Markovian stochastic volatility with stochastic correlation - joint calibration and consistency of SPX/VIX short-maturity smiles

Martin Forde Benjamin Smith*

November 8, 2025

Abstract

We show how to calibrate a general Markovian stochastic volatility model with stochastic correlation to the VIX implied volatility smile and the overall level, slope and curvature of the SPX smile in the $T \to 0$ limit. Explicit formulae are obtained for the asymptotic VIX smile for Heston and SABR-type models with mean reversion, and the Lewis CEV-p-model. We also discuss how the Bass martingale can be used to give an exact fit to a single VIX smile for $T > 0$. In the second half of the article, we derive a more involved integral equation for the correlation function $\rho(y)$ to be perfectly consistent with the short-maturity SPX and VIX smiles at all strikes (or all strikes in an interval) as $T \to 0$, and discuss consistency conditions between the wings of the two asymptotic smiles and how to avoid $|\rho(y)| > 1$ for the calibrated $\rho(y)$ in practice.

1 Stochastic volatility with stochastic correlation

1.1 Introduction

The theoretical value of the VIX index at time t is $\text{VIX}_t = \sqrt{-\frac{2}{\Delta}\mathbb{E}^{\mathbb{Q}}(\log \frac{S_{t+\Delta}}{S_t}|\mathcal{F}_t)}$ where S_t is the S&P 500 index value at time t, $\Delta = 30$ days, $(\mathcal{F}_t)_{t\geq 0}$ is the market/model filtration and $\mathbb{Q}$ is the pricing measure, so VIX_t^2 is effectively a rolling 30-day Variance swap rate. A VIX option is a European call or put option on VIX_T for some maturity T, and if we replace the spot value S_0 in the Black-Scholes formula with the VIX future price $\mathbb{E}^{\mathbb{Q}}(\text{VIX}_T)$ where $\mathbb{Q}$ is the pricing measure, we can define the implied volatility of a VIX call or put in the usual way by inverting the Black-Scholes formula. VIX options are very liquid in practice (although their bid/offer spreads are still comparatively high).

In this article, we first show how to calibrate a general Markovian stochastic volatility model to the VIX implied volatility smile as $T \to 0$ and the level, slope and curvature of the SPX asymptotic smile at-the-money. The instantaneous correlation between the Brownians for the model is a deterministic function of the instantaneous variance process Y (and hence stochastic), and our methodology requires that the observed SPX and VIX asymptotic smiles exhibit $H = \frac{1}{2}$-type behaviour as $T \to 0$, which is consistent with some recent empirical findings ([GL22], [Rom22b], [AIL22],[AIL22b]). We also give explicit formulae for the asymptotic VIX smile for Heston, Hull-White and CEV-p-type models, and we find that the SABR model produces a more realistic (i.e. increasing) asymptotic VIX smile when the mean reversion $\kappa > 0$ (as opposed to flat when $\kappa = 0$). The calibration is obtained via a power series expansion (in log-moneyness) to the eikonal equation for the geodesic distance for the model, and one can also go to higher order in the calibration. Hence the novel stochastic correlation feature of the model allows for decoupling between the SPX and VIX smiles in some sense, since $\rho'(y_0)$ provides an extra degree of freedom.

In the second part of the article, we derive a more involved integral equation for the correlation function $\rho(y)$ to be perfectly consistent with the short-maturity SPX and VIX smiles at all strikes (or all strikes in an interval), and discuss consistency conditions between the wings of the two asymptotic smiles and ways to avoid $|\rho(y)| > 1$ for the calibrated $\rho(y)$ in practice. These results should not be blindly applied for T away from zero in practice (nor indeed should any small-time large deviations result) since ideally one also has to account for the higher order heat kernel asymptotics in [AFLZ17] for the SPX smile and modify the main result in [GHLOW12] for the VIX smile. Rather the two main results are intended to show that having stochastic correlation is essentially the only way to decouple the SPX and VIX smiles as $T \to 0$ for an otherwise conventional two-dimensional Markov continuous stochastic volatility model where the drift and volatility of the volatility process have no S dependence, and to show that an exact theoretical solution to the joint calibration problem exists when $T \to 0$ limit (modulo having the calibrated $|\rho(y)| \leq 1$). For $T > 0$, in principle we can use the martingale optimal transport approaches in [GLOW22] and [Guy22] for the SPX-VIX calibration problem, but it is not a priori clear when a given set of SPX and VIX smiles are arbitrage-free (see below for more on this approach). As a by-product of our analysis, we also report the

*Department of Mathematics, King's College London, Strand, London, WC2R 2LS (Benjamin.Smith@kcl.ac.uk)

surprising behaviour that the point-to-line geodesic for this problem bends back on itself in the x-direction when the log-moneyness $x_1 > 0$ is sufficiently large, or equivalently when the correlation is sufficiently negative (for x_1 greater than some critical value), and we also discuss small-time VIX option pricing under local-stochastic volatility which requires a new VIX transversality condition.

[FGS21] show that the VIX implied volatility smile exhibits power-law skew under the rough Heston model in the small-maturity limit, and (up to a scaling factor) the asymptotic smile is essentially the same as for the spot smile but with the correlation ρ set to 1. [FS21] augment the rough Heston model with an additional CGMY jump process, and show that in principle one can simultaneously use the rough Heston parameters to fit the at-the-money VIX level and skew as $T \to 0$, and the CGMY parameters to fit the observed level, at-the-money correction and at-the-money skew of SPX options as $T \to 0$, and the drift of the V process can be made to be fully consistent with the initial observed variance curve structure; in this sense, infinite-activity jumps allows the SPX and VIX smiles to decouple as $T \to 0$, although in our experience (using an Adams scheme) the CGMY-Heston model does not calibrate well to two SPX smiles with maturity T_1, T_2 and a single VIX smile with maturity T_1 compared to the quadratic rough Heston model in [GR20] (see below for more on the latter), and adding jumps means we do not get a well defined smile as $T \to 0$ in the usual [FZ17] large deviations regime when the log-moneyness scales as $xT^{\frac{1}{2}-H}$. [**?**] consider an extended rough Heston model with jumps in the V process, for which one can still use a VIE for pricing but the VIE now has an additional term which is essentially of the same form as the Lévy-Khintchine formula.

$H = \frac{1}{2}$ models have staged something of a resurgence this decade, in part because some empirical evidence (cf. [GL22],[GL22b], [Rom22b], [AIL22], [AIL22b], [CGS22]) suggests that the SPX at-the-money skew does not always follow a power-law as $T \to 0$, and [Rom22b] suggests that SPX and VIX smiles may be better fit with a mixed (non-rough) Bergomi-type model with two stochastic factors (or extension thereof), one of which has large mean reversion. Markov models can also exhibit "fake roughness" when using non-parametric estimators for H based on realized p-variation or maximum likelihood methods, due to the fact that volatility cannot actually be observed but rather has to be approximated using sums of squares of realized log returns, so there is an additional microstructure noise effect at work where the (re-scaled) difference between realized and actual volatility tends to a sequence of i.i.d. Normals which are independent of everything else as the step size tends to zero (see Theorem 2.1 in [FTW22] and e.g. Cont&Das[CD22] for more on this, and related work on estimating Hurst exponents in [CHLRS22b] using wavelets, and [BFN22],[FTW22] using the Whittle asymptotic approximation for the (Toeplitz) covariance matrix of the (stationary) increments process $Y_j = B^H_{j+1} - B^H_j$ of fractional Brownian motion B^H (see section 5.5 on page 109 in [1] for details)[1] which is maximized to approximate the true maximum likelihood estimator for H. However, in our experience the Whittle approximation to the true covariance matrix of Y does not work well below around $H = 0.25$ (Y_j is the usual canonical process used for these type of problems, since we need the process to be stationary). Also, in our experience, GMM-type estimators of the type in [BCPV22] are of limited use in practice because in reality we do not have access to tens of millions of historical data points for a single stock/index which is typically required for decent ergodic convergence of estimators of e.g. $\mathbb{E}(IV_t IV_{t+\ell})$ (where IV_t is daily integrated variance here), e.g. in [BCPV22] they use 16yrs of data with 1 sec intervals which is 94 million time points.

Abi-Jaber et al.[AIL22] (pages 10-11) and [AIL22b] report strong fits to SPX and VIX options for a simple Markov model where the volatility is a time-dependent function (chosen so as to match a given variance curve term structure) multiplied by an increasing quintic polynomial of a fast mean-reverting Ornstein-Uhlenbeck process (which is close to the Fouque et al.[FPS00] fast mean-reverting regime) and in [AIL22b] they provide an analytic formula for pricing VIX options as a double integral with respect to a Gaussian density over $\mathbb{R} \times [T, T + \Delta]$ which can be approximated with Gaussian quadrature. A quintic function is used to ensure the VIX smile is increasing in strike, although their volatility function can be extremely non-homogenous over small time periods when using a flat variance curve term structure.

Another Markov (and essentially fast mean-reverting) model is proposed on pages 18-19 of Guyon[GL22] (see in the calibrated parameter values in Table 7 there); this model is time-homogenous and the volatility cannot go negative, and the correlation ρ for this model is -1 (so the model is complete) and has an additional state variable R_2 to incorporate additional stylized features (volatility clustering/spikes, positive-sloping VIX smiles Zumbach effect etc.), but this model does not allow for exact sampling of the VIX (unlike the model in [AIL22b] and the quadratic rough Heston model, see e.g. section 6.2 in [Rom22]), so nested Monte Carlo is required to price VIX options which is computationally expensive/error prone. From Eq 4.2 in [GL22], we see this model behaves likes just like the SABR model with $\beta = 1$ and $\rho = -1$ in the small-time fixed-strike limit (since drift terms do not affect small-time Freidlin-Wentzell asymptotics), hence the model does not have three free parameters to fit the at-the-money level, slope and curvature in the small-maturity limit because ρ is hard-wired to -1 (same also applies to the quadratic rough Heston model when the c parameter for the model is zero; for $c > 0$ the volatility has a non-zero lower bound, which is the price we pay for setting $\rho = -1$), but a more general model is considered in section 4.2 of [GL22]. Moreover, due to the massive calibrated vol-of-vol for the models in [AIL22],[AIL22b],[GL22] and large strikes considered, standard Monte Carlo methods (Euler, Cholesky etc) for lower strike options considered (i.e. the left wing of the smile) at small maturities (e.g. 1 month) lead to huge sample variance and bias for MC estimates unless a colossal number of time steps and sample paths are used (e.g. by running on a GPU) and if $\rho = -1$

[1] i.e. fractional Brownian noise, see e.g. [ST02])

or close to -1, even very slightly out-of-the-money calls have close to zero probability of expiring in-the-money which also causes problems for Monte Carlo, and importance sampling with a Girsanov change-of-measure typically doesn't work in practice to resolve this for calibrated parameter values. One could argue their models are in the fast mean-reverting large deviations regime of Fouque at al. [FFK12],[FFF10], for which the rate function for the log stock price in the asymptotic regime is the same as the rate function in [FK16] for the large-time regime, but with the contribution from the drift of the log stock price removed. ρ-values close or equal to -1 also mean we gain little or no benefit from the classic Renault-Touzi[RT96] conditioning trick for Monte Carlo which leads to very high sample variance for Monte Carlo.

To give a concrete example, using a standard antithetic Cholesky (or basic Euler) MC scheme with the usual Renault-Touzi conditioning trick and 1024 time steps, we estimate the standard deviation of the proportional pricing error for the put option with $\log \frac{K}{S_0} = -0.2$ in Figure 1 (for $T = 1$ month) in [AIL22] to be $\approx 6.64\% \times \sqrt{\frac{10^6}{N}}$ where N is the number of paths, and this number is slightly larger for the $T = 1$ month smile on page 19 in [GL22] for the leftermost strike there of $K = 0.75$ (see also code provided by the authors in [AIL22b]). We can obviously multiply this number by e.g. $\Phi^{-1}(.975) \approx 1.96$ to estimate a 95% confidence interval, but this analysis also ignores the bias which we see when we run the code with a fixed seed for different TimeStep values. For these reasons, we strongly advocate using e.g. an explicit or implicit ADI/Douglas finite difference scheme as opposed to Monte Carlo for the [AIL22b] model. The p-value and MLE method discussed in [F23] also show that neither of these models (or any other "in vogue" models - e.g. rough Bergomi-type models, and (discretized) rough and quadratic rough Heston models) are consistent with historical SPX time series (i.e. under the $\mathbb{P}$-measure). The p-value method extracts the underlying Brownian increments (residuals) implied by the data (either SPX time series and/or realized variance using e.g. 1 minute bins for a 6.5 trading day) and tests whether they are in fact i.i.d Normals using e.g. the well known Kolmogorov-Smirnov and Shapiro-Wilks normality tests. In particular the [GL22] model leads to volatility paths which are abnormally large and not consistent with SPX realized variance); rough Bergomi-type models (for which $\log V_t$ is of course Gaussian) do at least fit the data much better than discretized rHeston and qrHeston models, but there are still inconsistencies between H-values obtained from the p-value method and the MLE approach and right tail of $\log V_t$ is not Gaussian in practice. The p-value method applied to the continuous-time Markovian two-factor PDV model in [GL22], leads to V paths which are way too smooth to the naked eye, because the volatility is overly dominated by the unusual drift term.

Severe Monte Carlo problems also arise for rough models (rough Heston, mixed rough Bergomi, quadratic rough Heston etc.) for small H-values (e.g. .05 and below) using the usual hybrid/moment-matching schemes for realistic calibrated parameters, which is easily exposed by comparing the closed-form expressions for the third moment of the driftless log stock price against a Monte Carlo estimate for the third moment (see section 5.2 in [FGS21] for the rough Heston model, and section 4 in [FFGS22] for the rough Bergomi model). In our experience the quadratic rough Heston model provides a much better fit to e.g. a two SPX smiles and a single VIX smile (e.g. with kernel $K(t) = e^{-\lambda t} t^{\alpha - 1}$ and $\theta = 0$ we found that $\alpha = 0.5342$, $\lambda = 4.363$, $a = 0.2006$, $b - Z_0 = 0.09197$, $c = 0.001921$ when calibrated to the $T = 21/365$ and $T = 51/365$ 1st Aug 2018 SPX smiles and the $T = 21/365-$VIX smile given in Guyon[Guy21] using 7.5 million sample paths and 4096 time steps (ran on a GPU), and note that η can be set to 1 and $Z_0 = 0$ W.L.O.G. so the model is very parsimonious, but unfortunately we do not have an expression for the third moment of the driftless log stock price for this model so we can never say for sure how accurate Monte Carlo results are, and the asymptotic short-maturity implied volatility skew can flip sign (see [FS21]).

The recent article of Friz et al.[FSW22] makes some positive theoretical progress in addressing Monte Carlo problems for rough models, showing that the weak error rate for Monte Carlo is $\frac{1}{2}$ for $H < \frac{1}{6}$ and 1 for $H = \frac{1}{2}$, although based on practical experience we suspect the pre-factor in front of this error estimate blows up as $H \to 0$ for the aforementioned well known rough models, and is likely to grow at least linearly in the vol-of-vol parameter when $H = \frac{1}{2}$. One can obtain very good results for the rough Heston model for SPX and VIX options using a basic Adams scheme and Gaussian quadrature for the Fourier inversion (consistent with the exact theoretical value for the third moment of the driftless log stock price), but the rough Heston model can often produce unrealistic downward-sloping or humped-shaped VIX smiles away from $T = 0$ (mixed rough Bergomi models are better at avoiding this issue). One should also never use H-values as low as .01 for the quadratic rough Heston model particularly if one is using also using the finite-dimensional Markov approximation and neural network approximations on top of this as some authors have done with as little as 50,000 time steps (see e.g. footnote 2 on page 5 in [RZ22]), since the answers will typically be wildly inaccurate.

Guo et al. [GLW22],[GLOW22] (see also [HL19] and [Guy22]) show how to construct a generalized local/stochastic volatility model consistent with a finite number of European tradeable options at multiple maturities by minimizing a cost function over calibrated models which penalizes deviations from a standard reference model (e.g. Black-Scholes or Heston), and then re-casting the problem via dualization as an (unconstrained) minmax problem in terms of a non-linear HJB equation (so the cost function effectively regularizes the problem). If options at multiple maturities are used in the calibration set, the HJB equation unfortunately also includes Dirac source terms (but this can be avoided using a nested PDE, see [F23]), and this method is extended to include VIX options in section 3.3 in [GLOW22], by re-expressing V_t for the reference model in terms of $\mathbb{E}(\int_t^T \sigma_s^2 ds | \mathcal{F}_t)$ (this analysis in simplified in [F23] using that VIX_t^2 is just an affine function of V_t when the drift of V under the reference model has a Heston-type drift). This approach is mathematically rich and exciting albeit numerically intensive since it requires

numerically solving a non-linear HJB equation using very fiddly implicit policy-iteration finite difference schemes and then maximizing over the option weights vector. If path-dependent options are included in the calibration set we have the issue that we do not know whether such a consistent model exists to begin with (which is partly what motivated the current paper to address such consistency issues more explicitly as $T \to 0$). We also refer the reader to related results in [Guy21] for a two-period model, and formal results in [Guy22] for the continuous-time setting, using a similar approach but where the penalty function is now the relative entropy of the model from a reference stochastic volatility model. The Guyon approach allows one to compute the inner inf in the minmax duality problem explicitly but this trick only works if $\mathbb{E}(e^{\sum_{i=0}^n w_i(S_T - K_i^+)}) < \infty$ where $K_0 = 0$ and the w_i's are the optimal call option weights for the case when we only have finite tradeable call options, but this expression is infinite if $\sum_{i=0}^n w_i > 0$ for most models of interest (e.g. Black-Scholes, Markov and rough stochastic volatility models), because exponential moments of the stock price do not exist. One can circumvent this by using e.g. a Bachelier-type reference model, or just not using the trick and computing the inner inf by numerically solving the associated HJB equation as discussed above. The [GLOW22] methodology can in principle be generalized to rough reference model using a variational approach, but one ends up with an intractable non-standard FBSDE.

[Lew16] computes the asymptotic smile for the CEV(p)-vol class of stochastic volatility models with non-zero correlation, using a scaling solution to reduce the associated eikonal PDE to a non-linear ODE (see also [FJ11] for similar computations for a general uncorrelated local-stochastic volatily model but using geodesic as opposed to working directly with the eikonal equation), see also [GL14] and [Gul17] for geodesic computations for the Heston model, where the geodesics are shown to be shifts or translations of the standard cycloid $x = s - \sin s$, $y = 1 - \cos s$. Higher-order asymptotic estimates for implied volatility under local/stochastic voatility models are computed in [AFLZ17] (and formally in the earlier works [HL09], [Pau10]) using the heat kernel expansion, and in [BBF04] using viscosity solutions. See also [FG22] and [Fuk22] for another recent and interesting development on asymptotic expansions for SABR and rough Bergomi models, and [JMP21] who provide explicit small-time formulae for the at-the-money implied volatility, skew and curvature for SPX and VIX options for a two-factor rough Bergomi model (this model typically fits a single VIX option smile well but doesn't jointly fit SPX and VIX options well in our experience, see similar findings in [Guy21b] where the two-factor model is refereed to as a "skewed" rough Bergomi model).

1.2 The model

We work on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t \geq 0})$ throughout where the filtration $(\mathcal{F}_t)$ satisfies the usual conditions.

Consider a Markovian stochastic volatility model for a stock price process $S_t = e^{X_t}$, where

$$\left\{ \begin{array}{rcl} dX_t &=& -\frac{1}{2} Y_t dt + \sqrt{Y_t}(\bar{\rho}(Y_t) dW_t^1 + \rho(Y_t) dW_t^2)\,, \\ dY_t &=& \kappa(\theta - Y_t) dt + \alpha(Y_t) dW_t^2\,, \end{array} \right. \tag{1}$$

where W^1, W^2 are two independent standard Brownian motions with respect to $\mathcal{F}_t$, $Y_0 = y_0 > 0$, $\bar{\rho}(y) := \sqrt{1 - \rho(y)^2}$ and $-1 \leq \rho(y) \leq 0$ for all y, and we set $V_t = Y_t$ throughout. We further assume ρ is continuous, and that α differentiable and strictly increasing with

$$\alpha(y) \quad \sim \quad \nu_\infty y^p \tag{2}$$

as $y \to \infty$ with $p \in (0, 1]$ and $\kappa \geq 0$, and we assume κ and α are such that $Y = 0$ is unattainable, see e.g. usual Feller conditions discussed in e.g. [KS91] and [KT81]. We further assume that $X_0 = 0$ without loss of generality since the law of $X_t - X_0$ is independent of X_0.

Remark 1.1 Our localization arguments in the next subsection deal with the issue that $\alpha(y)$ may not be bounded or globally Lipschitz (which is somewhat restrictive in practice), and in particular we will not actually require a full LDP for (X, Y).

Remark 1.2 By a well known Girsanov argument (see e.g. Lemma 2.3 and related results in [AP07] and [LM07]), $\mathbb{E}(S_t) = \mathbb{P}^*(\tau_\infty > t)$ where under $\mathbb{P}^*$, Y has drift $\kappa(\theta - Y_t) + \alpha(Y_t)\sqrt{Y_t}\rho(Y_t)$ and τ_∞ is the explosion time for Y. But from our assumption that $\rho \leq 0$, Y cannot explode under $\mathbb{P}^*$ either. Thus $\mathbb{E}(S_t) = 1$, and (given that the model is Markov in the pair (S, Y)) this implies that S is a martingale.

1.3 Localization arguments

The localization arguments in the proof of the following lemma allow us to deal with the unbounded drift for Y and the unbounded diffusion coefficient for the pair (X, Y).

Lemma 1.1 *Let $\boldsymbol{g} = g_{ij}$ denote the Riemannian metric on $\mathbb{R} \times (0, \infty)$ equal to the inverse $(a^{ij})^{-1}$ of the diffusion coefficient for (X, Y) which has line element $ds^2 = \frac{1}{\bar{\rho}(y)^2}(\frac{1}{y}dx^2 - \frac{2\rho(y)}{\sqrt{y}\alpha(y)}dxdy + \frac{1}{\alpha(y)^2}dy^2)$. Then*

$$\lim_{T \to 0} T \log \mathbb{P}(X_T > x_1) \quad = \quad -I(x_1)\,,$$

where

$$I(x_1) \quad = \quad \frac{1}{2}(\inf_{f \in C_{(0,y_0)}([0,1]): f_1(1) = x_1} \int_0^1 \sqrt{\sum_{i,j=1}^{2} g_{ij} \frac{df^i}{dt}\frac{df^j}{dt}}dt)^2 \quad = \quad \frac{1}{2}d(x_1)^2$$

and $d(x_1)$ is the shortest distance from $(0, y_0)$ to the vertical line $\{x = x_1\}$ under the metric $\boldsymbol{g}$.

The proof of this lemma is deferred to Appendix F.

Remark 1.3 Note that we have not proved (nor do we need) an LDP for the pair of processes (X, Y) here.

1.4 The eikonal equation

From the argument on page 209 in [doC92], we know that the distance-minimizing geodesic γ from (x_0, y_0) to the line $\{x = x_1\}$, satisfies the *transversality* condition

$$\frac{\partial L}{\partial \dot{y}}_{(x_1, y_1^*)} \quad = \quad \mathbf{g}(\frac{d\gamma}{dt}, (0, 1))|_{(x_1, y_1^*)} \quad = \quad 0 \tag{3}$$

i.e. the shortest geodesic comes in perpendicular to the vertical line under the metric g_{ij} (see page 14 in [FJ11] for more details on this point). If $\rho = 0$, the shortest geodesic is also perpendicular in the usual Euclidean sense.

Using (27), we now give a self-contained geometric (as opposed to probabilistic) proof of two well know results which have been proved in [BBF04] using PDE methods with viscosity solutions[2]

Lemma 1.2 *Let $y_1^*(x_1)$ denote the y-value of the shortest geodesic at $x = x_1 > 0$ from $(0, y_0)$ to the vertical line $\{x = x_1\}$ (see discussion above about uniqueness). Then $d(x_1)$ is differentiable and $d'(x_1) = \frac{1}{\sqrt{y_1^*(x_1)}}$ (see also [BBF04]), and the geodesic distance $d(x, y) = d(x, y; x_1)$ from any point (x, y) (with $x < x_1$) to the vertical line with abscissa-value x_1 under the metric g_{ij} satisfies the eikonal PDE:*

$$y d_x^2 \; + \; 2\rho(y)\sqrt{y}\alpha(y)d_x d_y \; + \; \alpha(y)^2 d_y^2 \quad = \quad 1 \tag{4}$$

with $d(0, y) = 0$, for $x \neq x_1$ (see e.g. section 6 in [BBF04]). (4) holds for all x if we replace d with the signed geodesic distance (i.e. flip the sign of d when $x > x_1$, see also section 6 in [BBF04]).

The proof is deferred to Appendix B.

Remark 1.4 $d(-x_1)$ also satisfies (4) with the same boundary condition but with the sign of ρ reversed.

Remark 1.5 $y_1^*(x_1) = \lim_{t \to 0} \mathbb{E}(Y_t | X_t = x_1)$ is the *effective local volatility* (also known as the Markovian projection) as $t \to 0$ (see [BBF04]); we do not require this result in this article. For a general model of the form $dS_t = S_t Y_t^p dW_t, dY_t = \nu Y_t^p dB_t$ with $dW_t dB_t = -1dt$ and $p \in (0, 1)$ (which includes Heston and Hull-White models as special cases) in the small-time limit we essentially have that

$$X_t \quad \approx \quad -\frac{1}{\nu}(Y_t - Y_0)\,,$$

so in particular

$$y_1^*(x_1) \quad = \quad y_0 - \nu x_1 \tag{5}$$

for $x_1 \in (-\infty, \frac{y_0}{\nu}]$, see e.g.[Gath06], and $S_t = e^{X_t}$ is approximately either the exponential of a square root process or a GBM for $p = \frac{1}{2}$ and $p = 1$ respectively, and the $p = 1$ case also applies to the toy Markovian path-dependent volatility model introduced on page 18 in Guyon[GL22], since the diffusion term in Eq 4.3 there just corresponds to a SABR model with $p = 1$ and $\rho = -1$ in the small-time limit.

[2] This can also be proved by making an exponential transformation to the original backward Kolmogorov equation and then considering the small-noise limit to obtain a HJB equation for $\frac{1}{2}d(x_1)^2$, see e.g. Fleming&Soner[FS93]

1.5 Using $\alpha(.)$ and $\rho(.)$ to fit the asymptotic VIX and SPX smiles

If we now assume that

$$\rho(y) \;=\; \rho_0 \,+\, \rho_1(y-y_0) \,+\, O((y-y_0)^2)\,,$$

then by equating coefficients in (4), we can easily compute a power series solution of the form $d(x,y) = \frac{x}{\sqrt{y}} + g_2(y)x^2 + g_3(y)x^3 + O(x^4)$, and we find that the asymptotic implied volatility $\hat{\sigma}(x_1)$ for European options behaves like

$$\hat{\sigma}(x_1) \;=\; \frac{-x_1}{d(-x_1,y_0)} \;=\; \sqrt{y_0} \,+\, \frac{\rho_0\alpha(y_0)}{4y_0}x_1 \,+\, \frac{\alpha(y_0)}{48y_0^{\frac{5}{2}}}((2-7\rho^2+4y_0\rho_0\rho_1)\alpha(y_0)+4y_0\rho_0^2\alpha'(y_0))x_1^2 \,+\, O(x_1^3), \quad (6)$$

where $x_1 = \log\frac{K}{S_0} = \log K$ is the log-moneyness of the call/put option under consideration and K is the strike, where we have used the well known formula from [BBF02] for $\hat{\sigma}(x)$, and

$$y_1^*(x_1)^{\frac{1}{2}} \;=\; \sqrt{y_0} \,+\, \frac{\rho_0\alpha(y_0)}{2y_0}x_1 \,+\, O(x_1^2)\,, \quad (7)$$

and we note that the $O(x_1)$ term is twice the $O(x_1)$ term in (6), which is the familiar "$\frac{1}{2}$ skew" rule of thumb. We can also derive a recurrence relation for higher order terms $g_n(y)$, but the expressions are very fiddly so we omit the details.

If α is known, we can use (6) to choose y_0, ρ_0 and ρ_1 to fit a given/observed overall level, slope and convexity for I at $x_1 = 0$, i.e. fit behaviour of the form $\hat{\sigma}(x_1) = \sigma_0 + \sigma_1 x_1 + \sigma_2 x_1^2$; of course we need the value of the calibrated ρ_0 parameter:

$$\rho_0 \;=\; \frac{4\sigma_1 y_0}{\alpha(y_0)} \;=\; \frac{4\sigma_1\sigma_0^2}{\alpha(\sigma_0^2)} \quad (8)$$

to lie in $[-1,1]$, or else the model is mis-specified, we are also assuming that the limiting implied volatility $\lim_{T\to 0}\hat{\sigma}(x_1,T)$ exists and is finite and non-constant, which is not true for e.g. rough volatility or Lévy models for x_1 fixed in general, see e.g. [FG22]. The formula for ρ_1 is given by

$$\rho_1 \;=\; \frac{56y_0^2\sigma_1^2\alpha(y_0) - \alpha(y_0)^3 - 32y_0^3\sigma_1^2\alpha'(y_0)}{8y_0^2\sigma_1\alpha(y_0)^2} + \frac{3\sqrt{y_0}\sigma_2}{\sigma_1\alpha(y_0)} \quad (9)$$

(recall that $\sigma_0 = \sqrt{y_0}$) and we can go to higher order with this procedure as well, i.e. fit the coefficients of $\rho(y) = \sum_{k=0}^{n-1}\rho_k y^k$ to fit an observed asymptotic smile of the form $\hat{\sigma}(x_1) = \sum_{k=0}^{n}\sigma_k x_1^k$; ρ_k depends on $(\sigma_1,...,\sigma_k,\sigma_{k+1})$, and we find that ρ_k is uniquely determined and finite for all k so long as the skew term $\sigma_1 \neq 0$, and ρ_k is affine in σ_{k+1}.

Remark 1.6 (8) and (9) can also be used to make smart initial guesses for a calibrating a model of the form in (1) to multiple non-zero maturities.

1.6 Calibrating θ to the VIX future price

We first note that

$$\begin{aligned}\mathrm{VIX}_t^2 \;&=\; \frac{1}{\Delta}\mathbb{E}(\int_t^{t+\Delta} Y_u du|\mathcal{F}_t) \;=\; \frac{1}{\Delta}\int_t^{t+\Delta}\mathbb{E}(Y_u|\mathcal{F}_t)du\,, \;=\; \frac{1}{\Delta}\int_t^{t+\Delta}\mathbb{E}(Y_u|Y_t)du\,, \quad (10)\\ \mathbb{E}(Y_t) \;&=\; Y_0 \,+\, \mathbb{E}(\int_0^t \kappa(\theta - Y_u)du \,+\, \int_0^t \alpha(Y_u)dW_u) \;=\; Y_0 \,+\, \int_0^t \kappa(\theta - \mathbb{E}(Y_u))du\,.\end{aligned}$$

Setting $g(t) := \mathbb{E}(Y_t)$ we see that $g'(t) = \kappa(\theta - g(t))$ with initial condition $g(0) = y_0$, which has solution $g(t) = \theta + e^{-\kappa t}(Y_0 - \theta)$. For (10), we need to be able to compute $\mathbb{E}(Y_u|Y_t)$. But since $\mathbb{E}(Y_u|Y_t = y) = \mathbb{E}(Y_{u-t}|Y_0 = y)$, we see that $\mathbb{E}(Y_u|Y_t) = \theta + e^{-\kappa(u-t)}(Y_t - \theta)$ for $u \geq t$, so setting $t = T$ in (10) we see that

$$\mathrm{VIX}_T^2 \;=\; \frac{1}{\Delta}\int_T^{T+\Delta}(\theta + e^{-\kappa(u-T)}(Y_T - \theta))du\,.$$

We can compute the integral here explicitly since Y_T does not depend on u, and we obtain

$$\mathrm{VIX}_T^2 \;=\; F(Y_T)^2 \;=\; aY_T \,+\, b$$

for $\kappa > 0$, where

$$a \;=\; \frac{1-e^{-\kappa\Delta}}{\kappa\Delta} \quad , \quad b \;=\; \frac{\theta}{\kappa\Delta}(e^{-\kappa\Delta} + \kappa\Delta - 1) > 0 \quad (11)$$

and $F(y) = \sqrt{ay+b}$, so in particular

$$\mathrm{VIX}_0^2 \quad = \quad aY_0 \, + \, b \, . \tag{12}$$

If $\kappa = 0$ then $\mathrm{VIX}_T^2 = Y_T$ since Y is a martingale in this case, i.e. $a = 1$ and $b = 0$.

For $\kappa > 0$, since $V_T \geq b > 0$, we see that

$$\mathrm{VIX}_T \quad = \quad (\frac{1}{T}\int_T^{T+\Delta} \mathbb{E}_T(V_u)du)^{\frac{1}{2}} \quad \geq \quad \sqrt{b} \, ,$$

which implies that the VIX implied volatility is zero for strikes $k \leq \sqrt{b}$ at all strikes, since in this case $\mathbb{E}((\mathrm{VIX}_T - k)^+) = \mathbb{E}(\mathrm{VIX}_T) - k$, i.e. there is no time-value to the option.

Lemma 1.3 $\mathbb{E}(\mathrm{VIX}_T - \mathrm{VIX}_0) = \mathbb{E}(\mathrm{VIX}_T - \sqrt{aY_0 + b}) = o(\sqrt{T})$ *as* $T \to 0$.

The proof is deferred to Appendix C

Remark 1.7 Formally at least, we can sharpen this statement to the following:

$$\begin{aligned} \mathbb{E}(\mathrm{VIX}_T) \quad = \quad (P_T F)(y) \quad &= \quad (I \, + \, T\mathcal{A} \, + \, \frac{1}{2!}T^2\mathcal{A}^2 + ...)F(y)|_{y=y_0} \\ &= \quad \sqrt{ay_0 + b} \, + \, T(\frac{a(\theta - y)\kappa}{2\sqrt{ay_0+b}} - \frac{a^2\alpha(y_0)^2}{8(ay_0+b)^{\frac{3}{2}}}) \, + \, O(T^2) \, , \end{aligned} \tag{13}$$

where $\mathcal{A} = \kappa(\theta - y)\frac{\partial}{\partial y} + \frac{1}{2}\alpha(y)^2\frac{\partial^2}{\partial y^2}$ is the infinitesimal generator of Y and P_t is the associated semigroup.

Using this lemma, and noting that b is a linear function of θ, we can then calibrate θ to the *observed* asymptotic VIX future price as $T \to 0$ by equating said price with $\sqrt{ay_0 + b} = \sqrt{ay_0 + b(\theta)}$ (for κ given), and solving for θ.

1.7 Calibrating α to the short-maturity VIX smile

With κ chosen exogenously and θ calibrated as above, we will now show how α can be calibrated exactly to the observed small-time asymptotic behaviour of VIX options. We first note that for $k > \sqrt{b}$

$$\frac{1}{2}d_{\mathrm{VIX}}(k)^2 \quad := \quad -\lim_{T\to 0} T \log \mathbb{P}(\mathrm{VIX}_T > k) \quad = \quad -\lim_{T\to 0} T \log \mathbb{P}(Y_T > \frac{k^2 - b}{a}) \quad = \quad \frac{1}{2}d_Y(\frac{k^2-b}{a})^2 \tag{14}$$

where a and b are defined as in (11) (note that k is not log-moneyness here) where

$$d_Y(y) \quad := \quad \int_{y_0}^{y} \frac{du}{\alpha(u)} \, ,$$

so $d_Y'(y) = \frac{1}{\alpha(y)}$, and $d_Y(k) = d_{\mathrm{VIX}}(\sqrt{ak+b})$.

Corollary 1.4 *For* $\mathrm{VIX}_0 e^x > \sqrt{b}$*, we see that*

$$\lim_{T\to 0} T \log \mathbb{E}((\mathrm{VIX}_T - \mathrm{VIX}_0 e^x)^+) \quad = \quad -\frac{1}{2}d_{\mathrm{VIX}}(\mathrm{VIX}_0 e^x)^2 \, .$$

Proof. See Appendix D. ■

Corollary 1.5 *If* $\hat{\sigma}_{\mathrm{VIX}}(k,T)$ *denotes the implied volatility of a VIX call option with strike* $k > \sqrt{b}$*, then* $\hat{\sigma}_{\mathrm{VIX}}(x)$ *is given by the following [BBF02]-type formula:*

$$\hat{\sigma}_{\mathrm{VIX}}(x) \quad := \quad \lim_{T\to 0} \hat{\sigma}_{\mathrm{VIX}}(\mathrm{VIX}_0 e^x, T) \quad = \quad \frac{x}{d_{\mathrm{VIX}}(\mathrm{VIX}_0 e^x)} \tag{15}$$

for $x \in \mathbb{R}$*, and recall that* (12) *has to be satisfied.*

The proof is deferred to Appendix E.

From (15) and (14), we can then back out $\alpha(.)$ from the observed $\hat{\sigma}_{\mathrm{VIX}}(x)$.

In practice, to avoid $|\rho(y)| > 1$, we can consider a $\rho(y)$ of the form

$$\rho(y) \quad = \quad c_1 \tanh(a_1(y - y_0) + b_1) \, + \, \eta_1 \tag{16}$$

and match the parameters to the calibrated values for $\rho_0 = \rho(y_0)$ and $\rho_1 = \rho'(y_0)$, and exogenously choose $\rho(0)$ and $\lim_{y\to\infty} \rho(y)$ (see Figure 2 for a numerical example). The values of y_0, ρ_0 and ρ_1 here will themselves be obtained by matching the observed level, slope and convexity of the asymptotic implied volatility using the expansion in (6) (see Figure 2 below for a numerical calibration example).

1.8 Examples

1.8.1 Hull-White/SABR-type model

For a Hull-White type model with $\alpha(y) = \nu y$, $d_Y(y) = \frac{1}{\nu}\log\frac{y}{y_0}$, we find that

$$\hat{\sigma}_{\text{VIX}}(x) = \begin{cases} \frac{x\nu}{\log[\frac{-\theta+e^{2x}(\theta-y_0)+e^{\Delta\kappa}\theta(1-\Delta\kappa)+e^{2x+\Delta\kappa}(y_0+\theta(\Delta\kappa-1))}{(e^{\Delta\kappa}-1)y_0)}]} & (\kappa > 0) \\ \frac{x\nu}{\log\frac{e^{2x+\Delta\kappa}\Delta\kappa+e^{\Delta\kappa}(1-\Delta\kappa)-1}{e^{\Delta\kappa}-1}} \;=\; \frac{\nu-e^{-\Delta\kappa}\nu}{2\Delta\kappa} - \frac{(1-e^{-\Delta\kappa}-\Delta\kappa)\ \nu)x}{2\Delta\kappa} + O(x^2) & (\kappa > 0,\, \theta = y_0) \\ \frac{1}{2}\nu \quad (\kappa = 0)\,. \end{cases}$$

The $O(x)$ term for the $\kappa > 0, \theta = y_0$ case is positive for κ sufficiently large, and $\hat{\sigma}_{\text{VIX}}(x) \nearrow \frac{1}{2}\nu$ as $x \to \infty$, and $\hat{\sigma}_{\text{VIX}}(x)$ is concave in x, so this behaviour is more realistic than the Heston case discussed below (see plot of $\hat{\sigma}_{\text{VIX}}(x)$ in Figure 8).

1.8.2 The Heston model

For the standard Heston model $\alpha(y) = \nu\sqrt{y}$, $d_Y(y) = \frac{2}{\nu}(y - \sqrt{y_0})$ and setting $\theta = y_0$ for simplicity, the rate function for VIX_T is Gaussian, and from (15) we find that

$$\hat{\sigma}_{\text{VIX}}(x) = \begin{cases} \frac{x\nu}{2(\sqrt{\frac{y_0(e^{2x+\Delta\kappa}\Delta\kappa+e^{\Delta\kappa}(1-\Delta\kappa)-1)}{e^{\Delta\kappa}-1}}-\sqrt{y_0})} \;=\; \frac{1-e^{-\Delta\kappa}\nu}{2\sqrt{y_0}\Delta\kappa} + \frac{e^{-\Delta\kappa}(2+e^{-\Delta\kappa}(-2+\Delta\kappa))\ \nu}{4\sqrt{y_0}\Delta\kappa}x + O(x^2) & (\kappa > 0) \\ \frac{\frac{1}{2}\nu x}{\sqrt{y_0}\,(e^x-1)} \;=\; \frac{\nu}{\sqrt{y_0}}(\frac{1}{2} - \frac{1}{4}x + \frac{1}{24}x^2 + O(x^3)) & (\kappa = 0)\,. \end{cases} \tag{17}$$

For $\kappa = 0$, $\hat{\sigma}_{\text{VIX}}(x)$ here is convex and monotonically decreasing, i.e. a pronounced negative skew and no smile much like the spot implied volatility for a Bachelier model (see Figure 6), and note that this is not the kind of behaviour observed in practice where VIX smile are increasing in the strike. One can circumvent this issue by instead assuming that $\alpha(y) \sim \nu y^p$ for $p > 1$ as $y \to y_0$ (like the so-called $\frac{3}{2}$ model), but also impose that $\alpha(y) \sim \nu y^q$ for some $q \in [0, 1]$ as $y \to \infty$ to ensure the martingale property for Y is preserved and hence that $\text{VIX}_T^2 = Y_T$ when $\kappa = 0$ (see Figure 2 for a numerical example like this).

For $\kappa > 0$, $\hat{\sigma}_{\text{VIX}}(x)$ is monotonically increasing for $x < x^*$ for some x^* and decreasing for $x > x^*$, and in particular the at-the-money VIX skew is positive for κ sufficiently large.

1.8.3 Lewis CEV-p-type model

If $\alpha(y) = \nu y^p$ for $p \in (0, 1)$ and $\kappa = 0$, then

$$\hat{\sigma}_{\text{VIX}}(x) \;=\; \frac{(1-p)x\nu}{(e^x\sqrt{y_0})^{2p-2} - y_0^{1-p}} \qquad (p \in (0,1)) \tag{18}$$

and this formula holds asymptotically as $x \to \infty$ if $\alpha(y) \sim \nu y^p$ as $y \to \infty$.

Remark 1.8 We can also compute $\hat{\sigma}_{\text{VIX}}(x)$ explicitly for a mixed model where $\alpha(y) = \nu y^p + \xi y^q$ for $p, q > 0$ in terms of the confluent hypergeometric function (we omit details for the sake of brevity), which is useful for fitting VIX smiles in practice for $T \ll 1$.

1.9 The Bass martingale - an exact fit to a single VIX smile

Let $Y_t = V_t = \mathbb{E}(g(W_T)|W_t) = f(W_t, t)$, where W is a standard Brownian, and g is chosen so V_T has a given law, and since V is a martingale, we know that $V_T = \text{VIX}_T^2$ and

$$dY_t \;=\; f_x(W_t, t)dW_t \;=\; \alpha(Y_t, t)dW_t$$

where $\alpha(y, t) = f_x(f^{-1}(v, t), t)$, i.e. V is a time-inhomogenous diffusion. Thus if the prescribed law for $\sqrt{V_T}$ agrees with the law for VIX_T implied by VIX option prices at maturity T, then this model is consistent with that single-maturity VIX smile. One is then "locked in" to the spot price dynamics $dS_t = S_t\sqrt{V_t}dB_t$, where $dB_t dW_t = \rho dt$, so the only additional freedom one has to fit Europeans is to make the correlation stochastic. V here is the well known Bass martingale (see e.g. [BBHK20]) and it should be possible to extend this to multiple maturities, as [CH21] have done for standard European options).

2 An integral equation for $\rho(y)$ for an exact fit to the SPX/VIX smiles

From here on, we make the following assumption:

Assumption 2.1
(i) $-1 < \rho(y) \leq 0$
(ii) $\rho(y_0) < 0$
(iii) $\bar{\rho}(y_0)^2 y_0 > \bar{\rho}(y)^2 y$ *for all* $y \in [\bar{\rho}(y_0)^2 y_0, y_0)$, *and* $\bar{\rho}(y)^2 y$ *is non-decreasing in* y *for* $y \in [y_0, y_{\max}]$ *for some* $y_{\max} > y_0$
(iv) $\rho(y) \to \rho_\infty \in [-1, 0)$ *as* $y \to \infty$.

Remark 2.1 A necessary condition for the first part of Assumption iii) is that $\frac{d}{dy}(\bar{\rho}(y_0)^2 y_0 - \bar{\rho}(y)^2 y)|_{y=y_0} = \frac{d}{dy}(-\bar{\rho}(y)^2 y)|_{y=y_0} \leq 0$, which is equivalent to

$$\rho_0^2 + 2 y_0 \rho_0 \rho_1 \quad \leq \quad 1\,, \tag{19}$$

or (if $\rho_0 < 0$) $\rho_1 \geq -\frac{1-\rho_0^2}{2 y_0 |\rho_0|}$, which in turn puts an *upper bound* on the at-the-money convexity σ_2 in (9), assuming that $\sigma_1 < 0$ as well. See Remark 2.3 below for a discussion on the extent to which we can relax this condition.

For convenience, we now define

$$x_1^* \quad = \quad \int_{\bar{\rho}(y_0)^2 y_0}^{y_0} \left(\frac{\bar{\rho}(y)^2 y}{\alpha(y)\sqrt{\bar{\rho}(y_0)^2 y_0 - \bar{\rho}(y)^2 y}} - \frac{\rho(y)\sqrt{y}}{\alpha(y)}\right) dy\,.$$

Proposition 2.2 *(From Assumption 2.1ii) and iii)) $x_1^* > 0$, and for $x_1 \leq x_1^*$, x_1 and $y_1^*(x_1)$ are related via*

$$x_1 \quad = \quad \int_{y_1^*(x_1)}^{y_0} \left(\frac{\bar{\rho}(y)^2 y}{\alpha(y)\sqrt{y_1^*(x_1) - \bar{\rho}(y)^2 y}} - \frac{\rho(y)\sqrt{y}}{\alpha(y)}\right) dy\,. \tag{20}$$

For $x > x_1$, we define $y_c(.)$ via $y_1^*(x_1) = y_c(x_1)\bar{\rho}(y_c(x_1))^2$ (note y_c is uniquely defined by the second part of Assumption 2.1 iii)); then x_1 and $y_1^*(x_1)$ are related via

$$x_1 \quad = \quad \left(\int_{y_0}^{y_c(x_1)} + \int_{y_1^*(x_1)}^{y_c(x_1)}\right) \frac{\bar{\rho}(y)^2 y}{\alpha(y)\sqrt{y_1^*(x_1) - \bar{\rho}(y)^2 y}} dy + \int_{y_0}^{y_1^*(x_1)} \frac{\rho(y)\sqrt{y}}{\alpha(y)} dy \qquad (x_1 > x_1^*) \tag{21}$$

if $y_c(x_1) \leq y_{\max}$. $y_1^*(x_1)$ is continuous in x_1, and $y_1^*(x_1)$ is minimized at $x_1 = x_1^*$ where $y_1^*(x_1^*) := \bar{\rho}(y_0)^2 y_0$. The Riemannian distance from $(0, y_0)$ to $(x_1, y_1^*(x_1))$ for all $x_1 \in \mathbb{R}$ under the metric g_{ij} defined above is then given by

$$d(x_1) \quad = \quad \begin{cases} \sqrt{y_1^*(x_1)}\,\left|\int_{y_0}^{y_1^*} \frac{dy}{\alpha(y)\sqrt{y_1^*(x_1) - \bar{\rho}(y)^2 y}}\right| & (x_1 \leq x_1^*) \\ \sqrt{y_1^*(x_1)}\left(\int_{y_0}^{y_c(x_1)} + \int_{y_c(x_1)}^{\bar{\rho}(y_c(x_1))^2 y_c(x_1)}\right) \frac{dy}{\alpha(y)\sqrt{\bar{\rho}(y_c(x_1))^2 y_c(x_1) - \bar{\rho}(y)^2 y}} & (x_1 > x_1^*) \end{cases} \tag{22}$$

(assuming $y_c(x_1) \leq y_{\max}$ for the second formula), and if

$$y_1^* \quad \geq \quad \frac{y_0 \bar{\rho}(y_0)^2}{\rho(y_0)^2}$$

and $x_1 > 0$ and $y_c(x_1) \leq y_{\max}$, the shortest geodesic from $(0, y_0)$ to $\{x = x_1\}$ bends back on itself in the x-direction.

Remark 2.2 (21) and (22) generalize the two main two equations Eq 26 and 27 in [FJ11] for the case when $\rho = 0$ (see also section 12.10 in [Lew16] for similar computations for the specific case when $\alpha(y) = \xi y^p$ by solving the eikonal PDE rather than working with the underlying geodesics as we are here). Since $\bar{\rho}(y_0)^2 y_0 > 0$ and $y_c^*(x_1) < \infty$, we do not have to worry about these shortest geodesics hitting $y = 0$ or ∞.

The proof of the Proposition is given below Corollary 2.3.

Remark 2.3 The first part of Assumption iii) implies that $y_1^* - \bar{\rho}(y)^2 y > 0$ for all $y_1^* \in [\bar{\rho}(y_0)^2 y_0, y_0)$ and all $y \in (\bar{\rho}(y_0)^2 y_0, y_0)$, for which we have seen that (19) is a necessary condition, but this condition can be restrictive in practice. If $\bar{\rho}(y_0)^2 y_0 - \bar{\rho}(y)^2 y$ has a root in $[\bar{\rho}(y_0)^2 y_0, y_0)$ i.e. the first part of Assumption 2.1iii) is violated, let $y_{\min}$ denote the smallest y_1-value such that $y_1 - \bar{\rho}(y)^2 y \geq 0$ for all $y \in [\bar{\rho}(y_0)^2 y_0, y_0)$. If $y_{\min} - \bar{\rho}(y)^2 y$ has a minimum at $\hat{y}$ over $[\bar{\rho}(y_0)^2 y_0, y_0]$ which is a stationary point (i.e. $\frac{d}{dy}(\bar{\rho}(y)^2 y)|_{y=\hat{y}} = 0$) then $y_{\min} - \bar{\rho}(y)^2 y$ just touches zero at $\hat{y}$, but does not go below zero on this range. Then if $y_{\min} \in [\bar{\rho}(y_0)^2 y_0, y_0)$, this is the lowest attainable value for y_1^*, and from the differentiability of $\bar{\rho}(y)$, we know that

$$y_{\min} - \bar{\rho}(y)^2 y \quad = \quad const. \times (y - \hat{y})^2 + O((y - \hat{y})^3)\,, \tag{23}$$

so the pole in (34) at $y = \hat{y}$ is not integrable, so

$$\lim_{y_1^* \searrow y_{\min}} x(y_1^*) \quad = \quad +\infty \tag{24}$$

and hence there is a shortest geodesic from $(0, y_0)$ to $\{x = x_1\}$ for any $x_1 > 0$ all of which have $y'(0) < 0$ and no turning point along the geodesic, and in this case $y_1^*(x_1) = \frac{1}{d'(x_1)^2} \searrow y_{\min}$ as $x_1 \to \infty$. so the asymptotic implied volatility smile $\hat{\sigma}(x_1) = \frac{x_1}{d(x_1)}$ tends to a finite constant as $x_1 \to +\infty$ since $d(x_1) \sim x_1 d'(\infty)$ as $x_1 \to \infty$, and the Assumption (19) does not have to hold in this case.

Remark 2.4 If we let $\rho(y) \to -1$ in Proposition 2.2, then $\bar{\rho}(y) \to 0$ and $x_1^* \to x_1^{*,\rho \to -1} := \int_0^{y_0} \frac{\sqrt{y}}{\alpha(y)} dy$ and in the limit $\rho \to -1$, $y_1^*(x_1)$ satisfies

$$x_1 \quad = \quad \int_{y_1^*(x_1)}^{y_0} \frac{\sqrt{y}}{\alpha(y)} dy$$

and $d(x_1) \to \int_{y_0}^{y_1^*} \frac{dy}{\alpha(y)}$ for $x_1 \le x_1^*$, and $d(x_1) = \infty$ for $x_1 \ge x_1^*$, since such x_1 values are essentially unattainable. The geodesic equations in this case are

$$x'(t) \quad = \quad -\frac{\sqrt{y}}{\alpha(y)} y'(t) \quad , \quad y'(t) \quad = \quad \pm \alpha(y) \sqrt{y_1^*} \, ,$$

so $\frac{dy}{dx} = -\frac{\sqrt{y}}{\alpha(y)}$, which is a straight line with slope ν for the Heston case (see also the second figure in the top row of Figure 4). Note we have let $\rho(y) \to -1$ here, not just set $\rho(y) = -1$. We cannot directly do the latter using our g_{ij} metric since the line element ds^2 becomes singular in this case, but we can circumvent this issue by performing a simple one-dimensional Freidlin-Wentzell/geodesic analysis, i.e. we know that $X_t = -\int_0^t \sqrt{Y_s} dW_s^2 = -F(Y_t) +$ (drift term), for some F with $F'(y)\alpha(y) = \sqrt{y}$ and Y_t satisfies the LDP as $t \to 0$ with rate $J(y) = \int_{y_0}^{y} \frac{du}{\alpha(u)} du$, so X_t satisfies the LDP with rate $J(F^{-1}(y))$, and note F is linear in the Heston case.

Corollary 2.3 *From* (2)*, if* $y_{\max} = \infty$*, then* $\hat{\sigma}(x)$ *has tail behaviour*

$$\hat{\sigma}(x_1)^2 \quad \sim \quad \tilde{c}^{\pm}_{\nu_\infty, \rho_\infty, p} x_1^{\frac{2}{3-2p}} \tag{25}$$

as $x_1 \to \pm\infty$ *for some constants* $\tilde{c}^{\pm}_{\nu_\infty, \rho_\infty, p}$ *which are given in the proof. Note the condition* $y_{\max} = \infty$ *is only required for* $x_1 \to +\infty$ *limit.*

The proof is given below the proof of Proposition 2.2.

Remark 2.5 (25) essentially gives a **consistency condition** for the wings of the SPX and VIX asymptotic smiles to be consistent: we can either obtain p and ν_∞ from the tail behaviour of the observed $\sigma_{\text{VIX}}(x)$ function, so we are only free to choose ρ_∞, or we can exogenously choose p, ν_∞ and ρ_∞ to fit desired tail behaviour for $\hat{\sigma}(x_1)$ as $x_1 \to \pm\infty$ as in (25), which then imposes (2) and fixes the behaviour of $\hat{\sigma}_{\text{VIX}}(x)$ for $|x| \to \infty$ via (18) (this is appropriate if e.g. we only have an observed $\sigma_{\text{VIX}}(x)$ function over a finite interval so we need to extrapolate $\sigma_{\text{VIX}}(x)$ to $x = \pm\infty$.

We now have the following corollary for the calibration problem of solving for $\rho(.)$ for a given/observed $d(x_1) = \frac{x_1}{\hat{\sigma}(x_1)}$ function:

Remark 2.6 We can re-write (22) as

$$f(y_1) \quad := \quad \frac{d((y_1^*)^{-1}(y_1)}{\sqrt{y_1}} \quad = \quad \int_{y_0}^{y_1} \frac{dy}{\alpha(y)\sqrt{y_1 - y\bar{\rho}(y)^2)}} \tag{26}$$

where $(y_1^*)^{-1}(y_1)$ is the smallest non-negative root of $y_1^*(x_1) = y_1$, and for $x > x_1^*$ as

$$f(y_1) \quad = \quad \Big(\int_{y_0}^{y_c((y_1^*)^{-1}(y_1))} + \int_{y_0}^{y_1}\Big) \frac{dy}{\alpha(y)\sqrt{y_1 - y\bar{\rho}(y)^2)}} \, , \tag{27}$$

where in the second equation, the function $y_c(x_1)$ is defined implicitly in (38) in terms of $y_1^*(x_1)$ and $(y_1^*)^{-1}(y_1)$ is the largest non-negative root of $y_1^*(x_1) = y_1$. If $d(x_1) = \frac{x}{\hat{\sigma}(x)}$ is given or observed for all x_1 **in some interval** $[a, b]$ (and hence also $y_1^*(x_1)$, using that $d'(x_1) = \frac{1}{\sqrt{y_1^*(x_1)}}$), then (26) and (27) are non-standard **Abel-type integral equations** for the unknown function $\rho(y)$ on $[y_1^*(a), y_1^*(b)]$, assuming $\alpha(y)$ has already been calibrated to the asymptotic VIX smile $\hat{\sigma}_{\text{VIX}}$ or is given. Even if we can solve the integral equation, it does not give us $\rho(y)$ for $y \in (0, \bar{\rho}(y_0)^2 y_0 = \min_{x_2 \in \mathbb{R}} y_1^*(x_1))$, since the behaviour of α and ρ in this range does not affect $y_1^*(x_1)$ and $d(x_1)$.

Remark 2.7 The calibrated $\rho(y)$ must of course satisfy $|\rho(y)| \leq 1$ (if not then the model is mis-specified, and may suggest that the diffusion coefficient for the model should have x-dependence as well, e.g. let ρ also depend on x).

Proof. (of Proposition 2.2). We break the proof into multiple parts:

1. Deriving the geodesic equations. The Lagrangian is conserved along geodesics, i.e.

$$L \;=\; \frac{1}{2\bar{\rho}(y)^2}(\frac{1}{y}(\frac{dx}{dt})^2 \,-\, \frac{2\rho(y)}{\sqrt{y}\alpha(y)}\frac{dx}{dt}\frac{dy}{dt} \,+\, \frac{1}{\alpha(y)^2}(\frac{dy}{dt})^2) \;=\; E$$

for some constant E (see Appendix A for proof), and the x-component of the Euler-Lagrange equation is

$$\frac{d}{dt}(\frac{1}{2\bar{\rho}(y)^2}(\frac{2}{y}(\frac{dx}{dt}) \,-\, \frac{2\rho(y)}{\sqrt{y}\alpha(y)}\frac{dy}{dt})) \;=\; 0 \quad\Rightarrow\quad \frac{1}{y}(\frac{dx}{dt}) \,-\, \frac{\rho(y)}{\sqrt{y}\alpha(y)}\frac{dy}{dt} \;=\; \bar{\rho}(y)^2 K_1 \tag{28}$$

for some constant K_1 (where we have used our assumption that $\rho(y) \in (-1, 0]$), i.e. the (momentum) quantity on the left is a conserved quantity, so

$$\frac{dy}{dt} \;=\; \pm\alpha(y)\sqrt{2E - K_1^2 y + K_1^2 y\rho(y)^2} \;=\; \pm\alpha(y)\sqrt{2E - K_1^2 y\bar{\rho}(y)^2}\,. \tag{29}$$

2. Transversality condition. From (F-1) we know we have to compute the shortest distance to the vertical line $\{x = x_1\}$ for $x_1 \in \mathbb{R}$, and we first assume $x_1 > 0$. The transversality condition (after multiplying by $\bar{\rho}(y)^2$) is given by

$$0 \;=\; [0,1]\begin{bmatrix} \frac{1}{y} & -\frac{\rho(y)}{\alpha(y)\sqrt{y}} \\ -\frac{\rho(y)}{\alpha(y)\sqrt{y}} & \frac{1}{\alpha(y)^2} \end{bmatrix}\begin{bmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{bmatrix}|_{(x_1^*, y_1^*(x_1))} \;=\; (-\frac{\rho(y)}{\alpha(y)\sqrt{y}}\frac{dx}{dt} \,+\, \frac{1}{\alpha(y)^2}\frac{dy}{dt})|_{(x_1^*, y_1^*(x_1))} \;=\; 0\,. \tag{30}$$

(28) yields an expression for $\frac{dx}{dt}$, so the right hand side of (30) can be re-written as

$$\begin{aligned} -\frac{\rho(y)}{\alpha(y)\sqrt{y}}\, y[K_1\bar{\rho}(y)^2 + \frac{\rho(y)}{\sqrt{y}\alpha(y)}\frac{dy}{dt})] \,+\, \frac{1}{\alpha(y)^2}\frac{dy}{dt} \;&=\; -\frac{\rho(y)\sqrt{y}}{\alpha(y)}K_1\bar{\rho}(y)^2 - \frac{\rho(y)^2}{\alpha(y)^2}\frac{dy}{dt} \,+\, \frac{1}{\alpha(y)^2}\frac{dy}{dt} \\ &=\; \bar{\rho}(y)^2(-\frac{\rho(y)\sqrt{y}}{\alpha(y)}K_1 + \frac{1}{\alpha(y)^2}\frac{dy}{dt}) \;=\; 0 \end{aligned}$$

at $(x_1, y_1^*(x_1))$, so

$$y'(t) \;=\; \rho(y)\sqrt{y}\alpha(y)K_1$$

at $(x_1, y_1^*(x_1))$. But we also know that

$$y'(t) \;=\; \pm\alpha(y)\sqrt{2E - K_1^2 y + K_1^2 y\rho(y)^2}\,.$$

Combining these last two expressions, we find that

$$2E \;=\; \rho(y)^2 y_1^* K_1^2 \,-\, (-K_1^2 y_1^* + K_1^2 y_1^*\rho(y_1^*)^2) \;=\; K_1^2 y_1^*\,,$$

so

$$y_1^* \;=\; 2E/K_1^2 \tag{31}$$

(as for the zero correlation case discussed in [FJ11]), and

$$y'(t) \;=\; \pm\alpha(y)K_1\sqrt{y_1^* - y\bar{\rho}(y)^2}\,. \tag{32}$$

From the transversality condition in (30), since $\rho \leq 0$ by assumption, we see that

$$\frac{dy}{dx}|_{(x_1^*, y_1^*)} \;<\; 0\,, \tag{33}$$

irrespective of the sign of x_1. We now have to distinguish between two cases: shortest geodesics from $(0, y_0)$ to $\{x = x_1\}$ for which $\frac{dy}{dt} < 0$ at $t = 0$, or $\frac{dy}{dt} > 0$ at $t = 0$, plus the critical case where $\frac{dy}{dt} = 0$ at $t = 0$.

Recall also from (28) that $\frac{1}{y}(\frac{dx}{dt}) - \frac{\rho(y)}{\sqrt{y}\alpha(y)}\frac{dy}{dt} = \frac{1}{y}(\frac{dx}{dt}) - \pm\frac{\rho(y)}{\sqrt{y}}K_1\sqrt{y_1^* - y\bar{\rho}(y)^2} = \bar{\rho}(y)^2 K_1$, so (combined with (32)) we see that $\frac{dy}{dt}$ and $\frac{dx}{dt}$ are both proportional to K_1, so (with y_1^* fixed) we can set $K_1 = 1$ W.L.O.G since the specific choice of time parametrization is irrelevant and all we ultimately care about is $y(x)$ (note t will be negative when $x_1 < 0$ using this convention, see final comment in bold face about the case $x_1 < 0$ just before the case 3ii)

below). Moreover, the ODE for $y'(t)$ in (32) implies that if $y(t)$ tends to some constant $y_\infty > 0$ as $t \to \infty$ then $y'(t)$ converges to some negative constant as $t \to \infty$. If the derivative of a function converges to a non-zero constant as $t \to \infty$ then the function blows up i.e. tends to $\pm\infty$, so we have a contradiction, i.e. we must have that $y_\infty = 0$.

3i) Integral expression linking x_1 and $y_1^*(x_1)$ - the case $x_1 \le x_1^*$ where the geodesic has no turning point.

Let t_1 be such that $x(t_1) = x_1$ (and note that t_1 depends on the choice of K_1, see discussion above). We first investigate whether there are shortest geodesics from $(0, y_0)$ to $\{x = x_1\}$ for which $\frac{dy}{dt} \le 0$ along this geodesic until it hits $x = x_1$. For this to be the case, for $x_1 > 0$ we must have

$$\begin{aligned} x_1 &= \int_0^{t_1} \frac{dx}{dt} dt = \int_{y_0}^{y_1^*} y(K_1\bar\rho(y)^2 + \frac{\rho(y)}{\sqrt{y}\alpha(y)}\frac{dy}{dt})\frac{dt}{dy}dy \quad \text{(from (28))} \\ &= \int_{y_0}^{y_1^*} -\frac{y\bar\rho(y)^2}{\alpha(y)\sqrt{2E - y + y\rho(y)^2}}dy + \int_{y_0}^{y_1^*} \frac{\rho(y)\sqrt{y}}{\alpha(y)}dy \\ &\quad \text{(setting } K_1 = 1 \text{ here W.L.O.G, see above)} \\ &= \int_{y_0}^{y_1^*} -\frac{\bar\rho(y)^2 y}{\alpha(y)\sqrt{y_1^* - \bar\rho(y)^2 y}}dy + \int_{y_0}^{y_1^*} \frac{\rho(y)\sqrt{y}}{\alpha(y)}dy \\ &= \int_{y_1^*}^{y_0} (\frac{\bar\rho(y)^2 y}{\alpha(y)\sqrt{y_1^* - \bar\rho(y)^2 y}} - \frac{\rho(y)\sqrt{y}}{\alpha(y)})dy \end{aligned} \tag{34}$$

and $y_1^* = y_1^*(x_1) \le y_0$.

Since $\rho(y_0) \neq 0$ by assumption, we also note that

$$\begin{aligned} \frac{dx_1}{dy_1^*} &= -(\frac{\bar\rho(y_1^*)^2 y_1^*}{\alpha(y_1^*)\sqrt{y_1^* - \bar\rho(y_1^*)^2 y_1^*}} - \frac{\rho(y_1^*)\sqrt{y_1^*}}{\alpha(y_1^*)}) - \frac{1}{2}\int_{y_1^*}^{y_0} (\frac{\bar\rho(y)^2 y}{\alpha(y)(y_1^* - \bar\rho(y)^2 y)^{\frac{3}{2}}}dy \\ &= -\frac{\sqrt{y_1^*}}{\alpha(y_1^*)}(\frac{\bar\rho(y_1^*)^2}{|\rho(y_1^*)|} - \rho(y_1^*)) - \frac{1}{2}\int_{y_1^*}^{y_0} \frac{\bar\rho(y)^2 y}{\alpha(y)(y_1^* - \bar\rho(y)^2 y)^{\frac{3}{2}}}dy \\ &= -\frac{\sqrt{y_1^*}}{\alpha(y_1^*)\rho(y_1^*)}(-\bar\rho(y_1^*)^2 - \rho(y_1^*)^2) - \frac{1}{2}\int_{y_1^*}^{y_0} \frac{\bar\rho(y)^2 y}{\alpha(y)(y_1^* - \bar\rho(y)^2 y)^{\frac{3}{2}}}dy \\ &= \frac{\sqrt{y_1^*}}{\alpha(y_1^*)\rho(y_1^*)} - \frac{1}{2}\int_{y_1^*}^{y_0} \frac{\bar\rho(y)^2 y}{\alpha(y)(y_1^* - \bar\rho(y)^2 y)^{\frac{3}{2}}}dy\,, \end{aligned} \tag{35}$$

where the penultimate line follows since we are assuming $\rho \le 0$.

If $\bar\rho(y_0)^2 y_0 > \bar\rho(y)^2 y$ for all $y \in [\bar\rho(y_0)^2 y_0, y_0)$ (i.e. the first part of Assumption 2.1iii) holds) then $y_1^* - \bar\rho(y)^2 y > 0$ for all $y_1^* \in [\bar\rho(y_0)^2 y_0, y_0]$ and all $y \in [\bar\rho(y_0)^2 y_0, y_0)$, and we see that (35) is real and negative for $y_1^* \in (\bar\rho(y_0)^2 y_0, y_0]$ (since $\rho \le 0$), so in particular $y_1^*(x_1)$ is initially decreasing in x_1 as x_1 moves away from zero to the right, but at $y_1^* = \underline{y}_1^* := \bar\rho(y_0)^2 y_0$ we see that $\frac{dx_1}{dy_1^*} = -\infty$ i.e. $\frac{dy_1^*}{dx_1} = 0$ (since the integrand in (35) has a non-integrable singularity at $y = y_0$ when $y_1^* = \bar\rho(y_0)^2 y_0$), so $y_1^*(x_1)$ has a stationary point at $x_1 = x_1^* > 0$ defined by

$$x_1^* = \int_{\bar\rho(y_0)^2 y_0}^{y_0} (\frac{\bar\rho(y)^2 y}{\alpha(y)\sqrt{\bar\rho(y_0)^2 y_0 - \bar\rho(y)^2 y}} - \frac{\rho(y)\sqrt{y}}{\alpha(y)})dy\,, \tag{36}$$

which is the minimizer of $y_1^*(x)$ over all $x \in (-\infty, x_1^*]$, and from (32) we see that $\frac{dy}{dx} = 0$ at the point $(0, y_0)$. (34) is not defined (i.e. is not real-valued) for $y_1^* < \underline{y}_1^*$.

Thus we have shown that the shortest geodesic from $(0, y_0)$ to $\{x = x_1\}$ has $y'(0) < 0$ if $x \in [0, x_1^*)$. Moreover, $y_1^*(x_1)$ is initially *increasing* in x_1 as x_1 moves away from zero to the *left* and continues to increase, so the shortest geodesic from $(0, y_0)$ to $\{x = x_1\}$ will have $\frac{dy}{dt} \le 0$ for all $x_1 \le 0$, but now $y_1^* > y_0$, so the integral on the right cannot diverge in this case, so by $\frac{dy_1}{dx_1^*} < 0$.

By the same argument, (34) also holds for $x_1 < 0$, but now $y_1^*(x_1) > y_0$ and if we use the convention that $K_1 = 1$ as above then $t_1 < 0$. There is not a turning point along the geodesic for all $x_1 < 0$, since $y'(0) < 0$ and the geodesic also hits $x = x_1$ with negative slope.

3ii): The case $x > x_1^*$ where the geodesic has a turning point. For $x_1 > x_1^*$ (and note $x_1^* > 0$), again using Assumption 2.1iii), the shortest geodesic has a turning point at some $y = y_c > y_0$ before hitting the line $\{x = x_1\}$,

so in this case

$$\begin{aligned} x_1 &= \int_0^{t_1} \frac{dx}{dt}dt &= (\int_{y_0}^{y_c} + \int_{y_c}^{y_1^*}) y(\bar\rho(y)^2 + \frac{\rho(y)}{\sqrt{y}\alpha(y)}\frac{dy}{dt})\frac{dt}{dy}dy \quad \text{(from (28)) and setting } K_1 = 1 \text{ as above)} \\ &= (\int_{y_0}^{y_c} + \int_{y_1^*}^{y_c} y(\bar\rho(y)^2 + \frac{\rho(y)}{\sqrt{y}\alpha(y)}\frac{dy}{dt})|\frac{dt}{dy}|dy \\ &= (\int_{y_0}^{y_c} + \int_{y_1^*}^{y_c})\frac{y\bar\rho(y)^2}{\alpha(y)\sqrt{2E - y + y\rho(y)^2}}dy + (\int_{y_0}^{y_c} - \int_{y_1^*}^{y_c})\frac{\rho(y)\sqrt{y}}{\alpha(y)}dy \\ &= (\int_{y_0}^{y_c} + \int_{y_1^*}^{y_c})\frac{\bar\rho(y)^2 y}{\alpha(y)\sqrt{y_1^* - \bar\rho(y)^2 y}}dy + \int_{y_0}^{y_1^*}\frac{\rho(y)\sqrt{y}}{\alpha(y)}dy\,. \end{aligned} \tag{37}$$

$y'(t)$ will vanish iff $2E - \bar\rho(y)^2 y = 0$ i.e. iff $y_1^*(x_1) - \bar\rho(y)^2 y = 0$ at $y = y_c(x_1)$. So we set

$$y_1^*(x_1) \;:=\; y_c(x_1)\bar\rho(y_c(x_1))^2\,, \tag{38}$$

so we can re-write (37) as

$$x_1 = (\int_{y_0}^{y_c} + \int_{\bar\rho(y_c)^2 y_c}^{y_c})\frac{\bar\rho(y)^2 y}{\alpha(y)\sqrt{\bar\rho(y_c)^2 y_c - \bar\rho(y)^2 y}}dy + \int_{y_0}^{\bar\rho(y_c)^2 y_c}\frac{\rho(y)\sqrt{y}}{\alpha(y)}dy \tag{39}$$

and we solve for $y_c = y_c(x_1)$ for a given x_1, and then we can compute $y_1^*(x_1)$ using (38).

Since we are considering $x > x_1^*$, $\frac{dy}{dt} > 0|_{t=0}$ and we recall from (28) that $\frac{1}{y}(\frac{dx}{dt}) - \frac{\rho(y)}{\sqrt{y}\alpha(y)}\frac{dy}{dt} = \frac{1}{y}(\frac{dx}{dt}) - \frac{\rho(y)}{\sqrt{y}}\sqrt{y_1^* - y\bar\rho(y)^2} = \bar\rho(y)^2$. Then we see that $\frac{dx}{dt}|_{(0,y_0)} \le 0$ if and only if

$$y_1^* \;\ge\; \frac{y_0\bar\rho(y_0)^2}{\rho(y_0)^2} \tag{40}$$

(using that $\rho(y_0) \le 0$). This means for x_1 sufficiently large (or ρ sufficiently close to -1 and $x_1 > x_1^{*,\rho\to -1} = \int_0^{y_0}\frac{\sqrt{y}}{\alpha(y)}dy$, see Remark 2.4), the shortest geodesic from $(0, y_0)$ to $\{x = x_1\}$ bends back on itself in the x-direction (see first row of plots in Figure 4 for numerical examples of this phenomenon). Recall also however that as $\rho \to -1$, the distance to the line $\{x = x_1\}$ is infinite for $x_1 \ge \int_0^{y_0}\frac{\sqrt{y}}{\alpha(y)}dy$ (see Remark 2.4 above). To compute the y-value above y_0 where the geodesic returns to the y-axis, we have to solve (21) for this new "synthetic" y_0 value, and the choice of which formula applies in (21) depends on the particular parameters[3]. To compute the y-value where the geodesic turns around in the x-direction i.e. where $\frac{dx}{dt} = 0$, we have to find the root of $\frac{\rho(y)}{\sqrt{y}}\sqrt{y_1^* - y\bar\rho(y)^2} + \bar\rho(y)^2$.

From Corollary 2.3 we know that $y_1^*(x_1) \to \infty$ as $x_1 \to \infty$ there is a shortest geodesic from $(0, y_0)$ to $\{x = x_1\}$ for any $x_1 > 0$, since x_1 is monotonically increasing in y_c, and $x_1 \to \infty$ as $y_c \to \infty$.

4. Distance computations for $x \le x_1^*$ and $x \ge x_1^*$. For $x_1 \le x_1^*$, the distance from $(0, y_0)$ to $(x_1, y_1^*(x_1))$ is then given by

$$\begin{aligned} d(x_1) = \int_0^{t_1}\sqrt{2E}dt &= \int_0^{y_1^*}\sqrt{2E}\,|\frac{dt}{dy}|dy &= \sqrt{2E}\,|\int_{y_0}^{y_1^*}\frac{dy}{\alpha(y)\sqrt{2E - \bar\rho(y)^2 y}}| \\ &&= \sqrt{y_1^*(x_1)}\,|\int_{y_0}^{y_1^*}\frac{dy}{\alpha(y)\sqrt{y_1^* - \bar\rho(y)^2 y}}|\,. \end{aligned} \tag{41}$$

Similarly, for $x_1 > x_1^*$, the distance from $(0, y_0)$ to $(x_1, y_1^*(x_1))$ is then given by

$$\begin{aligned} d(x_1) = \int_0^{t_1}\sqrt{2E}dt &= \int_0^{y_1^*}\sqrt{2E}\,|\frac{dt}{dy}|dy &= \sqrt{2E}(\int_{y_0}^{y_c} + \int_{y_1^*}^{y_c})\frac{dy}{\alpha(y)\sqrt{2E - \bar\rho(y)^2 y}} \\ &&= (\int_{y_0}^{y_c} + \int_{y_1^*}^{y_c})\frac{dy}{\alpha(y)\sqrt{1 - \bar\rho(y)^2 y/y_1^*}} \\ &&= \sqrt{y_1^*(x_1)}(\int_{y_0}^{y_c} + \int_{y_1^*}^{y_c})\frac{dy}{\alpha(y)\sqrt{y_1^* - \bar\rho(y)^2 y}} \\ &&= \sqrt{y_1^*(x_1)}(\int_{y_0}^{y_c} + \int_{\bar\rho(y_c)^2 y_c}^{y_c})\frac{dy}{\alpha(y)\sqrt{\bar\rho(y_c)^2 y_c - \bar\rho(y)^2 y}}\,. \end{aligned}$$

■

[3] see left and right plots in the top of Figure 4 for numerical examples of both cases

Remark 2.8 We also mention in passing that if $-\bar\rho(y)^2 y$ is minimized at $y = \hat y = \bar\rho^2(y_0)y_0$ but this is not a stationary point and $y_{\min}$ is in the allowable range $(\bar\rho(y_0)^2 y, y_0)$, then we do not have the locally quadratic behaviour in (23), so the pole in (35) at $y = \hat y$ is integrable, and

$$\lim_{y_1^* \searrow y_{\min}} x(y_1^*) \;<\; \infty\,. \tag{42}$$

Proof. (of Corollary 2.3). Using the assumed tail behaviour $\alpha(y) \sim \nu y^p$ as $y \to \infty$ and $\rho(y) \to \rho_\infty \in [-1, 0)$ and (39), we see that

$$\begin{aligned} x_1 &\sim \bar\rho_\infty\Big(\int_{y_0}^{y_c} + \int_{\bar\rho_\infty^2 y_c}^{y_c}\Big)\frac{y}{\nu y^p\sqrt{y_c - y}}dy \;+\; \frac{\rho_\infty\bar\rho_\infty^2 y_c^{\frac32 - p}}{(\frac32 - p)\nu} \\ &= \frac{2y_c^{1-p}}{\nu}\big(\bar\rho\sqrt{y_c - y_0}\,{}_2F_1(\frac12, p-1, \frac32, 1 - \frac{y_0}{y_c}) + |\rho|\bar\rho\sqrt{y_c}\,{}_2F_1(\frac12, p-1, \frac32, \rho^2) \;+\; \frac{\rho_\infty\bar\rho_\infty^2 y_c^{\frac32 - p}}{(\frac32 - p)\nu} \\ &\sim \frac{2y_c^{1-p}}{\nu}\big[\bar\rho_\infty\sqrt{y_c}\,{}_2F_1(\frac12, p-1, \frac32, 1) \;+\; |\rho_\infty|\bar\rho_\infty\sqrt{y_c}\,{}_2F_1(\frac12, p-1, \frac32, \rho_\infty^2)\big] \;+\; \frac{\rho_\infty\bar\rho_\infty^2 y_c^{\frac32 - p}}{(\frac32 - p)\nu} \;=\; c^+_{\nu_\infty,\rho_\infty,p} y_c^{\frac32 - p} \end{aligned}$$

as $y_c \to \infty$, where $c^\pm_{\nu_\infty,\rho_\infty,p} = \frac{2\bar\rho_\infty}{\nu}({}_2F_1(\frac12, p-1, \frac32, 1) + {}_2F_1(\frac12, p-1, \frac32, \rho_\infty^2)|\rho_\infty|) \pm \frac{\rho_\infty\bar\rho_\infty^2}{(\frac32 - p)\nu}$ and ${}_pF_q$ is the generalized hypergeometric function[4], which implies that $y_1^*(x_1) = \bar\rho_\infty^2 y_c(x_1) \sim \bar\rho_\infty^2\big(\frac{x_1}{c^+_{\nu_\infty,\rho_\infty,p}}\big)^{\frac{1}{\frac32 - p}} \to \infty$ as $x_1 \to \infty$. Performing similar computations on (21) for $x_1 < 0$, we find that

$$\hat\sigma(x_1)^2 \;\sim\; \frac{4(p-1)^2}{(3-2p)^2}\Big(\frac{x_1}{c^\pm_{\nu_\infty,\rho_\infty,p}}\Big)^{\frac{2}{3-2p}}\bar\rho_\infty^2$$

as $x_1 \to \pm\infty$, and for the Heston case $p = \frac12$ this reduces to $\hat\sigma(x_1)^2 \sim \frac{\frac14\nu\bar\rho_\infty}{\arccos(\pm\rho_\infty)}|x|$ as $x_1 \to \pm\infty$ (which is consistent with the main result in Theorem 1.1 in [FJ09] for the Heston model). ■

[4] using Mathematica's definition

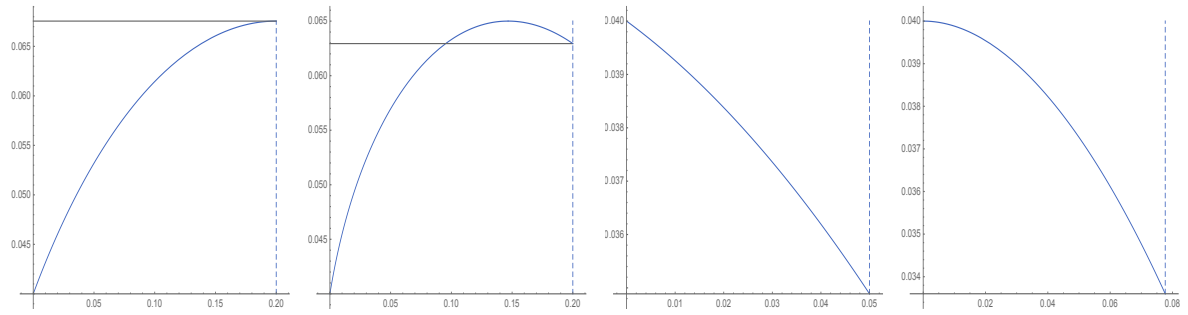

Figure 1: Here we have plotted the shortest geodesic (blue) to the vertical line $\{x = x_1\}$ for a (generalized) Heston model with $\rho(y) = 0$ (left) and $\rho(y) = -.4 + .4\tanh(25(y - y_0))$ (second from left), with $\alpha(y) = \nu\sqrt{y}$, $\nu = 0.4$, $y_0 = .04$ and $x_1 = 0.2$. For the second-from-right plot $\rho(y) = -.4$ and $x_1 = .05$ with all other parameters unchanged, and in this case we see there is no turning point. For the far right plot, the parameters are the same as the second-from-right plot but now x_1 is equal to the critical value $x_1^* = 0.0777161$. When $\rho(y)$ is constant, the numerical answers for $y_1^*(x_1)$ can be checked against the solution for the asymptotic implied volatility obtained via the Gärtner-Ellis theorem from the main Theorem 1.1 in [FJ09].

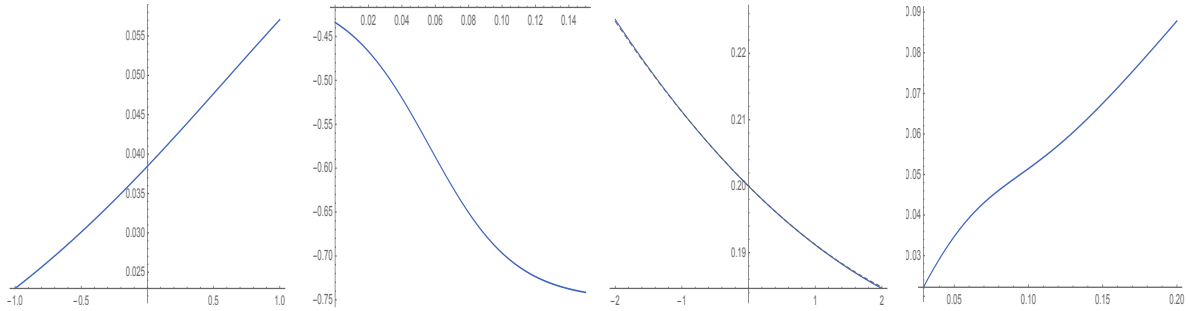

Figure 2: Here we have plotted the asymptotic VIX smile $\hat{\sigma}_{\text{VIX}}(x)$ (left) for $\alpha(y) = \frac{\nu y^{\frac{3}{2}}}{1+\frac{\nu}{\nu_\infty}y}$ with $\nu = \nu_\infty = .4$ and $\kappa = 0$, and assuming the observed asymptotic SPX smile $\hat{\sigma}(x) = \sigma_0 + \sigma_1 x + \sigma_2 x^2$ with $\sigma_0 = .2$, $\sigma_1 = -.01$, $\sigma_2 = .0012$, we find that that $\rho_0 = -.52$ and $\rho_1 = -3.11154$ and we have constructed a $\rho(y)$ function (second from left plot) consistent with ρ_0 and ρ_1 of the form in (16) and we have exogenously chosen the parameters in (16) so as to additionally impose that $\rho(\infty) = -.75$ and $\rho(0) = -.4$. In the second-from-right panel, we have plotted the induced SPX smile (in blue) obtained from the geodesic distance function $d_1(x_1)$ in (41), verses $\sigma_0 + \sigma_1 x + \sigma_2 x^2$ (grey dashed), and we see that both curves are in very close agreement as we would expect. In the final plot, we have plotted $\bar{\rho}(y)^2 y$ to verify that it is non-decreasing (see Assumption 2.1iii))

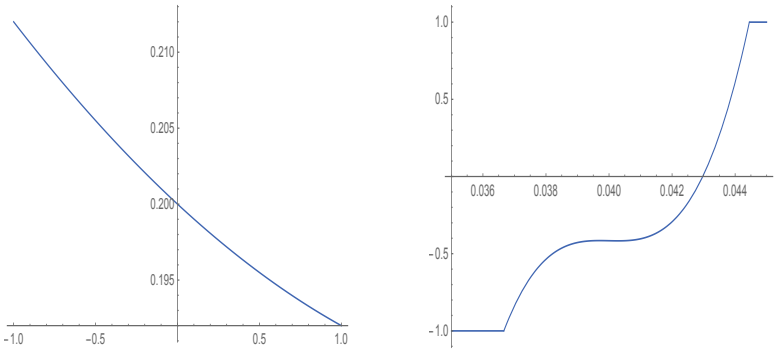

Figure 3: Here we have plotted a fictitious asymptotic SPX smile $\hat{\sigma}(x) = \sigma_0 + \sigma_1 x + \sigma_2 x^2 + \sigma_3 x^3 + \sigma_4 x^4$ with $\sigma_1 = -.01$, $\sigma_2 = .002$, $\sigma_3 = 0$ and $\sigma_4 = -.02$ (left) and assuming that $\alpha(y) = \frac{\nu y^{\frac{3}{2}}}{1+\frac{\nu}{\nu_\infty}y}$ with $\nu = \nu_\infty = .5$, we solve for $\rho(y) = \rho_0 + \rho_1(y - y_0) + \rho_2(y - y_0)^2 + \rho_3(y - y_0)^3 + \rho_4(y - y_0)^4$ (right) capped at -1 and 1, and (19) is satisfied. Note this $\rho(y)$ is only a Taylor series approximation to a true consistent ρ over a certain strict sub interval of $(0, \infty)$.

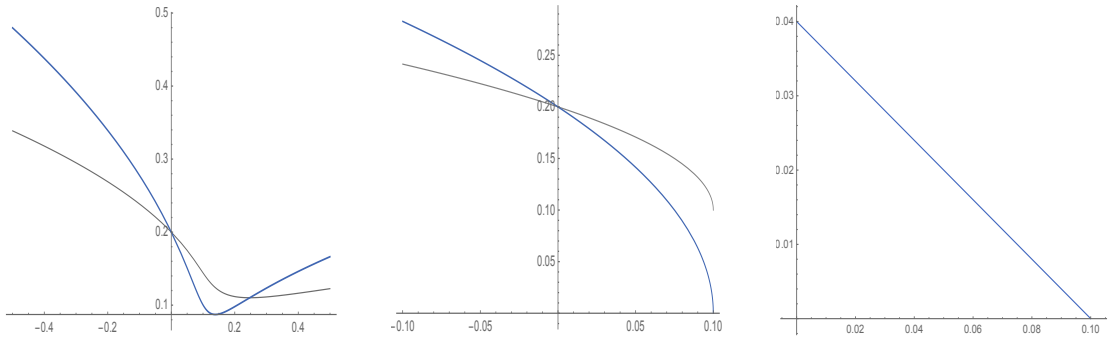

Figure 4: Here we have plotted $\hat{\sigma}(x_1)$ (grey) and $\sqrt{y_1^*(x_1)}$ (blue) for the standard Heston model with $y_0 = .04$, $\nu = .4$ and $\rho = -.4$ (left) and $\rho = -0.9$ (middle) (same parameters as the right plot in Figure 1), and $y_1^*(x_1)$ is minimized at $x_1 = x_1^* = 0.0777161$ on the left. For the middle plot, $y_1^*(x) = y_0 - \nu x_1$ which vanishes at $x_1 = y_0/\nu$ (see (5)) and taking the limit as $\rho \to -1$ in the main Theorem in [FJ11] we find that the rate function $\frac{1}{2}d(x_1)^2$ is the Legendre transform of $V(p)$ defined by $V(p) = \frac{y_0 p^2}{2+\nu p}$ for $p > -2/\nu$, and $V(p) = +\infty$ otherwise, and recall that the $\frac{1}{2}$-rule in (7) satisfied. On the right we have plotted the shortest geodesic to $x = y_0/\nu = .1$ for the case $\rho = -1$.

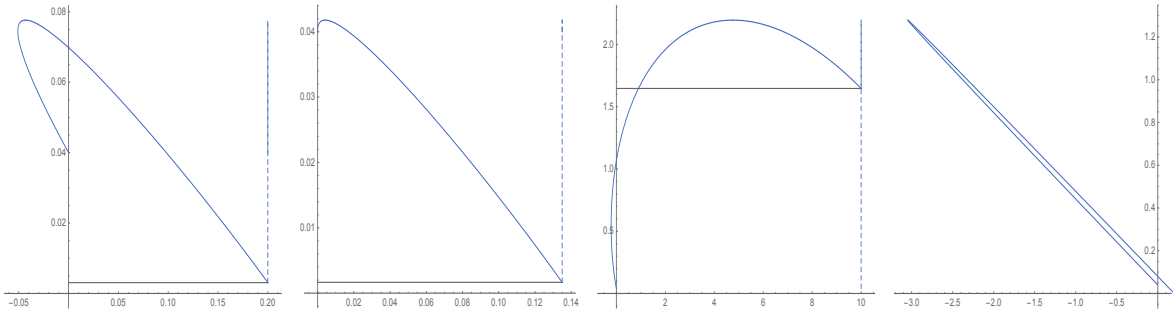

Figure 5: Here we have plotted the shortest geodesic (blue) to the vertical line $\{x = x_1\}$ for the Heston model with with $\alpha(y) = \nu\sqrt{y}$, $\nu = 0.4$, $y_0 = .04$, and $x_1 = 0.2$ (left) and $x_1 = .135$ (second from left), and for the second from right plot $\rho(y) = -.5$ and $x_1 = 7.5$ and for the right plot $\rho(y) = -.99995$ and $x_1 = 0.2$. In the first, third and fourth panels we see the extreme behaviour where the geodesic initially starts heading in the wrong direction, and the results are consistent with the rate function obtained via the Gärtner-Ellis theorem in the main Theorem 1.1 in [FJ09].

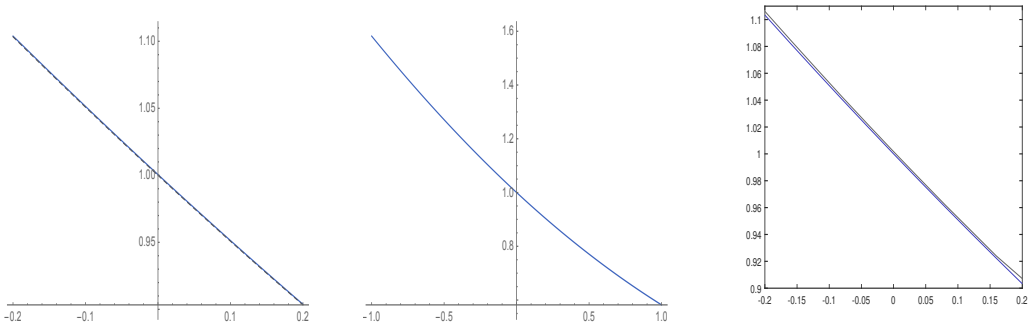

Figure 6: On the left we have plotted the asymptotic VIX implied volatility smile $\hat{\sigma}(x)$ in the $T \to 0$ limit for the standard Heston model (with $\kappa = 0$) as a function of x where $K = \sqrt{y_0}e^x$ (blue) verses the VIX implied volatility computed via numerical integration (grey dashed) over the known non-central χ-square density of V_T given above for $T = .001$, with $\kappa = 0$, $\nu = .4$ and $Y_0 = .04$, and we see that both curves are almost indistinguishable over this range of x-values. In the second panel, we have re-plotted the $\hat{\sigma}(x)$ over a wider range of x-values. The final panel again plots $\hat{\sigma}(x)$ (blue) verses the values obtained from Monte Carlo (grey) in Matlab with $T = .004$, 5million simulations, 1000 time steps and antithetic variables.

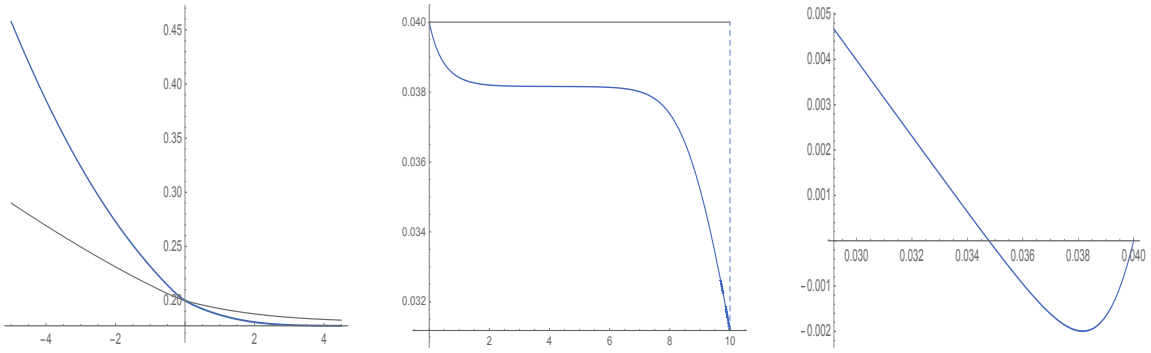

Figure 7: On the left here we have plotted $\sqrt{y_1^*(x_1)}$ (blue) and $\hat{\sigma}(x_1)$ (grey) for the non-standard case discussed in (24), with $\rho(y) = -0.575 + 0.175 \tanh(0.3252937830705757 - 489.56591415830565(y - 0.04))$ and the same y_0 and $\alpha(.)$ but with $\nu = .4$, and for this example $y_{\min} = .0311779$ and $\hat{y} = 0.0381605$, and in the middle we have plotted the shortest geodesic to $\{x = x_1\}$ for this non-standard case with $x_1 = 10$. On the right we have plotted $\bar{\rho}(y_0)^2 y_0 - \bar{\rho}(y)^2 y$ for $y \in [\bar{\rho}(y_0)^2 y_0, y_0]$, and we see that Assumption 2.1 is violated.

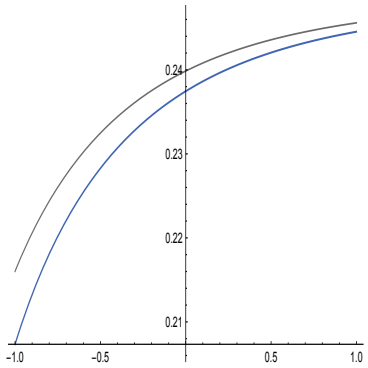

Figure 8: Here we have plotted the asymptotic VIX smile $\hat{\sigma}_{\text{VIX}}(x)$ in (17) for the mean-reverting Hull-White model with $\kappa = 1$, $\nu = 0.5$ and $\theta = y_0$ (blue) and $y_0 = .04, \theta = .05$ (grey).

3 Small-time VIX smiles under local-stochastic volatility

We now consider a local-stochastic volatility model of the form:

$$\left\{ \begin{array}{rcl} dS_t &=& S_t\sigma(S_t)\sqrt{Y_t}dW_t^1\,, \\ dY_t &=& \mu(S_t,Y_t)dt \,+\, \alpha(Y_t)dW_t^2\,, \end{array} \right.$$

where W^1, W^2 are independent Brownian motions, σ is bounded, Lipschitz and strictly monotonically increasing. We set $a(s) := \sigma(s)^2$ and assume α satisfy the same conditions as before and take $S_0 = 1$. Then

$$d(a(S_t)Y_t) \;\;=\;\; (\frac{1}{2}a''(S_t)S_t^2a(S_t)Y_t \,+\, a(S_t)\mu(S_t,Y_t))dt \,+\, dM_t\,,$$

for some local martingale M with respect to the filtration generated by S and Y. Hence if μ satisfies

$$\frac{1}{2}a''(s)a(s)s^2y \,+\, a(s)\mu(s,y) \;\;=\;\; a(s)(\frac{1}{2}a''(s)y \,+\, \mu(s,y)) \;\;=\;\; 0\,,$$

and assuming M is a true martingale, then $a(S_t)Y_t$ is a martingale with respect to the aforementioned filtration (this martingale feature will be needed in (44) below where we price VIX options). Note the unusual feature here that S is feeding back into the drift of Y here in general. Now let $X_t = g(S_t))$ where $g(s) = \int_{S_0}^s \frac{du}{u\sigma(u)}$, and set $a(S_t)Y_t = a(g^{-1}(X_t))Y_t = b(X_t)^2Y_t$, where $b(x) := \sigma(g^{-1}(x))$ and assume that b is decreasing, so

$$dX_t \;\;=\;\; \sqrt{Y_t}dW_t^1 \,+\, (\text{drift term}) \tag{43}$$

with $X_0 = 0$. The Lagrangian for (X,Y) is given by

$$L \;\;=\;\; \frac{1}{2}(\frac{1}{y}(\frac{dx}{dt})^2 + \frac{1}{\alpha(y)^2}(\frac{dy}{dt})^2)$$

and the x-component of the Euler-Lagrange equation for the geodesics under the metric induced by the inversion of the diffusion matrix for (X,Y) (i.e. with $g_{ij} = a_{ij}^{-1}$) is $\frac{d}{dt}(\frac{1}{y}\frac{dx}{dt}) = 0$. The usual transversality condition for European call options is $\frac{dy}{dt}|_{K,y_1^*} = 0$, leads to $\frac{1}{2}K_1^2y|_{x_1,y_1^*} = E$ (see [FJ11], or just set $\rho(y) = 0$ in the previous section). The price of a VIX call option with strike k under this model is then given by

$$\mathbb{E}((\mathbb{E}(\frac{1}{\Delta}\int_T^{T+\Delta} a(S_u)Y_u du|\mathcal{F}_T)^{\frac{1}{2}} - k)^+) \;\;=\;\; \mathbb{E}((\sigma(S_T)\sqrt{Y_T} - k)^+) \;\;=\;\; \mathbb{E}((b(X_T)\sqrt{Y_T} - k)^+) \tag{44}$$

since $a(S_t)Y_t$ is a martingale by construction, and formally applying the Freidlin-Wentzell LDP and the contraction principle we have

$$\lim_{T\to 0} T\log\mathbb{E}((b(X_T)\sqrt{Y_T} - k)^+) \;\;=\;\; -J(k)\,,$$

where $J(k) = \inf_{(x,y):b(x)\sqrt{y}=k} I(x,y) = \inf_{x>0} I(x, \frac{k^2}{b(x)^2})$, where $I(x,y) = \frac{1}{2}d(0,y_0;x,y)^2$ and $d(0,y_0;x,y)$ is the Riemannian distance under g_{ij} from $(0,y_0)$ to (x,y). Then the asymptotic VIX implied volatility $\hat{\sigma}(x)$ at strike $k = \sqrt{y_0}e^x$ satisfies

$$\hat{\sigma}(x) \;\;=\;\; \frac{x}{J(\sqrt{y_0}e^x)} \;\;=\;\; \frac{\log\frac{k}{\sqrt{y_0}}}{J(k)}\,. \tag{45}$$

So we now have to compute the shortest geodesic from $(0,y_0)$ to the *curve* $\psi_k(x) := \frac{k^2}{b(x)^2}$, which is increasing since we assume b is decreasing above. The tangent vector to this curve is $(1, -\frac{2k^2b'(x)}{b(x)^3})$, so the VIX call transversality condition is

$$0 \;\;=\;\; \left[1, -\tfrac{2k^2b'(x)}{b(x)^3}\right]\begin{bmatrix}\frac{1}{y} & 0\\ 0 & \frac{1}{\alpha(y)^2}\end{bmatrix}\begin{bmatrix}\frac{dx}{dt}\\ \frac{dy}{dt}\end{bmatrix} \;\;=\;\; \left[1, -\tfrac{2k^2b'(x)}{b(x)^3}\right]\begin{bmatrix}\frac{1}{y}\frac{dx}{dt}\\ \frac{1}{\alpha(y)^2}\frac{dy}{dt}\end{bmatrix} \;\;=\;\; \frac{1}{y}\frac{dx}{dt} - \frac{2k^2b'(x)}{b(x)^3\alpha(y)^2}\frac{dy}{dt} \;\;=\;\; 0 \tag{46}$$

$$\Rightarrow\;\; (\frac{dx}{dy} - \frac{2k^2b'(x)}{b(x)^3}\frac{y}{\alpha(y)^2})|_{(x^*,y^*)} \;\;=\;\; 0\,. \tag{47}$$

The geodesic equations are the same as in [FJ11] (or section 2 above with $\rho(y) \equiv 0$):

$$\frac{1}{2}(\frac{1}{y}(\frac{dx}{dt})^2 + \frac{1}{\alpha(y)^2}(\frac{dy}{dt})^2) \;\;=\;\; E \quad,\quad \frac{1}{y}\frac{dx}{dt} \;\;=\;\; K_1 \tag{48}$$

so

$$\frac{1}{2}(K_1^2 y + \frac{1}{\alpha(y)^2}(\frac{dy}{dt})^2) \quad = \quad E$$

or $\frac{dy}{dt} = \pm\alpha(y)\sqrt{2E - K_1^2 y}$. Plugging the last eq in (48) into the VIX call transversality condition (46) leads to $\frac{K_1 b(x)^3 \alpha(y)^2}{2k^2 b'(x)} = \frac{dy}{dt}$ and substituting this into the first equation in (48) we see that

$$\frac{1}{2}(K_1^2 y + (\frac{K_1 b(x)^3 \alpha(y)}{2k^2 b'(x)})^2)|_{(x_1^*, y_1^*)} \quad = \quad E\,,$$

so

$$(y + (\frac{b(x)^3\alpha(y)}{2k^2 b'(x)})^2)|_{(x_1^*,y_1^*)} \quad = \quad (y + (\frac{b(\psi_k^{-1}(y_1))^3\alpha(y)}{2k^2 b'(\psi_k^{-1}(y_1))})^2)|_{(x_1^*,y_1^*)} \quad =: \quad c_1(y_1^*) \quad = \quad \frac{2E}{K_1^2}\,.$$

Then proceeding as in section 2, we find that y_1^* has to solve

$$x_1^* \quad = \quad \psi_k^{-1}(y_1^*) \quad = \quad \int_{y_1^*(x_1)}^{y_0} \frac{y}{\alpha(y)\sqrt{c_1(y_1^*) - y}} dy\,, \tag{49}$$

and we solve for a different y_1^* for each VIX strike-value k.

$b(.)$ is decreasing by assumption, so (from the transversality condition in (30)) the shortest geodesic from $(0, y_0)$ to the curve $\psi_k(.)$ hits ψ_k with negative slope. Moreover, ψ_k is increasing since b is decreasing, so if $k > \sqrt{y_0}$ then ψ_k lies above $(0, y_0)$ and the shortest geodesic heads upwards to meet $\psi_k(.)$; conversely if $k < \sqrt{y_0}$ then ψ_k sits below $(0, y_0)$ and the shortest geodesic heads downwards to meet ψ_k (see Figure 9). When $k^2 = y_0$, then $k^2/b(0)^2 = y_0$ as well ($b(0) = 1$ here from the definition of $b()$ because $S_0 = 1$ by assumption), then $J(k) = 0$ because in this case $(0, y_0)$ is a point on the curve ψ_k.

We then have the corresponding formula for the shortest distance from $(0, y_0)$ to $\psi_k(.)$:

$$J(k) \quad = \quad \left| \int_{y_0}^{y_1^*(x_1)} \frac{1}{\alpha(y)\sqrt{1 - \frac{y}{c_1(y_1^*)}}} dy \right|\,, \tag{50}$$

and from this we can compute the asymptotic VIX implied volatility using (45).

3.1 The CEV and CEV-Heston cases

For the CEV case where $\sigma(s) = \delta s^{\beta-1}$ for $\beta \in (0, 1)$ and $\delta = 1$, $b(x) = (1 + x - x\beta)^{-1}$ is indeed decreasing and $\mu(s, y) = -\frac{1}{2}s^{-2+2\beta}y$. If we also impose that $\alpha(y) = \nu\sqrt{y}$ (which corresponds to an uncorrelated CEV-Heston type model with a feedback drift term), then (after some Mathematica computations which exploit the fact that the integrals in (49) and (50) can explicitly computed in terms of basic trigonometric functions for the Heston case, see also discussion in the introduction about geodesics being transformations of the standard cycloid) we obtain the small log-moneyness expansions:

$$J(k) \quad = \quad \frac{2(k - \sqrt{y_0})}{\sqrt{4y_0^2\bar{\beta}^2 + \nu^2}} - \frac{4(y_0^{\frac{3}{2}}\bar{\beta}^2(8y_0^2\bar{\beta}^2 + 3\nu^2)(k - \sqrt{y_0})^2}{(4y_0^2\bar{\beta}^2 + \nu^2)^{\frac{5}{2}}} + O((k - \sqrt{y_0})^3)\,, \tag{51}$$

$$\hat{\sigma}_{\text{VIX}}(x) \quad = \quad \frac{1}{2\sqrt{y_0}}\sqrt{4y_0^2\bar{\beta}^2 + \nu^2} + \frac{16y_0^4\bar{\beta}^4 + 4y_0^2\bar{\beta}^2\nu^2 - \nu^4}{4y_0(4y_0^2\bar{\beta}^2 + \nu^2)^{\frac{3}{2}}}(k - \sqrt{y_0}) + O((k - \sqrt{y_0})^2)\,, \tag{52}$$

when $S_0 = 1$, where $k = \sqrt{y_0}e^x$ as above and $\bar{\beta} = 1 - \beta$, and we find that the $O((k - \sqrt{y_0}))$ at-the-money skew term in the second equation is non-negative if and only if $\beta \le \beta^* := 1 - \frac{\nu\sqrt{\frac{1}{2}(\sqrt{5}-1)}}{2y_0}$, assuming $\beta^* > 0$ (see Figure 8).

The assumptions that $S_0 = 1$ and $\delta = 1$ are made W.L.O.G here, since for the general case

$$d(S_t/S_0) \quad = \quad \frac{1}{S_0}\delta S_t^\beta \sqrt{Y_t} dW_t \quad = \quad S_0^{\beta-1}\delta(\frac{S_t}{S_0})^\beta\sqrt{Y_t}dW_t \quad = \quad (\frac{S_t}{S_0})^\beta\sqrt{V_t}dW_t^1\,,$$
$$dV_t \quad = \quad d(\tilde{\delta}^2 Y_t) + (\text{drift term}) \quad = \quad \tilde{\delta}\nu\sqrt{V_t}dW_t^2 + (\text{drift term})\,,$$

where $V_t = \tilde{\delta}^2 Y_t$ and $\tilde{\delta} = (S_0^{\beta-1}\delta)^2$.

Note there is also a small log-moneyness expansion for European options in Theorem 4.1 in [FJ11] for a general uncorrelated local-stochastic volatility model. One should be able to generalize said Theorem and (52) to include non-zero ρ, and then choose y_0, ν, β and ρ to fit the overall level and slope of the SPX and VIX smiles at-the-money, but such computations will be messy since are not using a eikonal equation to compute (51), so we defer the details for future research.

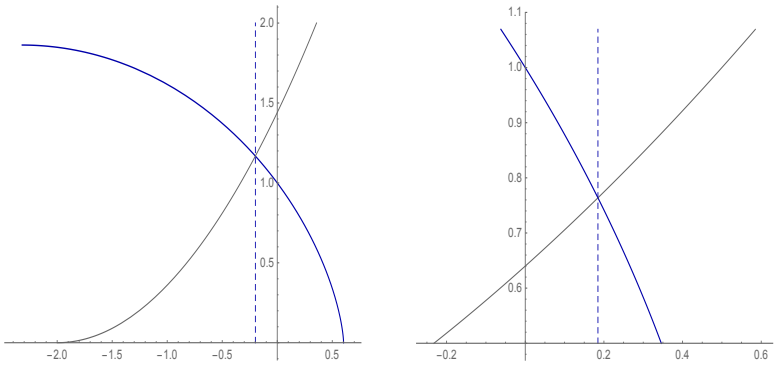

Figure 9: Here we have plotted the shortest geodesic (blue) to the curve $\psi_k(x) = \frac{k^2}{b(x)^2}$ (grey) for a CEV-Heston model with $\sigma(s) = s^{\beta-1}$, $\alpha(y) = \sqrt{y}$ for which $b(x) = (1 + x - x\beta)^{-1}$ with $\beta = \frac{1}{2}$, $y_0 = 1$, $S_0 = 1$ for $k = 1.2$ (left plot) and $k = .8$ (right plot), and since the metric is $ds^2 = \frac{1}{y}(dx^2 + dy^2)$ in this case, the shortest geodesic to ψ_k is perpendicular (in the usual Euclidean sense) to ψ_k at the hitting point (x_1^*, y_1^*). Note that $x_1^* < 0$ for the first case and the shortest geodesic goes up from $(0, y_0)$ to meet ψ_k, whereas in the second plot, $x_1^* > 0$ and the shortest geodesic goes down to meet ψ_k .

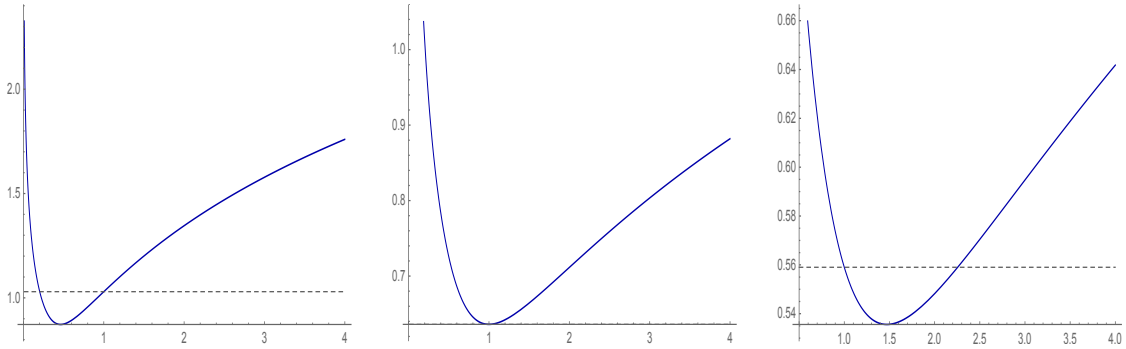

Figure 10: Here we have drawn the asymptotic VIX implied volatility smile using (15) as a function of the strike k for the same parameters as above except now $\beta = 0.1$, $\beta = \beta^* = 0.606924$ and $\beta = 0.75$ respectively, and note that $k = 1$ is the at-the-money strike value.

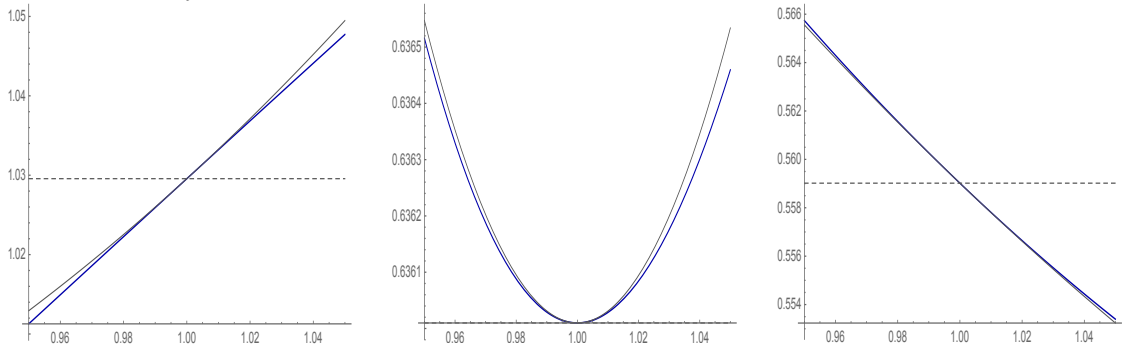

Figure 11: Here we have drawn the same three smiles again but now zoomed in around the at-the-money strike $k = \sqrt{y_0} = 1$ (blue) verses the at-the-money approximation obtained plugging (51) into (45) (grey) and the at-the-money skew flips from positive to negative as β goes above β^*.

References

[AIL22] E.Abi Jaber, C.Illand and X.Li (2022), Joint SPX–VIX calibration with Gaussian polynomial volatility models: deep pricing with quantization hints, arXiv preprint, `arxiv2212.08297`

[AIL22b] E.Abi Jaber, C.Illand and X.Li (2022), The quintic Ornstein-Uhlenbeck volatility model that jointly calibrates SPX&VIX smiles, to appear in `risk.net`

[AP07] L.B.G.Andersen and V.Piterbarg (2007), Moment Explosions in Stochastic Volatility Models, *Finance and Stochastics*, 11(1), 29-50.

[AFLZ17] J.Armstrong, M.Forde, M.Lorig and H.Zhang (2017), Small-time asymptotics under local-stochastic volatility with a jump-to-default: curvature and the heat kernel expansion, *SIAM J. Finan. Math.*, 8(1), 82-113.

[BBHK20] J.Backhoff-Veraguas, M.Beiglböck, M.Huesmann and S.Källblad (2020), Martingale Benamou–Brenier: A probabilistic perspective, *Ann.Probab.*, 48(5): 2258-2289.

[BL11] P.Baldi, and L.Caramellino (2011), General Freidlin–Wentzell Large Deviations and positive diffusions, *Statistics&Probability Letters*, 81(8), 1218-1229.

[1] [Ber94] J.Beran, Statistics for Long-Memory Processes (1994), Chapman and Hall, New York.

[BBF02] H.Berestycki, J.Busca and I.Florent (2002), Asymptotics and calibration of local volatility models, *Quantitative Finance*, 2, 61-69.

[BBF04] H.Berestycki, J.Busca and I.Florent (2004), Computing the Implied Volatility in Stochastic Volatility models, *Communications on Pure and Applied Mathematics*, Vol LVII, 1352-1373.

[BCPV22] A.E.Bolko, K.Christensen, M.S.Pakkanen and B.Veliyev (2022), A GMM approach to estimate the roughness of stochastic volatility, *Journal of Econometrics*, 235(2), 745–778, 2023.

[Bil99] P.Billingsley (1999), Convergence of probability measures, Wiley, second edition.

[BG22] F.Bourgey and J.Guyon (2022), Fast Exact Joint S&P 500/VIX Smile Calibration in Discrete and Continuous Time, SSRN preprint, `https://ssrn.com/abstract=4315084`

[BPS24] A.Bondi, S.Pulido and S.Scotti (2022), The rough Hawkes Heston stochastic volatility model, *Mathematical Finance*, 34(4), 1197-1241, 2024.

[BFN22] C.Bayer, M.Fukasawa, and S. Nakahara (2022), On the weak convergence rate in the discretization of rough volatility models, *SIAM J. Finan. Math.* 13(2).

[doC92] M.do Carmo (1992), Riemannian Geometry, Birkhäuser.

[CHLRS22b] C.Chong, M.Hoffmann, Y.Liu, and M.Rosenbaum, Grégoire Szymanski (2022), Statistical inference for rough volatility: Central limit theorems, preprint,

[CH21] C.Conze and P.Henry-Labordère (2021), Bass Construction with Multi-Marginals: Lightspeed Computation in a New Local Volatility Model, SSRN preprint,
`https://ssrn.com/abstract=3853085`

[CD22] R.Cont and P.Das (2022), Rough Volatility: Fact or Artefact?, arXiv preprint, `arxiv2203.13820`

[CGS22] C.Cuchiero, G.Gazzani, S.Svaluto-Ferro (2022), Signature-based models: theory and calibration, arXiv preprint, `arxiv2207.13136`

[DZ98] A.Dembo and O.Zeitouni (1998), Large Deviations Techniques and Applications, Jones and Bartlet publishers, Boston.

[Durr04] V.Durrleman (2004), "From implied to spot volatilities", PhD dissertation, Princeton University.

[Durr10] V.Durrleman (2010), From implied to spot volatilities, *Finance and Stochastics*, 14, 157–177.

[EFGR19] El Euch, O., M.Fukasawa, J.Gatheral, and M.Rosenbaum (2019), "Short-term at-the-money asymptotics under stochastic volatility models", *SIAM J. Finan. Math.*, 10(2), 491–511.

[FFF10] J.Feng, M.Forde, and J.P.Fouque (2010), Short maturity asymptotics for a fast mean-reverting Heston stochastic volatility model', *SIAM J. Finan. Math.*, 1(1), 126-141.

[FFK12] J.Feng, J.P.Fouque and R.Kumar (2012), Small-time asymptotics for fast mean-reverting stochastic volatility models, *Ann. Appl. Probab.*, Vol 22, No.4, 1541-1575.

[FS93] W.H.Fleming and H.M.Soner (1993), "Controlled Markov processes and viscosity solutions", Springer-Verlag, New York.

[FJ09] M.Forde and A.Jacquier (2009), Small-time asymptotics for implied volatility under the Heston model, *Int. J. Theor. Appl. Finance*, 12(6), 861-876.

[FJ11] M.Forde and A.Jacquier (2011), Small-time asymptotics for an uncorrelated Local-Stochastic volatility model, *Appl. Math. Finance*, 18(6), 517-535.

[FGS21] M.Forde, S.Gerhold and B.Smith (2021), Small-time, large-time and $H \to 0$ asymptotics for the Rough Heston model, *Mathematical Finance*, 31(1), 203-241.

[FFGS22] M.Forde, Fukasawa, M., S.Gerhold and B.Smith (2022), The Riemann-Liouville field and its GMC as $H \to 0$, and skew flattening for the rough Bergomi model, *Stat. Prob. Lett.*, Volume 181, Feb 2022.

[F23] M.Forde (2023), Statistical issues and calibration problems under rough and Markov volatility, presentation, King's College London, `https://nms.kcl.ac.uk/martin.forde/Talk.pdf`

[FK16] M.Forde, and R.Kumar (2016), Large-time option pricing using the Donsker-Varadhan LDP - correlated stochastic volatility with stochastic interest rates and jumps, *Ann. Appl. Probab.*, 26(6), 3699–3726.

[FSV21] M.Forde, M., B.Smith and L.Viitasaari (2021), Rough volatility and CGMY jumps with a finite history and the Rough Heston model - small-time asymptotics in the $k\sqrt{t}$ regime, *Quantitative Finance*, 21(4), 541-563, 21(4).

[FZ17] M.Forde, H.ZHang, Asymptotics for rough stochastic volatility models", *SIAM J. Finan. Math.*, 8(1), 114-145, 2017.

[FS21] M.Forde, and B.Smith (2021), Rough Heston with jumps - joint calibration to SPX/VIX level and skew as $T \to 0$, and issues with the quadratic rough Heston model, preprint.

[FPS00] J.P.Fouque, G.Papanicolaou, and K.R.Sircar (2000), Derivatives in financial markets with stochastic volatility, *Cambridge University Press.*

[FG22] M.Fukasawa and J.Gatheral (2022), A rough SABR formula, *Frontiers of Mathematical Finance*, 1(1), 81-97.

[FTW22] M.Fukasawa, T.Takabatake and R.Westphal (2022), Consistent estimation for fractional stochastic volatility model under high-frequency asymptotics (Is Volatility Rough?), *Mathematical Finance*, 32, 1086-1132.

[Fuk22] M.Fukasawa (2022), On asymptotically arbitrage-free approximations of the implied volatility, to appear in *Frontiers of Mathematical Finance.*

[FSW22] P.K.Friz, W.Salkeld and T.Wagenhofer (2022), Weak error estimates for rough volatility models, arXiv preprint, `arxiv2212.01591`

[GHLOW12] J.Gatheral, E. Hsu, E.P. Laurence, C.Ouyang and T.-H. Wang (2012), Asymptotics of implied volatility in local volatility models, *Mathematical Finance*, 22(4), 591-620.

[Gath06] J.Gatheral (2006), The Volatility Surface: A Practitioner's Guide, Wiley, New Jersey.

[GLOW22] I.Guo, G.Loeper, J.Obłój, S.Wang (2022), Optimal transport for model calibration, `risk.net`

[GLW22] I.Guo, G.Loeper, S.Wang (2022), Calibration of Local-Stochastic Volatility Models by Optimal Transport, *Mathematical Finance*, 32(1), 46-77.

[GL14] A.Gulisashvili and P.Laurence (2014), The Heston Riemannian distance function, *J.Math.Pures Appl.*, 101 303-329.

[GR20] J.Gatheral, and M.Rosenbaum (2020), "The quadratic rough Heston model and the joint S&P 500/VIX smile calibration problem", `risk.net`.

[Gul17] A.Gulisashvili (2017), Distance to the line in the Heston model, *Journal of Mathematical Analysis and Applications*, 450(1), 197-228.

[Guy21] J.Guyon (2021), Dispersion-Constrained Martingale Schrödinger Problems and the exact joint S&P 500/VIX Smile Calibration puzzle, SSRN, preprint, `https://ssrn.com/abstract=3853237`

[Guy21b] J.Guyon (2021), The Joint S&P 500/VIX Smile Calibration Puzzle Solved, presentation, NYU Courant Institute, `https://www.youtube.com/watch?v=pvq-rfajFRs`

[Guy22] J.Guyon, Dispersion-Constrained Martingale Schrödinger Bridges: Joint Entropic Calibration of Stochastic Volatility Models to S&P 500 and VIX Smiles, preprint, 2022

[GL22] J.Guyon and J.Lekeufack (2022), "Volatility is (mostly) path-dependent", SSRN preprint, https://ssrn.com/abstract=4174589

[GL22b] J.Guyon and J.Lekeufack (2022), "Does the Term-Structure of Equity At-the-Money Skew Really Follow a Power Law?", SSRN preprint, https://ssrn.com/abstract=4174538

[HL09] P.Henry-Labordère (2009), Analysis, Geometry, and Modelling in Finance: Advanced Methods in Option Pricing", Chapman & Hall.

[HL19] P.Henry-Labordère (2019), From (Martingale) Schrödinger Bridges to a new class of Stochastic Volatility Model, preprint.

[KT81] S.Karlin and H.Taylor (1981), A Second Course in Stochastic Processes, Academic Press, New York.

[KS91] I.Karatzas and S.Shreve (1991), Brownian motion and Stochastic Calculus, Springer-Verlag.

[Lew07] A.Lewis (2007), Geometries and Smile Asymptotics for a Class of Stochastic Volatility models, presentation at UCSB.

[Lew16] A.Lewis (2016), Advanced Smile Asymptotics: Geometry, Geodesics, and all that, (chapter 12 from *Option Valuation under Stochastic Volatility II*): 9798465205863: www.amazon.com.

[Lin01] V.Linetsky, Pricing and Hedging Path-Dependent Options Under the CEV Process (2001), *Management Science*, 47(7).

[LM07] P.L.Lions, and M.Musiela (2007), Correlations and bounds for stochastic volatility models, *Annales de l'Institut Henri Poincaré*, 24(1), 1-16.

[Pau10] L.Paulot (2010), "Asymptotic implied volatility at the second order with application to the SABR model", working paper.

[JMP21] Jacquier, A., Muguruza, A. and A.Pannier, "Rough Multifactor Volatility for SPX and VIX options", preprint, 2021

[RT96] E.Renault and N.Touzi (1996), Option hedging and implied volatilities in a stochastic volatility model, *Mathematical Finance*, 6(3), 279–302.

[Rom22] S.Rømer (2022), Hybrid multifactor scheme for stochastic Volterra equations with completely monotone kernels, preprint.

[Rom22b] S.Rømer (2022), Empirical analysis of rough and classical stochastic volatility models to the SPX and VIX markets, *Quantitative Finance*, 22(10), 1805-1838.

[RZ22] M.Rosenbaum and J.Zhang (2022), Deep calibration of the quadratic rough Heston model, risk.net.

[ST02] G.Samorodnitsky and M.S.Taqqu (2002), "Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance", CRC press.

[Var67] S.R.S.Varadhan (1967), On the behaviour of the fundamental solution of the heat equation with variable coefficients, *Comm. Pure Appl. Math.*, 20(2), 431-455.

[Var67b] S.R.S.Varadhan (1967), Diffusion processes in a small time interval, *Comm. Pure Appl. Math.*, 20, 659-685.

A The Lagrangian as a conserved quantity along geodesics

Take the total time derivative of L:

$$\frac{dL}{dt} = \frac{1}{2}\frac{d}{dt}[\sum_{i,j} g_{i,j}\dot{x}_i\dot{x}_j] = \frac{\partial L}{\partial \dot{x}_k}\ddot{x}_k + \frac{\partial L}{\partial x_k}\dot{x}_k = \frac{\partial L}{\partial \dot{x}_k}\ddot{x}_k + \frac{d}{dt}(\frac{\partial L}{\partial \dot{x}_k})\dot{x}_k = \frac{d}{dt}(\frac{\partial L}{\partial \dot{x}_k}\dot{x}_k)\,, \tag{A-1}$$

where we have used the Euler-Lagrange eq to obtain the third equality. But $L(x_1,..,x_n;\lambda\dot{x}_1,..,\lambda\dot{x}_n) = \lambda^2 L(x_1,..,x_n;\dot{x}_1,..,\dot{x}_n)$, i.e. L is homogenous of degree 2 in $(\dot{x}_1,..,\dot{x}_n)$. Thus, by differentiating wrt λ, and setting $\lambda = 1$, we obtain

$$\dot{x}_k\frac{\partial L}{\partial \dot{x}_k} = 2L\,.$$

Then taking the time-derivative of both sides, and using (A-1), we see that $\frac{dL}{dt} = 2\frac{dL}{dt}$, i.e. $\frac{dL}{dt} = 0$.

Now consider the related problem of computing the shortest distance in (F-1) (note the square root now inside the integral). Writing out the EL equations for this problem, we see that

$$\begin{aligned} & & \frac{d}{dt}(\frac{\partial L^{\frac{1}{2}}}{\partial \dot{x}}) &= \frac{\partial L^{\frac{1}{2}}}{\partial x}\,, \\ \Rightarrow \qquad & \frac{d}{dt}(\frac{1}{2}L^{-\frac{1}{2}}\frac{\partial L}{\partial \dot{x}}) &= \frac{1}{2}L^{-\frac{1}{2}}\frac{d}{dt}\frac{\partial L}{\partial \dot{x}} + \frac{\partial L}{\partial \dot{x}}\frac{d}{dt}(\frac{1}{2}L^{-\frac{1}{2}}) &= \frac{1}{2}L^{-\frac{1}{2}}\frac{\partial L}{\partial x}\,. \end{aligned}$$

But the second term in the middle expression is zero, so this simplifies to the same EL eq as above. Hence the function that minimizes $\int_0^1 L dt$ also minimizes $\int_0^1 \sqrt{L} dt$.

B Proof of Lemma 1.1

Let $\gamma(z,.)$ denote shortest geodesic from $(0,y_0)$ to the vertical line $\{x=z\}$. Consider the parametrized family of curves $f(s,t)=\gamma(x_1+s;t)$ with $f(s,1)=(x_1+s,y_1^*(x_1+s))$ so clearly $f(0,0)=(0,y_0)$. Then $V(1)=(1,\frac{\partial}{\partial x_1}y_1^*(x_1))$. Using Proposition 2.4, page 195 in [doC92][5]and that $\frac{D}{dt}\frac{d\gamma}{dt}=0$ for any geodesic[6], we have

$$\frac{1}{2}E'(0) = \frac{\partial}{\partial x_1}(\frac{1}{2}d(x_1)^2) = d(x_1)d'(x_1) = \mathbf{g}((1,\frac{\partial}{\partial x_1}y_1^*(x_1)),(\frac{dx}{dt},\frac{dy}{dt}))|_{(x_1,y_1^*)}\,. \tag{B-1}$$

From the transversality condition $-\frac{\rho(y)}{\alpha(y)\sqrt{y}}\frac{dx}{dt}+\frac{1}{\alpha(y)^2}\frac{dy}{dt}=0$ at (x_1,y_1^*), we see that

$$\begin{aligned} (\frac{dx}{dt},\frac{dy}{dt}) &= \alpha(y)\sqrt{2E-K_1^2 y\bar{\rho}(y)^2}\,(\frac{\sqrt{y}}{\alpha(y)\rho(y)},1) &= \alpha(y_1^*)d(x_1)\sqrt{1-\bar{\rho}(y_1^*)^2 y_1^*/y_1^*}\,(\frac{\sqrt{y_1^*}}{\alpha(y_1^*)\rho(y_1^*)},1) \\ & &= \alpha(y_1^*)d(x_1)|\rho(y_1^*)|\,(\frac{\sqrt{y_1^*}}{\alpha(y_1^*)\rho(y_1^*)},1) \end{aligned}$$

(since $\sqrt{2E}=d(x_1)$ and $\rho(y)\le 0$ by assumption) at (x_1,y_1^*), so (B-1) can be evaluated as

$$\left[1,\tfrac{\partial}{\partial x_1}y_1^*(x_1)\right]\frac{1}{\bar{\rho}(y)^2}\begin{bmatrix}\frac{1}{y} & -\frac{2\rho(y)}{\alpha(y)\sqrt{y}}\\ -\frac{2\rho(y)}{\alpha(y)\sqrt{y}} & \frac{1}{\alpha(y)^2}\end{bmatrix}\begin{bmatrix}\frac{dx}{dt}\\ \frac{dy}{dt}\end{bmatrix}|_{(x_1,y_1^*)} = \frac{d(x_1)}{\sqrt{y_1^*(x_1)}}$$

and the result follows after dividing by $d(x_1)$ and using the third equality in (B-1).

To prove the eikonal equation, we consider the variation with respect to y_0, and we obtain that

$$\frac{1}{2}E'(0) = \frac{\partial}{\partial y_0}(\frac{1}{2}d(x_1)^2) = d(x_1)d_{y_0}(x_1) = \mathbf{g}((0,1),(x'(0),y'(0))) = \frac{y'(0)-x'(0)\frac{\alpha(y_0)\rho(y_0)}{\sqrt{y_0}}}{\alpha(y_0)^2\bar{\rho}(y_0)^2}\,. \tag{B-2}$$

But

$$y'(0) = \alpha(y)\sqrt{2E-K_1^2 y\bar{\rho}(y_0)^2} = \alpha(y_0)d(x_1)\sqrt{1-y_0\bar{\rho}(y_0)^2/y_1^*} \tag{B-3}$$

using (31), and using that $\frac{1}{\bar{\rho}(y)^2}(\frac{1}{y}\frac{dx}{dt}-\frac{\rho(y)}{\sqrt{y}\alpha(y)}\frac{dy}{dt})=K_1$ we know that

$$x'(0) = y_0(\frac{\rho(y_0)}{\sqrt{y_0}\alpha(y_0)}y'(0) + K_1\bar{\rho}(y_0)^2)\,. \tag{B-4}$$

Substituting Eqs (B-3) and (B-4) into (B-2), and using that $y_1^*K_1^2=2E=d(x_1)^2$ we obtain:

$$\begin{aligned} \frac{1}{2}E'(0) &= \frac{1}{\alpha(y_0)}[\frac{1}{\bar{\rho}(y_0)^2}d(x_1)\sqrt{1-\frac{y_0\bar{\rho}(y_0)^2}{y_1^*}}(1-y_0(\frac{\rho(y_0)}{\sqrt{y_0}\alpha(y_0)}\frac{\alpha(y_0)\rho(y_0)}{\sqrt{y_0}})-\frac{1}{\bar{\rho}(y_0)^2}y_0K_1\bar{\rho}(y_0)^2\frac{\alpha(y_0)\rho(y_0)}{\sqrt{y_0}}] \\ &= \frac{1}{\alpha(y_0)}[d(x_1)\sqrt{1-\frac{y_0\bar{\rho}(y_0)^2}{y_1^*}}-\rho(y_0)\sqrt{\frac{y_0}{y_1^*}}d(x_1)]\,. \end{aligned} \tag{B-5}$$

Completing the square on the left hand side of the Eikonal equation we see that

$$\begin{aligned} yd_x^2+2\rho(y)\sqrt{y}\alpha(y)d_xd_y+\alpha(y)^2d_y^2 &= yd_x^2\bar{\rho}(y)^2+(\alpha(y)d_y+\rho(y)\sqrt{y}d_x)^2 \\ &= yd_x^2\bar{\rho}(y)^2+(\alpha(y)d_yd+\rho(y)\sqrt{y}d_xd)^2/d^2 \\ &= \frac{y}{y_1^*}\bar{\rho}(y)^2+(\alpha(y)d_yd+\rho(y)\sqrt{y}d_xd)^2/d^2\,, \end{aligned}$$

[5]Note that the Energy functional E as we define it here is one-half of the E that is used in [doC92]]

[6]where $\frac{D}{dt}$ denotes the covariant derivative, see Proposition 2.5 in [doC92]

where d is shorthand for $d(x_1)$ here. Using that $d_x = 1/\sqrt{y_1^*}$, we see that

$$\begin{aligned}\alpha(y)d_{y_0}d + \rho(y_0)\sqrt{y_0}d_x d &= d(x_1)\sqrt{1-\frac{y_0\bar{\rho}(y_0)^2}{y_1^*}} - \rho(y_0)\sqrt{\frac{y_0}{y_1^*}}d(x_1) + \rho(y)\sqrt{y_0}d(x_1)/\sqrt{y_1^*} \\ &= d(x_1)\sqrt{1-\frac{y_0\bar{\rho}(y_0)^2}{y_1^*}}\,,\end{aligned}$$

where we have used that $\frac{1}{2}E'(0) = dd_y$ (see above) and (B-5) in the first equality. Hence the left hand side of the eikonal equation equals 1, as required.

C Proof of Lemma 1.3

From Ito's lemma we know that $dY_t^2 = 2Y_t dY_t + \alpha(Y_t)^2 dt$ Then using the linear growth assumption on α and Jensen, we see that so

$$\begin{aligned}\frac{\partial}{\partial t}\mathbb{E}(Y_t^2) &= 2\mathbb{E}(Y_t\kappa(\theta - Y_t)) + \mathbb{E}(\alpha(Y_t)^2) &\leq 2\kappa\theta\mathbb{E}(Y_t) - 2\kappa\mathbb{E}(Y_t^2) + \bar{\nu}\mathbb{E}(Y_t^2)ds \\ &= 2\kappa\theta(\theta + (y_0-\theta)e^{-\kappa t}) - 2\kappa\mathbb{E}(Y_t^2) + \bar{\nu}\mathbb{E}(Y_t^2) \\ &\leq \bar{c} + \nu_1\mathbb{E}(Y_t^2)\end{aligned}$$

for some constant $\bar{c}$, where $\nu_1 := \bar{\nu} - 2\kappa$. Then from Gronwell's lemma we see that

$$\mathbb{E}(Y_t^2) \leq \frac{1}{\nu_1}(e^{\nu_1 t}(\bar{c} + y_0^2\nu_1) - \bar{c})$$

(recall that $Y_0 = y_0$). Thus

$$\mathbb{E}((Y_t - Y_0)^2) = \mathbb{E}(Y_t^2) - 2y_0\mathbb{E}(Y_t) + y_0^2 \leq \frac{1}{\nu_1}(e^{\nu t}(\bar{c} + y_0^2\nu_1) - \bar{c}) - 2y_0(\theta + (y_0-\theta)e^{-\kappa t}) + y_0^2 = O(t)\,,$$

so $\frac{1}{\sqrt{t}}\mathbb{E}((Y_t - Y_0)^2)^{\frac{1}{2}} = O(1)$ for t sufficiently small; hence the family of random variables $\Upsilon_t := \frac{Y_t - Y_0}{\sqrt{t}}$ is U.I. for $t \in [0, t^*]$ for some $t^* > 0$. Moreover, if we let $Z_t^\varepsilon = \varepsilon^{-\frac{1}{2}}(Y_{\varepsilon t} - Y_0)$ then $Y_{\varepsilon t} = Y_0 + \sqrt{\varepsilon}Z_t^\varepsilon$ and

$$dZ_t^\varepsilon = \varepsilon^{-\frac{1}{2}}dY_{\varepsilon t} = \varepsilon^{-\frac{1}{2}}(\kappa(\theta - Y_{\varepsilon t})\varepsilon dt + \alpha(Y_{\varepsilon t})dW_{\varepsilon t}) = \kappa(\theta - Y_{\varepsilon t})\sqrt{\varepsilon}dt + \alpha(Y_{\varepsilon t})dB_t^\varepsilon\,,$$

where $B_t^\varepsilon := W_{\varepsilon t}/\sqrt{\varepsilon}$ is another Brownian motion, and we can further write Z^ε as

$$Z_t^\varepsilon = \int_0^t \kappa(\theta - Y_{\varepsilon u})\sqrt{\varepsilon}du + \tilde{W}_{\int_0^t \alpha(Y_{\varepsilon u})^2 du}$$

for some other Brownian motion $\tilde{W}$, using the usual Dambis-Dubins-Schwarz time-change.

Since α is increasing, $\alpha(Y_{\varepsilon t})^2 \leq \alpha(\bar{Y}_{\varepsilon t})^2 < \infty$ a.s., where $\bar{Y}_t = \max_{0\leq s\leq t} Y_s$ Hence by the dominated convergence theorem, $\int_0^1 \alpha(Y_{\varepsilon t})^2 dt \to \alpha(Y_0)^2$ a.s. and hence also in probability. Then from a similar argument in the footnote on page 4 of [EFGR19], $Z_1^\varepsilon \xrightarrow{w} \alpha(Y_0)\tilde{W}_1 = \alpha(Y_0)\bar{Z}$ as $\varepsilon \to 0$, where $\bar{Z} \sim N(0,1)$, so $\Upsilon_t := (Y_t - Y_0)/\sqrt{t} \xrightarrow{w} \alpha(Y_0)\bar{Z}$ as $t \to 0$, where $\bar{Z} \sim N(0,1)$ as $t \to 0$, as we would expect.

Now define $\tilde{Y}_t := \frac{1}{\sqrt{Y_t} + \sqrt{Y_0}}$ (note this is not the same $\tilde{Y}$ as appears in section 2). Then $\tilde{Y}_t$ is a continuous function of Y_t and $Y_t \to \tilde{Y}_0 := \frac{1}{2\sqrt{Y_0}}$ a.s. as $t \to 0$ so (by the continuous mapping theorem) $\tilde{Y}_t \to \tilde{Y}_0$ in probability, and clearly $\tilde{Y}_t \leq \frac{1}{\sqrt{Y_0}}$. Note that $\frac{\sqrt{Y_t} - \sqrt{Y_0}}{\sqrt{t}} = \Upsilon_t\tilde{Y}_t$ (recall that $\Upsilon_t = (Y_t - Y_0)/\sqrt{t}$), and from above we know that $\Upsilon_t \xrightarrow{w} \alpha(Y_0)\bar{Z}$. From the general standard result that if $X_n \xrightarrow{w} X$ and $Y_n \to c$ (a constant) in probability, then $(X_n, Y_n) \xrightarrow{w} (X, c)$, we see that $(\Upsilon_t, \tilde{Y}_t)$ tends weakly to $(\alpha(Y_0)\bar{Z}, \tilde{Y}_0)$, and from the continuous mapping theorem $\Upsilon_t\tilde{Y}_t$ tends weakly to $\alpha(Y_0)\bar{Z}\tilde{Y}_0$. Moreover, $\tilde{Y}_t$ is uniformly bounded so $\Upsilon_T\tilde{Y}_t$ is also U.I. Then by Theorem 3.5 in Billingsley[Bil99], $\mathbb{E}(\Upsilon_t\tilde{Y}_t) = \mathbb{E}(\frac{\sqrt{Y_t} - \sqrt{Y_0}}{\sqrt{t}}) \to \tilde{Y}_0\mathbb{E}(Z) = 0$.

D Proof of Corollary 1.4

(i) Lower bound. For any $\delta > 0$, we have

$$\mathbb{E}(\mathrm{VIX}_T - \mathrm{VIX}_0 e^x)^+) \geq \delta\mathbb{P}(\mathrm{VIX}_T > \mathrm{VIX}_0 e^x + \delta)\,.$$

Then from (14) we see that

$$\liminf_{T\to 0} T\log\mathbb{E}(\mathrm{VIX}_T - e^x)^+ \;\geq\; \liminf_{T\to 0} T\log\mathbb{P}(\mathrm{VIX}_T > \mathrm{VIX}_0 e^x + \delta) \;=\; -\frac{1}{2} d_{\mathrm{VIX}}(\mathrm{VIX}_0 e^x + \delta))^2\,.$$

Letting $\delta \to 0$ and using the continuity of d_{VIX}, we obtain the desired lower bound.

(ii) Upper bound. We note that for $q > 1$, we have

$$\mathbb{E}(\mathrm{VIX}_T - e^x)^+ \;=\; \mathbb{E}((\mathrm{VIX}_T - e^x)^+ 1_{\mathrm{VIX}_T \geq e^x}) \;\leq\; \mathbb{E}[((\mathrm{VIX}_T - e^x)^+)^q]^{1/q}\,\mathbb{E}(1_{\mathrm{VIX}_T \geq e^x})^{1-1/q}.$$

Thus

$$\begin{aligned} T\log\mathbb{E}(\mathrm{VIX}_T - e^x)^+ &\leq \frac{T}{q}\log\left[\mathbb{E}[((\mathrm{VIX}_T - e^x)^+)^q]\right] + T(1-\frac{1}{q})\log\mathbb{P}(\mathrm{VIX}_T \geq e^x) \\ &\leq \frac{T}{q}\log\mathbb{E}(\mathrm{VIX}_T^q) + T(1-\frac{1}{q})\log\mathbb{P}(\mathrm{VIX}_T \geq e^x). \end{aligned} \tag{C-1}$$

But using that $\mathrm{VIX}_T^q) = (aY_T + b)^{\frac{1}{2}q}$ and the same approach as (13), we find that

$$\mathbb{E}(Y_T^q) \;=\; y_0^q + T\frac{1}{2} q y^{q-2}(2y\kappa(\theta - y) + (q-1)y^{2p}\xi^2) + O(T^2)\,,$$

so $T\log\mathbb{E}(Y_T^q) \to 0$ as $T \to 0$. If we then take $\lim_{q\to\infty}\limsup_{t\to 0}$ on both sides of (C-1), we see that

$$\limsup_{t\to 0} t\log\mathbb{E}(\mathrm{VIX}_T - S_0 e^x)^+ \;\leq\; -d_{\mathrm{VIX}}(\mathrm{VIX}_0 e^x)$$

as required. The case $x < 0$ follows by similar arguments.

E Proof of Corollary 1.5

For convenience, we let $J(x) := \frac{1}{2}(d_{\mathrm{VIX}}(\mathrm{VIX}_0 e^x)^2$. Let $C^{\mathrm{BS}}(S, K, \sigma, T)$ denote the usual Black-Scholes call option formula with zero interest rate and dividend. Then can easily verify that for any $b \in \mathbb{R}$

$$\lim_{T\to 0} T\log C^{\mathrm{BS}}(\mathrm{VIX}_0 + b\sqrt{T}, \mathrm{VIX}_0 e^x, \sigma, T) \;=\; -\frac{x^2}{2\sigma^2}$$

so from Lemma 1.3 (and using that C^{BS} is monotonic in its first argument) we see that

$$\lim_{T\to 0} T\log C^{\mathrm{BS}}(\mathbb{E}(\mathrm{VIX}_T), \mathrm{VIX}_0 e^x, \sigma, T) \;=\; -\frac{x^2}{2\sigma^2}\,.$$

For any $\delta \in (0, J(x))$, we can then choose σ so that $-J(x) = -\frac{x^2}{2\sigma^2} - \delta$. Then from Corollary 2.3

$$\begin{aligned} -J(x) &= \limsup_{T\to 0} T\log\mathbb{E}((\mathrm{VIX}_T - \mathrm{VIX}_0 e^x)^+) \\ &= \limsup_{T\to 0} T\log C^{\mathrm{BS}}(\mathbb{E}(\mathrm{VIX}_T), \mathrm{VIX}_0 e^x, \hat{\sigma}_{\mathrm{VIX}}(x,T), T) \quad \text{(by definition of } \hat{\sigma}_{\mathrm{VIX}}(x,T)) \\ &< \lim_{T\to 0} T\log C^{\mathrm{BS}}(\mathbb{E}(\mathrm{VIX}_T), \mathrm{VIX}_0 e^x, \sigma, T) \;=\; -\frac{x^2}{2\sigma^2}\,. \end{aligned}$$

Since $C^{\mathrm{BS}}(.)$ is monotonically increasing in the σ argument, we see that $\limsup_{T\to 0}\hat{\sigma}_{\mathrm{VIX}}(x,T) \leq \sigma$. Finally we let $\delta \to 0$, and we proceed similarly for the lower bound.

F Proof of Lemma 1.1

Let $f_a(y) = (y_0 e^{-a} \vee (y \wedge y_0 e^a)^{\frac{1}{2}}$ and $\alpha_a(y) = y_0 e^{-a} \vee (\alpha(y) \wedge y_0 e^a)$ for $a > y_0$, and consider the following re-scaled variant of the model in (1):

$$\begin{cases} d\hat{X}_t^\varepsilon \;=\; \sqrt{\varepsilon} f_a(\hat{Y}_t^\varepsilon)(\bar{\rho}(\hat{Y}_t^\varepsilon) dW_t^1 + \rho(\hat{Y}_t^\varepsilon) dW_t^2)\,, \\ d\hat{Y}_t^\varepsilon \;=\; \sqrt{\varepsilon}\alpha_a(\hat{Y}_t^\varepsilon) dW_t^2 \end{cases}$$

for $\varepsilon > 0$, and the corresponding version with drift:

$$\begin{cases} d\tilde{X}_t^\varepsilon \;=\; -\frac{1}{2}\varepsilon f_a(\tilde{Y}_t^\varepsilon)^2 dt + \sqrt{\varepsilon} f_a(\tilde{Y}_t^\varepsilon)(\bar{\rho}(\tilde{Y}_t^\varepsilon) dW_t^1 + \rho(\tilde{Y}_t^\varepsilon) dW_t^2)\,, \\ d\tilde{Y}_t^\varepsilon \;=\; \varepsilon\kappa(\theta - \tilde{Y}_t^\varepsilon) dt + \sqrt{\varepsilon}\alpha_a(\tilde{Y}_t^\varepsilon) dW_t^2 \;=\; \sqrt{\varepsilon}\alpha_a(\tilde{Y}_t^\varepsilon)(\sqrt{\varepsilon}\frac{\kappa}{\alpha_a(\tilde{Y}_t^\varepsilon)}(\theta - \tilde{Y}_t^\varepsilon) dt + dW_t^2) \end{cases}$$

with $\hat{X}_0^\varepsilon = \tilde{X}_0^\varepsilon = 0$ and $\hat{Y}_0^\varepsilon = \tilde{Y}_0^\varepsilon = Y_0 = y_0$. Let $\bar{X}$ and $\underline{X}$ denote the running maximum and minimum respectively of a generic process X, and let $\hat{A}_\varepsilon := \{\underline{\hat{Y}}_1^\varepsilon > y_0 e^{-a}\} \cap \{\bar{\hat{Y}}_1^\varepsilon < y_0 e^a\}$. Then

$$\begin{aligned}
\mathbb{P}(X_\varepsilon > x_1) &\leq \mathbb{E}(1_{X_\varepsilon > x_1} 1_{\{\underline{Y}_\varepsilon > y_0 e^{-a}\} \cap \{\bar{Y}_\varepsilon < y_0 e^a\}}) + \mathbb{P}((\{\underline{Y}_\varepsilon > y_0 e^{-a}\} \cap \{\bar{Y}_\varepsilon < y_0 e^a\})^c) \\
&= \mathbb{E}(1_{\tilde{X}_\varepsilon > x_1} 1_{\{\underline{\tilde{Y}}_\varepsilon > y_0 e^{-a}\} \cap \{\bar{\tilde{Y}}_\varepsilon < y_0 e^a\}}) + \mathbb{P}((\{\underline{Y}_\varepsilon > y_0 e^{-a}\} \cap \{\bar{Y}_\varepsilon < y_0 e^a\})^c) \\
&= \mathbb{E}(1_{\tilde{X}_1^\varepsilon > x_1} 1_{\{\underline{\tilde{Y}}_1^\varepsilon > y_0 e^{-a}\} \cap \{\bar{\tilde{Y}}_1^\varepsilon < y_0 e^a\}}) + \mathbb{P}((\{\underline{Y}_1^\varepsilon > y_0 e^{-a}\} \cap \{\bar{Y}_1^\varepsilon < y_0 e^a\})^c) \text{ (since } (X_{\varepsilon(.)}, Y_{\varepsilon(.)}) \sim (\tilde{X}_\varepsilon, \tilde{Y}_\varepsilon)) \\
&= \mathbb{E}(e^{\int_0^1 \sqrt{\varepsilon}\gamma(\hat{Y}_t^\varepsilon) dW_t^2 - \frac{1}{2}\sqrt{\varepsilon}\int_0^1 \gamma(\hat{Y}_t^\varepsilon)^2 dt} (1_{\hat{X}_1^\varepsilon > x_1} 1_{\hat{A}_\varepsilon} + 1_{\hat{A}_\varepsilon^c})) \qquad \text{(by Girsanov)} \\
&= \mathbb{E}(e^{\sqrt{\varepsilon}(\Gamma(\hat{Y}_t^\varepsilon) - \frac{1}{2}\int_0^1 \Gamma''(\hat{Y}_t^\varepsilon)\alpha(\hat{Y}_t^\varepsilon)^2 dt - \frac{1}{2}\int_0^1 \gamma(\hat{Y}_t^\varepsilon)^2 dt)} (1_{\hat{X}_1^\varepsilon > x_1} 1_{\hat{A}_\varepsilon} + 1_{\hat{A}_\varepsilon^c})) \qquad \text{(from Ito's formula)} \\
&\leq e^{\sqrt{\varepsilon}\sup_{y \in [\frac{1}{a}, a]} (\Gamma(y) + \frac{1}{2}|\Gamma''(y)\alpha(y)^2| + \frac{1}{2}|\gamma(y)^2| dt)} \mathbb{P}(\hat{X}_1^\varepsilon > x_1) + \mathbb{P}(\hat{A}_\varepsilon^c),
\end{aligned}$$

where $\gamma(y) = \frac{\kappa}{\alpha_a(y)}(\theta - y)$ and $\Gamma'(y)\alpha(y) = \gamma(y)$. Similarly

$$\mathbb{P}(X_\varepsilon > x_1) \geq \mathbb{E}(e^{(\ldots)} 1_{\hat{X}_1^\varepsilon > x_1} 1_A) = \mathbb{E}(e^{(\ldots)} 1_{\hat{X}_1^\varepsilon > x_1} 1_A) \geq e^{(\ldots)} \mathbb{P}(\hat{X}_1^\varepsilon > x_1) - \mathbb{P}(\hat{A}_\varepsilon^c),$$

where the Girsanov factor inside the expectations here is the same as above, and the final $e^{(\ldots)}$ term is now $e^{\sqrt{\varepsilon}\inf_{y \in [\frac{1}{a}, a]} (\Gamma(y) - \frac{1}{2}\int_0^1 \Gamma''(y) dt - \frac{1}{2}\int_0^1 \gamma(y)^2 dt)}$.

The diffusion coefficient for $\hat{X}^1$ is bounded and Lipschitz on $[y_0 e^{-a}, y_0 e^a]$ (since α is differentiable), so from standard Freidlin-Wentzell theory (see e.g. Theorem 6.3 in [Var67]), we know that $(\hat{X}^\varepsilon)$ satisfies the large deviation principle (LDP) as $\varepsilon \to 0$ with lower semi-continuous rate function

$$J(f) = \frac{1}{2}\int_0^1 \sum_{i,j=1}^2 g_{ij} \frac{df^i}{dt}\frac{df^j}{dt} dt$$

for $f \in C_{(0,y_0)}([0,1]; \mathbb{R} \times (0,\infty))$ where g_{ij}^a is equal to the inverse of the diffusion coefficient for $(\hat{X}^1, \hat{Y}^1)$ which (for x, y fixed and a sufficiently large) also has line element $ds^2 = \frac{1}{\bar{\rho}(y)^2}(\frac{1}{y}dx^2 - \frac{2\rho(y)}{\sqrt{y}\alpha(y)}dxdy + \frac{1}{\alpha(y)^2}dy^2)$, i.e. the same as for the original pair of processes (X, Y). Then from the contraction principle $\hat{X}_1^\varepsilon$ satisfies the LDP with good rate function

$$I_a(x_1) = \inf_{f \in C_{(0,y_0)}([0,1]): f_1(1) = x_1} J(f) = \frac{1}{2}(\inf_{f \in C_{(0,y_0)}([0,1]): f_1(1) = x_1} \int_0^1 \sum_{i,j=1}^2 \sqrt{g_{ij}^a \frac{df^i}{dt}\frac{df^j}{dt}} dt)^2 = \frac{1}{2} d_a(x_1^2), \tag{F-1}$$

(see e.g. Appendix A to see why both expressions are equal) where $d_a(x_1)$ is the shortest distance from $(0, y_0)$ to the vertical line $\{x = x_1\}$ under the metric g_{ij}^a, and $d_a(.)$ is continuous.

Moreover, the *two-sided maximum* $\max_{0 \leq t \leq 1} |\log \frac{\hat{Y}_t}{y_0}|$ is also a continuous functional of $\hat{Y}$ under the sup norm metric, so by the contraction principle $\max_{0 \leq t \leq 1} |\log \frac{\hat{Y}_t^\varepsilon}{y_0}|$ satisfies the LDP as $\varepsilon \to 0$ with good rate function $\Lambda(a) := \inf_{\phi \in C_{y_0}([0,1]): \underline{\phi}(1) \leq y_0 e^{-a} \text{ or } \bar{\phi}(1) \geq y_0 e^a} J(\phi)$ on $\mathbb{R}$, so

$$\lim_{\varepsilon \to 0} \varepsilon \log \mathbb{P}(\hat{A}_\varepsilon) = -\Lambda(a),$$

and from the goodness of the rate function we know that $\Lambda(a) > I_a(x)$ for a sufficiently large, so the $\mathbb{P}(\hat{A}_\varepsilon)$ term decays faster than the $\mathbb{P}(X_\varepsilon > x_1)$ term above. Thus, putting everything together, we arrive at

$$\lim_{\varepsilon \to 0} \varepsilon \log \mathbb{P}(X_\varepsilon > x_1) = \lim_{\varepsilon \to 0} \varepsilon \log \mathbb{P}(\hat{X}_1^\varepsilon > x_1) = -I_a(x_1)$$

(recall that $\hat{X}$ depends on a). Our geodesic computations in section 2 show that $I_a(x_1) = I(x_1)$ for a sufficiently large because the shortest geodesic from $(0, y_0)$ to $\{x = x_1\}$ stays inside $[y_0 e^{-a}, y_0 e^a]$ for a sufficiently large. Thus we have effectively argued away the effect of the non-zero drift of (X, Y) and the unbounded coefficients, which may also be non-Lipschitz outside this interval.