

$$\mathfrak{H}^1 \ni \mathfrak{h} \mapsto -I\left(\varphi_1^{\mathfrak{h}}\right) + \frac{1}{2}\|\mathfrak{h}\|_{\mathfrak{H}}^2 .$$

(Step 4) The functional derivative of this map at $\mathfrak{h}^0$ must hence be zero. In particular, for all $\mathfrak{h} \in \mathfrak{H}$,

$$\begin{aligned} 0 &\equiv -I'\left(\varphi_1^{\mathfrak{h}^0}\right)\left\langle D\varphi_1\left(\mathfrak{h}^0\right), \mathfrak{h}\right\rangle + \left\langle \mathfrak{h}^0, \mathfrak{h}\right\rangle \\ &= -I'(x)\left\langle D\varphi_1\left(\mathfrak{h}^0\right), \mathfrak{h}\right\rangle + \left\langle \mathfrak{h}^0, \mathfrak{h}\right\rangle . \end{aligned}$$

(Step 5) With $\mathfrak{h}^0 = (h^x, f^x)$ and $\mathfrak{h} = (h, f)$

$$\begin{aligned} \left\langle D\varphi_1\left(\mathfrak{h}^0\right), \mathfrak{h}\right\rangle &= \left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} \int_0^1 \sigma(\widehat{f^x} + \varepsilon\widehat{f}) d\left(\overline{\rho}h^x + \rho f^x + \varepsilon\left(\overline{\rho}h + \rho f\right)\right) \\ &= \int_0^1 \sigma(\widehat{f^x}) d\left(\overline{\rho}h + \rho f\right) + \int_0^1 \sigma'(\widehat{f^x})\widehat{f} d\left(\overline{\rho}h^x + \rho f^x\right) \end{aligned}$$

By continuous extension, replace $\mathfrak{h} = (h, f)$ by (W, B) above and note that

$$\left\langle D\varphi_1\left(\mathfrak{h}^0\right), (W, B)\right\rangle = g_1$$

since indeed $g_1 = \int_0^1 \sigma(\widehat{f}_t) d\left(\overline{\rho}W_t + \rho B_t\right) + \sigma'(\widehat{f}_t)\widehat{B}_t d\left(\overline{\rho}h_t + \rho f_t\right)$. Hence

$$\int_0^1 \dot{h}^x dW + \int_0^1 \dot{f}^x dB = I'(x)\, g_1. \qquad \square$$

3 Monte Carlo pricing under rough stochastic volatility

This chapter is organized as follows. In Section 3.1, we set the scene, introduce notation and state assumptions. In Section 3.2, we present the novel Monte Carlo pricing scheme. In Section 3.3, we collect the results of our numerical experiments in which we look at numerical strong and weak rates of convergence of different objects of interest.

3.1 Exposition and assumptions

Throughout this chapter, we shall be working on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ satisfying the *usual conditions* and supporting two independent Brownian motions under the pricing measure $\mathbb{P}$. We consider a finite time horizon $T < \infty$ and assume that the asset price process $S = (S_t)_{t\in[0,T]}$ has been normalized without loss of generality such that spot $S_0 = 1$ and the risk-free rate $r = 0$.

We are interested in stochastic volatility models of the form

$$\frac{\mathrm{d}S_t}{S_t} = f(\widehat{W}_t, t)\mathrm{d}B_t, \quad t \in [0, T] \tag{3.1.1}$$

where B is a Brownian motion. Following Bayer et al. (2016), for fixed Hurst index $H \in (0, \frac{1}{2})$ and another Brownian motion W, the process $\widehat{W} = (\widehat{W}_t)_{t\in[0,T]}$ is a Volterra or Riemann-Liouville (RL) fractional Brownian motion (fBM) with integral representation

$$\widehat{W}_t := \int_0^t K(s,t)\,\mathrm{d}W_s = \int_0^t \sqrt{2H}\,|t-s|^{H-\frac{1}{2}}\,\mathrm{d}W_s, \quad s \in [0,t].$$

Here, we work with a RL fBM but other choices such as the classical fBM by Mandelbrot and Van Ness (1968) for suitably modified kernel K are also possible. The leverage effect $\mathrm{d}\langle B, W\rangle_t = \rho \mathrm{d}t$ between stock and volatility drivers is incorporated by working with a 2D standard Brownian motion $\left(W, \overline{W}\right)$ and setting

$$B := \rho W + \overline{\rho}\,\overline{W} \equiv \rho W + \sqrt{1-\rho^2}\,\overline{W}.$$

Note that, in contrast even to many classical stochastic volatility models such as Heston (1993), the stochastic volatility is explicitly given and no rough / stochastic differential equation needs to be solved. In the terminology of Bayer et al. (2017), this class of rough stochastic volatility models is called *simple*. Throughout this chapter, we further assume that $f : \mathbb{R}^2 \to \mathbb{R}_+$ is smooth in its first argument and such that the asset price is a (local) martingale, as required by modern financial theory. We remark that the function f admits in particular an explicit time dependence such that for example the *rough Bergomi* model (Bayer et al., 2016) with non-constant forward variance curve is also covered by our framework.

We are interested in pricing a European Call with spot $S_0 > 0$, strike $K > 0$ and time to maturity $T > 0$ in the context of a *simple* stochastic volatility model. Let the Black-Scholes pricing function for said option be given by

$$C_{\mathrm{BS}}\left(S_0, K, \sigma^2 T\right) := \mathbb{E}\left[S_0 \exp\left(\sigma\sqrt{T}Z - \frac{\sigma^2}{2}T\right) - K\right]^+$$

where Z denotes a standard normal random variable. By an elementary conditioning argument, introduced in a Markovian context by Romano and Touzi (1997); Willard (1997), on the sample path $W_{[0,T]}$, the fair price of a European Call in the context of a simple rough volatility model can be reduced to the expectation of a Black-Scholes (Black & Scholes, 1973) formula with random inputs. More precisely, one can show that

$$C_{\mathrm{RV}}(K,T) := \mathbb{E}\left[C_{BS}\left(\exp\left(\rho\mathscr{I} - \frac{\rho^2}{2}\mathscr{V}\right), K, \overline{\rho}^2\mathscr{V}\right)\right] \tag{3.1.2}$$

with the bivariate object $(\mathscr{I}, \mathscr{V})$ defined as follows

$$(\mathscr{I}, \mathscr{V}) := (\mathscr{I}(T), \mathscr{V}(T)) = \left(\int_0^T f(\widehat{W}_r, r)\mathrm{d}W_r, \int_0^T f^2(\widehat{W}_r, r)\mathrm{d}r\right). \tag{3.1.3}$$

From a numerical perspective, this is a step forward as it avoids any simulation of the second Brownian motion $\overline{W}$. In order to sample from (3.1.3), an immediate idea would be to sample the two-dimensional Gaussian process $(W, \widehat{W})_{t\in[0,T]}$ on some fine (equidistant) discretization grid and then approximate $(\mathscr{I}, \mathscr{V})$ by an Euler discretization of the respective integrals (3.1.3). Recall from the introduction that for the joint simulation of $(W, \widehat{W})_{t\in[0,T]}$, exact and approximate numerical schemes have been proposed in the literature (Bayer et al., 2016; Bennedsen et al., 2017; McCrickerd & Pakkanen, 2018; Horvath et al., 2017). A remaining problem however with that approach is that the convergence of an Euler approximation of $\mathscr{I}$ is very slow since $\widehat{W}$ has little regularity when H is small (recall that Gatheral et al. (2018) report H as low as 0.05). In fact, Neuenkirch and Shalaiko (2016) show (in a slightly different setting) that the strong rate for the standard Euler scheme (or, more precisely, left-point rule) is no better than H in general even when the fractional process is exactly simulated.

3.2 Monte Carlo pricing

The aim of this section is to propose a novel approximation scheme for the bivariate object $(\mathscr{I}, \mathscr{V})$ as defined in (3.1.3). For fixed $N \in \mathbb{N}$, let us first define a level N Haar grid to be one with step size given by $\varepsilon = 2^{-N}$. With $\phi := \mathbf{1}_{[0,1)}$ the so-called father wavelet, the Haar system given by the set of functions

$$\left\{\phi_{l,N} = 2^{N/2}\,\phi(\cdot\, 2^N - l) = 2^{N/2}\mathbf{1}_{[l2^{-N},(l+1)2^{-N})} \mid l \in \mathbb{Z}\right\}$$

then forms an orthonormal basis of $L^2(\mathbb{R})$. While the Paley-Wiener-Zygmund theorem states that Brownian motion is almost surely nowhere differentiable (in a classical sense), it does have a derivative in a *distributional* or *generalized* sense. This derivative is given by White Noise which we shall denote $\dot{W}$.

The basis of our approach is a Karhunen-Loève-style approximation of White Noise. Mathematically, for N fixed and with $\{Z_l\}_{l\in\mathbb{Z}}$ some standard iid normal random variables, let $\dot{W}^\varepsilon$ be an approximation of White noise given by

$$\dot{W}^{\varepsilon}(t)=\sum_{l=-\infty}^{\infty} Z_l\phi_{l,N}(t)=\sum_{l=0}^{\lceil t2^N\rceil-1} Z_l 2^{N/2}\mathbf{1}_{[l2^{-N},(l+1)2^{-N})}(t), \quad t\in[0,T]. \tag{3.2.1}$$

By integration against against the Kernel $K(s,t)$, this then induces an approximation of the Riemann-Liouville fBm $\widehat{W}$ given by

$$\widehat{W}^{\varepsilon}(t)=\int_0^t K(s,t)\,\dot{W}^{\varepsilon}(s)\mathrm{d}s=\sum_{l=0}^{\lceil t2^N\rceil-1} Z_l\,\widehat{e}_l^{\varepsilon}(t), \quad t\in[0,T] \tag{3.2.2}$$

where for all $l\in\left[0,\left\lceil t2^N\right\rceil-1\right]$, the functions $\widehat{e}_l^{\varepsilon}$ are given by

$$\widehat{e}_l^{\varepsilon}(t)=\frac{\sqrt{2H}2^{N/2}}{H+1/2}\left(|t-l2^{-N}|^{H+1/2}-|t-\min((l+1)2^{-N},t)|^{H+1/2}\right). \tag{3.2.3}$$

Recall that it is our fundamental interest to devise an approximation scheme for the bivariate object $(\mathscr{I},\mathscr{V})$ defined in (3.1.3). Having constructed a joint approximation $\left(\dot{W}^{\varepsilon},\widehat{W}^{\varepsilon}\right)$ of $\left(\dot{W},\widehat{W}\right)$, a mathematical analysis by Bayer et al. (2017) however reveals that

$$\mathscr{I}^{\varepsilon}=\int_0^T f(\widehat{W}^{\varepsilon}(t),t)\dot{W}^{\varepsilon}(t)\mathrm{d}t \nrightarrow \mathscr{I}=\int_0^T f(\widehat{W}_r,r)\mathrm{d}W_r, \quad \varepsilon\to 0. \tag{3.2.4}$$

This does not come as a surprise as even when considering a standard Brownian motion ($H=1/2$), a well-known result by Wong and Zakai (1965) states that $\mathscr{I}^{\varepsilon}$ converges to the Stratonovich version of $\mathscr{I}$ which is given by $\mathscr{I}$ plus an Itô-Stratonovich correction term. The latter is given by the quadratic covariation, defined (whenever possible) as the limit, in probability, of

$$\sum_{[u,v]\in\pi}(f(\widehat{W}_v)-f(\widehat{W}_u))(W_v-W_u), \tag{3.2.5}$$

along any sequence (π) of partitions with mesh-size tending to zero. But, disregarding trivial situations, this limit does not exist for $H<1/2$. For instance, when $f(x)=x$ fractional scaling immediately gives divergence (at rate $H-1/2$) of the above bracket approximation.

Bayer et al. (2017) argue that the theory of *regularity structures* (Hairer, 2014) provides a convenient, yet technically advanced framework to address this issue and to *renormalize* the approximative integral $\mathscr{I}^\varepsilon$ such that it converges to the desired Itô integral $\mathscr{I}$ in the limit.

Definition 3.2.1 (Bayer et al. (2017))**.** For the specific case of the Haar basis, let us first define a renormalization object $\mathscr{C}^\varepsilon(t)$ which can be one of

$$\mathscr{C}^\varepsilon(t) = \begin{cases} \frac{2^N\sqrt{2H}}{H+1/2}|t - \left\lfloor t2^N \right\rfloor 2^{-N}|^{H+1/2}, & H \in \left(0, \frac{1}{2}\right] \\ \frac{\sqrt{2H}}{(H+1/2)(H+3/2)} 2^{N(1/2-H)}, & H > \frac{1}{4}. \end{cases} \tag{3.2.6}$$

where the second expression only valid for $H > 1/4$ is the time average of the first equation. One approach towards renormalization of $\mathscr{I}^\varepsilon$ is then given by

$$\widetilde{\mathscr{I}^\varepsilon}(t) := \int_0^t f(\widehat{W}^\varepsilon(r), r)\, \mathrm{d}W^\varepsilon(r) - \int_0^t \mathscr{C}^\varepsilon(r)\partial_1 f(\widehat{W}^\varepsilon(r), r)\, \mathrm{d}r \tag{3.2.7}$$

for $t \in [0, T]$.

The next theorem then provides quantitative estimates for the convergence of $\widetilde{\mathscr{I}^\varepsilon}$ towards the desired Itô integral $\mathscr{I}$. More precisely, a strong rate for the integral approximation is given below.

Theorem 3.2.2 (Bayer et al. (2017))**.** *Let f be a smooth and bounded function with bounded derivatives. Alternatively, let f and its derivatives be of exponential growth. With $\widetilde{\mathscr{I}^\varepsilon}$ as defined in Definition 3.2.1, for any $\delta \in (0, 1)$ and any $p < \infty$, there exists C such that*

$$\left\| \sup_{t\in[0,T]} \left| \widetilde{\mathscr{I}^\varepsilon}(t) - \int_0^t f(\widehat{W}(r), r)\mathrm{d}W(r) \right| \right\|_{L^p} \leq C\varepsilon^{\delta H}, \tag{3.2.8}$$

Remark 3.2.3. With regards to the mentioned results of Neuenkirch and Shalaiko (2016), this shows that the scheme described in Definition 3.2.1 is almost optimal with a strong rate of almost H.

Remark 3.2.4. Readers interested in the derivation of the correction terms are advised to study the original paper. In this work, we exclusively provide the numerical counterpart to the theoretical results obtained in (Bayer et al., 2017).

Substituting the terms of the renormalized integral approximation defined in (3.2.7) by the respective expressions (3.2.1) and (3.2.6) for the approximation of the White Noise and the non-constant renormalization object, we arrive at the following expression which has been transformed into a form more convenient for simulation:

$$\widetilde{\mathscr{I}}^{\varepsilon} = \sum_{l=0}^{\lceil T2^N \rceil - 1} \int_{l2^{-N}}^{(l+1)2^{-N}} \Big[Z_l 2^{N/2} f(\widehat{W}^{\varepsilon}(t), t) - \frac{\sqrt{2H}2^N}{H+1/2} |t - l2^{-N}|^{H+1/2} \partial_1 f(\widehat{W}^{\varepsilon}(t), t) \Big] \mathrm{d}t \quad (3.2.9)$$

and similarly for the other integral (which does not exhibit convergence issues)

$$\mathscr{V}^{\varepsilon} = \sum_{l=0}^{\lceil T2^N \rceil - 1} \int_{l2^{-N}}^{(l+1)2^{-N}} f^2(\widehat{W}^{\varepsilon}(t), t) \mathrm{d}t. \quad (3.2.10)$$

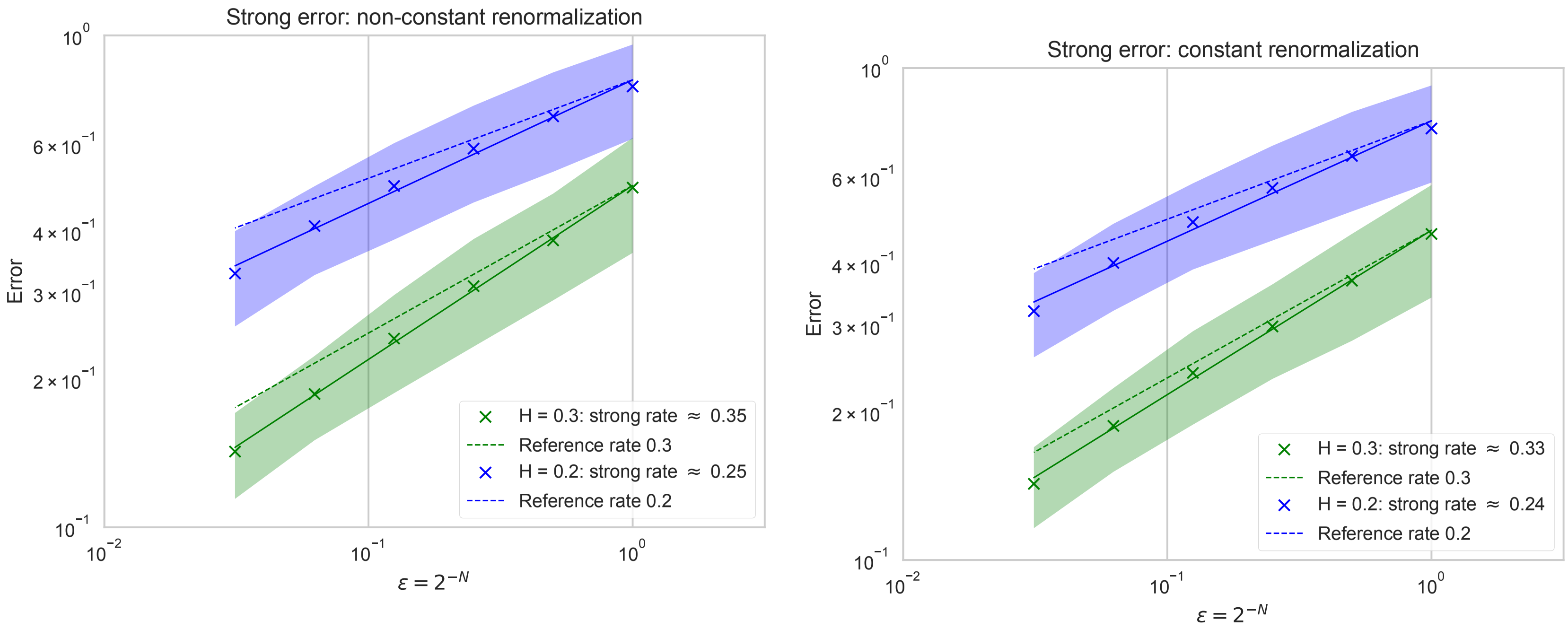
A vectorized numerical procedure for computing Monte Carlo samples of $\left(\widetilde{\mathscr{I}}^{\varepsilon}, \mathscr{V}^{\varepsilon}\right)$ based on (3.2.9), (3.2.10) is then collected in Algorithm 1.

3.3 Numerical results

In this subsection, we will discuss strong convergence of the approximative object $\widetilde{\mathscr{I}}^{\varepsilon}$ to the actual object of interest $\mathscr{I}$ as well as weak convergence of the option price as the Haar grid interval size $\varepsilon \to 0$. Specifically, we will be looking at Monte Carlo estimates of our errors, that is, in order to approximate some quantity $\mathbb{E}[X]$ for some random variable X, we will instead be looking at $\frac{1}{M}\sum_{i=1}^{M} X_i$ where the X_i are M *iid* samples drawn from the same distribution as X. In other words, we need to generate M realisations of the bivariate stochastic object $\left(\widetilde{\mathscr{I}}^{\varepsilon}, \mathscr{V}^{\varepsilon}\right)$, a task that can be vectorized as described below, thus avoiding expensive looping through realizations.[1] Without loss of generality, we set time to maturity $T = 1$.

Algorithm 1: Simulation of M samples of $(\widetilde{\mathscr{I}}^{\varepsilon}, \mathscr{V}^{\varepsilon})$

```
Parameters: # Monte Carlo simulations $M$, Haar grid level $N$,
            # discretisation points of trapezoidal rule in each Haar
            subinterval $d$
Output: $M$ samples of bivariate object $(\widetilde{\mathscr{I}}^{\varepsilon}, \mathscr{V}^{\varepsilon})$
initialize $\widetilde{\mathscr{I}}^{\varepsilon} = \mathscr{V}^{\varepsilon} = \mathbf{0} \in \mathbb{R}^{M}$;
simulate array $\mathbf{Z} \in \mathbb{R}^{M \times \lceil T2^{N} \rceil}$ of iid standard normals;
for each Haar subinterval $[l2^{-N}, (l+1)2^{-N})$ where
  $l \in \{0, \dots, \lceil T2^{N} \rceil - 1\}$ do
    choose discretization grid $\mathcal{D}^{l}$ with $d$ points on the Haar subinterval;
    evaluate functions $\widehat{e}_{k}^{\varepsilon}$ defined in (3.2.3) for $k = 0, \dots, l$ on $\mathcal{D}^{l}$ to
      obtain $\widehat{\mathbf{e}}^{\varepsilon} \in \mathbb{R}^{(l+1) \times d}$;
    compute $\widehat{\mathbf{W}}^{\varepsilon} = \mathbf{Z}^{*} \times \widehat{\mathbf{e}}^{\varepsilon} \in \mathbb{R}^{M \times d}$ where $\mathbf{Z}^{*} \in \mathbb{R}^{M \times (l+1)}$ is the
      truncation of $\mathbf{Z}$ to its first $l+1$ columns such that $\widehat{\mathbf{W}}^{\varepsilon}$ is an
      approximation of the fBM on $\mathcal{D}^{l}$;
    evaluate integrands from equations (3.2.9, 3.2.10) on $\mathcal{D}^{l}$ using $\widehat{\mathbf{W}}^{\varepsilon}$
      and the last column of $\mathbf{Z}^{*}$;
    approx. respective integrals on subinterval by trapezoidal rule ;
    add obtained estimates to running sums $\widetilde{\mathscr{I}}^{\varepsilon}$ and $\mathscr{V}^{\varepsilon}$;
end
return $\widetilde{\mathscr{I}}^{\varepsilon}, \mathscr{V}^{\varepsilon}$
```



(a) Non-constant renormalization

(b) Constant renormalization

Figure 3.1: **Numerical strong rates of convergence of the renormalized integral approximation $\widetilde{\mathscr{I}}^{\varepsilon}$.** Depicted are strong errors on a log-log-scale as $\varepsilon \to \varepsilon' = 2^{-8}$, obtained through $M = 10^5$ Monte Carlo samples with trapezoidal rule delta of $\Delta = 2^{-17}$. Solid lines visualize empirical rates of convergence obtained by least squares regression, dashed lines provide visual reference rates. Shaded colour bands show interpolated 95% confidence levels based on normality of MC estimator.

3.3.1 Strong convergence of the renormalized integral

With the help of several Monte Carlo experiments we can verify part (i) of Theorem 3.2.2 numerically as follows. We fix $p = 2$, i.e. look at convergence in $L^2(\Omega)$, and choose $f(x,t) = \exp(x)$ because this closely resembles models such as the *rough Bergomi* model by Bayer et al. (2016). Note that for simplicity we have excluded an explicit time dependence. We are concerned with Monte Carlo approximations of

$$\|\widetilde{\mathscr{I}}^{\varepsilon} - \int_0^1 \exp(\widehat{W}_t)\mathrm{d}W_t\|_{L^2(\Omega)}$$

and in line with Theorem 3.2.2, we expect an error almost of order ε^H.

Remark 3.3.1. We also checked the simplest non-trivial choice, $f(x,t) = x$ but here the discretization error is overshadowed by the Monte Carlo error, even for very coarse grids.

As has been mentioned before, $\left(W, \widehat{W}\right)$ is a two-dimensional Gaussian process with known covariance structure, it is therefore possible to use the Cholesky algorithm (Bayer et al., 2016) to simulate the joint paths exactly on some grid and then use standard Riemann sums to approximate the integral. The value obtained in this way could serve as a reference value for our scheme. However for strong convergence we need both objects to be based on the same stochastic sample. For this reason, we find it easier to construct a reference value by the wavelet-based scheme itself, i.e. we simply pick some $\varepsilon' \ll \varepsilon$ and consider

$$\|\widetilde{\mathscr{I}}^{\varepsilon} - \widetilde{\mathscr{I}}^{\varepsilon'}\|_{L^2(\Omega)} \tag{3.3.1}$$

as $\varepsilon \to \varepsilon'$. As can be seen in Figures 3.1a and 3.1b, both renormalization approaches stated in (3.2.6) are consistent with a theoretical strong rate of almost H across the full range of $0 < H < \frac{1}{2}$.

[1] Documented Python 3 code has been made available at the URL https://www.github.com/RoughStochVol.

3.3.2 Weak convergence

An analysis that – for suitable test functions $\varphi : \mathbb{R} \to \mathbb{R}$ – yields a weak rate of convergence for

$$\left| \mathbb{E}\left[\varphi\left(\widetilde{\mathscr{I}}^{\varepsilon}\right)\right] - \mathbb{E}\left[\varphi\left(\int_0^1 \exp(\widehat{W}_t)\mathrm{d}W_t\right)\right]\right|, \quad \varepsilon \to 0$$

remains an open problem.[2] Hence, we shall perform two numerical experiments in this section. First, picking $\varphi(x) = x^2$, Ito's isometry conveniently yields

$$\mathbb{E}\left[\left(\int_0^1 \exp(\widehat{W}_t)\mathrm{d}W_t\right)^2\right] = \int_0^1 \mathbb{E}\left[\exp\left(2\widehat{W}_t\right)\right]\mathrm{d}t = \int_0^1 \exp\left(2t^{2H}\right)\mathrm{d}t \tag{3.3.2}$$

which can be approximated numerically. So we can consider

$$\left|\mathbb{E}\left[\left(\widetilde{\mathscr{I}}^{\varepsilon}\right)^2\right] - \int_0^1 \exp\left(2t^{2H}\right)\mathrm{d}t\right|, \quad \varepsilon \to 0. \tag{3.3.3}$$

Our preliminary results indicate that for both renormalization approaches the weak rate seems to be around the strong rate H.

In another experiment, we pick a simplified version of the *rough Bergomi* model (Bayer et al., 2016) where the instantaneous variance is given by

$$f^2(x) = \sigma_0^2 \exp(\eta x)$$

with σ_0 and η denoting spot volatility and volatility of volatility respectively. Let $C_{\mathrm{RV}}^{\varepsilon}(K,T)$ denote the approximation of the call price (3.1.2) based on $\left(\widetilde{\mathscr{I}}^{\varepsilon}, \widetilde{\mathscr{V}}^{\varepsilon}\right)$, fix some $\varepsilon' \ll \varepsilon$ and consider

$$\left|C_{\mathrm{RV}}^{\varepsilon}(K,1) - C_{\mathrm{RV}}^{\varepsilon'}(K,1)\right|, \quad \varepsilon \to \varepsilon'. \tag{3.3.4}$$

The empirical results displayed in Figure 3.2 indicate a weak rate of $2H$ across the full range of $0 < H < \frac{1}{2}$.

[2] Of course, for φ Lipschitz, strong convergence implies weak convergence with the same rate or better.

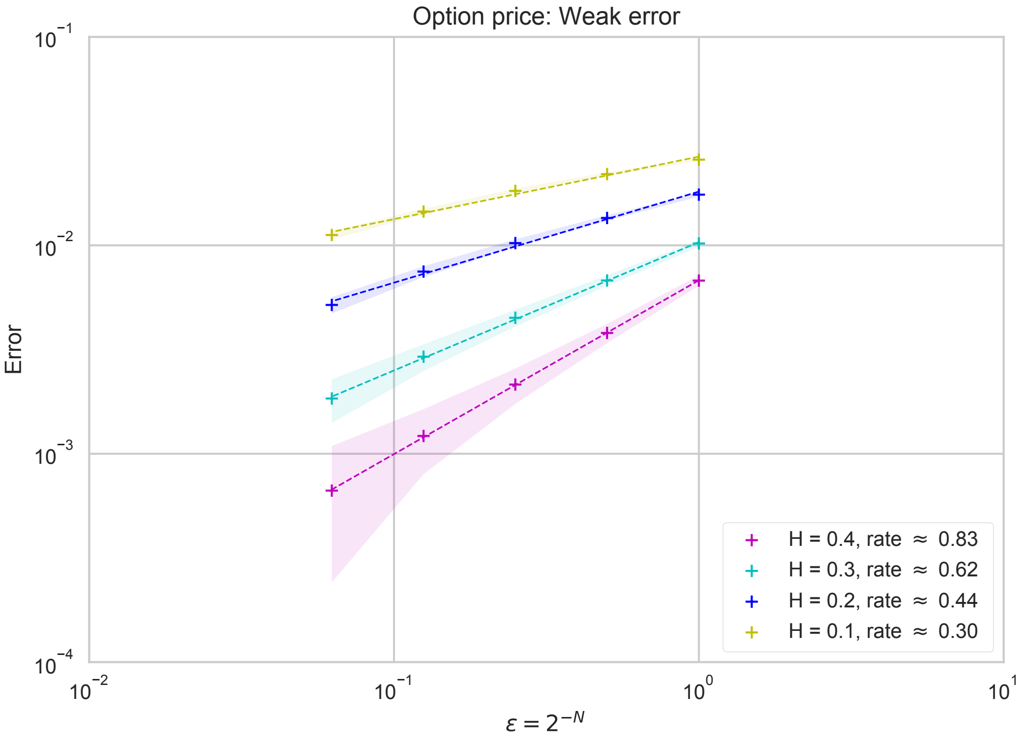


Figure 3.2: **Numerical weak rates of convergence of the approximative option price.** Depicted are weak errors on a log-log-scale as $\varepsilon \to \varepsilon' = 2^{-8}$, obtained through $M = 10^5$ MC samples with spot $S_0 = 1$, strike $K = 1$, correlation $\rho = -0.8$, spot vol $\sigma_0 = 0.2$, vvol $\eta = 2$ and trapezoidal rule delta $\Delta = 2^{-17}$. Dashed lines represent LS estimates for rate estimation, shaded colour bands show confidence levels based on normality of the MC estimator.

4 Deep calibration of rough volatility models

This chapter is organized as follows. In Section 4.1, we set the scene, introduce notation and revisit some important machinery that lies at the core of our proposed calibration scheme. In Section 4.2, we state the model calibration objective and introduce *deep calibration*, our approach of combining the established Levenberg-Marquardt calibration algorithm with neural network regression to enable the efficient calibration of (rough) stochastic volatility models. In Section 4.3, we outline practical intricacies of our approach, ranging from considerations related to generating synthetic, tailored *labeled data* for training, validation and testing to tricks of the trade when training neural networks and performing hyperparameter optimization. Finally, in Section 4.4, we collect the results of our numerical experiments.

4.1 Background

We now set the scene and introduce notation. Throughout the chapter, we shall be working on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ satisfying the *usual conditions* and supporting two (or more) independent Brownian motions under the pricing measure $\mathbb{P}$. We consider a finite time horizon $T < \infty$ and assume the asset price process $S = (S_t)_{t\in[0,T]}$ has been without loss of generality normalized such that spot $S_0 = 1$ and risk-free rate $r = 0$. We define *moneyness* $M := K/S_0$ and *log moneyness* $m := \log(M) = \log(K)$.

4.1.1 Construction of a model IV surface

The concept of an *implied volatility surface* is an important idea and tool central to the theory of modern option pricing. In the introduction, we saw how such a surface arises from market prices of liquid European Call options on the S&P 500 Index *SPX* (cf. Figure 1.1). We now formalize the construction of such a surface from model prices. In a first step, we define the pricing function that maps model & option parameters (and possibly external market information) to the fair price of a European option at time $t = 0$.

Definition 4.1.1 (Pricing map)**.** Consider a (rough) stochastic volatility (market) model for an asset S with model parameters $\mu \in \mathcal{M} \subseteq \mathbb{R}^m$ and possibly incorporated market information $\xi \in \mathcal{E} \subseteq \mathbb{R}^k$. The fair price of a European Call option at time $t = 0$ is then given by

$$\mathbb{E}\left[S_T(\mu,\xi) - M\right]^+$$

where $(M,T) \in \Theta \subseteq \mathbb{R}^2$ denote moneyness and time to maturity respectively. Letting

$$\mathcal{I} := \{(\mu,\xi) \times (M,T) \mid \mu \in \mathcal{M}, \xi \in \mathcal{E}, (M,T)^T \in \Theta\} \subseteq \mathbb{R}^{m+k+2} \tag{4.1.1}$$

be the pricing input space, we then define the pricing map $P_0 : \mathcal{I} \to \mathbb{R}_+$ by

$$(\mu,\xi) \times (M,T) \mapsto \mathbb{E}\left[S_T(\mu,\xi) - M\right]^+ . \tag{4.1.2}$$

Example 4.1.2. In the rough Bergomi model by Bayer et al. (2016), the dynamics for the asset price process S and the instantaneous variance process $v = (v_t)_{t\in[0,T]}$ are given by

$$\begin{aligned} \frac{\mathrm{d}S_t}{S_t} &= \sqrt{v_t}\mathrm{d}\left(\rho W_t + \sqrt{1-\rho^2}W_t^{\perp}\right) \\ v_t &= \xi_0(t)\exp\left(\eta W_t^H - \frac{1}{2}\eta^2 t^{2H}\right), \quad t \in [0,T]. \end{aligned}$$

Here, $\left(W, W^{\perp}\right) = \left(W_t, W_t^{\perp}\right)_{t\in[0,T]}$ are two independent Brownian motions and $\rho \in (-1, 1)$ is a constant correlation parameter introducing the *leverage effect* – the empirically observed anti correlation between stock and volatility movements – at the driving noise level. The parameter $\eta > 0$ denotes volatility of variance and $\xi_0(t) : \mathbb{R}_+ \to \mathbb{R}_+$ given by $\xi_0(t) = \mathbb{E}(v_t), t \in [0, T]$ is a so-called *forward variance curve* which may be recovered from market information (Bayer et al., 2016). Moreover, W^H is a *Riemann-Liouville* fractional Brownian motion given by

$$W_t^H = \sqrt{2H} \int_0^t (t-s)^{H-\frac{1}{2}} \mathrm{d}W_s, \quad t \in [0, T]$$

with Hurst parameter $H \in (0, 1)$. By Kolmogorov, sample paths of W^H are locally almost surely H-ε Hölder for $\varepsilon > 0$. With respect to Definition 4.1.1, hence $\mu = (H, \eta, \rho)$ and $\xi = \xi_0$.

Example 4.1.3. In the Heston model (Heston, 1993), with independent Brownian motions W and $W^{\perp}$ and model parameters ρ, η defined as in Example 4.1.2, the dynamics of the asset price S and the instantaneous variance process $v = (v_t)_{t\in[0,T]}$ starting from spot variance $v_0 > 0$ follow

$$\begin{aligned}
\frac{\mathrm{d}S_t}{S_t} &= \sqrt{v_t}\mathrm{d}\left(\rho W_t + \sqrt{1-\rho^2} W_t^{\perp}\right) \\
\mathrm{d}v_t &= \lambda(\overline{v} - v_t)\mathrm{d}t + \eta\sqrt{v_t}\mathrm{d}W_t, \quad t \in [0, T].
\end{aligned}$$

Here, $\overline{v} > 0$ is the long-run average variance and $\lambda > 0$ is the speed of mean reversion. Feller's condition $2\lambda\overline{v} > \eta^2$ ensures that $v_t > 0$ for $t \geq 0$. In this model, we thus have $\mu = (\lambda, \overline{v}, v_0, \rho, \eta)$ and no market information is incorporated into the model.

Let $\mathrm{BS}(M, T, \sigma)$ denote the Black-Scholes price of a European Call with moneyness M, time to maturity T and assumed constant volatility σ of the underlying and let $Q(M, T)$ be the corresponding market price. The BS implied volatility $\sigma_{\mathrm{iv}}(M, T)$ corresponding to $Q(M, T)$ satisfies

$$Q(M, T) - \mathrm{BS}(M, T, \sigma_{\mathrm{iv}}(M, T)) \overset{!}{=} 0.$$

and the map $(M, T) \mapsto \sigma_{\mathrm{iv}}(M, T)$ is called a *volatility surface*.

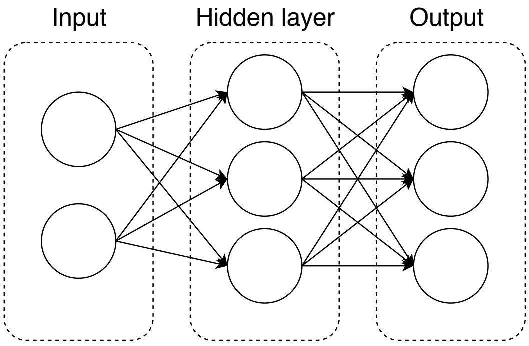


Figure 4.1: **Schematic of a fully-connected neural network (FCNN).** Depicted FCNN has a single *hidden layer* consisting of three neurons and may learn to represent a subset of general functions $f : \mathbb{R}^2 \to \mathbb{R}^3$. In the directed acyclic graph above, vertices denote nodes and directed edges describe the flow of information from one node to the next. If number of hidden layers higher than one (typically dozens or hundreds of layers), a neural network is considered *deep*.

Definition 4.1.4 (IV map)**.** Let μ, ξ, M, T be defined as in Definition 4.1.1. The Black-Scholes IV $\sigma_{\text{iv}}(\mu, \xi, M, T)$ corresponding to the theoretical model price $P_0\left(\mu, \xi, M, T\right)$ satisfies

$$P_0\left(\mu, \xi, M, T\right) - \text{BS}(M, T, \sigma_{\text{iv}}(\mu, \xi, M, T)) \overset{!}{=} 0. \tag{4.1.3}$$

The function $\varphi : \mathcal{I} \to \mathbb{R}_+$ given by

$$(\mu, \xi, M, T) \mapsto \sigma_{\text{iv}}(\mu, \xi, M, T) \tag{4.1.4}$$

is what we call the *implied volatility map*.

4.1.2 Regression with neural networks

Given a data set $\mathcal{D} = \left\{(x_i, y_i) : x_i \in \mathbb{R}^d, y_i \in \mathbb{R}\right\}_{i=1}^n$ of variables x_i and corresponding scalar, continuous response variables y_i, the statistical procedure of estimating the relationship between these variables is commonly called *regression analysis*. Here, we will introduce neural networks and outline their prowess as a regression tool.

The atomic building block of every neural network is a *node*, a functional that performs a weighted sum of its (multi-dimensional) inputs, adds a bias term and then composes the linearity with a scalar non-linear function α :

$\mathbb{R} \to \mathbb{R}$ that is identical across the network. Formally, for some input $x \in \mathbb{R}^d, d \in \mathbb{N}$, the output of an individual node is given by

$$y = \alpha\left(w^T x + b\right) \in \mathbb{R}$$

where $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are individual *weight* and *bias* terms. An *artificial neural network* is then a collection of many such nodes, grouped into non-overlapping sets called *layers* together with a rule of how the information flows between the layers.

Over the years different architectural styles have been developed to suit the specific needs of different domains such as speech, text or vision. Arguably the simplest neural network topology not adapted to any particular domain is that of a *fully-connected neural network* (FCNN). An FCNN consists of sequentially ordered so-called *dense layers* followed by a linear output layer. Any two nodes of a dense layer act independently of each other and do not share weights and biases. Their input is given by the output of all nodes in the previous layer – or all input features if it is the first layer – and their output serves as an input to all nodes in the following layer, see Figure 4.1 for a depiction of a small example.

FCNNs serve as powerful regression tools because they are able to represent large families of functions. In his *Universal Approximation Theorem*, Hornik (1991) proves that FCNNs can approximate continuous functions on $\mathbb{R}$ arbitrarily well.

Theorem 4.1.5 (Universal Approximation Theorem)**.** *Let $N(\alpha)$ denote the space of functions that a fully connected neural network with activation function $\alpha : \mathbb{R} \to \mathbb{R}$, a single hidden layer with a finite number of neurons $l \in \mathbb{N}$ and a linear output layer can represent, i.e.*

$$N(\alpha) = \left\{ f : \mathbb{R}^d \to \mathbb{R} \mid f(x) = \sum_{i=1}^{l} w_i \alpha\left(\sum_{j=1}^{d} \overline{w}_j^{(i)} x_j + b^{(i)}\right) + b_i \right.$$
$$\left. \textit{for some } w, b \in \mathbb{R}^l \textit{ and } \overline{w}^{(i)}, \overline{b}^{(i)} \in \mathbb{R}^d, 1 \le i \le l \right\}$$

where $w, b \in \mathbb{R}^l$ are weights and biases of the output layer and $\overline{w}^{(i)}, \overline{b}^{(i)} \in \mathbb{R}^d, 1 \le i \le l$ are the weights and biases of the l individual neurons in the *hidden layer. Assuming the activation function* $\alpha : \mathbb{R} \to \mathbb{R}$ *is non-constant, unbounded and continuous,* $N(\alpha)$ *is dense in* $C(X)$ *for compact* $X \subseteq \mathbb{R}$ *in the uniform topology, i.e. for any* $f \in C(X)$ *and arbitrary* $\varepsilon > 0$*, there is* $g \in N(\alpha)$ *such that*

$$\sup_{x \in X} |f(x) - g(x)| < \varepsilon.$$

The *Rectified Linear Unit (ReLU)* nonlinearity $\alpha : \mathbb{R} \to \mathbb{R}_+$ given by $\alpha(x) := \max(0, x)$ fulfills the conditions of being non-constant, unbounded and continuous and so in theory ReLU FCNNs allow for approximation of continuous functions to arbitrary accuracy. However, the reason the ReLU has become a de facto standard in recent years (LeCun, Bengio, & Hinton, 2015) is that in comparison to first generation nonlinearities such as the *sigmoid* or *tanh*, ReLU networks are superior in terms of their *algorithmic learnability*, see more in Section 4.3.

Remark 4.1.6. Over the years, various alternative activation functions have been proposed such as Leaky ReLU (He, Zhang, Ren, & Sun, 2015), ELU (Clevert, Unterthiner, & Hochreiter, 2015) or lately the SiLU (Elfwing, Uchibe, & Doya, 2018; Ramachandran, Zoph, & Le, 2017). To date, none of these activation functions have been shown to consistently outperform ReLUs (Ramachandran et al., 2017), so a systematic comparison of the effect of different activation functions on training results has been left for future research.

4.2 Calibration of option pricing models

The implied volatility map $\varphi : \mathcal{I} \to \mathbb{R}_+$ defined in (4.1.4) formalizes the influence of model parameters on an option pricing model's implied volatility surface. *Calibration* describes the procedure of tweaking model parameters to fit a model surface to an empirical IV surface obtained by transforming liquid European option market prices to Black-Scholes IVs (cf. Figure 1.1). A mathematically convenient approach consists of minimizing the weighted squared differences between market and model IVs of $N \in \mathbb{N}$ plain vanilla European options.

Proposition 4.2.1 (Calibration objective)**.** *Consider a (rough) stochastic* *volatility model with model parameters $\mu \in \mathcal{M} \subseteq \mathbb{R}^m$ and embedded market information $\xi \in \mathcal{E} \subseteq \mathbb{R}^k$ (recall Def. 4.1.1). Suppose the* market *IV quotes of N European options with moneyness $M^{(i)}$ and time to maturity $T^{(i)}$ are given by*

$$\mathbf{Q} := \left(Q\left(M^{(1)}, T^{(1)}\right), \ldots, Q\left(M^{(N)}, T^{(N)}\right)\right)^T \in \mathbb{R}^N$$

and analogously the model *IV quotes of the same options under said pricing model are given by*

$$\boldsymbol{\varphi}\left(\mu, \xi\right) := \left(\varphi\left(\mu, \xi, M^{(1)}, T^{(1)}\right), \ldots, \varphi\left(\mu, \xi, M^{(N)}, T^{(N)}\right)\right)^T \in \mathbb{R}^N.$$

Given market quotes $\boldsymbol{Q}$ and market information ξ, we define the residual $\boldsymbol{R}(\mu) : \mathcal{M} \to \mathbb{R}^N$ between market and model IVs by

$$\boldsymbol{R}(\mu) := \boldsymbol{\varphi}(\mu, \xi) - \mathbf{Q}$$

so that the calibration objective becomes

$$\mu^\star = \underset{\mu \in \mathcal{M}}{\arg\min} \|\mathbf{W}^{\frac{1}{2}} \boldsymbol{R}(\mu)\|_2^2 = \underset{\mu \in \mathcal{M}}{\arg\min} \|\mathbf{W}^{\frac{1}{2}} \left[\boldsymbol{\varphi}(\mu, \xi) - \mathbf{Q}\right]\|_2^2 := \Psi\left(\mathbf{W}, \xi, \boldsymbol{Q}\right) \tag{4.2.1}$$

where $\boldsymbol{W} = diag\left[w_1, \ldots, w_N\right] \in \mathbb{R}^{N \times N}$ is a diagonal matrix of weights and $\|\cdot\|_2$ denotes the standard Euclidean norm.

Since $\boldsymbol{R}(\mu) : \mathcal{M} \to \mathbb{R}^N$ is non-linear in the parameters $\mu \in \mathcal{M} \subseteq \mathbb{R}^m$ and $N > m$, the optimization objective (4.2.1) is an example of an overdetermined non-linear least squares problem, usually solved numerically using iterative solvers such as the de-facto standard Levenberg-Marquardt (LM) algorithm (Levenberg, 1944; Marquardt, 1963).

Proposition 4.2.2 (LM calibration)**.** *Suppose $\boldsymbol{R} : O \to \mathbb{R}^N$ is twice continuously differentiable on an open set $O \subseteq \mathbb{R}^m$ and $N > m$. Let $\boldsymbol{J} : O \to \mathbb{R}^{N \times m}$ denote the Jacobian of $\boldsymbol{R}$ with respect to the model parameters $\mu \in \mathbb{R}^m$, i.e.*

its components are given by

$$[\boldsymbol{J}_{ij}]_{\substack{1\leq i\leq N,\\ 1\leq j\leq m}} = \left[\frac{\partial \boldsymbol{R}_i(\mu)}{\partial \mu_j}\right]_{\substack{1\leq i\leq N,\\ 1\leq j\leq m}} = \left[\frac{\partial \boldsymbol{\varphi}_i\,(\mu,\xi)}{\partial \mu_j}\right]_{\substack{1\leq i\leq N,\\ 1\leq j\leq m}}.$$

With regards to the objective in (4.2.1), the algorithm starts with an initial parameter guess $\mu_0 \in \mathbb{R}^m$ and then at each iteration step with current parameter estimate $\mu_k \in \mathbb{R}^m, k \in \mathbb{N}$, the parameter update $\Delta_\mu \in \mathbb{R}^m$ solves

$$\left[\boldsymbol{J}(\mu_k)^T \boldsymbol{W} \boldsymbol{J}(\mu_k) + \lambda I_m\right] \Delta_\mu = \boldsymbol{J}(\mu_k)^T \boldsymbol{W} \boldsymbol{R}(\mu_k) \tag{4.2.2}$$

where $I_m \in \mathbb{R}^{m\times m}$ denotes the identity and $\lambda \in \mathbb{R}$.

It is hence necessary that the *normal equations* (4.2.2) be quickly and accurately solved for the iterative step Δ_μ. In a general (rough) stochastic volatility setting this is problematic: The true implied volatility map $\varphi : \mathcal{I} \to \mathbb{R}_+$ as well as its Jacobian $\boldsymbol{J} :\ O\ \to\ \mathbb{R}^{N\times m}$ are unknown in analytical form. In the absence of an analytical expression for Δ_μ, an immediate remedy is:

(I) Replace the (theoretical) true pricing map $P_0 : \mathcal{I} \to \mathbb{R}_+$ defined in (4.1.2) by an efficient numerical approximation $\widetilde{P}_0 : \mathcal{I} \to \mathbb{R}_+$ such as Monte Carlo, Fourier Pricing or similar means. This gives rise to an approximate implied volatility map $\widetilde{\varphi} : \mathcal{I} \to \mathbb{R}_+$.

(II) Apply finite-difference methods to $\widetilde{\varphi} : \mathcal{I} \to \mathbb{R}_+$ to compute an approximate Jacobian $\widetilde{\boldsymbol{J}} :\ O\ \to\ \mathbb{R}^{N\times m}$.

In many (rough) stochastic volatility models such as *rough Bergomi*, expensive Monte Carlo simulations have to be used to approximate the pricing map. In a common calibration scenario where the normal equations (4.2.2) have to be solved frequently, the approach outlined above thus renders calibration prohibitively expensive.

4.2.1 Deep calibration

In a first step, we use the approximate implied volatility map $\widetilde{\varphi} : \mathcal{I} \to \mathbb{R}_+$ to synthetically generate a large and as accurate as computationally feasible set

Algorithm 2: Deep calibration (LM combined with NN regression)

```
Input: Implied vol map $\boldsymbol{\varphi}_{\text{NN}}$ and its Jacobian $\boldsymbol{J}_{\text{NN}}$, market quotes $\boldsymbol{Q}$,
       market info $\xi$
Parameters: Lagrange multiplier $\lambda_0 > 0$, maximum number of
            iterations $n_{\text{max}}$, minimum tolerance of step norm $\varepsilon_{\text{min}}$,
            bounds $0 < \beta_0 < \beta_1 < 1$
Result: Calibrated model parameters $\mu^\star$
initialize model parameters $\mu = \mu_0$ and step counter $n = 0$;
compute $\boldsymbol{R}(\mu) = \boldsymbol{\varphi}_{\text{NN}}(\mu, \xi) - \mathbf{Q}$ and $\boldsymbol{J}_{\text{NN}}(\mu)$ and solve normal
  equations (4.2.2) for $\Delta_\mu$;
while $n < n_{max}$ and $\|\Delta_\mu\|_2 > \varepsilon$ do
    compute relative improvement $c_\mu = \frac{\|\boldsymbol{R}(\mu)\|_2 - \|\boldsymbol{R}(\mu+\Delta_\mu)\|_2}{\|\boldsymbol{R}(\mu)\|_2 - \|\boldsymbol{R}(\mu)+\boldsymbol{J}_{\text{NN}}(\mu)\Delta_\mu\|_2}$ with
      respect to predicted improvement under linear model;
    if $c_\mu \leq \beta_0$ then reject $\Delta_\mu$, set $\lambda = 2\lambda$;
    if $c_\mu \geq \beta_1$ then accept $\Delta_\mu$, set $\mu = \mu + \Delta_\mu$ and $\lambda = \frac{1}{2}\lambda$;
    compute $\boldsymbol{R}(\mu)$ and $\boldsymbol{J}_{\text{NN}}(\mu)$ and solve normal equations (4.2.2) for
      $\Delta_\mu$;
    set $n = n + 1$;
end
```

of labeled data

$$\mathcal{D} := \left\{\left(x^{(i)}, \widetilde{\varphi}\left(x^{(i)}\right)\right) \mid x \in \mathcal{I}\right\}_{i=1}^{n} \in (\mathcal{I} \times \mathbb{R}_+)^n, \quad n \in \mathbb{N}.$$

Here, it is sensible to trade computational savings for an increased numerical accuracy since the expensive data generation only has to be performed once. Using the sample input-output pairs $\mathcal{D}$, a ReLU FCNN is trained to approximate $\widetilde{\varphi} : \mathcal{I} \to \mathbb{R}_+$, in other words, we use a ReLU FCNN to regress response variables $\widetilde{\varphi}\left(x^{(i)}\right) = \widetilde{\varphi}\left(\mu^{(i)}, \xi^{(i)}, M^{(i)}, T^{(i)}\right)$ on explanatory variables $\left(\mu^{(i)}, \xi^{(i)}, M^{(i)}, T^{(i)}\right)$. We denote this function that the network is now able to represent by $\varphi_{\text{NN}} : \mathcal{I} \to \mathbb{R}_+$. With respect to the repeated solving of the normal equations (4.2.2), the benefit of this new approach is twofold:

(I) Evaluations of $\varphi_{\text{NN}} : \mathcal{I} \to \mathbb{R}_+$ amount to forward runs of a trained ReLU FCNN. Computationally, forwards runs come down to highly optimized and parallelizable matrix-matrix multiplications combined with element-wise comparison operations – recall the ReLU activation is given by $\alpha(\cdot) = \max(0, \cdot)$ – both of which are fast.

(II) In order to perform *backpropagation*, the standard training algorithm for neural networks, industrial grade machine learning software libraries such as Google Inc.'s Tensorflow (Abadi et al., 2016) ship with built-in implementations of *automatic differentiation* (Baydin, Pearlmutter, Radul, & Siskind, 2015). This may easily be exploited to quickly compute approximative Jacobians $\boldsymbol{J}_{\text{NN}} : O \to \mathbb{R}^{N \times m}$ accurate to machine precision.

It is also important to stress that trained networks can be efficiently stored, moved and loaded, so training results can be shared and deployed quickly.

Remark 4.2.3. Hernandez (2017) calibrates the Hull and White (1990) short-rate model by directly learning calibrated model parameters from market data, i.e. the total calibration routine Ψ in (4.2.1). Extending his approach to equity models necessitates a network topology that allows to learn from empirical IV point clouds. Here, adaptations of Convolutional Neural Networks (CNNs) invented for computer vision problems might be worthwhile to explore.

4.3 Neural network training

While theoretically easy to understand, the training of neural networks in practice often becomes a costly and most importantly time–consuming exercise full of potential pitfalls. To this end, we outline here the approach taken in this chapter, briefly mentioning important *tricks of the trade* that have been utilized to facilitate or accelerate the training of the ReLU FCNN networks.

4.3.1 Generation of synthetic labeled data

The ability of a neural network to learn the implied volatility map φ to a high degree of accuracy critically hinges upon the provision of a large and accurate labeled data set

$$\mathcal{D} = \left\{ \left(\mu^{(i)}, \xi^{(i)}, M^{(i)}, T^{(i)}, \widetilde{\varphi}\left(\mu^{(i)}, \xi^{(i)}, M^{(i)}, T^{(i)} \right) \right) \right\}_{i=1}^{n} \in (\mathcal{I} \times \mathbb{R}_{+})^{n}, \quad n \in \mathbb{N}.$$

Table 4.1: Marginal priors of model parameters μ for synthetically generating $\mathcal{D}$. The continuous uniform distribution on the interval bounded by $a_i, b_i \in \mathbb{R}$ is denoted by $\mathcal{U}[a_i, b_i]$ and $\mathcal{N}_{\text{trunc}}[a_i, b_i, \lambda, \sigma]$ stands for the normal distribution with mean $\lambda \in \mathbb{R}$ and standard deviation $\sigma \in \mathbb{R}_+$, truncated to the interval $[a_i, b_i]$ with $a_i, b_i \in \mathbb{R}$.

Heston		rough Bergomi	
Parameter	Marginal	Parameter	Marginal
η	$\mathcal{U}[0,5]$	η	$\mathcal{N}_{\text{trunc}}[1,4,2.5,0.5]$
ρ	$\mathcal{U}[-1,0]$	ρ	$\mathcal{N}_{\text{trunc}}[-1,-0.5,-0.95,0.2]$
λ	$\mathcal{U}[0,10]$	H	$\mathcal{N}_{\text{trunc}}[0.01,0.5,0.07,0.05]$
$\overline{v}$	$\mathcal{U}[0,1]$	v_0	$\mathcal{N}_{\text{trunc}}[0.05,1,0.3,0.1]^2$
v_0	$\mathcal{U}[0,1]$		

Knowledge of the parametric dependence structure $\widetilde{\varphi} : \mathcal{I} \to \mathbb{R}_+$ between inputs and corresponding outputs allows us to address these requirements adequately. First, trading computational savings for increased numerical accuracy, we ensure that $\|\widetilde{\varphi} - \varphi\|_\infty < \varepsilon$ for ε small. Second, we can sample an arbitrarily large set of labeled data $\mathcal{D}$, allowing the network to learn the underlying dependence structure $\widetilde{\varphi}$ – rather than noise present in the training set – and generalize well to unseen test data. In the numerical tests in Section 4.4, we draw $n = |\mathcal{D}| = 10^6$ iid sample inputs from a to be specified sampling distribution $\mathcal{G}$ on $\mathcal{I}$ and compute the corresponding outputs as follows: For Heston, we use the Fourier pricing method implemented in the open-source quantitative finance library *QuantLib* (Ametrano et al., 2015) which makes use of the well-known fact that the characteristic function of the log asset price is known. For *rough Bergomi*, we use a self-coded, parallelized implementation of a slightly improved version of the Monte Carlo scheme proposed by McCrickerd and Pakkanen (2018). Black-Scholes IVs are inverted from option prices using a publicly available implementation of the implied volatility solver by Jäckel (2015). The full dataset $\mathcal{D}$ is then randomly shuffled and partitioned into training, validation and test sets $\mathcal{D}_{train}, \mathcal{D}_{valid}$ and $\mathcal{D}_{test}$ of sizes $n_{\text{train}}, n_{\text{valid}}$ and n_{test} respectively.[1]

An important advantage of being able to synthetically generate labeled data is the freedom in choosing the sampling distribution $\mathcal{G}$ on $\mathcal{I}$. Prior to calibra-

[1] The code has been made available at https://github.com/roughstochvol.

tion, little is known about the interplay of model parameters and particular model parameter regions of highest interest to be learned accurately. Consequently, we assume zero prior knowledge of the (joint) relevance of model parameters μ in the Heston experiment in Section 4.4. An ad-hoc approach is to sample individual model parameters independently of each other from the uniformly continuous marginal distributions collected in Table 4.1. A similar reasoning also applies in the rough Bergomi experiment in Section 4.4, except that here we do assume some prior marginal distributional knowledge and use truncated normal marginals instead of uniform marginals.

On the other hand, it is reasonable to increase the number of samples in option parameter regions with high liquidity since these are given more weight by the calibration objective (4.2.1) and as such require to be more accurate. To that end, we postulate a joint distribution of moneyness and time to maturity based on liquidity and estimate it using a weighted Gaussian kernel density estimation (wKDE) (Scott, 2015): Let L_i denote the market liquidity of an option $i, i \in \mathbb{N}$, with time to maturity $T^{(i)}$ and moneyness $M^{(i)}$. We proxy liquidities by inverse bid-ask spreads of traded European Call Options on SPX and then run a wKDE on samples $\left\{\left(M^{(i)}, T^{(i)}\right)\right\}_{i=1}^{n}$ with weights $\{L_i\}_{i=1}^{n}$ and a smoothing bandwidth. In a similar vein, one may also derive a multivariate distribution $\mathcal{K}_\xi$ of external market data $\xi \in \mathbb{R}^k$.

With regards to the individual marginals collected in Table 4.1, the sampling distribution $\mathcal{G}_{\text{Heston}}$ on $\mathcal{I} \subseteq \mathbb{R}^{m+k+2}$ is given by

$$\mathcal{G}_{\text{Heston}} := \mathcal{U}^{\otimes m}[a_i, b_i] \otimes \mathcal{K}_\xi \otimes \mathcal{K}_{(M,T)} \tag{4.3.1}$$

and analogously for the rough Bergomi model, we have

$$\mathcal{G}_{\text{rBergomi}} := \mathcal{N}_{\text{trunc}}^{\otimes m}[a_i, b_i, \lambda_i, \sigma_i] \otimes \mathcal{K}_\xi \otimes \mathcal{K}_{(M,T)}. \tag{4.3.2}$$

4.3.2 Backpropagation and hyper parameter optimization

Consider a ReLU FCNN with $L \in \mathbb{N}$ hidden layers as described in Section 4.1.2. Let $n_l, 1 \leq l \leq L$ denote the number of nodes of the hidden layers and $\mathcal{S}_{h_{\text{model}}}$ the function space spanned by such a network with model hyper parameters $h_{\text{model}} = (L, n_1, \dots, n_L)$. Let $X : \Omega \to \mathcal{I}$ denote a random input and consider h_{model} fixed. Then the fundamental objective of neural network training is to learn a function that minimizes the generalization error:

$$f^{\star}_{h_{\text{model}}} = \operatorname*{arg\,min}_{f_{h_{\text{model}}} \in \mathcal{S}_{h_{\text{model}}}} \|f_{h_{\text{model}}}(X) - \widetilde{\varphi}(X)\|^2_{L^2(\Omega)}, \quad X \sim \mathcal{G} \tag{4.3.3}$$

where $\mathcal{G} \in \{\mathcal{G}_{\text{Heston}}, \mathcal{G}_{\text{rBergomi}}\}$, depending on experiment. In many calibration scenarios, $\widetilde{\varphi}$ is a Monte-Carlo approximation to φ, so $\widetilde{\varphi}(\cdot) = \varphi(\cdot) + \varepsilon$ for ε some homoskedastic error with $\mathbb{E}(\varepsilon) = 0$ and $\operatorname{Var}(\varepsilon) = \sigma^2 > 0$. The MSE loss in (4.3.3) admits the well-known bias-variance decomposition

$$\|f_{h_{\text{model}}}(X) - \widetilde{\varphi}(X)\|^2_{L^2(\Omega)} = \left(\mathbb{E}\left[f_{h_{\text{model}}}(X) - \varphi(X)\right]\right)^2 + \operatorname{Var}\left[f_{h_{\text{model}}}(X)\right] + \sigma^2 \tag{4.3.4}$$

where in addition to a bias and variance term we also have the variance of the sample error, the irreducible error. The empirical analogue to (4.3.3) relevant for practical training is given by

$$f^{\star}_{h_{\text{model}}} \approx \operatorname*{arg\,min}_{f_{h_{\text{model}}} \in \mathcal{S}_{h_{\text{model}}}} \frac{1}{n_{\text{valid}}} \sum_{i=1}^{n_{\text{valid}}} \left[f_{h_{\text{model}}}\left(x^{(i)}\right) - \widetilde{\varphi}\left(x^{(i)}\right)\right]^2 \tag{4.3.5}$$

where $\left(x^{(i)}, \widetilde{\varphi}\left(x^{(i)}\right)\right) \in \mathcal{D}_{\text{valid}}$. The optimization in the function space $\mathcal{S}_{h_{\text{model}}}$ corresponds to a high-dimensional nonlinear optimization in the space of network weights and biases, similarly to (4.2.1) typically addressed by gradient-based schemes. *Backpropagation* (Goodfellow et al., 2016), a specific form of *reverse-mode automatic differentiation* (Baydin et al., 2015) in the context of neural networks, prevails as the go-to approach to iteratively compute gradients of the empirical MSE loss with respect to weights and biases of all nodes in the network. The gradients are then often used in the well-known *Mini-Batch Gradient Descent* (Goodfellow et al., 2016) optimization algorithm, a variant of which called *Adam* (Kingma & Ba, 2014) we use in our experiments. *Adam* incorporates momentum to prevent the well-known zigzagging of *Gradient Descent* in long and sharp valleys of the error surface and adaptively modifies a given global step size for each component of the gradient individually to speed up the optimization process. It in turn has its own optimization hyper parameters $h_{\text{opt}} = (\delta, \beta)$ where δ denotes the mentioned global learning rate and β denotes the mini-batch size used. In the following, we denote the learning algorithm *Adam* mapping training data $\mathcal{D}_{\text{train}}$ to a local minimizer

$f^\star_{h_{\text{model}}}$ of (4.3.5) by $\mathcal{A}_{h_{\text{opt}}} : \mathcal{I}^n \to \mathcal{S}_{h_{\text{model}}}$.

Up to know, we treated the hyper parameters $(h_{\text{opt}}, h_{\text{model}})$ as fixed whereas in reality they may be varied and have a crucial influence on the training outcome. Indeed, for all other variables besides $(h_{\text{opt}}, h_{\text{model}})$ fixed, let us define a hyper parameter response function $\mathcal{H}$ by

$$\mathcal{H}(h_{\text{opt}}, h_{\text{model}}) := \frac{1}{n_{\text{valid}}} \sum_{i=1}^{n_{\text{valid}}} \left[\left[\mathcal{A}_{h_{\text{opt}}} \left(\mathcal{D}_{\text{train}} \right) \right]_{h_{\text{model}}} \left(x^{(i)} \right) - \widetilde{\varphi} \left(x^{(i)} \right) \right]^2 . \tag{4.3.6}$$

In practice, it then turns out the real challenge in training neural networks to high accuracy lies in the additional (outer) optimization over hyper parameters:

$$(h^\star_{\text{opt}}, h^\star_{\text{model}}) = \underset{(h_{\text{opt}}, h_{\text{model}})}{\arg\min} \; \mathcal{H}(h_{\text{opt}}, h_{\text{model}}). \tag{4.3.7}$$

The scope of effect of hyper parameters h_{model} and h_{opt} does not overlap: The former determines the capacity of $\mathcal{S}_{h_{\text{model}}}$, the latter governs which local minimizer $f^\star_{h_{\text{model}}}$ the optimization algorithm $\mathcal{A}_{h_{\text{opt}}}$ converges to and the speed with which this happens. This allows us to treat their optimization separately. A coarse grid search reveals that adding additional layers beyond 4 hidden layers does not consistently reduce errors on the validation set. Rather, networks become harder to train as evidenced by errors fluctuating more wildly on the validation set. We suspect this is a consequence of what Ioffe and Szegedy (2015) call *internal covariate shift*: First-order methods such as Gradient Descent are blind to changes in the weights and biases of the layers feeding into a given layer and so with deeper networks the propagating and magnifying effects of changes in one layer to subsequent layers worsen and slow down the training. On the other hand, our locally available compute resources max out at $4096 = 2^{12}$ nodes per layers, so we fix $h_{\text{model}} = (4) \times (4096)^4$. Each evaluation of the hyper parameter response function $\mathcal{H}$ in (4.3.7) requires a ReLU FCNN to be fully trained from scratch which is a very costly operation in terms of time and computing resources. Moreover, gradient-based optimization approaches are ruled out by the fact that gradients of $\mathcal{H}$ with respect to h_{opt} are unavailable (after all, batch sizes are discrete). In our experiments,

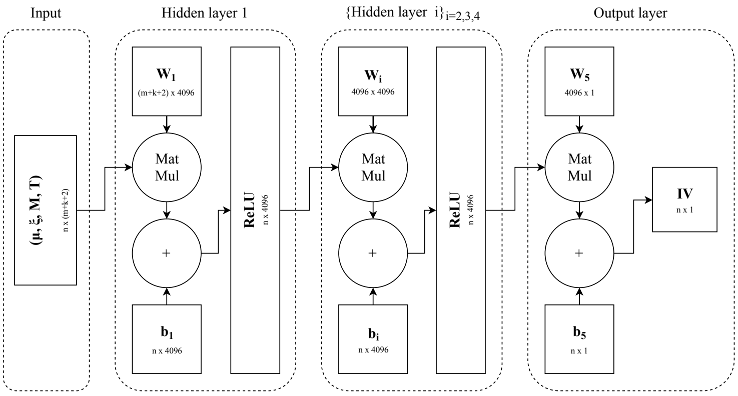


Figure 4.2: **Schematic of IV ReLU FCNN**. Depiction of 4-layer ReLU FCNN used to learn IV maps. It consists of $2^{12} = 4096$ nodes at each hidden layer. Note the output layer is a linear layer with no activation function. Rectangles denote tensors and circles denote operations, *MatMul* is matrix multiplication. The number of parallel IV calculations is given by $n \in \mathbb{N}$.

we explore the use of Gaussian Regression (Rasmussen & Williams, 2006; Snoek, Larochelle, & Adams, 2012) which is an adaptive gradient-free minimization algorithm. Postulating a surrogate Gaussian model for $\mathcal{H}$, it takes existing function evaluations into account and – balancing exploitation and exploration – iteratively proposes the next most promising candidate input in terms of information gain. As is common in applied sciences, we use a Matérn Kernel for the covariance function of the Gaussian model and the Lower Confidence Bound (LCB) acquisition function.

4.3.2.1 Tricks of the trade

Feature scaling or **preconditioning** is a standard preprocessing technique applied to input data of optimization algorithms in order to improve the speed of optimization. After the data set $\mathcal{D}$ has been partitioned into training, validation and test sets, we compute the sample mean $\overline{x}_{\text{train}} \in \mathbb{R}^{m+k+2}$ of the inputs across the training set and the corresponding sample standard deviation $s_{\text{train}} \in \mathbb{R}^{m+k+2}$. For each input $x^{(i)} \in \mathcal{I}$ from $\mathcal{D}$, its standardized version $\widehat{x}^{(i)}$ is given by

$$\widehat{x}^{(i)} := \frac{x^{(i)} - \overline{x}_{\text{train}}}{s_{\text{train}}}, \quad 1 \leq i \leq n.$$

where the operations are defined componentwise. We then use these standardized inputs $\widehat{x}^{(i)}$ – which have zero offset and unit scale – for training and prediction. It is important to stress that *all* n inputs from the complete set $\mathcal{D}$ are standardized using the *training* mean and standard deviation, including those of the validation and test sets.

Weight initialization is an important precursor to the iterative optimization process of *Adam*. Initialization is a delicate task that may speed up or hamper the training process all together: If within (but not necessarily across) all layers, weights and biases of all nodes are identical, then the same is true for their outputs and the partial derivative of the loss with respect to their weights and biases, impeding any learning on the part of the optimizer. To *break the symmetry*, it is standard procedure to draw weights from a symmetric probability distribution centered at zero. Suppose $w_{ij}^{(l)}$ denotes the weight of node $i, 1 \leq i \leq n_l$ in layer $1 \leq l \leq L$ being multiplied with the output of node $j, 1 \leq j \leq n_{l-1}$ in layer $l-1$ and n_0 denotes the number of network inputs. He et al. (2015) suggest the weights and biases be independently drawn as follows

$$w_{ij}^{(l)} \sim \mathcal{N}\left(0, \frac{2}{n_{l-1}}\right), \quad b_i^{(l)} = 0.$$

Adapting an argument by Glorot and Bengio (2010) for linear layers to ReLU networks, they can show that – under some assumptions – this ensures that, at least at initialization, input signals and gradients do not get magnified exponentially during forward or backward passes of backpropagation.

Regularization in the context of regression – be it deterministic in the case of L^2 or L^1 or stochastic in the form of Dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) – describes a set of techniques aimed at modifying a training algorithm so as to reduce overfitting on the training set. With regards to (4.3.4), the conceptual idea is that a modified optimizer allows to trade an increased bias of the estimator for an over proportional decrease in its variance, effectively reducing the MSE overall. In our experiments, we only regularize in time in the form of *early stopping*: While optimizing the weights and biases on the training set, we periodically check the performance on the validation set and save the model if a new minimum error is reached. When the error on the validation set begins to stall, training

Table 4.2: Reference model parameters $\mu^\dagger$ for Heston and rough Bergomi. Obtained from (Gatheral, 2011) and (Bayer et al., 2016) respectively.

Heston		rough Bergomi	
Parameter	Value	Parameter	Value
η	0.3877	η	1.9
ρ	-0.7165	ρ	-0.9
λ	1.3253	H	0.07
$\overline{v}$	0.0354	v_0	0.01
v_0	0.0174		

is stopped.

Batch normalization (BN) devised by Ioffe and Szegedy (2015) is very popular technique to facilitate and accelerate the training of deeper networks by addressing the mentioned *internal covariate shift*. It alters a network's topology by inserting normalization operations between linearities and non-linearities of each dense layer, effectively reducing the dependence of each node's input on the weights and biases of all nodes in previous layers. Our numerical experiments confirm a strongly regularizing effect of BN as is well-known in the literature, reducing the expressiveness of our networks considerably and hence leading to worse performance. Despite its success in allowing to train deeper networks, we hence decided to turn it off.

4.4 Numerical experiments

Here, we examine the performance of our approach by applying it to the option pricing models recalled in Section 4.1: First, we consider the Heston model as a test case and then the rough Bergomi model as a representative from the class of *rough* stochastic volatility models. Specifically, we look at the speed and accuracy of the learned implied volatility map $\varphi_{\mathrm{NN}} : \mathcal{I} \to \mathbb{R}_+$. A systematic comparison of performance metrics between existing methods and our approach has been left for future research.

The Gaussian hyper parameter optimization and individual network training runs are performed on a local CPU-only compute server. Unless otherwise stated, all computations and performance measures referenced in this section are performed on a standard early 2015 Apple Mac Book with a 2.9 GHz Intel

Core i5 CPU with no GPU used.

4.4.1 The Heston model

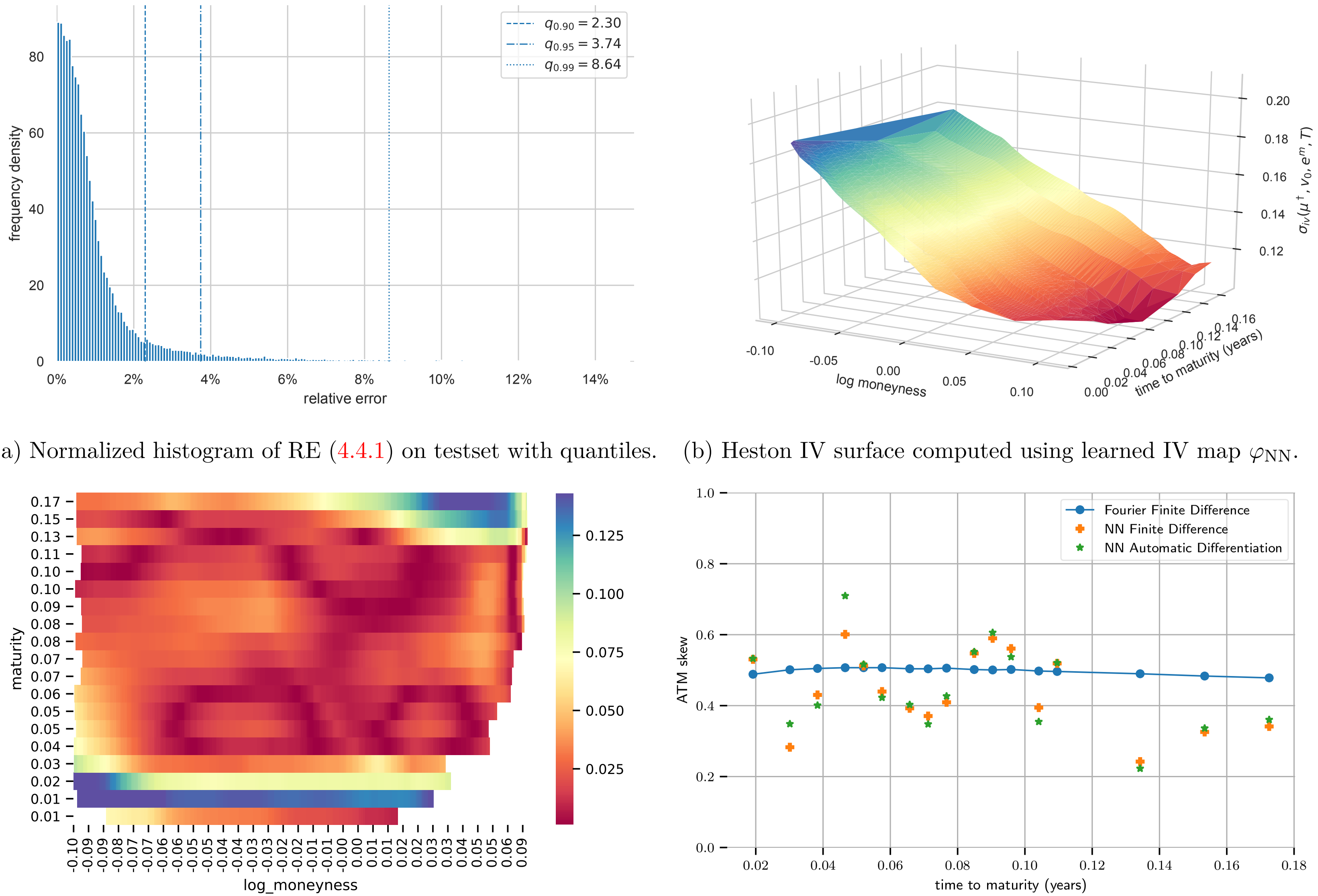
Following the approach outlined in Section 4.3.1, we estimate $\mathcal{K}_{(M,T)}$ using SPX Option Price data[2] from 15th February 2018. Empirically, we observe that a majority of the liquidity as proxied by inverse bid-ask spreads is concentrated in the small region given by $-0.1 \leq m \leq 0.28$ and $\frac{1}{365} \leq T \leq 0.2$ which is why for this test case we exclusively learn the IV map on this bounded domain. The size of the labeled set data $\mathcal{D}$ is $n = 990000$ of which we allocate $n_{\text{train}} = 900000$ samples to the training set and $n_{\text{valid}} = n_{\text{test}} = 45000$ to test and validation sets.

Single evaluations of the learned implied volatility map $\varphi_{\text{NN}} : \mathcal{I} \to \mathbb{R}_+$ and the associated Jacobian $\boldsymbol{J}_{\text{NN}} : O \to \mathbb{R}^{N \times m}$ are extremely fast with about 36ms on average to compute both together, making this neural network based approach at least competitive with existing Fourier-based schemes. To determine the accuracy of φ_{NN}, we define

$$\text{RE}(\mu, m, T) := \frac{|\varphi_{\text{NN}}(\mu, v_0, e^m, T) - \widetilde{\varphi}(\mu, v_0, e^m, T)|}{\widetilde{\varphi}(\mu, v_0, e^m, T)} \tag{4.4.1}$$

to be the relative error of the output of φ_{NN} with respect to that of a Fourier-based reference map $\widetilde{\varphi}$ for model parameters μ, option parameters (m, T) and fixed spot variance v_0. Figure 4.3a shows a normalized histogram of relative errors on the test set where μ and (M, T) are allowed to vary across samples, demonstrating that empirically, φ_{NN} approximates $\widetilde{\varphi}$ with a high degree of accuracy. In typical pricing or calibration scenarios, we are interested in the accuracy of φ_{NN} for some fixed model parameters μ which is why in Figures 4.3b, 4.3c and 4.3d, we fix $\mu = \mu^\dagger$ with $\mu^\dagger$ the reference model parameters in Table 4.2. In Figure 4.3b, we compute an IV point cloud using φ_{NN}, interpolate it using a (not necessarily arbitrage-free) Delaunay triangulation and recover a characteristic Heston-like model IV surface. Indeed, as the heatmap of interpolated relative errors in Figure 4.3c shows, these are small across most of the IV surface, with increased relative errors only for times on

[2] Option prices for SPX Weeklys can be retrieved from a publicly available database at www.cboe.com/DelayedQuote/QuoteTableDownload.aspx.



(a) Normalized histogram of RE (4.4.1) on testset with quantiles.

(b) Heston IV surface computed using learned IV map φ_{NN}.

(c) Interpolated heatmap of $\mathrm{RE}(\mu^\dagger, m, T)$ (4.4.1) for varying m, T.

(d) Approximations to the ATM volatility skew.

Figure 4.3: **Accuracy of the IV map φ_{NN} as learned by the Heston ReLU FCNN.**

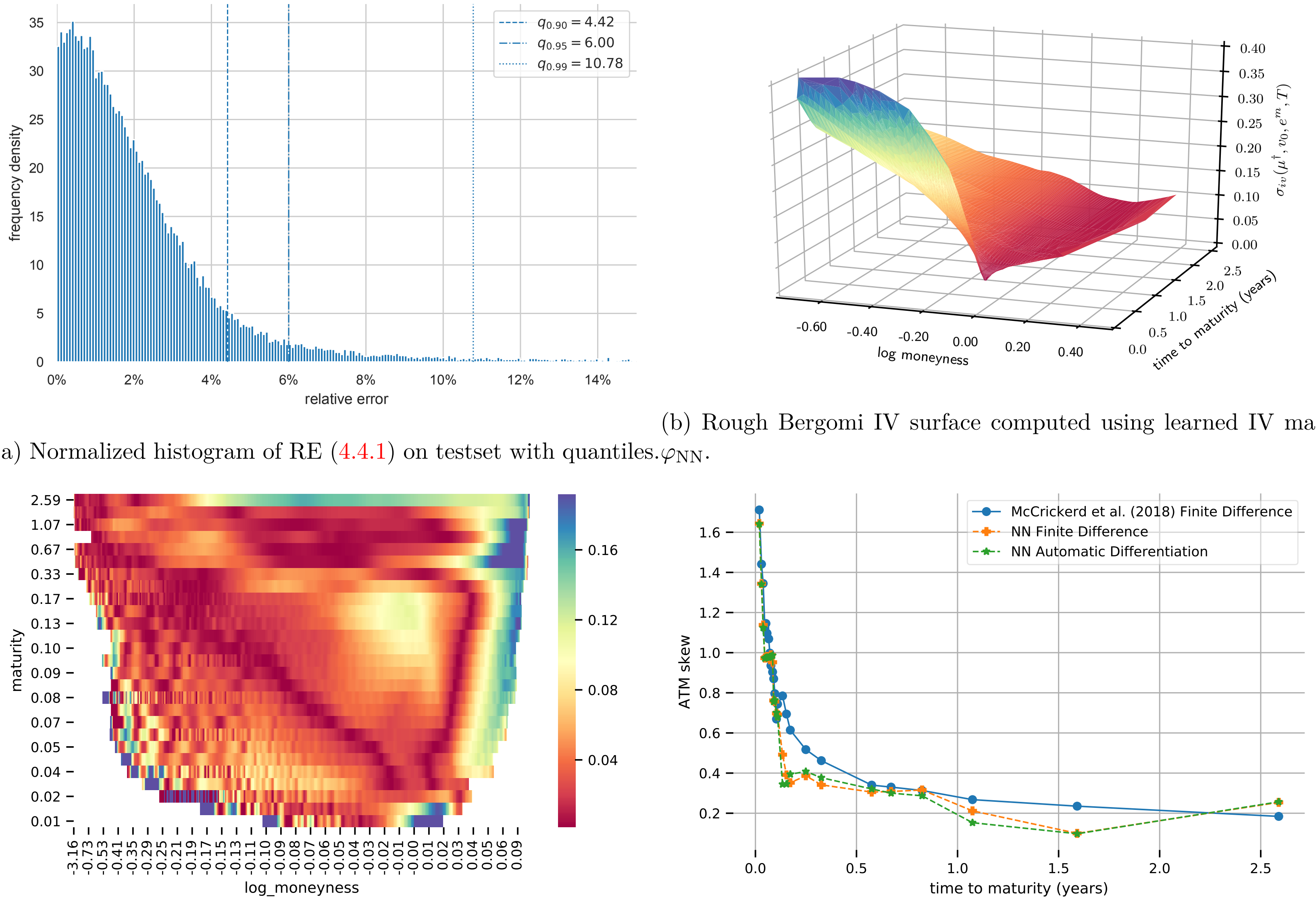
the short and long end which may be attributed to less training because of less liquidity. Finally, in Figure 4.3d, we plot three different approximations to the Heston ATM volatility skew for small times: A reference skew in blue obtained by a finite difference approximation using $\widetilde{\varphi}$, another skew in orange obtained by the same method but applied to φ_{NN} and finally the exact ATM skew of φ_{NN} in green, available by automatic differentiation. As is to be expected, φ_{NN} recovers the characteristic flat behaviour for short times, the general drawback of bivariate diffusion models such as Heston.

4.4.2 The rough Bergomi model

For simplicity, we consider the rough Bergomi model as introduced in Example 4.1.2 with a flat forward variance curve $\xi_0(t) = v_0 \in \mathbb{R}_+$ for $t \geq 0$. For the remainder of this work, we shall consider v_0 an additional model parameter. Again, following the approach outlined in Section 4.3.1, we estimate $\mathcal{K}_{(M,T)}$ using SPX Option Price data, this time from 19th May 2017[3]. We do not restrict the option parameter region considered and learn the whole surface with parameter bounds given by $-3.163 \leq m \leq 0.391$ and $0.008 \leq T \leq 2.589$. Of the one million synthetic data pairs sampled, 90% are allocated to the training set and 5% to validation and test sets respectively.

Recall that in this experiment we use the same network topology as in the Heston example. As is to be expected, the speed of single evaluations of the learned rough Bergomi IV map $\varphi_{\text{NN}} : \mathcal{I} \to \mathbb{R}_+$ and the associated Jacobian $\boldsymbol{J}_{\text{NN}} : O \to \mathbb{R}^{N\times m}$ are hence of the same order with about 36ms to compute both objects together, beating state of the art methods by magnitudes. Intuitively, the non-Markovian nature of rough Bergomi manifests itself in an increased model complexity and so it is unsurprising that the general accuracy of the rough Bergomi IV map φ_{NN} on the rough Bergomi test set is lower than its counterpart on the Heston test set (cf. 4.4a). On the other hand, for fixed model parameters $\mu = \mu^{\dagger}$ (cf. Table 4.2), the implied volatility map φ_{NN} recovers the characteristic rough Bergomi model IV surface (Figure 4.4b) with low relative error across most of the liquid parts of the IV surface (Figure 4.4c). It also exhibits the striking power law behaviour of the ATM volatility skew near zero (Figure 4.4d).

[3] Thanks to Jim Gatheral for providing us with this data set.



(a) Normalized histogram of RE (4.4.1) on testset with quantiles.

(b) Rough Bergomi IV surface computed using learned IV map φ_{NN}.

(c) Interpolated heatmap of $\text{RE}(\mu^\dagger, m, T)$ (4.4.1) for varying m, T.

(d) Approximations to the ATM volatility skew.

Figure 4.4: **Accuracy of the IV map φ_{NN} as learned by the rough Bergomi ReLU FCNN.**

On the contrary, measuring the accuracy of the neural-network enhanced Levenberg-Marquardt scheme introduced in Section 4.2.1 is not a straightforward task. To see why, consider the small-time asymptotic formula for the BS implied volatility σ_{iv} of rough stochastic volatility models as derived by Bayer et al. (2018). With scaling parameter $\beta < \frac{2}{3}H$, their expansion applied to our setting yields

$$\sigma_{\text{iv}}(e^{k_t}, t) = \sqrt{v_0} + \frac{1}{2}\rho\eta\, C(H)kt^{\beta} + \mathcal{O}(t) \tag{4.4.2}$$

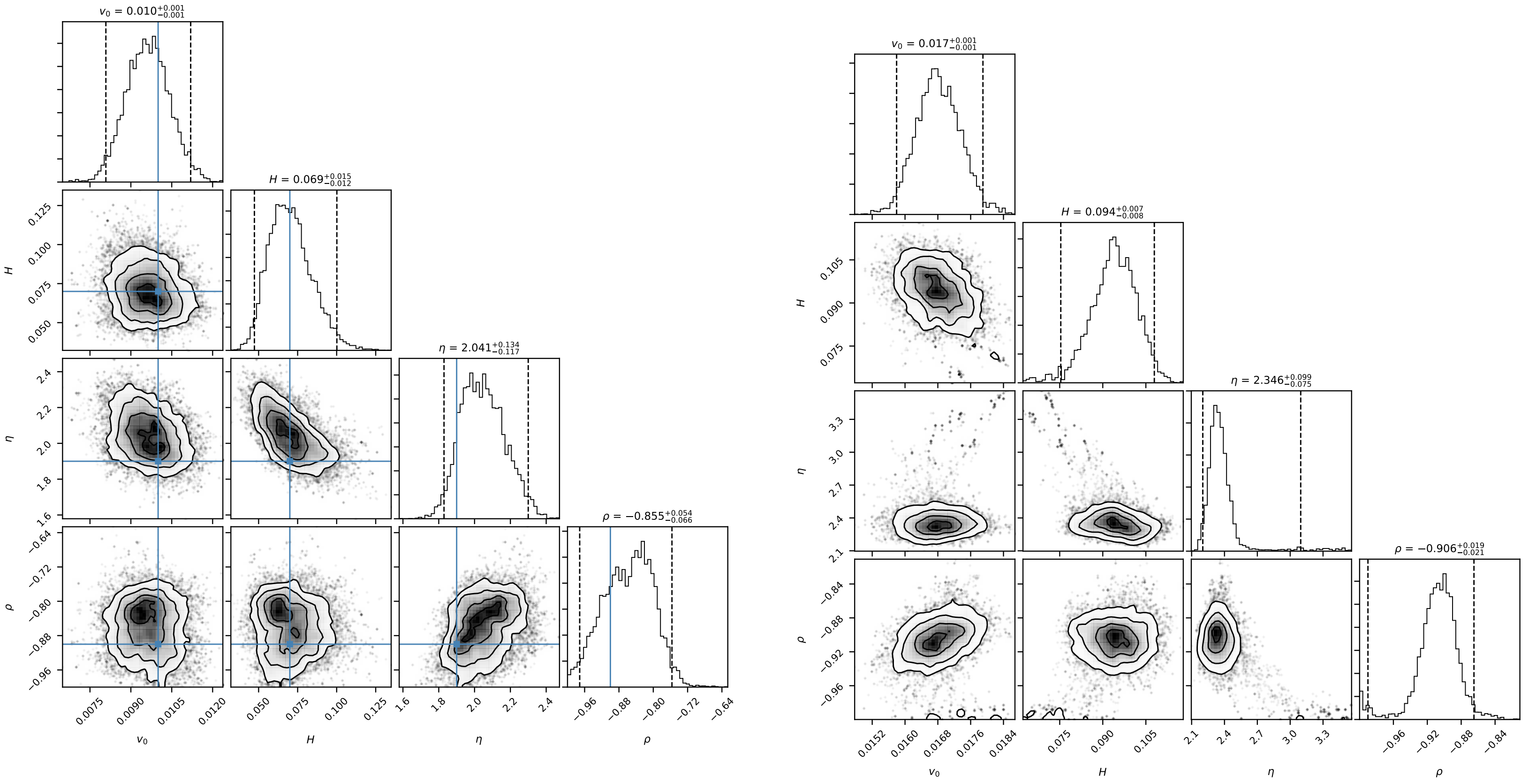
for small times $t \to 0$, time-scaled log moneyness $k_t = kt^{\frac{1}{2}-H+\beta}$ and constant $C(H)$ depending on H. Hence, at least for small times, all three model parameters enter multiplicatively either directly (ρ and η) or indirectly (H) into the second term in (4.4.2) which corrects the crude estimate given by spot volatility. A decrease in $|\rho|$ could hence for example be offset by an adequate increase in η and still yield the same IV. Mathematically speaking, for fixed moneyness and time to maturity, it is thus to be expected that the map φ_{NN} is non-injective in its model parameters on large parts of its model parameter input domain. Quantifying the accuracy of the deep calibration scheme by computing any form of distance between true and calibrated model parameters in model parameter space is hence nonsensical.

4.4.2.1 Bayesian parameter inference

Intuitively, we are interested in *quantifying the uncertainty* about model parameter estimates obtained by calibrating with the approximative IV map φ_{NN}. To this end, we switch to a Bayesian viewpoint and treat model parameters μ as random variables. The fundamental idea behind Bayesian parameter inference is to update prior beliefs $p(\mu)$ formalised in (4.3.2) with the likelihood $p(\boldsymbol{y} \mid \mu)$ of observing a given IV point cloud $\boldsymbol{y} \in \mathbb{R}^N$ to deduce a posterior (joint) distribution $p(\mu \mid \boldsymbol{y})$ over model parameters μ.

Formally, for pairs $\left(M^{(i)}, T^{(i)}\right)$ of moneyness & time to maturity, let an IV point cloud to calibrate against be given by

$$\boldsymbol{y} = \left[y_1\left(M^{(1)}, T^{(1)}\right), \dots, y_N\left(M^{(N)}, T^{(N)}\right)\right]^T \in \mathbb{R}^N$$



(a) Bayes calibration against synthetic IV surface computed for model parameters $\mu^\dagger$. Solid vertical blue lines indicate true parameter values.

(b) Liquidity-weighted Bayes calibration against SPX market IV surface from 19th May 2017. Liquidity proxies given by inverse bid-ask-spreads.

Figure 4.5: **1d- and 2d-projections of 4d Bayesian posterior over rBergomi model parameters.** On diagonal, univariate histograms of model parameters with titles stating median ± the delta to 2.5% and 97.5% quantiles. Dashed vertical lines indicate those quantiles. Off-diagonal, 2d histograms of 2d projections of MCMC samples together with isocontours from 2d Gaussian KDE.

and analogously, collect model IVs for model parameters μ as follows

$$\boldsymbol{\varphi}_{\mathrm{NN}}(\mu) = \left[\varphi_{\mathrm{NN}}\left(\mu, M^{(1)}, T^{(1)}\right), \ldots, \varphi_{\mathrm{NN}}\left(\mu, M^{(N)}, T^{(N)}\right)\right]^T \in \mathbb{R}^N.$$

We perform a liquidity-weighted nonlinear Bayes regression. Mathematically, for heteroskedastic sample errors $\sigma_i > 0, i = 1, \ldots, N$, we postulate

$$\boldsymbol{y} = \boldsymbol{\varphi}_{\mathrm{NN}}(\mu) + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}\left(0, \operatorname{diag}[\sigma_1^2, \ldots, \sigma_N^2]\right)$$

so that for some diagonal weight matrix $\boldsymbol{W} = \operatorname{diag}[w_1, \ldots, w_N] \in \mathbb{R}^{N\times N}$, the liquidity-weighted residuals are distributed as follows

$$\boldsymbol{W}^{\frac{1}{2}}\left[\boldsymbol{y} - \boldsymbol{\varphi}_{\mathrm{NN}}(\mu)\right] \sim \mathcal{N}\left(0, \operatorname{diag}[w_1\sigma_1^2, \ldots, w_N\sigma_N^2]\right).$$

In other words, we assume that the joint likelihood $p(\boldsymbol{y} \mid \mu)$ of observing data $\boldsymbol{y}$ is given by a multivariate normal. In absence of an analytical expression for the posterior (joint) probability $p(\mu|\boldsymbol{y}) \propto p(\boldsymbol{y}|\mu)p(\mu)$, we approximate it numerically using MCMC techniques (Foreman-Mackey, Hogg, Lang, & Goodman, 2013) and plot the one- and two-dimensional projections of the four-dimensional posterior by means of an MCMC plotting library (Foreman-Mackey, 2016).

We perform two experiments. First, fixing $\mu = \mu^\dagger$, we generate a synthetic IV point cloud

$$\boldsymbol{y}_{\mathrm{synth}} = \left[\widetilde{\varphi}\left(\mu^\dagger, M^{(1)}, T^{(1)}\right), \ldots, \widetilde{\varphi}\left(\mu^\dagger, M^{(N)}, T^{(N)}\right)\right] \in \mathbb{R}^N$$

using the reference method $\widetilde{\varphi}$. Next, we perform a non-weighted Bayesian calibration against the synthetic surface and collect the numerical results in Figure 4.5a. If the map φ_{NN} is sufficiently accurate for calibration, the computed posterior should attribute a large probability mass around $\mu^\dagger$. The results in Figure 4.5a are quite striking in several ways: (1) From the univariate histograms on the diagonal it is clear that the calibration routine has identified sensible model parameter regions covering the true values. (2) Histograms are unimodal and its peaks close or identical to the true parameters. (3) The isocontours of the 2d Gaussian KDE in the off-diagonal pair plots for (η, H) and (η, ρ) show exactly the behaviour expected from the reasoning in the last section: Since increases or decreases in one of η, H or ρ can be offset by adequate changes in the others with no impact on the calculated IV, the Bayes posterior cannot discriminate between such parameter configurations and places equal probability on both combinations. This can be seen by the diagonal elliptic probability level sets.

In a second experiment, we want to check whether the inaccuracy of φ_{NN} allows for a successful calibration against market data. To this end, we perform a liquidity-weighted Bayesian regression against SPX IVs from 19th May 2017. For bid and ask IVs $a_i > 0$ and $b_i > 0$ respectively, we proxy the IV of the mid price by $m_i := \frac{a_i+b_i}{2}$. With spread defined by $s_i = a_i - b_i \geq 0$, all options with $s_i/m_i \geq 5\%$ are removed because of too little liquidity. Weights are chosen to be $w_i = \frac{m_i}{a_i - m_i} \geq 0$, effectively taking inverse bid-ask spreads as a proxy for liquidity. Finally, σ_i are proxied by a fractional of the spread s_i. The numerical results in Figure 4.5b further confirm the accuracy of φ_{NN}: (1) As can be seen on the univariate histograms on the diagonal, the Bayes calibration has again identified sensible model parameter regions in line with what is to expected. (2) Said histograms are again unimodal with peaks at or close to values previously reported by Bayer et al. (2016). (3) Quite strikingly, at a first glance, the effect of the diagonal probability level sets in the off-diagonal plots as documented in Figure 4.5a cannot be confirmed here. However, the scatter plots in the diagrams do reveal some remnants of that phenomenon.

Acknowledgments

The authors thank Jim Gatheral and Peter K. Friz for some helpful discussions and suggestions and acknowledge financial support through DFG Research Grants BA5484/1 and FR2943/2.

Bibliography

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., . . . Isard, M. (2016). Tensorflow: a system for large-scale machine learning. In *OSDI* (Vol. 16, pp. 265–283).

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

Ametrano, F., Ballabio, L., Bianchetti, M., Césaré, N., Eddelbuettel, D., Firth, N., . . . Marchioro, M. (2015). *Quantlib: A free/open-source library for quantitative finance.*

Azencott, R. (1982). Formule de Taylor stochastique et développement asymptotique d'intégrales de Feynman. In *Seminar on Probability, XVI, Supplement* (Vol. 921, pp. 237–285). Springer, Berlin-New York.

Azencott, R. (1985). Petites perturbations aléatoires des systemes dynamiques: développements asymptotiques. *Bulletin des sciences mathématiques*, *109*(3), 253–308.

Baudoin, F., & Ouyang, C. (2015). On Small Time Asymptotics for Rough Differential Equations Driven by Fractional Brownian Motions. In P. K. Friz, J. Gatheral, A. Gulisashvili, A. Jacquier, & J. Teichmann (Eds.), *Large Deviations and Asymptotic Methods in Finance* (pp. 413–438). Springer International Publishing.

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., & Siskind, J. M. (2015). *Automatic Differentiation in Machine Learning: a Survey.*

Bayer, C., Friz, P. K., Gassiat, P., Martin, J., & Stemper, B. (2017). A regularity structure for rough volatility. *ArXiv e-prints.*

Bayer, C., Friz, P. K., & Gatheral, J. (2016). Pricing under rough volatility. *Quantitative Finance*, *16*(6), 887-904.

Bayer, C., Friz, P. K., Gulisashvili, A., Horvath, B., & Stemper, B. (2018). Short-time near-the-money skew in rough fractional volatility models. *Quantitative Finance*, 1-20. doi: 10.1080/14697688.2018.1529420

Bayer, C., & Stemper, B. (2018). Deep calibration of rough stochastic volatility models. *ArXiv e-prints*.

Ben Arous, G. (1988). Methods de Laplace et de la phase stationnaire sur l'espace de Wiener. *Stochastics*, *25*(3), 125–153.

Bennedsen, M., Lunde, A., & Pakkanen, M. S. (2016). Decoupling the short- and long-term behavior of stochastic volatility. *ArXiv e-prints*.

Bennedsen, M., Lunde, A., & Pakkanen, M. S. (2017). Hybrid scheme for Brownian semistationary processes. *Finance and Stochastics*, *21*(4), 931–965.

Berestycki, H., Busca, J., & Florent, I. (2004). Computing the implied volatility in stochastic volatility models. *Communications on Pure and Applied Mathematics*, *57*(10), 1352–1373.

Bismut, J.-M. (1984). *Large deviations and the Malliavin calculus* (Vol. 45). Birkhäuser Boston, Inc., Boston, MA.

Black, F., & Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, *81*(3), 637-654.

Carlen, E., & Kree, P. (1991). Lp estimates on iterated stochastic integrals. *The Annals of Probability*, 354–368.

Cass, T., & Friz, P. K. (2010). Densities for rough differential equations under Hörmander's condition. *Annals of Mathematics*, 2115–2141.

Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). *ArXiv e-prints*.

Decreusefond, L. (2005). Stochastic integration with respect to Volterra processes. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, *41*(2), 123 - 149.

Deuschel, J.-D., Friz, P. K., Jacquier, A., & Violante, S. (2014a). Marginal density expansions for diffusions and stochastic volatility II: Applications. *Comm. Pure Appl. Math.*, *67*(2), 321–350.

Deuschel, J.-D., Friz, P. K., Jacquier, A., & Violante, S. (2014b). Marginal density expansions for diffusions and stochastic volatility I: Theoretical foundations. *Comm. Pure Appl. Math.*, *67*(1), 40–82.

Deuschel, J.-D., & Stroock, D. W. (1989). *Large deviations* (Vol. 137). Academic press Boston.

Elfwing, S., Uchibe, E., & Doya, K. (2018). Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*.

Euch, O. E., & Rosenbaum, M. (2018). The characteristic function of rough Heston models. *Mathematical Finance*.

Forde, M., & Jacquier, A. (2009). Small-time asymptotics for implied volatility under the Heston model. *International Journal of Theoretical and Applied Finance*, *12*(06), 861–876.

Forde, M., & Zhang, H. (2017). Asymptotics for rough stochastic volatility models. *SIAM Journal on Financial Mathematics*, *8*(1), 114-145.

Foreman-Mackey, D. (2016). corner. py: Scatterplot matrices in python. *The Journal of Open Source Software*, *1*.

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. (2013). emcee: the MCMC hammer. *Publications of the Astronomical Society of the Pacific*, *125*(925), 306.

Friz, P. K., & Gassiat, P. (2018). *Martingality and moments for lognormal rough volatility.*

Friz, P. K., Gerhold, S., & Pinter, A. (2018). Option pricing in the moderate deviations regime. *Mathematical Finance*, *28*(3), 962–988.

Friz, P. K., & Hairer, M. (2014). *A course on rough paths.* Springer.

Fukasawa, M. (2011). Asymptotic analysis for stochastic volatility: martingale expansion. *Finance and Stochastics*, *15*(4), 635–654.

Fukasawa, M. (2017). Short-time at-the-money skew and rough fractional volatility. *Quantitative Finance*, *17*(2), 189-198.

Gao, K., & Lee, R. (2014). Asymptotics of implied volatility to arbitrary order. *Finance and Stochastics*, *18*(2), 349–392.

Gatheral, J. (2011). *The volatility surface: a practitioner's guide.* John Wiley & Sons.

Gatheral, J., Jaisson, T., & Rosenbaum, M. (2018). Volatility is rough. *Quantitative Finance*, 1–17.

Glasserman, P. (2003). *Monte Carlo methods in Financial Engineering* (Vol. 53). Springer Science & Business Media.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics* (pp. 249–256).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* (Vol. 1). MIT Press Cambridge.

Guennoun, H., Jacquier, A., Roome, P., & Shi, F. (2014). Asymptotic behaviour of the fractional Heston model. *ArXiv e-prints.*

Gulisashvili, A. (2017). Large deviation principle for Volterra type fractional stochastic volatility models. *ArXiv e-prints.*

Hagan, P. S., Kumar, D., Lesniewski, A. S., & Woodward, D. E. (2002). Managing smile risk. *The Best of Wilmott*, *1*, 249–296.

Hairer, M. (2014). A theory of regularity structures. *Inventiones mathematicae*, *198*(2), 269–504.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1026–1034).

Hernandez, A. (2017). Model calibration with neural networks. *Risk.*

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, *6*(2), 327–343.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, *4*(2), 251–257.

Horvath, B., Jacquier, A., & Muguruza, A. (2017). Functional central limit theorems for rough volatility. *ArXiv e-prints.*

Hull, J., & White, A. (1990). Pricing interest-rate-derivative securities. *The Review of Financial Studies*, *3*(4), 573–592.

Hull, J., & White, A. (1993). One-factor interest-rate models and the valuation of interest-rate derivative securities. *Journal of Financial and Quantitative Analysis*, *28*(2), 235–254.

Inahama, Y. (2013, 01). Laplace approximation for rough differential equation driven by fractional brownian motion. *Ann. Probab.*, *41*(1), 170–205.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ArXiv e-prints.*

Jäckel, P. (2015). Let's be rational. *Wilmott*, *2015*(75), 40–53.

Jacquier, A., Pakkanen, M. S., & Stone, H. (2017). Pathwise large deviations for the Rough Bergomi model. *ArXiv e-prints.*

Jourdain, B. (2004). Loss of martingality in asset price models with lognormal stochastic volatility. *International Journal of Theoretical and Applied Finance*, *13*, 767–787.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *ArXiv e-prints*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems* (pp. 1097–1105).

Lamperti, J. (1962). Semi-stable stochastic processes. *Transactions of the American Mathematical Society*, *104*(1), 62–78.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436.

Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, *2*(2), 164–168.

Lions, P.-L., & Musiela, M. (2007). Correlations and bounds for stochastic volatility models. In *Annales de l'institut henri poincare (c) non linear analysis* (Vol. 24, pp. 1–16).

Mandelbrot, B. B., & Van Ness, J. W. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Review*, *10*(4), 422–437.

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, *11*(2), 431–441.

McCrickerd, R., & Pakkanen, M. S. (2018). Turbocharging Monte Carlo pricing for the rough Bergomi model. *Quantitative Finance*, 1–10.

Medvedev, A., & Scaillet, O. (2003). A simple calibration procedure of stochastic volatility models with jumps by short term asymptotics. *SSRN*(477441). Preprint.

Medvedev, A., & Scaillet, O. (2007). Approximation and calibration of short-term implied volatilities under jump-diffusion stochastic volatility. *Review of Financial Studies*, *20*(2), 427–459.

Mijatović, A., & Tankov, P. (2016). A new look at short-term implied volatility in asset price models with jumps. *Mathematical Finance*, *26*(1), 149–183.

Muhle-Karbe, J., & Nutz, M. (2011, 12). Small-time asymptotics of option prices and first absolute moments. *J. Appl. Probab.*, *48*(4), 1003–1020.

Neuenkirch, A., & Shalaiko, T. (2016). The order barrier for strong approximation of rough volatility models. *arXiv preprint arXiv:1606.03854*.

Olver, F. W. (2010). *Nist Handbook of Mathematical Functions.* Cambridge University Press.

Osajima, Y. (2007). The asymptotic expansion formula of implied volatility for dynamic SABR model and FX hybrid model. *SSRN*(965265).

Osajima, Y. (2015). General Asymptotics of Wiener Functionals and Application to Implied Volatilities. In P. K. Friz, J. Gatheral, A. Gulisashvili, A. Jacquier, & J. Teichmann (Eds.), *Large deviations and asymptotic methods in finance* (pp. 137–173). Cham: Springer International Publishing.

Pham, H. (2010). Large deviations in finance. *Third SMAI European Summer School in Financial Mathematics.*

Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. *ArXiv e-prints.*

Rasmussen, C. E., & Williams, C. K. (2006). *Gaussian process for machine learning.* MIT Press.

Romano, M., & Touzi, N. (1997). Contingent claims and market completeness in a stochastic volatility model. *Mathematical Finance*, *7*(4), 399–412.

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization.* John Wiley & Sons.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., . . . Bolton, A. (2017). Mastering the Game of Go without Human Knowledge. *Nature*, *550*(7676), 354.

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv e-prints.*

Sin, C. A. (1998, 03). Complications with stochastic volatility models. *Adv. in Appl. Probab.*, *30*(1), 256–268.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems* (pp. 2951–2959).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., & Anguelov, D. (2015).

Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–9).

Willard, G. A. (1997). Calculating prices and sensitivities for path-independent derivatives securities in multifactor models. *The Journal of Derivatives*, *5*(1), 45–61.

Wong, E., & Zakai, M. (1965). On the convergence of ordinary integrals to stochastic integrals. *Ann. Math. Statist.*, *36*, 1560–1564.