WILEY MATHEMATICAL FINANCE An International Journal of Mathematics, Statistics and Financial Economics

ORIGINAL ARTICLE

Option pricing in the moderate deviations regime

Peter Friz[1] | Stefan Gerhold[2] | Arpad Pinter[2]

[1]TU and WIAS Berlin, Germany

[2]TU Wien, Austria

Correspondence
Peter Friz, Institut für Mathematik, Technische Universität Berlin, Strasse des 17. Juni 136, 10623 Berlin, Germany.
Email: friz@math.tu-berlin.de

Abstract

We consider call option prices close to expiry in diffusion models, in an asymptotic regime ("moderately out of the money") that interpolates between the well-studied cases of at-the-money and out-of-the-money regimes. First and higher order small-time moderate deviation estimates of call prices and implied volatilities are obtained. The expansions involve only simple expressions of the model parameters, and we show how to calculate them for generic local and stochastic volatility models. Some numerical computations for the Heston model illustrate the accuracy of our results.

KEYWORDS

asymptotics, implied volatility, moderate deviations, option pricing

1 | INTRODUCTION

Consider a European call option struck at K with remaining time to expiry $t > 0$ and no-arbitrage price[1] $C(K, t)$. Today's price of the underlying, the spot value S_0, is known and fixed. Discrete option data are available from the market, typically quoted in (Black–Scholes) implied volatilities; see Figure 1.1. Many option pricing models have been proposed to combine reasonable dynamics for the underlying, small number of parameters and acceptable fits to the data. However, with the notable exception of the Black–Scholes model, closed-form expressions for call prices are scarce, and approximate pricing formulae have been proposed as substitute: often used to improve calibration, but also toward a better quantitative understanding of a given model. (A classic reference in this context is Gatheral, 2006.)

More specifically, small-maturity approximations of option prices have been studied extensively in recent years. Starting with Carr and Wu (2003), it was understood that the asymptotic behavior of $C(K, t)$ as $t \downarrow 0$ exhibits very different behavior in the respective cases $K > S_0$ ("out-of-the-money")
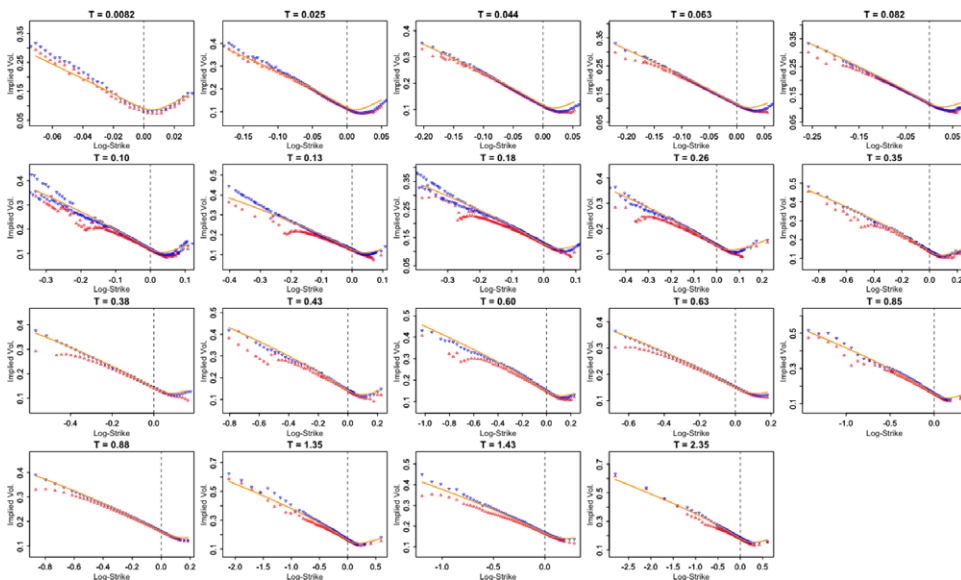
FIGURE 1.1 SPX volatility smiles as of August 14, 2013 (courtesy of J. Gatheral). Strikes of options with small remaining time to maturity ($T = 0.0082$) are about $e^{0.02} - 1 \approx 2\%$ around the money (spot); good data for a later time $T = 0.26$ already have a range of $\approx 30\%$. The highest maturity $T = 2.35$ has a range of $\approx 65\%$ around the money [Color figure can be viewed at wileyonlinelibrary.com]

and $K = S_0$ ("at-the-money"). We argue that there is a significant asymptotic regime in between, namely,

$$\sqrt{t} \ll K - S_0 \ll 1.$$

It has received little attention, and, to the best of our knowledge, none at all in the classical diffusion case. The aim of the present paper is to fill this gap. This "moderately out-of-the-money" regime in fact reflects the reality of quoted option prices: As seen in Figure 1.1, the range of strikes tends to concentrate "around-the-money" as time to expiry becomes small. At the same time, the regime offers excellent analytic tractability.

To put our results into perspective, we recall some well-known facts on option price approximations close to expiry. We write $c(k, t)$ for the normalized call price as a function of log-moneyness $k = \log(K/S_0)$

$$C\left(S_0 e^k, t\right)/S_0 = c(k, t). \tag{1.1}$$

In general, $c(k, t)$ depends tacitly on S_0, the (fixed) spot value.[2] We start with the at-the-money (ATM) regime $k = 0$. In the Black–Scholes model, writing $c(k, t) = c_{BS}(0, t; \sigma)$ with volatility parameter $\sigma > 0$, we have the following ATM call price behavior

$$c_{BS}(0, t; \sigma) \sim \frac{\sigma\sqrt{t}}{\sqrt{2\pi}}, \qquad t \downarrow 0.$$

From Muhle-Karbe and Nutz (2011), the same is actually true in a generic semimartingale model with diffusive component (with spot volatility $\sigma_0 = \sqrt{v_0} > 0$),

$$c(0, t) \sim \frac{\sigma_0\sqrt{t}}{\sqrt{2\pi}}, \qquad t \downarrow 0, \tag{1.2}$$

and this translates to the generic ATM implied variance formula (even in presence of jumps, as long as $v_0 > 0$)

$$\sigma_{\text{imp}}^2(0, t) = v_0 + o(1), \qquad t \downarrow 0.$$

(We use the notation $\sigma_{\text{imp}}(k, t)$ for the Black–Scholes implied volatility with log-moneyness k and maturity t.) Higher order terms in t will be model dependent. For instance, in the Heston case, with variance dynamics $dV = -\kappa(V - \bar{v})dt + \eta\sqrt{V}dW$, implied variance has the ATM expansion

$$\sigma_{\text{imp}}^2(0, t) = v_0 + a(0)t + o(t), \qquad t \downarrow 0,$$

$$a(0) = -\frac{\eta^2}{12}\left(1 - \frac{\rho^2}{4}\right) + \frac{v_0\rho\eta}{4} + \frac{\kappa}{2}\left(\bar{v} - v_0\right). \tag{1.3}$$

This is corollary 4.4 in Forde, Jacquier, and Lee (2012), and we note that $a(0)$ has no easy interpretation in terms of the model parameters.

Relaxing $k = 0$ to $k_t = o(\sqrt{t})$ amounts to what we dub "almost ATM" (AATM) regime.[3] (In particular, $k_t \sim t^\beta$ is in the AATM regime if and only if $\beta > 1/2$.) Again for generic semimartingale models with diffusive component and spot volatility $\sigma_0 > 0$, it is easy to see from Caravenna and Corbetta (2016) and Muhle-Karbe and Nutz (2011) that the ATM asymptotics (1.2) imply the almost ATM asymptotics

$$c(k_t, t) \sim \frac{\sigma_0\sqrt{t}}{\sqrt{2\pi}}, \qquad k_t = o\left(\sqrt{t}\right), \qquad t \downarrow 0.$$

This fails when k_t ceases to be $o(\sqrt{t})$. Indeed, for $k_t = \theta\sqrt{t}$ with constant factor $\theta > 0$, we have, from Caravenna and Corbetta (2016) and Muhle-Karbe and Nutz (2011),

$$c(k_t, t) \sim \mathbb{E}\left[N\left(-\theta, \sigma_0^2\right)^+\right]\sqrt{t}, \qquad t \downarrow 0,$$

where $N(-\theta, \sigma_0^2)$ stands for a Gaussian random variable with mean $-\theta$ and variance σ_0^2. This, too, holds true in the stated semimartingale generality. In any case, the proof is based on the Lévy case with nonzero diffusivity v_0, and the result follows from comparison results, which imply that the difference is negligible to first order. For a thorough discussion of the regime $k = O(\sqrt{t})$ in the (local) diffusion case, see Pagliarani and Pascucci (2017).

Beyond this regime, call price asymptotics change considerably. For instance, take an additional slowly diverging factor $\log(1/t)$,

$$k_t = \theta\sqrt{t\log(1/t)}.$$

Even in the Black–Scholes model, we now lose the $\sqrt{t}$-behavior of call prices described above and in fact

$$c_{\text{BS}}(k_t, t; \sigma) = t^{\frac{1}{2} + \frac{\theta^2}{2\sigma^2}}\ell(t),$$

for some slowly varying function $\ell(t)$, see Mijatović and Tankov (2016). On the other hand, in a genuine out-of-the-money (OTM) situation, with $k_t \equiv k > 0$ fixed, option values are exponentially small in diffusion models, and we are in the realm of *large deviations theory*. For instance,

$$c_{\text{BS}}(k, t; \sigma) = \exp\left(-\frac{\Lambda_{\text{BS}}(k)}{t}(1 + o(1))\right), \qquad k > 0 \text{ fixed}, \ t \downarrow 0,$$

TABLE 1.1 Asymptotic behavior of short-maturity call options, $t \downarrow 0$

	ATM (At-the-Money)	AATM (Almost At-the-Money)	MOTM (Moderately Out-of-the-Money)	OTM (Out-of-the-Money)
Process type	$K = S_0$	$\log \frac{K}{S_0} \sim (const)t^\beta$ $\beta > 1/2$	$\log \frac{K}{S_0} \sim (const)t^\beta$ $0 < \beta < 1/2$	$\log \frac{K}{S_0} \equiv k > 0$
Black–Scholes	$O(\sqrt{t})$, elementary	$O(\sqrt{t})$, elementary	$\exp(-\frac{const}{t^{1-2\beta}})$, elementary	$\exp(-\frac{const}{t})$, elementary
Stochastic or local volatility (diffusion model)	$O(\sqrt{t})$,	$O(\sqrt{t})$,	$\exp(-\frac{const}{t^{1-2\beta}})$	$\exp(-\frac{const}{t})$
Jump diffusion/general semimartingale with diff. component	$O(\sqrt{t})$	$O(\sqrt{t})$	$O(t)$ in finite variation Lévy models[4]	$O(t)$, see Bentata and Cont (2012)

with $\Lambda_{\mathrm{BS}}(k) = \frac{1}{2}k^2/\sigma^2$ in the Black–Scholes model. Similar results appear in the literature, with different levels of mathematical rigor, for other and/or generic diffusion models; see Berestycki, Busca, and Florent (2002), Carr and Wu (2003), Forde and Jacquier (2009), and Paulot (2015). Table 1.1 summarizes first-order call price asymptotics in various models and regimes.

Throughout the paper, we reserve the term OTM for *fixed* OTM log-strike $k > 0$, to distinguish this regime from the *moderately* out-of-the-money regime that we now define. Our basic observation is that for

$$k_t \sim (const)t^\beta, \qquad t \downarrow 0, \tag{1.4}$$

the cases of $\beta > \frac{1}{2}$, resp. $\beta = 0$, are covered by the before-discussed AATM, resp. OTM, results. This leaves open a significant gap, namely, $\beta \in (0, \frac{1}{2})$, which we call moderately out-of-the-money (MOTM). We have a threefold interest in this MOTM regime,

$$k_t \sim (const)t^\beta, \qquad t \downarrow 0, \qquad \text{for } \beta \in \left(0, \frac{1}{2}\right). \tag{1.5}$$

(i) First, it is related to the *reality of quoted (short-dated) option prices*, where strikes of option price data with acceptable bid–ask spreads tend to accumulate "around the money," as illustrated in Figure 1.1. To account for this accumulation, we consider strikes that move closer to the money as expiry shrinks, and the simplest way to do so is to consider strikes of the order $k = O(t^\beta)$ for some $\beta > 0$. There is no reason why quoted strikes should always be almost ATM ($\beta > 1/2$), which effectively means an extreme concentration around the money; we are thus led to study the regime (1.5).

(ii) The second reason is *mathematical convenience*. In contrast to the genuine OTM regime (large deviation regime) in which the rate function $\Lambda(k)$ is notoriously difficult to analyze—often related to geodesic distance problems—MOTM naturally comes with a quadratic rate function and, most remarkably, higher order expansions are always explicitly computable in terms of the model parameters. The terminology *moderately* OTM (MOTM) is in fact in reference to *moderate deviations theory*, which effectively interpolates between the central limit and large deviations regimes.[5] This also identifies the AATM regime as bordering the central limit regime, where

asymptotics are precisely those of the Black–Scholes model, which in turn is the rescaled Gaussian (in log-coordinates) limit of a general semimartingale model with diffusive component.

(iii) Finally, our third point is that MOTM expansions naturally involve quantities very familiar to practitioners, notably, spot (implied) volatility, implied volatility skew, and so on.

In the Black–Scholes model, it is easy to check that we have the MOTM asymptotics

$$c_{\mathrm{BS}}(k_t \equiv \theta t^\beta, t; \sigma) = \exp\left(-\frac{\theta^2}{2\sigma^2 t^{1-2\beta}}(1 + o(1))\right), \qquad t \downarrow 0.$$

Loosely speaking, our main results (Theorems 2.3 and 2.5 below) assert that such relations (even of higher order) are true in great generality for diffusion models, and that all quantities are computable and then related to implied volatility expansions. We note in passing that, for Lévy models, the regime (1.5) has been studied by Mijatović and Tankov (2016); then, call prices decay algebraically rather than exponentially. For recent related results on *fractional* stochastic volatility models, see Forde and Zhang (2017) and Guennoun, Jacquier, and Roome (2014). Guillin (2003), who considers small-noise moderate deviations of diffusions, should also be mentioned here; however, in Guillin (2003), the dynamics depend on a "fast" random environment (with motivation from physics, and no obvious financial interpretation), and the nondegeneracy assumption (D) is not satisfied in our context. The recent related paper by Gao and Wang (2016) contains a *first-order* moderate deviation principle (MDP) for diffusions under classical regularity assumptions from SDE theory. The main difference to the bulk of our results is that we develop *higher order* expansions, until Section 6 where we revisit first-order MOTM estimates from a moment-generating function perspective. However, in this case the underlying models (e.g., Heston) fall immediately outside the scope of Gao and Wang (2016), because of the typical square-root structure of coefficients. (The matter is discussed in more detail at the beginning of Section 6.)

To round off the introduction, we briefly recall some background on moderate deviations. Consider the classical setting of a centered i.i.d. sequence $(X_n)_{n \geq 1}$ with finite exponential moments. Then the empirical means $\hat{X}_n := n^{-1} \sum_{k=1}^n X_k$ converge to zero (law of large numbers, LLN), and this is quantified by an LDP according to Cramér's classical theorem: $\mathbb{P}[\hat{X}_n > x] = \exp(-I(x)n + o(n))$ decreases exponentially as $n \to \infty$ for fixed $x > 0$, governed by a rate function $I(x) = \sup_{y \in \mathbb{R}}(yx - \log \mathbb{E}[e^{yX_1}])$. On the other hand, by the CLT, $\sqrt{n}\hat{X}_n = n^{-1/2} \sum_{k=1}^n X_k$ has a Gaussian limit law. Moderate deviations cover intermediate scalings, i.e., $\sqrt{na_n}\hat{X}_n$ with $a_n \to 0$ and $na_n \to \infty$. It turns out (theorem 3.7.1 in Dembo & Zeitouni, 1998) that, for any such sequence $a_n > 0$, the family $\sqrt{na_n}\hat{X}_n$ satisfies an LDP with speed $1/a_n$ and qquadratic rate function. (A natural scaling family is given by $a_n = n^{2\beta-1}$, with parameter $\beta > 0$, so that one considers $\sqrt{na_n}\hat{X}_n = n^{\beta-1} \sum_{k=1}^n X_k$. Interpolation between LDP, with LLN scaling, and CLT scaling then amounts to considering $0 < \beta < 1/2$.) This is sometimes called an *MDP*. Formally, an MDP is thus just a certain LDP with appropriate scaling and speed function. Still, the terminology is often useful because of the trichotomy

CLT   –   MDP for a range of scalings, with quadratic rate function   –   genuine LDP,

which occurs in many situations, not just for i.i.d. sequences of random variables. For references to some other classical moderate deviations results (on empirical measures); e.g., see sections 6.7 and 7.4 of Dembo and Zeitouni (1998). Several authors have investigated moderate deviations in actuarial risk theory; see, e.g., Fu and Shen (2017) and references therein.

The rest of the paper is organized as follows. Section 2 contains our main results, which translate asymptotics for the transition density of the underlying into MOTM call price asymptotics. The

corresponding proofs are presented in Section 3. Section 4 and the Appendix give the implied volatility expansion resulting from our call price approximations. Section 5 applies our main results to standard examples, namely, generic local volatility models (Section 5.1), generic stochastic volatility models (Section 5.2), and the Heston model (Section 5.3). As usual, the square-root degeneracy of the Heston model makes it difficult to apply results for general stochastic volatility models, so we verify the validity of our results—if formally applied to Heston—by a direct "affine" analysis. Finally, in Section 6 we present a second approach to MOTM estimates, which employs the Gärtner–Ellis theorem from large deviations theory. Throughout we take zero rates, a natural simplification in view of our short-time consideration. Also, w.l.o.g. we normalize spot to $S_0 = 1$.

2 | MOTM OPTION PRICES VIA DENSITY ASYMPTOTICS

We consider a *general stochastic volatility* model, i.e., a positive martingale $(S_t)_{t\geq0}$ with dynamics

$$dS_t = S_t\sigma_t dW_t,$$

and started (w.l.o.g.) at $S_0 = 1$. We assume that the stochastic volatility process $(\sigma_t)_{t\geq0}$ itself is an Itô-diffusion, started at some deterministic value σ_0, called *spot volatility*. Recall that in any such stochastic volatility model, the local (or effective) volatility is defined by

$$\sigma_{\text{loc}}^2(t, K) := \mathbb{E}\left[\sigma_t^2 | S_t = K\right].$$

As is well known, the equivalent local volatility model

$$d\tilde{S}_t = \tilde{S}_t\sigma_{\text{loc}}(t, \tilde{S}_t)dW_t$$

has the property that $\tilde{S}_t = S_t$ (in law) for all fixed times. See Brunick and Shreve (2013) for precise technical conditions under which this holds true.[6] As a consequence, European option prices $C(K,t)$ match in both models. Recall also Dupire's formula in this context

$$\sigma_{\text{loc}}^2(K, t) = \frac{\partial_t C(K, t)}{\frac{1}{2}K^2\partial_{KK}C(K, t)}, \qquad t > 0, K > 0. \tag{2.1}$$

We now state our two crucial conditions.

Assumption 2.1. *For all $t > 0$, S_t has a continuous pdf $K \mapsto q(K, t)$, which behaves asymptotically as follows for small time:*

$$q(K, t) \sim e^{-\frac{\Lambda(k)}{t}}t^{-1/2}\gamma(k), \qquad t \downarrow 0, \tag{2.2}$$

uniformly for $K = e^k$ in some neighborhood of $S_0 = 1$. The energy function Λ is smooth in some neighborhood of zero, with $\Lambda(0) = \Lambda'(0) = 0$. Moreover, $\lim_{k\to0}\gamma(k) = \gamma(0) > 0$.

Assumption 2.2. *For $t \downarrow 0$ and $K \to S_0 = 1$, the local volatility function of $(S_t)_{t\geq0}$ converges to spot volatility*

$$\lim_{\substack{K \to S_0 \\ t\downarrow0}} \sigma_{\text{loc}}(K, t) = \sigma_0. \tag{2.3}$$

The latter assumption is fairly harmless (in diffusion models; see the beginning of Section 5.2). The first assumption is potentially (very) difficult to check, but fortunately we can rely on substantial recent progress in this direction; see Deuschel, Friz, Jacquier, and Violante (2014a, 2014b) and Osajima (2015). We shall see in Section 5.2 that both assumptions indeed hold in generic stochastic volatility models. Let us also note the fundamental relation between spot volatility σ_0 (actually equal to implied spot volatility $\sigma_{\text{imp}}(0, 0)$ here) and the Hessian of the energy function $\Lambda = \Lambda(k)$,

$$\sigma_0 = \Lambda''(0)^{-1/2}.$$

(This is well known [see Durrleman, 2004] and also follows from Proposition 2.4 below.) Now we state our main result. We slightly generalize the log-strikes considered in (1.5), replacing the constant factor by an arbitrary slowly varying function ℓ.

Theorem 2.3. *Under Assumptions 2.1 and 2.2, consider MOTM calls, in the sense that log-strike is*

$$k_t = t^\beta \ell(t), \qquad t > 0, \tag{2.4}$$

where $\ell > 0$ varies slowly at zero and $\beta \in (0, \frac{1}{2})$.

 (i) *The call price satisfies the moderate deviation estimate*

$$c(k_t, t) = \exp\left(-\frac{\Lambda''(0)}{2} \frac{k_t^2}{t} (1 + o(1)) \right)$$

$$= \exp\left(-\frac{1}{2\sigma_0^2} \frac{k_t^2}{t} (1 + o(1)) \right), \qquad t \downarrow 0. \tag{2.5}$$

(ii) *If we restrict β to $(0, \frac{1}{3})$, then the following moderate second-order expansion holds true*

$$c(k_t, t) = \exp\left(-\frac{1}{2}\Lambda''(0)\frac{k_t^2}{t} - \frac{1}{6}\Lambda'''(0)\frac{k_t^3}{t} (1 + o(1)) \right)$$

$$= \exp\left(-\frac{1}{2\sigma_0^2} \frac{k_t^2}{t} \left(1 - \frac{S}{\sigma_0^2} k_t(1 + o(1)) \right) \right), \qquad t \downarrow 0, \tag{2.6}$$

with spot-variance σ_0^2, equal to $\sigma_{\text{imp}}^2(0, 0)$, and implied variance skew $S = \frac{\partial}{\partial k}|_{k=0}\sigma_{\text{imp}}^2(k, 0)$.

In particular, for $\ell \equiv \theta > 0$, we have the (first-order) expansion

$$t^{1-2\beta} \log c(\theta t^\beta, t) \sim -\frac{\theta^2}{2\sigma_0^2}, \qquad t \downarrow 0,$$

exhibiting a qquadratic rate function $\theta \mapsto \theta^2/2\sigma_0^2$, typical of *moderate deviation problems.*[7]

In a nutshell, (2.5) says that inserting the time-dependent log-strike (2.4) into the fixed-strike OTM/LD approximation $c(k, t) = \exp(-\Lambda(k)/t (1 + o(1)))$ yields a correct result, upon Taylor expanding Λ. Mind, however, that this needs a proof using the specifics of our situation, in light of the fact that validity of a large deviation principle does not automatically imply an MDP.

The quantities $\Lambda''(0), \Lambda'''(0), \ldots$ appearing above are *always* computable from the initial values and the diffusion coefficients of the stochastic volatility model. This is in stark contrast to the OTM

regime, where one needs the function $\Lambda(\cdot)$, which is in general not available in closed form (with some famous exceptions, like the SABR model). We quote the following result on N-factor models from Osajima (2015) and refer to Section 5.2 for detailed calculations in a two-factor stochastic volatility model.

Proposition 2.4. *Assume that* $(\log S, \sigma^1, \dots, \sigma^{N-1})$ *is Markov, started at* $(0, \bar{\sigma}_0)$ *with* $\bar{\sigma}_0 \in \mathbb{R}^{N-1}$ *and* $\bar{\sigma}_0^1 > 0$, *with stochastic volatility* $\sigma \equiv \sigma^1$, *where the generator has (nondegenerate) principal part* $\sum a^{ij} \partial_{ij}$ *in the sense that* a^{-1} *defines a Riemannian metric. Then*

$$\Lambda(k) = \frac{1}{2b_1} k^2 - \frac{b_2}{3b_1^3} k^3 + \left(-\frac{b_3}{4b_1^4} + \frac{b_2^2}{2b_1^5} \right) k^4 + O(k^5), \qquad k \to 0,$$

where the coefficients are given by

$$b_1 = \int_0^1 a^{11}(t, \bar{\sigma}_0) dt,$$

$$b_2 = \frac{3}{2} \int_0^1 (V a^{11})(t, \bar{\sigma}_0) dt,$$

$$b_3 = 2 \int_0^1 (V^2 a^{11})(t, \bar{\sigma}_0) dt + \frac{1}{2} \int_0^1 \Gamma\left(a^{11}, a^{11} \right)(t, \bar{\sigma}_0) dt,$$

using the functions

$$(V f)(t, x) = \sum_{i=1}^N a^{1i}(t, x) \int_t^1 \frac{\partial f}{\partial x_i}(s, x) ds,$$

$$\Gamma(f, g)(t, x) = \sum_{i,j=1}^N a^{ij}(t, x) \left(\int_t^1 \frac{\partial f}{\partial x_i}(s, x) ds \right) \left(\int_t^1 \frac{\partial g}{\partial x_j}(s, x) ds \right).$$

Proof. See Osajima (2015, theorem 1(1), with $T = 1$). $\qquad \square$

The following result presents a higher order expansion in the MOTM regime. It yields an asymptotically equivalent expression for call prices (and not just logarithmic asymptotics).

Theorem 2.5. *Under the assumptions of Theorem 2.3, the logarithm of the call price has the refined MOTM expansion*

$$\log c(k_t, t) = - \sum_{m=2}^{\lfloor 1/\beta \rfloor} \frac{\Lambda^{(m)}(0)}{m!} \frac{k_t^m}{t}$$

$$+ \left(2\beta - \frac{3}{2} \right) \log \frac{1}{t} - 2 \log \ell(t) + \log \left(\gamma(0) v_0^2 \right) + o(1), \qquad t \downarrow 0. \qquad (2.7)$$

This can be expressed equivalently as

$$c(k_t, t) \sim \gamma(0) v_0^2 \frac{t^{3/2 - 2\beta}}{\ell(t)^2} \exp\left(- \sum_{m=2}^{\lfloor 1/\beta \rfloor} \frac{\Lambda^{(m)}(0)}{m!} \frac{k_t^m}{t} \right), \qquad t \downarrow 0.$$

If $1/\beta$ is not an integer, then k_t^m/t tends to infinity for $m = \lfloor 1/\beta \rfloor$, of order $t^{\beta \lfloor 1/\beta \rfloor - 1}$ (up to a slowly varying factor). If, on the other hand, $1/\beta$ is an integer, then the last summand of the sum $\sum_{m=2}^{\lfloor 1/\beta \rfloor}$

in (2.7) is of order $\ell(t)$, which means that the following term $\log(1/t)$ may be asymptotically larger. The upper summation limit $\lfloor 1/\beta \rfloor$ thus ensures that no irrelevant (i.e., $o(1)$) terms are contained in the sum. Note that $\lfloor 1/\beta \rfloor = 2$ for $\beta \in (\frac{1}{3}, \frac{1}{2})$, and $\lfloor 1/\beta \rfloor \geq 3$ for $\beta \in (0, \frac{1}{3})$, and so (2.7) is consistent with (2.5), resp. (2.6).

The passage from the derivatives of the energy function to ATM derivatives of the implied volatility in the short time limit is best conducted via the BBF formula that was proved in Berestycki et al. (2002). (That said, theses relations are also a direct consequence of our expansions, as is pointed out in Section 4.) In this regard, we have:

Theorem 2.6. *Suppose that Λ is a function with the properties required in Assumption 2.1, with $\Lambda''(0) = \sigma_0^{-2} = v_0^{-1}$, and that the Berestycki–Busca–Florent formula $\sigma_{\mathrm{imp}}^2(0, k) = k^2/2\Lambda(k)$ holds. Then the small-time ATM implied variance skew and curvature, respectively, relate to Λ via*

$$S := \left.\frac{\partial}{\partial k}\right|_{k=0} \sigma_{\mathrm{imp}}^2(k, 0) = -\frac{1}{3}\frac{\Lambda'''(0)}{\Lambda''(0)^2} \tag{2.8}$$

and

$$C := \left.\frac{\partial^2}{\partial k^2}\right|_{k=0} \sigma_{\mathrm{imp}}^2(k, 0) = \frac{\frac{2}{3}\Lambda'''(0)^2 - \frac{1}{2}\Lambda''''(0)\Lambda''(0)}{3\Lambda''(0)^3}. \tag{2.9}$$

Proof. By the BBF formula and our smoothness assumptions on Λ,

$$\sigma_{\mathrm{imp}}^2(k, 0) = \frac{k^2}{2\Lambda(k)}$$

$$= k^2\left(\Lambda''(0)k^2 + \frac{1}{3}\Lambda'''(0)k^3 + \frac{1}{12}\Lambda''''(0)k^4 + O(k^5)\right)^{-1}$$

$$= \frac{1}{\Lambda''(0)} - \frac{1}{3}\frac{\Lambda'''(0)}{\Lambda''(0)^2}k + \left(\frac{\frac{1}{9}\Lambda'''(0)^2 - \frac{1}{12}\Lambda''''(0)\Lambda''(0)}{\Lambda''(0)^3}\right)k^2 + O(k^3), \qquad k \to 0.$$

This implies (2.8) and (2.9). $\qquad\qquad\square$

Proposition 2.4 combined with Theorem 2.6 allows to compute skew and curvature (and higher derivatives of the implied volatility smile, if desired) directly from the coefficients of a general stochastic volatility model. Related formulae for "general" (even non-Markovian) models also appear in the work of Durrleman (theorem 3.1.1. in Durrleman, 2004; see also Durrleman, 2010). While not written in the setting of general Markovian diffusion models, and hence not in terms of the energy function Λ, they inevitably give the same results if applied to given parametric stochastic volatility models (see section 3.1 in Durrleman, 2004). However, Durrleman's work comes with some (seemingly) uncheckable assumptions, the drawbacks of which are discussed in section 2.6 of Durrleman (2004).

3 | PROOFS OF THE MAIN RESULTS

Proof of Theorem 2.3. As the density of S_t satisfies $q = \partial_{KK}C$, we have, by Dupire's formula (2.1),

$$C(K, t) = \int_0^t \partial_s C(K, s)ds = \int_0^t \frac{1}{2}q(K, s)K^2\sigma_{\mathrm{loc}}^2(K, s)ds.$$

Then, for $K_t = e^{k_t}$ with $k_t \downarrow 0$ as stated, we apply Assumption 2.2 as follows:

$$C(K_t, t) = \int_0^t \frac{1}{2} q(K_t, s) K_t^2 \sigma_{\text{loc}}^2(K_t, s) ds$$

$$\sim \frac{\sigma_0^2}{2} \int_0^t q(K_t, s) ds, \qquad t \downarrow 0. \tag{3.1}$$

And then, using local uniformity of our density expansion (2.2), as $t \downarrow 0$,

$$C(K_t, t) \sim \frac{\sigma_0^2 \gamma(0)}{2} \int_0^t e^{-\frac{\Lambda(k_t)}{s}} s^{-1/2} dt \tag{3.2}$$

$$= \frac{\sigma_0^2 \gamma(0)}{2} t \int_0^1 e^{-\frac{\Lambda(k_t)}{xt}} (xt)^{-1/2} dx$$

$$= \frac{\sigma_0^2 \gamma(0)}{2} t^{1/2} \int_0^1 e^{-\frac{\Lambda(k_t)}{xt}} x^{-1/2} dx. \tag{3.3}$$

Because Λ is smooth at zero, and using the fact that $\Lambda(0) = \Lambda'(0) = 0$, we have

$$\frac{\Lambda(k_t)}{t} \sim \frac{1}{2} \Lambda''(0) \frac{k_t^2}{t} \to \infty \qquad \text{as } t \downarrow 0.$$

For small t, the integrand in (3.3) is thus concentrated near $x = 1$, and by the Laplace method (theorem 3.7.1 in Olver, 1974)

$$\int_0^1 e^{-\frac{\Lambda(k_t)}{xt}} x^{-1/2} dx \sim \frac{t}{\Lambda(k_t)} \exp\left(-\frac{\Lambda(k_t)}{t}\right), \qquad t \downarrow 0. \tag{3.4}$$

Therefore,

$$C(K_t, t) \sim \frac{\sigma_0^2 \gamma(0)}{2} \frac{t^{3/2}}{\Lambda(k_t)} \exp\left(-\frac{\Lambda(k_t)}{t}\right) \sim v_0^2 \gamma(0) \frac{t^{3/2}}{k_t^2} \exp\left(-\frac{\Lambda(k_t)}{t}\right), \qquad t \downarrow 0, \tag{3.5}$$

which implies (recall the notation c resp. C from (1.1))

$$-\log c(k_t, t) = \frac{\Lambda(k_t)}{t} - \log \frac{t^{3/2}}{k_t^2} + O(1)$$

$$= \frac{1}{t} \left(\frac{1}{2} \Lambda''(0) k_t^2 + \frac{1}{6} \Lambda'''(0) k_t^3 + O\left(k_t^4\right)\right) + O\left(\log \frac{k_t^2}{t^{3/2}}\right), \qquad t \downarrow 0. \tag{3.6}$$

To prove (i) and (ii), we thus need to argue that k_t^2/t dominates $\log(k_t^2 t^{-3/2})$ if $\beta \in (0, \frac{1}{2})$, and that k_t^3/t dominates $\log(k_t^2 t^{-3/2})$ if $\beta \in (0, \frac{1}{3})$. For $m \in \{2, 3\}$, we calculate

$$\frac{k_t^m/t}{\left|\log\left(k_t^2 t^{-3/2}\right)\right|} = \frac{t^{m\beta-1} \ell(t)^m}{\left|\log(t^{2\beta-3/2} \ell(t)^2)\right|}$$

$$= \frac{t^{m\beta-1} \ell(t)^m}{\left|\left(2\beta - \frac{3}{2}\right)\log t + 2\log \ell(t)\right|}.$$

From proposition 1.3.6 (i) in Bingham, Goldie, and Teugels (1987), we know that $\log \ell(t) = o(\log t)$, and so

$$\frac{k_t^m / t}{|\log(k_t^2 t^{-3/2})|} \sim \frac{t^{m\beta - 1}\ell(t)^m}{|(2\beta - \frac{3}{2})\log t|}, \qquad t \downarrow 0.$$

This tends to infinity for $m = 2$ and $\beta \in (0, \frac{1}{2})$, and for $m = 3$ and $\beta \in (0, \frac{1}{3})$, as desired. $\qquad \square$

Inspecting the preceding proof, it is easy to see that we can expand $\log c(k_t, t)$ further.

Proof of Theorem 2.5. Taking logs in (3.5) yields

$$\log c(k_t, t) = -\frac{\Lambda(k_t)}{t} + \log\frac{t^{3/2}}{k_t^2} + \log(\gamma(0)v_0^2) + o(1), \qquad t \downarrow 0.$$

Then (2.7) follows by Taylor expanding Λ. Note that $k_t^m / t = o(1)$ for $m \geq \lfloor 1/\beta \rfloor + 1$. $\qquad \square$

4 | IMPLIED VOLATILITY

Corollary 4.1. *Under the assumptions of Theorem 2.3, let $k_t = t^\beta \ell(t)$ with $\beta \in (0, \frac{1}{3})$ and $\ell > 0$ slowly varying. Then the implied volatility has the MOTM expansion*

$$\sigma_{\text{imp}}(k_t, t) = \sigma_0 - \frac{1}{6}\sigma_0^3 \Lambda'''(0) k_t (1 + o(1)), \qquad t \downarrow 0. \tag{4.1}$$

Proof. We use our main result (Theorem 2.3) in conjunction with a transfer result of Gao and Lee (2014). As the call price tends to zero, we are in case "(−)" of Gao and Lee (2014) (defined on p. 354 of that paper). The notation L, V of Gao and Lee (2014) means $L = -\log c(k_t, t)$ resp. $V = t^{1/2}\sigma_{\text{imp}}(k_t, t)$, the dimensionless implied volatility. Then corollary 7.2 of Gao and Lee (2014) implies that

$$V = \frac{k_t}{\sqrt{2L}}\left(1 + O(t^{1-2\beta-\varepsilon})\right) + O(t^{5/2-4\beta-\varepsilon}), \qquad t \downarrow 0. \tag{4.2}$$

Here, $\varepsilon > 0$ denotes an arbitrarily small constant that serves to eat up slowly varying functions in O-estimates (see proposition 1.3.6 (v) in Bingham et al., 1987). By part (ii) of Theorem 2.3, we have

$$2L = \frac{1}{\sigma_0^2}\frac{k_t^2}{t} + \frac{\Lambda'''(0)}{3}\frac{k_t^3}{t}(1 + o(1)), \qquad t \downarrow 0.$$

Inserting this into (4.2) gives

$$\sigma_{\text{imp}}(k_t, t) = t^{-1/2}k_t\left(\sigma_0\frac{t^{1/2}}{k_t} - \frac{\sigma_0^3\Lambda'''(0)}{6}t^{1/2}(1 + o(1))\right) + O(t^{2-4\beta-\varepsilon}), \qquad t \downarrow 0,$$

which yields (4.1). $\qquad \square$

We have no doubt that Corollary 4.1 is true for the whole MOTM regime, i.e., for all $\beta \in (0, \frac{1}{2})$, under very mild assumptions (Assumption A.1 in the Appendix). For any *fixed* $\beta \in (0, \frac{1}{2})$, one can compute the implied volatility expansion using the results of Gao and Lee (2014). However, for β close

to $\frac{1}{2}$, more and more terms are needed for the intermediate computations, and there does not seem to be a simple pattern that would allow for a general proof. The details are discussed in the Appendix, where we push the range of β for which (4.1) is proven rigorously to $0 < \beta < \frac{3}{7} \approx 0.429$. Note that the expansion in Theorem 2.5 becomes *finer* (i.e., contains more explicit terms) if β is close to zero. Suppose, on the other hand, that β is very close to $\frac{1}{2}$: Then the summands $m > \lfloor 1/\beta \rfloor = 2$ in (2.7), which are related to ATM derivatives of implied variance by Theorem 2.6 (see also paragraph (iii) in the Introduction), disappear into the $o(1)$-term of (2.7).

Corollary 4.1 has some interesting consequences. Under the sheer assumption that implied volatility has a first-order Taylor expansion for small maturity and small log-strike of the form

$$\sigma_{\mathrm{imp}}(k, t) = \sigma_0 + \partial_k \sigma_{\mathrm{imp}}(0, 0)\, k + o(k) + O(t), \qquad t \downarrow 0,\ k = o(1); \tag{4.3}$$

then of course in the MOTM regime, we have $t \ll k_t$, and so the k-term dominates the $O(t)$-term, which in turn identifies the implied variance skew as

$$S = \lim_{t \downarrow 0} \frac{2\sigma_0}{k_t} \left(\sigma_{\mathrm{imp}}(k_t, t) - \sigma_0 \right). \tag{4.4}$$

On the other hand, Corollary 4.1 now implies that the right-hand side of (4.4) equals $-\frac{1}{3}\sigma_0^4 \Lambda'''(0)$. We have thus arrived at an alternative proof of the skew representation (2.8) in terms of the energy function, without using the BBF formula. The curvature and higher order derivatives of the ATM smile can be dealt with similarly, if desired.

5 | EXAMPLES

5.1 | Generic local volatility models

Clearly, Assumption 2.2 is satisfied for any local volatility model, assuming continuity of the local volatility function. We now discuss Assumption 2.1 and show how to compute our MOTM expansions. First consider the time-homogeneous local volatility model

$$dS_t = \sigma(S_t) S_t dW_t, \qquad S_0 = 1, \tag{5.1}$$

where the deterministic function σ is C^2 on $(0, \infty)$. An expansion of the pdf $q(\cdot, t)$ of S_t has been worked out in Gatheral, Hsu, Laurence, Ouyang, and Wang (2012). They assume growth conditions on σ and its derivatives, which can be alleviated by the principle of not feeling the boundary (appendix A of Gatheral et al., 2012). Proposition 2.1 of Gatheral et al. (2012) states that

$$q(e^k, t) \sim \frac{e^{-k} u_0(0, k)}{\sqrt{2\pi t}} \exp\left(-\frac{\Lambda(k)}{t} \right), \qquad t \downarrow 0, \tag{5.2}$$

uniformly in k, where the energy function is given by (cf. Varadhan, 1967)

$$\Lambda(k) = \frac{1}{2} \left(\int_0^k \frac{dx}{\sigma(e^x)} \right)^2,$$

and

$$u_0(0, k) = \sigma(1)^{1/2} \sigma(e^k)^{-3/2} e^{-k/2}. \tag{5.3}$$

(Recall that we normalize spot to $S_0 = 1$ throughout.) This shows that Assumption 2.1 is satisfied, with

$$\gamma(0) = \frac{1}{\sqrt{2\pi}\sigma(1)}. \tag{5.4}$$

To evaluate the expansions from Theorem 2.3, we compute the derivatives of Λ

$$\Lambda'(k) = \frac{1}{\sigma(e^k)} \int_0^k \frac{dx}{\sigma(e^x)}, \qquad \Lambda''(k) = \frac{1}{\sigma(e^k)^2} - \frac{e^k \sigma'(e^k)}{\sigma(e^k)^2} \int_0^k \frac{dx}{\sigma(e^x)},$$

$$\Lambda'''(k) = -\frac{3e^k \sigma'(e^k)}{\sigma(e^k)^3} + \left(\frac{2e^{2k}\sigma'(e^k)^2}{\sigma(e^k)^3} - \frac{\sigma(e^k)''}{\sigma(e^k)^2} \right) \int_0^k \frac{dx}{\sigma(e^x)},$$

which yield

$$\Lambda''(0) = \frac{1}{\sigma(1)^2} = \frac{1}{\sigma(S_0)^2},$$

$$\Lambda'''(0) = -\frac{3\sigma'(1)}{\sigma(1)^3} = -\frac{3\sigma'(S_0)}{\sigma(S_0)^3}. \tag{5.5}$$

Alternatively, these expressions can be obtained from Proposition 2.4. As the assumptions of Theorem 2.3 are satisfied, we obtain the following MOTM call price estimates, where $k_t = \theta t^\beta$ and $\theta > 0$

$$c(k_t, t) = \exp\left(-\frac{\theta^2}{2\sigma(1)^2 t^{1-2\beta}} (1 + o(1)) \right), \qquad \beta \in (0, \tfrac{1}{2}), \ t \downarrow 0,$$

$$c(k_t, t) = \exp\left(-\frac{\theta^2}{2\sigma(1)^2 t^{1-2\beta}} - \frac{\sigma'(1)}{2\sigma(1)^3} \frac{\theta^3}{t^{1-3\beta}} (1 + o(1)) \right), \qquad \beta \in (0, \tfrac{1}{3}), \ t \downarrow 0.$$

Recall from Theorem 2.6 that we denote by $\mathcal{S}$ the (limiting small-time ATM) implied *variance* skew, and so the implied *volatility* skew is given by $\mathcal{S}/2\sigma_0$, which equals $\mathcal{S}/2\sigma(1) = \mathcal{S}/2\sigma(S_0)$ in model (5.1). From (2.8) and (5.5), we find that the *local* skew $\sigma'(1) = \sigma'(S_0)$ equals twice the implied volatility skew,

$$\sigma'(S_0) = 2 \times \frac{\mathcal{S}}{2\sigma(S_0)},$$

as observed in remark 5.2 of Henry-Labordère (2009). Generic time-inhomogeneous local volatility models

$$dS_t = \sigma(S_t, t) S_t dW_t$$

can be treated very similarly, using the heat kernel expansion in section 3 of Gatheral et al. (2012), itself taken from Yosida (1953).

5.2 | Generic stochastic volatility models

We now discuss the results of Section 2 in generic stochastic volatility models. Rigorous conditions under which stochastic volatility models satisfy Assumption 2.1 can be found in Deuschel et al. (2014a) and Osajima (2015). The function Λ is given by the Riemannian metric associated to the model: $2\Lambda(k)$ is the squared geodesic distance from $(S_0 = 1, \sigma_0)$ to $\{(K, \sigma) : \sigma > 0\}$ with $K = e^k$. Theorem 2.2

in Berestycki, Busca, and Florent (2004) gives conditions under which Assumption 2.2, concerning convergence of local volatility, is true.

Now we describe how the expressions appearing in the expansions from Theorem 2.3 can be computed explicitly in a generic two-factor stochastic volatility model

$$dS_t = S_t\sqrt{V_t}dW_t, \qquad S_0 = 1,$$

$$dV_t = (\dots)dt + \eta\sqrt{V_t}v(V_t)dZ_t, \qquad V_0 = v_0 > 0, \tag{5.6}$$

where $v : \mathbb{R} \to \mathbb{R}$ and $d\langle W, Z\rangle_t = \rho\, dt$. The Heston model ($v(v) \equiv 1$) and the 3/2-model ($v(v) = v$; see Lewis, 2000) are special cases. The infinitesimal generator L of the stochastic process (S, V), neglecting first-order terms, can be written as

$$Lf \approx \frac{1}{2}\operatorname{Tr}\left(\begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} D^2 f\right), \qquad f \in C^2(\mathbb{R}^2),$$

where $D^2 f$ denotes the Hessian matrix of f, and the coefficient matrix $g = (g_{ij})$ is given by

$$g = \begin{pmatrix} v & \rho\eta v v(v) \\ \rho\eta v v(v) & \eta^2 v v(v)^2 \end{pmatrix}.$$

We define the constants $b_1 = g_{11}|_{v=v_0} = v_0$ and $b_2 = \frac{3}{4}\sum_{i=1}^2 g_{1i}\partial_i g_{11}|_{v=v_0} = \frac{3}{4}\rho\eta v_0 v(v_0)$. If we assume that the coefficients in (5.6) are nice enough to justify application of the marginal density expansion obtained in Deuschel et al. (2014a) or part (2) of theorem 1 in Osajima (2015), then we get the desired small-time density expansion (2.2). Moreover, thanks to Proposition 2.4,

$$\Lambda(k) = \frac{1}{2b_1}k^2 - \frac{b_2}{3b_1^3}k^3 + O(k^4)$$

as $k \to 0$. Therefore, the quantities $\Lambda''(0) = v_0^{-1} = \sigma_0^{-2}$ and $\Lambda'''(0) = -\frac{3}{2}\rho\eta v(v_0)/v_0^2$ can easily be computed, as well as the small-time ATM implied variance skew

$$S = -\frac{v_0^2}{3}\Lambda'''(0) = \frac{\rho\eta}{2}v(v_0).$$

Thus, all quantities appearing in our expansions (Theorem 2.3, Corollary 4.1) have very simple expressions in terms of the model parameters.

5.3 | The Heston model

This section contains an application of the results of Sections 2 and 4 to the familiar case of the Heston model, where many explicit "affine" computations are possible. At the beginning of Section 5.2, we recalled some general results implying our Assumptions 2.1 and 2.2. The Heston model is not covered by these results, but satisfies Assumptions 2.1 and 2.2 nevertheless, and thus Theorems 2.3 and 2.5 are applicable to Heston. We will explain how both assumptions can be verified rigorously by a dedicated analysis; full details would involve rather dull repetition of arguments that are found in the literature in a very similar form, and are therefore omitted. The model dynamics are

$$dS_t = S_t\sqrt{V_t}dW_t, \qquad S_0 = 1,$$

$$dV_t = -\kappa(V - \bar{v})\, dt + \eta\sqrt{V_t}dZ_t, \qquad V_0 = v_0 > 0,$$

where $\bar{v}, \kappa, \eta > 0$, and $d\langle W, Z\rangle_t = \rho\, dt$ with $\rho \in (-1, 1)$. According to Forde and Jacquier (2009), the first-order OTM (large deviations) behavior of the call prices is

$$t \log c_{\text{He}}(k, t) \sim -\Lambda_{\text{He}}(k), \qquad k > 0 \text{ fixed, } t \downarrow 0, \tag{5.7}$$

where Λ_{He} is the (not explicitly available) Legendre transform of

$$\Gamma(p) = \frac{v_0 p}{\eta(\bar{\rho} \cot(\eta\bar{\rho}p/2) - \rho)} = \frac{v_0 p}{\eta\left(\bar{\rho}\left(\frac{1}{\eta\bar{\rho}p/2} + O(p)\right) - \rho\right)} = \frac{v_0 p}{\frac{1}{p/2} - \eta\rho + O(p)}$$

$$= \frac{v_0 p^2 / 2}{1 - p\eta\rho/2 + O(p^2)} = \frac{v_0 p^2}{2}\left(1 + p\eta\rho/2 + O(p^2)\right), \qquad p \to 0. \tag{5.8}$$

(We use the standard notation $\bar{\rho}^2 = 1 - \rho^2$.) This expansion implies

$$\Gamma''(0) = v_0 = \sigma_0^2. \tag{5.9}$$

The locally uniform density asymptotics (2.2) hold, as seen from an easy modification of the arguments in Forde et al. (2012). There, the Fourier representation of the call price was analyzed by the saddle point method to obtain a refinement of (5.7). Proceeding analogously for the Fourier representation of the pdf of S_t, we get the density approximation

$$q_{\text{He}}(e^k, t) = \frac{e^{-k}}{2\pi t} \int_{-\infty - ip^*(k)}^{\infty - ip^*(k)} \text{Re}\left(e^{iku/t} \phi_t(-u/t)\right) du$$

$$= \exp\left(-\frac{\Lambda_{\text{He}}(k)}{t}\right) \frac{U(p^*(k))}{\sqrt{2\pi\Gamma''(k)}} t^{-1/2} \left(1 + o(1)\right), \qquad t \downarrow 0,$$

locally uniformly in k, where ϕ_t is the characteristic function of $X_t = \log S_t$, and p^* and U are defined on p. 693 of Forde et al. (2012). (Note that Forde et al., 2012, use the notation Λ, Λ^* instead of our $\Gamma, \Lambda_{\text{He}}$.) From (5.9) and the fact that $U(p^*(0)) = U(0) = 1$, we see that the factor $\gamma(k)$ from (2.2) converges to

$$\gamma(0) = \frac{1}{\sqrt{2\pi}\sigma_0}, \tag{5.10}$$

as $k \to 0$.

To verify Assumption 2.2 (convergence of local volatility), the Dupire formula (2.1) can be subjected to an analysis similar to De Marco, Friz, and Gerhold (2013) and Friz and Gerhold (2015). More precisely, $\partial_{KK}C(K, t)$ in the numerator of (2.1) is the pdf of S_t, the analysis of which we have just described. Virtually the same saddle point approach can be applied to the numerator $\partial_t C(K, t)$, yielding convergence of the quotient to σ_0^2.

We now calculate our MOTM asymptotic expansions for the Heston model. The Legendre transform Λ_{He} is given by $\Lambda_{\text{He}}(k) = \sup_x\{kx - \Gamma(x)\}$ with maximizer $x^* = x^*(k)$. From general facts on Legendre transforms,

$$\Lambda_{\text{He}}''(k) = \frac{1}{\Gamma''(x^*(k))}.$$

We have $x^*(0) = 0$, which implies

$$\Lambda_{\text{He}}''(0) = \frac{1}{\Gamma''(0)} = \frac{1}{v_0}.$$

From Theorem 2.3, with $k_t = \theta t^\beta$ and $\theta > 0$, we then obtain the MOTM call price estimate

$$c_{He}(k_t, t) = \exp\left(-\frac{\theta^2}{2v_0 t^{1-2\beta}}(1 + o(1))\right), \qquad t \downarrow 0. \tag{5.11}$$

As for the second-order expansion, from the expansion (5.8) of Γ, we clearly see that

$$\Gamma'''(0) = \frac{3}{2}v_0\eta\rho.$$
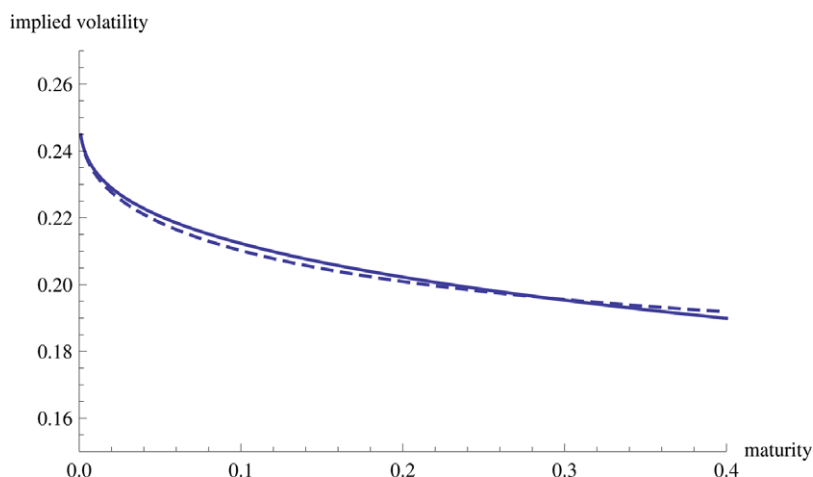
On the other hand, a general Legendre computation gives

$$\Lambda'''_{He}(k) = -\left(\frac{1}{\Gamma''(x^*(k))}\right)^2 \Gamma'''(x^*(k))\,(x^*)'(k) = -(\Lambda''_{He}(k))^3\,\Gamma'''(x^*(k)).$$

Therefore,

$$\Lambda'''_{He}(0) = -\frac{3}{2}\frac{\eta\rho}{v_0^2},$$

in accordance with the expression for generic two-factor models, found in Section 5.2. For $\beta \in (0, \frac{1}{3})$, Theorem 2.3 (ii) thus implies the second-order expansion

$$c_{He}(k_t, t) = \exp\left(-\frac{\theta^2}{2v_0 t^{1-2\beta}} + \frac{\eta\rho}{4v_0^2}\frac{\theta^3}{t^{1-3\beta}}(1 + o(1))\right), \qquad t \downarrow 0. \tag{5.12}$$

FIGURE 5.1 Illustration of our implied volatility expansion for the Heston model, with $\ell \equiv \theta = 0.4$ and $\beta = 0.3$. Thus, log-strike equals $k_t = 0.4\,t^{0.3}$. The model parameters are $\bar{v} = 0.0707$, $\kappa = 0.6067$, $\eta = 0.2928$, $\rho = -0.7571$, $v_0 = 0.0654$ (i.e., $\sigma_0 = 0.2557$), and $S_0 = 1$. The horizontal axis is time. The dashed curve is the exact MOTM implied volatility $\sigma_{imp}(k_t, t)$. The solid curve is the approximation $\sigma_0 + \frac{\eta\rho}{4\sigma_0}k_t$ on the right-hand side of (5.15) [Color figure can be viewed at wileyonlinelibrary.com]

By Theorem 2.5 and (5.10), we obtain the following refined call price expansions, as $t \downarrow 0$:

$$\log c_{\text{He}}(k_t, t) = -\frac{1}{2\sigma_0^2}\frac{k_t^2}{t} + \left(\frac{3}{2} - 2\beta\right)\log t + \log\frac{\sigma_0^3}{\sqrt{2\pi}} + o(1), \qquad \beta \in \left(\frac{1}{3}, \frac{1}{2}\right), \tag{5.13}$$

$$\log c_{\text{He}}(k_t, t) = -\frac{1}{2\sigma_0^2}\frac{k_t^2}{t} + \frac{\eta\rho}{4v_0^2}\frac{k_t^3}{t} + \left(\frac{3}{2} - 2\beta\right)\log t + \log\frac{\sigma_0^3}{\sqrt{2\pi}} + o(1), \qquad \beta \in \left(\frac{1}{4}, \frac{1}{3}\right). \tag{5.14}$$

From the relation (2.8) between implied variance skew and $\Lambda'''(0)$, we get the explicit expression $S_{\text{He}} = \eta\rho/2$ for the skew. This agrees with Gatheral (2006, p. 35). The implied volatility expansion (4.4) becomes

$$\sigma_{\text{imp}}(k_t, t) = \sigma_0 + \frac{\eta\rho}{4\sigma_0}k_t\left(1 + o(1)\right), \qquad t \downarrow 0. \tag{5.15}$$

Figure 5.1 shows a good fit of this approximation, even for maturities that are not very small.

6 | OTHER APPROACHES AT MOTM ASYMPTOTICS

In a recent paper, Gao and Wang, 2016 study small noise sample-path MDPs for SDE solutions, and specialize to the small-time regime (corollary 4.1.2 in Gao and Wang, 2016). Their asymptotic regime is in fact slightly more general than (2.4), allowing for (in our notation) any k_t satisfying $\sqrt{t} \ll k_t \ll 1$ as $t \downarrow 0$. (In the financial context, this offers no useful additional flexibility; it allows, e.g., switching between two regimes $k_t = t^{\beta_1}$ and $k_t = t^{\beta_2}$ infinitely often as $t \downarrow 0$.) However, Gao and Wang (2016) impose the assumptions of linearly bounded and locally Lipschitz coefficients. These are the typical assumptions for small-noise LDPs in the literature, but they are rarely satisfied in stochastic volatility models. In particular, their results are not directly applicable to the Heston model. The paper by Cai and Wang (2015) is also worth mentioning here: It presents moderate deviations for the CIR process (the Heston variance process) and a generalization, where the exponent in the dynamics is not necessarily 1/2. The paper uses estimates tied to the (generalized) CIR stochastic differential equation.

In this section, we discuss a different approach at small-time moderate deviations. While yielding only first-order results, its conditions are usually easy to check for models with explicit characteristic function. Assumptions 2.1 and 2.2 are not in force here. Recall that in the classical setting of sequences of i.i.d. random variables, a moderate deviation analogue of Cramér's theorem can be deduced by applying the Gärtner–Ellis theorem to an appropriately rescaled sequence (see Dembo & Zeitouni 1998, section 3.7). The MD short-time behavior of diffusions can be subjected to a similar analysis. Consider the log-price $X_t = \log S_t$ with $X_0 = 0$ and mgf (moment-generating function)

$$M(p, t) := \mathbb{E}\left[e^{pX_t}\right]. \tag{6.1}$$

Assumption 6.1. *For all $\beta \in (0, \frac{1}{2})$, the rescaled mgf satisfies*

$$\lim_{t\downarrow 0} t^{1-2\beta}\log M(t^{\beta-1}p, t) = \frac{1}{2}\sigma_0^2 p^2, \qquad p \in \mathbb{R}. \tag{6.2}$$

We expect that this assumption holds for diffusion models in considerable generality. It is easy to check that (6.2) holds for the Heston model, either by its explicit characteristic function, or, more elegantly, from the associated Riccati equations; see Pinter (2017) for details. Thus, the results of the present section provide an alternative proof of the first order MOTM behavior (5.11) of Heston call prices.

Heuristically, Assumption 6.1 can be derived from the density asymptotics in Assumption 2.1, which in turn hold in quite general diffusion settings (see Deuschel et al., 2014a, 2014b).

$$M(t^{\beta-1}p, t) = \int e^{t^{\beta-1}px} q(x,t) dx \approx \int \exp\left(t^{\beta-1}px - \frac{\Lambda(x)}{t}\right) dx \tag{6.3}$$

$$\approx \int \exp\left(t^{\beta-1}px - \frac{\Lambda''(0)x^2}{2t}\right) dx \tag{6.4}$$

$$= \exp\left(t^{\beta-1}px - \frac{x^2}{2\sigma_0^2 t}\Bigg|_{x=\sigma_0^2 pt^\beta} (1+o(1))\right) \tag{6.5}$$

$$= \exp\left(\tfrac{1}{2}\sigma_0^2 p^2 t^{2\beta-1}(1+o(1))\right), \qquad t \downarrow 0. \tag{6.6}$$

In (6.3), we ignored that the density expansion (2.2) might not be valid globally in space; this might be made rigorous by estimating $q(x,t)$ by a Freidlin–Wentzell LD argument for x sufficiently large. As for (6.4), we can expect concentration near $x \approx 0$, because $\Lambda(x)$ increases with $|x|$. Finally, (6.5), and thus (6.6), follows from a (rigorous) application of the Laplace method. If (6.6) is correct, then (6.2) clearly follows.

The critical moment of S_t is defined by

$$p_+(t) := \sup\{p \geq 0 : M(p,t) < \infty\}.$$

It is obvious that

$$\lim_{t\downarrow 0} t^{1-\beta} p_+(t) = \infty \tag{6.7}$$

is necessary for (6.2); i.e., $p_+(t)$ must grow faster than $t^{\beta-1}$ as $t \downarrow 0$. In the Heston model, e.g., the critical moment is of order $p_+(t) \sim (const)/t \gg t^{\beta-1}$ for small t, as follows from inverting (6.2) in Keller-Ressel (2011). On the other hand, we do not expect our results to be of much use in the presence of jumps. Indeed, suppose that (6.1) is the mgf of an exponential Lévy model. Then, $p_+(t) \equiv p_+$ does not depend on t, and is finite for most models used in practice. Therefore, (6.7) cannot hold, and so Assumption 6.1 is not satisfied. The Merton jump diffusion model is one of the few Lévy models of interest that have $p_+ = \infty$, but it is easy to check that it does not satisfy (6.2), either.

After this discussion of Assumption 6.1, we now give an asymptotic estimate for the distribution function of X_t (put differently, MOTM *digital call* prices) in Theorem 6.2. Then, we translate this result to MOTM *call* prices in Theorem 6.3. If desired, higher order terms in (6.2) will give refined asymptotics in Theorem 6.2, using Gulisashvili and Teichmann's (2015) recent refinement of the Gärtner–Ellis theorem. (Further work will be required to translate the resulting expansions into call price asymptotics.) For other asymptotic results on option prices using the Gärtner–Ellis theorem, see, e.g, Forde and Jacquier (2009, 2011).

Theorem 6.2. *Under Assumption 6.1 (and without any further assumptions on our model), for $k_t = \theta t^\beta$ with $\beta \in (0, \frac{1}{2})$ and $\theta > 0$, we have a first-order MD estimate for the cdf of X_t:*

$$\mathbb{P}[X_t \geq k_t] = \exp\left(-\frac{1}{2\sigma_0^2}\frac{k_t^2}{t}(1+o(1))\right), \qquad t \downarrow 0. \tag{6.8}$$

Proof. Define

$$Z_t := t^{-\beta} X_t, \qquad \text{with mgf} \qquad M_Z(s,t) = \mathbb{E}[e^{sZ_t}],$$

and

$$a_t := t^{1-2\beta} = o(1), \qquad t \downarrow 0.$$

Then, (6.2) is equivalent to

$$\Gamma_Z(p) := \lim_{t \downarrow 0} a_t \log M_Z(p/a_t, t) = \tfrac{1}{2}\sigma_0^2 p^2, \qquad p \in \mathbb{R}.$$

As Γ_Z is finite and differentiable on $\mathbb{R}$, the Gärtner–Ellis theorem (theorem 2.3.6 in Dembo & Zeitouni 1998) implies that $(Z_t)_{t \geq 0}$ satisfies an LDP as $t \downarrow 0$, with rate a_t and good rate function Λ_Z, the Legendre transform of Γ_Z. Trivially, Λ_Z is qquadratic, too:

$$\Lambda_Z(x) = \sup_{p \in \mathbb{R}}(px - \Gamma_Z(p))$$

$$= \sup_{p \in \mathbb{R}}\left(px - \tfrac{1}{2}\sigma_0^2 p^2\right) = \frac{x^2}{2\sigma_0^2}, \qquad x \in \mathbb{R}.$$

Now fix $\theta > 0$. Applying the lower estimate of the LDP to (θ, ∞) yields

$$\liminf_{t \downarrow 0} a_t \log \mathbb{P}[Z_t \geq \theta] \geq \liminf_{t \downarrow 0} a_t \log \mathbb{P}[Z_t > \theta]$$

$$\geq -\Lambda_Z(\theta) = -\frac{\theta^2}{2\sigma_0^2},$$

and applying the upper estimate to $[\theta, \infty)$ yields

$$\limsup_{t \downarrow 0} a_t \log \mathbb{P}[Z_t \geq \theta] \leq -\frac{\theta^2}{2\sigma_0^2},$$

and so

$$\lim_{t \downarrow 0} a_t \log \mathbb{P}[Z_t \geq \theta] = -\frac{\theta^2}{2\sigma_0^2}.$$

This is the same as (6.8). ◻

As in the LD/OTM regime, first-order cdf asymptotics translate readily into call price asymptotics. The proof of the following result is similar to Pham (2010, p. 30f; concerning the LD regime) and Caravenna and Corbetta (2016, theorem 1.5). In the MD/MOTM regime, one can replace the condition (1.19) of Caravenna and Corbetta (2016) by a mild condition on the moments of the model.

Theorem 6.3. *Let $S = e^X$ be a continuous positive martingale. Assume that, for all $p \geq 1$, its pth moment explodes at a positive time (infinity included).[8] By this we mean that there is a positive $t^*(p)$ such that the mgf $\mathbb{E}[\exp(pX_t)]$ is finite for all $t \in [0, t^*(p)]$. Let $v_0 = \sigma_0^2 > 0$. Then the following are equivalent*

(i) *For $k_t = \ell(t)t^\beta$, with $\beta \in (0, \frac{1}{2})$ and $\ell > 0$ slowly varying at zero, it holds that*

$$\mathbb{P}[X_t \geq k_t] = \exp\left(-\frac{1}{2v_0}\frac{k_t^2}{t}(1 + o(1))\right), \qquad t \downarrow 0.$$

(ii) *Under the assumptions of (i), we have*

$$c(k_t, t) = \exp\left(-\frac{1}{2v_0}\frac{k_t^2}{t}(1 + o(1))\right), \qquad t \downarrow 0. \tag{6.9}$$

Proof. First assume (i). Let $\varepsilon > 0$ and define $\tilde{k}_t = (1 + \varepsilon)k_t$. Then

$$c(k_t, t) \geq \mathbb{E}[(e^{X_t} - e^{k_t})^+ \mathbf{1}_{\{X_t \geq \tilde{k}_t\}}]$$

$$\geq (e^{\tilde{k}_t} - e^{k_t})^+ \mathbb{P}[X_t \geq \tilde{k}_t]. \tag{6.10}$$

The first factor is

$$\left(e^{\tilde{k}_t} - e^{k_t}\right)^+ = \left(\tilde{k}_t - k_t + O\left(k_t^2\right)\right)^+ = \varepsilon k_t + O\left(k_t^2\right), \qquad t \downarrow 0.$$

For the second factor in (6.10), we apply (i) with $\tilde{k}_t$.

$$\lim_{t \downarrow 0} \frac{t}{\tilde{k}_t^2} \log \mathbb{P}[X_t \geq \tilde{k}_t] = -\frac{1}{2v_0}.$$

Therefore,

$$\liminf_{t \downarrow 0} \frac{t}{k_t^2} \log c(k_t, t) \geq \lim_{t \downarrow 0} \frac{t}{k_t^2}\left(-\frac{1}{2v_0}\frac{\tilde{k}_t^2}{t}(1 + o(1))\right)$$

$$= -\frac{(1 + \varepsilon)^2}{2v_0}.$$

Now let $\varepsilon \downarrow 0$ to get the desired lower bound for $c(k_t, t)$.

As for the upper bound, we let $p > 1$ and note that, by definition of $p \mapsto t^*(p)$, we have $\mathbb{E}[S_t^{p+1}] < \infty$ for all $t \in [0, t^*(p + 1)]$. Define $\bar{S}_t = \sup_{0 \leq u \leq t} S_u$ for $t \geq 0$. By Doob's inequality (theorem 3.8 in Karatzas & Shreve, 1991), we have

$$\mathbb{P}\left[\bar{S}_{t^*(p+1)} \geq s\right] \leq \frac{\mathbb{E}\left[S_{t^*(p+1)}^{p+1}\right]}{s^{p+1}}, \qquad s > 0.$$

Hence $\bar{S}_{t^*(p+1)}$ has a finite pth moment,

$$\mathbb{E}\left[(\bar{S}_{t^*(p+1)})^p\right] = p\int_0^\infty s^{p-1}\mathbb{P}\left[\bar{S}_{t^*(p+1)} \geq s\right]ds < \infty.$$

By the dominated convergence theorem and the continuity of S, we thus conclude

$$\lim_{t \downarrow 0} \mathbb{E}\left[S_t^p\right] = S_0^p. \tag{6.11}$$

Now let $1/p + 1/q = 1$ and apply Hölder's inequality.

$$c(k_t, t) = \mathbb{E}\left[(e^{X_t} - e^{k_t})^+ \mathbf{1}_{\{X_t \geq k_t\}}\right]$$

$$\leq \mathbb{E}[((e^{X_t} - e^{k_t})^+)^p]^{1/p} \, \mathbb{P}[X_t \geq k_t]^{1/q}$$

$$\leq \mathbb{E}\left[S_t^p\right]^{1/p} \, \mathbb{P}[X_t \geq k_t]^{1/q}.$$

By (6.11) and (i), we obtain

$$\limsup_{t\downarrow 0} \frac{t}{k_t^2} \log c(k_t, t) \leq \frac{1}{q} \limsup_{t\downarrow 0} \frac{t}{k_t^2} \log \mathbb{P}[X_t \geq k_t] = -\frac{1}{2qv_0}.$$

Now let $p \uparrow \infty$, i.e., $q \downarrow 1$. The same argument yields the lower bound of the implication (ii) $\Rightarrow$ (i). The remaining upper bound of (ii) $\Rightarrow$ (i) is shown very similarly to the lower bound of the implication (i) $\Rightarrow$ (ii). $\qquad\square$

In the light of the general MDP result by Gao and Wang (2016) quoted at the beginning of this section, it might be worth noting that Theorem 6.3 holds, with virtually the same proof, if the assumption $k_t = \ell(t)t^\beta$ is replaced by $\sqrt{t} \ll k_t \ll 1$.

ACKNOWLEDGMENTS

ENDNOTES

[1] As we focus on stochastic volatility models, which are in general incomplete, it is understood that call prices are computed w.r.t. some fixed pricing measure.

[2] There is no spot dependence of the normalized call price in the Black–Scholes model. This holds true, more generally, whenever dynamics for the log-price $X = \log(S/S_0)$ are specified without further spot dependence; this includes the Heston model and many other stochastic volatility models.

[3] The term "almost ATM" seems new, but this regime was considered by a number of authors including Caravenna and Corbetta (2016) and Muhle-Karbe and Nutz (2011).

[4] This is also true for Lévy models with a Brownian component and a finite variation jump part. For Lévy models with jump part of infinite variation, the call price still decays algebraically, but slower than $O(t)$. See theorem 1 and proposition 2 in Mijatović and Tankov (2016).

[5] We write CLT for central limit theorem and LDP for large deviation principle. For readers unfamiliar with moderate deviations, we recall some of the basics toward the end of the Introduction.

[6] The situation is very different with jumps; see Friz, Gerhold, and Yor (2014).

[7] Recall that the MD rate function for a centered i.i.d. sequence $(X_n)_{n\geq 1}$ is given by $\theta \mapsto \theta^2/(2\mathrm{Var}(X_1))$. This is the "moderate" version of Cramér's theorem (theorem 3.7.1 in Dembo and Zeitouni, 1998; see also the Introduction of the present paper).

[8] See Keller-Ressel (2011) for a discussion of moment explosion in affine stochastic volatility models.

REFERENCES

Bentata, A., & Cont, R. (2012). Short-time asymptotics for marginal distributions of semimartingales. Preprint, available at https://arxiv.org/abs/1202.1302

Berestycki, H., Busca, J., & Florent, I. (2002). Asymptotics and calibration of local volatility models. *Quantitative Finance*, *2*(1), 61–69.

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.

Bingham, N. H., Goldie, C. M., & Teugels, J. L. (1987). *Regular variation, Encyclopedia of mathematics and its applications* (Vol. 27). Cambridge, UK: Cambridge University Press.

Brunick, G., & Shreve, S. (2013). Mimicking an Itô process by a solution of a stochastic differential equation. *Annals of Applied Probability*, *23*(4), 1584–1628.

Cai, Y., & Wang, S. (2015). Central limit theorem and moderate deviation principle for CKLS model with small random perturbation. *Statistics & Probability Letters*, *98*, 6–11.

Caravenna, F., & Corbetta, J. (2016). General smile asymptotics with bounded maturity. *SIAM Journal on Financial Mathematics*, *7*(1), 720–759. https://doi.org/10.1137/15M1031102

Carr, P., & Wu, L. (2003). What type of process underlies options? A simple robust test. *Journal of Finance*, *58*(6), 2581–2610.

De Marco, S., Friz, P., & Gerhold, S. (2013). Rational shapes of local volatility. *Risk*, *2*, 82–87.

Dembo, A., & Zeitouni, O. (1998). *Large deviations techniques and applications, Stochastic Modelling and Applied Probability* (Vol. 38, 2nd ed.). New York: Springer-Verlag.

Deuschel, J. D., Friz, P. K., Jacquier, A., & Violante, S. (2014a). Marginal density expansions for diffusions and stochastic volatility I: Theoretical foundations. *Communications on Pure and Applied Mathematics*, *67*(1), 40–82.

Deuschel, J. D., Friz, P. K., Jacquier, A., & Violante, S. (2014b). Marginal density expansions for diffusions and stochastic volatility II: Applications. *Communications on Pure and Applied Mathematics*, *67*(2), 321–350.

Durrleman, V. (2004). *From implied to spot volatilities* (PhD thesis). Princeton, NJ: Princeton University.

Durrleman, V. (2010). From implied to spot volatilities. *Finance and Stochastics*, *14*(2), 157–177.

Forde, M., & Jacquier, A. (2009). Small-time asymptotics for implied volatility under the Heston model. *International Journal of Theoretical and Applied Finance*, *12*(6), 861–876.

Forde, M., & Jacquier, A. (2011). Small-time asymptotics for an uncorrelated local-stochastic volatility model. *Applied Mathematical Finance*, *18*(6), 517–535.

Forde, M., Jacquier, A., & Lee, R. (2012). The small-time smile and term structure of implied volatility under the Heston model. *SIAM Journal on Financial Mathematics*, *3*(1), 690–708.

Forde, M., & Zhang, H. (2017). Asymptotics for rough stochastic volatility models. *SIAM Journal on Financial Mathematics*, *8*(1), 114–145. https://doi.org/10.1137/15M1009330

Friz, P., & Gerhold, S. (2015). Extrapolation analytics for Dupire's local volatility. In *Large deviations and asymptotic methods in finance. Springer Proceedings in Mathematics* (Vol. 110, pp. 273–286). Cham, Switzerland: Springer.

Friz, P. K., Gerhold, S., & Yor, M. (2014). How to make Dupire's local volatility work with jumps. *Quantitative Finance*, *14*(8), 1327–1331.

Fu, K.-A., & Shen, X. (2017). Moderate deviations for sums of dependent claims in a size-dependent renewal risk model. *Communication in Statistics—Theory and Methods*, *46*(7), 3235–3243.

Gao, F., & Wang, S. (2016). Asymptotic behaviors for functionals of random dynamical systems. *Stochastic Analysis and Applications*, *34*(2), 258–277.

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

Gatheral, J. (2006). *The volatility surface, a practitioner's guide*. Hoboken, NJ: Wiley.

Gatheral, J., Hsu, E. P., Laurence, P., Ouyang, C., & Wang, T.-H. (2012). Asymptotics of implied volatility in local volatility models. *Mathematical Finance*, *22*(4), 591–620.

Guennoun, H., Jacquier, A., & Roome, P. (2014). Asymptotic behaviour of the fractional Heston model. Preprint, available at https://arxiv.org/abs/1411.7653

Guillin, A. (2003). Averaging principle of SDE with small diffusion: Moderate deviations. *Annals of Probability*, *31*(1), 413–443.

Gulisashvili, A., & Teichmann, J. (2015). The Gärtner-Ellis theorem, homogenization, and affine processes. In *Large deviations and asymptotic methods in finance*. *Springer Proceedings in Mathematics and Statistics* (Vol. 110, pp. 287–320). Cham, Switzerland: Springer.

Henry-Labordère, P. (2009). *Analysis, geometry, and modeling in finance*. *Chapman & Hall/CRC financial mathematics series*. Boca Raton, FL: CRC Press.

Karatzas, I., & Shreve, S. E. (1991). *Brownian motion and stochastic calculus*. *Graduate texts in mathematics* (Vol. 113, 2nd ed.). New York: Springer-Verlag.

Keller-Ressel, M. (2011). Moment explosions and long-term behavior of affine stochastic volatility models. *Mathematical Finance*, *21*(1), 73–98.

Lewis, A. L. (2000). *Option valuation under stochastic volatility*. Newport Beach, CA: Finance Press.

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. https://doi.org/10.1111/mafi.12055

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

Olver, F. W. J. (1974). *Asymptotics and special functions*. New York: Academic Press.

Osajima, Y. (2015). General asymptotics of Wiener functionals and application to implied volatilities. In *Large deviations and asymptotic methods in finance*. *Springer Proceedings in Mathematics and Statistics* (Vol. 110, pp. 137–173). Cham, Switzerland: Springer.

Pagliarani, S., & Pascucci, A. (2017). The exact Taylor formula of the implied volatility. *Finance and Stochastics*, *21*, 661–718. https://doi.org/10.1007/s00780-017-0330-x

Paulot, L. (2015). Asymptotic implied volatility at the second order with application to the SABR model. In *Large deviations and asymptotic methods in finance*. *Springer Proceedings in Mathematics and Statistics* (Vol. 110, pp. 37–69). Cham, Switzerland: Springer.

Pham, H. (2010, August). *Large deviations in finance*. Lecture notes for the third SMAI European Summer School in Financial Mathematics, Paris.

Pinter, A. (2017). *Small-time asymptotics, moment explosion and the moderate deviations regime* (PhD thesis). TU Wien, Austria.

Tehranchi, M. (2016). Uniform bounds for Black-Scholes implied volatility. *SIAM Journal on Financial Mathematics*, *7*, 893–916.

Varadhan, S. R. S. (1967). Diffusion processes in a small time interval. *Communications on Pure and Applied Mathematics*, *20*, 659–685.

Yosida, K. (1953). On the fundamental solution of the parabolic equation in a Riemannian space. *Osaka Journal of Mathematics*, *5*, 65–74.

APPENDIX: IMPLIED VOLATILITY FOR $\beta \geq \frac{1}{3}$

Assumption A.1. *We refine Assumptions 2.1 and 2.2 as follows*

(i) *Additionally to Assumption 2.1, the relative error in (2.2) is $O(t)$, locally uniformly w.r.t. k.*

(ii) *Convergence of local volatility (see (2.3)) can be refined to*

$$\sigma_{\text{loc}}(K, t) = \sigma_0 + O(t), \qquad t \downarrow 0,$$

uniformly for bounded K.

(iii) *In (2.2), $\gamma(k)$ satisfies*

$$\gamma(0) = \frac{1}{\sqrt{2\pi}\sigma_0}.$$

While *proving* Assumption A.1 for stochastic volatility models would go well beyond the scope of the present paper, there are good reasons to believe that it holds in reasonable generality. It does hold for local volatility models, which satisfy (i) according to proposition 2.1 of Gatheral et al. (2012). For stochastic volatility models, the approach of Deuschel et al. (2014a) suggests that the relative error term in (2.2) has a full expansion in (integer) powers of t, which would imply (i).

Part (ii) is clear in local volatility models, just assuming smoothness of the local volatility function. In stochastic volatility models, it should be possible to refine the results of Berestycki et al. (2004) (convergence to σ_0) to a Taylor expansion.

Part (iii) of Assumption A.1 is true for the Heston model (see (5.10)) and generic local volatility models (see (5.4)); the gist of the saddle point argument we applied for Heston, and the resulting expression (5.10), are not tied to that model.

Theorem A.2. *Under Assumption A.1, the statement of Corollary 4.1 holds for $\beta \in (0, \frac{3}{7})$.*

To simplify notation in the following proof, we write $f \sim_\ell g(t)$ for two functions f and g, if $f(t)/g(t) \sim \ell(t)$ as $t \downarrow 0$ for some slowly varying function ℓ. We will use this for functions of algebraic growth order, which are then "almost" asymptotically equivalent. The index ℓ in "$\sim_\ell$" is a generic symbol and does not stand for any concrete slowly varying function.

Proof of Theorem A.2. We start by improving the call price expansion from Theorem 2.5, taking into account the asymptotic errors that were made in obtaining it. By part (ii) of Assumption A.1, the relative error in (3.1) is $O(t)$. Part (ii) of Assumption A.1 shows the same for (3.2). The relative error in (3.5) is $O(k_t)$, as seen from

$$\Lambda(k_t)^{-1} = \frac{2v_0}{k_t^2}(1 + O(k_t)), \qquad t \downarrow 0.$$

The only remaining source of error is the application of the Laplace method in (3.4). Here, it does not suffice to state the relative error; we have to do a little better than in the proof of Theorem 2.5. By the higher order extension of the Laplace method (theorem 3.8.1 in Olver, 1974), and because $\Lambda(k_t)/t \to \infty$ as $t \downarrow 0$, we have the integral expansion

$$\int_0^1 \exp\left(-\frac{\Lambda(k_t)}{tx}\right) x^{-1/2}\, dx = \int_1^\infty \exp\left(-\frac{\Lambda(k_t)}{t}x\right) x^{-3/2}\, dx$$

$$= \exp\left(-\frac{\Lambda(k_t)}{t}\right) \frac{t}{\Lambda(k_t)} \left(1 - \frac{3}{2}\frac{t}{\Lambda(k_t)} + O\left(\frac{t^2}{\Lambda(k_t)^2}\right)\right),$$

with error term $t^2/\Lambda(k_t)^2 \sim_\ell t^{2(1-2\beta)}$. Therefore, our MOTM call price approximation becomes

$$c(k_t, t) = \frac{\gamma(0)\sigma_0^2}{2} \frac{t^{3/2}}{\Lambda(k_t)} \exp\left(-\frac{\Lambda(k_t)}{t}\right) \left(1 - \frac{3}{2}\frac{t}{\Lambda(k_t)} + O(t^{2(1-2\beta)-\varepsilon})\right),$$

where $\varepsilon > 0$ is arbitrarily small. The Taylor expansion $\Lambda(k) = \frac{1}{2}\Lambda''(0)k^2 + \frac{1}{6}\Lambda'''(0)k^3 + O(k^4)$ implies

$$L_t := -\log c(k_t, t) = \frac{\Lambda(k_t)}{t} - \log\left(\frac{\gamma(0)\sigma_0^2}{2} \frac{t^{3/2}}{\Lambda(k_t)}\right) + \frac{3}{2}\frac{t}{\Lambda(k_t)} + O(t^{2(1-2\beta)-\varepsilon})$$

$$= \frac{1}{2}\Lambda''(0)\frac{k_t^2}{t} + \log\left(\frac{k_t^2}{t^{3/2}\sigma_0^4\gamma(0)}\right) + \frac{3}{\Lambda''(0)}\frac{t}{k_t^2} + \frac{1}{6}\Lambda'''(0)\frac{k_t^3}{t} + O(t^{\min\{2(1-2\beta),\beta\}-\varepsilon}).$$

We now translate the refined call price expansion to implied volatility asymptotics. In the proof of Corollary 4.1, we used corollary 7.2 of Gao and Lee (2014) to achieve the transfer. This would suffice for the interval $\beta \in (0, \frac{2}{5})$, too, but for the larger interval $\beta \in (0, \frac{3}{7})$ we have to take a closer look at the (arbitrary order) asymptotic machinery developed in Gao and Lee (2014). Any unexplained terminology and notation in what follows is as in Gao and Lee (2014). Using proposition 5.6, lemma 5.8, and example 5.13 of Gao and Lee (2014) yields the following estimates in our MOTM regime:

$$|G_-(k, \phi(k, L)) - V| = O\left(\frac{k}{L^{3/2}}\left(\frac{\Psi}{L^P} + \frac{1}{L^N}\right)\right), \tag{A.1}$$

$$|G_-(k, \phi(k, L)) - G_-(k, \phi(k, \hat{L}))| = O\left(\frac{k}{L^{1/2}}\frac{|L - \hat{L}|}{L}\right), \tag{A.2}$$

$$\left|G_-(k, \phi(k, \hat{L})) - \frac{k}{\sqrt{2\phi(k, \hat{L}) + k}}\right| = O\left(\frac{k}{L^{1/2}}\frac{k^2}{L^2}\right), \tag{A.3}$$

with integers $N, P \geq 1$, $L = -\log c(k_t, t)$, an approximation $\hat{L}$ of L, dimensionless implied volatility $V := t^{1/2}\sigma_{\text{imp}}(k_t, t)$, and error estimate Ψ. We suppress the time dependence of $k, L, \hat{l}, V$, and Ψ, in order to keep the notation of Gao and Lee (2014). Note that, in the MOTM regime, $k/L \to 0$ as $t \downarrow 0$.

We have $k \sim_\ell t^\beta$ and $L \sim_\ell t^{2\beta-1}$. The factor $k/L^{1/2} \sim_\ell t^{1/2}$ in (A.1)–(A.3) corresponds to the $t^{1/2}$-term of the dimensionless implied volatility V. The error term $k^2/L^2 \sim_\ell t^{2(1-\beta)}$ in (A.3) is of negligible order. Therefore, we have to deal with the iteration scheme error $(\Psi/L^P + 1/L^N)/L$ in (A.1) and the approximation error $|L - \hat{L}|/L$ in (A.2).

We now define the iteration scheme, following Gao and Lee (2014). The choice $N = 2$ and $P = 2$ suffices for our needs. Define a 2-ply regular iteration scheme $H := \{h, \eta_1, \eta_2\}$ via the sublog functions

$$h(\kappa, \lambda) := \alpha(\kappa, \lambda) - \frac{3}{2\lambda}$$

$$\eta_1(\kappa, \lambda) := \frac{3}{2\lambda}$$

$$\eta_2(\kappa, \lambda) := \frac{3}{2}\left(\log\left(1 + \frac{\alpha(\kappa, \lambda)}{\lambda}\right) - \frac{\alpha(\kappa, \lambda)}{\lambda}\right) + \frac{3}{2\lambda}\left(\left(1 + \frac{\alpha(\kappa, \lambda)}{\lambda}\right)^{-1} - 1\right),$$

using the auxiliary function

$$\alpha(\kappa, \lambda) := -\frac{3}{2}\log \lambda + \log\left(\frac{\kappa}{4\sqrt{\pi}}\right).$$

The corresponding implied volatility approximation function is given by

$$\phi_H(\kappa, \lambda) := \lambda + \alpha(\kappa, \lambda) - \frac{3}{2\lambda}(\alpha(\kappa, \lambda) + 1).$$

The 2-residual Ψ of H yields a sufficiently small iteration scheme error in (A.1):

$$\frac{1}{L}\left(\frac{\Psi}{L^2} + \frac{1}{L^2}\right) \sim_\ell t^{\min\{3(1-2\beta),\beta+2(1-2\beta)\}}. \tag{A.4}$$

Define $\kappa_t := \frac{1}{2\sigma_0^2}\frac{k_t^2}{t}$ for $t > 0$ and the constant $C := \frac{1}{3}\sigma_0^2\Lambda'''(0)$. By inserting $\Lambda''(0) = \sigma_0^{-2}$ and $\gamma(0) = (\sqrt{2\pi}\sigma_0)^{-1}$ (see part (iii) of Assumption A.1), we get the approximation

$$\hat{L}_t := \frac{1}{2\sigma_0^2}\frac{k_t^2}{t} + \ell_0(t) + 3\sigma_0^2\frac{t}{k_t^2} + \frac{1}{6}\Lambda'''(0)\frac{k_t^3}{t}$$

$$= \kappa_t\left(1 + \ell_0(t)\kappa_t^{-1} + \frac{3}{2}\kappa_t^{-2} + Ck_t\right),$$

where $\ell_0(t) := \log(\sqrt{2\pi}k_t^2/(\sigma_0^3 t^{3/2}))$ is slowly varying. Thus, the approximation error in (A.2) is

$$\frac{|L - \hat{L}|}{L} \sim_\ell t^{\min\{3(1-2\beta),1-\beta\}}. \tag{A.5}$$

It remains to put together all ingredients. By (A.1), (A.2), and (A.3), combined with (A.4) and (A.5), the following approximation of the dimensionless implied volatility $V_t := t^{1/2}\sigma_{\text{imp}}(k_t, t)$ holds

$$\left|\frac{k_t}{\sqrt{2\phi_H(k_t, \hat{L}_t) + k_t}} - V_t\right| \sim_\ell t^{1/2+\min\{3(1-2\beta),1-\beta\}}.$$

Further calculations show

$$\frac{k_t}{\sqrt{2\phi_H(k_t, \hat{L}_t) + k_t}}$$

$$= \frac{k_t}{\sqrt{2\hat{L}_t}}\left(1 + \frac{k_t}{2\hat{L}_t} + \frac{1}{\hat{L}_t}\alpha(k_t, \hat{L}_t) - \frac{3}{2\hat{L}_t^2}(\alpha(k_t, \hat{L}_t) + 1)\right)^{-1/2}$$

$$= \frac{k_t}{\sqrt{2\hat{L}_t}}\left(1 - \frac{1}{2\hat{L}}\alpha(k_t, \hat{L}) + \frac{3}{4\hat{L}^2}\left(\frac{1}{2}\alpha(k_t, \hat{L}_t)^2 + \alpha(k_t, \hat{L}) + 1\right) + O(t^{\min\{3(1-2\beta),1-\beta\}-\varepsilon})\right), \tag{A.6}$$

because $k_t/\hat{L}_t \sim_\ell t^{1-\beta}$, $1/\hat{L}_t^3 \sim_\ell t^{3(1-2\beta)}$, and $\alpha(k_t, \hat{L}_t) \sim_\ell 1$. Expansion of the appearing functions yields

$$\alpha(k_t, \hat{L}_t) = -\ell_0(t) - \frac{3}{2}\ell_0(t)\kappa_t^{-1} + O(t^{\min\{2(1-2\beta),\beta\}-\varepsilon}),$$

$$\alpha(k_t, \hat{L}_t)^2 = \ell_0(t)^2 + O(t^{\min\{1-2\beta,\beta\}-\varepsilon}),$$

$$\frac{1}{\hat{L}_t} = \kappa_t^{-1} - \ell_0(t)\kappa_t^{-2} + O(t^{\min\{3(1-2\beta),1-\beta\}-\varepsilon}),$$

$$\frac{1}{\hat{L}_t^2} = \kappa_t^{-2} + O(t^{\min\{3(1-2\beta),\beta+2(1-2\beta)\}-\varepsilon}),$$

$$\frac{1}{\sqrt{2\hat{L}}} = (2\kappa_t)^{-1/2}\left(1 - \frac{1}{2}\ell_0(t)\kappa_t^{-1} + \frac{3}{8}\ell_0(t)^2\kappa_t^{-2} - \frac{3}{4}\kappa_t^{-2} - \frac{1}{2}Ck_t + O(t^{\min\{3(1-2\beta),2\beta\}-\varepsilon})\right).$$

Putting these formulas back into (A.6), we get

$$\frac{k_t}{\sqrt{2\phi_H(k_t, \hat{L}_t) + k_t}} = \sigma_0 t^{1/2}\left(1 - \frac{1}{2}Ck_t + O(t^{\min\{3(1-2\beta),2\beta,1-\beta\}-\varepsilon})\right) \tag{A.7}$$

$$= t^{1/2}(\sigma_0 - \frac{1}{6}\sigma_0^3\Lambda'''(0)k_t) + O(t^{1/2+\min\{3(1-2\beta),2\beta,1-\beta\}-\varepsilon}).$$

For the second-order expansion of the implied volatility to be correct, the error term should be negligible compared to k_t, which amounts to $t^{\min\{3(1-2\beta),2\beta,1-\beta\}} = o(k_t)$. This is true if and only if $\min\{3(1-2\beta),2\beta,1-\beta\} > \beta$, which is equivalent to our assumption $\beta \in (0, \frac{3}{7})$. $\qquad\square$

For larger β, closer to $\frac{1}{2}$, the whole analysis has to be refined. A more precise iteration scheme H has to be chosen, so that the iteration scheme error in (A.4) gets smaller. Moreover, a better log-price approximation $\hat{L}_t$ has to be taken into account, using even more terms of the Laplace expansion, in order to decrease the approximation error in (A.5). It should thus be possible to reduce the error in (A.7) to

$$O(t^{\min\{n(1-2\beta),2\beta,1-\beta\}-\varepsilon}), \qquad t \downarrow 0,$$

where $n \in \mathbb{N}$ can be arbitrarily large. In this fashion, for any *fixed* n, it should be straightforward to provide a proof of the second-order approximation (4.1) of the implied volatility for $\beta < \frac{n}{2n+1}$. That is, we have a clear procedure for any $n > 2$. For small n, say $n = 3, 4, \ldots$, this can be implemented by hand, and larger values (say, $n = 17$) are still feasible with the aid of Mathematica or similar software. In practice, as the calculations in each proof will be tied to that specific value of n, very large n remains out of reach. Here one would need a new idea to provide an argument for general n, which would then prove (4.1) for all $\beta \in (0, \frac{1}{2})$. At this moment, despite some effort, the details of such a construction elude us. Still, we believe that Assumption A.1 suffices to treat the whole interval, i.e., that Theorem A.2 holds with $\frac{3}{7}$ replaced by $\frac{1}{2}$.

Note that the paper by Tehranchi (2016), which presents uniform (nonasymptotic) bounds on implied volatility, does not seem to be applicable here: For $\beta > \frac{1}{3}$, the lower bound of proposition 4.6 in Tehranchi (2016) is not tight enough, as it yields a second-order term that is asymptotically larger than the second-order term k_t in (4.1).