Small-time and large-time smile behaviour for the Rough Heston model

Martin Forde*  Stefan Gerhold†  Benjamin Smith‡

June 24, 2019

Abstract

We characterize the asymptotic small-time and large-time implied volatility smile for the popular Rough Heston model introduced by Jaisson & Rosenbaum [JR15]. We show that the asymptotic short-maturity smile scales in qualitatively the same way as a general rough stochastic volatility model (cf. [FZ17], [FGP18a] et al.), and the rate function is equal to the Fréchet-Legendre transform of a simple transformation of the solution to the same Volterra integral equation (VIE) that appears in [ER19], but with the drift and mean reversion terms removed. The solution to this VIE satisfies a space-time scaling property which means we only need to solve this equation for the moment values of $p = 1$ and $p = -1$ so the rate function can be efficiently computed using an Adams scheme or a power series, and we compute a power series in the log-moneyness variable for the asymptotic implied volatility which yields tractable expressions for the vol skew and convexity which is useful for calibration purposes. We later derive formal asymptotics for the small-time moderate deviations regime (previously considered in [FGP18b], [BFCHS18]) and a formal saddlepoint approximation for call options in the large deviations regime which goes to higher order than previous works for rough models. Our higher order expansion captures the effect of both drift terms, and at leading order is of qualitatively the same form as the higher order expansion for a general model which appears in [FGP18a]. The limiting asymptotic smile in the large-maturity regime is obtained via a stability analysis of the fixed points of the VIE, and turns out to be the same as for the standard Heston model in [FJ11] (for which there is a well known closed-form formula in terms of the SVI parametrization given in [GJ11]).

1 Introduction

[JR15] introduced the Rough Heston stochastic volatility model and show that the model arises naturally as the large-time limit of a high frequency market microstructure model driven by two nearly unstable self-exciting Poisson processes (otherwise known as Hawkes process) with a Mittag-Leffler kernel which drives buy and sell orders (a Hawkes process is a generalized Poisson process where the intensity is itself stochastic and depends on the jump history via the kernel). The microstructure model captures the effects of endogeneity of the market, no-arbitrage, buying/selling asymmetry and the presence of metaorders. [ER18] show that the characteristic function of the log stock price for the Rough Heston model is the solution to a fractional Riccati equation which is non-linear (see also [EFR18] and [ER18]), and the variance curve for the model evolves as $dξ(t) = 1/(\kappa - \kappa_t) dV_t$, where $\kappa_t$ is the kernel for the $V_t$ process itself multiplied by a Mittag-Leffler function (see Proposition 2.2 below for a proof of this). Theorem 2.1 in [ER18] shows that a Rough Heston model conditioned on its history up to some time is still a Rough Heston model, but with a time-dependent mean reversion level $\theta(t)$ which depends on the history of the $V$ process. Using Fréchet derivatives, [ER18] also show that one can replicate a call option under the Rough Heston model if we have tradeable variance swaps at all maturities. More generally, we can replicate any Malliavin differentiable contingent claim under any two-dimensional Rough Stochastic volatility model with dynamic trading in the stock and a dynamic trading in a forward variance contract, using the Clark-Ocone formula for two-dimensional Brownian motion (explicit calculations in this respect are much easier for e.g. the Rough Bergomi and fractional Stein models than the Rough Heston model, since the latter is defined implicitly).

[EGR18] derive a quick and dirty (albeit useful) trick for approximating the Rough Heston model with a standard Heston model with the vol-of-vol parameter appropriately re-scaled, which comes from matching the second moment of the integrated variance for the two models. [GR18] propose a global Padé-type rational function approximation to the true solution of the Rough Heston VIE which asymptotically agrees with the true solution at small and large maturities, and option pricing via Fourier inversion using this approximation is reported as being fast and accurate.

[GR18] consider the more general class of affine forward variance (AFV) models of the form $dξ_h(t) = \kappa_h(u - t) \sqrt{V_t} dW_t$ (for which the Rough Heston model is a special case). They show that AFV models arise naturally as the weak limit of a so-called affine forward intensity (AFI) model, where order flow is driven by two generalized Hawkes-type process with an arbitrary jump size distribution, and we exogenously specify the evolution of the conditional expectation of the intensity at different maturities in the future, akin to a variance curve model. The weak limit here involves letting the jump size tend to zero as the jump intensity tends to infinity in a certain way, and one can argue that an AFI model

---

*Dept. Mathematics, King’s College London, Strand, London, WC2R 2LS (Martin.Forde@kcl.ac.uk)
†TU Wien, Financial and Actuarial Mathematics, Wiedner Hauptstraße 8/105-1, A-1040 Vienna, Austria (sgerhold@fam.tuwien.ac.at)
‡Dept. Mathematics, King’s College London, Strand, London, WC2R 2LS (Benjamin.Smith@kcl.ac.uk)
is more realistic than the bivariate Hawkes model in [ER19], since the latter only allows for jumps of a single magnitude (which correspond to buy/sell orders). Using martingale arguments (which do not require considering a Hawkes process as in the aforementioned El Euch & Rosenbaum articles) they show that the mgf of the log stock price for the affine variance model satisfies a convolution Riccati equation, or equivalently is a non-linear function of the solution to a VIE. Formally at least, one can also compute the next order term associated with the [GK18] convergence result, which we can view as an expansion around the limiting AFV model; the correction term satisfies a linear VIE and Fourier inversion has to be applied to the correction term for e.g. pricing a call option.

[GGP18] use comparison principle arguments for VIEs to show that the moment explosion time for the Rough Heston model is finite if and only if it is finite for the standard Heston model. [GGP18] also establish upper and lower bounds for the explosion time, and show that the critical moments are finite for all maturities, and formally derive refined tail asymptotics for the Rough Heston model using Laplace’s method. A recent talk by M. Keller-Ressel (joint work with Majid) states an alternate upper bound for the moment explosion time for the Rough Heston model, based on a comparison with a (deterministic) time-change of the standard Heston model, which they claim is usually sharper than the bound in [GGP18].

Corollary 7.1 in [FGP18a] provides a sharp small-time expansion in the [FZ17] large deviations regime (valid for x-values in some interval) for a general class of Rough Stochastic volatility models using regularity structures, which provides the next order correction to the leading order behaviour obtained in [FZ17], and some earlier intermediate results in Bayer et al. [BFGHS18], [EFGR18] derive a higher order Edgeworth expansion for implied volatility in the central limit theorem regime where the log moneyness scales as k√t as t → 0 for a certain class of models which includes the Rough Bergomi model with a non-flat initial variance curve structure, using a Fourier transform approach with an asymptotic expansion of the characteristic function of the log stock price. This complements the lower order expansion in [Fuk17] for a more general model, and [FSV19] derive formal small-time Edgeworth expansions for the Rough Heston model by solving a nested sequence of linear VIEs. The implied vol expansions in [EFGR18] and [FSV19] both include an additional O(T^{2H}) term, which itself contains an at-the-money, convexity and higher order correction term, which are important effects to capture in practice.

In this article, we establish small-time and large-time large deviation principles for the Rough Heston model, via the solution to a VIE, and we translate these results into asymptotic estimates for call options and implied volatility. The solution to the VIE satisfies a certain scaling property which means we only have to solve the VIE for the moment values of the solution to a VIE, and we translate these results into asymptotic estimates for call options and implied volatility. The mean reversion term and the drift of the log stock price, and we discuss practical issues and limitations of this result.

2 Rough Heston and other variance curve models - basic properties

In this section, we recall the definition and basic properties and origins of the Rough Heston model, and more general affine and non-affine forward variance models. Most of the results in this section are given in various locations in [ER18, ER19] and [GK18], but for pedagogical purposes we found it instructive to collate them together in one place.

Let (Ω, ℱ, ℙ) denote a probability space with filtration (ℱt)_{t≥0} which satisfies the usual conditions, and consider the Rough Heston model for a log stock price process Xt introduced in [JR10]:

\[ dX_t = -\frac{1}{2} V_t dt + \sqrt{V_t} dB_t \]

\[ V_t = V_0 + \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} \lambda (\theta - V_s) ds + \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} \nu \sqrt{V_s} dW_s \]

(1)

for \( \alpha \in (\frac{1}{2}, 1) \), \( \theta > 0 \), \( \lambda \geq 0 \) and \( \nu > 0 \), where W, B are two \( \mathcal{F}_t \)-Brownian motions with correlation \( \rho \in (-1, 1) \). We assume \( X_0 = 0 \) and zero interest rate without loss of generality, since the law of \( X_t - X_0 \) is independent of \( X_0 \).

2.1 Computing \( E(V_t) \)

Proposition 2.1

\[ E(V_t) = V_0 - (V_0 - \theta) \int_0^t f^{\alpha, \lambda}(s) ds \]

where \( f^{\alpha, \lambda}(t) := \lambda t^{\alpha-1} E_{\alpha, \alpha}(-\lambda t^\alpha) \), and \( E_{\alpha, \beta}(z) := \sum_{n=0}^{\infty} \frac{z^n}{\Gamma(n+1, \beta)} \) denotes the Mittag-Leffler function.
If \( \lambda \) (see \([ER19]\)). For Proposition 2.2 we substitute this expression into (4) to get:

\[
\kappa \xi \text{ and substituting the expression for \( \xi \) is the expectation of (1) and using that the expectation of the stochastic integral term is zero, we see that}
\]

\[
\mathbb{E}(V_t) = V_0 + \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} \lambda (\theta - \mathbb{E}(V_s)) ds.
\]

(3)

Let \( k(t) := \frac{M_{t-1}^\alpha}{\Gamma(\alpha)} \) and \( f(t) := \mathbb{E}(V_t) - \theta \). Then we can re-write (3) as

\[
f(t) = (V_0 - \theta) - k \ast f(t).
\]

(4)

where \( \ast \) denotes convolution. Now define the resolvent \( r(t) \) as the unique function which satisfies \( r = k - k \ast r \). Then we claim that

\[
f(t) = (V_0 - \theta) - r \ast (V_0 - \theta).
\]

To verify the claim, we substitute this expression into (4) to get:

\[
(V_0 - \theta) - k \ast [(V_0 - \theta) - r \ast (V_0 - \theta)] = (V_0 - \theta) - (V_0 - \theta) \ast (k - k \ast r)(t)
\]

\[
= (V_0 - \theta) - (V_0 - \theta) \ast r(t)
\]

so \((V_0 - \theta) - k \ast f(t) = (V_0 - \theta) - (V_0 - \theta) \ast r(t) = f(t)\), which is precisely the integral equation we are trying to solve. Taking Laplace transform of both sides of \( k - k \ast r = r \) we obtain \( \hat{r} = k - k \hat{r} \), which we can re-arrange as

\[
\hat{r} = \frac{k}{1+k} = \frac{\lambda z^{-\alpha}}{1+\lambda z^{-\alpha}} = \frac{\lambda}{z^\alpha + \lambda}
\]

and the inverse Laplace transform of \( \hat{r} \) is \( r(t) = \lambda t^{\alpha-1} E_{\alpha,\alpha}(-\lambda t^\alpha) \). ■

### 2.2 Computing \( \mathbb{E}(V_u | F_t) \)

Now let \( \xi_{u}(u) := \mathbb{E}(V_u | F_t) \). Then \( \xi_{u}(u) \) is an \( F_t \)-martingale, and

\[
\xi_{u}(u) = V_0 + \frac{1}{\Gamma(\alpha)} \int_0^u (u-s)^{\alpha-1} \lambda (\theta - \mathbb{E}(V_s | F_t)) ds + \frac{1}{\Gamma(\alpha)} \int_0^t (u-s)^{\alpha-1} \nu \sqrt{V_s} dW_s.
\]

If \( \lambda = 0 \), we can re-write this expression as

\[
d\xi_{u}(u) = \frac{1}{\Gamma(\alpha)} (u-t)^{\alpha-1} \sqrt{V_t} dW_t.
\]

**Proposition 2.2** (see \([ER19]\)). For \( \lambda > 0 \)

\[
d\xi_{u}(u) = \kappa(u-t) \sqrt{V_t} dW_t = \kappa(u-t) \sqrt{\xi_{t}(t)} dW_t
\]

(5)

where \( \kappa \) is the inverse Laplace transform of \( \hat{\kappa}(z) = \frac{\nu z^{-\alpha}}{1+\nu z^{-\alpha}} \), which is given explicitly by

\[
\kappa(x) = \nu x^{\alpha-1} E_{\alpha,\alpha}(-\lambda x^\alpha) \sim \frac{1}{\Gamma(\alpha)} \nu x^{\alpha-1}
\]

as \( x \to 0 \) (see also page 6 in \([GK18]\) and page 29 in \([ER18]\)).

**Proof.** See Appendix A. ■

**Remark 2.1** From \([3]\), we see that \( \xi_{t}(.) \) is Markov in \( \xi_{t}(.) \). However \( V \) is not Markov in itself.

### 2.3 Evolving the variance curve

We simulate the variance curve at time \( t > 0 \) using

\[
\xi_{t}(u) = \xi_{0}(u) + \int_0^t \kappa(u-s) \sqrt{V_s} dW_s
\]

and substituting the expression for \( \xi_{0}(t) = \mathbb{E}(V_t) \) in \([2]\) and the expression for \( \kappa(t) \) in Proposition 2.2 (which are both expressed in terms of the Mittag-Leffler function).
2.4 The characteristic function of the log stock price

From Corollary 3.1 in [ER19] (see also Section 5 in [GGP18]), we know that for all $t \geq 0$
\[ \mathbb{E}(e^{i\alpha X_t}) = \mathbb{E}(e^{i\alpha X_t}) \]
\[ = e^\mu t + \alpha \mathbb{E}(X_0) + \int_0^t \mathbb{E}(X_{s-}) \nu(s) ds + \mathbb{E}(X_0) \mathbb{E}(X_{t-}) + \mathbb{E}(X_0) \frac{\nu(\alpha \xi)}{\nu(\alpha \mu)}(\xi)\]
\[ = e^{\mu t} + \alpha \mathbb{E}(X_0) + \int_0^t \mathbb{E}(X_{s-}) \nu(s) ds + \mathbb{E}(X_0) \mathbb{E}(X_{t-}) + \mathbb{E}(X_0) \frac{\nu(\alpha \xi)}{\nu(\alpha \mu)}(\xi)\]
\[ = e^{\mu t} + \alpha \mathbb{E}(X_0) + \int_0^t \mathbb{E}(X_{s-}) \nu(s) ds + \mathbb{E}(X_0) \mathbb{E}(X_{t-}) + \mathbb{E}(X_0) \frac{\nu(\alpha \xi)}{\nu(\alpha \mu)}(\xi)\]
for $p$ in some open interval $I \supset [0, 1]$, where $f(p, t)$ satisfies
\[ D^\alpha f(p, t) = \frac{1}{2} (p^2 - p) + (p \rho \nu - \lambda) f(p, t) + \frac{1}{2} \nu^2 f(p, t)^2 \] with initial condition $f(p, 0) = 0$, where $I^\alpha f$ denotes the fractional integral operator of order $\alpha$ (see e.g. page 16 in [ER19] for definition) and $D^\alpha$ denotes the fractional derivative operator of order $\alpha$ (see page 17 in [ER19] for definition).

2.5 The generalized time-dependent Rough Heston model and fitting the initial variance curve

If we now replace the constant $\theta$ with a time-dependent function $\theta(t)$, then
\[ \mathbb{E}(V_t) = V_0 + \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} \lambda(\theta(s) - \mathbb{E}(V_s)) ds, \]
which we can re-arrange as
\[ \mathbb{E}(V_t) - V_0 + \lambda I^\alpha \mathbb{E}(V_t) = \lambda I^\alpha \theta(t) \] so to make this generalized model consistent with a given initial variance curve $\mathbb{E}(V_t)$, we set
\[ \theta(t) = \frac{1}{\lambda} D^\alpha (\mathbb{E}(V_t) - V_0 + \lambda I^\alpha \mathbb{E}(V_t)) = \frac{1}{\lambda} D^\alpha (\mathbb{E}(V_t) - V_0) + \mathbb{E}(V_t) \]
(see also Remark 3.2, Theorem 3.2 and Corollary 3.2 in [ER18]).

2.6 Other affine and non-affine variance curve models

We can also consider other models which are not Rough Heston but for which [3] is still satisfied, and models of this form are known as affine forward variance (AFV) models (see [GK19] for an excellent treatise on such models and how to obtain the Rough Heston model as the limit of a market microstructure model driven by a generalized Hawkes process in the small-jump, high jump intensity limit). We can of course integrate (5) and set $u = t$ to get
\[ V_t = \xi_0(t) + \int_0^t \kappa(t-s) \sqrt{V_{s-}} dW_s \]
which generalizes the Rough Heston model. Another well known (and non-affine) variance curve model is the Rough Bergomi model, for which $d\xi_t(u) = \eta(u-t) dW_t$ or the standard Bergomi model $d\xi_t(u) = \eta e^{-\lambda(u-t)} \xi_t(u) dW_t$.

2.7 Microstructural foundations of the Rough Heston model

The canonical $n$-dimensional Hawkes process is a generalized Poisson process $(N_t)_{t \geq 0}$ with stochastic intensity given by $\lambda_t = \mu_t + \int_0^t \phi(t-s) dN_s$. Such processes are useful for modelling contagion in finance, and $N_t$ can also be interpreted as a branching process where immigrants arrive at a rate $\mu_t$, immigrants give birth to children at a rate $\phi(t)$, and children give birth to further children at a rate $\phi(t)$. With this interpretation, the average number of descendants of a particular immigrant is the $L^1$ norm of the kernel $||\phi||$, which is also the average proportion of the population who are children as opposed to immigrants. If this norm is $\geq 1$ the population explodes and hence $||\phi|| < 1$ is known as the stability criterion.

Using a standard conditioning argument, one can show that the characteristic function (C.F.) of $N_t$ is:
\[ \mathbb{E}(e^{i\alpha N_t}) = e^{\int_0^t (C(a,t)-s) \mu(s) ds} \]
where $C(a, t)$ satisfies the non-linear integral equation $C(a, t) = e^{i\alpha t + \int_0^t \phi(s)C(a, t-s) ds}$. Moreover, with a certain choice of parameters these processes can generate price processes which display observed and conjectured stylized features of financial time series such as market endogeneity and buy/sell asymmetry, and one can control the proportion of orders that are so called “metaorders”. The intensity is chosen to satisfy:
\[ \lambda_T = \left( \begin{array}{c} \lambda_{T-}^+ \\ \lambda_{T-}^- \end{array} \right) = \tilde{\mu}_T \left( \begin{array}{c} 1 \\ 1 \end{array} \right) + \int_0^t \phi_T(t-s) dN_s^T \]
where \( \beta > 0, \alpha \in (\frac{1}{2}, 1), \lambda > 0, \mu > 0, \xi > 0, a_T = 1 - \lambda T^{-\alpha}, \phi^T = \varphi^T \chi, \chi = \frac{1}{\beta + 1} \left( \begin{array}{l} \beta \\ \beta \\ \end{array} \right), \varphi = f^{\alpha, 1}, \bar{\mu}_T(t) = \mu t^{-\alpha - 1} + \xi t^{-\alpha - 1}(1 - \int_0^t \varphi^T(s)ds) - \int_0^t \varphi^T(s)ds), \) where \( f^{\alpha, 1} \) is the Mittag-Leffler density function defined in the appendix of [ER19]. Returning to the branching interpretation, the exogenous orders are the immigrants and the endogenous orders are the children. The fact that \( 1 > a_T \to 1 \) is what gives the interpretation of a highly endogenous market and we say our Hawkes process is “nearly unstable” (see also section 5.4 in [FS18] for details on this point). From this one can define the following rescaled processes for \( t \in [0, 1] \):

\[
X_t^T = \frac{1 - a_T}{T^{\alpha}} N^T_t, \quad A_t^T = \frac{1 - a_T}{T^{\alpha}} \int_0^t \lambda_s^T ds, \quad Z_t^T = \sqrt{\frac{T^\alpha}{1 - a_T}} (X_t^T - A_t^T)
\]

Building on [JR16], [ER19] show that the processes \((A_t^T, X_t^T, Z_t^T)_{t \in [0, 1]}\) converges in law under the Skorokhod topology to

\[
A_t = X_t = \int_0^t Y_s ds \left( \begin{array}{l} 1 \\ 1 \\ \end{array} \right), \quad Z_t = \int_0^t \sqrt{Y_s} \left( \begin{array}{l} dB_1^x \\ dB_2^x \end{array} \right)
\]

and \( Y \) is the unique solution of the rough stochastic differential equation

\[
Y_t = \xi + \frac{1}{\Gamma(\alpha)} \int_0^t (t - s)^{-\alpha - 1}(1 - Y_s)ds + \lambda \sqrt{\frac{1 + \beta^2}{4\mu(1 + \beta^2)}} \frac{1}{\Gamma(\alpha)} \int_0^t (t - s)^{-\alpha - 1} \sqrt{Y_s} dB_s,
\]

where \( B = \frac{B_1^x + B_2^x}{\sqrt{1 + \beta^2}} \) and \((B_1^x, B_2^x)\) is 2-dimensional Brownian motion. As a corollary, for \( \theta > 0 \), if \( V = \theta Y \) and

\[
P_t^T = \sqrt{\frac{\theta}{2}} \frac{1 - a_T}{T^{\alpha}} (N_{T^{\alpha}}^{T, +} - N_{T^{\alpha}}^{T, -}) - \frac{\theta}{2} \frac{1 - a_T}{T^{\alpha}} N_{T^{\alpha}}^{T, +}
\]

then we have similar convergence in distribution of \( P_t^T \) to \( P_t = \int_0^t \sqrt{V_s} dW_s - \frac{1}{2} \int_0^t V_s ds \), so the Rough Heston model is recovered. The time scale \( T \) is of the order of the reciprocal of the price tick size, hence as \( T \to \infty \), the price moves more frequently with smaller size. The equality \( \mu_T = \mu > 0 \) in Assumption 2.2 in [ER19] is quite intuitive in this scaling, meaning that exogenous moves of size 1/T occur \( \mu_T \) times on average in a unit time interval, maintaining a non-zero contribution from exogenous activity in the limit.

## 3 Small-time asymptotics

### 3.1 The small-time LDP

To simplify calculations, we make the following assumption throughout this section:

**Assumption 3.1** \( \lambda = 0 \).

**Remark 3.1** The formal higher order Laplace asymptotics in subsection 3.3 indicate that \( \lambda \) will not affect the leading order small-time asymptotics, i.e. \( \lambda \) will not affect the rate function, as we would expect from previous works on small-time asymptotics for rough stochastic volatility models. The assumption that \( \lambda = 0 \) is relaxed in the next section where we consider large-time asymptotics.

We now state the main small-time result in the article (recall that \( \alpha = H + \frac{1}{2} \)):

**Theorem 3.2** For the Rough Heston model defined in [1], we have

\[
\lim_{t \to 0} t^{2H} \log \mathbb{E}(e^{t \frac{X_t}{t^{1-H}}}) = \lim_{t \to 0} t^{2H} \log \mathbb{E}(e^{t \frac{X_t}{t^{1-H}}}) = \begin{cases} \tilde{\Lambda}(p) & \text{if } T^*(p) > 1 \\ +\infty & \text{if } T^*(p) \leq 1 \end{cases}
\]

where \( \tilde{\Lambda}(p) := V_0 \Lambda(p), \Lambda(p) := \Lambda(p, 1), \Lambda(p, t) := I^{1-\alpha} \psi(p, t) \) and \( \psi(p, t) \) satisfies the Volterra differential equation

\[
D^\alpha \psi(p, t) = \frac{1}{2} p^2 + pp^\alpha \psi(p, t) + \frac{1}{2} \nu^2 \psi(p, t)^2
\]

with initial condition \( \psi(p, 0) = 0 \), where \( T^*(p) > 0 \) is the explosion time for \( \psi(p, t) \) which is finite for all \( p \neq 0 \) (assuming \( \nu > 0 \)). Moreover, the scaling relation in Lemma 3.3 and its Corollary 3.4 inside the main proof below shows that \( \Lambda(p) = |p|^{\frac{2H}{\alpha}} \Lambda(\text{sgn}(p), |p|^{\frac{1}{\alpha}}) \), so in fact we only need to solve (10) for \( p = \pm 1 \), and we can re-write (10) in more familiar form as

\[
\lim_{t \to 0} t^{2H} \log \mathbb{E}(e^{t \frac{X_t}{t^{1-H}}}) = \lim_{t \to 0} t^{2H} \log \mathbb{E}(e^{t \frac{X_t}{t^{1-H}}}) = \begin{cases} \tilde{\Lambda}(p) & p \in (p_-, p_+) \\ +\infty & p \notin (p_-, p_+) \end{cases}
\]

where \( p_\pm = \pm (T^*(\pm 1))^{\alpha} \), so \( p_+ > 0 \) and \( p_- < 0 \). Then \( X_t / t^{\frac{1}{2} - H} \) satisfies the LDP as \( t \to 0 \) with speed \( t^{-2H} \) and good rate function \( I(x) \) equal to the Fenchel-Legendre transform of \( \Lambda \).
Proof. We first consider the following family of re-scaled Rough Heston models:

\[ dX_t^\varepsilon = -\frac{1}{2}\varepsilon V_t^\varepsilon dt + \sqrt{\varepsilon V_t^\varepsilon} dB_t, \quad V_t^\varepsilon = V_0 + \frac{\varepsilon^\alpha}{\Gamma(\alpha)} \int_0^t (t-s)^{H-\frac{1}{2}} \lambda (\theta - \gamma V_s^\varepsilon) ds + \frac{\varepsilon^H}{\Gamma(\alpha)} \int_0^t (t-s)^{H-\frac{3}{2}} \nu \sqrt{V_s^\varepsilon} dW_s \quad (12) \]

with \( X_0^\varepsilon = 0 \), where \( H = \alpha - \frac{1}{2} \in (0, \frac{1}{2}] \). Then from Appendix [A] we know that

\[ (X_t^\varepsilon, V_t^\varepsilon) \overset{(d)}{=} (X_{\varepsilon t}(\cdot), V_{\varepsilon t}(\cdot)) \quad (13) \]

(note this actually holds for all \( \lambda > 0 \), but from here on we set \( \lambda = 0 \)). Proceeding along similar lines to Theorem 4.1 in [EZ14], we let \( \tilde{X}_t^\varepsilon \) denote the solution to

\[ d\tilde{X}_t^\varepsilon = \sqrt{\varepsilon} \sqrt{V_t^\varepsilon} dB_t \quad (14) \]

with \( \tilde{X}_0^\varepsilon = 0 \). From Eq 8 in [ERS], we know that

\[ \mathbb{E}(e^{\varepsilon \tilde{X}_t^\varepsilon}) = \mathbb{E}_{\varepsilon}^\circ (e^{\frac{\varepsilon}{2} \int_0^t V_s^\varepsilon ds}) \]

where \( \mathbb{Q}_p \) is defined as in [ERS], but under \( \mathbb{Q}_p \) the value of the mean reversion speed changes from zero to \( \bar{\lambda} = \rho \nu v \), so

\[ \mathbb{E}(e^{\varepsilon \tilde{X}_t^\varepsilon}) = e^{\bar{\lambda} t^\alpha - \gamma (p,t)} \]

on some non-empty interval \([0, T^\ast(p)]\), where

\[ D^\alpha g(p,t) = \frac{1}{2} p^2 + \rho \nu v g(p,t) + \frac{1}{2} \nu^2 g(p,t)^2 \]

with \( g(p,0) = 0 \). Existence and uniqueness of solutions to these kind of fractional differential equation (FDE) is standard (as is their equivalence to VIEs), see [GGP18] for details and references. Now define \( g_\varepsilon(p,t) := e^{1-\alpha} g(p,\varepsilon t) \) and setting \( s = \varepsilon u \), we see that

\[ I^{1-\alpha} g_\varepsilon(p,t) = \frac{1}{\Gamma(\alpha)} \int_0^t (t-u)^{\alpha-1} g(p,\varepsilon u) du = \frac{1}{\Gamma(1-\alpha)} \int_0^t (t-s/\varepsilon)^{-\alpha} \varepsilon^{-\alpha} g(p,s) ds \]

\[ = \varepsilon^{\alpha} \frac{1}{\Gamma(1-\alpha)} \int_0^t (\varepsilon t - s) -\alpha \varepsilon^{-\alpha} g(p,s) ds \]

\[ = (I^{1-\alpha} g(p,\cdot))(\varepsilon t) \quad (15) \]

and

\[ \varepsilon^{\alpha} I^1 g_\varepsilon(p,t) = \varepsilon^\alpha \int_0^t \varepsilon^{1-\alpha} g(p,\varepsilon u) du = \varepsilon^\alpha \int_0^{\varepsilon t} \varepsilon^{-\alpha} g(p,s) ds = (I^1 g(p,\cdot))(\varepsilon t) \quad (16) \]

Thus when \( \lambda = 0 \), replacing \( g(p,t) \) with \( g_\varepsilon(p,t) \) is tantamount to changing the maturity \( t \) to \( \varepsilon t \) (as opposed to \( t \)). Combining this observation with the results of Section 5 of [GGP18], we see that

\[ \mathbb{E}(e^{\varepsilon \tilde{X}_t^\varepsilon}) = \mathbb{E}(e^{\varepsilon \tilde{X}_{\varepsilon t}^\varepsilon}) = e^{\bar{\lambda} t^\alpha - \gamma (p,t)} \quad (17) \]

on some non-empty interval \([0, T^\ast(p)]\). Moreover

\[ g_\varepsilon(p,t) = e^{1-\alpha} \frac{1}{\Gamma(\alpha)} \int_0^{\varepsilon t} (\varepsilon t - s)^{\alpha-1} \left( \frac{1}{2} p^2 + \rho \nu \varepsilon^{-\alpha} g(p,s) + \frac{1}{2} \nu^2 g(p,s)^2 \right) ds \]

\[ = e^{1-\alpha} \frac{1}{\Gamma(\alpha)} \int_0^{\varepsilon t} (\varepsilon t - u)^{\alpha-1} \left( \frac{1}{2} p^2 + \rho \nu \varepsilon^{\alpha-1} g(p,u) + \frac{1}{2} \nu^2 g(p,u)^2 \right) du \]

\[ = \frac{1}{\Gamma(\alpha)} \int_0^t (t-u)^{\alpha-1} \left( \frac{1}{2} p^2 + \rho \nu \varepsilon^{\alpha-1} g(p,u) + \frac{1}{2} \nu^2 g(p,u)^2 \right) du \quad (18) \]

so we see that \( g_\varepsilon(p,t) \) satisfies

\[ D^\alpha g_\varepsilon(p,t) = \frac{1}{2} \varepsilon p^2 + \varepsilon^\alpha \rho \nu v g_\varepsilon(p,t) + \frac{1}{2} \varepsilon^{2H} \nu^2 g_\varepsilon(p,t)^2 \quad (19) \]

with initial condition \( g_\varepsilon(p,0) = 0 \). Now set

\[ g_\varepsilon \left( \frac{p}{\varepsilon^\alpha}, t \right) = \frac{\psi(p,t)}{\varepsilon^{2H}} \quad (20) \]

\[ ^1 \text{In fact this relationship clearly holds for any function } g \]
Then setting $p \mapsto \frac{p}{\varepsilon}$, and substituting for $g_e(\frac{x}{\varepsilon}, t)$ in (19) and multiplying by $\varepsilon^{2H}$, we find that

$$D^\alpha \psi(p, t) = \frac{1}{2} p^2 + pp\nu \psi(p, t) + \frac{1}{2} \nu^2 \psi(p, t)^2$$

(21)

with $\psi(p, 0) = 0$. Moreover, from Propositions 3.2 and 3.4 in [GGP18], we know that $\psi(p, t)$ blows up at some finite time $T^*(p) > 0$, so $\Lambda(0) = \frac{1}{2} p^2 + pp\nu \psi(p, t) + \frac{1}{2} \nu^2 \psi(p, t)^2$ has no real roots (i.e. case A or B in the [GGP18] classification). Moreover, $\psi(p, t)$ is independent of $\varepsilon$ so $T^*_\varepsilon(\frac{x}{\varepsilon})$ (where $T^*_\varepsilon(p)$ is defined above) is equal to $T^*(p)$. Thus we see that

$$E(e^{\frac{xt}{\varepsilon}} X^*_t) = e^\Lambda(p, t)$$

for all $t \in [0, T^*(p))$, which we can re-write as $E(e^{\frac{xt}{\varepsilon}} X^*_t) = e^{\Lambda(p, t)}$. Thus we see that

$$\lim_{\varepsilon \to 0} \varepsilon^{2H} \log E(e^{\frac{xt}{\varepsilon}} X^*_t) = V_0 I^{1-\alpha} \psi(p, t) = \tilde{\Lambda}(p, t)$$

and $\Lambda(p) := \Lambda(p, 1) < \infty$ if and only if $T^*(p) > 1$. Thus $E(e^{\frac{tx}{\varepsilon}} X^*_t) = \infty$ if $T^*(p) \leq 1$, and otherwise (by (23)) we see that

$$\lim_{t \to 0} \varepsilon^{2H} \log E(e^{\frac{tx}{\varepsilon}} X^*_t) = \lim_{\varepsilon \to 0} \varepsilon^{2H} \log E(e^{\frac{tx}{\varepsilon}} X^*_t) = V_0 I^{1-\alpha} \psi(p, 1) = \tilde{\Lambda}(p).$$

Lemma 3.3 We have the scaling relation for $t \in [0, T^*(p)]$:

$$\Lambda(p, t) = t^{-2H} \Lambda(p t^\alpha, 1) = t^{-2H} \Lambda(p t^\alpha).$$

(23)

Proof. See Appendix C (as a sanity check we note that [23] is satisfied by the function $\Lambda(p, t) = \frac{1}{2} p^2 t$, i.e. the solution when $\nu = 0$.)

Corollary 3.4

$$\Lambda(q) = t^{2H} \Lambda(|q| t^{\frac{1}{\alpha}}, t) = (t_q^*)^{2H} \Lambda(1, t_q^*) = |q|^{2H} \Lambda(|\text{sgn}(q) |q|^{-\frac{1}{\alpha}})$$

(24)

where we have set $p = 1 = |\frac{q}{\psi}|$. and $t_q^* = |\frac{q}{\psi}|$.

Remark 3.2 This implies that $\Lambda(p) \to \infty$ as $p \to p_\pm := \pm (T^*(\pm 1))^\alpha$, and more generally

$$p T^*(p)^\alpha = 1_{p > 0} p_+ + 1_{p < 0} p_-.$$  

(25)

To prove the LDP, we first prove the corresponding LDP for $\tilde{X}_t$. From Lemma 2.3.9 in [DZ98], we know that $\lim_{t \to 0} t^{2H} \log E(e^{\frac{tx}{\varepsilon}} \tilde{X}_t) = \Lambda(p) = \Lambda(p, 1) = I^{1-\alpha} \psi(p, t)|_{t=1}$ is convex in $p$, and from (21) we know that

$$\frac{d}{dt} \Lambda(p, t) = \frac{1}{2} p^2 + pp\nu \psi(p, t) + \frac{1}{2} \nu^2 \psi(p, t)^2$$

which shows that $\Lambda(p, t)$ is also differentiable in $t$, and thus from the scaling property in [23], $\Lambda(p) = \Lambda(p, 1)$ is differentiable in $p$. We also know that $\psi(p, t) \to \infty$ as $t \to T^*(p)$ (see Propositions 3.2 and 3.4 in [GGP18]), so $\Lambda(p, t) = I^{1-\alpha} \psi(p, t)$ also explodes at $T^*(p)$ by Lemma 3.8 in [GGP18]. Then from Corollary 3.4 we know that $\Lambda(p) = p^{2H} \Lambda(\text{sgn}(p), |p|^{-\frac{1}{\alpha}})$, so $\Lambda(p) \to \infty$ as $p \to p_\pm = \pm (T^*(\pm 1))^\alpha$, and (by convexity and differentiability) $\Lambda$ is also essentially smooth, so by the Gärtner-Ellis theorem from large deviations theory (see Theorem 2.3.6 in [DZ98]), $\tilde{X}_t/\varepsilon^{1-H}$ satisfies the LDP as $\varepsilon \to 0$ with speed $\varepsilon^{-2H}$ and rate function $I(x)$.

Moreover, using that

$$E(e^{\frac{xt}{\varepsilon}} f_0^{H \varepsilon} V_s^* ds) = E(e^{\frac{xt}{\varepsilon}} f_0^{H \varepsilon} V_s^* ds) = e^{V_0 I^{1-\alpha} \phi(p, t)}$$

for $p \in (-\infty, \frac{1}{2} p^2)$ (and infinity otherwise), where $\tilde{p}$ is the value of $p_+$ for $p = 0$ and $D^\alpha \phi(p, t) = p + \frac{1}{2} \nu^2 \psi(p, t)^2$ with $\psi(p, t) = 0$ (see also [22] and Theorem 3.2 in [FR18]), we find that

$$J(p) = \lim_{\varepsilon \to 0} \varepsilon^{2H} \log E(e^{\frac{tx}{\varepsilon}} f_0^{H \varepsilon} V_s^* ds) = \lim_{\varepsilon \to 0} \varepsilon^{2H} \log E(e^{\frac{tx}{\varepsilon}} f_0^{H \varepsilon} V_s^* ds)$$

so (again using part a) of the Gärtner-Ellis theorem in Theorem 2.3.6 in [DZ98], $A_\varepsilon := \int_0^1 V_s^* ds$ satisfies the upper bound LDP as $\varepsilon \to 0$ with speed $\varepsilon^{-2H}$ and good rate function $J^*$ equal to the FL transform of $J$. But we also know that

$$X_1 - X_1^* = -\frac{1}{2} \varepsilon A_\varepsilon$$

and for any $a > 0$ and $\delta > 0$,

$$\mathbb{P}(\frac{X_1}{\varepsilon^2 H} - \frac{X_1^*}{\varepsilon^{2 H}} > \delta) = \mathbb{P}(\frac{1}{2} \varepsilon^{H/2} A_\varepsilon > \delta) = \mathbb{P}(A_\varepsilon > \frac{2\delta}{\varepsilon^{H/2}}) \leq \varepsilon^{-H/2} \mathbb{P}(A_\varepsilon > a) \leq e^{-\frac{\inf_{a > 0} J^*(a) - \delta}{\varepsilon^{H/2}}}$$
for any $\varepsilon$ sufficiently small, where we have use the upper bound LDP for $A_{\varepsilon}$ to obtain the final inequality. Thus

$$
\lim_{\varepsilon \to 0} \varepsilon^{2H} \log \mathbb{P}(\frac{X_t}{\varepsilon^{1-H}} - \frac{\hat{X}_t}{\varepsilon^{1-H}} > \delta) \leq J(a)
$$

but $a$ is arbitrary and (from Lemma 2.3.9 in [DZ98]), $J$ is a good rate function, so in fact

$$
\lim_{\varepsilon \to 0} \varepsilon^{2H} \log \mathbb{P}(\frac{X_t}{\varepsilon^{1-H}} - \frac{\hat{X}_t}{\varepsilon^{1-H}} > \delta) = -\infty.
$$

Thus $\frac{X_t}{\varepsilon^{1-H}}$ and $\frac{\hat{X}_t}{\varepsilon^{1-H}}$ are exponentially equivalent in the sense of Definition 4.2.10 in [DZ98], so (by Theorem 4.2.13 in [DZ98]) $\frac{X_t}{\varepsilon^{1-H}}$ satisfies the same LDP as $\frac{\hat{X}_t}{\varepsilon^{1-H}}$. $\blacksquare$

### 3.2 Asymptotics for call options and implied volatility

**Corollary 3.5** We have the following limiting behaviour for the price of a European call option with maturity $t$ and log-strike $t^{1-H}x$, with $x > 0$ fixed:

$$
\lim_{t \to 0} t^{2H} \log \mathbb{E}((e^{X_t} - e^{t^{1-H}})^+) = -I(x).
$$

**Proof.** The lower estimate follows from the exact same argument used in Appendix C in [PZ17] (see also Theorem 6.3 in [FGP18]). The proof of the upper estimate is the same as in Theorem 6.3 in [FGP18]. $\blacksquare$

**Corollary 3.6** For $x \neq 0$ fixed, the implied volatility satisfies

$$
\hat{\sigma}(x) := \lim_{t \to 0} \hat{\sigma}_t(t^{1-H}x) = \frac{|x|}{\sqrt{2I(x)}}.
$$

**Proof.** Follows from Corollary 7.2 in [GL14]. See also the proof of Corollary 4.1 in [FGP18] for details on this, but the present situation is simpler, as we only require the leading order term here. $\blacksquare$

### 3.3 Series expansion for the asymptotic smile and calibration

Proceeding as in Lemma 7.2 in [GGP18], we can compute a fractional power series for $\psi(p, t)$ (and hence $\Lambda(p, t)$) and then using (24), we find that

$$
\hat{\Lambda}(p) = 2V_0 \frac{\rho \nu}{\nu^2} \sum_{n=1}^{\infty} a_n(1)p^{1+n} \frac{\Gamma(\alpha n + 1)}{\Gamma(2 + (n - 1)\alpha)}
$$

where the $a_n = a_n(u)$ coefficients are defined (recursively) as in [GGP18] except for our application here (based on (21)) we have to set $\lambda = 0$, and $c_1 = \frac{1}{2}u^2$ instead of $\frac{1}{2}u(u - 1)$ (note this series will have a finite radius of convergence). Using the Lagrange inversion theorem, we can then derive a power series for $I(x)$ which takes the form

$$
\hat{\sigma}(x) = \sqrt{V_0} + \frac{\rho \nu}{2\Gamma(2 + \alpha)\sqrt{V_0}} x^{\alpha} + \frac{\Gamma(1 + 2\alpha) + 2\rho^2 \Gamma(1 + \alpha)^2(2 - 3\frac{\Gamma(2+2\alpha)}{\Gamma(2+\alpha)})}{8V_0^2 \Gamma(1 + \alpha)^2 \Gamma(2 + \alpha)} x^2 + O(x^3).
$$

(compare this to Theorem 3.6 in [BGH18] for a general class of rough models and Theorem 4.1 in [FJ11] for a Markovian local-stochastic volatility model). We can re-write this expansion more concisely in dimensionless form as

$$
\hat{\sigma}(x) = \sqrt{V_0} [1 + \frac{\rho}{2\Gamma(2 + \alpha)} x^\alpha + \frac{\Gamma(1 + 2\alpha) + 2\rho^2 \Gamma(1 + \alpha)^2(2 - 3\frac{\Gamma(2+2\alpha)}{\Gamma(2+\alpha)})}{8\Gamma(1 + \alpha)^2 \Gamma(2 + \alpha)} x^2 + O(x^3)]
$$

where the dimensionless quantity $z = \frac{ux}{V_0}$.

**Remark 3.3** In principle one can use (27) to calibrate $V_0$, $\rho$ and $\nu$ to observed/estimated values of $\hat{\sigma}(0)$, $\hat{\sigma}’(0)$ and $\hat{\sigma}''(0)$ (i.e. the short-end implied vol level, skew and convexity respectively).

#### 3.3.1 Wing behaviour of the rate function

From Eq 3.2 in [RO96], we expect that $\psi(p, t) \approx \frac{\text{const.}}{(T^*(p) - t)^{2\alpha - 1}}$ as $t \to T^*(p)$ and thus $\Lambda(p, t) = T^*(p) - t)^{2\alpha - 1}$ as $t \to T^*(p)$. Assuming this is consistent with the $p$-asymptotics, then (by (23)) we have

$$
\Lambda(p) = \Lambda(p, 1) \approx \frac{\text{const.}}{(T^*(p) - 1)^{2\alpha - 1}} = \frac{\text{const.}}{(T^*(p) - 1)^{2\alpha - 1}} \sim \frac{\text{const.}}{(p_+ - p)^{2\alpha - 1}} (p \to p_+)
$$

so $p^*(x)$ in $I(x) = \sup_{p}(px - V_0 \Lambda(p))$ satisfies $p^*(x) \approx p_+ - \text{const.} \cdot x^{1/2\alpha}$, so $I(x) \approx \text{const.} \cdot x^{1/2\alpha}$ as $x \to \infty$. 
3.4 Higher order Laplace asymptotics

If we now relax the assumption that \( \lambda = 0 \), and work with the original \( X^\varepsilon \) process in \([12]\) (as opposed to the driftless \( \tilde{X}^\varepsilon \) process in \([14]\)), then we know that

\[
\mathbb{E}(e^{pX^\varepsilon_t}) = \mathbb{E}(e^{pX^\varepsilon_{t+}}) = e^{\tilde{V}_0 t^{1-\alpha} g_\varepsilon(p,t) + \varepsilon^\alpha \lambda \theta t^2 g_\varepsilon(p,t)}
\]

for \( t \) in some non-empty interval \([0, T^*(p)]\), where now \( g_\varepsilon(p,t) \) satisfies

\[
D^\alpha g_\varepsilon(p,t) = \frac{1}{2} \varepsilon^2 (p^2 - p) + (pp\nu - \lambda)\varepsilon^\alpha g_\varepsilon(p,t) + \frac{1}{2} \varepsilon^2 \nu^2 g_\varepsilon(p,t)^2
\]

with initial condition \( g_\varepsilon(p,0) = 0 \). Setting

\[
g_\varepsilon(p, t) = \frac{\psi_\varepsilon(p,t)}{\varepsilon^{2H}}
\]

and setting \( p \mapsto \frac{p}{\varepsilon^{\alpha}} \), and substituting for \( g_\varepsilon\left(\frac{p}{\varepsilon^{\alpha}}, t\right) \) in \([25]\) and multiplying by \( \varepsilon^{2H} \) as before, we find that

\[
D^\alpha \psi_\varepsilon(p,t) = \frac{1}{2} \nu^2 + pp\nu \psi_\varepsilon(p,t) + \frac{1}{2} \nu^2 \psi_\varepsilon(p,t)^2 - \varepsilon^\alpha (\frac{1}{2} p + \lambda \psi_\varepsilon(p,t))
\]

with \( \psi(p,0) = 0 \). If we now formally try a higher order series approximation of the form \( \psi_\varepsilon(p,t) = \psi(p,t) + \varepsilon^{\frac{\alpha}{2} + H} \psi_1(p,t) \), we find that \( \psi_1(p,t) \) must satisfy

\[
D^\alpha \psi_1(p,t) = -\frac{1}{2} p - \lambda \psi(p,t) + pp\nu \psi_\varepsilon(p,t) + \nu^2 \psi(p,t) \psi_\varepsilon(p,t)
\]

with \( \psi_1(p,0) = 0 \), which is a linear FDE for \( \psi_1(p,t) \).

Remark 3.4 Setting \( \psi_1(p,t) = \sum_{n=1}^{\infty} b_n(p) t^{\alpha n} \) we see that

\[
\sum_{n=1}^{\infty} \frac{n a \Gamma(na)}{\Gamma(1 + (n - 1) \alpha)} b_n(p) t^{(n-1)\alpha} = -\frac{1}{2} p - \lambda \sum_{n=1}^{\infty} \bar{a}_n(p) t^{\alpha n} + pp\nu \sum_{n=1}^{\infty} b_n(p) t^{\alpha n} + \nu^2 \sum_{n=1}^{\infty} \bar{a}_n(p) t^{\alpha n} \sum_{m=1}^{\infty} b_m(p) t^{\alpha m}
\]

where \( \bar{a}_n(p) = \frac{\alpha}{\nu} a_n(p) \), and we have set \( \lambda = 0 \) and \( c_1 = \frac{1}{2} p^2 \) in computing the \( a_n(p) \) coefficients, so

\[
a \Gamma(\alpha) b_1(p) = -\frac{1}{2} p \quad , \quad \frac{na \Gamma((n + 1) \alpha)}{\Gamma(1 + na)} b_{n+1}(p) = -\left(\lambda a_n(p) + pp\nu b_n(p) + \nu^2 \sum_{k=1}^{n-1} a_k(p) b_{n-k}(p) \right)
\]

so we have fractional power series for \( \psi_1(p,t) \) on some finite radius of convergence.

Returning now to the main calculation, we see that if \( p_\varepsilon(x) \) denotes the density of \( \frac{X^\varepsilon_x}{\varepsilon^{2H}} \), then

\[
p_\varepsilon\left(\frac{x}{\varepsilon^{2H}}\right) \sim \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{ikx}{\varepsilon^{2H}}} e^{\frac{\varepsilon^\alpha}{\varepsilon^{2H}} (F(k) + \frac{\alpha \nu}{2H} G(k) + \frac{\alpha \nu}{2H} \theta F_1(k) + \frac{1}{2} \lambda \psi_\varepsilon(p,t))} dk
\]

where \( F(k) := \tilde{V}_0 I^{1-\alpha} \psi_\varepsilon(i k, 1), G(k) := \tilde{V}_0 I^{1-\alpha} \psi_\varepsilon(i k, 1), F_1 := I^1 \psi_\varepsilon(i k, 1) \) and \( G_1 := I^1 \psi_\varepsilon(i k, 1) \). The saddlepoint \( k^*(x) = ip^*(x) \) of \( \tilde{F}(k) = -ikx + F(k) \) satisfies \( \tilde{F}'(k^*) = 0 \) which always falls on the imaginary axis (and in our case \( p^*(x) \) in \((0, p_+)\) when \( x > 0 \) and \( p^*(x) < 0 < (p_-, 0) \) when \( x < 0 \), and

\[
\tilde{F}(k) \approx \tilde{F}(k^*) + \frac{1}{2} \tilde{F}''(k^*)(k - k^*)^2 = \tilde{F}(k^*) - \frac{1}{2} \frac{\tilde{F}''(k^*)}{(k^*)^2} (k - k^*)^2
\]

(recall that \( \tilde{F}(p) = F(-ip) \) and \( p^* = ik^* \in (p_-, p_+) \). Then proceeding along similar lines to \([11,12]\) and using Laplace’s method we have

\[
p_\varepsilon\left(\frac{x}{\varepsilon^{2H}}\right) \sim \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{\varepsilon^\alpha}{\varepsilon^{2H}} (F(k) + \frac{\alpha \nu}{2H} G(k) + \frac{\alpha \nu}{2H} \theta F_1(k) + \frac{1}{2} \lambda \psi_\varepsilon(p,t))} dk
\]

\[
\sim \frac{1}{2\pi} e^{-\frac{1}{2H} G(k^*) + \frac{1}{2H} \lambda \theta F_1(k^*)} e^{\frac{1}{2H} \lambda \theta F_1(k^*)} \int_{-\infty}^{\infty} e^{-\frac{1}{2H} \frac{\lambda \theta F_1(k^*)}{2H} (k - k^*)^2} dk
\]

\[
\sim \frac{e^{\frac{1}{2H} \frac{\lambda \theta F_1(k^*)}{2H}}}{\sqrt{2\pi \lambda \theta F_1(k^*)}} \left[1 + e^{-\frac{1}{2H} (G(k^*) + \lambda \theta F_1(k^*)) + O(\varepsilon^{(1-2H)\wedge 2H})} \right]
\]
$G_1$ as well, 2nd term in denominator if $H \in (0, \frac{1}{2})$, where the $O(\varepsilon^{2H})$ part of the error terms comes from the next order term in Theorem 7.1 in chapter 4 in [Olv74], and the $\varepsilon^{(1-2H)}$ term comes from the 2nd order term in expanding the exponential. Then letting $z = \frac{k}{\varepsilon}$, we see that

$$C_\varepsilon(x) = E((e^{X^*} - \varepsilon^{\frac{x}{2-H}})^+) = \frac{1}{2\pi} e^{\frac{x}{2-H}} \int_{-\frac{ip}{\varepsilon} - \infty}^{\frac{ip}{\varepsilon} + \infty} \text{Re}(e^{iZ}X^*) dz$$

$$= \frac{1}{2\pi} e^{\frac{x}{2-H}} \int_{-\frac{ip}{\varepsilon} - \infty}^{\frac{ip}{\varepsilon} + \infty} \text{Re}(e^{i\frac{\varepsilon e^{\frac{x}{2-H}}}{\varepsilon^{1-2H}}}) E(e^{i\frac{Z}{\varepsilon^{1-2H}}}) dz$$

$$\sim \frac{\varepsilon^{-\alpha}}{2\pi} e^{\frac{x}{2-H}} \int_{-\frac{ip}{\varepsilon} - \infty}^{\frac{ip}{\varepsilon} + \infty} \text{Re}(e^{i\frac{X}{\varepsilon^{2-3H}}(-\frac{\varepsilon^{2\alpha}}{k^2} - \frac{2\varepsilon^3}{k^3} + O(\varepsilon^{4\alpha}) E(e^{i\frac{X}{\varepsilon^{2-3H}}}))} dk$$

$$\sim \varepsilon^{\frac{x}{2-H}} \frac{H e^{-\frac{1}{\varepsilon^{1-2H}}}}{\varepsilon^{2-3H}} [1 + \varepsilon^{\frac{x}{2-H}} (x + G(k^*) + \lambda \theta F_1(k^*)) + O(\varepsilon^{(1-2H)^2})].$$

The $\varepsilon$-dependence of the leading order term here is exactly the same as in Corollary 7.1 in the recent article of Friz et al. [FGP18a] (in [FGP18a] $\varepsilon^2 = t$ whereas here $\varepsilon = t$) which deals with a general class of rough stochastic volatility models (which excludes Rough Heston).

More generally, we can formally substitute a fractional power series of the form $\psi_n(p,t) = \sum_{n=0}^{\infty} \psi_n(p,t) \varepsilon^{(n+1)\alpha}$ (where $\psi_0(p,t) := \psi(p,t)$), and we find that $(\psi_n)_{n \geq 1}$ satisfies a nested sequence of linear fractional differential equations:

$$D^\alpha \psi_1(p,t) = \frac{1}{2} p - \lambda \psi_0(p,t) + p \rho \nu \psi_1(p,t) + \nu^2 \psi_0(p,t) \psi_1(p,t)$$

$$D^{2\alpha} \psi_2(p,t) = -\lambda \psi_1(p,t) + 2 p \rho \nu \psi_2(p,t) + \nu^2 \psi_0(p,t) \psi_2(p,t) + \frac{1}{2} \nu^2 \psi_1(p,t)^2$$

$$D^{n\alpha} \psi_n(p,t) = -\lambda \psi_{n-1}(p,t) + p \rho \nu \psi_n(p,t) + \frac{1}{2} \nu^2 \sum_{k=0}^{n} \psi_k(p,t) \psi_{n-k}(p,t) + \varepsilon \psi_n(p,t)^2$$

with $\psi_n(p,0) = 0$, and in principle we can then compute fractional power series expansions for each $\psi_n(p,t)$ of the form $\psi_n(p,t) = \sum_{m=1}^{n} \mathfrak{a}_{n,m} \rho^m$, as in Remark 33 above.

### 3.4.1 Using these approximations in practice

(31) is of little use in practice, since the leading order Laplace approximation ignores the variation of the function $\frac{1}{p}$ in the integrand, and even if we partially take account of this effect by going to next order with Laplace’s method using the formula in Theorem 7.1 in chapter 4 in [Olv74] (which we have checked and tried), it still frequently gives a worse estimate that the leading order estimate $\tilde{\sigma}(x)$ because the higher order error terms being ignored are too large, and since $H$ is usually very small in practice, $t^H$ converges very slowly to zero. If we instead compute an approximate call price using the Fourier integral along the horizontal contour going through the saddlepoint in [30] (using e.g. the NIntegrate command in Mathematica) and use our higher order asymptotic estimate $\psi(ik,t) + \varepsilon^{\frac{x}{2-H}} \psi_1(ik,t)$ for $\log E(e^{i\frac{X}{\varepsilon^{2-3H}}})$, and then compute the exact implied volatility associated with this price (which avoids the problems with the Laplace approximation), then (for the parameters we considered) we found this approximation to be an order of magnitude closer to the Monte Carlo value than the leading order approximation $\tilde{\sigma}(x)$ (see graph and tables below). See [LK07] for more on computing the optimal contour of integration for such problems.

In principle we can use Corollary 7.2 in Gao&Lee [GL14] to translate (31) into an asymptotic estimate for implied volatility, for which we obtain a cumbersome expression which shows that $\hat{\sigma}_1(x) = \tilde{\sigma}(x) + O(t^{2H} \log t)$, but again in practice we have found this approximation to be of little practical use since the error terms which are ignored are typically too large.

### 3.5 Small-time moderate deviations

Inspired by [BFGHST18], if we replace $\mathfrak{a}_{n,1}$ with

$$g_\varepsilon(p, t) = \frac{\psi_1(p,t)}{\varepsilon^{2-3H}}$$

where $q = \frac{1}{2} - H + \beta$, then we find that

$$D^\alpha \psi_1(p,t) = \frac{1}{2} p - p \frac{\varepsilon^{\frac{1}{2-H}} + \beta}{\varepsilon^{2-3H}} + p \rho \nu \psi_1(p,t) - \varepsilon^{\frac{1}{2-3H} + 4\beta} \lambda \psi_1(p,t) + \frac{1}{2} \varepsilon^{-4H + 6\beta} \nu^2 \psi_1(p,t)^2$$

and we see that all non constant terms on the right hand side are $o(1)$ as $\varepsilon \to 0$ if $\beta \in (\frac{1}{2}H, H)$ and $H \in (0, \frac{1}{2})$. Following similar calculations as above, we formally obtain that $\lim_{\varepsilon \to 0} \varepsilon^{2H-3\beta} \log E(e^{i\frac{X}{\varepsilon^{2-3H}}}) = V_0 I^{1-\alpha} I^{(1/\beta)} = \frac{1}{2} V_0 p^2$ for all $p \in \mathbb{R}$, which (modulo some rigour) implies that $X_1/t^\beta$ satisfies the LDP with speed $\frac{1}{2} V_0 p^2$ and Gaussian rate function $I(x) = \frac{1}{2} x^2 / V_0$. Note that $\beta = H$ corresponds to the central limit or Edgeworth regime, see [FSV19] for details.
Figure 1: Here we have plotted the quadratic function $G(p, w)$ as a function of $w$ for the four cases described in [GGP18]. In cases A and B there are no roots and the solution $\psi(p, t)$ to (21) increases without bound whereas in cases C and D we have a stable fixed point (the lesser of the two roots) and an unstable root, so a solution starting at the origin increases (decreases) until it reaches the stable fixed fixed point. For Case D we have also drawn the curve arising from the reflection transformation used in the proof in Appendix D.

Figure 2: Here we have solved for the solution $f(p, t)$ to (11) numerically by discretizing the VIE with 2000 time steps, and plotted $f(p, t)$ a function of $t$ and the corresponding quadratic function $G(p, w)$ as a function of $w$ with $p$ fixed. In the first case $\alpha = .75$, $\lambda = 2$, $\rho = -0.1$, $\nu = .4$ and $p = 2$ and $f(p, t)$ tends to a finite constant, and in the second case $\alpha = .75$, $\lambda = 1$, $\rho = 0.1$, $\nu = 1$ and $p = 5$ and we see that $f(p, t)$ has an explosion time at some $T^*(p) \approx 0.4$. 
Figure 3: On the left we have plotted $\Lambda(p)$ using an Adams scheme to numerically solve the VIE in (21) with 2000 time steps combined with Corollary 3.4, for $\alpha = 0.75$, $V_0 = 0.04$, $\nu = 0.15$, $\rho = -0.02$, and we find that $p_+ = T^*(1) \approx 34.5$ and $p_- = T^*(-1) \approx 33.25$. On the right we have plotted the corresponding asymptotic small-maturity smile $\hat{\sigma}(x)$ (in blue) versus the higher order approximation using Eq (30) (red “+” signs), and the smile points obtained from a simple Euler-type Monte Carlo scheme with maturity $T = 0.00005$, $10^5$ simulations and 1000 time steps in Matlab (grey crosses), Matlab and Mathematica code available on request. We did not use the Adams scheme to compute $\hat{\sigma}(x)$; rather have used the first 15 terms in the series expansion for $\bar{\Lambda}(p)$ in subsection 3.3 and then numerically computed its Fenchel-Legendre transform and used this to compute $I(x)$ and hence $\hat{\sigma}(x)$. We see that the Monte Carlo and higher order smile points can barely be distinguished by the naked eye. For $|x|$ small, we have found this method of computing $\hat{\sigma}(x)$ to be far superior to using an Adams scheme, since the numerical computation of the fractional integral $I^{1-\alpha} f(p,t)$ for $|t| \ll 1$ can lead to numerical artefacts when computing the FL transform of $\bar{\Lambda}(p,1)$ close to $x = 0$.

Figure 4: On the left here we have the same plot as above but with $T = 0.005$ and for the right plot $T = 0.005$ and $\alpha = 0.6$ (i.e. $H = 0.1$), and again we see that the higher order approximation makes a significant improvement over the leading order smile. Of course we would not expect such close agreement for smaller values of $\alpha$, or larger values of $T$, $|x|$ or $|\rho|$, e.g. $\rho = -0.65$ reported in e.g. [GR18], but the point here is really just to verify the correctness of the asymptotic formula in (20), and give a starting point for other authors/practitioners who wish to test refinements/variants of our formula. We have not repeated numerical results for the large-time case at the current time, since it is intuitively fairly clear that our large maturity formula is correct (since it just boils down to computing the stable fixed point of the VIE) and for maturities $\approx 30$ years with a small step-size, the code would take a prohibitively long time to give good results given that each simulation takes $O(N^2)$ for a rough model (where $N$ is the number of time steps), and it is difficult to verify the formula numerically even for the standard Heston model.
| x  | σ(x) | Higher order T = .00005 | Monte Carlo T = .00005 | Higher order T = .005 | Monte Carlo T = .005 |
|-----|------|-----------------------|-------------------------|-----------------------|---------------------|
| -0.10 | 20.2068% | 20.2023% | 20.2020% | 20.1615% | 20.1589% |
| -0.08 | 20.141% | 20.1364% | 20.1363% | 20.0953% | 20.0931% |
| -0.06 | 20.0869% | 20.0822% | 20.0824% | 20.0407% | 20.0388% |
| -0.04 | 20.045% | 20.0404% | 20.0407% | 19.9986% | 19.9968% |
| -0.02 | 20.016% | 20.0113% | 20.0119% | 19.9693% | 19.9676% |
| 0.00 | 0.00000% | - | 19.9942% | - | 19.9513% |
| 0.02 | 19.9973% | 19.9926% | 19.9921% | 19.9503% | 19.9509% |
| 0.04 | 20.0079% | 20.0033% | 20.0029% | 19.9610% | 19.9613% |
| 0.06 | 20.0316% | 20.0270% | 20.0266% | 19.9850% | 19.9850% |
| 0.08 | 20.068% | 20.0634% | 20.0629% | 20.0218% | 20.0213% |
| 0.10 | 20.1166% | 20.1120% | 20.1114% | 20.0709% | 20.0699% |

Table of numerical results corresponding to the right plot in Figure 3 and the left plot in Figure 4.

4 Large-time asymptotics

In this section, we derive large-time large deviation asymptotics for the Rough Heston model, and we begin making the following assumption throughout this section:

**Assumption 4.1** \( \lambda > 0, \rho \leq 0. \)

Recall that \( f(p, t) \) in (6) satisfies

\[
D^a f(p, t) = H(p, f(p, t))
\]

subject to \( f(p, 0) = 0, \) where \( H(p, w) := \frac{1}{2}p^2 - \frac{1}{2}p(pw - \lambda)w + \frac{1}{2}\nu^2w^2. \) We write

\[
U(p) := \frac{1}{\nu^2} [\lambda - pw - \sqrt{\lambda^2 - 2\lambda p\nu + \nu^2(p(1-p^2))]}
\]

for the smallest root of \( H(p, .) \), and note that \( U(p) \) is real if and only if \( p \in [\underline{p}, \bar{p}] \), where

\[
\underline{p} := \frac{\nu - 2\lambda p - \sqrt{4\lambda^2 + \nu^2 - 4\lambda \nu}}{2\nu(1 - \rho^2)}, \quad \bar{p} := \frac{\nu - 2\lambda p + \sqrt{4\lambda^2 + \nu^2 - 4\lambda \nu}}{2\nu(1 - \rho^2)}.
\]

**Proposition 4.2**

\[
V(p) := \lim_{t \to \infty} \frac{1}{t} \log \mathbb{E}(e^{bX_t}) = \begin{cases} \lambda U_1(p) & p \in [\underline{p}, \bar{p}], \\ +\infty & p \notin [\underline{p}, \bar{p}]. \end{cases}
\]

**Proof.** [GGPTS] show that the explosion time for the Rough Heston model \( T^*(p) < \infty \) if and only if \( T^*(p) < \infty \) for the corresponding standard Heston model (i.e. the case \( \alpha = 1 \)).

From the usual quadratic solution formula \( x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \), we know that \( H(p, .) \) has two distinct real roots (or a single root) if and only if

\[
(\lambda - pw)^2 \geq (p^2 - p)\nu
\]

which is the same as the condition \( e_1(p) \geq 0 \) in condition C) in [GGPTS]. We have already imposed that \( \rho \leq 0 \) and \( \lambda > 0 \), so clearly \( \lambda > \rho \sigma \) and we note that \( \underline{p}, \bar{p} \) are the zeros of \( e_1(p) \).

We now have to verify that under our assumptions that \( \lambda > 0 \) and \( \rho \leq 0 \), \( T^*(p) < \infty \) if and only \( e_1(p) < 0 \). We have two cases to consider to verify this claim:

- Suppose \( e_1(p) \geq 0 \). Then case B in [GGPTS] is impossible by definition, and \( p \in [\underline{p}, \bar{p}] \), and Eq (3.5) in [FJ11] is satisfied. Eq (3.4) in [FJ11] is

\[
\lambda \geq \rho \sigma p
\]

in our current notation, and by the assertion on p.769 in [FJ11] that “(3.4) is implied by (3.5)”, we see that it holds, which is equivalent to \( e(p) < 0 \). Therefore, case A is impossible. So we are in the non-explosive cases C or D of the [GGPTS] classification. Case C is by definition equivalent now to \( e_1(p) > 0 \), and by an easy calculation this is equivalent to \( U_1(p) > 0 \).

- Suppose \( e_1(p) < 0 \). By definition we are not in case C. And we have \( p \notin [\underline{p}, \bar{p}] \), but from p.769 in [FJ11], we know the interval \([0, 1]\) is strictly contained in \([\underline{p}, \bar{p}]\). Hence, case D is also impossible, and we are in the explosive cases A or B.
Hence our claim is verified. We can now re-write \( f(p,t) \) in integral form as

\[
f(p,t) = \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} H(p,f(p,s))ds.\]

Clearly, we have \( H(p,w) \searrow 0 \) as \( w \nearrow U_1(p) \). Assume to begin with that \( U_1(p) > 0 \) (this is case C in the \([17]\) classification). Then from the proof of Proposition 3.6 in \([17]\), we know that \( 0 \leq f(p,t) \leq U_1(p) \).

Moreover, \( w^* = U_1(p) \) is the smallest root of \( H(p,w) \), so \( H(p,w) \geq H_\delta := H(p,U_1(p) - \delta) \) for \( w \leq U_1(p) - \delta \) and \( \delta \in (0,U_1(p)) \); hence we must have

\[
\frac{H_\delta}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} 1_{f(p,s) \leq U_1(p) - \delta} ds < U_1(p)
\]

for all \( t > 0 \). This implies that \( \frac{H_\delta}{\Gamma(\alpha)} (t-1)^{\alpha-1} \int_1^t 1_{f(p,s) \leq U_1(p) - \delta} ds < U_1(p) \), or equivalently

\[
t - 1 - \int_1^t 1_{f(p,s) > U_1(p) - \delta} ds \leq \frac{\Gamma(\alpha)}{H_\delta} U_1(p)(t-1)^{1-\alpha}.
\]

Then we see that

\[
\frac{1}{t} \int_0^t f(p,s)ds \geq \frac{1}{t} \int_1^t f(p,s)ds \geq \frac{1}{t} \int_1^t f(p,s)1_{f(p,s) > U_1(p) - \delta} ds
\]

\[
\geq \frac{1}{t} (U_1(p) - \delta)(t - 1 - \frac{\Gamma(\alpha)}{H_\delta} U_1(p)(t-1)^{1-\alpha})
\]

\[
\geq U_1(p) - 2\delta
\]

for \( t \) sufficiently large. Thus \( U_1(p) - 2\delta \leq \frac{1}{t} \int_0^t f(p,s)ds \leq U_1(p) \), so \( \frac{1}{t} \int_0^t f(p,s)ds \to U_1(p) \) as \( t \to \infty \). Then using that

\[
\log \mathbb{E}(e^{pX_t}) = V_0 t^{1-\alpha} f(p,t) + \lambda \theta I f(p,t)
\]

and that \( f(p,t) \) is bounded, the result follows. We proceed similarly for the case \( U_1(p) < 0 \) (i.e. case D in the \([17]\) classification, see also Lemma 4.3). ■

**Corollary 4.3** \( X_t/t \) satisfies the LDP as \( t \to \infty \) with speed \( t \) and rate function \( V^*(x) \) equal to the Fenchel-Legendre transform of \( V(p) \), as for the standard Heston model.

**Proof.** Since \( U'_1(p) \to +\infty \) as \( p \to \bar{p} \) and \( U''_1(p) \to -\infty \) as \( p \to \bar{p} \), the function \( \lambda U_1(p) \) is essentially smooth; so the stated LDP follows from the Gärtner-Ellis theorem in large deviations theory. ■

**Remark 4.1** We can easily add stochastic interest rates into this model by modelling the short rate \( r_t \) by an independent Rough Heston process, and proceeding as in \([16]\) (we omit the details), see also \([11]\).

Note that we have not proved that \( f(p,t) \to U_1(p) \), but to establish the leading order behaviour in Proposition 4.2 this is not necessary, rather we only needed to show that \( F^{1-\alpha} f(p,t) \sim tU_1(p) \). Nevertheless, this convergence would be required to go to higher order, so for completeness we prove this property as well, as a special case of the following general result:

**Lemma 4.4** Consider functions \( G(y) \) and \( K(z) \) which satisfy the following:

- \( G(y) \) is analytic and increasing on \([0,y_0]\) and decreasing on \([y_0,\infty)\) where \( y_0 \geq 0 \);
- \( G(0) \geq 0 \);
- \( K(z) \) is positive, continuous and strictly decreasing for \( z > 0 \);
- \( \int_0^t K(z)dz \) is finite for each \( t > 0 \) and diverges as \( t \to \infty \);
- \( K(z+\alpha)/K(z) \) is strictly increasing in \( z \) for each fixed \( \alpha \) greater than zero.

Then the solution to \( y(t) = \int_0^t K(t-s)G(y(s))ds \) is monotonically increasing, and if \( G \) has at least one positive root then \( y(t) \) converges to the smallest positive root of \( G \) as \( t \to \infty \).

**Proof.** See Appendix D. ■

This lemma can be applied to both cases C and D. As shown in \([17]\), the solution in case C is bounded between zero and the smallest positive root of \( G \) (denoted \( a \) in that paper) so \( G \) need only satisfy the conditions of the above lemma on the interval \([0,a]\) which it does with \( y_0 = 0 \). For case D, multiplying the defining integral equation by \(-1\) and applying the transformations \(-y(t) \to y(t)\) and \(-G(-y(t)) \to G(y(t))\) (see final plot in Figure 3) we recover an integral equation of the desired form (again \( G \) need only satisfy the conditions of the lemma over the corresponding interval \([0,a]\)).
4.1 Asymptotics for call options and implied volatility

Corollary 4.5 We have the following large-time asymptotic behaviour for European put/call options in the large-time, large log-moneyness regime:

\[ -\lim_{t \to \infty} \frac{1}{t} \log \mathbb{E}(S_t - S_0 e^{x t})^+ = V^*(x) - x \quad (x \geq \frac{1}{2} \bar{\theta}), \]

\[ -\lim_{t \to \infty} \frac{1}{t} \log (S_0 - \mathbb{E}(S_t - S_0 e^{x t})^+) = V^*(x) - x \quad (-\frac{1}{2} \bar{\theta} \leq x \leq \frac{1}{2} \bar{\theta}), \]

\[ -\lim_{t \to \infty} \frac{1}{t} \log (\mathbb{E}(S_0 e^{x t} - S_t)^+) = V^*(x) - x \quad (x \leq -\frac{1}{2} \bar{\theta}), \]

where \(\bar{\theta} = \frac{\lambda \alpha}{\lambda - \rho \nu}\).

Proof. See Corollary 2.4 in [FJ11].]

Corollary 4.6 We have the following asymptotic behaviour in the large-time, large log-moneyness regime, where \(\hat{\sigma}_t(kt)\) is the implied volatility of a European put/call option with strike \(S_0 e^{kt}\):

\[ \hat{\sigma}_\infty(x)^2 = \lim_{t \to \infty} \hat{\sigma}_t^2(x t) = \frac{\omega_1}{2} (1 + \omega_2 \rho x + \sqrt{(\omega_2 x + \rho)^2 + \rho^2}) \]

where

\[ \omega_1 = \frac{4 \lambda \theta}{\nu^2 \rho^2} \left[ \sqrt{(2 \lambda - \rho \nu)^2 + \nu^2 \rho^2} - (2 \lambda - \rho \nu) \right], \quad \omega_2 = \frac{\nu}{\lambda \theta}. \]

Proof. See Proposition 1 in [GJ11] (note that for the Rough Heston model \(\lambda\) has to be replaced with \(\frac{1}{1/\alpha}\) and \(\nu\) replaced with \(\frac{1}{(\alpha)^2}\), but the effect of the \(\alpha\) here cancels out in the final formula for \(\hat{\sigma}_\infty(k)\).]

4.2 Higher order large-time behaviour

We can formally try going to higher order; indeed, using the ansatz \(f(p, t) = U_1(p)t + U_2(p)t^{-\alpha}(1 + o(1))\) for \(p \in \mathbb{P},\), and we find that

\[ U_2(p) = -\frac{U_1(p)}{(\lambda - U_1(p)\nu^2 - p \rho \nu)\Gamma(1 - \alpha)} \]

but if we try and go higher order again, the fractional derivative on the left hand side of (7) does not exist. Using the same approach as in [FJ11], one should be able to use this to compute a higher order large-time saddlepoint approximation for call options. For the sake of brevity, we defer the details of this for future work.

References

[BFGHS18] Bayer, C., P.K.Friz, A.Gulisashvili, B.Horvath, B.Stemper, “Short-Time Near-The-Money Skew In Rough Fractional Volatility Models”, to appear in Quantitative Finance.

[DZ98] Dembo, A. and O.Zeitouni, “Large deviations techniques and applications”, Jones and Bartlet publishers, Boston, 1998.

[Dur10] Durrleman, V., “From Implied to Spot Volatilities”, Finance and Stochastics, 14(2):157-17, 2010.

[EFGGR18] El Euch, O., M.Fukasawa, J.Gatheral and M.Rosenbaum, “Short-term at-the-money asymptotics under stochastic volatility models”, to appear in SIAM Journal on Financial Mathematics.

[EFR18] El Euch, O., M.Fukasawa, and M.Rosenbaum, “The microstructural foundations of leverage effect and rough volatility”, Finance and Stochastics, 12 (6), p. 241-280, 2009.

[EGR18] El Euch, O., Gatheral, J. and M.Rosenbaum, “Roughening Heston”, to appear in Mathematical Finance.

[ER18] El Euch, O. and M.Rosenbaum, “Perfect hedging in Rough Heston models”, The Annals of Applied Probability, 28 (6), 3813-3856, 2018.

[ER19] El Euch, O. and M.Rosenbaum, “The characteristic function of Rough Heston models”, Mathematical Finance, 29(1), 3-38, 2019.

[F11] Forde, M., Large-time asymptotics for an uncorrelated stochastic volatility model, “Statistics&Probability Letters’, 81(8), 1230-1232, 2011.

[FJ11] Forde, M. and A.Jacquier, “The Large-maturity smile for the Heston model”, Finance and Stochastics, 15, 755-780, 2011.
A Computing the kernel for the Rough Heston variance curve

Let $Z_t = \int_0^t \sqrt{V_s} dW_s$, and we recall that

$$V_t = V_0 + \frac{1}{\Gamma(\alpha)} \int_0^t (t - s)^{\alpha - 1} \lambda(\theta - V_s) ds + \frac{1}{\Gamma(\alpha)} \int_0^t (t - s)^{\alpha - 1} \nu \sqrt{V_s} dW_s = \tilde{\xi}_0(t) - \frac{\lambda}{\nu} (\varphi * V) + \varphi * dZ$$

where $*$ denotes the convolution of two functions, $\varphi * dZ = \int_0^t \varphi(t-s) dZ_s$ and $\tilde{\xi}_0(t) = V_0 + \frac{\lambda}{\Gamma(\alpha)} \int_0^t (t - s)^{\alpha - 1} \lambda \theta ds = V_0 + \frac{\lambda}{\alpha \Gamma(\alpha)} t^\alpha$, and $\varphi(t) = \frac{1}{\Gamma(\alpha)} t^\alpha$. Now define $\kappa$ to be the unique function which satisfies

$$\kappa = \varphi - \frac{\lambda}{\nu} (\varphi * \kappa).$$

(A-1)
Such a $\kappa$ exists and is known as the resolvent of $\varphi$. Then we see that

$$V_t - \frac{\lambda}{\nu} \kappa * V_t = \xi_0(t) - \frac{\lambda}{\nu} \varphi * V + \varphi * dZ - \frac{\lambda}{\nu} \kappa * [\xi_0(t) - \frac{\lambda}{\nu} \varphi * V + \varphi * dZ]$$

$$= \xi_0(t) - \frac{\lambda}{\nu} (\varphi - \frac{\lambda}{\nu} \kappa * \varphi) * V + (\varphi - \frac{\lambda}{\nu} \kappa * \varphi) * dZ$$

$$= \xi_0(t) - \frac{\lambda}{\nu} \kappa * V + \kappa * dZ$$

where $\xi_0(t) = \tilde{\xi}_0(t) - \frac{1}{\nu} \kappa * \tilde{\xi}_0(t)$, and we have used (A-1) in the final line. Cancelling the $-\frac{1}{\nu} \kappa * V$ terms, we see that

$$V_t = \xi_0(t) + \kappa * dZ = \xi_0(t) + \int_0^t \kappa(t-s) \sqrt{V_s} dW_s$$

and thus

$$\Rightarrow \xi_t(u) = \mathbb{E}(V_u|\mathcal{F}_t) = \xi_0(u) + \int_0^t \kappa(u-s) \sqrt{V_s} dW_s$$

and thus

$$d\xi_t(u) = \kappa(u-t) \sqrt{V_t} dW_t$$

i.e. the correct $\kappa$ function is the solution to (A-1). If we take the Laplace transform of (A-1), we get

$$\hat{\kappa}(z) = \hat{\varphi}(z) - \frac{\lambda}{\nu} \hat{\varphi}(z) \hat{\kappa}(z).$$

(A-2)

and (A-2) is just an algebraic equation now, which we can solve explicitly to get $\hat{\kappa}(z) = \frac{\hat{\varphi}(z)}{1 + \frac{\lambda}{\nu} \hat{\varphi}(z)}$. But we know that $\varphi(t) = \frac{\nu}{\Gamma(\alpha)} t^\alpha$ whose Laplace transform is $\hat{\varphi}(z) = \nu z^{-\alpha}$, so $\hat{\kappa}(z)$ evaluates to

$$\hat{\kappa}(z) = \frac{\nu z^{-\alpha}}{1 + \lambda z^{-\alpha}}.$$

Then the inverse Laplace transform of $\hat{\kappa}(z)$ is given by

$$\kappa(x) = \nu x^{\alpha-1} E_{\alpha,\alpha}(-\lambda x^\alpha).$$

B The re-scaled model

We first let

$$dX^*_\epsilon = -\frac{1}{2} \epsilon V^*_\epsilon dt + \sqrt{\epsilon} \sqrt{V^*_\epsilon} dW_t$$

$$V^*_\epsilon - V_0 = \epsilon^{\gamma-1} \frac{\gamma}{\Gamma(\alpha)} \int_0^{\epsilon t} (t-s)^{H-\frac{\gamma}{2}} \lambda(\theta - V^*_{s/\epsilon}) ds + \frac{\epsilon H}{\Gamma(\alpha)} \int_0^{\epsilon t} (t-s)^{H-\frac{1}{2}} \nu \sqrt{V^*_s} dW_s$$

$$\mathrel{\overset{(d)}{=}} \epsilon^{\gamma-1} \frac{\gamma}{\Gamma(\alpha)} \int_0^{\epsilon t} (t-s)^{H-\frac{\gamma}{2}} \lambda(\theta - V^*_{s/\epsilon}) ds + \frac{\epsilon H}{\Gamma(\alpha)} \int_0^{\epsilon t} (t-s)^{H-\frac{1}{2}} \nu \sqrt{V^*_s} dW_s$$

$$= \epsilon^{\gamma-1} \frac{\gamma}{\Gamma(\alpha)} \int_0^{\epsilon t} (t-u)^{H-\frac{\gamma}{2}} \lambda(\theta - V^*_{u/\epsilon}) du + \frac{\epsilon H}{\Gamma(\alpha)} \int_0^{\epsilon t} (t-u)^{H-\frac{1}{2}} \nu \sqrt{V^*_u} dW_u.$$
D Proof of monotonicity of the solution for a general class of Volterra integral equations

Recall that $y(t)$ satisfies

$$y(t) = \int_0^t K(t-s)G(y(s))ds$$

One can easily verify that the kernel used for the Rough Heston model satisfies the stated properties in Lemma 4.4.

In the classical case $K(t) \equiv 1$ the integral eq clearly reduces to an ODE, and it is well known that the solution of this is at least continuously differentiable on the domain of existence. In the following it will be assumed that the solution $y(t)$ is analytic for $t > 0$. This is proved for the kernel relevant to the Rough Heston model in [GGP18] (Theorem 6), see also the end of page 14 in [GGP18].

What follows is a natural extension of the technique used in [MW51] (Theorem 8). Using the properties of convolution and differentiating under the integral sign, we have:

$$y(t) = \int_0^t K(t-s)G(y(s))ds = \int_0^t K(s)G(y(t-s))ds$$

$$y'(t) = K(t)G(0) + \int_0^t K(s)G'(y(t-s))y'(t-s)ds$$

$$y'(t) = K(t)G(0) + \int_0^t K(t-s)G'(y(s))y'(s)ds$$

$$G(0) > 0 \text{ so } y'(t) \to +\infty \text{ as } t \to 0^+ \text{ and since } G(y) \text{ is increasing for } y \leq y_0 \text{ we have that } y'(t) > 0 \text{ until } y(t) \text{ reaches } y_0 \text{ i.e. the solution increases. For } y \geq y_0 \text{, } G(y) \text{ is decreasing and suppose that } y(t) \text{ ceases to be increasing at some point.}$$

This implies (assuming a continuous derivative) the existence of a $t_0$ and an interval $I = [t_0, t_1]$ such that $y'(t_0) = 0$ and $y(t_1) < 0$ for all $t_1 \in I$ (if $y(t)$ and hence $y'(t)$ is analytic then the zeros of the derivative are isolated and a sufficiently small interval $I$ exists). Using the integral equation for $y'(t)$:

$$y'(t_0) = K(t_0)G(0) + \int_0^{t_0} K(t_0-s)G'(y(s))y'(s)ds = 0 \quad (D-4)$$

$$y'(t_1) = K(t_1)G(0) + \int_0^{t_0} K(t_1-s)G'(y(s))y'(s)ds + \int_{t_0}^{t_1} K(t_1-s)G'(y(s))y'(s)ds$$

We can re-write the kernels in the first and second terms of the expression for $y'(t_1)$ as:

$$K(t_1) = \frac{K(t_1)}{K(t_0)}K(t_0), \quad K(t_1-s) = \frac{K(t_1-s)}{K(t_0-s)}K(t_0-s)$$

and we can easily check that the quotient in the second expression here decreases monotonically from $K(t_1)/K(t_0)$ to zero.

By the mean value theorem for definite integrals there exists a $\tau \in (t_0, t_1)$ such that:

$$\int_0^{t_0} \frac{K(t_1-s)}{K(t_0-s)}K(t_0-s)G'(y(s))y'(s)ds = \frac{K(t_1-s)}{K(t_0-s)}K(t_0)G(0)$$

$$= -\frac{K(t_1-s)}{K(t_0-s)}K(t_0)G(0) \quad (D-5)$$

where the second equality follows from (D-4). Substituting this into our expression for $y'(t_1)$:

$$y'(t_1) = \frac{K(t_1)}{K(t_0)}K(t_0)G(0) + \frac{K(t_1-s)}{K(t_0-s)}K(t_0)G(0)$$

$$= \frac{K(t_1)}{K(t_0)}K(t_0)G(0) - \int_0^{t_0} K(t_0-s)G'(y(s))y'(s)ds + \int_{t_0}^{t_1} K(t_1-s)G'(y(s))y'(s)ds$$

$$= K(t_0)G(0)\left[\frac{K(t_1)}{K(t_0)} - \frac{K(t_1-s)}{K(t_0-s)}\right] + \int_{t_0}^{t_1} K(t_1-s)G'(y(s))y'(s)ds > 0 \quad (D-6)$$

and we have used (D-4) in the second line. But this is a contradiction so the solution remains increasing.

As discussed elsewhere in this paper, when studying the Rough Heston model, the non-linearity in the integral equation has the generic form $G(y) = (y - \theta_1)^2 + \theta_2$ i.e. a quadratic with positive leading coefficient (for simplicity set to 1 here) and minimum of $\theta_2$ obtained at $y = \theta_1$. Depending on the values of $\{\theta_1, \theta_2\}$ the following cases due to [GGP18] are distinguished:

- (C) $G(0) > 0$, $\theta_1 > 0$ and $\theta_2 < 0$
Case C is already in the form considered here with \( y_0 = 0 \). In case D, applying the transformation \( y(t) \to -y(t) \) and \(-G(-y(t)) \to G(y(t)) \) (reflecting in the \( x \) and then \( y \) axis) yields a function \( G(y) \) which is a quadratic with negative leading coefficient and thus increases until it reaches its maximum after which it decreases which is of the type considered here.

As shown in Propositions 3.6 and 3.7 in [GGP18], solutions to this integral equation must be bounded between 0 and \( a \) where \( a \) is the first positive root of \( G(y) \), and monotonicity of \( y(t) \) implies that \( y(t) \to a \) as \( t \to \infty \) (since if \( y(t) \) were to tend to a constant \( c_1 \) with \( 0 < c_1 \leq a \), then \( G(y(t)) \) will be bounded below by some \( G^* > 0 \), so

\[
y(t) = \int_0^t K(t-s)G(y(s))ds \geq G^* \int_0^t K(t-s)ds \to \infty
\]

which contradicts the boundedness of \( y(t) \).