Marginal density expansions for diffusions and stochastic volatility

J.D. Deuschel, P.K. Friz, A. Jacquier, S. Violante
TU Berlin, TU and WIAS Berlin, TU Berlin, Imperial College

Abstract

Density expansions for hypoelliptic diffusions $\left(X^1,\dots,X^d\right)$ are revisited. In particular, we are interested in density expansions of the projection $\left(X_T^1,\dots,X_T^l\right)$, at time $T>0$, with $l\leq d$. Global conditions are found which replace the well-known "not-in-cutlocus" condition known from heat-kernel asymptotics; cf. G. Ben Arous [8]. Our small noise expansion allows for a "second order" exponential factor. Applications include tail and implied volatility asymptotics in some correlated stochastic volatility models; in particular, we solve a problem left open by A. Gulisashvili and E.M. Stein (2009).

Keywords: Laplace method on Wiener space, generalized density expansions in small noise and small time, sub-Riemannian geometry with drift, focal points, stochastic volatility, implied volatility, large strike and small time asymptotics for implied volatility

1 Introduction

Given a multi-dimensional hypoelliptic diffusion process $\mathrm{X}_t = \left(X_t^1,\dots,X_t^d : t\geq 0\right)$, started at $\mathrm{X}_0 = \mathrm{x}_0$, we are interested in the behaviour of the probability density function $f = f(\mathrm{y},t)$ of the projected (in general non-Markovian) process

$$\mathrm{Y}_t := \Pi_l \circ \mathrm{X}_t := \left(X_t^1,\dots,X_t^l\right),\ \ 1\leq l\leq d.$$

Both short time asymptotics and tail asymptotics, in presence of some scaling, can be derived from the small noise problem

$$d\mathrm{X}_t^\varepsilon = b\left(\varepsilon,\mathrm{X}_t^\varepsilon\right)dt + \varepsilon\sigma\left(\mathrm{X}_t^\varepsilon\right)dW_t,\quad \text{with } \mathrm{X}_0^\varepsilon = \mathrm{x}_0^\varepsilon \in \mathbb{R}^d.$$

Our main technical result, based on the Laplace method on Wiener space following Ben Arous [8, 9], is a density expansion for $\mathrm{Y}_t^\varepsilon := \Pi_l \circ \mathrm{X}_t^\varepsilon$ of the form, for $\mathrm{x}_0, \mathrm{y}, T$ fixed,

$$f^\varepsilon(\mathrm{y},T) = e^{-c_1/\varepsilon^2} e^{c_2/\varepsilon}\varepsilon^{-l}\left(c_0 + O(\varepsilon)\right)\ \text{ as } \varepsilon\downarrow 0. \tag{1}$$

Leaving definitions and precise statements to the main text below (cf theorem 9) let us briefly mention our key assumptions
(i) a strong Hörmander condition at all points (or in fact, a weak Hörmander condition at x_0 and

an explicit controllability condition);
(ii) existence of at most finitely many minimizers in the control problem which govern the leader order behaviour;
(iii) invertibility of the deterministic Malliavin covariance matrix along the minimizers;
(iv) a global condition on $\mathrm{x}_0 \in \mathbb{R}^d$, $\mathrm{y} \in \mathbb{R}^l$ which we call *non-focality;* motivated by geometric terminology [46, 12].

Conditions (i)-(iii) will not surprise the reader familiar with the works [8, 9, 11, 49]. However, condition (iv)[1] which guarantees *non-degeneracy* of the minimizers (cf. proposition 7), appears to be new in the context of density expansions, to the best of our knowledge, even in the Riemannian case. It forms the essence of what is needed to extend the well-known *point-point* concept of *non-conjugacy* (crucial part of the "$\notin$ *cut-locus condition"* familiar from heat kernel expansions) to a (sub-Riemannian, with drift) *point-subspace* setting. A simple example where (iv) *and* (1) fails, is given in section 4.2. We emphasize that our applications require us to introduce and characterize non-focality in a control-theoretic generality.

As far as the expansion (1) is concerned, we draw attention to (in the context of density expansions) the somewhat unusual *second order exponential factor* present when $c_2 \neq 0$. As was understood in the context of the general Laplace method on Wienerspace, [5, 9], this has to do with allowing drift vector field b (and in the present paper also: starting point) depend on ε in first order; the special case that arises from considering short time asymptotics - the small noise parameter ε is then introduced by Brownian scaling - always leads to $c_2 = 0$. It is interesting to note the work of Kusuoka–Stroock [36], concerning precise asymptotics for Wiener functionals (in the small noise limit), see also [42, 35] for recent applications to projected diffusions, was set up as expansion in ε^2. This is enough to cover the model case of short time expansions, but cannot yield an expansion of the type (1) with $c_2 \neq 0$. A similar remark applies to the small noise expansions for projected diffusions due to Takanobu–Watanabe [49].

One of our main motivations comes from recent density expansions by Gulisashvili–Stein. In [27, Theorem 2.1] they prove that the stock-price in the *uncorrelated* Stein–Stein stochastic volatility model admits a density with expansion[2]

$$B_1 s^{-B_3} e^{B_2\sqrt{\log s}} (\log s)^{-\frac{1}{2}} \left(1 + O\left(\log s\right)^{-\frac{1}{2}}\right) \text{ as } s \uparrow \infty$$

with explicitly computable constants; asymptotic formulae of the implied volatility in the large (similar: small) strike regime are then obtained as corollaries. When writing this expansion in terms of log-price $Y = \log S$, it indeed has the form (1) with $y = \log s = 1/\varepsilon^2$. More generally, we can show from rather general and robust principles that the tail behaviour of $Y_T \in \mathbb{R}^1$ for fixed $T > 0$, subject to a certain scaling with parameter $\theta \in \{1, 2\}$ in the full Markovian specification of the model, has the form

$$f(y, T) = e^{-c_1 y^{2/\theta}} e^{c_2 y^{1/\theta}} y^{\frac{1}{\theta}-1} \left(c_0 + O\left(1/y^{1/\theta}\right)\right) \text{ as } y \uparrow \infty. \tag{2}$$

It is worth mentioning that such an expansion leads immediately to call price and then (Black–Scholes) implied volatility expansions in the large strike regime, cf. [27, 29]; in the case $\theta = 2$

[1] More precisely, $\mathrm{x}_0 \in \mathbb{R}^d$ must not be focal for the submanifold $N_y := (\mathrm{y}, \cdot) \subset \mathbb{R}^d$. The classical example here is of course $(0, 0) \in \mathbb{R}^2$ which is focal for unit circle $S^1 \subset \mathbb{R}^2$.

[2] Strictly speaking, their O-term is $\log s$ with power $-1/4$; the authors have informed us, however, that a closer look at their argument indeed gives power $-1/2$.

typical for stochastic volatility,

$$\sigma_{BS}(k,T)^2 T = (\beta_1 k + \beta_2 + o(1))^2 \text{ as log-strike } k \to \infty;$$
$$\beta_1 = \sqrt{2}\left(\sqrt{c_1} - \sqrt{c_1 - 1}\right),$$
$$\beta_2 = c_2\sqrt{2}\left(1/\sqrt{c_1 - 1} - 1/\sqrt{c_1}\right).$$

(Small strike asymptotics are similar and will not be discussed here.) The leading order behaviour described by $\beta_1 = \beta_1(c_1)$ is well understood [37, 7]; the second order behaviour is given by $\beta_2 = \beta_2(c_1, c_2)$. Further terms in this expansion are in principle possible [29]; in particular, the next term would involve c_0. When applied to the Stein–Stein stochastic volatility model,[3] the afore-mentioned scaling indeed leads to a small-noise, hypoelliptic diffusion problem with non-vanishing second order exponential factor, as is handled by our main theorem. We then solve a problem left open in the afore-mentioned work [27, Theorem 2.1] in that we are able to compute the expansion in the correlated case. The importance of allowing for correlation in stochastic volatility models is well-documented, e.g. [20, 38], and evidence from estimation of parametric stochastic volatility models suggests correlation parameter $\rho \approx -0.7$ or $\rho \approx -0.8$ for S&P 500, for instance; a finding fairly robust across models and time periods [1]. With this in mind, we shall focus on the case $-1 < \rho \leq 0$ in our explicit analysis and derive explicit expressions for c_1, c_2. (In principle, the Laplace method on which we rely yields an explicit expression for c_0, cf. [9, Thm 4, p 135], [35].)

Density expansions of diffusions in the small noise regime seem to go back (at least) to [34]; density expansions for projected diffusions in the small noise regime (which include the short time regime), with applications to implied volatility expansions, were recently considered by Y. Osajima [42], based on work with S. Kusuoka [35]. We partially improve on these results. First, as was already mentioned, $c_2 = 0$ in these works which makes any expansion of the form (2) out of reach. Additionally, in comparison with [42] we do not assume x_0 near $(y, \cdot)$, nor ellipticity of the problem. In further contrast to (the general results in) [35, 36] we provide a checkable, finite-dimensional criterion that guarantees that the crucial infinite-dimensional non-degeneracy assumption, left as such in [35, 36], is actually satisfied. On the other hand, these authors give explicit formulae for c_0 which we (presently) do not.

Finally, our expansion (1) leads to short time expansion for projected diffusion densities, under global conditions on $(\mathrm{x}_0, \mathrm{y})$, of the form

$$f(\mathrm{y}, t) \sim c_0(\mathrm{x}_0, \mathrm{y}) \frac{1}{t^{l/2}} \exp\left(-\frac{d^2(\mathrm{x}_0, \mathrm{y})}{2t}\right) \text{ as } t \downarrow 0. \tag{3}$$

When $l = d$, such expansions go back to classical works ranging from Molchanov [40] to Ben Arous [8]. The leading order behaviour $2t \log f(\mathrm{y}, t) \sim -d^2(\mathrm{x}_0, \mathrm{y})$ is due to Varadhan [50]. The case $l < d$, in particular our global condition on $(\mathrm{x}_0, \mathrm{y})$, appears to be new. That said, expansions of this form have appeared in [49, 30, 42]; the last two references aimed at implied volatility expansions. In the context of a time-homogenuous local volatility models ($l = d = 1$), the expansion (3) holds trivially without any conditions on $(\mathrm{x}_0, \mathrm{y})$; the resulting expansion was derived (with explicit constant c_0) in [22]. Subject to mild technical conditions on the diffusion coefficient, they show how to deduce first a call price and then an implied volatility expansion in the short time (to maturity) regime:

$$\sigma_{BS}(k,t) = k/d(\mathrm{x}_0, k) + c(\mathrm{x}_0, k)\, t + O\left(t^2\right) \text{ as } t \downarrow 0;$$

[3] In fact, the leading order behaviour of the density was discussed with large deviations methods in [16, p40, p265].

where $d(\mathrm{x}_0, k)$ is a point-point distance and $c(\mathrm{x}_0, k)$ is explicitly given. The celebrated Berestycki–Busca–Florent (BBF) formula [13] asserts that $\sigma_{BS}(k,t) \sim k/d(\mathrm{x}_0,k)$ as $t \downarrow 0$, is in fact valid in generic stochastic volatilty models, $d(\mathrm{x}_0,k)$ is then understood as point-hyperplane distance. In fact, $k/d(\mathrm{x}_0,k)$ arose as initial condition of a non-linear evolution equation for the entire implied volatility surface. As briefly indicated in [13, Sec 6.3] this can be used for a Taylor expansion of $\sigma_{BS}(k,t)$ in t. Such expansions have also been discussed, based on heat kernel expansions on Riemannian manifolds by [15, 31, 43], not always in full mathematical rigor. Some mathematical results are given in [42], assuming ellipticity and *close-to-the-moneyness* $|k| << 1$; see also forthcoming work by Ben Arous–Laurence [10]. We suspect that our formula (3), potentially applicable far-from-the-money, will prove useful in this context and shall return to this in future work.

It should be noted, that the BBF formula alone can be obtained from soft large deviation arguments, cf. [44, Sec. 3.2.1] and the references therein. In a similar spirit, the Varadhan-type formula $2t \log f(\mathrm{y},t) \sim -d^2(\mathrm{x}_0,\mathrm{y})$, when $l < d$, can be shown, without any conditions on $(\mathrm{x}_0,\mathrm{y})$ by large deviation methods, only relying on the existence of a reasonable density, cf. [51, Sec 5, Rmk 2.9].

As a final note, we recall that the (in general, non-Markovian) $\mathbb{R}^l$-valued Itô-process $(\mathrm{Y}_t : t \geq 0)$ admits - subject to some technical assumptions [28, 45] - a *Markovian (or Gyöngy) projection*. That is, a time-inhomogeneous Markov diffusion $(\tilde{\mathrm{Y}}_t : t \geq 0)$ with matching time-marginals i.e $\mathrm{Y}_t = \tilde{\mathrm{Y}}_t$ (in law) for every fixed $t \geq 0$. In a financial context, when $l = 1$, this process is known as (Dupire) local volatility model and various authors [13, 15, 31, 10] have used this as an important intermediate step in computing implied volatility in stochastic volatility models. Since all our expansions (small noise, tail, short time) are relative to such time-marginals they may also be viewed as expansions for the corresponding Markovian projections.

Acknowledgement: JDD, PKF and AJ would like to thank MATHEON for financial support. PKF would like to thank G. Ben Arous for pointing out conceptual similarities in [24, 8] and several discussions thereafter. It is also a pleasure to thank F. Baudoin, J.P. Gauthier, A. Gulisashvili and P. Laurence for their interest and feedback.

2 The main result and its corollaries

Consider a d-dimensional diffusion $(\mathrm{X}^\varepsilon_t)_{t\geq 0}$ given by the stochastic differential equation

$$d\mathrm{X}^\varepsilon_t = b(\varepsilon, \mathrm{X}^\varepsilon_t)\, dt + \varepsilon\sigma(\mathrm{X}^\varepsilon_t)\, dW_t, \quad \text{with } \mathrm{X}^\varepsilon_0 = \mathrm{x}^\varepsilon_0 \in \mathbb{R}^d \tag{4}$$

and where $W = (W^1, \ldots, W^m)$ is an m-dimensional Brownian motion. Unless otherwise stated, we assume $b : [0,1) \times \mathbb{R}^d \to \mathbb{R}^d$, $\sigma = (\sigma_1, \ldots, \sigma_m) : \mathbb{R}^d \to L(\mathbb{R}^m, \mathbb{R}^d)$ and $\mathrm{x}^\cdot_0 : [0,1) \to \mathbb{R}^d$ to be smooth, bounded with bounded derivatives of all orders. Set $\sigma_0 = b(0,\cdot)$ and assume that, for every multiindex α, the drift vector fields $b(\varepsilon,\cdot)$ converges to σ_0 in the sense[4]

$$\partial^\alpha_x b(\varepsilon,\cdot) \to \partial^\alpha_x b(0,\cdot) = \partial^\alpha_x \sigma_0(\cdot) \text{ uniformly on compacts as } \varepsilon \downarrow 0. \tag{5}$$

We shall also assume that

$$\partial_\varepsilon b(\varepsilon,\cdot) \to \partial_\varepsilon b(0,\cdot) \quad \text{uniformly on compacts as } \varepsilon \downarrow 0 \tag{6}$$

[4] If (4) is understood in Stratonovich sense, so that dW is replaced by $\circ dW$, the drift vector field $b(\varepsilon,\cdot)$ is changed to $\tilde{b}(\varepsilon,\cdot) = b(\varepsilon,\cdot) - (\varepsilon^2/2)\sum_{i=1}^m \sigma_i \cdot \partial\sigma_i$. In particular, σ_0 is also the limit of $\tilde{b}(\varepsilon,\cdot)$ in the sense of (5) .

and

$$\mathrm{x}_0^\varepsilon = \mathrm{x}_0 + \varepsilon \mathrm{x}_0' + o(\varepsilon) \text{ as } \varepsilon \downarrow 0. \tag{7}$$

Later applications to stochastic volatility models aside, the (technical) main result of this paper is a density expansion in ε of the $\mathbb{R}^l$-valued (in general, non-Markovian) projection[5]

$$\mathrm{Y}_T^\varepsilon := \Pi_l \circ \mathrm{X}_T^\varepsilon := \left(X_T^{\varepsilon,1}, \ldots, X_T^{\varepsilon,l}\right) \in \mathbb{R}^l;$$

where Π_l denotes the projection $\left(x^1, \ldots, x^d\right) \mapsto \left(x^1, \ldots, x^l\right)$, for fixed $l \in \{1, \ldots, d\}$ and $T > 0$. Of course, we need to guarantee that Y_T^ε indeed admits a density. To this end, we make the *standing assumption* that *weak Hörmander condition* holds at x_0,

$$\text{span}\left[\sigma_i : 1 \le i \le m; \ [\sigma_j, \sigma_k] : 0 \le j, k \le m; ...\right]_{\mathrm{x}_0} = \mathcal{T}_{\mathrm{x}_0}\mathbb{R}^d; \tag{H}$$

that is the linear span of $\sigma_1, \ldots, \sigma_m$ and all Lie brackets of $\sigma_0, \sigma_1, \ldots, \sigma_m$ is full. Since this condition is "open" it also holds, thanks to (5), for $\varepsilon > 0$ small enough, with σ_0 and x_0 replaced by $b(\varepsilon, \cdot)$ (or $\tilde{b}(\varepsilon, \cdot)$, cf. previous footnote) and x_0^ε, respectively. It then is a classical result (due to Hörmander, Malliavin) that the $\mathbb{R}^d$-valued r.v. X_T^ε admits a (smooth) density for all times $T > 0$ and so does its $\mathbb{R}^l$-valued projection Y_T^ε. We denote the probability density of Y_T^ε by

$$f_\varepsilon(\cdot, T) \equiv f_\varepsilon(\mathrm{y}, T) \text{ with } \mathrm{y} \in \mathbb{R}^l.$$

In theorem 9 below, it will be assumed that $\mathcal{K}_\mathrm{y}$ is non-empty, where for fixed $\mathrm{a} \in \mathbb{R}^l$ we define[6]

$$\mathcal{K}_{x_0,T;\mathrm{a}} = \mathcal{K}_\mathrm{a} := \{\mathrm{h} \in H : \Pi_l \circ \phi_T(h) = \mathrm{a}\}. \tag{8}$$

Here, H denotes the Cameron-Martin space, i.e. absolutely continuous paths with derivative in $L^2([0,T], \mathbb{R}^m)$, and $\phi_T(\mathrm{h})$ denotes the time-T solution to the controlled ordinary differential equation

$$d\phi_t^\mathrm{h} = \sigma_0\left(\phi_t^\mathrm{h}\right) dt + \sum_{i=1}^m \sigma_i\left(\phi_t^\mathrm{h}\right) d\mathrm{h}_t^i, \ \phi_0^\mathrm{h} = \mathrm{x}_0 \in \mathbb{R}^d. \tag{9}$$

Let us also define $\phi_{T \leftarrow t}(\mathrm{h}) := \phi_T(\mathrm{h}) \circ (\phi_t(\mathrm{h}))^{-1}$; at occasions we shall also write $\phi_T^\mathrm{h}(\mathrm{x}_0)$ resp. $\phi_{T \leftarrow t}^\mathrm{h}(\mathrm{x}_0)$ instead of $\phi_T(\mathrm{h})$ resp. $\Phi_{T \leftarrow t}(\mathrm{h})$. We note that $\phi_T(\mathrm{h})$ is a diffeomorphism, as function of $\mathrm{x}_0 \in \mathbb{R}^d$, and denotes its differential by $\Phi_{T \leftarrow t}(\mathrm{h})$.

A well-known sufficient condition for $\mathcal{K}_\mathrm{a} \neq \varnothing$ is the *strong Hörmander condition*[7]

$$\forall x \in \mathbb{R}^d : \text{Lie}\left[\sigma_1, \ldots, \sigma_m\right]|_x = \mathcal{T}_x\mathbb{R}^d \cong \mathbb{R}^d. \tag{H1}$$

Whenever $\mathcal{K}_\mathrm{a} \neq \varnothing$, it makes sense to define the *energy* and the set of *minimizers*

$$\Lambda_{x_0,T}(\mathrm{a}) := \Lambda(\mathrm{a}) := \inf\left\{\frac{1}{2}\|\mathrm{h}\|_H^2 : \mathrm{h} \in \mathcal{K}_\mathrm{a}\right\}, \tag{10}$$

$$\mathcal{K}_\mathrm{a}^{\min} := \left\{\mathrm{h}_0 \in \mathcal{K}_\mathrm{a} : \frac{1}{2}\|\mathrm{h}_0\|_H^2 = \Lambda(\mathrm{a})\right\}.$$

[5] Due to the non-Markovianity of the problem, PDE techniques are poorly suited to study the density of Y_T^ε.

[6] In later applications to tail asymptotics, when $l = 1$, we have $\mathrm{a} = 1$ and $\varepsilon := 1/y$ or $1/y^2$ as $y \uparrow \infty$. For this reason we prefer to keep a and y at this stage separated.

[7] A weak Hörmander type condition which ensures $\mathcal{K}_\mathrm{a} \neq \varnothing$ is found in [33].

In words, $\Lambda(\mathrm{a})$ is the minimal energy required to go in time T from $\mathrm{x}_0 \in \mathbb{R}^d$ to the "target" submanifold

$$N := N_{\mathrm{a}} := \left\{ \mathrm{x} \in \mathbb{R}^d : \Pi_l(\mathrm{x}) = \left(x^1, \ldots, x^l\right) = \mathrm{a} \right\}.$$

Elements of $\mathcal{K}_{\mathrm{a}}^{\min}$ will be called *minimizers* or *minimizing controls*. A standard weak-compactness argument shows that $\mathcal{K}_{\mathrm{a}} \neq \varnothing$ already implies that $\mathcal{K}_{\mathrm{a}}^{\min}$ is non-empty. (Throughout the paper, we shall only be concerned with the situation that $\mathcal{K}_{\mathrm{a}}^{\min}$ contains one or finitely many minimizers.)

Following Bismut [14] it will be crucial that $\mathcal{K}_{\mathrm{a}}$ enjoys a (Hilbert) manifold structure, locally around (each) $\mathrm{h}_0 \in \mathcal{K}_{\mathrm{a}}^{\min}$. As is well-known, this can be guaranteed by assuming invertibility of $C_{x_0,T;\mathrm{a}}(\mathrm{h}_0)$, the deterministic Malliavin matrix given by

$$C_{x_0,T;\mathrm{a}}(\mathrm{h}) := C(\mathrm{h}) := \langle D\phi_T(\mathrm{h}), D\phi_T(\mathrm{h}) \rangle_H \in \mathrm{Lin}\left(\mathcal{T}^*_{\mathrm{x}_T}\mathbb{R}^d \to \mathcal{T}_{\mathrm{x}_T}\mathbb{R}^d\right) \cong \mathbb{R}^{d\times d}$$

where $\mathrm{x}_T := \phi_T(\mathrm{h})$ and D will always denote the (H-valued) Fréchet derivative of some function depending on $\mathrm{h} \in H$. We can also view $C(\mathrm{h})$ as (positive semi-definite) quadratic form on $\mathcal{T}^*_{\mathrm{x}_T}\mathbb{R}^d$, in coordinates

$$\langle C(\mathrm{h})\, p, p\rangle = \sum_{i=1}^{d} \left\langle D\phi_T^i(\mathrm{h}), p_i \right\rangle^2 \text{ where } p = p_i dx^i \in \mathcal{T}^*_{\mathrm{x}_T}\mathbb{R}^d.$$

In fact, large parts of our analysis only rely on non-degeneracy of $C(\mathrm{h})$ restricted to $\mathbb{R}^{l\times l}$ but we find it more convenient to deal with the "full" matrix $C(\mathrm{h}_0)$. The following condition will cover most of our applications[8].

Proposition 1 *Assume* $\mathrm{h} \in H$ *and*

$$\exists t \in [0,T] : span\left[\sigma_1, \ldots, \sigma_m\right]|_{\mathrm{x}_t} = \mathcal{T}_{\mathrm{x}_t}\mathbb{R}^d$$

where $\mathrm{x}_t := \phi_t(\mathrm{h})$. *Then* $C(\mathrm{h})$ *is invertible.*

Proof. We have the well-known formula, for any $k \in H$,

$$\langle D\phi_T(\mathrm{h}), k\rangle_H = D\phi_T(\mathrm{h})[k] = \int_0^T \sum_{j=1}^{m} \Phi_{T\leftarrow t}(\mathrm{h})\, \sigma_j(\mathrm{x}_t)\, \dot{k}_t^j dt \in \mathcal{T}_{\mathrm{x}_T}\mathbb{R}^d$$

When pairing this with $p = p_i dx^i \in \mathcal{T}^*_{\mathrm{x}_T}\mathbb{R}^d$, we have

$$\langle \langle p, D\phi_T(\mathrm{h})\rangle, k\rangle_H = \int_0^T \sum_{j=1}^{m} \langle p, \Phi_{T\leftarrow t}(\mathrm{h})\, \sigma_j(\mathrm{x}_t)\rangle\, \dot{k}_t^j dt \in \mathcal{T}_{\mathrm{x}_T}\mathbb{R}^d$$

and it easily follows that

$$\|\langle p, D\phi_T(\mathrm{h})\rangle\|_H^2 = \int_0^T \sum_{j=1}^{m} \langle p, \Phi_{T\leftarrow t}(\mathrm{h})\, \sigma_j(\mathrm{x}_t)\rangle^2 dt = \int_0^T \sum_{j=1}^{m} \left\langle \left(\Phi_{T\leftarrow t}(\mathrm{h})\right)^* p, \sigma_j(\mathrm{x}_t)\right\rangle^2 dt.$$

By assumption $\mathrm{span}[\sigma_1, \ldots, \sigma_m]|_{\mathrm{x}_t} = \mathcal{T}_{\mathrm{x}_t}\mathbb{R}^d$ for some $t \in [0,T]$, and this clearly remains valid in a small enough open interval containing t which is enough to conclude $\left(\Phi_{T\leftarrow t}(\mathrm{h})\right)^* p \equiv 0$. By non-degeneracy of the (co-)tangent flow, this implies $p = 0$ and so $C(\mathrm{h})$ is non-degenerate, as claimed. ■

[8] A sufficient condition for "$C(\mathrm{h})$ is invertible for every $h \neq 0$" in a strictly sub-elliptic setting is given as condition (H2) by [14]; although much stronger than Hörmander's condition, it does apply to examples such as the 3-dimensional Heisenberg group.

Remark 2 (Tangent space of $\mathcal{K}_{\mathrm{a}}$) *Assume $C(\mathrm{h})$ is invertible. Then $\mathcal{K}_{\mathrm{a}}$ enjoys a (Hilbert) manifold structure, locally around* h *and*

$$\mathcal{T}_{\mathrm{h}}\mathcal{K}_{\mathrm{a}} \cong \ker D\left(\Pi_l \circ \phi_T\right)(\mathrm{h}) =: H_0.$$

Moreover,

$$H_0 = \left\{ k \in H : \sum_{j=1}^{m} \langle p, \Phi_{0\leftarrow t}(\mathrm{h})\, \sigma_j(\mathrm{x}_t)\rangle \dot{k}_t^j dt = 0 \forall p \in \operatorname{span}\left[dx^1, \ldots, dx^l\right]|_{\mathrm{x}_T} \subset \mathcal{T}^*_{\mathrm{x}_T}\mathbb{R}^d \right\}$$

We now introduce the *Hamiltonian*

$$\begin{aligned}\mathcal{H}(x,p) &:= \langle p, \sigma_0(x)\rangle + \frac{1}{2}\sum_{i=1}^{m} \langle p, \sigma_i(x)\rangle^2 \\ &= \langle p, \sigma_0(x)\rangle + \frac{1}{2}\left\langle p, \left(\sigma\sigma^T\right)(x)\, p\right\rangle\end{aligned}$$

and $\mathrm{H}_{t\leftarrow 0} = \mathrm{H}_{t\leftarrow 0}(x_0, p_0)$ as the flow associated to the vector field $(\partial_p\mathcal{H}, -\partial_x\mathcal{H})$ on $\mathcal{T}^*\mathbb{R}^d$. (Under which $\mathcal{H}$ is invariant; it follows that $H_{\cdot\leftarrow 0}$ does not explode.)

Remark 3 *Our setup here is tied to the SDE (4), driven by m independent Brownians $W^1, \ldots, W^m$. Many stochastic models, notably in finance, are written in terms of correlated Brownians, i.e. with a non-trivial correlation matrix $\Omega = \left(\omega^{i,j} : 1 \le i, j \le m\right)$, where $d\left\langle W^i, W^j\right\rangle_t = \omega^{i,j} dt$. The Hamiltonian then becomes*

$$\mathcal{H}(x,p) = \langle p, \sigma_0(x)\rangle + \frac{1}{2}\left\langle p, \left(\sigma\Omega\sigma^T\right)(x)\, p\right\rangle. \tag{11}$$

The following propositions generalize the respective results in Bismut's book [14] (see also Ben Arous [8, Theorems 1.15 and 1.1.8]) from a drift-free ($\sigma_0 \equiv 0$), point-to-point setting ($x_0 \in \mathbb{R}^d$ to $y \in \mathbb{R}^d$) to a point-to-subspace setting ($x_0 \in \mathbb{R}^d$ to $(y, \cdot) \in \mathbb{R}^l \oplus \mathbb{R}^{d-l}$) with drift vector field σ_0. Note that the Bismut setting [14, Chapter I] is recovered by taking zero drift, $\sigma_0 \equiv 0$, and $l = d$.

Proposition 4 *If (i) $\mathrm{h}_0 \in \mathcal{K}_{\mathrm{a}}^{\min}$ is a minimizing control and (ii) the deterministic Malliavin covariance matrix $C(\mathrm{h}_0)$ is invertible then there exists a unique $\mathrm{p}_0 = \mathrm{p}_0(\mathrm{h}_0) \in \mathcal{T}^*_{\mathrm{x}_0}\mathbb{R}^d$, in fact*[9]

$$\mathrm{p}_0 \in \left(\Phi_{0\to T}(\mathrm{h}_0)\right)^* \operatorname{span}\left\{dx^1, \ldots, dx^l\right\}|_{\phi_T^h(\mathrm{x}_0)},$$

such that

$$\phi_t^{\mathrm{h}_0}(\mathrm{x}_0) = \pi \mathrm{H}_{t\leftarrow 0}(\mathrm{x}_0, \mathrm{p}_0),\ 0 \le t \le T \tag{12}$$

(π denotes the projection from $\mathcal{T}^*\mathbb{R}^d$ onto $\mathbb{R}^d$; in coordinates $\pi(\mathrm{x}, \mathrm{p}) = \mathrm{x}$).
Moreover, $(\mathrm{x}(t), \mathrm{p}(t)) := H_{t\leftarrow 0}(\mathrm{x}_0, \mathrm{p}_0)$ solves the Hamiltonian ODEs in $\mathcal{T}^*\mathbb{R}^d \cong \mathbb{R}^d \oplus \mathbb{R}^d$

$$\begin{pmatrix} \dot{\mathrm{x}} \\ \dot{\mathrm{p}} \end{pmatrix} = \begin{pmatrix} \partial_p\mathcal{H}(\mathrm{x}(t), \mathrm{p}(t)) \\ -\partial_x\mathcal{H}(\mathrm{x}(t), p(t)) \end{pmatrix}, \tag{13}$$

[9] The (global) coordinate chart $\left(x^1, \ldots, x^d\right)$ of $\mathbb{R}^d$ induces coordinates co-vectors fields (or one-forms) $\left(dx^1, \ldots, dx^d\right)$.

the minimizing control $\mathrm{h}_0 = \mathrm{h}_0(\cdot)$ *is recovered by*

$$\dot{\mathrm{h}}_0 = \begin{pmatrix} \langle \sigma_1(\mathrm{x}(\cdot)), \mathrm{p}(\cdot) \rangle \\ \cdots \\ \langle \sigma_m(\mathrm{x}(\cdot)), \mathrm{p}(\cdot) \rangle \end{pmatrix} \tag{14}$$

and with $C := \mathcal{H}(\mathrm{x}(t), \mathrm{p}(t))$ *independent of* $t \in [0,T]$,

$$\Lambda(\mathrm{y}) = \frac{1}{2}\|\mathrm{h}_0\|_H^2 = \frac{TC}{2} - \frac{1}{2}\int_0^T \langle \sigma_0(\mathrm{x}(t)), \mathrm{p}(t) \rangle \, dt. \tag{15}$$

At last, crucial for actual computations, $(\mathrm{x}(t), \mathrm{p}(t)) = H_{t \leftarrow 0}(\mathrm{x}_0, \mathrm{p}_0)$ *satisfies the Hamiltonian ODEs (13) as boundary value problem, subject to the following initial -, terminal - and transversality conditions,*

$$\begin{aligned} &\mathrm{x}(0) = \mathrm{x}_0 \in \mathbb{R}^d \\ &\mathrm{x}(T) = (\mathrm{y}, \cdot) \in \mathbb{R}^l \oplus \mathbb{R}^{d-l} \\ &\mathrm{p}(T) = (\cdot, 0) \in \mathbb{R}^l \oplus \mathbb{R}^{d-l}. \end{aligned} \tag{16}$$

Proof. The key remark, due to Bismut [14, Chapter I], is that under the assumption "$\exists C(\mathrm{h}_0)^{-1}$" the set $\mathcal{K}_{\mathrm{a}}^{\min}$ can be described by Hamilton–Jacobi theory. It then suffices to adapt the arguments of Bismut, as done in the drift-free case by Takanobu–Watanabe, [49]. Let us note that the additional drift vector field σ_0 is trivially incorporated in their setting by adding a 0th component to the controls, i.e. $h^0(t) = t$. The boundary conditions - in particular, transversality, have not been pointed out explicitly in [49] although are implicitly contained in their formulation. In fact, formal application of Pontryagin's maximum principle leads precisely to the above boundary value problem; care is necessary, however, since without assuming invertibility of $C(\mathrm{h}_0)$, one can be in the so-called "strictly abnormal" case; the above approach is then not possible. ■

Remark 5 *Assume there exists a (smooth) map* $(\mathrm{a}-\varepsilon, \mathrm{a}+\varepsilon) \ni \mathrm{y} \mapsto \mathrm{h}_0(\mathrm{y}) \in \mathcal{K}_{\mathrm{y}}^{\min}$. *Then*

$$\Lambda(\mathrm{y}) = \frac{1}{2}\|\mathrm{h}_0(\mathrm{y})\|^2 \implies \partial_{\mathrm{y}}\Lambda(\mathrm{a}) = \langle \mathrm{h}_0(\mathrm{a}), \partial_{\mathrm{y}}\mathrm{h}_0(\mathrm{a}) \rangle_H.$$

On the other hand, we know from $\Pi_l \phi_T(\mathrm{h}_0(\mathrm{y})) = \mathrm{y}$ *that*

$$(\Pi_l)_* D\phi_T(\mathrm{h}_0(\mathrm{a}))[\partial_{\mathrm{y}}\mathrm{h}_0(\mathrm{a})] = \mathrm{Id} \text{ on } \mathbb{R}^l;$$

Using $\mathrm{h}_0(\mathrm{y}) = D\phi_T(\mathrm{h}_0(\mathrm{y}))^* \mathrm{p}(T) = D\phi_T(\mathrm{h}_0(\mathrm{y}))^* (\Pi_l)_*^* q(\mathrm{a})$, *where* $(\Pi_l)_*^* q(\mathrm{a}) \equiv (q(\mathrm{a}), 0)$, *we have*

$$\langle \mathrm{h}_0(\mathrm{a}), \partial_{\mathrm{y}}\mathrm{h}_0(\mathrm{a}) \rangle_H = \langle q(\mathrm{a}), (\Pi_l)_* D\phi_T(\mathrm{h}_0(\mathrm{y}))\, \partial_{\mathrm{y}}\mathrm{h}_0(\mathrm{a}) \rangle_H.$$

It thus follows that the derivatives of the energy are given in terms of $q(\mathrm{a})$,

$$\begin{pmatrix} \partial_{\mathrm{y}^1}\Lambda(\mathrm{a}) \\ \cdots \\ \partial_{\mathrm{y}^l}\Lambda(\mathrm{a}) \end{pmatrix} = \begin{pmatrix} q_1(\mathrm{a}) \\ \cdots \\ q_l(\mathrm{a}) \end{pmatrix}. \tag{17}$$

This can be a useful short when computing the energy from the Hamiltonian system. If $\#\mathcal{K}_{\mathrm{a}}^{\min} = 1$ *for some* a, *and our non-degeneracy condition (ND) as introduced below is met, the existence of such a map* $\mathrm{h}_0(\cdot)$ *can be shown along the lines of [14, Thm 1.26]. We shall not rely on formula (17) in the sequel although will find it confirmed in various examples.*

Remark 6 (How to compute optimal controls h_0) *Proposition 4 - as it stands - requires* h_0 *to be a minimizer and then, subject to condition (ii), provides us with some information about* $\phi^{\mathrm{h}_0}(\mathrm{x}_0), (\mathrm{x},\mathrm{p}) \equiv H_{\cdot\leftarrow 0}(\mathrm{x}_0,\mathrm{p}_0)$ *and in particular allows us to reconstruct* h_0 *from the Hamiltonian flow* (x,p)*, cf. equation (14). That said, we can consider any solution to the boundary valued problem (13),(16), say* $(\hat{\mathrm{x}},\hat{\mathrm{p}})$*, and define a (possible non-minimizing) control path* $\hat{h_0}$ *via (14) i.e.*

$$\hat{\mathrm{h}}_0^i = \int_0^{\cdot} \langle \sigma_i(\hat{\mathrm{x}}(t)), \hat{\mathrm{p}}(t)\rangle \, dt, \quad i = 1,\dots,m.$$

From (13),

$$d\hat{\mathrm{x}}_t = \partial_p \mathcal{H}(\hat{\mathrm{x}}_t,\hat{\mathrm{p}}_t)\,dt = \sigma_0(\hat{\mathrm{x}}_t)\,dt + \sum_{i=1}^m \sigma_i(\hat{\mathrm{x}}_t)\langle \sigma_i(\hat{\mathrm{x}}_t),\hat{\mathrm{p}}_t\rangle\,dt$$

and so relation (12) remains valid i.e. $\phi_t^{\hat{\mathrm{h}}_0}(\mathrm{x}_0) = \hat{\mathrm{x}}_t$*. It follows that the boundary conditions valid for* $\hat{x}$ *(namely,* $\hat{x}_0 = \mathrm{x}_0, \Pi_l \hat{x}_T = \mathrm{y}$*) are also valid for* $\phi^{\hat{\mathrm{h}}_0}(\mathrm{x}_0)$ *and hence* $\hat{h_0} \in \mathcal{K}_{\mathrm{y}}$*. While we do not if* $\hat{h_0} \in \mathcal{K}_{\mathrm{y}}^{\min}$*, proposition 4 guarantees that every minimizer* $h_0 \in \mathcal{K}_{\mathrm{y}}^{\min}$ *can be found be the above procedure. We thus have the following* ***recipe****:*
(i) *Argue a priori that* $C(\mathrm{h}_0)$ *is invertible (or ignore and check in the end).*
(ii) *Solve Hamiltonian ODEs as boundary value problem, cf. (13),(16). Characterize all solution via the (non-empty!) set*

$$\{\hat{\mathrm{p}}_0 : \mathrm{H}_{t\leftarrow 0}(\mathrm{x}_0,\hat{\mathrm{p}}_0) \equiv (\hat{\mathrm{x}}_t,\hat{\mathrm{p}}_t) \ \textit{satisfies (13),(16)}\ \};$$

(iii) *For each such* $\hat{p}_0$*, compute* $\|\hat{h_0}\|_H^2$ *where* $\hat{h_0}$ *is given by*

$$\hat{\mathrm{h}}_0^i = \int_0^{\cdot} \langle \sigma_i(\hat{\mathrm{x}}_t),\hat{\mathrm{p}}_t\rangle\,dt, \quad i = 1,\dots,m;$$

(iv) *The minimizing* h_0 *are precisely those elements in* $\{\hat{h_0} : \textit{as constructed in (ii),(iii)}\}$ *which minimize energy* $\|\hat{h_0}\|_H^2$*. In particular then,*

$$\Lambda(\mathrm{y}) = \frac{1}{2}\|\mathrm{h}_0\|_H^2.$$

The following proposition is crucial.

Proposition 7 *Under the assumptions of the proposition 4, in particular* $\mathrm{h}_0 \in \mathcal{K}_a^{\min}$ *with associated* $\mathrm{p}_0 = \mathrm{p}_0(\mathrm{h}_0) \in \mathcal{T}_{x_0}^*\mathbb{R}^d$*, the following are equivalent:*
(iii) $\mathrm{h}_0 \in \mathcal{K}_{\mathrm{a}}$ *is a non-degenerate minimum of the energy* $I := \frac{1}{2}\|\cdot\|_H^2$ *restricted to the Hilbert manifold* $\mathcal{K}^a$*; i.e.*

$$I''(\mathrm{h}_0)[k,k] > 0 \qquad \forall 0 \neq k \in H_0 \cong \mathcal{T}_{\mathrm{h}_0}\mathcal{K}_{\mathrm{a}}$$

(iii') x_0 *is* ***non-focal*** *for* $N = (\mathrm{a},\cdot)$ *along* h_0 *in the sense that, with* $(\mathrm{x}_T,\mathrm{p}_T) := H_{0\to T}(\mathrm{x}_0,\mathrm{p}_0(\mathrm{h}_0)) \in \mathcal{T}^*\mathbb{R}^d$*,*

$$\partial_{(\mathfrak{q},\mathfrak{z})}|_{(\mathfrak{q},\mathfrak{z})=(0,0)} \pi H_{0\leftarrow T}(\mathrm{x}_T + (0,\mathfrak{z}), \mathrm{p}_T + (\mathfrak{q},0))$$

is non-degenerate (as $d \times d$ *matrix; here we think of* $(\mathfrak{q},\mathfrak{z}) \in \mathbb{R}^l \times \mathbb{R}^{d-l} \cong \mathbb{R}^d$ *and recall that* π *denotes the projection from* $\mathcal{T}^*\mathbb{R}^d$ *onto* $\mathbb{R}^d$*; in coordinates* $\pi(x,p) = x$*).*

Proof. Let us give a quick proof of (iii') $\Longrightarrow$ (iii) in the Riemannian setting, the general (sub-Riemannian, with drift) case is new and full proof is given in the next section. Since $\mathrm{h}_0 \in \mathcal{K}_a^{\min}$ we know that $I''(\mathrm{h}_0)$ must be positive semi-definite. In particular, the index of h_0, relative to the point-submanifold problem $\mathrm{x}_0 \times N$, is zero. By the *Morse index theorem* [46, 12], there cannot be any focal point along the $(\mathrm{x}_0 \times N)$-geodesic

$$\{\pi H_{T\to t}(\mathrm{x}_T, \mathrm{p}_T) : t \in (0,T]\}.$$

Condition (iii') guarantees that this extends to $t = 0$, i.e. there is no focal point along

$$\{\pi H_{T\to t}(\mathrm{x}_T, \mathrm{p}_T) : t \in [0,T]\}.$$

We can then use [46, lemma 2.9 (b)] to conclude that $I''(\mathrm{h}_0)$ is positive definite. ■

Definition 8 (Condition (ND); generalized $\notin$ cut-locus condition) *We say that* $\{\mathrm{x}_0\} \times N_{\mathrm{a}}$ *where* $N_{\mathrm{a}} := (\mathrm{a}, \cdot) := \left\{\mathrm{x} \in \mathbb{R}^d : \Pi_l \mathrm{x} = \mathrm{a} \in \mathbb{R}^l\right\}$ *satisfies condition (ND) if*
(i) $1 \le \#\mathcal{K}_a^{\min} < \infty$,
(ii) the deterministic Malliavin covariance matrix $C(\mathrm{h})$ *is invertible,* $\forall \mathrm{h} \in \mathcal{K}_a^{\min}$;
(iii) x_0 *is not focal for* N_{a} *along* h, *for any* $h \in \mathcal{K}_a^{\min}$.

When $\sigma_0 \equiv 0$ and $l = d$, i.e. $N = \{\mathrm{y}\}$, and $\#\mathcal{K}_a^{\min} = 1$, condition (ND) says precisely that $(\mathrm{x}_0, \mathrm{y})$ is not contained in the sub-Riemannian cut-locus in the sense of Ben Arous [8]; extending the usual Riemannian meaning. In this sense our (global) condition (ND) is effectively a generalization of the well-known "$\notin$ cut-locus" condition in the context of heat-kernel expansions.

Theorem 9 ***(Small noise)*** *Let* (X^ε) *be the solution process to*

$$d\mathrm{X}_t^\varepsilon = b(\varepsilon, \mathrm{X}_t^\varepsilon)\, dt + \varepsilon\sigma(\mathrm{X}_t^\varepsilon)\, dW_t, \quad \text{with } \mathrm{X}_0^\varepsilon = \mathrm{x}_0^\varepsilon \in \mathbb{R}^d.$$

Assume $b(\varepsilon, \cdot) \to \sigma_0(\cdot)$ *in the sense of (5), (6), and* $\mathrm{X}_0^\varepsilon \equiv \mathrm{x}_0^\varepsilon \to \mathrm{x}_0$ *as* $\varepsilon \to 0$ *in the sense of (7). Assume the weak Hörmander condition (H) at* $\mathrm{x}_0 \in \mathbb{R}^d$. *Fix* $\mathrm{y} \in \mathbb{R}^l$, $N_{\mathrm{y}} := (\mathrm{y}, \cdot)$ *and assume that* $\{\mathrm{x}_0\} \times N_{\mathrm{y}}$ *satisfies (ND), i.e. the generalized* $\notin$ *cut-locus condition (in particular then,* $\#\mathcal{K}_{\mathrm{y}}^{\min} \ge 1$*). Then the energy*

$$\Lambda(\mathrm{y}) = \inf\left\{\frac{1}{2}\|\mathrm{h}\|_H^2 : \mathrm{h} \in \mathcal{K}_{\mathrm{y}}\right\} = \frac{1}{2}\|\mathrm{h}_0\|_H^2.$$

is smooth in a neighbourhood of y *provided* $\#\mathcal{K}_{\mathrm{y}}^{\min} = 1$*; otherwise i.e. when* $\#\mathcal{K}_{\mathrm{y}}^{\min} > 1$, *we assume so.*[10]
Fix x_0, y *and* $T > 0$. *Then there exists* $c_0 = c_0(\mathrm{x}_0, \mathrm{y}, T) > 0$ *such that*

$$\mathrm{Y}_T^\varepsilon = \Pi_l \mathrm{X}_T^\varepsilon = \left(X_T^{\varepsilon,1}, \ldots, X_T^{\varepsilon,l}\right), \quad 1 \le l \le d$$

admits a density with expansion

$$f_\varepsilon(\mathrm{y}, T) = e^{-\frac{\Lambda(\mathrm{y})}{\varepsilon^2}} e^{\frac{\max\left\{\Lambda'(\mathrm{y})\cdot \hat{\mathrm{Y}}_T(\mathrm{h}_0) : \mathrm{h}_0 \in \mathcal{K}_{\mathrm{a}}^{\min}\right\}}{\varepsilon}} \varepsilon^{-l}\left(c_0 + O(\varepsilon)\right) \text{ as } \varepsilon \downarrow 0.$$

[10] It will not be true in general, when $\#\mathcal{K}_{\mathrm{y}}^{\min} > 1$, that $\Lambda(\mathrm{y})$ is automatically smooth near y. To wit consider, $\mathcal{K}_{\mathrm{y}}^{\min} = \{\mathrm{h}_0(\mathrm{y}), \tilde{\mathrm{h}}_0(\mathrm{y})\}$. Then $\Lambda(\mathrm{y}) = \min\left(\frac{1}{2}\|\mathrm{h}_0(\mathrm{y})\|_H^2, \frac{1}{2}\|\tilde{\mathrm{h}}_0(\mathrm{y})\|_H^2\right)$ and even if $\|\mathrm{h}_0(\cdot)\|_H^2$ and $\|\tilde{\mathrm{h}}_0(\cdot)\|_H^2$ are smooth near y, this need not be the case for the minimum.

Here $\hat{\mathrm{Y}} = \hat{\mathrm{Y}}(\mathrm{h}_0) = \left(\hat{Y}^1, \ldots, \hat{Y}^l\right)$ *is the projection,* $\hat{\mathrm{Y}} = \Pi_l \hat{\mathrm{X}}$*, of the solution to the following (ordinary) differential equation*

$$d\hat{\mathrm{X}}_t = \left(\partial_x b\left(0, \phi_t^{\mathrm{h}_0}(\mathrm{x}_0)\right) + \partial_x \sigma(\phi_t^{\mathrm{h}_0}(\mathrm{x}_0))\dot{h}_0(t)\right)\hat{\mathrm{X}}_t dt + \partial_\varepsilon b\left(0, \phi_t^{\mathrm{h}_0}(\mathrm{x}_0)\right) dt, \tag{18}$$
$$\hat{\mathrm{X}}_0 = \partial_\varepsilon|_{\varepsilon=0}\, \mathrm{x}_0^\varepsilon.$$

Remark 10 *The assumption* $\mathcal{K}_{\mathrm{y}} \neq \varnothing$*, implicit through* $\#\mathcal{K}_{\mathrm{y}}^{\min} = 1$ *in the statement of the above theorem, is known to be necessary for the existence of a positive density; in presence of (H) and invertibility of* $C(\mathrm{h})$*, some* $\mathrm{h} \in \mathcal{K}_{\mathrm{y}}$ *it is actually sufficient; [11]. The strong Hörmander condtion at all points (H1) is well-known to ensure* $\mathcal{K}_{\mathrm{y}} \neq \varnothing$*; a less well-known and subtle condition of weak-Hörmander type is given in [33].*

Remark 11 *When applied to small time expansions, the weak Hörmander condition in the above theorem automatically reduces to the strong Hörmander condition at* x_0*; indeed, the "drift" vector field in the weak condition will be the limit of* ε^2 *times the original drift vector fields; plainly this is zero and therefore does not figure in the span.*

Proof. Assume $\#\mathcal{K}_a^{\min} = 1$ and see remark 12 below for the reduction of $\#\mathcal{K}_a^{\min} < \infty$ to this case. The basic remark is that $f_\varepsilon(\mathrm{y}, T)$ is the Fourier inverse of its characteristic function,

$$\mathbb{E}\left[\exp\left(i\xi \cdot \mathrm{Y}_T^\varepsilon\right)\right] = \mathbb{E}\left[\exp\left(i(\xi, 0) \cdot X_T^\varepsilon\right)\right]$$

where we write $(\xi, 0) = \left(\xi^1, \ldots, \xi^l, 0, \ldots, 0\right) \in \mathbb{R}^d$. In other words, it suffices to *restrict* the c.f. of X_T^ε, the full (Markovian) process evaluated at time T, to obtain the c.f. of Y_T^ε. The density is then obtained by Fourier-inversion. When X_T^ε is affine the c.f. is analytically described by ODEs; (approximate) saddle points are easy to compute and the Fourier inversion - after shifting the contour through the saddle point - becomes a finite-dimensional Laplace method which leads to the desired expansion of $f_\varepsilon(\mathrm{y}, T)$; in essence, this approach was carried out by Friz et al. in [24]. In our present situation, of course, X does not enjoy any affine structure, but - following Ben Arous [8], who considers the "point-to-point" case $l = d$; a similar approach works and ultimately boils down to Laplace method on Wiener space [9]. The differences to the setting of [8], aside from (i) allowing for $l < d$, is that (ii) our drift-term does not vanish of order ε^2 (which is typical when aiming for short time asymptotics; cf. also proposition 15 below) and (iii) that the starting point is allowed to depend on ε. In fact, (ii),(iii) are responsible for the additional exponential $\exp\{(...)/\varepsilon\}$ factor in our expansion (Such a factor was already seen in the general context of Laplace method on Wiener space [9].) Also, (ii) implies that the limiting vector field $\sigma_0 = \lim_\varepsilon b(\varepsilon, \cdot)$ affects the leading order behaviour in that the energy $\Lambda(\mathrm{y})$ has no geometric interpretation as square of some (sub)Riemannian point-subspace distance. In particular, if we want to implement the strategy of [8] we are forced to revisit the meaning of all geometric concepts (cut-locus, geodesics, conjugate points ...) upon which the work [8] is based. The key observation is that essentially all geometric concepts channel through the (non-geometric, but infinite-dimensional) condition (iii) of proposition 7 into the application of Laplace's method. Now, the whole point of proposition 7 was to provide check-able conditions for x_0, y to satisfy (iii). Having made these part of our assumption we are in fact ready to proceed along the lines of Ben Arous [8].

Fix y and note that for any C^∞-bounded function $\mathrm{z} \mapsto F(\mathrm{z})$ on $\mathbb{R}^l$, by Fourier inversion,

$$f_\varepsilon(\mathrm{y},T)e^{-F(\mathrm{y})/\varepsilon^2} = \frac{1}{(2\pi)^l}\int_{\mathbb{R}^l}\mathbb{E}\left[\exp\left(i\xi\cdot(\mathrm{Y}_T^\varepsilon-\mathrm{y})-\frac{F(\mathrm{Y}_T^\varepsilon)}{\varepsilon^2}\right)\right]d\xi \tag{19}$$

$$= \frac{1}{(2\pi\varepsilon)^l}\int_{\mathbb{R}^l}\mathbb{E}\left[\exp\left(i\zeta\cdot\left(\frac{\mathrm{Y}_T^\varepsilon-\mathrm{y}}{\varepsilon}\right)-\frac{F(\mathrm{Y}_T^\varepsilon)}{\varepsilon^2}\right)\right]d\zeta.$$

$$= \frac{1}{(2\pi\varepsilon)^l}\int_{\mathbb{R}^l}\mathbb{E}\left[\exp\left(i\left(\zeta,0\right)\cdot\left(\frac{\mathrm{X}_T^\varepsilon-(\mathrm{y},0)}{\varepsilon}\right)\right)e^{-\frac{F(\Pi_l\mathrm{X}_T^\varepsilon)}{\varepsilon^2}}\right]d\zeta. \tag{20}$$

In particular, the last integrand can be computed, as asymptotic expansion in ε for fixed ζ, by Laplace method in Wiener space, cf. [8], [9], based on the full (Markovian) process X_T^ε. We pick F (for fixed y) such that $F(\cdot)+\Lambda(\cdot)$ has minimum at y, i.e.

$$\Lambda(\mathrm{y}) = \inf\left\{F(\mathrm{z})+\Lambda(\mathrm{z}) : \mathrm{z}\in\mathbb{R}^l\right\}$$

and such that this minimum is non-degenerate; a natural candidate for $F(\mathrm{z})$ would then be given (at least for z near y) by

$$\mathrm{z}\mapsto\lambda\left|\mathrm{z}-\mathrm{y}\right|^2-\Lambda(\mathrm{z}),\ \text{some}\ \lambda>0;$$
$$\text{or}\ \mathrm{z}\mapsto\lambda\left|\mathrm{z}-\mathrm{y}\right|^2-[\Lambda(\mathrm{z})-\Lambda(\mathrm{y})],$$

since adding constants is irrelevant here (recall that y is kept fix). The trouble with the above candidate is their potential lack of (global) smoothness of Λ; even in the classical Riemannian setting Λ will not be smooth at the cut-locus. On the other hand, $\Lambda(\cdot)$ is smooth near y in case $\#\mathcal{K}_a^{\min}=1$; this is seen exactly as in [14, Thm 1.26]. (In the case $1<\#\mathcal{K}_a^{\min}$, smoothness of $\Lambda(\cdot)$ near y was in fact part of our assumptions.) It is thus natural to localize the above candidates around y which leads us to define F, at least in a neighbourhood of y, by

$$F(\mathrm{z}) = \lambda\left|\mathrm{z}-\mathrm{y}\right|^2-\left[\frac{\partial}{\partial\mathrm{y}}\Lambda(\mathrm{y})(\mathrm{y}-\mathrm{z})+\frac{1}{2}\frac{\partial^2}{\partial\mathrm{y}^2}\Lambda(\mathrm{y})(\mathrm{y}-\mathrm{z},\mathrm{y}-\mathrm{z})\right];$$

a routine modification of F, away from y, then guarantees C^∞-boundedness of F. (Since $F(\mathrm{y})=0$ with this last choice of F, the l.h.s. of (19) is actually precisely $f_\varepsilon(\mathrm{y},T)$) Non-degeneracy of the minimum y of F entails that the functional $H\ni\mathrm{h}\mapsto F\left(\phi_T^{\mathrm{h}}(\mathrm{x}_0)\right)+\frac{1}{2}\left\|\mathrm{h}\right\|_H^2$ has a non-degenerate minimum at $\mathrm{h}_0\in H$. (The argument is identical to [8, Thm 2.6] and makes crucial use of proposition 7.) The Laplace method is then applicable: we replace εdW by $\varepsilon dW+d\mathrm{h}_0$ in (4) and call the resulting diffusion process Z^ε. The integrand of (20) can then be expressed in terms with X^ε replaced by Z^ε; of course at the price of including the Girsanov factor

$$\mathcal{G} := \exp\left(-\frac{1}{\varepsilon}\int_0^T\dot{\mathrm{h}}_0(t)\,dW_t-\frac{1}{2\varepsilon^2}\int_0^T\left|\dot{\mathrm{h}}_0(t)\right|^2dt\right)=\exp\left(-\frac{1}{\varepsilon}\int_0^T\dot{\mathrm{h}}_0(t)\,dW_t-\frac{1}{\varepsilon^2}\Lambda(\mathrm{y})\right).$$

A stochastic Taylor expansion of Z^ε, noting right away that

$$F(\Pi_l\mathrm{Z}_T^\varepsilon)\,|_{\varepsilon=0}=F\left(\Pi_l\phi_T^h(\mathrm{x}_0)\right)=F(\mathrm{y})=0,$$

then leads to (cf. [9, Lemme 1.43])

$$\begin{aligned}
&\exp\left(-\frac{1}{\varepsilon^2}F\left(\Pi_l \mathrm{Z}_T^\varepsilon\right)\right)\\
&=\exp\left(-\frac{1}{\varepsilon^2}\left[F\left(\mathrm{y}\right)-\varepsilon\int_0^T \dot{h}_0\left(t\right)dW_t-\varepsilon\Pi_l\hat{\mathrm{X}}_T\cdot\partial_{\mathrm{y}}\Lambda\left(\mathrm{y}\right)+O\left(\varepsilon^2\right)\right]\right)\\
&=\exp\left(\frac{1}{\varepsilon}\int_0^T \dot{\mathrm{h}}_0\left(t\right)dW_t+\frac{1}{\varepsilon}\left(\hat{\mathrm{Y}}_T\right)\cdot\partial_{\mathrm{y}}\Lambda\left(\mathrm{y}\right)+O\left(1\right)\right). \qquad (21)
\end{aligned}$$

Putting things together, we have, using $F\left(\mathrm{y}\right)=0$, and noting cancellation of $\int_0^T \dot{\mathrm{h}}_0\left(t\right)dW_t$ in (21) with the identical term in the Girsanov factor $\mathcal{G}$,

$$\begin{aligned}
f_\varepsilon(\mathrm{y},T)&=\frac{1}{\left(2\pi\varepsilon\right)^l}\int_{\mathbb{R}^l}\mathbb{E}\left[\mathcal{G}\times\exp\left(i\left(\zeta,0\right)\cdot\left(\frac{\mathrm{Z}_T^\varepsilon-\left(\mathrm{y},0\right)}{\varepsilon}\right)\right)e^{-\frac{F\left(\Pi_l\mathrm{Z}_T^\varepsilon\right)}{\varepsilon^2}}\right]d\zeta\\
&=\frac{1}{\varepsilon^l}\exp\left(-\frac{1}{\varepsilon^2}\Lambda\left(\mathrm{y}\right)\right)\exp\left(\frac{1}{\varepsilon}\left(\hat{\mathrm{Y}}_T\right)\cdot\partial_{\mathrm{y}}\Lambda\left(\mathrm{y}\right)\right)\\
&\quad\times\underbrace{\frac{1}{\left(2\pi\right)^l}\int_{\mathbb{R}^l}\mathbb{E}\left[\exp\left(i\left(\zeta,0\right)\cdot\left(\frac{\mathrm{Z}_T^\varepsilon-\left(\mathrm{y},0\right)}{\varepsilon}\right)\right)\exp\left(O\left(1\right)\right)\right]d\zeta}_{=:c_0} \qquad (22)
\end{aligned}$$

where $O\left(1\right)$ denotes the term, bounded as $\varepsilon\downarrow 0$, from (21). What is left to show, of course, is that c_0, i.e. the final factor in the above expression, is indeed a strictly positive and finite real number. But since our analysis is based on the full Markovian process X_T (resp. Z_T^ε after change of measure), the arguments of [8, Lemme (3.25)] apply with essentially no changes. In particular, one uses large deviations as in [8, Lemme (3.25)]) and, crucially, non-degeneracy of the minimizer $\mathrm{h}_0\in H$, guaranteed by proposition 7. Finally, integrating the asymptotic expansion with respect to $\zeta\in\mathbb{R}^l$ is justified using the estimates of [8, Lemme 3.48], obtained using Malliavin calculus techniques. At last one sees $c_0>0$, as in [8, p. 330]. ■

Remark 12 (Finitely many multiple minimizers) *The case* $1<\#\mathcal{K}_a^{\min}<\infty\in\{2,3,\dots\}$ *is handled as in [9]. If*

$$\mathcal{K}_a^{\min}=\left\{h_0^{(1)},\dots,h_0^{(n)}\right\},$$

and invertibility of the Malliavin matrix as well as non-focality holds along each of these, the expansion for $f_\varepsilon\left(\mathrm{y},T\right)$ *as given in theorem 9 remains valid. Indeed, after localization around each of these n minimizers,*

$$\begin{aligned}
f_\varepsilon\left(y,T\right)&=\left(\sum_{\mathrm{h}_0\in\mathcal{K}_{\mathrm{a}}^{\min}}e^{-\frac{\Lambda(y)}{\varepsilon^2}}e^{\frac{\Lambda'(\mathrm{y})\cdot\hat{\mathrm{Y}}_T(\mathrm{h}_0)}{\varepsilon}}\varepsilon^{-l}c_0\left(\mathrm{h}_0\right)\right)\left(1+O\left(\varepsilon\right)\right)\\
&\sim\textit{(const)}\ e^{-\frac{\Lambda(y)}{\varepsilon^2}}e^{\max\left\{\frac{\Lambda'(\mathrm{y})\cdot\hat{\mathrm{Y}}_T(\mathrm{h}_0)}{\varepsilon}:\mathrm{h}_0\in\mathcal{K}_{\mathrm{a}}^{\min}\right\}}\varepsilon^{-l}
\end{aligned}$$

where $\hat{\mathrm{Y}}_T\left(\mathrm{h}_0\right)$ *denotes the solution of (18).*

Remark 13 (Localization) *The assumptions on the coefficients b, σ in theorem 9 (smooth, bounded with bounded derivatives of all orders) are typical in this context (cf. Ben Arous [8, 9] for instance) but rarely met in practical examples from finance. This difficulty can be resolved by a suitable localization which we now outline.Set $\tau_R := \inf\left\{t \in [0,T] : \sup_{s\in[0,t]} |X_s^\varepsilon| \geq R\right\}$ and assume*

$$\mathbb{P}\left[\tau_R \leq T\right] \lesssim e^{-J_R/\varepsilon^2} \text{ as } \varepsilon \downarrow 0$$

with $J_R \to \infty$ as $R \to \infty$ by this we mean, more precisely,

$$\lim_{R\to\infty} \limsup_{\varepsilon\to 0} \varepsilon^2 \log \mathbb{P}\left[\tau_R \leq T\right] = -\infty. \tag{23}$$

In that case, we can pick R large enough so that $\Lambda(y) < J_R$, uniformly for ε near $0+$, and can expect that the behaviour beyond some big ball of radius R will not influence the expansion. In particular, if the coefficients b, σ are smooth, but fail to be bounded resp. have bounded derivatives, we can modify them outside a ball of radius R such as to have this property; call $\tilde{b}, \tilde{\sigma}$ these new coefficients and $\tilde{X}^\varepsilon$ the associated diffusion. To illustrate the localization, consider $l = 1$, i.e. $Y_T^\varepsilon \equiv X_T^{\varepsilon,1}$, and the distribution function for Y_T^ε. Clearly, one has the two-sided estimates

$$\mathbb{P}\left[Y_T^\varepsilon \geq y; \tau_R > T\right] \leq \mathbb{P}\left[Y_T^\varepsilon \geq y\right] \leq \mathbb{P}\left[Y_T^\varepsilon \geq y; \tau_R > T\right] + \mathbb{P}\left[\tau_R \leq T\right],$$

and similar for $\tilde{Y}_T^\varepsilon \equiv \tilde{X}_T^{\varepsilon,1}$. Since $\mathbb{P}\left[Y_T^\varepsilon \geq y; \tau_R > T\right] = \mathbb{P}\left[\tilde{Y}_T^\varepsilon \geq y; \tau_R > T\right]$ it then follows

$$\left|\mathbb{P}\left[Y_T^\varepsilon \geq y\right] - \mathbb{P}\left[\tilde{Y}_T^\varepsilon \geq y\right]\right| \leq \mathbb{P}\left[\tau_R \leq T\right] \lesssim e^{-J_R/\varepsilon^2}.$$

In particular, any expansion for $\tilde{Y}_T^\varepsilon$ of the form

$$\mathbb{P}\left[\tilde{Y}_T^\varepsilon \geq y\right] = e^{-c_1/\varepsilon^2} e^{c_2/\varepsilon^2} \varepsilon^{-l} c_0 \left(1 + O(\varepsilon)\right)$$

leads, upon taking R large enough so that $J_R > c_1$, to the same expansion for $\mathbb{P}\left[Y_T^\varepsilon \geq y\right]$. With more work of routine type, this localization also be employed for the density expansion in theorem 9.

2.1 Corollary on tail expansions

We have the following application to tail behaviour of, say, the first component (i.e. $l = 1$ here) of a diffusion processes at a fixed time T. The scaling assumption below is met in a number of stochastic volatility models.

Corollary 14 ***(Tail behaviour)*** *Assume $\mathrm{x}_0^\varepsilon \to 0 \in \mathbb{R}^d$ as $\varepsilon \to 0$ and some diffusion process X^ε, started at x_0^ε, satisfies the assumptions of theorem 9 with $\mathrm{x}_0 = 0$ and $N = (1, \cdot) \subset \mathbb{R} \times \mathbb{R}^{d-1}$; in particular, $\{0\} \times (1, \cdot)$ is assumed to satisfy condition (ND). Assume also **θ-scaling** by which we mean the scaling relation*

$$Y_T^\varepsilon \overset{(law)}{=} \varepsilon^\theta Y_T \text{ where } \quad Y \equiv \Pi_1 X$$

for some $\theta \geq 1$. Then the probability density function of Y_T has the expansion

$$f(y) = e^{-c_1 y^{\frac{2}{\theta}}} e^{c_2 y^{\frac{1}{\theta}}} y^{\frac{1}{\theta}-1} \left(\alpha_0 + O\left(1/y^{1/\theta}\right)\right) \text{ as } y \to \infty \tag{24}$$

where

$$c_1 = \Lambda(1)$$
$$c_2 = \hat{Y}_T \Lambda'(1) = \frac{2\hat{Y}_T}{\theta}\Lambda(1)$$

In particular, when $\theta = 1$ we have a Gaussian tail behaviour of the precise form

$$f(y) = e^{-\Lambda(1)y^2} e^{2\hat{Y}_T \Lambda(1) y} \left(c_0 + O(1/y)\right);$$

while $\theta = 2$ leads to the exponential tail of the precise form

$$f(y) = e^{-\Lambda(1)y} e^{\hat{Y}_T \Lambda'(1)\sqrt{y}} y^{-1/2} \left(c_0 + O(1/\sqrt{y})\right).$$

Proof. Let f_ε denote the density of Y_T^ε. Since $f\left(y/\varepsilon^\theta\right) = \varepsilon^\theta f_\varepsilon(y)$ we can take $y = 1$ and $\varepsilon^\theta = y^{-1}$ in the theorem below. Another observation is that the assumed scaling implies

$$\Lambda_0(y) = y^{2/\theta}\Lambda_0(1)$$

and hence $\Lambda_0'(1) = \frac{2}{\theta}\Lambda_0(1)$. The rest is obvious. ■

2.2 Corollary on short time expansions

Finally, we have the following application to short time asymptotics. Note that for $l < d$, the projection of X is non-Markovian and there is no Fokker-Planck equation that describes the evolution of f. In particular, there is no direct PDE approach that leads to the expansion below.

Corollary 15 ***(Short time)*** *Consider $dX_t = b(X_t)\,dt + \sigma(X_t)\,dW$, started at $X_0 = x_0 \in \mathbb{R}^d$, with C^∞-bounded vector fields such that the strong Hörmander condition holds,*

$$\forall x \in \mathbb{R}^d : Lie\left[\sigma_1, \ldots, \sigma_m\right]|_x = \mathcal{T}_x \mathbb{R}^d. \tag{H1}$$

For fixed $l \in \{1, \ldots, d\}$ assume $\{\mathrm{x}_0\} \times N_\mathrm{y}$, where $N_\mathrm{y} := (\mathrm{y}, \cdot)$ for some $\mathrm{y} \in \mathbb{R}^l$, satisfies condition (ND). Let $f(t, \cdot) = f(t, y)$ be the density of $Y_t = \left(X_t^1, \ldots, X_t^l\right)$. Then

$$f(t, y) \sim (const)\frac{1}{t^{l/2}} \exp\left(-\frac{d^2(x_0, y)}{2t}\right) \text{ as } t \downarrow 0$$

where $d(x_0, y)$ is the sub-Riemannian distance, based on $(\sigma_1, \ldots, \sigma_m)$, from the point x_0 to the affine subspace N_y.

Proof. After Brownian scaling, we apply the theorem with $T = 1, \varepsilon^2 = t$ so that

$$b(\varepsilon, \cdot) = \varepsilon^2 b(\cdot) \to \sigma_0(\cdot) \equiv 0;$$

which explains why there is no drift vector field in the present Hörmander condition H1. Also $x_0^\varepsilon = x_0$ here. The identification of the energy with $1/2$ times the square of the sub-Riemannian (or: control - , Carnot-Caratheodory -) distance from x to Σ_y is classical. At last, the unique ODE solution to (18) is then given by $\hat{Y} \equiv 0$ and there is no $\exp\{(...)/\varepsilon\}$ factor. ■

3 Non-focality and infinite-dimensional non-degeneracy

In the present section only, we write h (rather than h_0) for a fixed element in $K_{\mathrm{a}}^{\min}$.

Recall $\mathcal{T}_{\mathrm{h}}\mathcal{K}_{\mathrm{a}} = \ker D\left(\Pi_l \phi_T\right)(\mathrm{h}) =: H_0$. Since $\mathrm{h} \in \mathcal{K}_{\mathrm{a}}^{\min}$, it is critical in the sense that

$$I'(\mathrm{h}) = DI(h) = 0 \text{ on } \mathcal{T}_{\mathrm{h}}\mathcal{K}_{\mathrm{a}} = H_0.$$

Also recall $\mathrm{x}_T = \phi_{T\leftarrow 0}^{\mathrm{h}}(\mathrm{x}_0) = \phi_T(\mathrm{h})$, notation used when x_0 is fixed. Given

$$q \in \operatorname{span}\left\{dx^1|_{\mathrm{x}_T}, \ldots, dx^l|_{\mathrm{x}_T}\right\}$$

with $1 \leq l \leq d$ we shall write[11]

$$(q,0) \in \operatorname{span}\left\{dx^1|_{\mathrm{x}_T}, \ldots, dx^d|_{\mathrm{x}_T}\right\} = \mathcal{T}^*_{\mathrm{x}_T}\mathbb{R}^d$$

for q "viewed" as element in $\mathcal{T}^*_{\mathrm{x}_T}\mathbb{R}^d$. We can describe H_0 as the set of those $k = \left(k^1, \ldots, k^m\right) \in H$ such that, for any $q \in \operatorname{span}\left\{dx^1|_{\mathrm{x}_T}, \ldots, dx^l|_{\mathrm{x}_T}\right\}$,

$$\int_0^T \left\langle (q,0), \Phi_{T\leftarrow t}^{\mathrm{h}} \sigma_i\left(\phi_{t\leftarrow T}^{\mathrm{h}}(\mathrm{x}_T)\right)\right\rangle \dot{k}_t^i dt = 0;$$

where, of course, $\left\{x^1, \ldots, x^d\right\}$ denotes the standard coordinate chart of $\mathbb{R}^d$ and we tacitly use Einstein's summation convention.We recall our standing assumption that the deterministic Malliavin covariance matrix $C(\mathrm{h})$ is invertible.

Lemma 16 *The linear map $\tilde{\rho}_{\mathrm{h}} : \operatorname{span}\left\{dx^1|_{\mathrm{x}_T}, \ldots, dx^l|_{\mathrm{x}_T}\right\} \to H$ given by*

$$\tilde{\rho}_{\mathrm{h}}(q) := \begin{pmatrix} \int_0^{\cdot} \left\langle (q,0), \Phi_{T\leftarrow t}^{\mathrm{h}} \sigma_1\left(\phi_{t\leftarrow T}^{\mathrm{h}}(\mathrm{x}_T)\right)\right\rangle dt \\ \cdots \\ \int_0^{\cdot} \left\langle (q,0), \Phi_{T\leftarrow t}^{\mathrm{h}} \sigma_m\left(\phi_{t\leftarrow T}^{\mathrm{h}}(\mathrm{x}_T)\right)\right\rangle dt \end{pmatrix}$$

for $i = 1, \ldots, m$ and $t \in [0,T]$ is one-one with range $H_0^{\perp}$.

Proof. Since H_0 is the set of those $k \in H$ such that, for any $q \in \operatorname{span}\left\{dx^1|_{\mathrm{x}_T}, \ldots, dx^l|_{\mathrm{x}_T}\right\}$,

$$\int_0^T \left\langle (q,0), \Phi_{T\leftarrow t}^{\mathrm{h}} \sigma_i\left(\phi_{t\leftarrow T}^{\mathrm{h}}(\mathrm{x}_T)\right)\right\rangle \dot{k}_t^i dt = 0$$

we see that H_0 is the orthogonal complement in H of

$$\left\{\tilde{\rho}_{\mathrm{h}}(q) : q \in \operatorname{span}\left\{dx^1|_{\mathrm{x}_T}, \ldots, dx^l|_{\mathrm{x}_T}\right\}\right\};$$

i.e. $H_0^{\perp}$ is the range of $\tilde{\rho}_{\mathrm{h}}$. Invertibility of the deterministic Malliavin matrix (along h) then implies $\ker \tilde{\rho}_{\mathrm{h}} = \{0\}$ which shows that $\tilde{\rho}_{\mathrm{h}}$ is one-one (and also that $H_0^{\perp}$ has dimension l). ■

[11] In fancy notation, $(q,0) = (\Pi_l)_*^* q$ where $(\Pi_l)_*^*$ is the adjoint of $(\Pi_l)_* : T_{\mathrm{x}_T}\mathbb{R}^d \to T_{\Pi_l \mathrm{x}_T}\mathbb{R}^l$, the differential of the projection map $\Pi_l : \left(x^1, \ldots, x^d\right) \to \left(x^1, \ldots, x^l\right)$

Lemma 17 *For each minimizer* $\mathrm{h} \in \mathcal{K}_{\mathrm{a}}^{\min}$*, there exists a unique* $q = q(\mathrm{h}) \in \operatorname{span}\left\{dx^1|_{\mathrm{x}_T}, \ldots, dx^l|_{\mathrm{x}_T}\right\}$ *s.t.*

$$\mathrm{h} = D\phi_T(\mathrm{h})^* [(q, 0)].$$

(Recall $D\phi_T(\mathrm{h}) : H \to \mathcal{T}_{\mathrm{x}_T}\mathbb{R}^d$*; its adjoint then maps* $\mathcal{T}^*_{\mathrm{x}_T}\mathbb{R}^d \to H$ *where we identify* H^* *with* H*.)*

Proof. By assumption, h is a minimizer, and so its differential $I'(\mathrm{h})$ is 0 on $T_{\mathrm{h}}\mathcal{K}_a \equiv H_0$. It follows that for every $k \in H_0$,

$$\langle dI(\mathrm{h}), k\rangle = \int_0^T \sum_{i=1}^m \dot{h}^i_t \dot{k}^i_t dt = 0$$

so that h is in the orthogonal complement of H_0. It follows that there exists a (unique, thanks to invertibility of the deterministic Malliavin matrix along h)

$$q = q(\mathrm{h}) \in \operatorname{span}\left\{dx^1|_{\mathrm{x}_T}, \ldots, dx^l|_{\mathrm{x}_T}\right\}$$

such that $\mathrm{h} = \hat{\rho}_{\mathrm{h}}(q)$. It follows that

$$\dot{h}^i_t = \left\langle (q, 0), \Phi^{\mathrm{h}}_{T\leftarrow t}\sigma_i\left(\phi^{\mathrm{h}}_{t\leftarrow T}(\mathrm{x}_T)\right)\right\rangle.$$

It remains to see that, for any $k \in H$,

$$\langle k, \mathrm{h}\rangle_H = \left\langle k, D\phi_T(\mathrm{h})^* [(q, 0)]\right\rangle_H = \langle (q, 0), D\phi_T(\mathrm{h})[k]\rangle,$$

but this follows immediately from the computation

$$\begin{aligned}
\langle k, \mathrm{h}\rangle_H &= \langle k, \hat{\rho}_{\mathrm{h}}(q)\rangle_H \\
&= \int_0^T \dot{k}^i_t \left\langle (q, 0), \Phi^{\mathrm{h}}_{T\leftarrow t}\sigma_i\left(\phi^{\mathrm{h}}_{t\leftarrow T}(\mathrm{x}_T)\right)\right\rangle dt \\
&= \left\langle (q, 0), \int_0^T \Phi^{\mathrm{h}}_{T\leftarrow t}\sigma_i\left(\phi^{\mathrm{h}}_{t\leftarrow T}(\mathrm{x}_T)\right)\dot{k}^i_t dt\right\rangle.
\end{aligned}$$

■

Lemma 18 $I''(\mathrm{h})$ *is a bilinear form on* H_0 *given by*

$$\begin{aligned}
I''(\mathrm{h})[k, l] &= \langle k, l\rangle_H - \left\langle (q(\mathrm{h}), 0), D^2\phi_T(h)[k, l]\right\rangle \\
&= \langle k, l\rangle_H - \left(q(\mathrm{h}), D^2\psi_T(h)[k, l]\right)
\end{aligned}$$

where $(q(\mathrm{h}), 0) \in \mathcal{T}^*_{\mathrm{x}_T}\mathbb{R}^d$ *was constructed lemma 17. In particular, an element* $k \in H_0$ *is in the null-space* $\mathcal{N}(\mathrm{h})$ *of* $I''(\mathrm{h})$*,*

$$\begin{aligned}
k \in \mathcal{N}(\mathrm{h}) &:= \{k \in H_0 : I''(\mathrm{h})[k, k] = 0\} \\
&= \{k \in H_0 : I''(\mathrm{h})[k, \cdot] \equiv 0 \text{ on } H_0\}.
\end{aligned}$$

if and only if (identifying H^* *with* H*)*

$$\langle k, \cdot\rangle_H - \left(p_T, D^2\psi(h)[k, \cdot]\right) \in H_0^{\perp}.$$

Proof. Take a smooth curve $c:(-\varepsilon,\varepsilon)\to K_a$ s.t. $c(0)=\mathrm{h}, \dot{c}(0)=k$. Then

$$I''(h)[k,k]=|k|_H^2+\langle \mathrm{h}, \ddot{c}(0)\rangle .$$

From the previous lemma

$$\begin{aligned} I''(\mathrm{h})[k,k] &= |k|_H^2+\left\langle (q,0), D\phi_T(\mathrm{h})\left[\ddot{c}(0)\right]\right\rangle \\ &= |k|_H^2+\left\langle q, D\psi_T(\mathrm{h})\left[\ddot{c}(0)\right]\right\rangle \end{aligned}$$

On the other hand, since $\psi_T(c(t))=\Pi_l\phi_T(c(t))\equiv a$ for $t\in(-\varepsilon,\varepsilon)$ we have

$$\begin{aligned} 0 &= \frac{d^2}{dt^2}\psi_T(c(t))|_{t=0} \\ &= \frac{d}{dt}D\psi_T(c(t))[\dot{c}(t)]|_{t=0} \\ &= D^2\psi_T(\mathrm{h})[k,k]+D\psi_T(\mathrm{h})\left[\ddot{c}(0)\right] \end{aligned}$$

and hence

$$\begin{aligned} I''(\mathrm{h})[k,k] &= |k|_H^2-\left\langle q, D^2\psi_T(\mathrm{h})[k,k]\right\rangle \\ &= |k|_H^2-\left\langle (q,0), D^2\phi_T(\mathrm{h})[k,k]\right\rangle . \end{aligned}$$

The characterization of elements in $\mathcal{N}(\mathrm{h})$ is then clear. Let us just remark that $\mathcal{N}(\mathrm{h})$ is indeed equal to the space $\{k\in H_0 : I''(\mathrm{h})[k,\cdot]\equiv 0 \text{ on } H_0\}$ as is easily seen from the fact that $I''(\mathrm{h})$ is positive semi-definite, since h is (by assumption) a minimizer. ■

If U is a vector field on $\mathbb{R}^d$ we define the push-forward, under the diffeomorphism $\left(\phi^{\mathrm{h}}_{s\leftarrow T}\right)^{-1}$, by

$$\left(\phi^{\mathrm{h}}_{s\leftarrow T}\right)^{-1}_* U(z) := \left(\Phi^{\mathrm{h}}_{s\leftarrow T}\right)^{-1} U\left(\phi^{\mathrm{h}}_{s\leftarrow T}(z)\right)\in \mathcal{T}_z\mathbb{R}^d$$

We shall then need the following known formula, cf. [14, 1.21] combined with trivial time reparameterization $t\rightsquigarrow T-t$;

$$D\left(\phi^{\mathrm{h}}_{t\leftarrow T}\right)^{-1}_* U(z)[k]=\int_t^T\left[\left(\phi^{\mathrm{h}}_{s\leftarrow T}\right)^{-1}_*\sigma_j, \left(\phi^{\mathrm{h}}_{t\leftarrow T}\right)^{-1}_* U\right](z)\,\dot{k}^j_s ds. \tag{25}$$

Lemma 19 *For $k,l\in H$ we have, with $\mathrm{x}_T=\phi_T(\mathrm{h})$,*

$$\begin{aligned} D^2\phi_T(\mathrm{h})[k,l] &= \int_0^T\int_t^T\left[\left(\phi^{\mathrm{h}}_{s\leftarrow T}\right)^{-1}_*\sigma_j, \left(\phi^{\mathrm{h}}_{t\leftarrow T}\right)^{-1}_*\sigma_i\right](\mathrm{x}_T)\,\dot{k}^j_s\dot{l}^i_t ds dt \\ &\quad +\int_0^T \Phi^{\mathrm{h}}_{T\leftarrow t}\partial\sigma_i\left(\phi^{\mathrm{h}}_{t\leftarrow T}(\mathrm{x}_T)\right)\Phi^{\mathrm{h}}_{t\leftarrow T}D\phi_T(\mathrm{h})[k]\,\dot{l}^i_t dt. \end{aligned}$$

Proof. Clearly

$$D\phi_T(\mathrm{h})[l]=\int_0^T \Phi^{\mathrm{h}}_{T\leftarrow t}\sigma_i\left(\phi^{\mathrm{h}}_{t\leftarrow T}(\mathrm{x}_T)\right)\dot{l}^i_t dt$$

where $\phi_T(\mathrm{h}) = \mathrm{x}_T$. Perturbing h implies

$$\phi_T(\mathrm{h}+\varepsilon k) = \mathrm{x}_T + \varepsilon D\phi_T(\mathrm{h})[k] + o(\varepsilon)$$

and then

$$D\phi_T(\mathrm{h}+\varepsilon k)[l] = \int_0^T \Phi_{T\leftarrow t}^{\mathrm{h}+\varepsilon k}\sigma_i\left(\phi_{t\leftarrow T}^{\mathrm{h}+\varepsilon k}(\mathrm{x}_T + \varepsilon D\phi_T(\mathrm{h})[k] + o(\varepsilon))\right)\dot{l}_t^i dt.$$

Taking derivatives then leads us to[12]

$$\begin{aligned} D^2\phi_T(\mathrm{h})[k,l] = &\int_0^T D\left\{\Phi_{T\leftarrow t}^{\mathrm{h}}\sigma_i\left(\phi_{t\leftarrow T}^{\mathrm{h}}(\mathrm{x}_T)\right)\right\}[k]\,\dot{l}_t^i dt \\ &+ \int_0^T \Phi_{T\leftarrow t}^{\mathrm{h}}\partial\sigma_i\left(\phi_{t\leftarrow T}^{\mathrm{h}}(\mathrm{x}_T)\right)\Phi_{t\leftarrow T}^{\mathrm{h}} D\phi_T(\mathrm{h})[k]\,\dot{l}_t^i dt. \end{aligned}$$

The proof is then finished using (25). ■

Given $k \in H_0$, set

$$\begin{pmatrix} 0 \\ \eta \end{pmatrix} := D\phi_T(\mathrm{h})[k] \tag{26}$$

where the notation is meant to suggest that

$$\eta \in \mathcal{T}_{\mathrm{x}_T}N_{\mathrm{a}} \text{ where } N_{\mathrm{a}} = (\mathrm{a},\cdot) \subset \mathbb{R}^l \times \mathbb{R}^{d-l} \cong \mathbb{R}^d.$$

Proposition 20 *Elements $k \in \mathcal{N}(\mathrm{h}) \subset H_0$ are characterized by (inhomogeneous, linear "backward") Volterra equation*[13]

$$\begin{aligned} \dot{k}_t^i = &\left\langle (q(\mathrm{h}),0), \int_t^T \left[\left(\phi_{s\leftarrow T}^{\mathrm{h}}\right)_*^{-1}\sigma_j, \left(\phi_{t\leftarrow T}^{\mathrm{h}}\right)_*^{-1}\sigma_i\right](\mathrm{x}_T)\,\dot{k}_s^j ds\right\rangle \\ &+ \left\langle (q(\mathrm{h}),0), \Phi_{T\leftarrow t}^{\mathrm{h}}\partial\sigma_i\left(\phi_{t\leftarrow T}^{\mathrm{h}}(\mathrm{x}_T)\right)\Phi_{t\leftarrow T}^{\mathrm{h}}\begin{pmatrix}0\\ \eta\end{pmatrix}\right\rangle \\ &+ \left\langle (\theta,0), \Phi_{T\leftarrow t}^{\mathrm{h}}\sigma_i\left(\phi_{t\leftarrow T}^{\mathrm{h}}(\mathrm{x}_T)\right)\right\rangle. \end{aligned}$$

where

$$\eta = \eta(k) \in \mathrm{span}\{\partial_{l+1}|_{\mathrm{x}_T}, \ldots, \partial_d|_{\mathrm{x}_T}\} = \mathcal{T}_{\mathrm{x}_T}N_{\mathrm{a}}$$

is given by (26) and

$$\theta = \theta(k) \in \mathrm{span}\left\{dx^1|_{\mathrm{x}_T}, \ldots, dx^l|_{\mathrm{x}_T}\right\} = \mathcal{T}_{\mathrm{x}_T}^* N_{\mathrm{a}}^{\perp}.$$

Remark 21 *When $k \in \mathcal{N}(\mathrm{h})$ is also in $H_1 = \ker D\phi_T(\mathrm{h})$ (which is always true in the point-point setting!) we have $\eta = 0$; the equation for k simplifies accordingly and matches precisely the Bismut's equation [14, 1.65].*

[12] It should be noted that the term $D\phi_T(\mathrm{h})[k]$ is zero for $k \in H_1 = \ker D\phi_T(\mathrm{h})$; in particular the second summand will vanish when $D^2\phi_T(\mathrm{h})[\cdot,\cdot]$ is restricted to H_1 i.e. when considering the point-point case $l = d$.

[13] ... which takes the usual form upon reparameterizing time $\tau \leftarrow T - t$ ...

Remark 22 *It is an important step in our argument to single out η. In fact, we must not use*

$$\begin{pmatrix} 0 \\ \eta \end{pmatrix} = \int_0^T \Phi^{\mathrm{h}}_{T\leftarrow s}\sigma_j\left(\phi^{\mathrm{h}}_{s\leftarrow T}\left(\mathrm{x}_T\right)\right)\dot{k}^j_s ds$$

as integral term for $\dot{k}$ in the above integral equation for $\dot{k}$. Indeed, doing so would lead to a Fredholm integral equation (of the second kind) for $\dot{k}$ whereas it will be crucial for the subsequent argument to have a Volterra structure. (Solutions to such Volterra equations are unique; the same is not true for Fredholm integral equations.)

Proof. For fixed $k \in H_0$, we write

$$\begin{pmatrix} 0 \\ \eta \end{pmatrix} := D\phi_T\left(\mathrm{h}\right)[k].$$

With slight abuse of notation (Riesz!) the previous result then implies that

$$\begin{aligned}\left\{D^2\phi_T\left(\mathrm{h}\right)[k,\cdot]\right\}^i_t = \int_t^T &\left[\left(\phi^{\mathrm{h}}_{s\leftarrow T}\right)^{-1}_* \sigma_j, \left(\phi^{\mathrm{h}}_{t\leftarrow T}\right)^{-1}_* \sigma_i\right]\left(\mathrm{x}_T\right)\dot{k}^j_s ds \\ &+\Phi^{\mathrm{h}}_{T\leftarrow t}\partial\sigma_i\left(\phi^{\mathrm{h}}_{t\leftarrow T}\left(\mathrm{x}_T\right)\right)\Phi^{\mathrm{h}}_{t\leftarrow T}\begin{pmatrix} 0 \\ \eta \end{pmatrix}.\end{aligned} \tag{27}$$

On the other hand, for $k \in \mathcal{N}\left(\mathrm{h}\right)$, we know that

$$\langle k,\cdot\rangle_H - \left\langle\left(q\left(\mathrm{h}\right),0\right), D^2\phi_T\left(h\right)[k,\cdot]\right\rangle \in H_0^{\perp} = \mathrm{range}\left(\tilde{\rho}_{\mathrm{h}}\right).$$

Hence, recalling

$$\tilde{\rho}_{\mathrm{h}}\left(\theta\right) = \left\langle\left(\theta,0\right), \Phi^{\mathrm{h}}_{T\leftarrow t}\sigma_i\left(\phi^{\mathrm{h}}_{t\leftarrow T}\left(\mathrm{x}_T\right)\right)\right\rangle,$$

it follows from (27) that

$$\begin{aligned}\dot{k}^i_t = &\left\langle\left(q\left(\mathrm{h}\right),0\right), \int_t^T \left[\left(\phi^{\mathrm{h}}_{s\leftarrow T}\right)^{-1}_* \sigma_j, \left(\phi^{\mathrm{h}}_{t\leftarrow T}\right)^{-1}_* \sigma_i\right]\left(\mathrm{x}_T\right)\dot{k}^j_s ds\right\rangle \\ &+\left\langle\left(q\left(\mathrm{h}\right),0\right), \Phi^{\mathrm{h}}_{T\leftarrow t}\partial\sigma_i\left(\phi^{\mathrm{h}}_{t\leftarrow T}\left(\mathrm{x}_T\right)\right)\Phi^{\mathrm{h}}_{t\leftarrow T}\begin{pmatrix} 0 \\ \eta \end{pmatrix}\right\rangle \\ &+\left\langle\left(\theta,0\right), \Phi^{\mathrm{h}}_{T\leftarrow t}\sigma_i\left(\phi^{\mathrm{h}}_{t\leftarrow T}\left(\mathrm{x}_T\right)\right)\right\rangle\end{aligned}$$

■

Remark 23 *If we introduce the orthogonal complement H_2 so that*

$$H_0 = H_1 \oplus H_2 \text{ (orthogonal)}$$

the map

$$k \mapsto D\phi_T\left(h\right)[k] = \begin{pmatrix} 0 \\ \eta \end{pmatrix} \mapsto \eta$$

is a bijection from $H_2 \to \mathcal{T}_{\mathrm{x}_T}N_{\mathrm{a}}$.

3.1 Jacobi variation

Again, the starting point is the formula

$$\begin{aligned}\dot{h}_t^i &= \left\langle \mathrm{p}_T, \Phi^{\mathrm{h}}_{T\leftarrow t}\sigma_i\left(\phi^{\mathrm{h}}_{t\leftarrow T}\left(\mathrm{x}_T\right)\right)\right\rangle \\ &= \left\langle \left(q(\mathrm{h}),0\right), \Phi^{\mathrm{h}}_{T\leftarrow t}\sigma_i\left(\phi^{\mathrm{h}}_{t\leftarrow T}\left(\mathrm{x}_T\right)\right)\right\rangle\end{aligned}$$

where we recall

$$\mathrm{p}_T = \left(q(\mathrm{h}),0\right),\ \mathrm{x}_T \in (\mathrm{a},\cdot) \equiv N_{\mathrm{a}}.$$

We keep p_T and x_T fixed and note that the Hamiltonian (backward) dynamics are such that

$$\Pi H_{t\leftarrow T}\left(\mathrm{x}_T,\mathrm{p}_T\right) = \phi^{\mathrm{h}}_{t\leftarrow T}\left(\mathrm{x}_T\right)$$

Replace p_T by $\mathrm{p}_T+\varepsilon(\theta,0)$ above, x_T by $\mathrm{x}_T+\varepsilon\begin{pmatrix}0\\ \eta\end{pmatrix}$ and write $h\left(\varepsilon\right)$ for the according control[14] which satisfies the relation

$$\dot{h}\left(\varepsilon\right)_t^i = \left\langle \mathrm{p}_T+\varepsilon\left(\theta,0\right), \Phi^{h(\varepsilon)}_{T\leftarrow t}\sigma_i\left(\phi^{h(\varepsilon)}_{t\leftarrow T}\left(\mathrm{x}_T+\varepsilon\begin{pmatrix}0\\ \eta\end{pmatrix}\right)\right)\right\rangle$$

Define the *Jacobi type variation*

$$g := \partial_{(\theta,\eta)}h := \frac{\partial h\left(\varepsilon\right)}{\partial\varepsilon}|_{\varepsilon=0}$$

so that

$$\begin{aligned}\dot{g}_t^i = &\left\langle \mathrm{p}_T, D\left\{\Phi^{\mathrm{h}}_{T\leftarrow t}\sigma_i\left(\phi^{\mathrm{h}}_{t\leftarrow T}\left(\mathrm{x}_T\right)\right)\right\}[g]\right\rangle \\ &+\left\langle \mathrm{p}_T, \Phi^{\mathrm{h}}_{T\leftarrow t}\partial\sigma_i\left(\phi^{\mathrm{h}}_{t\leftarrow T}\left(\mathrm{x}_T\right)\right)\Phi^{\mathrm{h}}_{t\leftarrow T}\begin{pmatrix}0\\ \eta\end{pmatrix}\right\rangle \\ &+\left\langle \left(\theta,0\right), \Phi^{\mathrm{h}}_{T\leftarrow t}\sigma_i\left(\phi^{h}_{t\leftarrow T}\left(\mathrm{x}_T\right)\right)\right\rangle.\end{aligned}$$

With $\mathrm{p}_T = (q(\mathrm{h}),0)$ and formula (25) we see that $\dot{g}$ satisfies the identical (inhomogeneous, linear backward[15] Volterra equation) as the one given for $\dot{k}$ in proposition 20. By basic uniqueness theory for such Volterra equations we see that $\dot{g}=\dot{k}$ as elements in $L^2\left([0,T],\mathbb{R}^m\right)$, and hence $g=k$ as elements in H.

Proposition 24 *Let $k\in\mathcal{N}\left(\mathrm{h}\right)\subset H_0$ with associated parameters*

$$\begin{aligned}\theta &\in \mathrm{span}\left\{dx^1|_{\mathrm{x}_T},\ldots,dx^l|_{\mathrm{x}_T}\right\} = \mathcal{T}^*_{\mathrm{x}_T}N_{\mathrm{a}}^{\perp} \\ \eta &\in \mathrm{span}\left\{\partial_{l+1}|_{\mathrm{x}_T},\ldots,\partial_d|_{\mathrm{x}_T}\right\} = \mathcal{T}_{\mathrm{x}_T}N_{\mathrm{a}}\end{aligned}$$

[14]... which can be constructed explicitly from the Hamiltonian (backward) flow

$$\left(x_t\left(\varepsilon\right),p_t\left(\varepsilon\right)\right) := H_{t\leftarrow T}\left(\mathrm{x}_T+\varepsilon\begin{pmatrix}0\\ \eta\end{pmatrix},\mathrm{p}_T+\varepsilon(\theta,0)\right)$$

and the usual formula $\dot{h}\left(\varepsilon\right)_t^i = \left(p_t\left(\varepsilon\right),V_i\left(x_t\left(\varepsilon\right)\right)\right)$.

[15]Trivial reparameterization $t\rightsquigarrow T-t$ will bring it in standard "forward" form.

provided by proposition 20. (In particular, η is given by $D\phi_T(\mathrm{h})[k]$, cf. (26).) Then k can be written in terms of a Jacobi type variation

$$k = \partial_{(\theta,\eta)} h.$$

Conversely, any Jacobi type variation, with $\theta \in \mathcal{T}^*_{\mathrm{x}_T} N_{\mathrm{a}}^{\perp}, \eta \in \mathcal{T}_{\mathrm{x}_T} N_{\mathrm{a}}$ yields an element in $\mathcal{N}(\mathrm{h})$.

Proof. The first part follows from the above discussion and it only remains to prove the converse part. Since we have seen that every Jacobi type variation $g := \partial_{(\theta,\eta)} h$ satisfies the appropriate Volterra equation, cf. proposition 20, we only need to check

$$\begin{pmatrix} 0 \\ \eta \end{pmatrix} = D\phi_T(\mathrm{h})[g]\, x$$

and we leave this as an easy exercise to the reader. ■

Recall that we say that x_0 is *non-focal* for $(\mathrm{a}, \cdot) \equiv N_{\mathrm{a}}$ along h if for all $\theta \in \mathcal{T}^*_{\mathrm{x}_T} N_{\mathrm{a}}^{\perp}$, $\eta \in \mathcal{T}_{\mathrm{x}_T} N_{\mathrm{a}}$

$$\partial_{\varepsilon}|_{\varepsilon=0} \Pi H_{0\leftarrow T} \left(\mathrm{x}_T + \varepsilon \begin{pmatrix} 0 \\ \eta \end{pmatrix}, \mathrm{p}_T + \varepsilon(\theta, 0) \right) = 0 \implies (\theta, \eta) = 0.$$

In the point-point setting (i.e. $l = d$ so that $\theta \in \mathcal{T}^*_{\mathrm{x}_T} \mathbb{R}^d$, $\eta = 0$) the criterion reduces to

$$\partial_{\varepsilon}|_{\varepsilon=0} \Pi H_{0\leftarrow T}(\mathrm{x}_T, \mathrm{p}_T + \varepsilon\theta) = 0 \implies \theta = 0;$$

disregarding time reparameterization $t \leftarrow T - t$ and the fact that our setup allows for a non-zero drift vector field, this is precisely Bismut's non-conjugacy condition [14, p.50].

Corollary 25 *The point x_0 is non-focal for $(\mathrm{a}, \cdot) \equiv N_{\mathrm{a}}$ along h if and only if $I''(\mathrm{h})$, i.e. the second derivative of $\|\cdot\|^2_H\Big|_{\mathcal{K}_{\mathrm{a}}}$ at the minimizer h, viewed as quadratic form on $H_0 = \ker D(\Pi_l \phi_T)(\mathrm{h})$, is non-degenerate, i.e.*

$$\mathcal{N}(\mathrm{h}) \equiv \{0\}.$$

Proof. "⇒": Take $k \in \mathcal{N}(\mathrm{h})$; from proposition 24

$$k = \partial_{(\theta,\eta)} h \equiv \partial_{\varepsilon}|_{\varepsilon=0} h(\varepsilon)$$

for suitable $\theta \in \mathcal{T}^*_{\mathrm{x}_T} N_{\mathrm{a}}^{\perp}, \eta \in \mathcal{T}_{\mathrm{x}_T} N_{\mathrm{a}}$; in fact,

$$\begin{pmatrix} 0 \\ \eta \end{pmatrix} = D\phi^{\mathrm{h}}_{T\leftarrow 0}(\mathrm{x}_0)[k].$$

The criterion says that if

$$\partial_{\varepsilon}|_{\varepsilon=0} \Pi H_{0\leftarrow T} \left(\mathrm{x}_T + \varepsilon \begin{pmatrix} 0 \\ \eta \end{pmatrix}, \mathrm{p}_T + \varepsilon(\theta, 0) \right) = \partial_{\varepsilon}|_{\varepsilon=0} \left(\phi^{h(\varepsilon)}_{0\leftarrow T} \right) \left(\mathrm{x}_T + \varepsilon \begin{pmatrix} 0 \\ \eta \end{pmatrix} \right)$$

equals zero then (θ,η) must be zero. But this is indeed the case here since

$$\begin{aligned}
&\partial_\varepsilon|_{\varepsilon=0}\left(\phi^{h(\varepsilon)}_{0\leftarrow T}\right)\left(\mathrm{x}_T+\varepsilon\begin{pmatrix}0\\ \eta\end{pmatrix}\right)\\
&=D\left\{\phi^{\mathrm{h}}_{0\leftarrow T}\left(\mathrm{x}_T\right)\right\}\left[\partial_\varepsilon|_{\varepsilon=0}h\left(\varepsilon\right)\right]+\Phi^{\mathrm{h}}_{0\leftarrow T}\begin{pmatrix}0\\ \eta\end{pmatrix}\\
&=D\left\{\phi^{\mathrm{h}}_{0\leftarrow T}\left(\mathrm{x}_T\right)\right\}\left[k\right]+\Phi^{\mathrm{h}}_{0\leftarrow T}D\phi^{\mathrm{h}}_{T\leftarrow 0}\left(\mathrm{x}_0\right)\left[k\right]\\
&=D\left\{\phi^{\mathrm{h}}_{0\leftarrow T}\circ\phi^{\mathrm{h}}_{T\leftarrow 0}\left(\mathrm{x}_0\right)\right\}\left[k\right]\\
&=0.
\end{aligned}$$

We thus conclude that the directional derivative $\partial_{(\theta,\eta)}h$, which of course depends linearly on (θ,η), vanishes. It then follows that $k=\partial_{(\theta,\eta)}h=0$ which is what we wanted to show.
"$\Leftarrow$": Assume there exists $(\theta,\eta)\neq 0$ so that

$$\partial_\varepsilon|_{\varepsilon=0}\Pi H_{0\leftarrow T}\left(\mathrm{x}_T+\varepsilon\begin{pmatrix}0\\ \eta\end{pmatrix},\mathrm{p}_T+\varepsilon\left(\theta,0\right)\right)=0.$$

Then $k:=\partial_{(\theta,\eta)}h$ yields an element in the null-space $\mathcal{N}\left(\mathrm{h}\right)$. We need to see that k is non-zero. Assume otherwise, i.e. $k=0$. Then $D\phi^{\mathrm{h}}_{T\leftarrow 0}\left(\mathrm{x}_0\right)\left[k\right]=0$ and hence also $\eta=0$. From the Volterra equation for k we see that

$$0=\left\langle\left(\theta,0\right),\Phi^{\mathrm{h}}_{T\leftarrow t}\sigma_i\left(\phi^{h}_{t\leftarrow T}\left(\mathrm{x}_T\right)\right)\right\rangle=\tilde{\rho}_{\mathrm{h}}\left(\left(\theta,0\right)\right).$$

But $\ker\tilde{\rho}_{\mathrm{h}}$ was seen to be trivial and so $\theta=0$; in contradiction to assumption $(\theta,\eta)\neq 0$. ■

4 Applications

4.1 Scaled Ornstein-Uhlenbeck

As a warmup, consider a scaled one-dimensional Ornstein-Uhlenbeck process of the form

$$dY^\varepsilon_t=\left(\alpha\varepsilon+\beta Y^\varepsilon_t\right)dt+\gamma\varepsilon dW_t,\ \ Y^\varepsilon_0=\varepsilon y_0\in\mathbb{R}.$$

Clearly, $Y^\varepsilon_T\sim N\left(\varepsilon\mu,\varepsilon^2\sigma^2\right)$ with

$$\mu=y_0e^{\beta T}x_0e^{\beta T}+\frac{\alpha}{\beta}\left(e^{\beta T}-1\right),\ \sigma^2=\frac{\gamma^2}{2\beta}\left(e^{2\beta T}-1\right) \tag{28}$$

and so Y^ε_T has density of the form

$$f^\varepsilon\left(y,T\right)=\frac{1}{\varepsilon\sigma\sqrt{2\pi}}e^{-\frac{(y-\varepsilon\mu)^2}{2\varepsilon^2\sigma^2}}\equiv\varepsilon^{-1}e^{-c_1/\varepsilon^2}e^{c_2/\varepsilon}\left(c_0+O\left(\varepsilon\right)\right), \tag{29}$$

with constants c_i implicitly defined by (29). We leave it to the reader to check that the same density, with explicit expressions for c_1,c_2 is obtained from our main theorem.

4.2 Elliptic example with degeneracy

Consider the small noise problem for the stochastic differential equation

$$\begin{aligned} dY^\varepsilon &= \varepsilon dW^1 + \theta Z^\varepsilon \varepsilon dW^2,\ Y_0^\varepsilon = 0; \\ dZ^\varepsilon &= \varepsilon dW^2,\ Z_0^\varepsilon = 0; \end{aligned}$$

where $\theta \in [0,1]$, say. Note that it could be immediately rephrased as short-time problem $\left(T = 1, t = \varepsilon^2\right)$. We are in an elliptic (Riemannian) setting. (In fact, $\mathbb{R}^2$ with the induced metric has zero-curvature and empty cut-locus.) Clearly, Y_T^ε admits a density, say $f^\varepsilon(y)$ at time $T = 1$. Considering the point $y = 1$, for instance, it is not hard to see that

$$\varepsilon^2 \log f^\varepsilon(1) \sim -\frac{1}{2} \text{ as } \varepsilon \downarrow 0.$$

At least when $\theta = 0$ it is obvious from $Y_T^\varepsilon \sim N\left(0, \varepsilon^2 T\right)$ that one has the expansion

$$f^\varepsilon(1) = \varepsilon^{-1} e^{-\frac{1}{2\varepsilon^2}} \left(c_0 + O(\varepsilon)\right)$$

for some (easy to compute) $c_0 > 0$. Interestingly, the general situation is much more involved. Exploiting the fact that Y_T^ε can be written as the independent sum of a Gaussian and a (non-centered) Chi-square random-variable, $f^\varepsilon(y)$ is given by a convolution integral and a direct (tedious) analysis shows that

$$f^\varepsilon(1) = \begin{cases} \varepsilon^{-1} e^{-\frac{1}{2\varepsilon^2}} \left(c_0 + O(\varepsilon)\right) & \text{when } \theta \in [0,1) \\ \varepsilon^{-3/2} e^{-\frac{1}{2\varepsilon^2}} \left(c_0 + O(\varepsilon)\right) & \text{when } \theta = 1 \end{cases}. \tag{30}$$

While the energy is equal to $1/2$, no matter the value $\theta \in [0,1]$, we see the appearance of an atypical algebraic factor $\varepsilon^{-3/2}$ in the case $\theta = 1$.

With a view towards applying our theorem 9: we have vector fields σ_1, σ_2 of the form ∂_y, $\theta z \partial_y + \partial_z$. One checks without difficulty that $\mathrm{h}_0(t) = (t, 0)$ is the (unique) element in $\mathcal{K}_a^{\min}$, for any $\theta = [0,1]$. In particular, the "most-likely" arrival point is $(1,0) \in (1, \cdot)$. (Minimizers and energy start to look different when $\theta > 1$, our focus on $\theta \in [0,1]$ is pure algebraic convenience.) In the case $\theta = 1$, the explicit "backward" and projected Hamiltonian flow is

$$\begin{aligned} &\pi H_{0 \leftarrow T}\left((y_T, z_T), (p_T, q_T)\right) \\ &= \begin{pmatrix} y_T + \frac{1}{2}\left(p_T z_T T + q_T T - z_T\right)^2 - \left(p_T T + \frac{1}{2} z_T^2\right) \\ z_T - q_T T - p_T z_T T \end{pmatrix}. \end{aligned}$$

From this expression, it is then easy to check that $(0,0)$ is focal for $(1, \cdot)$. (Proposition 7 then implies that the Hessian of the energy at h_0 is degenerate. In fact, a simple computation shows that in this example the null-space of $I''(\mathrm{h}_0)$ is given by $\mathcal{N}(\mathrm{h}_0) = [\mathrm{k}]$ where $\mathrm{k} = \mathrm{k}(t) = (0,t) \in \mathcal{T}_{\mathrm{h}_0} \mathrm{K}_1 \setminus \{0\}$.) It follows that one must not apply theorem 9 here, and indeed, the predication of the theorem (algebraic factor ε^{-1}) would be false in the case $\theta = 1$, as we know from (30). On the other hand, one checks without trouble that for $\theta < 1$ the situation is non-focal, all our assumptions are then met, and so theorem 9 yields the correct expansion, in agreement with (30).

4.3 Hypoelliptic Gaussian example

We are interested in the *tail behaviour* of Y_T where

$$\begin{aligned} dY &= Zdt, \quad Y_0 = y_0, \\ dZ &= dW_t, \quad Z_0 = z_0. \end{aligned}$$

Of course, Y_T is Gaussian with mean $\mu = y_0 + z_0 T$ and variance

$$\sigma^2 := \mathbb{V}[Y_T] = \mathbb{E}\left[\int_0^T \int_0^T W_s W_t ds dt\right] = 2\left[\int_{0<s<t<T} s ds dt\right] = \int_0^T t^2 dt = T^3/3.$$

We are *not* looking here at the short time behaviour of Y_t as $t \downarrow 0$: Indeed the condition H1 is not satisfied here and indeed the log density of Y_t is proportional to $1/\sigma^2 = O\left(t^{-3}\right)$ as $t \downarrow 0$ which is not at all the behaviour described in corollary (15). Instead, let us fix $T > 0$ and note that the density of Y_T is of the form

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \sim (\text{const}) e^{-\frac{3}{2T^3}\left(y^2 - 2\mu y\right)} \equiv (\text{const}) e^{-c_1 y^2 + c_2 y} \text{ as } y \uparrow \infty. \tag{31}$$

We now illustrate how this follows from corollary 14.

Scaling: Set $Y^\varepsilon := \varepsilon Y$ and similar for Z. Then

$$\begin{aligned} dY_t^\varepsilon &= Z_t^\varepsilon dt, \quad Y_0^\varepsilon = \varepsilon y_0, \\ dZ_t^\varepsilon &= \varepsilon dW_t, \quad Z_0^\varepsilon = \varepsilon z_0. \end{aligned}$$

In other words, the (first) component of interest to us scales with $\theta = 1$. It remains to **check the assumptions**. With $\sigma_0 = z\partial_y$ and $\sigma_1 = \partial_z$ we have $[\sigma_0, \sigma_1] = \partial_y$ which not only implies the weak Hörmander's condition H but a stronger "Bismut H2 type" condition which implies [14, Thm 1.10] invertibility of C_T^{h,x_0} for all $h \neq 0$. We are interested in paths going from $\lim_{\varepsilon\to 0}\left(Y_0^\varepsilon, Z_0^\varepsilon\right) = (0,0)$ to $N = (a, \cdot)$ with $a = 1$ and it is easy to see that this is possible upon replacing W by a suitable Cameron-Martin path; in other words,

$$\mathcal{K}_a \neq \varnothing.$$

(Cf. [33] for an abstract criterion that applies in this example). Since $h \equiv 0$ will never stir us from $(0,0)$ to N we only need to check that $(0,0) \times N$ satisfies condition (ND). To this end, we note that the Hamiltonian in the present setting is

$$\mathcal{H}\left((y,z);(p,q)\right) = pz + \frac{1}{2}q^2;$$

The Hamiltonian flow $\mathrm{H}_{t\leftarrow 0} = \mathrm{H}_{t\leftarrow 0}(y,z,p,q)$ turns out to be an easily computable linear map. The details of computing the minimizing control $h = h(t)$ then follow the recipe given in remark 6. In particular, we find $\mathrm{p}_0 = (p_0, q_0) = (3/T, 3/T)$ and minimizing control $h_0 = h_0(t) = 3(T-t)/T^3$ in the notation of proposition 4. In particular then,

$$c_1 := \Lambda(1) = \frac{1}{2}\int_0^T \left|\dot{h}(t)\right|^2 dt = \frac{3}{2T^3}$$

in agreement with (31). For the second order constant c_2, we need to compute $\hat{Y}_T$ where

$$d\hat{Y}_t = \hat{Z}_t dt, \quad \hat{Y}_0 = y_0, \quad d\hat{Z}_t = 0, \quad \hat{Z}_0 = z_0.$$

This leads immediately to $\hat{Y}_T = y_0 + z_0 T =: \mu$ and then $c_2 = \left(2\hat{Y}_T/\theta\right)\Lambda(1) = 2\mu c_1$, again in agreement with the Gaussian computation.

4.4 Lévy's Area

Following a similar discussion in [49], we consider

$$\sigma_1 \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ -y/2 \end{pmatrix}, \sigma_2 \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ z/2 \end{pmatrix};$$

the resulting diffusion

$$d\mathrm{X}_t = \sigma_1(\mathrm{X}_t)\, dW_t^1 + \sigma_2(\mathrm{X}_t)\, dW_t^2, \quad \mathrm{X}_0 = \mathrm{x}_0 \in \mathbb{R}^3$$

is known as *Brownian motion on the* 3*-dimensional Heisenberg group*; it can also be viewed as the *Brownian rough path* (e.g. [19] and the references therein) associated to the 2-dimensional standard Brownian motion $\left(W^1, W^2\right)$. Set $\mathrm{X}^\varepsilon := \delta_\varepsilon \mathrm{X}$ where $\delta_\varepsilon(x,y,z) = \left(\varepsilon x, \varepsilon y, \varepsilon^2 z\right)$ is the dilation operator on the Heisenberg group. Then

$$d\mathrm{X}_t^\varepsilon = \sigma_1(\mathrm{X}_t^\varepsilon)\, \varepsilon dW_t^1 + \sigma_2(X_t^\varepsilon)\, \varepsilon dW_t^2$$

which is also the form relevant to study short time asymptotics. Since the Hamiltonian is explicitly available [23], our criterion (ND) can be checked directly[16]. In particular, we so recover all "non-degenerate" projected Lévy area expansion results of [49]: in the notation of that paper, section 7, we cover their cases $(\mathrm{I})_1$,$(\mathrm{I})_2$,$(\mathrm{III})_1$,$(\mathrm{III})_2$,$(\mathrm{III})_3$. The main difference, comparing the approach [49] with ours, is that our criterion (ND) bypasses the involved analysis, carried out by hand in [49], of the infinite-dimension Hessian of the energy at the minimizer. On the other hand, our approach (presently) does not deal with degenerate minima, and we do not cover their cases $(\mathrm{I})_3$, (II), $(\mathrm{III})_4$; all of which are, of course, ruled out by violating condition (ND).

Let us conclude with an application to Lévy's area not considered in [49]. We are interested in the tail-behaviour of the density, say $f = f(z)$, of Lévy's area Z_t at unit time,

$$Z_1 := \frac{1}{2}\int_0^1 \left(W_s^1 dW_s^2 - W_s^2 dW_s^1\right).$$

One expects $-\log f(z) \sim c^* z$ for some $c^* > 0$ since Z_1 is an element of the second Wiener Itô-chaos which entails[17]

$$E\left[\exp(cZ_1)\right] < \infty \text{ for some } 0 < c < c^* \tag{32}$$

$$E\left[\exp(cZ_1)\right] = +\infty \text{ for some } c > c^*. \tag{33}$$

No explicit density is available but a c.f. is known (Lévy's formula). Density expansions could then be obtained using saddle point methods, for instance. We now illustrate how such expansion

[16] Bismut's H2 condition guarantees invertibility of the deterministic Malliavin matrix for any $h \neq 0$.

[17] There are examples of r.v.s with (32),(33) with density f for this fails!

follows from corollary 14. To this end we exploit scaling, $\mathrm{X}^\varepsilon := \delta_\varepsilon \mathrm{X}$. By trivial relabelling of the coordinates (1 versus 3), our tail result shows, after verification of condition (ND), that Lévy's area Z_1 admits a density of the form

$$f(z) = e^{-\Lambda_0(1)z} z^{-1/2} \left(c_0 + O\left(1/\sqrt{z}\right)\right)$$

where $\Lambda_0(1)$ is seen to be the Carnot-Caratheodory distance of $(0,0,0)$ to the hyperplane $N = (\cdot, \cdot, 1)$.

4.5 Black-Scholes

The Black-Scholes (BS) model, written in log-language is an example where the above theorem is applicable with $\theta = 1$. Indeed, $Y := \log S$ satisfies, with fixed Black-Scholes volatility $\sigma > 0$

$$dY_t = -\frac{\sigma^2}{2} dt + \sigma dW_t, \quad Y_0 = y_0 = \log S_0.$$

Of course, $Y_t \sim N\left(y_0 - \sigma^2 t/2, \sigma^2 t\right)$ and the explicit Gaussian density

$$f_{\mathrm{BS}}(t,y) = \frac{1}{\sqrt{2\pi\sigma^2 t}} \exp\left\{ -\frac{\left(y - \left(y_0 - \sigma^2 t/2\right)\right)^2}{2\sigma^2 t} \right\}$$

immediately yields short time resp. tail expansions,

$$f_{\mathrm{BS}}(t,y) \sim (\mathrm{const})\, t^{-1/2} \exp\left(-\frac{\left(\frac{y-y_0}{\sigma}\right)^2}{2t} \right) \text{ as } t \downarrow 0; \text{ any } y \in \mathbb{R} \tag{34}$$

$$f_{\mathrm{BS}}(T,y) \sim (\mathrm{const}) \exp\left(-\frac{1}{2\sigma^2 T} y^2 \right) \exp\left(\frac{y_0 - \sigma^2 T/2}{2\sigma^2 T} y \right) \text{ as } y \to \infty; \text{ any } T > 0. \tag{35}$$

We derive now both expansions from general theory, i.e. with aid of corollary 15 resp 14. The short time limit corresponds to a flat Riemannian situation, in particular the cutlocus is empty, which is enough to guarantee (ND); the remaining computations to derive (34) from corollary 15 are left to the reader and we focus on the (more interesting) case of tail asymptotics. Corollary 14 applies with $\theta = 1$, and (rescaled) starting point $\varepsilon y_0 \to 0$. Condition (ND) needs to be checked; the relevant Hamiltonian is

$$\mathcal{H}(y,p) = -\frac{\sigma^2}{2} yp + \frac{\sigma^2 p^2}{2}, \qquad \text{for all } (y,p) \in \mathbb{R}^2$$

and the Hamilton ODEs are

$$\dot{y}_t = -\frac{\sigma^2}{2} y_t + \sigma^2 p_t, \quad \dot{p}_t = \frac{\sigma^2}{2} p_t,$$

with boundary conditions $y_0 = 0$ and $y_T = a$. We first solve the ODE for p and we obtain $p_t = p e^{\sigma^2 t/2}$, for some $p \in \mathbb{R}$. We can then deduce the solution for the path y as a function of p:

$$y_t = \frac{1 - e^{-\sigma^2 t/2}}{\sigma^2/2} \left(\frac{\sigma^2 p}{2} \right).$$

The boundary condition $y_T = a$ implies that

$$p = p_0 := \frac{2}{\sigma^2} \frac{\left(-\frac{\sigma^2}{2} y_0\right) e^{-\sigma^2 T/2} + \sigma^2 a/2}{e^{\sigma^2 T/2} - e^{-\sigma^2 T/2}}.$$

In particular, $\partial_p y_T|_{p_0} = 1 - e^{-\sigma^2 T/2} > 0$, and hence invertible, for $T, \sigma > 0$. En passant, we also deduce the optimal control $h_0(t) = \sigma p_t$, and get the correct leading order factor

$$c_1 := \frac{1}{2} \|h_0\|^2 = \frac{1}{2} \int_0^T h_0(t)^2 dt = \frac{p_0^2}{\sigma^2} \left(e^{\sigma T} - 1\right) = \frac{1}{2\sigma^2 T}.$$

With the hint $\hat{Y}_t = y_0 + \left(-\frac{\sigma^2}{2}\right) t$ we leave it to the reader to verify that $c_2 = \left(y_0 - \sigma^2 T/2\right) / \left(2\sigma^2 T\right)$. Frequently, one chooses $y_0 = 0$ in this context (which amounts to normalize spot price to unit).

4.6 The Stein-Stein model

For given parameters, $a \geq 0, b < 0, c > 0, \sigma_0 \geq 0, \rho = d \left\langle W^1, W^2 \right\rangle / dt$, the Stein–Stein model expresses log-price Y, under the forward measure, via[18]

$$\begin{aligned} dY &= -\frac{1}{2} Z^2 dt + Z dW^1, \; Y(0) = y_0 = 0 \\ dZ &= (a + bZ)\, dt + c dW^2, \;\; Z(0) = \sigma_0 > 0. \end{aligned} \tag{36}$$

We will be interested in the behaviour, and in particular the tail-behaviour, of the probability density function of Y_T. In fact, there is no loss of generality to consider $T = 1$. Applying Brownian scaling, it is a straight-forward computation to see that the pair $\left(\tilde{Y}, \tilde{Z}\right)$ given by

$$\tilde{Y}(t) := Y(tT), \;\; \tilde{Z}(t) := Z(tT)\, T^{1/2}$$

satisfies the same parametric SDE form as Stein-Stein, but with the following parameter substitutions

$$a \leftarrow \tilde{a} \equiv aT^{3/2}, b \leftarrow \tilde{b} \equiv bT, c \leftarrow \tilde{c} \equiv cT, \sigma_0 \leftarrow \tilde{\sigma}_0 \equiv \sigma_0 T^{1/2}.$$

In particular then, $Y_T = Y_T(a, b, c, \sigma_0, \rho)$ has the same law as $Y_1\left(\tilde{a}, \tilde{b}, \tilde{c}, \tilde{\sigma}_0, \rho\right)$.

4.6.1 The case of zero-correlation

For the moment, we shall follow [27] in assuming the Brownians to be uncorrelated,

$$d\left\langle W^1, W^2 \right\rangle_t = \rho dt \text{ with } \rho = 0.$$

Recall their main result, a density expansion for Y_T of the form

$$(*) : f(y) = e^{-c_1 y} e^{c_2 y^{1/2}} y^{-1/2} \left(c_3 + O\left(y^{-1/2}\right)\right) \text{ as } y \to \infty. \tag{37}$$

Scaling: Setting

$$Y_\varepsilon := \varepsilon^2 Y, \; Z_\varepsilon := \varepsilon Z$$

[18] Sometimes the Stein–Stein model is written with $|Z| dW^1$ rather than $Z dW^1$. In the zero correlation case this does not make a difference to the law of the process. In fact, there is a recent tendency in the finance community to use the form $Z dW^1$ which we analyze here, cf. [39].

yields the small noise problem

$$dY_\varepsilon = -\frac{1}{2}Z_\varepsilon^2 dt + Z_\varepsilon \varepsilon dW^1,\ Y_\varepsilon(0) = 0 =: y_0\ \forall \varepsilon > 0 \tag{38}$$
$$dZ_\varepsilon = (a\varepsilon + bZ_\varepsilon)\,dt + c\varepsilon dW^2,\ \ Z_\varepsilon(0) = \varepsilon\sigma_0 \to 0 =: z_0 \text{ as } \varepsilon \downarrow 0.$$

Our corollary 14, assuming its application to be justified, then gives the correct expansion (37), namely

$$f(y) = e^{-c_1 y} e^{c_2 y^{1/2}} y^{-1/2}\left(c_3 + O\left(y^{-1/2}\right)\right),$$

and also identifies the constants $c_1 = \Lambda(1)$, $c_2 = \hat{Y}_T \Lambda'(1)$. (The leading order constant $c_1$is in agreement with both [27] and [16, p40].)

Remark 26 *Corollary 14 relies on an application of theorem 9 to (38); let us note straight away that the coefficients here are smooth but unbounded. With a view towards the earlier remark on localization, and in particular (23), we note here that, due to the particular structure of the SDE, it suffices to localize such as to make σ bounded; e.g. by stopping it upon leaving a big ball of radius R. This amounts to, cf. (23), to shows that*

$$\lim_{R\to\infty} \limsup_{\varepsilon\to 0} \varepsilon^2 \log \mathbb{P}\left[|\sigma_\varepsilon|_{\infty;[0,T]} \geq R\right] = -\infty.$$

But since $\mathbb{P}\left[|\sigma_\varepsilon|_{\infty;[0,T]} \geq R\right] = \mathbb{P}\left[|\sigma|_{\infty;[0,T]} \geq R/\varepsilon\right]$ and σ is a Gaussian process, this is an immediate consequence of Fernique's estimate.

We postpone the justification that we may indeed apply corollary 14 (which involves an analysis of the Hamiltonian ODEs) and proceed in showing how further qualitative information about the expansion can be obtained without much computations.
Some information on c_1: From the "theorem"

$$c_1 := \Lambda(1) = \inf\left\{\frac{1}{2}\|\mathrm{h}\|_H^2 : \phi_0^{\mathrm{h}} = (0,0), \phi_T^{\mathrm{h}} \in (1,\cdot)\right\}$$

where $d\phi_t^{\mathrm{h},1} = -\frac{1}{2}\left|\phi_t^{\mathrm{h},2}\right|^2 dt + \phi_t^{\mathrm{h},2} d\mathrm{h}^1, d\phi_t^{\mathrm{h},2} = b\phi_t^{\mathrm{h},2}dt + cd\mathrm{h}^2$. If then follows *a priori* that

$$c_1 = c_1(b,c;T) \text{ but not on } a, \sigma_0.$$

The same is true for $\mathrm{h}^* := \mathrm{h}_0$ and $\phi^* := \phi^{\mathrm{h}_0}$of course.
Some information on c_2: First, $\Lambda'(1) = c_1$ also only depends on the parameters b, c, T (but not on a, σ_0). It remains to analyze the factor $\hat{Y}_T$ where $\left(\hat{Y}_t, \hat{Z}_t : t \geq 0\right)$ solves the ODE

$$d\hat{Y}_t = \left(-\phi_t^{*,2} + \mathrm{h}_t^{*,1}\right)\hat{Z}_t dt, \quad \hat{Y}_0 = 0$$
$$d\hat{Z}_t = b\hat{Z}_t dt + a dt, \quad \hat{Z}_0 = \sigma_0.$$

Since $\hat{Z}_t = \sigma_0 e^{bt} + a\int_0^t e^{b(t-s)}ds$ it follows that $\hat{Z}_T$ is linear in σ_0, a with coefficients depending on b and T. Furthermore, noting that

$$\hat{Y}_T = \int_0^T \left(-\phi_t^{*,2} + \mathrm{h}_t^{*,1}\right)\hat{Z}_t dt$$

a similar statement is true for $\hat{Y}_T$ and then $c_2 = \Lambda'(1) \times \hat{Y}_T^1$. Namely, for constants $C_i = C_i(b, c; T)$

$$c_2 = C_1(b, c; T)\,\sigma_0 + C_2(b, c; T)\,a.$$

It is interesting to compare this with the Heston result [24] where the constant c_2 also depends linearly on spot-vol $\sigma_0 = \sqrt{v_0}$.

Solving the Hamiltonian ODEs and computing c_1 After replacing εdW by a control $d\mathrm{h}$, and taking $\varepsilon \downarrow 0$ elsewhere in (38), we have to consider the controlled ordinary differential equation

$$\begin{aligned} dy &= -\frac{1}{2}z^2 dt + z dh^1,\ y_0 = 0 \\ dz &= bz dt + c dh^2,\ \ z_0 = 0, \end{aligned} \tag{39}$$

minimizing the energy, $\frac{1}{2}\int_0^T \left|\dot{\mathrm{h}}_t\right|^2 dt$ subject to $y_T = \mathrm{a} \equiv 1 > 0$.

According to general theory, we now write out the Hamiltonian associated to (39),

$$\begin{aligned} &\mathcal{H}\left(\begin{pmatrix} y \\ z \end{pmatrix}, \begin{pmatrix} p \\ q \end{pmatrix}\right) \\ &= \begin{pmatrix} -\frac{1}{2}z^2 \\ bz \end{pmatrix} \cdot \begin{pmatrix} p \\ q \end{pmatrix} + \frac{1}{2}\left|\begin{pmatrix} z \\ 0 \end{pmatrix} \cdot \begin{pmatrix} p \\ q \end{pmatrix}\right|^2 + \frac{1}{2}\left|\begin{pmatrix} 0 \\ c \end{pmatrix} \cdot \begin{pmatrix} p \\ q \end{pmatrix}\right|^2 \\ &= -\frac{1}{2}z^2 p + bzq + \frac{1}{2}\left(z^2 p^2 + c^2 q^2\right). \end{aligned} \tag{40}$$

The Hamiltonian ODEs then become

$$\begin{aligned} \begin{pmatrix} \dot{y}_t \\ \dot{z}_t \end{pmatrix} &= \begin{pmatrix} z_t^2\left(p_t - \frac{1}{2}\right) \\ bz_t + c^2 q_t \end{pmatrix} \\ \begin{pmatrix} \dot{p}_t \\ \dot{q}_t \end{pmatrix} &= \begin{pmatrix} 0 \\ p_t z_t (1 - p_t) - bq_t \end{pmatrix}. \end{aligned} \tag{41}$$

Trivially, $p_t \equiv p_0$ which we shall denote by p from here on. As it turns out there is a simple expression for the energy. Although we shall ultimately take $\mathrm{a} \equiv 1$ it is convenient to carry out the following analysis for general $\mathrm{a} > 0$.

Lemma 27 *For any $\mathrm{h}_0 \in \mathcal{K}_{\mathrm{a}}^{\min}$, and in fact any h_0 given by (14), i.e.*

$$\dot{\mathrm{h}}_0(t) = \begin{pmatrix} p z_t \\ q_t c \end{pmatrix} \tag{42}$$

where $(y, z; p, q)$ satisfies (41), subject to boundary conditions $(y_0, z_0) = (0, 0)$ and $y_T = \mathrm{a}, q_T = 0$, we have

$$\Lambda(\mathrm{a}) = \frac{1}{2}\int_0^T \left|\dot{\mathrm{h}}_0(t)\right|^2 dt = p\mathrm{a}.$$

In particular, we see that

$$p \geq 0.$$

Remark 28 *In fact, linearity in* a *of (50) also follows immediately from the fact that the Stein-Stein model satisfies θ-scaling with $\theta = 2$ in the sense of corollary 14. Indeed, it was seen in the proof of that corollary that the rate function $\Lambda(\mathrm{a})$ scales like $\mathrm{a}^{2/\theta} = \mathrm{a}$. This already implies that p does not depend on* a*. This is also consistent with the principle $\partial_{\mathrm{a}}\Lambda(\mathrm{a}) = p_T$ pointed out in remark 5.*

Proof. We give an elegant argument based on the Hamiltonian ODEs. The idea is to express $\left|\dot{\mathrm{h}}_0(t)\right|^2$ as a time-derivative which then allows for immediate integration over $t \in [0,T]$. Indeed,

$$\begin{aligned}\left|\dot{\mathrm{h}}_0(t)\right|^2 &= p^2 z_t^2 + c^2 q_t^2 \\ &= p^2 z_t^2 + \partial_t(z_t q_t) - z_t^2\left(p^2 - p\right) \\ &= 2pz_t^2(p - 1/2) + \partial_t(z_t q_t) \\ &= 2p\dot{y}_t + \partial_t(z_t q_t)\end{aligned}$$

where we used the ODEs for z, q as given in (41). It follows that

$$\int_0^T \left|\dot{\mathrm{h}}_0(t)\right|^2 dt = 2p(y_T - y_0) + (z_T q_T - z_0 q_0)$$

and we conclude with the initial/terminal/transversality conditions $y_0 = z_0 = 0$, $y_T = \mathrm{a}$ and $q_T = 0$. ■

Lemma 29 (Partial Hamiltonian Flow) *Consider (41) as initial value problem, with initial data $(y_0, z_0) = (0,0)$ and (p, q_0). Assume*[19]

$$\chi_p^2 := c^2 p(p-1) - b^2 \geq 0. \tag{43}$$

Then the explicit solution is given by

$$\begin{aligned} y_t &= \frac{q_0^2 c^4(2p_0 - 1)}{8\chi_p^3}\left(2\chi_p t - \sin\left(2\chi_p t\right)\right), \\ z_t &= \frac{q_0 c^2}{\chi_p}\sin\left(\chi_p t\right), \\ p_t &\equiv p, \\ q_t &= q_0\left(\cos\left(\chi_p t\right) - \frac{b}{\chi_p}\sin\left(\chi_p t\right)\right). \end{aligned} \tag{44}$$

Remark 30 *The given solutions remain valid when $\chi_p^2 < 0$; it suffices to consider χ_p as purely imaginary; then, if desired, rewrite as $\cos\left(\chi_p t\right) = \cosh\left(\left|\chi_p\right| t\right)$ etc. Below, we shall solve (41) as boundary value problem, subject to $(y_0, z_0) = (0,0)$, $y_T = \mathrm{a} > 0$ and $q_T = 0$; we shall see then that (43) is always satisfied and in fact $\chi_p^2 > 0$.*

[19] All explicit solutions given in (44) are even functions of χ_{p_0} and have a removable singularity for $\chi_{p_0} = 0$. By convention we shall always assume $\chi_{p_0} \geq 0$ although the sign of χ_{p_0} does not matter.

Proof. Let us first remark that the path $(p_t)_{t\geq 0}$ is constant, $p_t = p$ for all $t \in [0,T]$. From the Hamiltonian ODEs, the couple $(z_t, q_t)_{t\geq 0}$ solves a linear ODE in $\mathbb{R}^2$, so that the solution must be a linear function of $(z_0, q_0) = (0, q_0)$. Indeed, a simple computation gives

$$q_t = q_0 \left(\cos\left(\chi_p t\right) - \frac{b}{\chi_p} \sin\left(\chi_p t\right) \right) \qquad \text{and} \qquad z_t = \frac{q_0 c^2}{\chi_p} \sin\left(\chi_p t\right),$$

Elementary integration ("$2\int_0^t \sin^2 = t - \cos\sin t$") then gives $(y_t)_{t\geq 0}$ by direct integration; indeed

$$y_t = \left(p - \frac{1}{2}\right) \int_0^t z_s^2 ds = \frac{q_0^2 c^4 \left(2p-1\right)}{8\chi_p^3} \left(2\chi_p t - \sin\left(2\chi_p t\right)\right).$$

This proves the lemma. ■

For the next proposition we recall the standing assumptions $T > 0$, $b \leq 0$ (which models mean-reversion) and $\mathrm{a} > 0$.

Proposition 31 *The ensemble of solutions to the Hamilton ODEs as boundary value problem*

$$(y_0, z_0) = (0,0) \text{ and } y_T = \mathrm{a}, q_T = 0$$

with $\mathrm{a} = 1 > 0$ *are characterized by inserting, for any* $k \in \{1, 2, ...\}$ *and any choice of sign in (46) below,*

$$p = p_k = \frac{1}{2}\left(1 + \sqrt{1 + \frac{4b^2}{c^2} + \frac{4r_k^2}{c^2 T^2}}\right), \tag{45}$$

$$q_{0,k}^{\pm} = \pm \frac{2}{c^2} \sqrt{\frac{2r_k^3 \,\mathrm{a}}{\left(2p_{0,k}^{+} - 1\right) T^3 \left(2r_k - \sin\left(2r_k\right)\right)}} \tag{46}$$

in (44). Here $\{r_k : k = 1, 2, \dots\}$ *denotes the set of (increasing) strictly positive roots to*

$$r\cos(r) - bT\sin(r) = 0.$$

Remark 32 *As the proof will show, p as given in (45) is the unique positive root to*

$$c^2 p\left(p-1\right) - b^2 = \left(\frac{r_{0,k}}{T}\right)^2;$$

in particular, assumption (43) in the previous lemma is met.

Proof. By assumption and (44),

$$0 = q_T = q_0 \left(\cos\left(\chi_p T\right) - \frac{b}{\chi_p} \sin\left(\chi_p T\right) \right). \tag{47}$$

At this stage, χ_p could be a complex number (when $\chi_p^2 < 0$). Let us note straight away that we must have $q_0 \neq 0$ for otherwise $(y_t)_{t\geq 0}$ - which depends linearly on q_0 as is seen explicitly in (44) - would be identically equal to zero in contradiction with $y_T = \mathrm{a} > 0$. Let us also note that

$\chi_p \neq 0$ for otherwise (47), which has a removable singularity at $\chi_p = 0$, leads to the contradiction $0 = 1 - bT$.(Recall $b \leq 0, T > 0$.) But then $r := \chi_p T$ is a root, i.e. maps to zero, under the map

$$r \in \mathbb{C} \mapsto r \cos r - bT \sin r = r \left(\cos r - \frac{bT}{r} \sin r \right). \tag{48}$$

A complex analysis lemma [27, Lemma 4] asserts that this map, provided

$$-bT \geq 0, \tag{49}$$

has only real roots; it follows that χ_p is real and so $\chi_p^2 \geq 0$; actually $\chi_p^2 > 0$, since we already noted that $\chi_p \neq 0$. Note that (47), and in fact all further expressions involving χ_p, are unchanged upon changing sign of χ_p, we shall agree to take $\chi_p > 0$ as the positive square-root of χ_p^2. In particular, (47) is equivalent to the existence of $\chi_p > 0$ such that

$$\chi_p T \cos\left(\chi_p T\right) - bT \sin\left(\chi_p T\right) = 0.$$

It follows that $\chi_p T \in \{r_k : k = 0, 1, 2, \dots\}$, the set of zeros of (48) written in increasing order. We deduce that, for each $k = 0, 1, 2, \dots$ there is a choice of p arising from

$$\chi_p^2 = c^2 p \left(p - 1\right) - b^2 = \left(\frac{r_k}{T}\right)^2.$$

For each k, there is a negative solution, say $p = p_k^- < 0$ which we may ignore thanks to lemma 27, and a positive solution, namely

$$p = p_k^+ = \frac{1}{2} \left(1 + \sqrt{1 + \frac{4b^2}{c^2} + \frac{4r_k^2}{c^2 T^2}} \right) > 1.$$

We now exploit $y_T = \mathrm{a}$. From the explicit expression of y_t given in (44) we get

$$\begin{aligned} \mathrm{a} = y_T &= \frac{q_0^2 c^4 \left(2p - 1\right)}{8 \chi_p^3} \left(2\chi_p T - \sin\left(2\chi_p T\right)\right) \\ &= \frac{q_0^2 c^4 \left(2p - 1\right) T^3}{8 r_k^3} \left(2r_k - \sin\left(2r_k\right)\right) \end{aligned}$$

and thus

$$q_0^2 = \frac{8 r_k^3}{c^4 \left(2p - 1\right) T^3 \left(2r_k - \sin\left(2r_k\right)\right)} \mathrm{a}.$$

It follows that, for each $k \in \{1, 2, \dots\}$, we can take

$$\begin{aligned} p = p_k^+ &= \frac{1}{2} \left(1 + \sqrt{1 + \frac{4b^2}{c^2} + \frac{4}{c^2} \left(\frac{r_k}{T}\right)^2} \right) \\ q_0 = q_{0,k}^{\pm} &= \pm \frac{2}{c^2} \sqrt{\frac{2 r_k^3 \, \mathrm{a}}{\left(2p_k^+ - 1\right) T^3 \left(2r_k - \sin\left(2r_k\right)\right)}} \end{aligned}$$

and any such choice in (44) leads to a solution of the boundary value problem. ■

So far, we have for each $k \in \{1, 2, \dots\}$ two choices of (p, q_0), depending on the sign in (46) so that the resulting Hamiltonian ODE solutions, started from $(y_0, z_0) = (0, 0)$ and $(p\,, q_0)$, describe *all* possible solutions of the boundary value problem given by the Hamiltonian ODEs with mixed initial/terminal data

$$(y_0, z_0) = (0, 0) \text{ and } y_T = \mathrm{a}, q_T = 0.$$

It remains to see which choice (or choices) lead to minimizing controls; i.e. $\mathrm{h}_0 \in \mathcal{K}_{\mathrm{a}}^{\min}$. But this is easy since we know from lemma 27 that, for any $p \in \left\{p_k^+ : k = 1, 2, \dots\right\}$,

$$\frac{1}{2}\int_0^T \left|\dot{\mathrm{h}}_0\left(t\right)\right|^2 dt = p\mathrm{a}.$$

Since p_k^+ is plainly (strictly) increasing in $k \in \{1, 2, \dots\}$, we see that the energy is minimal if and only if $p = p_1^+$. On the other hand, we are left with two choices for q_0, namely $q_{0,1}^+$ and $q_{0,1}^-$. Using (42) we then see that there are *two* minimizing controls,

$$\mathcal{K}_{\mathrm{a}}^{\min} = \left\{\mathrm{h}_0^+, \mathrm{h}_0^-\right\},$$

given by

$$\dot{\mathrm{h}}_0^{\pm}\left(t\right) = \begin{pmatrix} p\frac{q_0 c^2}{\chi_p}\sin\left(\chi_p t\right) \\ cq_0\left(\cos\left(\chi_p t\right) - \frac{b}{\chi_p}\sin\left(\chi_p t\right)\right) \end{pmatrix} \text{ with } (p, q_0) \leftarrow \left(p_1^+, q_{0,1}^+\right) \text{ resp. } \left(p_1^+, q_{0,1}^-\right).$$

Of course, $\mathrm{h}_0^{\pm}$ stands for h_0^+ resp. h_0^- depending on the chosen substitution above. In (y, z)-coordinates, note that both h_0^+ and h_0^- have identical y-components; their z-components only differ by a flipped sign due to $q_{0,1}^- = -q_{0,1}^+$. (This reflects a fundamental symmetry in our problem which is in fact invariant under $(y, z) \mapsto (y, -z)$). We summarize our finds in stating that

$$\Lambda\left(\mathrm{a}\right) = \frac{1}{2}\|\mathrm{h}_0^+\|_H^2 = \frac{1}{2}\|\mathrm{h}_0^-\|_H^2 = p_1^+\mathrm{a} \tag{50}$$

and upon taking $\mathrm{a} = 1$ we have computed the leading order constant

$$c_1 = \Lambda\left(1\right) = p_1^+ = \frac{1}{2}\left(1 + \sqrt{1 + \frac{4b^2}{c^2} + \frac{4}{c^2}\left(\frac{r_1}{T}\right)^2}\right)$$

where we recall that r_1 is the first strictly positive root of the equation $r\cos(r) - bT\sin(r) = 0$.

Computing c_2 According to general theory, cf. equation (18), we need to compute certain ODEs for each minimizer, $\mathrm{h}_0^+ = (h_{0,\cdot}^{+,1}, h_{0,\cdot}^{+,2})$ resp. $\mathrm{h}_0^- = (h_{0,\cdot}^{-,1}, h_{0,\cdot}^{-,2})$, exhibited in the previous section. For ease of notation we shall write $\left(p, q_0^{\pm}\right)$ instead of $\left(p_1^+, q_{0,1}^+\right)$ resp.$\left(p_1^+, q_{0,1}^-\right)$ in this section. Related to equation (38) we then have to consider the following ODE along h_0^+ (and then along h_0^-)

$$\begin{aligned} \frac{d}{dt}\begin{pmatrix}\hat{Y}_t \\ \hat{Z}_t^2\end{pmatrix} &= \left\{\begin{pmatrix}0 & -z_t^+ \\ 0 & b\end{pmatrix} + \begin{pmatrix}0 & 1 \\ 0 & 0\end{pmatrix}\dot{h}_{0,t}^{+,1}\right\}\begin{pmatrix}\hat{Y}_t \\ \hat{Z}_t^2\end{pmatrix} + \begin{pmatrix}0 \\ a\end{pmatrix} \\ &= \begin{pmatrix}0 & (p-1)\, z_t^+ \\ 0 & 0\end{pmatrix}\begin{pmatrix}\hat{Y}_t \\ \hat{Z}_t^2\end{pmatrix} + \begin{pmatrix}0 \\ a\end{pmatrix} \\ \text{with } \begin{pmatrix}\hat{Y}_0 \\ \hat{Z}_0^2\end{pmatrix} &= \begin{pmatrix}0 \\ \sigma_0\end{pmatrix}. \end{aligned}$$

Here, we used the fact that $\dot{h}_0^{+,1} = pz_t^+$, z_t^+ indicates the chosen sign of $q_{0,1}$ upon which it depends, cf. (46). The ODE along $\mathrm{h}_0^-$for $\hat{Y} = \hat{Y}^-$ is similar, with $z_t^+, \dot{h}_{0,t}^{+,1}$ replaced by $z_t^- = -z_t^+, \dot{h}_{0,t}^{-,1} = -\dot{h}_{0,t}^{+,1}$ respectively. We can solve these ODEs explicitly. In a first step (regardless of the chosen sign for z, h_0)

$$\hat{Z}_t = \begin{cases} \sigma_0 e^{bT} + \frac{a}{b}\left(e^{bt} - 1\right) & \text{for } b < 0 \\ \sigma_0 + at & \text{for } b = 0 \end{cases}$$

and since

$$\hat{Y}_T^{\pm} = (p-1)\int_0^T z_t^{\pm}\hat{Z}_t dt$$

we see that $\hat{Y}_T^- = -\hat{Y}_T^+$.In fact, under the (usual) model parameter assumptions $a > 0, \sigma_0 > 0$ we see that $\hat{Z}_t > 0$. We then note that

$$z_t^{\pm}/q_0^{\pm} = \frac{c^2}{\chi_p}\sin\left(\chi_p t\right) \geq 0 \text{ for } t \in [0,T]\,;$$

indeed we saw that $\chi_p T \in [\pi/2, \pi)$ which implies $\chi_p t \in [0, \pi)$ and hence $\sin\left(\chi_p t\right) \geq 0$. In particular, given that $q_0^+ > 0$ and $p > 1$we see that $\hat{Y}_T^+ > 0$ (and then $\hat{Y}_T^- < 0$). It follows that

$$\begin{aligned} c_2 \,:= c_2^+ &= \Lambda'(1) \times \hat{Y}_T^{+,1} \\ &= p(p-1)\int_0^T z_t^+ \hat{Y}_t^2 dt \end{aligned} \tag{51}$$

whereas the contribution from $c_2^- = \Lambda'(1) \times \hat{Y}_T^{-,1}$ is exponentially smaller and will not figure in the expansion (cf. remark 12). In fact, given the explicit form of $t \mapsto z_t^+$ resp. $\hat{Y}_t^2$ in terms of $\sin(.)$ and $\exp(.)$, it is clear that the integration in (51) can be carried out in closed form. In doing so, one exploits a cancellation due to

$$-\chi_p \cos\left(\chi_p T\right) + b\sin\left(\chi_p T\right) = 0$$

and also the equality $\chi_p^2 + b^2 = c^2 p(p-1)$, one is led to

$$c_2 = q_0^+ \left\{ \sigma_0 + a\frac{\tan\left(\chi_p T/2\right)}{\chi_p} \right\}.$$

It is possible, of course, to substitute the explicitly known quantities q_0^+, χ_p but this does not yield additional insight.

4.6.2 The case of non-zero correlation

We consider again the SDE (36) with diffusion matrix

$$\sigma = (\sigma_1, \sigma_2) = \begin{pmatrix} z & 0 \\ 0 & c \end{pmatrix}$$

but now allow for correlation ρ between W^1, W^2; we thus have the non-trivial correlation matrix

$$\Omega = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \implies \sigma\Omega\sigma^T = \begin{pmatrix} z^2 & \rho cz \\ \rho cz & c^2 \end{pmatrix}.$$

In view of financial applications [20] it makes sense to focus on the case $\rho \in (-1, 0]$. This will also prove convenient in our analysis below, although there is no doubt that the case $\rho > 0$, less interesting in practice, could also be handled within the present framework.

The Hamiltonian becomes, cf. (11),

$$\begin{aligned}\mathcal{H}\left(\begin{pmatrix} y \\ z \end{pmatrix}, \begin{pmatrix} p \\ q \end{pmatrix}\right) &= -\frac{1}{2}z^2 p + bzq + \frac{1}{2}\left(z^2p^2 + c^2q^2\right) + \rho czpq \\ &= -\frac{1}{2}z^2 p + \tilde{b}zq + \frac{1}{2}\left(z^2p^2 + c^2q^2\right)\end{aligned}$$

with

$$\tilde{b} := \tilde{b}_p := b + \rho cp$$

Noting $\partial_{(y,z)}\tilde{b} = (0,0)', \partial_{(p,q)}\tilde{b} = (\rho c, 0)'$. The Hamiltonian equations for $\dot{z}, \dot{p}, \dot{q}$, are thus identical as in the uncorrelated case, one just has to replace b by $\tilde{b}$. (In particular, p_t is again seen to be constant and we denote its value by p.) The Hamiltonian equation for $\dot{y} = \partial_p \mathcal{H}$ has, in comparison to the uncorrelated case, an additional term, namely $\left(\partial_p \tilde{b}\right) z_t q_t = \rho c z_t q_t$. In summary, the Hamiltonian ODEs are

$$\begin{aligned}\begin{pmatrix} \dot{y}_t \\ \dot{z}_t \end{pmatrix} &= \begin{pmatrix} z_t^2\left(p_t - \frac{1}{2}\right) + \rho c z_t q_t \\ \tilde{b}z_t + c^2 q_t \end{pmatrix} \\ \begin{pmatrix} \dot{p}_t \\ \dot{q}_t \end{pmatrix} &= \begin{pmatrix} 0 \\ p_t z_t\left(1 - p_t\right) - \tilde{b}q_t \end{pmatrix}.\end{aligned}$$

The following lemma is then obvious (only y requires a computation, due to the additional term in the Hamiltonian ODEs).

Lemma 33 (Partial Hamiltonian Flow, correlated case) *Consider the above Hamiltonian ODEs as initial value problem, with initial data $(y_0, z_0) = (0, 0)$ and (p, q_0) and assume*

$$\chi_p^2 := c^2 p\left(p - 1\right) - \tilde{b}_p^2 \geq 0. \tag{52}$$

Then the explicit solution for z, p, q are then identical to the uncorrelated case, one just has to replace b by $\tilde{b}_p$ throughout. The explicit solution for y is modified to

$$y_t = \frac{q_0^2 c^2}{8\chi_p^3}\left[\left(c^2\left(2p-1\right) - 2\rho c\tilde{b}_p\right)\left(2\chi_p t - \sin\left(2\chi_p t\right)\right) + 2\rho c\chi_p\left(1 - \cos\left(2\chi_p t\right)\right)\right]. \tag{53}$$

In our explicit analysis of the uncorrelated case (more precisely, in solving the coupled ODEs $\dot{z}_t = bz_t + c^2 q_t, \dot{q}_t = p_t z_t\left(1 - p_t\right) - bq_t$) we made use of the (model) assumption $b \leq 0$, cf. (49). Conveniently, this remains true when $\rho \in (-1, 0]$. Indeed, the following lemma shows we must have $p \geq 0$, so that (with $\rho \leq 0, c > 0$)

$$\tilde{b} = b + \rho cp \leq 0. \tag{54}$$

Lemma 34 *Let* $\mathrm{a} > 0$. *Then* $\Lambda\left(\mathrm{a}\right) = p\mathrm{a}$ *and therefore* $p \geq 0$.

Proof. We saw in the proof of lemma 27 that, in the uncorrelated case, as a direct consequence of the Hamiltonian ODEs,

$$p^2 z_t^2 + c^2 q_t^2 = 2p\dot{y}_t + \partial_t \left(z_t q_t \right).$$

The correlated case has the identical Hamiltonian ODEs provided we substitute

$$b \leftarrow \tilde{b} \text{ and } \dot{y} \leftarrow \dot{y} - \rho c z_t q_t.$$

We therefore have

$$\begin{aligned} \left| \dot{\mathrm{h}}_0 \left(t \right) \right|^2 &= \left(p \quad q_t \right) \begin{pmatrix} z^2 & \rho c z \\ \rho c z & c^2 \end{pmatrix} \begin{pmatrix} p \\ q_t \end{pmatrix} = p^2 z_t^2 + c^2 q_t^2 + 2\rho c p z_t q_t \\ &= 2p \left(\dot{y}_t - \rho c z_t q_t \right) + \partial_t \left(z_t q_t \right) + 2\rho c p z_t q_t = 2p\dot{y}_t + \partial_t \left(z_t q_t \right) \end{aligned}$$

and then conclude with the boundary data, exactly as in lemma 27. ■

As already noted, $\tilde{b} \leq 0$ allows to recycle all closed form expressions for z, q obtained in the uncorrelated case - it suffices to replace b by $\tilde{b}$. In particular, for some yet unknown p, q_0 which may and will depend on ρ,

$$\begin{aligned} z_t &= \frac{q_0 c^2}{\chi_p} \sin\left(\chi_p t \right), \\ q_t &= q_0 \left(\cos\left(\chi_p t \right) - \frac{\tilde{b}}{\chi_p} \sin\left(\chi_p t \right) \right) \end{aligned}$$

where $\chi_p^2 := c^2 p \left(p - 1 \right) - \tilde{b}^2$ is seen to be positive as in the "uncorrelated" argument. Also, $q_0 \neq 0$, seen as in the "uncorrelated" case. Transversality, $q_T = 0$, then implies

$$\chi_p \cos\left(\chi_p T \right) - \tilde{b} \sin\left(\chi_p T \right) = 0. \tag{55}$$

Introducing $r := \chi_p T$ the gives the equation

$$r \cot r = \left(b + \rho c p \right) T. \tag{56}$$

On the other hand, from the very definition of χ_p, we know

$$\left(r/T \right)^2 = c^2 p \left(p - 1 \right) - \left(b + \rho c p \right)^2. \tag{57}$$

In the uncorrelated case, these two equations were effectively decoupled; in particular, $r \cot r = bT$ lead to $r \in \left\{ r_k^+ : k = 1, 2, \dots \right\} \subset (0, \infty)$, written in increasing order. Since p^+ was seen to be monotonically increasing in r, cf. equation (45), and we were looking for the minimal p, corresponding to the minimal energy (cf. lemma 34), we were led to seek the first positive root r_1^+. (In fact, $r_1^+ \in (\pi/2, \pi)$ as we will also find in the "correlated" discussion below.)

The correlated case is a little more complicated and we start in expressing p in equation (56) in terms of r. Indeed, the quadratic equation (57) shows

$$p^{\pm}(r) = \frac{1}{2\left(1 - \rho^2\right)} \left\{ \left(1 + 2\rho \frac{b}{c} \right) \pm \sqrt{\left(1 + 2\rho \frac{b}{c} \right)^2 + 4\left(1 - \rho^2\right) \left[\frac{b^2}{c^2} + \frac{r^2}{c^2 T^2} \right]} \right\}, \tag{58}$$

where $p^-(r) < 0$ (and hence can be ignored in view of lemma 34) and $p^+(r) > 0$. We now look for

r which satisfies the equation

$$r\cot r = \left(b + \rho c p^+(r)\right) T$$

It is elementary to see that $r\cot r$ is non-negative on $[0, \pi/2]$ and then maps $[\pi/2, \pi)$ strictly monotonically to $(-\infty, 0]$. On the other hand, the map $r \mapsto (b + \rho c p^+(r))T$ is ≤ 0 for all r; in particular,

there will be a first intersection with the graph of $r \mapsto r\cot r$ in $[\pi/2, \pi)$, say at $r = r_1^+$. Since $p^+(r)$ is plainly strictly increasing in r, the minimal p must equal to

$$p_1^+ := p^+\left(r_1^+\right).$$

We then proceed as in the uncorrelated case, and determine q_0 from the boundary condition $y_T = \mathrm{a} > 0$ where y is now given by (53). This leads to $q_0 \in \left\{q_{0,1}^+, q_{0,1}^-\right\}$ where

$$q_{0,1}^{\pm} = \pm\frac{2}{c}\sqrt{\frac{2r^3\,\mathrm{a}}{T^3\left(\left(c^2(2p-1) - 2\rho c\tilde{b}\right)(2r - \sin(2r)) + 2\rho c r/T\,(1-\cos(2r))\right)}}$$

where $r = r_1^+$ and $p = p_1^+$. Again, we have *two* minimizing controls, $\mathcal{K}_{\mathrm{a}}^{\min} = \left\{\mathrm{h}_0^+, \mathrm{h}_0^-\right\}$. We now have

$$\dot{\mathrm{h}}_0(t) = \begin{pmatrix} z_t\sqrt{1-\rho^2} & 0 \\ \rho z_t & c \end{pmatrix}\begin{pmatrix} p \\ q_t \end{pmatrix} \tag{59}$$

instead of (42) and of course lemma 33 implies that z_t and q_t are fully and explicitly determined for each choice of (p, q_0). In particular for $(p, q_0) \leftarrow \left(p_1^+, q_{0,1}^+\right)$ resp. $\left(p_1^+, q_{0,1}^-\right)$ we so obtain h_0^+ resp. h_0^- which can be written explicitly by simple substitution. Moreover, and again as in the uncorrelated case,

$$\Lambda(\mathrm{a}) = \frac{1}{2}\|\mathrm{h}_0^+\|_H^2 = \frac{1}{2}\|\mathrm{h}_0^-\|_H^2 = p_1^+\mathrm{a} \tag{60}$$

and upon taking $\mathrm{a} = 1$ we have computed the leading order constant

$$c_1 = \Lambda(1) = p_1^+ \equiv p^+\left(r_1^+\right)$$

where we recall that r_1^+ is the first intersection point of $r \mapsto r\cot r$ with $(b + \rho c p^+(r))T$ and $p^+(\cdot)$

was given in (58).

At last, we turn to the computation of the second-order exponential constant, c_2. As in the uncorrelated case, we ease notation by writing $\left(p, q_0^{\pm}\right)$ instead of $\left(p_1^+, q_{0,1}^+\right)$ resp. $\left(p_1^+, q_{0,1}^-\right)$ for the rest of this section. Again, we have to consider ODEs for $\left(\hat{Y}_t, \hat{Z}_t\right)$, for each minimizer, $\mathrm{h}_0^+ = (h_{0,\cdot}^{+,1}, h_{0,\cdot}^{+,2})$ and $\mathrm{h}_0^- = (h_{0,\cdot}^{+,1}, -h_{0,\cdot}^{+,2})$. Recall from (59) that, with $\bar{\rho} = \sqrt{1-\rho^2}$,

$$\dot{\mathrm{h}}_0^+(t) = \begin{pmatrix} p\bar{\rho} z_t^+ \\ \rho p z_t^+ + c q_t^+ \end{pmatrix};$$

where $(\cdot)^{\pm}$ indicates the chosen sign of $q_0 \in \left\{q_{0,1}^+, q_{0,1}^-\right\}$ which determines the choice of minimizer. We first determine $\hat{Y}_T = \hat{Y}_T\left(\mathrm{h}_0^+\right)$ from the ODE

$$\begin{aligned}\frac{d}{dt}\begin{pmatrix}\hat{Y}_t\\ \hat{Z}_t^2\end{pmatrix} &= \left\{\begin{pmatrix}0 & -z_t^+\\ 0 & b\end{pmatrix} + \begin{pmatrix}0 & \bar{\rho}\\ 0 & 0\end{pmatrix}\dot{h}_{0,t}^{+,1} + \begin{pmatrix}0 & \rho\\ 0 & 0\end{pmatrix}\dot{h}_{0,t}^{+,2}\right\}\begin{pmatrix}\hat{Y}_t\\ \hat{Z}_t\end{pmatrix} + \begin{pmatrix}0\\ a\end{pmatrix}\\ &= \begin{pmatrix}0 & (p-1)\, z_t^+ + \rho c q_t^+\\ 0 & b\end{pmatrix}\begin{pmatrix}\hat{Y}_t\\ \hat{Z}_t\end{pmatrix} + \begin{pmatrix}0\\ a\end{pmatrix}\\ \text{with } \begin{pmatrix}\hat{Y}_0\\ \hat{Z}_0^2\end{pmatrix} &= \begin{pmatrix}0\\ \sigma_0\end{pmatrix}.\end{aligned}$$

This already shows that we have the identical (closed form) ODE solution for $\hat{Z}_t$ as in the uncorrelated case. On the other hand, the form of $\hat{Y}_T$ now exhibits an additional term as is seen in

$$\hat{Y}_T = (p-1)\int_0^T z_t^+ \hat{Z}_t dt + \rho c \int_0^T q_t^+ \hat{Z}_t dt.$$

Since q_t^+ is essentially of the same trigonometric form as z_t^+, it is clear that the explicit computations of the uncorrelated case extend. In the end, one finds without too much difficulties

$$c_2^+ = \Lambda'(1) \times \hat{Y}_T\left(\mathrm{h}_0^+\right) = q_0^+\left\{\sigma_0 + a\frac{\tan\left(\chi_p T/2\right)}{\chi_p}\right\}.$$

A similar computation along h_0^- gives $c_2^+ = \Lambda'(1) \times \hat{Y}_T\left(\mathrm{h}_0^-\right)$ in explicit form and c_2 is identified as $\max\left(c_2^+, c_2^-\right)$.

4.6.3 Checking non-degeneracy, zero and non-zero correlation

We now check the non-degeneracy condition (ND), as introduced in definition 8, which of course is the ultimate justification that an expansion of the form (37) with the constants computed above holds true. Again, focus is on the case of correlation parameter $\rho \in (-1, 0]$. We saw in the previous sections (for $\rho = 0$, then $\rho \leq 0$) that $\#K_{\mathrm{a}}^{\min} = \#\left\{\mathrm{h}_0^+, \mathrm{h}_0^-\right\} = 2$, whenever $\mathrm{a} > 0$. (In fact, we apply this with $\mathrm{a} = 1$.)

Secondly, a look at (39) reveals that the *degenerate region* is $\{(y, z) : z = 0\}$, the complement of which is elliptic. Clearly, no controlled path which reaches $y_T = \mathrm{a} > 0$ can stay in the degenerate region for all times $t \in [0, T]$; after all, this would entail $dy = 0$ and hence $y_T = 0$. We conclude the any ODE solution driven by $h \in \mathcal{K}_a$ must intersect the region of ellipticity; but this already implies non-degeneracy of the corresponding (deterministic) Malliavin covariance matrix.

At last, we check non-focality and focus on h_0^+, the other case being similar. We have to check non-degeneracy of the Jacobian of the map $\pi H_{0\leftarrow T}(\mathrm{a}, \cdot; *, 0)$, evaluated at $\cdot = z_T, * = p_T$ after differentiation, where z_T, p_T are obtained form the Hamiltonian flow at time T, cf. lemma 33, with time 0 initial data $\left(0, 0; p_1^+, q_{0,1}^+\right)$.

With some abuse of notation, write

$$\begin{pmatrix}y_0\\ z_0\end{pmatrix} \equiv \begin{pmatrix}y_0(z,p)\\ z_0(z,p)\end{pmatrix} \equiv \pi H_{0\leftarrow T}(\mathrm{a}, z; p, 0).$$

Our non-degeneracy condition requires us to show that

$$\det \left(\begin{array}{cc} \frac{\partial y_0}{\partial p} & \frac{\partial y_0}{\partial z} \\ \frac{\partial z_0}{\partial p} & \frac{\partial z_0}{\partial z} \end{array} \right) \Bigg|_* \neq 0 \tag{61}$$

where $(...)\,|_*$ indicates evaluation $(...)\,|_{(p,z)=(p^+,z_T)}$ in the sequel. This implies in particular that all expressions which are formulated in terms of the solutions to the Hamiltonian flows, reduced to the corresponding expressions identified in proposition 31, for $\rho = 0$, resp. in section 4.6.2 for $\rho \leq 0$. For instance, $(y_0, z_0)\,|_* = (0,0)\,, y_T|_* = \mathrm{a}, z|_* = z_T \neq 0$, $\chi_p T|_* \in [\pi/2, \pi)$ and so.

Since $(z_\cdot, q_\cdot)$ solves a linear ODE, we can compute

$$\begin{aligned} z_0\,(z,p) &= \left(1\,0\right)\, e^{-T \begin{pmatrix} \tilde{b}_p & c^2 \\ p\,(1-p) & -\tilde{b}_p \end{pmatrix}} \begin{pmatrix} z \\ 0 \end{pmatrix} \\ &= \frac{z}{\chi_p} \left(\chi_p \cos\left(\chi_p T\right) - \tilde{b}_p \sin\left(\chi_p T\right) \right). \end{aligned}$$

We first note that $\partial z_0/\partial z|_*$ is zero; indeed, this follows from (55). Our next claim is $\partial y_0/\partial z|_* \neq 0$. Indeed, from the structure of the Hamilton ODEs,

$$y_0 - \mathrm{a} = -\int_0^T \dot{y}_t dt = z^2\,(...)$$

where $(\cdots)$ does not depend on z. As a result $\partial y_0/\partial z|_* = 2z\,(...)\,|_* = 2\frac{y_0 - \mathrm{a}}{z}|_* = -2\mathrm{a}/z_T \neq 0$.

It remains to check that $\partial z_0/\partial p|_* \neq 0$. To this end, recall, as a consequence of the transversality condition, see (55), that $\chi_p \cos\left(\chi_p T\right) - \tilde{b}_p \sin\left(\chi_p T\right)|_* = 0$. It follows that

$$\partial z_0/\partial p|_* = \left\{ \frac{z}{\chi_p} \frac{\partial}{\partial p} \left(\chi_p \cos\left(\chi_p T\right) - \tilde{b}_p \sin\left(\chi_p T\right) \right) \right\}_*$$

and since $z/\chi_p|_* \neq 0$, it will be enough to show (strict) negativity of $\frac{\partial}{\partial p}\,(...)\,|_*$ above. By scaling, there is no loss of generality in taking $T = 1$ and we shall do so from here on. Then

$$\begin{aligned} &\frac{\partial}{\partial p} \left(\chi_p \cos\left(\chi_p\right) - \tilde{b}_p \sin\left(\chi_p\right) \right) \\ &= \chi_p' [\left(1 - \tilde{b}_p\right) \cos\left(\chi_p\right) - \chi_p \sin\left(\chi_p\right)] - \rho c \sin\left(\chi_p\right). \end{aligned}$$

Since $\tilde{b}_p|_* \leq 0$ and $\chi_p|_* \in [\pi/2, \pi)$ we see that $[...]\,|_* < 0$. Given that $\chi_p'|_* > 0$, this already settles the negativity claim in the zero-correlation case. In the case $-1 < \rho < 0$, we use (55) to write

$$\begin{aligned} &\frac{\partial}{\partial p} \left(\chi_p \cos\left(\chi_p\right) - \tilde{b}_p \sin\left(\chi_p\right) \right) |_* \\ &= \chi_p' [\left(1 - \tilde{b}_p\right) \frac{\tilde{b}_p \sin\left(\chi_p\right)}{\chi_p} - \chi_p \sin\left(\chi_p\right)] - \rho c \sin\left(\chi_p\right) |_*. \end{aligned}$$

After division by $\sin\left(\chi_p\right)/\chi_p|_* > 0$, we have, using $\tilde{b}_p = b + \rho cp \leq 0$, $b \leq 0$ and again $\chi_p'|_* > 0$,

$$\begin{aligned}
&\chi_p'[\left(1-\tilde{b}_p\right)\tilde{b}_p - \chi_p^2] - \rho c\chi_p|_* \\
\leq &\chi_p'[(1-\rho cp)\,\rho cp - \chi_p^2] - \rho c\chi_p|_* \\
\leq &-\rho c\left(\chi_p - p\chi_p'\right)|_*.
\end{aligned}$$

With $-\rho c > 0$, it will then be sufficient to show strict negativity of $\chi_p - p\chi_p'|_*$. To this end note that the definition, $\chi_p^2 = c^2 p\,(p-1) - \tilde{b}^2$, implies

$$\begin{aligned}
2\chi_p\chi_p' &= c^2\,(2p-1) - 2\tilde{b}\,(\rho c) \\
\chi_p p\chi_p' &= c^2 p\,(p-1/2) - \tilde{b}\,(\rho cp) \\
&= \chi_p^2 + \frac{c^2 p}{2} + b\tilde{b} > \chi_p^2
\end{aligned}$$

whenever $c^2 p/2 + b\tilde{b} > 0$ which is surely the case upon evaluation $...|_*$.

We conclude that $\partial z_0/\partial p|_* \neq 0$, and then validity of (61), for any parameter set $\rho \in (-1,0], b \leq 0, c > 0, T > 0$. In other words, we have completed the check of our non-degeneracy condition.

4.7 Comments on Heston [32] and Lions–Musiela [38]

We recall from [27, 24] that the density of log-stock price Y_T in the Heston model,

$$\begin{aligned}
dY &= -V/2 + \sqrt{V}dW^1,\ X\,(0) = x_0 = 0 \\
dV &= (a+bV)\,dt + c\sqrt{V}dW^2,\ V\,(0) = v_0 > 0,
\end{aligned}$$

with $a \geq 0, b \leq 0, c > 0$ and correlation $\rho \in (-1,0]$ has the form

$$f\,(y) = e^{-c_1 y}e^{c_2\sqrt{y}}y^{-3/4+a/c^2}\left(c_3 + O\left(1/\sqrt{y}\right)\right)\ \text{as}\ y \to \infty;$$

with explicitly computable $c_1 = C_1\,(b,c,\rho,T)$ and $c_2 = \sqrt{v_0} \times C_2\,(b,c,\rho,T)$, both do not depend on a. While **scaling** with $\theta = 2$,

$$Y_\varepsilon := \varepsilon^2 Y, \quad V_\varepsilon := \varepsilon^2 V$$

indeed yields a small noise problem, namely

$$\begin{aligned}
dY^\varepsilon &= -V^\varepsilon/2 + \sqrt{V^\varepsilon}\varepsilon dW^1,\ X\,(0) = x_0 = 0 \\
dV^\varepsilon &= \left(a\varepsilon^2 + bV^\varepsilon\right)dt + c\sqrt{V^\varepsilon}\varepsilon dW^2,\ V\,(0) = v_0\varepsilon^2 > 0.
\end{aligned}$$

The algebraic factor $y^{-3/4+a/c^2}$ in the above expansion then contradicts the expected factor; cf. (24)

$$y^{\frac{1}{\theta}-1} = y^{-1/2}.$$

There is no contradiction here, of course. Rather, we see an explicit example where "formal" application of a theorem to a model which is short of the required regularity leads to wrong conclusion (at least at the fine level of algebraic factors). Remark that one can trace the origin of this unexpected $y^{-3/4+a/c^2}$ factor to the behaviour of the one-dimensional variance process V; also known as Feller - or Cox-Ingersoll-Ross diffusion. Curiously then *even a large deviation principle for* V^ε

as given above presently lacks justification, despite the recent advances in [17], [6]. Clearly then, we are not anywhere near in obtaining the Heston tail result of [27, 24] with the present methods.

However, in the special case when $a = c^2/4$ it is an easy exercise to see that the Heston model can be realized as Stein-Stein model (take $V = Z^2$, where Z is the volatility component of the Stein-Stein model), the resulting expressions are then seen to be consistent with those obtained in [24] and, in particular, $y^{-3/4+a/c^2} = y^{-1/2}$.

Another class of non-smooth, non-affine stochastic vol model with "$\theta = 2$"-scaling, introduced by Lions-Musiela [38]. For $\delta \in [1/2, 1]$ and $\gamma = 1 - \delta$ they consider the 2-dimensional diffusion

$$\begin{aligned} dY &= -\frac{1}{2} Z^{2\delta} dt + Z^{\delta} d\tilde{W}_1, \;\; Y_0 = 0 \\ dZ &= bZdt + cZ^{\gamma} dW_2, \quad Z(0) = z_0 > 0. \end{aligned}$$

And indeed with $Y_\varepsilon = \varepsilon^2 Y$ and $Z_\varepsilon = \varepsilon^{1/\delta} Z$ this becomes a small noise problem;

$$\begin{aligned} dY_\varepsilon &= -\frac{1}{2} Z_\varepsilon^{2\delta} dt + Z_\varepsilon^{\delta} \varepsilon dW, \quad Y_\varepsilon(0) = 0 \\ dZ_\varepsilon &= bZ_\varepsilon dt + cZ_\varepsilon^{\gamma} \varepsilon dZ, \;\; Z_\varepsilon(0) = \varepsilon^{1/\delta} z_0. \end{aligned}$$

In their paper they establish exponential moments of Y_T in the precise sense (32),(33). It is tempting to use corollary 14, at least to leading large deviation order, to obtain the exponential tail of Z for models that scale with $\theta = 2$. Of course, as was discussed in the Heston case, such a "formal" application can be wrong. Further work, building on [17], [6], will be necessary to deal with such degenerate models directly.

References

[1] Y. Ait-Sahalia, Closed-form likelihood expansions for multivariate diffusions. *The Annals of Statistics* 2008, Vol. 36, No. 2, 906–937

[2] M. Avellaneda, D. Boyer-Olson, J.Busca, P. Friz: Application of large deviation methods to the pricing of index options in finance, *Comptes Rendus de l'Académie des Sciences - Series I - Mathematique* (2003).

[3] M. Avellaneda, D. Boyer-Olson, J.Busca, P. Friz: Reconstructing volatility, RISK (2004).

[4] R. Azencott. Formule de Taylor stochastique et développement asymptotique d'intégrales de Feynmann. *Séminaire de Probabilités XVI; Supplément: Géométrie différentielle stochastique. Lecture notes in Mathematics*, 921, 237-285, 1982.

[5] R. Azencott. Petites perturbations aléatoires des systèmes dynamiques: développements asymptotiques. *Bulletin des sciences mathématiques.* vol. 109, no3, pp. 253-308, 1985.

[6] P. Baldi and L. Caramellino. General Freidlin-Wentzell large deviations and positive diffusions. Forthcoming in *Statistics and Probability Letters*, 2011.

[7] Benaim, Friz: Regular Variation and Smile Asymptotics, Math. Finance Vol. 19 no 1. (2009), 1-12

[8] G. Ben Arous. Développement asymptotique du noyau de la chaleur hypoelliptique hors du cut-locus. *Annales Scientifiques de l'Ecole Normale Supérieure*, 4 (21): 307-331, 1988.

[9] G. Ben Arous. Methods de Laplace et de la phase stationnaire sur l'espace de Wiener. *Stochastics*, 25: 125-153, 1988.

[10] G. Ben Arous, P. Laurence: Second order expansion for implied volatility in two factor local-stochastic volatility models and applications to the dynamic Sabr model. Preprint 2010.

[11] G. Ben Arous, R. Léandre: Décroissance exponentielle du noyau de la chaleur sur la diagonale (I), Probab. Th. Re1. Fields 90, 175 202 (1991)
[12] R. Bishop, R. Crittenden, Geometry of Manifolds, Academic Press 1964.
[13] H. Berestycki, J. Busca, and I. Florent. Computing the implied volatility in stochastic volatility models. Communications on Pure and Applied Mathematics, 57(10):1352-1373, 2004.
[14] J.M. Bismut. Malliavin Calculus and Large Deviations. 1984
[15] P Bourgade and O Croissant. Heat kernel expansion for a family of stochastic volatility models : -geometry; arXiv:cs.CE/0511024, 2005.
[16] J.D. Deuschel and D.W. Stroock. Large Deviations. Volume 342 of AMS/Chelsea Series. 2000
[17] C. Donati-Martin, A. Rouault, M. Yor and M. Zani. Large deviations for squares of Bessel and Ornstein-Uhlenbeck processes. *PTRF*, 129(2), 261-289.
[18] M. Freidlin and A.D. Wentzell. Random perturbations of dynamical systems. Grundlehren der Mathematischen Wissenschaften (Second edition ed.). New York: Springer-Verlag, 1998.
[19] Friz, Peter; Victoir, Nicolas; Multidimensional Stochastic Processes as Rough Paths. Theory and Applications, Cambridge Studies of Advanced Mathematics Vol. 120, 670 p., Cambridge University Press
[20] Jim Gatheral. The Volatilty Surface. Wiley Finance, 2006.
[21] Gatheral, Jim; Further Developments in Volatility Derivatives Modeling. Presentation 2008. Available on www.math.nyu.edu/fellows_fin.../gatheral/FurtherVolDerivatives2008.pdf
[22] Gatheral, Jim; Hsu, Elton P.; Laurence, Peter; Ouyang, Cheng; Wang, Tai-Ho. Asymptotics of Implied Vol in Local Vol Models. *Math. Finance*, 2011 to appear.
[23] Gaveau B.: Principe de moindre action, propagation de la chaleur et estimées sous-elliptiques sur certains groupes nilpotents. Acta. Math. 139 (1977), 95- 153.
[24] P. Friz, S. Gerhold, A. Gulisashvili and S. Sturm. Refined implied volatility expansions in the Heston model. Quant. Finance, Volume 11, Issue 8, 1151-1164, 2011.
[25] Jim Gatheral and Antoine Jacquier, Convergence of Heston to SVI, Quant. Finance 2011, Volume 11, Issue 8, 2011.
[26] R. Giambo, F. Giannoni, P. Piccione and D. V. Tausk, Morse theory for normal geodesics in sub-Riemannian manifolds with codimension one distributions, Topological Methods in Nonlinear Analysis, Journal of the Julius Schauder Center Volume 21, 2003, 273–291
[27] A. Gulisashvili and E. Stein. Asymptotic Behavior of the Stock Price Distribution Density and Implied Volatility in Stochastic Volatility Models, Applied Mathematics & Optimization, Volume 61, Number 3, 287-315, DOI: 10.1007/s00245-009-9085-x
[28] I. Gyöngy. Mimicking the one-dimensional marginal distributions of processes having an Itô differential. Probab. Theory Relat. Fields, 71(4):501–516, 1986
[29] K. Gao and R. Lee. Asymptotics of implied volatility to arbitrary order. Preprint available at http://ssrn.com/abstract=1768383, 2011.
[30] Patrick Hagan, Andrew Lesniewski, and Diana Woodward; Probability Distribution in the SABR Model of Stochastic Volatility. Working Paper 2005. Available on lesniewski.us/working.html
[31] Henry-Labordère P, Analysis, geometry and modeling in finance, Chapman and Hill/CRC, 2008.
[32] Heston S. 1993. A closed-form solution for options with stochastic volatility, with application to bond and currency options. Review of Financial Studies 6, 327–343.
[33] Jurdjevic, Kupka; Polynomial Control Systems; Math. Ann. 272, 361-368 (1985)
[34] Yu. I. Kifer, "On the asymptotics of the transition probability density of processes with small

diffusion", Teor. Veroyatnost. i Primenen., 21:3 (1976), 527–536
[35] Shigeo Kusuoka and Yasufumi Osajima: A remark on the asymptotic expansion of density function of Wiener functionals. UTMS Preprint 2007-18.
[36] Shigeo Kusuoka and Daniel W. Stroock, Precise asymptotics of certain Wiener functionals. Journal of Functional Analysis, Volume 99, Issue 1, July 1991, Pages 1-74.
[37] Roger Lee, "The Moment Formula for Implied Volatility at Extreme Strikes " Mathematical Finance, vol 14 issue 3 (July 2004), 469-480.
[38] P.L. Lions and M. Musiela. Correlations and bounds for stochastic volatility models. *Ann. I.H. Poincaré*, 24, 2007, 1-16.
[39] Alex Lipton and Artur Sepp, Stochastic volatility models and Kelvin waves. 2008, *J. Phys. A: Math. Theor. 41.*
[40] S A Molchanov, "Diffusion processes and Riemannian geometry", Russ. Math. Surv., 1975, 30 (1), 1–63.
[41] R. Montgomery. A Tour of SubRiemannian Geometries, their Geodesics and Applications, Volume 91 of Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2002.
[42] Osajima, Yasufumi, General Asymptotics of Wiener Functionals and Application to Mathematical Finance (July 25, 2007). Available at SSRN: http://ssrn.com/abstract=1019587
[43] Paulot, Louis, Asymptotic Implied Volatility at the Second Order with Application to the SABR Model (June 3, 2009). Available at SSRN: http://ssrn.com/abstract=1413649
[44] Huyen Pham, Large deviations in Finance, 2010, Third SMAI European Summer School in Financial Mathematics.
[45] V. Piterbarg, Markovian projection method for volatility calibration; available at SSRN: http://ssrn.com/abstract=906473, 2006.
[46] Sakai, T.: Riemannian Geometry, AMS, 1992.
[47] Seierstad, A. and Sydsaeter, K.: Optimal Control Theory with Economic Applications. (Advanced Textbooks in Economics, 24). North- Holland Amsterdam, 1987
[48] Stein, E. M., and J. C. Stein, 1991, "Stock Price Distributions with Stochastic Volatility: An Analytic Approach," Review of Financial Studies, 4, 727-752.
[49] Takanobu S. Watanabe S.: Asymptotic expansion formulas of the Schilder type for a class of conditional Wiener functional integration. In "Asymptotics problems in probability theory: Wiener functionals and asymptotics". K.D. Elworthy N. Ikeda edit. Pitman. Res. Notes. Math. Series. 284 (1993), 194-241.
[50] Varadhan, S. R. S., On the behavior of the fundamental solution of the heat equation with variable coefficients. Communications on Pure and Applied Mathematics, 20: 431–455. 1967
[51] Varadhan, S.R.S.: Lectures on large deviations, available at http://math.nyu.edu/faculty/varadhan/LDP.html, 2010.