Arbitrage-free SABR

Article *in* Wilmott · January 2014

DOI: 10.1002/wilm.10290

CITATIONS
23

READS
7,961

4 authors, including:

Patrick S Hagan
Scotiabank
100 PUBLICATIONS **2,065** CITATIONS
SEE PROFILE

Andrew Lesniewski
City University of New York - Bernard M. Baruch College
94 PUBLICATIONS **1,898** CITATIONS
SEE PROFILE

ARBITRAGE FREE SABR

PATRICK S. HAGAN AND DEEP KUMAR
PATHAGAN1954@YAHOO.COM
MATHEMATICS INSTITUTE, 24-29 ST. GILES, OXFORD, UK OX1 3LB, UK
AVM LTD, BOCA RATON, FL

Abstract. Smile risk is often managed using the explicit implied vol formulas developed for the SABR model. These asymptotic formulas are not exact, and this can lead to arbitrage for low strike options. Here we provide an alternate method for pricing options under the SABR model: We use asympotic techniques to reduce the SABR model from two dimensions to one dimension. This leads to an effective one dimensional forward equation for the probability density which has the same asymptotic order of accuracy as the explicit implied vol formulas. We obtain arbitrage-free option prices by numerically solving this PDE. The implied volatilities obtained from the numerical solutions closely match the explicit implied volatility curves, apart from ultra-low strikes. For very low strikes, the implied absolute (normal) vol dips downwards, closely matching market observations. We also show how negative rates can be accomodated by replacing the F^β factor with $(F+a)^\beta$.

1. The SABR model. European option prices are often quoted by using the normal model. In this model the forward asset price $\tilde{F}(t)$ follows the process

$$d\tilde{F} = \sigma_N d\tilde{W}, \tag{1.1a}$$

and the (forward) price of European calls and puts works out to be

$$V^N_{call} = [f-K]\, N\left(\frac{f-K}{\sigma_N\sqrt{\tau_{ex}}}\right) + \sigma_N\sqrt{\tau_{ex}}\, G\left(\frac{f-K}{\sigma_N\sqrt{\tau_{ex}}}\right), \tag{1.1b}$$

$$V^N_{put} = [K-f]\, N\left(\frac{K-f}{\sigma_N\sqrt{\tau_{ex}}}\right) + \sigma\sqrt{\tau_{ex}}\, G\left(\frac{f-K}{\sigma_N\sqrt{\tau_{ex}}}\right). \tag{1.1c}$$

Here $N(\cdot)$ is the cumulative normal distribution and $G(\cdot)$ is the Gaussian density. Both V^N_{call} and V^N_{put} are increasing functions of the volatility σ_N. Consequently, European option prices can be quoted by stating the implied normal vol (*aka* absolute vol or bps per year vol), the unique value of σ_N that yields the option's dollar price when substituted into these formulas.

Alternatively, in Black's model the forward asset price is modeled by

$$d\tilde{F} = \sigma_B \tilde{F} d\tilde{W}. \tag{1.2}$$

Here, too, there is a one-to-one relation between the European option price and the Black (log normal) vol σ_B, so option prices can be quoted in terms of σ_B.Implied Black vols σ_B and implied normal vols σ_N are equivalent, and the mapping between σ_B and σ_N is well understood, so in this article we focus on the normal vols σ_N.

If the normal model correctly described the behavior of the asset price, then the same implied volatility σ_N would correctly price options with different strikes K and times-to-expiry τ_{ex}. In practice, matching market prices requires substantially different implied vols for options with different strikes and expiries, $\sigma_N = \sigma_N(K, \tau_{ex})$. For a given expiry τ_{ex}, the implied vol σ_N as a function of K is the options's smile, or skew. Handling the smile and skew judiciously is critical to an options desk, since the risks of options at all different strikes K have to be consolidated before the risks can be hedged efficiently. Although offsetting risks of options with different expiries is less common, handling the volatility surface (the dependence of σ_N on both K and τ_{ex}) is also be important for correctly pricing path dependent options, such as mid-curve options.

The need for handling smile and skew risk effectively led to the development of the SABR models[1-SABR]. These are models of the form

$$d\tilde{F} = \tilde{A}C(\tilde{F})d\tilde{W}_1, \tag{1.3a}$$

$$d\tilde{A} = \nu\tilde{A}d\tilde{W}_2, \tag{1.3b}$$

with

$$d\tilde{W}_1 d\tilde{W}_2 = \rho dt, \tag{1.3c}$$

where most commonly $C(F)$ is taken to be F^β. In [1-SABR], singular perturbation techniques were used to analyze this class of models in the small volatility regime. To carry out the expansion systematically, a small parameter ε was introduced,

$$d\tilde{F} = \varepsilon\tilde{A}C(\tilde{F})d\tilde{W}_1, \tag{1.4a}$$

$$d\tilde{A} = \varepsilon\nu\tilde{A}d\tilde{W}_2, \tag{1.4b}$$

$$d\tilde{W}_1 d\tilde{W}_2 = \rho dt. \tag{1.4c}$$

The model was analyzed in the $\varepsilon << 1$ limit, and then ε was set to 1 in the final result[1]. This analysis was used to obtain explict formulas for the implied vols of European options under the SABR model. There are now several variants of these formulas [1-5, SABR variants], all correct through $O(\varepsilon^2)$, but our favorite is

$$\sigma_N(K) = \frac{\varepsilon\alpha\,(f-K)}{\int_K^f \frac{df'}{C(f')}} \cdot \left(\frac{\zeta}{x(\zeta)}\right) \cdot \left\{1 + \left[g\alpha^2 + \tfrac{1}{4}\rho\nu\alpha\frac{C(f)-C(K)}{f-K} + \frac{2-3\rho^2}{24}\nu^2\right]\varepsilon^2\tau_{ex} + \cdots\right\}. \tag{1.5a}$$

Here $f = \tilde{F}(0), \alpha = \tilde{A}(0)$ are today's value of the foward price and the volatility, τ_{ex} is the time to exercise,

$$\zeta = \frac{\nu}{\alpha}\int_K^f \frac{df'}{C(f')}, \qquad x(\zeta) = \log\left(\frac{\sqrt{1-2\rho\zeta+\zeta^2}-\rho+\zeta}{1-\rho}\right), \tag{1.5b}$$

and the geometric factor is

$$g = \log\left(\frac{1}{f-K}\int_K^f \frac{\sqrt{C(f)C(K)}}{C(f')}df'\right) \Bigg/ \left(\int_K^f \frac{1}{C(f')}df'\right)^2. \tag{1.5c}$$

Although there are simpler formulas, this one seems to be the most robust.

The classic SABR model is the special case $C(F) = F^\beta$. For this case the implied vol formula reduces to

$$\sigma_N(K) = \frac{\varepsilon\alpha\,(1-\beta)\,(f-K)}{f^{1-\beta}-K^{1-\beta}} \cdot \left(\frac{\zeta}{x(\zeta)}\right) \cdot \left\{1 + \left[g\alpha^2 + \tfrac{1}{4}\rho\nu\alpha\frac{f^\beta-K^\beta}{f-K} + \frac{2-3\rho^2}{24}\nu^2\right]\varepsilon^2\tau_{ex} + \cdots\right\}, \tag{1.6a}$$

where

$$\zeta = \frac{\nu}{\alpha}\frac{f^{1-\beta}-K^{1-\beta}}{1-\beta}, \qquad x(\zeta) = \log\left(\frac{\sqrt{1-2\rho\zeta+\zeta^2}-\rho+\zeta}{1-\rho}\right), \tag{1.6b}$$

[1] Although this appears inconsistent, it is equivalent to non-dimensionalizing the problem, expanding in the low volatility regime, and then re-writing the answers in terms of the original dimensioned variables.

and

$$g = \frac{(1-\beta)^2}{\left(f^{1-\beta} - K^{1-\beta}\right)^2} \log\left((fK)^{\beta/2} \frac{f^{1-\beta} - K^{1-\beta}}{(1-\beta)(f-K)} \right). \tag{1.6c}$$

Two problems have developed using these explicit implied vol formulas. First, for some low strike, long dated options, the explict implied vol formulas can lead to arbitrageable prices; this is discussed in the next section. Second, the SABR model has a barrier wherever $C(F) = 0$; this is at $F = 0$ in the classical SABR model. There is an $O(\varepsilon)$ thick region next to this boundary, a region in which the original asymptotic analysis does not pertain. In the current, ultra-low rate environment, this region has a substantial influence on pricing, especially with the advent of zero strike options.

In this article we address both problems. We use singular pertubation methods to show that the *reduced density,*

$$Q(T,F)dF = \text{Prob}\left\{ F < \tilde{F}(T) < F + dF \;\middle|\; \tilde{F}(t) = f,\ \tilde{A}(t) = \alpha \right\}, \tag{1.7}$$

satisfies the *effective forward equation*

$$Q^c_T = \tfrac{1}{2}\varepsilon^2\alpha^2 \left[\left(1 + 2\varepsilon\rho\nu z + \varepsilon^2\nu^2 z^2\right) e^{\varepsilon^2\rho\nu a\Gamma(F)(T-t)} C^2(F) Q^c \right]_{FF} \qquad \text{for } T > t, \tag{1.8a}$$

where $z(F)$ and $\Gamma(F)$ are given by

$$z(F) \equiv \frac{1}{\varepsilon\alpha}\int_f^F \frac{df'}{C(f')}, \qquad \Gamma(F) \equiv \frac{C(F) - C(f)}{F - f}. \tag{1.8b}$$

This reduction is not exact, but is accurate through $O(\varepsilon^2)$, the same accuracy as the explicit implied vol formulas. Our approach is to solve the PDE numerically to obtain the probability density $Q(\tau_{ex}, F)$ at the exercise date τ_{ex}; the call and put prices are then obtained by integrating to find the expected value of the payoffs.

To simplify notation, let

$$D(F) = \sqrt{1 + 2\varepsilon\rho\nu z(F) + \varepsilon^2\nu^2 z^2(F)}\, e^{\frac{1}{2}\varepsilon^2\rho\nu a\Gamma(F)(T-t)} C(F). \tag{1.9}$$

Then the effective forward equation is

$$Q^c_T = \tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F) Q^c \right]_{FF} \qquad \text{for } T > t. \tag{1.10a}$$

Note that $D(F) = 0$ where, and only where, $C(F) = 0$.

Solving the problem numerically requires a finite domain, $F_{\min} < F < F_{\max}$. The appropriate lower boundary $F_{\min}$ is usually the barrier, where $C(F_{\min}) = 0$, and thus $D(F_{\min}) = 0$. In some situations it can make sense to use other boundaries. For example, one may wish to assume that the forward $\tilde{F}(T)$ cannot go below zero, in which case the boundary has to be set at $F_{\min} = 0$, regardless of whether $C(0) = 0$. Or the barrier may be more than 5 or 6 standard deviations below the forward, so there is no reason to extend the grid all the way to the barrier. The upper barrier $F_{\max}$ simply needs to be large enough so that there is a negligible probability of $\tilde{F}(\tau_{ex})$ reaching or exceeding $F_{\max}$; usually this is 4 to 6 standard deviations above the forward f. See Appendix D, which gives F in terms of the number of standard deviations above (or below) f.

The appropriate boundary conditions are investigated in Appendix B. There it is found that we must use absorbing boundary conditions,

$$D^2(F)Q^c \to 0 \qquad \text{as } F \to F^+_{\min}, \qquad D^2(F)Q^c \to 0 \qquad \text{as } F \to F^-_{\max}, \tag{1.10b}$$

in order for $\tilde{F}(T)$ to be a Martingale. Clearly the appropriate initial condition is

$$Q^c(T,F) \to \delta(F-f) \qquad \text{as } T \to t. \tag{1.10c}$$

Equations 1.10a - 1.10c form a well posed problem for the density $Q^c(T,F)$. Since the boundaries are absorbing, the probability density will develop δ-functions at the boundaries $F_{\min}$, $F_{\max}$ in addition to the continuous density $Q^c(T,F)$. Crudely speaking,

$$Q(T,F) = \begin{cases} Q^L(T)\delta(F-F_{\min}) & \text{at } F = F_{\min} \\ Q^c(T,F) & \text{for } F_{\min} < F < F_{\max} \\ Q^R(T)\delta(F-F_{\max}) & \text{at } F = F_{\max} \end{cases}. \tag{1.11}$$

The rate at which probability accumulates at $F_{\min}$ and $F_{\max}$ is determined by the flux reaching the barriers from the interior:

$$\frac{dQ^L}{dT} = \lim_{F \to F_{\min}^+} \tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q^c\right]_F, \tag{1.12a}$$

$$\frac{dQ^R}{dT} = -\lim_{F \to F_{\max}^+} \tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q^c\right]_F. \tag{1.12b}$$

We shall see that this ensures that the combined probability in $Q^L(T)$, $Q(T,F)$, and $Q^R(T)$ totals 1. Clealy the initial condition is

$$Q^L(T) \to 0, \qquad Q^R(T) \to 0 \qquad \text{as } T \to t. \tag{1.12c}$$

Of course if the probability $Q^R(T)$ at the upper boundary is significant, $F_{\max}$ should be increased.

At first glance, the delta function at $F_{\min}$ looks unusual. However, consider a situation in which the $C(F)$ has been modified to prevent the forward F from becoming negative, so that the effective volatility

$$\tilde{C}(F) = \begin{cases} C(F) & \text{for } F > \eta \\ C(F_0)F/\eta & \text{for } 0 < F < \eta \end{cases} \tag{1.13}$$

is used in place of $C(F)$. This situation is analyzed in Appendix B. We find that for $\eta \ll 1$, there is a thin boundary layer in $0 < F < \eta$ which has very high densities $Q(T,F)$. In the limit $\eta \to 0$, the total probability in $0 < F < \eta$ is $Q^L(T)$, and yet $C^2(F)Q(T,F)$ goes to zero at we approach η from above:

$$\lim_{F \to \eta^+} C^2(F)Q(T,F) \to 0, \qquad \int_0^\eta Q(T,F)dF \to Q^L(T) \qquad \text{as } \eta \to 0. \tag{1.14}$$

We believe that this is representative of the general situation: when the forward $\tilde{F}(T)$ is near enough to the boundary, other mechanisms come into play; after all, there must be some reason that $\tilde{F}(T)$ doesn't cross the boundary. For example, interest rates in zero, near zero, and even slightly negative rate environments, can be expected to behave differently than rates in more moderate regimes. If we put in a detailed model for these boundary mechanisms, and looked on a fine enough scale, then the delta functons should be resolved into structured boundary layers. Yet the total probability in the layer should match $Q^L(T)$, as this is required by conservation, and as we move away from the boundary, $Q(T,F)$ should transition into the solution of 1.10a with the absorbing boundary condition 1.10b, as this is required for $\tilde{F}(T)$ to be a Martigale.

We use a moment-preserving Crank-Nicholson scheme to solve the PDE 1.10a - 1.10c numerically for $t < T < \tau_{ex}$, while simultaneously integrating the ODE's 1.12a - 1.12c. Once we have obtained

$Q^L(\tau_{ex}), Q^c(\tau_{ex}, F)$, and $Q^R(\tau_{ex})$, we can obtain the option prices for all strikes K by integration:

$$V_{call}(\tau_{ex}, K) = \int_K^{F_{\max}} (F-K)\, Q^c(\tau_{ex}, F) dF + (F_{\max} - K)\, Q^R(\tau_{ex}), \tag{1.15a}$$

$$V_{put}(\tau_{ex}, K) = (K - F_{\min}) Q^L(\tau_{ex}) + \int_{F_{\min}}^{KI} (K-F)\, Q^c(\tau_{ex}, F) dF. \tag{1.15b}$$

If we wish, we can then find the implied normal vol that matches these prices at each K. That is, by solving the PDE once, we obtain the smile for all K at τ_{ex}.

We shall find that the implied volatilities obtained from these numerical solutions closely match the explicit implied volatility formulas, apart from ultra-low strikes. For very low strikes, the implied absolute (normal) vol dips downwards, closely matching market behavior.

We also show show how negative rates can be accomodated by replacing F^β with $(F+a)^\beta$ to move the barrier below zero.

2. Arbitrage using the explicit formulas for the SABR model. The explicit implied vol formulas make the SABR model easy to implement, calibrate, and use. *These implied volatility formulas are usually treated as if they are exactly correct*, even though they are derived from an expansion which requires that $\varepsilon\alpha\sqrt{\tau_{ex}}$, $\varepsilon\nu\sqrt{\tau_{ex}}$ and $|F-K|/\varepsilon\alpha\sqrt{\tau_{ex}}$ be not too large. The unstated argument is that *instead of treating these formulas as an (accurate) approximation to the SABR model, they should be treated as the* exact solution *to some other model which is well approximated by the SABR model.*

This is a valid viewpoint as long as the option prices obtained using the explicit formulas 1.5a-1.5c for $\sigma_N(K)$ are arbitrage free. There are two key requirements for these prices to be arbitrage free [6, Dupire]. The first is call-put parity, which holds automatically since we are using the same implied vol $\sigma_N(K)$ for both calls and puts. The second is that the probability density implied by the call and put prices needs to be positive. To explore this, note that the call and put values can be written quite generally as

$$V_{call}(\tau_{ex}, K) = \int_K^\infty (F-K)\, Q(\tau_{ex}, F) dF, \qquad V_{put}(K) = \int_{-\infty}^K (K-F)\, Q(\tau_{ex}, F) dF, \tag{2.1}$$

where $Q(\tau_{ex}, F)$ is the probability density at the exercise date (including any delta functions). Clearly,

$$\frac{\partial^2}{\partial K^2} V_{call}(K) = \frac{\partial^2}{\partial K^2} V_{put}(K) = Q(\tau_{ex}, F) \geq 0. \tag{2.2}$$

For the explicit implied vol formulas 1.5a-1.5c to represent an arbitrage free model, then, we need

$$\frac{\partial^2}{\partial K^2} V^N_{call}(f, K, \sigma_N(K), \tau_{ex}) = \frac{\partial^2}{\partial K^2} V^N_{put}(f, K, \sigma_N(K), \tau_{ex}) \geq 0 \qquad \text{for all } K. \tag{2.3}$$

That is, there cannot be a "butterfly arbitrage."

It is not terribly uncommon for this requirement to be violated for very low strike options for sufficiently large τ_{ex}. The problem does not appear to be the quality of the call and put prices obtained from the explicit implied vol formulas, because these usually remain quite accurate. Rather, the problem seems to be that implied volatility curves are not a stable representation of option prices for low strike options. It is very easy to find nearly identical, reasonable looking, volatility curves $\sigma_N(K)$, for which some of the curves are arbitrage free and others violate the arbitrage-free constraint of eq. 2.3.

This is illustrated in Figure 2.1. There the smile $\sigma_N(K)$ obtained from the explicit formula 1.5a-1.5c is graphed (*explicit)* along with a very similar smile (*arb free*) obtained from the arbitrage free procedure. These two smiles look very similar, and lead to nearly identical option prices, as shown in Figure 2.2. Differentiating

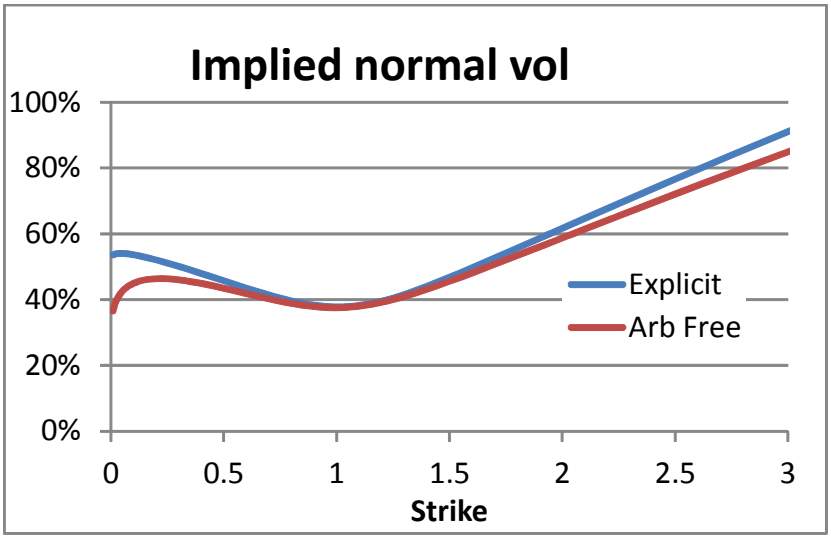

FIG. 2.1. *The implied normal vol for the SABR model for* $\alpha = 35\%$, $\beta = 0.25$, $\rho = -10\%$, *and* $\nu = 100\%$. *Shown are* $\sigma_N(K)$ *from the explicit formula and from the arb-free approach for* $\tau_{ex} = 1yr$.

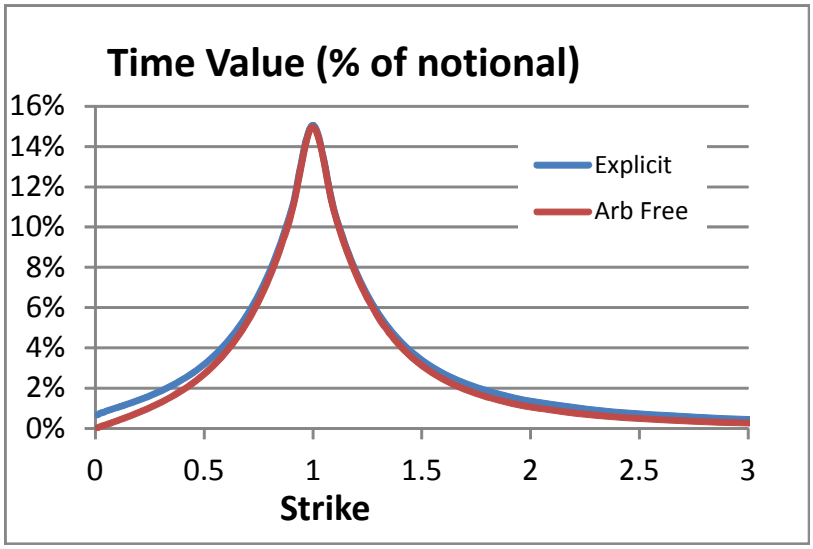

FIG. 2.2. *Call and put values from the SABR model for* $\alpha = 35\%$, $\beta = 0.25$, $\rho = -10\%$, $\nu = 100\%$. *Shown are the put prices (for* $K < f$*) and call prices (for* $K > f$*) obtained from the explicit formulas and the arb-free approach for* $\tau_{ex} = 1yr$.

these option prices with respect to the strike K yields the probability densities shown in Figure 2.3. The "explicit" smile $\sigma_N(K)$ leads to negative probabilities for ultra low strikes, and so is not arbitrage free, whilst the second curve has only positive probabilities, and is arbitrage free.

Using implied Black (log normal) vols $\sigma_B(K)$ instead of implied normal vols does not help. Figure 2.4 compares the implied Black vols from the explict formulas for $\sigma_N(K)$ with those from the arb free approach. There is no obvious way to discern that one curve leads to arbitrage free prices, while the other does not.

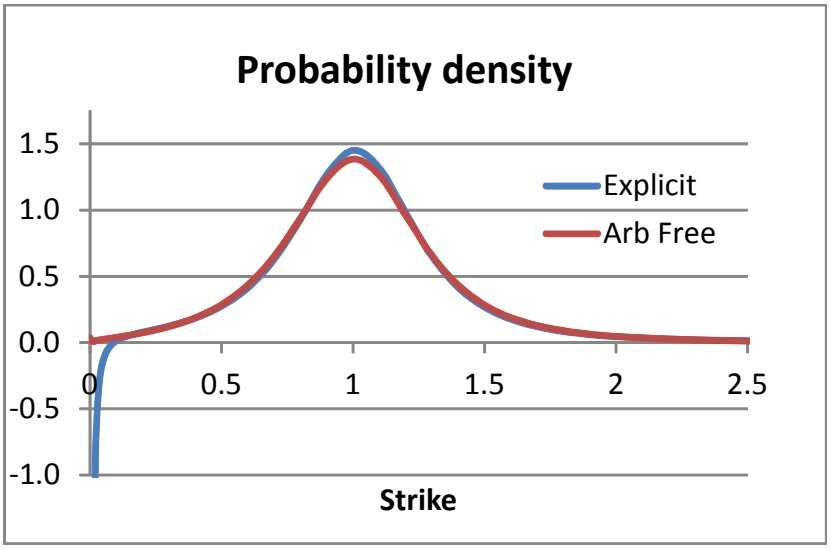

FIG. 2.3. *Probaiity density for the SABR model for* $\alpha = 35\%$, $\beta = 0.25$, $\rho = -10\%$, *and* $\nu = 100\%$. *Shown are the densities obtained from the explicit formulas for* $\sigma_N(K)$ *and from the arb free approach for* $\tau_{ex} = 1yr$.

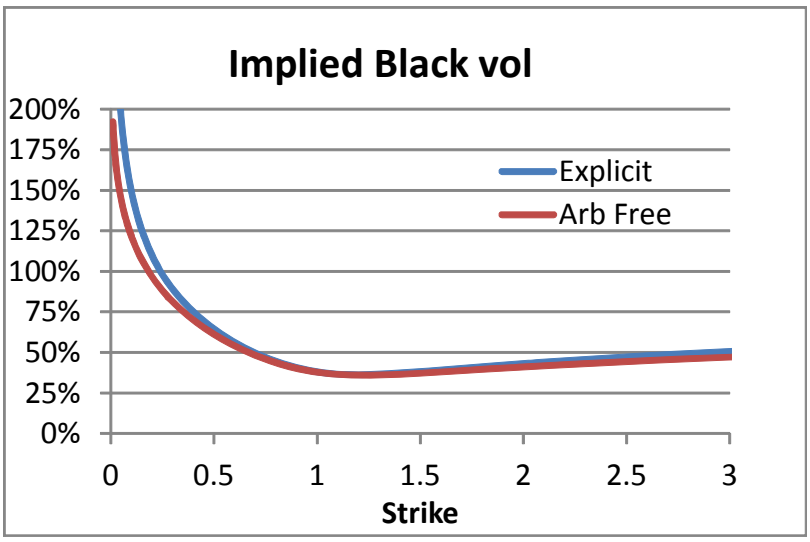

FIG. 2.4. *The implied log normal volatility from the SABR model at* $\tau_{ex} = 1yr$ *for* $\alpha = 35\%$, $\beta = 25\%$, $\rho = -10\%$, $\nu = 100\%$. *Shown are the implied volatilities from the option prices obtained from the explicit normal curve* $\sigma_N(K)$ *and the arb free curve.*

3. Arbitrage free pricing.

3.1. The effective forward equation. Here we present an alternative pricing approach which is arbitrage free and retains the $O(\varepsilon^2)$ accuracy of the original SABR analysis. We believe that variations of this approach have been used by other firms [7, Anderson, 8, the French team].

Consider the probability density that $\tilde{F}(T) = F$ and $\tilde{A}(T) = A$ at time T, given that we start at $\tilde{F}(t) =$

f and $\tilde{A}(t) = \alpha$ at time t:

(3.1a)

$$p(t, f, \alpha; T, F, A)dFdA = \text{Prob}\left\{ F < \tilde{F}(T) < F + dF,\ A < \tilde{A}(T) < A + dA \;\middle|\; \tilde{F}(t) = f,\ \tilde{A}(t) = \alpha \right\}.$$

This density p satisfies the Fokker-Planck equation (the forwards Kolmogorov equation),

$$p_T = \tfrac{1}{2}\varepsilon^2 A^2 \left[C^2(F)p\right]_{FF} + \varepsilon^2\rho\nu \left[A^2 C(F)p\right]_{FA} + \tfrac{1}{2}\varepsilon^2\nu^2 \left[A^2 p\right]_{AA} \qquad \text{for } T > t. \tag{3.1b}$$

In Appendix A we define the reduced (marginal) probability density,

$$Q(T, F) = \int_0^\infty p(t, f, \alpha; T, F, A)dA, \tag{3.2}$$

which is the probability density that $\tilde{F}(T) = F$ at time T, regardless of the value of $\tilde{A}(T)$. Although Q is also a function of the backwards variables t, f, α, for clarity we have omitted explicitly showing this dependence.

In appendix A we use singular perturbation techniques to analyze the Fokker-Planck equation 3.1b. This analysis shows that the marginal density satisfies the PDE

$$Q^c_T = \tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q^c\right]_{FF} \qquad \text{for } T > t, \tag{3.3a}$$

through $O(\varepsilon^2)$, where

$$D^2(F) = \left[1 + 2\varepsilon\rho\nu z(F) + \varepsilon^2\nu^2 z^2(F)\right] e^{\varepsilon^2\rho\nu a\Gamma(F)(T-t)} C^2(F), \tag{3.3b}$$

and where$z(F)$ and $\Gamma(F)$ are defined by

$$z(F) \equiv \frac{1}{\varepsilon\alpha}\int_f^F \frac{df'}{C(f')}, \qquad \Gamma(F) \equiv \frac{C(F) - C(f)}{F - f}. \tag{3.3c}$$

This effective forward equation reduces the dimensionality from two space dimensions (F and A) to one space dimension (F only), while retaining the same order of accuracy as the original SABR analysis [1]. Note that we have added a c superscript to denote that this is the continuous part of the density.

For the special case of $C(F) = F^\beta$, we have

$$z(F) \equiv \frac{F^{1-\beta} - f^{1-\beta}}{\varepsilon\alpha\,(1-\beta)}, \qquad \Gamma(F) \equiv \frac{F^\beta - f^\beta}{F - f}. \tag{3.4}$$

3.2. Boundary conditions. The SABR model has an innate barrier where $C(F) = 0$, or, equivalently, where $D(F) = 0$. Traditionally this barrier is at $F = 0$, as it is for $C(F) = F^\beta$, but currently there is an active debate over whether rates can be negative, and if so, how negative they can become. Appendix E explores alternative models for $C(F)$, such as $(F + b)^\beta$, which puts the barrier at $-b$.

We solve the PDE numerically, which requires a finite domain $F_{\min} < F < F_{\max}$. It is natural to place the lower boundary at the barrier, but not essential, and there may well be situations in which a different boundary makes more sense. So we do not assume that $F_{\min}$ is necessarily at the barrier. $F_{\max}$ should be chosen to be large enough so that the boundary doesn't affect the pricing appreciably.

Boundary conditions are examined in Appendix B. Since there may be a net probablity current at the boundaries, and we are not considering models in which the forward can leave the domain, we need to allow for probability accumulating at the boundaries. I.e., we need to allow δ-functions in the probabiliy density at $F_{\min}$ and $F_{\max}$:

$$Q(T, F) = \begin{cases} Q^L(T)\delta(F - F_{\min}) & \text{at } F = F_{\min} \\ Q^c(T, F) & \text{for } F_{\min} < F < F_{\max} \\ Q^R(T)\delta(F - F_{\max}) & \text{at } F = F_{\max} \end{cases}. \tag{3.5}$$

Here the superscript c is being used to denote the continuous part of the density.

The total probability has to be 1,

$$Q^L(T) + \int_{F_{\min}}^{F_{\max}} Q^c(T,F)dF + Q^R(T) = 1. \tag{3.6}$$

for all T, so

$$\frac{d}{dT}\left\{ Q^L(T) + \int_{F_{\min}}^{F_{\max}} Q^c(T,F)dF + Q^R(T) \right\} = 0. \tag{3.7}$$

Substituting 3.3a for Q_T^c and integrating leads to

$$\frac{dQ^L}{dT} + \tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q^c\right]_F\Big|_{F_{\min}}^{F_{\max}} + \frac{dQ^R(T)}{dT} = 0. \tag{3.8}$$

Thus, conservation of probability requires that

$$\frac{dQ^L}{dT} = \lim_{F\to F_{\min}^+} \tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q^c\right]_F, \tag{3.9a}$$

$$\frac{dQ^R}{dT} = \lim_{F\to F_{\max}^-} -\tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q^c\right]_F. \tag{3.9b}$$

Similarly, for $\tilde{F}(T)$ to be a Martingale, we need its expected value to be fixed:

$$E\left\{ \tilde{F}(T)\Big|\ \tilde{F}(t) = f,\ \tilde{A}(t) = \alpha \right\} = F_{\min}Q_L(T) + \int_{F_{\min}}^{\infty} FQ^c(T,F)dF + F_{\max}Q^R(T) = f. \tag{3.10}$$

Therefore,

$$F_{\min}\frac{dQ_L(T)}{dT} + \int_{F_{\min}}^{F_{\max}} FQ_T^c(T,F)dF + F_{\max}\frac{dQ^R(T)}{dT} = 0. \tag{3.11}$$

Substituting 3.3a for Q_T^c and integrating by parts twice and using equations 3.9a, 3.9b leads to

$$D^2(F)Q^c\big|_{F_{\min}}^{F_{\max}} = 0. \tag{3.12}$$

So we must require absorbing boundary conditions,

$$D^2(F)Q^c \to 0 \qquad \text{as } F \to F_{\min}^+, \tag{3.13a}$$

$$D^2(F)Q^c \to 0 \qquad \text{as } F \to F_{\max}^-, \tag{3.13b}$$

to ensure that the $\tilde{F}(T)$ to be a Martingale.

In summary, we solve

$$Q_T^c = \tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q^c\right]_{FF} \qquad \text{for } F_{\min} < F < F_{\max}, \tag{3.14a}$$

with the boundary condition

$$D^2(F)Q^c \to 0 \qquad \text{as } F \to F_{\min}^+, \qquad D^2(F)Q^c \to 0 \qquad \text{as } F \to F_{\max}^-, \tag{3.14b}$$

for $t < T < \tau_{ex}$. The probabilities at the boundary are given by

$$\frac{dQ^L}{dT} = \lim_{F \to F^+_{\min}} \tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q^c\right]_F \tag{3.14c}$$

$$\frac{dQ^R}{dT} = \lim_{F \to F^-_{\max}} -\tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q^c\right]_F \tag{3.14d}$$

and the initial conditions are

$$Q^L(0) = 0, \qquad Q^c(T,F) \to \delta(F-f) \qquad Q^R(0) = 0, \qquad \text{as } T \to t^+. \tag{3.14e}$$

3.3. Option pricing. In Appendix C we sketch out a Crank-Nicholson scheme [NumRecInC] for solving 3.14a - 3.14e for $Q^c(T,F)$, $Q^L(T)$, and $Q^R(T)$. This scheme conserves probability and the expected value of $\tilde{F}(T)$, so equations 3.6 and 3.10 remain exactly true for the numerical solution. Once we have solved these equations numerically, the call and put prices are obtained by integrating:

$$V_{call}(\tau_{ex},K) = f - K \qquad \text{for } K < F_{\min} \tag{3.15a}$$

$$V_{call}(\tau_{ex},K) = \int_K^{F_{\max}} (F-K)Q^c(\tau_{ex},F)dF + (F_{\max}-K)\,Q^R(\tau_{ex}) \quad \text{for } F_{\min} < K < F_{\max} \tag{3.15b}$$

$$V_{call}(\tau_{ex},K) = 0 \qquad \text{for } K > F_{\max} \tag{3.15c}$$

$$V_{put}(\tau_{ex},K) = 0 \qquad \text{for } K < F_{\min} \tag{3.15d}$$

$$V_{put}(\tau_{ex},K) = \int_{F_{\min}}^K (K-F)Q^c(\tau_{ex},F)dF + (K-F_{\min})\,Q^L(\tau_{ex}) \quad \text{for } F_{\min} < K < F_{\max} \tag{3.15e}$$

$$V_{put}(\tau_{ex},K) = K - f \qquad \text{for } K > F_{\max} \tag{3.15f}$$

The conservation of probablity 3.6 and Martingale proprerty 3.10 show that call-put parity holds exactly for the numerical solution:

$$V_{call}(\tau_{ex},K) - V_{put}(\tau_{ex},K) = f - K. \tag{3.16}$$

Since the effective forward equation has only one space dimension, solving the PDE is essentially instantaneous. Moreover, solving these equations for $t < T < \tau_{ex}$ yields the option prices for all strikes K. Thus the implied normal vols $\sigma_N(K)$ for all strikes K can be obtained by solving the PDE once.

The maximum principle for parabolic equations [Protter & Weinberg] guarantees that $Q^c(T,F) \geq 0$ for $F_{\min} < F < F_{\max}$, and that $Q^L(T)$ and $Q^R(T)$ are increasing, and hence positive. For fine enough grids, the numerical solutions will also be non-negative. Since call-put parity is also satisfied, the numerical option prices are arbitrage free [6, Dupire].

4. Discussion.

4.1. Results. Singular perturbation techniques can be used to analyze the effective forward equation,

$$Q^c_T = \tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q^c\right]_{FF} \qquad \text{for } T > t, \tag{4.1a}$$

where

$$D(F) = \left(1 + 2\varepsilon\rho\nu z + \varepsilon^2\nu^2 z^2\right) e^{\varepsilon^2\rho\nu a\Gamma(F)(T-t)} C^2(F), \tag{4.1b}$$

with

$$z(F) \equiv \frac{1}{\varepsilon\alpha}\int_f^F \frac{df'}{C(f')}, \qquad \Gamma(F) \equiv \frac{C(F)-C(f)}{F-f}. \tag{4.1c}$$

From this analysis we can obtain explicit option prices, and these prices can be used to find explicit formulas for the implied volatility $\sigma_N(K)$. Away from the boundaries $F_{\min}$ and $F_{\max}$, this analysis would lead to the same explicit implied vol formulas 1.5a-1.5c as before, at least if we continue to work through $O(\varepsilon^2)$. Thus, away from the boundaries, the "arbitrage free" implied volatility (obtained by numerically solving the effective forward equation) should match the explicit implied volatility formulas to within $O(\varepsilon^2)$. This is indeed our experience. Even in relatively extreme cases, such as Figure 2.1, the "arbitrage free" and explicit implied volatility smiles match closely, except when the strike K or the forward f gets too close to the boundary at $F_{\min}$.

For values of F within $O(\varepsilon)$ of the lower boundary $F_{\min}$, a non-negligible percentage of paths that would have reached F hit the boundary. This creates an $O(\varepsilon)$ thick boundary layer at $F_{\min}$, in which the explicit implied volatity formulas do not pertain. Instead, as the strike approaches $F_{\min}$, the implied normal volatility $\sigma_N(K)$ bends towards zero. See Figure 2.1

Figure 4.1 shows the effect of the boundary layer on the normal volatility of *at-the-money* options. As today's forward f decreases to within $O(\varepsilon)$ of $F_{\min}$, an increasing percentage of the paths reach the boundary before the expiry date, which reduces the ATM volatility. Seeing the "knee" in this graph, one might naively believe that the market switches from a normal regime to a log normal regime when the forward is sufficiently small. This is an illusion; this graph comes from the SABR model with $\beta = 0$ and $\rho = 0$, which is a stochastic *normal* model. The reduction is caused solely by the boundary.

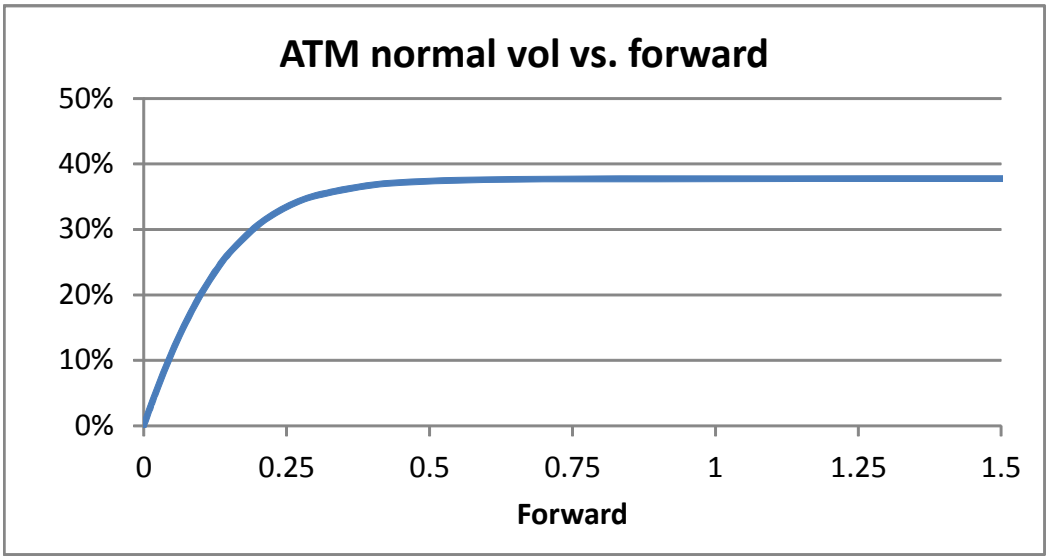

FIG. 4.1. *The at-the-money implied normal vol* $\sigma_N(K)$ *at* $\tau_{ex} = 1yr$ *for the arbitrage free SABR model with* $\alpha = 35\%$, $\beta = 0$, $\rho = 0$,*and* $\nu = 100\%$.

Below are the smiles $\sigma_N(K)$ obtained for different values of the forward f, using the same SABR parameters in each case.

Market data for at-the-money swaption volatilities exhibit this "knee": Historical studies comparing ATM normal vols with forward rates show that when the forward rate is above a critical value, the ATM

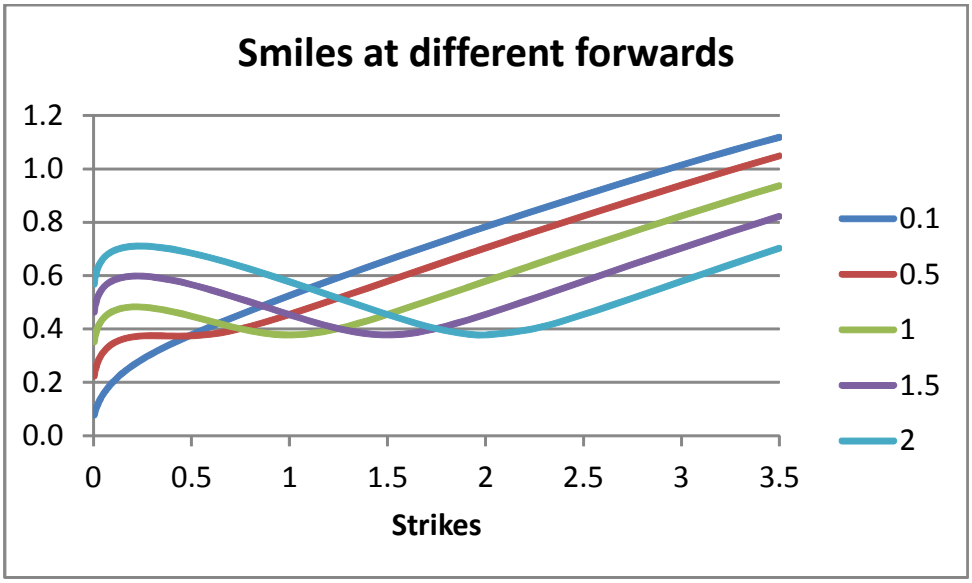

FIG. 4.2. *The smiles $\sigma_N(K)$ at $\tau_{ex} = 1yr$ for different values of the forward f for the SABR model with $\alpha = 35\%$, $\beta = 0$, $\rho = 0\%$, and $\nu = 100\%$.*

normal vols are reasonably constant; for forward rates below the critical value, the ATM vols decrease linearly with the rate. See figure **??**.

Figure caption: The normal vols for 1Y into 1Y at the money swaptions vs the forward swap rate. Shown is the historic data for 2002 through 2012 for USD, GBP, EUR, and JPY swaptions. Also shown is the implied volatility obtained from the SABR model for $\alpha = 65\%$, $\beta = 0.25$, $\rho = 0\%$, and $\nu = 75\%$.

Typically this knee is explained as the market switching from normal to log normal behaviour in ultra-low rate environments. This belief is often reinforced by using the explicit implied vol formulas to calibrate the SABR model. Calibrating the explicit implied vol formulas to observed smiles can lead to relatively high values of β and/or ρ for low forward rates. Since high values of β and ρ push up the high strike vols, this can create significant mis-pricing for instruments which are sensitive to *high* strikes, like constant maturity caps, floors, and swaps. To counter this, some firms have chosen to use

$$C(F) = \begin{cases} F & \text{for } F < F_0 \\ F_0 & \text{for } F > F_0 \end{cases} \tag{4.2}$$

in place of F^β. We believe that our approach provides a more natural explanation, since the knee occurs automatically, without requiring gross changes between the SABR parameters for low and moderate rate environments.

4.2. Hedging. The coefficients in the effective forward equation 4.1a - 4.1c depend on the current forward f as well as F. This means that
· there is no obvious "effective backwards equation" equivalent to the effective forward equation;
· the effective forward equation is *not* the Fokker-Planck equation (forward Kolmogorov equation) for some one dimensional Ito process.
I.e., there is *not* an effective one dimensional local volatility model corresponding to the effective forward equation. This should have been anticipated as the SABR model was created because of perceived shortcomings of local volatility models[1, SABR].

When the SABR model is calibrated to market data, it is very difficult to distinguish between β and ρ; both control the skew. If we fix β, say, and calibrate the rest of the parameters, the quality of the fit is usually pretty much independent of the particular value of β chosen. This is illustrated in Figure 4.3. There we have chosen $\beta = 0$, $\beta = \frac{1}{2}$, and $\beta = 1$, and calibrated the SABR model to the same market data for all three cases, which yields the following set of SABR parameters.

α	31.8%	32.9%	35.1%
β	0	0.5	1
ρ	−18.3%	−45.5%	−64.4%
ν	0.777	0.867	0.985

Although the tail of the smiles are somewhat different, all three sets of parameters seem to fit the actual market data.

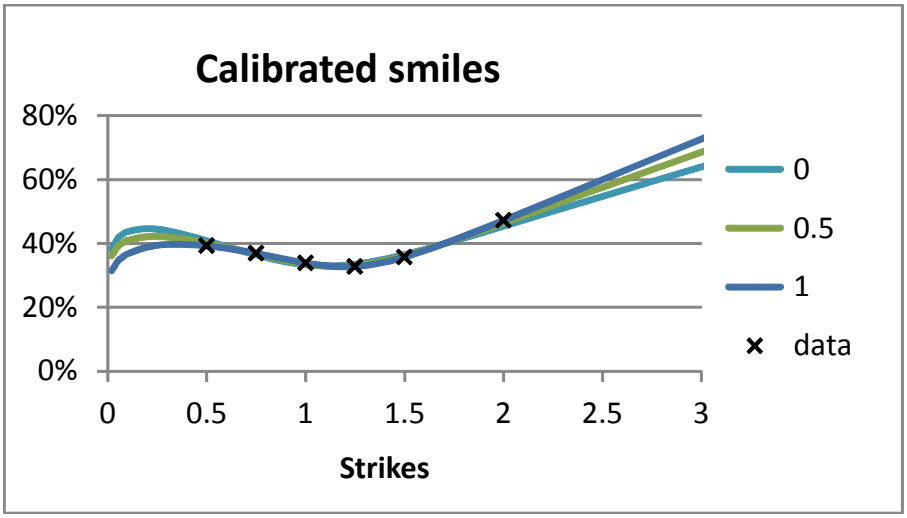

FIG. 4.3. *The SABR model calibrated to the same market data for $\beta = 0$, $\beta = \frac{1}{2}$, and $\beta = 1$. Because ρ can largely compensate for β, all three fits are well within market noise*

The conventional delta risk is calculated by shifting the current forward f and keeping the current volatility α fixed:

$$f \to f + \Delta f, \qquad \alpha \to \alpha. \tag{4.3}$$

If the delta risk is calculated in the conventional way, then the delta depends on the particular value of β used. This is shown in Figure 4.4. There we have calculated the conventional delta of a call option as a function of the strike K for the same three sets of SABR parameters. Even though all three sets lead to essential the same smile (especially for strikes which are not too extreme), we see that the different choices of β have led to different values of delta. This means that if we hedge our positions using the conventional delta, choosing a poor β may lead to a poor hedge, even though it may lead to a superb fit of the market data.

This issue led to an alternative approach for calculating delta hedges. Since $\tilde{F}$ and $\tilde{A}$ are correlated, when $\tilde{F}$ changes then $\tilde{A}$ changes as well, at least on average. It is argued that accounting for this shift should result in a better hedge [Bartlett]:

$$f \to f + \Delta f, \qquad \alpha \to \alpha + \Delta\alpha. \tag{4.4}$$

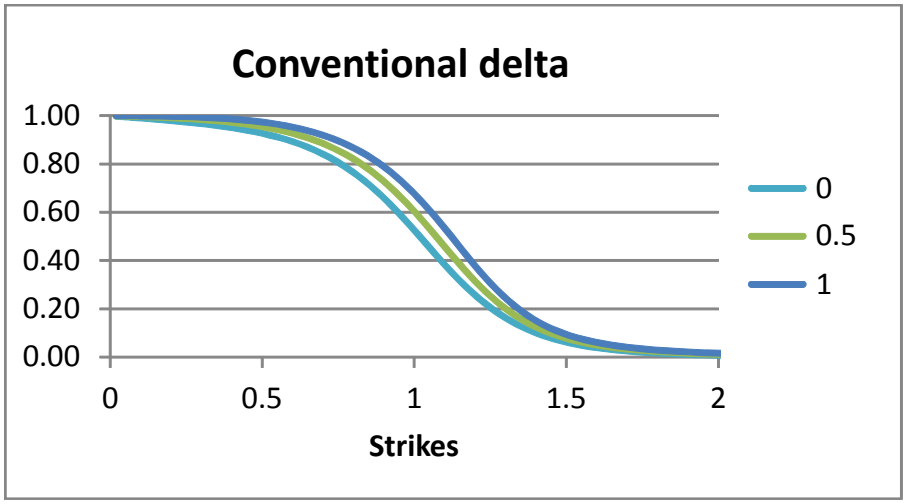

FIG. 4.4. *Conventional delta, $\partial V/\partial F$, as a function of the strike K for $\beta = 0$, $\beta = \frac{1}{2}$, and $\beta = 1$. For each β, the other three parameters α, ρ, and ν are chosen to match the arb free SABR smile to the market data.*

To compute the amount of the α shift, we re-write the SABR model as

$$d\tilde{F} = \tilde{A}C(\tilde{F})d\tilde{W}_1, \tag{4.5a}$$

$$d\tilde{A} = \nu\tilde{A}\left\{\rho d\tilde{W}_1 + \sqrt{1-\rho^2}d\tilde{Z}\right\}, \tag{4.5b}$$

where $d\tilde{W}_1$ and $d\tilde{Z}$ are independent. This implies that

$$d\tilde{A} = \frac{\rho\nu}{C(\tilde{F})}d\tilde{F} + \sqrt{1-\rho^2}\nu\tilde{A}d\tilde{Z} \tag{4.6}$$

Therefore, changes in $\tilde{A}(T)$ can be split into two independent components, one caused by the changes in $\tilde{F}(T)$, and one due to the idiosyncratic changes in the volatility $\tilde{A}(T)$. Accordingly, the delta hedge should be calculated with respect to the scenario [Bartlett]

$$f \to f + \Delta f, \qquad \alpha \to \alpha + \frac{\rho\nu}{C(f)}\Delta f. \tag{4.7}$$

Figure 4.5 shows this alternative delta as a function of the strike K for the same three sets of SABR parameters used above. We calculated these deltas by simply bumping the f and α values input into the pricing code according to 4.7. We see that this alternative delta is nearly independent of the particular value of β chosen. This has proven true for all cases we have investigated: as long as the SABR model fits market data decently, the alternative delta is nearly independent of the particular values of β or ρ used in the fitting. Apparently this new delta depends mostly on the actual market smile, and not how the smiles and skews are represented in the model.

It is generally accepted that hedges based on the alternative delta are superior to the conventional delta hedge when a desk is only using delta to hedge. When a desk is hedging both delta and vega, which delta hedge is used is irrelevent provided one doesn't double count when putting on the vega hedge.

Appendix A. Derivation of the effective forward equation.

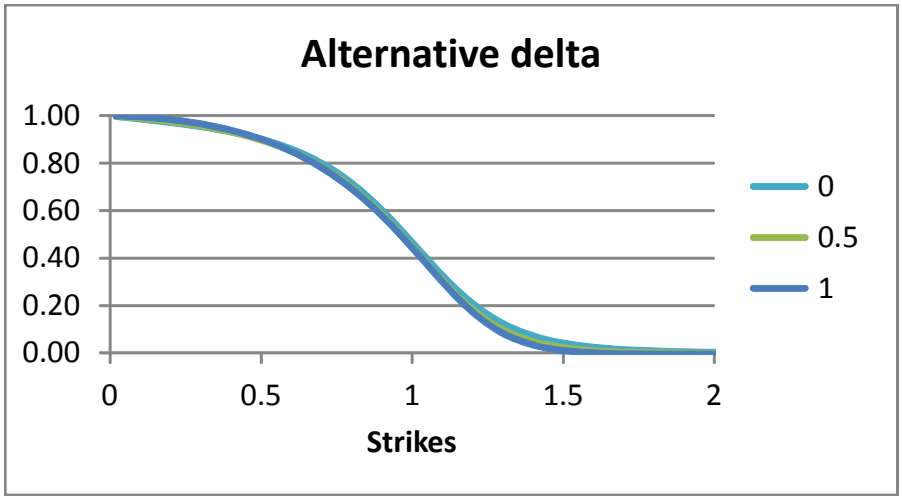

FIG. 4.5. *Alternative delta, $\partial V/\partial f + [\rho\nu/C(f)]\,\partial V/\partial\alpha$ for $\beta = 0$, $\beta = \frac{1}{2}$, and $\beta = 1$. For each β, the other three parameters α, ρ, and ν are chosen to match the arb free SABR smile to the market data.*

Here we analyze the SABR model

$$d\tilde{F} = \varepsilon \tilde{A} C(\tilde{F}) d\tilde{W}_1, \tag{A.1a}$$

$$d\tilde{A} = \varepsilon \nu \tilde{A} d\tilde{W}_2, \tag{A.1b}$$

$$d\tilde{W}_1 d\tilde{W}_2 = \rho dt, \tag{A.1c}$$

in the limit $\varepsilon \ll 1$, using singular perturbation methods to derive the effective forward equation.

Recall that $p(t, f, \alpha; T, F, A)$ is the probability density that $\tilde{F}(T) = F$ and $\tilde{A}(T) = A$ at time T, given that $\tilde{F}(t) = f$ and $\tilde{A}(t) = \alpha$ at time t. Define the moments

$$\begin{aligned} Q^{(k)}(t, f, \alpha; T, F) &= \int_0^\infty A^k p(t, f, \alpha; T, F, A) dA \\ &= E\left\{ A^k \delta(\tilde{F}(T) - F) \;\middle|\; \tilde{F}(t) = f,\ \tilde{A}(t) = \alpha \right\}. \end{aligned} \tag{A.2a}$$

Clearly the zeroeth moment $Q^{(0)}$ is the probability density of being at F at time T,

$$Q(T, F) = Q^{(0)}(t, f, \alpha; T, F), \tag{A.2b}$$

regardless of the value of $\tilde{A}(T)$.

The density p satisfies the Fokker-Planck equation

$$p_T = \tfrac{1}{2}\varepsilon^2 \left[C^2(F)A^2 p\right]_{FF} + \varepsilon^2 \rho\nu \left[C(F)A^2 p\right]_{FA} + \tfrac{1}{2}\varepsilon^2 \nu^2 \left[A^2 p\right]_{AA} \qquad \text{for all } T > t, \tag{A.3a}$$

with the initial condition

$$p = \delta(F - f)\delta(A - \alpha) \qquad \text{for all } T \to t^+. \tag{A.3b}$$

Now

$$\int_0^\infty \left[C(F)A^2 p\right]_{FA} dA = \left[C(F)A^2 p\right]_F \Big|_{A=0}^{A=+\infty} = 0, \tag{A.4a}$$

$$\int_0^\infty \left[A^2 p\right]_{AA} dA = \left[A^2 p\right]_A \Big|_{A=0}^{A=+\infty} = 0, \tag{A.4b}$$

for all F. This just states that there is no probability flux across the boundaries at $A = 0$ and $A = \infty$; i.e., that probability is conserved. Integrating the Fokker-Planck equation across all A now yields

$$Q_T^{(0)} = \tfrac{1}{2}\varepsilon^2 \left[C^2(F)Q^{(2)}\right]_{FF} \qquad \text{for } T > t. \tag{A.5}$$

I.e., the evolution of the reduced density $Q^{(0)}$ depends on the second moment $Q^{(2)}$.

Under the SABR model these moments satisfy the backward Kolmogorov equation

$$Q_t^{(k)} + \tfrac{1}{2}\varepsilon^2\alpha^2 C^2(f)Q_{ff}^{(k)} + \varepsilon^2\rho\nu\alpha^2 C(f)Q_{f\alpha}^{(k)} + \tfrac{1}{2}\varepsilon^2\nu^2\alpha^2 Q_{\alpha\alpha}^{(k)} = 0 \qquad \text{for } t < T, \tag{A.6a}$$

subject to the condition

$$Q^{(k)}(t, f, \alpha; T, F) \to \alpha^k \delta(F - f) \qquad \text{as } t \to T^-. \tag{A.6b}$$

We will successively transform this equation order-by-order until all the α derivatives are negligibly small. This effectively reduces the problem from two dimensions (f and α) to one dimension (f only). Instead of constructing explicit asymptotic solutions to the resulting one dimensional problem, as was done in the original SABR paper [1, SABR], here we seek to write $Q^{(2)}$ in terms of $Q^{(0)}$. This provides the "constituitive law" needed to close the "conservation law" A.5, which is then the effective forward equation. Throughout we work through $O(\varepsilon^2)$, neglecting higher order terms.

Since the backwards equation is autonomous, the moments $Q^{(k)}$ only depend on

$$\tau = T - t, \tag{A.7}$$

and not on T or t seperately. We first change independent variables from f to

$$z = \frac{1}{\varepsilon\alpha}\int_f^F \frac{df'}{C(f')}. \tag{A.8a}$$

For clarity, we also introduce

$$B(\varepsilon\alpha z) = C(f). \tag{A.8b}$$

Then

$$\frac{\partial}{\partial f} \longrightarrow \frac{-1}{\varepsilon\alpha C(f)}\frac{\partial}{\partial z} = \frac{-1}{\varepsilon\alpha B(\varepsilon\alpha z)}\frac{\partial}{\partial z}, \qquad \frac{\partial}{\partial \alpha} \longrightarrow \frac{\partial}{\partial \alpha} - \frac{z}{\alpha}\frac{\partial}{\partial z} \tag{A.9a}$$

$$\frac{\partial^2}{\partial f^2} \longrightarrow \frac{1}{\varepsilon^2\alpha^2 B^2(\varepsilon\alpha z)}\left\{\frac{\partial^2}{\partial z^2} - \varepsilon\alpha\frac{B'(\varepsilon\alpha z)}{B(\varepsilon\alpha z)}\frac{\partial}{\partial z}\right\}, \tag{A.9b}$$

$$\frac{\partial^2}{\partial f \partial \alpha} \longrightarrow \frac{1}{\varepsilon\alpha B(\varepsilon\alpha z)}\left\{-\frac{\partial^2}{\partial z \partial \alpha} + \frac{z}{\alpha}\frac{\partial^2}{\partial z^2} + \frac{1}{\alpha}\frac{\partial}{\partial z}\right\}, \tag{A.9c}$$

$$\frac{\partial^2}{\partial \alpha^2} \longrightarrow \frac{\partial^2}{\partial \alpha^2} - \frac{2z}{\alpha}\frac{\partial^2}{\partial z \partial \alpha} + \frac{z^2}{\alpha^2}\frac{\partial^2}{\partial z^2} + \frac{2z}{\alpha^2}\frac{\partial}{\partial z}, \tag{A.9d}$$

and

$$\delta(f-F)=\delta(\varepsilon\alpha z C(F))=\frac{1}{\varepsilon\alpha B(0)}\delta(z). \tag{A.9e}$$

The backwards equation now becomes

$$\begin{aligned} Q^{(k)}_\tau = {} & \tfrac{1}{2}\left(1+2\varepsilon\rho\nu z+\varepsilon^2\nu^2z^2\right)Q^{(k)}_{zz}-\tfrac{1}{2}\varepsilon\alpha\frac{B'(\varepsilon\alpha z)}{B(\varepsilon\alpha z)}Q^{(k)}_z \\ & +\left(\varepsilon\rho\nu+\varepsilon^2\nu^2z\right)\left(-\alpha Q^{(k)}_{\alpha z}+Q^{(k)}_z\right)+\tfrac{1}{2}\varepsilon^2\nu^2\alpha^2Q^{(k)}_{\alpha\alpha} \qquad \text{for } \tau>0, \end{aligned} \tag{A.10a}$$

with the initial condition

$$Q^{(k)}(\tau,z,\alpha)\to\frac{\alpha^{k-1}}{\varepsilon B(0)}\delta(z) \qquad \text{as } \tau\to 0^+. \tag{A.10b}$$

Accordingly, we define $\hat{Q}^{(k)}(\tau,z,\alpha)$ by

$$Q^{(k)}(\tau,z,\alpha)=\frac{\alpha^{k-1}}{\varepsilon B(0)}\hat{Q}^{(k)}(\tau,z,\alpha). \tag{A.11}$$

Then $\hat{Q}^{(k)}(\tau,z,\alpha)$ satisfies

$$\begin{aligned} \hat{Q}^{(k)}_\tau = {} & \tfrac{1}{2}\left(1+2\varepsilon\rho\nu z+\varepsilon^2\nu^2z^2\right)\hat{Q}^{(k)}_{zz}-\tfrac{1}{2}\varepsilon\alpha\frac{B'(\varepsilon\alpha z)}{B(\varepsilon\alpha z)}\hat{Q}^{(k)}_z \\ & -\left(\varepsilon\rho\nu+\varepsilon^2\nu^2z\right)(k-2)\,\hat{Q}^{(k)}_z-\left(\varepsilon\rho\nu+\varepsilon^2\nu^2z\right)\alpha\hat{Q}^{(k)}_{\alpha z} \\ & +\tfrac{1}{2}\varepsilon^2\nu^2\left\{\alpha^2\hat{Q}^{(k)}_{\alpha\alpha}+2(k-1)\alpha\hat{Q}^{(k)}_\alpha+(k-1)(k-2)\hat{Q}^{(k)}\right\} \qquad \text{for } \tau>0, \end{aligned} \tag{A.12a}$$

with

$$\hat{Q}^{(k)}(\tau,z,\alpha)\to\delta(z) \qquad \text{as } \tau\to 0^+. \tag{A.12b}$$

To leading order in $O(\varepsilon)$,the equation and initial condition for $\hat{Q}^{(k)}(\tau,z,\alpha)$ are

$$\hat{Q}^{(k)}_\tau = \tfrac{1}{2}\hat{Q}^{(k)}_{zz} \tag{A.13a}$$

$$\hat{Q}^{(k)}(\tau,z,\alpha)\to\delta(z) \qquad \text{as } \tau\to 0^+, \tag{A.13b}$$

which are independent of α to leading order. Therefore, if we *were* to expand

$$\hat{Q}^{(k)}(\tau,z,\alpha)=\hat{Q}^{(k)}_0(\tau,z)+\varepsilon\hat{Q}^{(k)}_1(\tau,z,\alpha)+\varepsilon^2\hat{Q}^{(k)}_2(\tau,z,\alpha)+\cdots, \tag{A.14}$$

then the leading term $\hat{Q}^{(k)}_0(\tau,z)$ would not depend on α, as indicated. Consequently, the terms $\varepsilon^2\nu^2z\alpha\hat{Q}^{(k)}_{\alpha z}$, $\varepsilon^2\nu^2\alpha^2\hat{Q}^{(k)}_{\alpha\alpha}$ and $\varepsilon^2\nu^2\alpha\hat{Q}^{(k)}_\alpha$ are actually no larger than $O(\varepsilon^3)$. Since we are only working through $O(\varepsilon^2)$, we drop these terms, obtaining

$$\begin{aligned} \hat{Q}^{(k)}_\tau = {} & \tfrac{1}{2}\left(1+2\varepsilon\rho\nu z+\varepsilon^2\nu^2z^2\right)\hat{Q}^{(k)}_{zz}-\tfrac{1}{2}\varepsilon\alpha\frac{B'(\varepsilon\alpha z)}{B(\varepsilon\alpha z)}\hat{Q}^{(k)}_z \\ & -\left(\varepsilon\rho\nu+\varepsilon^2\nu^2z\right)(k-2)\,\hat{Q}^{(k)}_z-\varepsilon\rho\nu\alpha\hat{Q}^{(k)}_{\alpha z} \\ & +\tfrac{1}{2}\varepsilon^2\nu^2(k-1)(k-2)\hat{Q}^{(k)} \qquad \text{for } \tau>0. \end{aligned} \tag{A.15}$$

We now define $H^{(k)}(\tau,z,\alpha)$ by

$$\hat{Q}^{(k)} = \sqrt{B(\varepsilon\alpha z)/B(0)}H^{(k)} \tag{A.16a}$$

Then

$$\hat{Q}_z^{(k)} = \sqrt{B(\varepsilon\alpha z)/B(0)}\left\{H_z^{(k)} + \tfrac{1}{2}\varepsilon\alpha\frac{B'}{B}H^{(k)}\right\}, \tag{A.16b}$$

$$\hat{Q}_{zz}^{(k)} = \sqrt{B(\varepsilon\alpha z)/B(0)}\left\{H_{zz}^{(k)} + \varepsilon\alpha\frac{B'}{B}H_z^{(k)} + \varepsilon^2\alpha^2\left(\tfrac{1}{2}\frac{B''}{B} - \tfrac{1}{4}\frac{B'B'}{B^2}\right)H^{(k)}\right\}, \tag{A.16c}$$

$$\hat{Q}_{\alpha z}^{(k)} = \sqrt{B(\varepsilon\alpha z)/B(0)}\left\{H_{\alpha z}^{(k)} + \tfrac{1}{2}\varepsilon\alpha\frac{B'}{B}H_\alpha^{(k)} + \tfrac{1}{2}\varepsilon z\frac{B'}{B}H_z^{(k)} + \tfrac{1}{2}\varepsilon\frac{B'}{B}H^{(k)} + O(\varepsilon^2)\right\}. \tag{A.16d}$$

So

$$\begin{aligned} H_\tau^{(k)} = {} & \tfrac{1}{2}\left(1+2\varepsilon\rho\nu z+\varepsilon^2\nu^2z^2\right)H_{zz}^{(k)} - (k-2)\left(\varepsilon\rho\nu+\varepsilon^2\nu^2 z\right)H_z^{(k)} + \tfrac{1}{2}\varepsilon^2\rho\nu\alpha z\frac{B'}{B}H_z^{(k)} \\ & + \left\{\tfrac{1}{2}\varepsilon^2\nu^2(k-1)(k-2) - \tfrac{1}{2}\varepsilon^2\rho\nu\alpha(k-1)\frac{B'}{B} + \varepsilon^2\alpha^2\left(\tfrac{1}{4}\frac{B''}{B} - \tfrac{3}{8}\frac{B'B'}{B^2}\right)\right\}H^{(k)} \\ & -\varepsilon\rho\nu\alpha H_{\alpha z}^{(k)} - \tfrac{1}{2}\varepsilon^2\rho\nu\alpha^2\frac{B'}{B}H_\alpha^{(k)}, \end{aligned} \tag{A.17a}$$

with

$$H^{(k)}(\tau,z,\alpha) \to \delta(z) \qquad \text{as } \tau\to 0^+. \tag{A.17b}$$

Inspection of eqs. A.17a, A.17b shows that if we were to expand

$$H^{(k)}(\tau,z,\alpha) = H_0^{(k)}(\tau,z) + \varepsilon H_1^{(k)}(\tau,z) + \varepsilon^2 H_2^{(k)}(\tau,z,\alpha) + \cdots, \tag{A.18}$$

then $H^{(k)}$ would be independent of α until $O(\varepsilon^2)$, as indicated. Thus, the last two terms in eq. A.17a are both asymptotically smaller then $O(\varepsilon^2)$, and can be neglected. Hence, we can write

$$\begin{aligned} H_\tau^{(k)} = {} & \tfrac{1}{2}\left(1+2\varepsilon\rho\nu z+\varepsilon^2\nu^2z^2\right)H_{zz}^{(k)} - (k-2)\left(\varepsilon\rho\nu+\varepsilon^2\nu^2 z\right)H_z^{(k)} + \tfrac{1}{2}\varepsilon^2\rho\nu\alpha z\frac{B'}{B}H_z^{(k)} \\ & + \left\{\tfrac{1}{2}\varepsilon^2\nu^2(k-1)(k-2) - \tfrac{1}{2}\varepsilon^2\rho\nu\alpha(k-1)\frac{B'}{B} + \varepsilon^2\alpha^2\left(\tfrac{1}{4}\frac{B''}{B} - \tfrac{3}{8}\frac{B'B'}{B^2}\right)\right\}H^{(k)}, \end{aligned} \tag{A.19a}$$

with

$$H^{(k)}(\tau,z,\alpha) \to \delta(z) \qquad \text{as } \tau\to 0^+, \tag{A.19b}$$

through $O(\varepsilon^2)$.

There are no longer any derivatives with respect to α in eq. A.19a. Therefore α can be treated as a parameter instead of as a variable. I.e., the problem has been reduced to one spatial dimension, at least through $O(\varepsilon^2)$. In [1, SABR] we constructed explicit asymptotic solutions to eqs. A.19a, A.19b. Here we take a different approach: We note that for $k=2$,

$$\begin{aligned} H_\tau^{(2)} = {} & \tfrac{1}{2}\left(1+2\varepsilon\rho\nu z+\varepsilon^2\nu^2z^2\right)H_{zz}^{(2)} + \tfrac{1}{2}\varepsilon^2\rho\nu\alpha z\frac{B'}{B}H_z^{(2)} \\ & -\tfrac{1}{2}\varepsilon^2\rho\nu\alpha\frac{B'}{B}H^{(2)} + \varepsilon^2\alpha^2\left(\tfrac{1}{4}\frac{B''}{B} - \tfrac{3}{8}\frac{B'B'}{B^2}\right)H^{(2)}, \end{aligned} \tag{A.20a}$$

whilst the equation for the $k = 0$ can be written as

$$H^{(0)}_\tau = \tfrac{1}{2}\left[\left(1 + 2\varepsilon\rho\nu z + \varepsilon^2\nu^2 z^2\right) H^{(0)}\right]_{zz} + \tfrac{1}{2}\varepsilon^2\rho\nu\alpha z\frac{B'}{B}H^{(0)}_z + \tfrac{1}{2}\varepsilon^2\rho\nu\alpha\frac{B'}{B}H^{(0)} + \varepsilon^2\alpha^2\left(\tfrac{1}{4}\frac{B''}{B} - \tfrac{3}{8}\frac{B'B'}{B^2}\right)H^{(0)}. \tag{A.20b}$$

Both satisfy the same initial condition

$$H^{(0)} \to \delta(z), \qquad H^{(2)} \to \delta(z) \qquad \text{as } \tau \to 0^+. \tag{A.20c}$$

If the term $+\frac{1}{2}\varepsilon^2\rho\nu\alpha\,(B'/B)\,H^{(0)}$ in eq. A.20b had the opposite sign, then $\left(1 + 2\varepsilon\rho\nu z + \varepsilon^2\nu^2 z^2\right) H^{(0)}$ would satisfy the equation for $H^{(2)}$, namely eq. A.20a, through $O(\varepsilon^2)$. Since it also satisfies the initial condition for $H^{(2)}$, we could conclude that $\left(1 + 2\varepsilon\rho\nu z + \varepsilon^2\nu^2 z^2\right) H^{(0)}$ and $H^{(2)}$ are identical, at least through $O(\varepsilon^2)$. Pursuing this idea, we consider

$$U(\tau, z, \alpha) = \left(1 + 2\varepsilon\rho\nu z + \varepsilon^2\nu^2 z^2\right) e^{\varepsilon^2\rho\nu\alpha\Gamma\tau} H^{(0)}(\tau, z, \alpha), \tag{A.21a}$$

where Γ matches $-B'(\varepsilon\alpha z)/B(\varepsilon\alpha z)$ to leading order,

$$\Gamma = -\frac{B'(\varepsilon\alpha z)}{B(\varepsilon\alpha z)}\left\{1 + O(\varepsilon)\right\}. \tag{A.21b}$$

A precise choice of Γ will be made later. Since $\varepsilon^2\Gamma_z = O(\varepsilon^3)$ and $\varepsilon^2\Gamma_{zz} = O(\varepsilon^3)$, eq. A.20b shows that U satisfies

$$U_\tau = \tfrac{1}{2}\left(1 + 2\varepsilon\rho\nu z + \varepsilon^2\nu^2 z^2\right) U_{zz} + \tfrac{1}{2}\varepsilon^2\rho\nu\alpha z\frac{B'}{B}U_z - \tfrac{1}{2}\varepsilon^2\rho\nu\alpha\frac{B'}{B}U + \varepsilon^2\alpha^2\left(\tfrac{1}{4}\frac{B''}{B} - \tfrac{3}{8}\frac{B'B'}{B^2}\right)U, \tag{A.22a}$$

through $O(\varepsilon^2)$, with

$$U \to \delta(z) \qquad \text{as } \tau \to 0^+. \tag{A.22b}$$

This is identical to the PDE and initial condition for $H^{(2)}$, so uniqueness allows us to conclude that U and $H^{(2)}$ are the same through $O(\varepsilon^2)$:

$$H^{(2)}(\tau, z, \alpha) = H^{(0)}(\tau, z, \alpha)e^{\varepsilon^2\rho\nu\alpha\Gamma\tau}\left(1 + 2\varepsilon\rho\nu z + \varepsilon^2\nu^2 z^2\right). \tag{A.23}$$

We now chase back through the transformations, and noting that $B'(\varepsilon\alpha z)/B(\varepsilon\alpha z)$ is $-C'(F)$, we obtain

$$Q^{(2)}(t, f, \alpha; T, F) = \alpha^2 Q^{(0)}(t, f, \alpha; T, F)\left\{1 + 2\varepsilon\rho\nu z + \varepsilon^2\nu^2 z^2\right\} e^{+\varepsilon^2\rho\nu\alpha\Gamma(T-t)} \tag{A.24a}$$

through $O(\varepsilon^2)$, where

$$\Gamma = C'(F)\left\{1 + O(\varepsilon)\right\}. \tag{A.24b}$$

Substituting equation A.24a into the conservation equation, A.5, we obtain

$$Q^{(0)}_T = \tfrac{1}{2}\varepsilon^2\alpha^2\left[\left(1 + 2\varepsilon\rho\nu z + \varepsilon^2\nu^2 z^2\right) e^{+\varepsilon^2\rho\nu\alpha\Gamma(T-t)} C^2(F) Q^{(0)}\right]_{FF} \qquad \text{for } T > t. \tag{A.25}$$

This is the effective forward equation for $Q(T,F) \equiv Q^{(0)}(t,f,\alpha;T,F)$.

This is the effective forward equation. In the derivation we could have used $\Gamma = C'(F)$ or $C'(f)$ or even $C'\left([f+F]/2\right)$ or any other reasonable choice without losing the $O(\varepsilon^2)$ accuracy of the final result. However, higher order analysis suggests that using an average value is more accurate than $C'(f)$ or $C'(F)$. This is why we have chosen

$$\Gamma = \frac{C(F) - C(f)}{F - f} \tag{A.26}$$

in the text. The examples seem to indicate that this is a good choice.

Appendix B. Boundary conditions.

To simplify notation, let

$$D^2(F) = \left(1 + 2\varepsilon\rho\nu z + \varepsilon^2\nu^2 z^2\right) e^{\varepsilon^2\rho\nu a\Gamma(F)(T-t)} C^2(F), \tag{B.1}$$

so the effective forward equation is

$$Q_T = \tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q\right]_{FF} \qquad \text{for } F_{\min} < F < F_{\max}. \tag{B.2}$$

Note that $C(F)$ can be zero where, and only where, $D(F) = 0$. The change in the total probability in any interval $F_1 < F < F_2$ is

$$\frac{d}{dT}\int_{F_1}^{F_2} Q(T,F)dF = \tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q\right]_F\Big|^{F_2} - \tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q\right]_F\Big|^{F_1}, \tag{B.3}$$

so clearly the probability flux at any point F is

$$J(F) = -\tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q\right]_F. \tag{B.4}$$

It is natural to place the lower boundary at the barrier, where $C(F)$, and hence $D(F)$, is zero. So let us first consider the cases where $D(F_{\min}) = 0$, with

$$D(F) \sim \text{const} \cdot (F - F_{\min})^{\beta} \qquad \text{as } F \to F_{\min}. \tag{B.5}$$

Barriers have been studied extensively in stochastic processes and PDEs [Feller, others, Hagan $\beta - \eta$]. If $0 < \beta < \frac{1}{2}$, it is known that $F_{\min}$ is a regular boundary. Paths can both enter and leave the barrier, and it is theoretically possible for probability to diffuse through the barrier, reaching the "forbidden region" $F < F_{\min}$. We do not consider any models in which paths reach the region $F < F_{\min}$ below the boundary. Any flux of probability from the interior $F > F_{\min}$ to the boundary $F_{\min}$ must accumulate as a delta function at the boundary:

$$Q(T,F) = Q^L(T)\delta(F - F_{\min}) \qquad \text{at } F = F_{\min}. \tag{B.6a}$$

Conservation requires that the accumulation of probability balances the flux,

$$\frac{dQ^L}{dT} = \lim_{F \to F_{\min}^+} \tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q\right]_F. \tag{B.6b}$$

For regular boundaries we also need to prescribe a boundary condition at $F_{\min}$. Typically this boundary condition would be absorbing,

$$D^2(F)Q(T,F) \to 0 \qquad \text{as } F \to F_{\min}^+, \tag{B.7a}$$

no flux,

$$\left[D^2(F)Q\right]_F \to 0 \qquad \text{as } F \to F^+_{\min}, \tag{B.7b}$$

or even mixed

$$\left[D^2(F)Q\right]_F - \gamma D^2(F)Q(T,F) \to 0 \qquad \text{as } F \to F^+_{\min}. \tag{B.7c}$$

To determine the correct boundary condition, note that the expected value of $\tilde{F}(T)$ has to be constant,

$$E\left\{\tilde{F}(T)\middle|\ \tilde{F}(t) = f,\ \tilde{A}(t) = \alpha\right\} = F_{\min}Q^L(T) + \int_{F_{\min}}^{\infty} FQ(T,F)dF = f, \tag{B.8}$$

for $\tilde{F}(T)$ to be a Martingal. Therefore,

$$\frac{d}{dT}\left\{F_{\min}Q^L(T) + \int_{F_{\min}}^{\infty} FQ(T,F)dF\right\} = 0. \tag{B.9}$$

Substituting B.2 for Q_T and B.6b for $Q^L(T)$, and integrating by parts shows that

$$\begin{aligned}\frac{d}{dT}\left\{F_{\min}Q^L(T) + \int_{F_{\min}}^{\infty} FQ(T,F)dF\right\} &= \tfrac{1}{2}\varepsilon^2\alpha^2 \int_{F_{\min}}^{\infty} \left[D^2(F)Q\right]_F dF \\ &= \lim_{F\to F^+_{\min}} -\tfrac{1}{2}\varepsilon^2\alpha^2 D^2(F)Q(T,F).\end{aligned} \tag{B.10}$$

Therefore the requirement that $\tilde{F}(T)$ be a Martingale means that $Q(T,F)$ must satisfy absorbing boundary conditions at $F_{\min}$,

$$\lim_{F\to F^+_{\min}} D^2(F)Q(T,F) = 0. \tag{B.11}$$

If $\frac{1}{2} < \beta < 1$, then $F_{\min}$ is an exit boundary. In this case some paths reach the barrier in finite time, but no paths leave the barrier. Therefore there is a finite amount of probability at the boundary $F_{\min}$,

$$Q(T,F) = Q^L(T)\delta(F - F_{\min}) \qquad \text{at } F = F_{\min}, \tag{B.12a}$$

and it accumulates according to the flux,

$$\frac{dQ^L}{dT} = \lim_{F\to F^+_{\min}} \tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q\right]_F. \tag{B.12b}$$

For exit boundaries, the probability density automatically satisfies the absorbing boundary condition

$$D^2(F)Q(T,F) \to 0 \qquad \text{as } F \to F^+_{\min}, \tag{B.12c}$$

as is well known [Feller]. Theoretically no boundary condition is needed at $F_{\min}$, since the absorbing boundary condition occurs automatically. In practice, however, the absorbing boundary condition should be applied explicitly when solving the effective forward equation numerically, since most numerical finite difference schemes engender a slight amount of numerical dispersion, meaning that $D(F_{\min})$ is effectively slightly positive. Even if we could develop and employ a dispersion-free finite difference scheme, applying this boundary condition would be redundant, and not lead to any contradictions.

If $\beta \geq 1$, the barrier at $F_{\min}$ is an *inaccessible*, or *natural,* boundary. No paths can reach the boundary, and the probability and flux both go to zero near the boundary,

$$D^2(F)Q(T,F) \to 0, \qquad \tfrac{1}{2}\left[D^2(F)Q\right]_F \to 0 \qquad \text{as } F \to F^+_{\min}. \tag{B.13}$$

In theory, no delta function is needed at $F_{\min}$, since no paths reach the boundary. In practice, due to numerical dispersion, small amounts of probability reach the boundary. By incorporating a delta function at the boundary,

$$Q(T,F) = Q^L(T)\delta(F - F_{\min}) \qquad \text{at } F = F_{\min}, \tag{B.14a}$$

one can keep track of this probability,

$$\frac{dQ^L}{dT} = \lim_{F \to F^+_{\min}} \tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q\right]_F. \tag{B.14b}$$

This ensures that probability is exactly conserved. Even if we had a dispersion-free numerical scheme, it would just result in $Q^L(T)$ being exactly zero, so the δ function would be redundant, but not erroneous. Similarly, we keep the absorbing boundary conditions at $F_{\min}$, even though it should be satisfied automatically.

In summary, whenever there is a barrier at $F_{\min}$, we use absorbing boundary bounary conditions

$$D^2(F)Q(T,F) \to 0 \qquad \text{as } F \to F^+_{\min}, \tag{B.15a}$$

and use a delta function

$$Q(T,F) = Q^L(T)\delta(F - F_{\min}) \qquad \text{at } F = F_{\min}, \tag{B.15b}$$

with

$$\frac{dQ^L}{dT} = \lim_{F \to F^+_{\min}} \tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q\right]_F, \tag{B.15c}$$

in all cases.

There may be situations in which it makes sense to place the boundary $F_{\min}$ at a point where $C(F_{\min}) \neq 0$. For example, one may wish to put the boundary at $F_{\min} = 0$, regardless of whether $C(F_{\min})$ is zero or not. In this case we must still use B.15a - B.15c at the boundary: Since we are not allowing any paths to go below $F_{\min}$, we have to allow a delta function at $F_{\min}$ with the probability at $F_{\min}$ increasing according to B.15c, and to preserve the Martingale property of $\tilde{F}(T)$, we need to use the absorbing boundary condition B.15a

The upper boundary $F_{\max}$ should be set high enough so it has no appreciable effect on option prices; typically setting $F_{\max}$ to be roughly 4 to 6 standard deviations above the forward suffices. This requires

$$z\left(F_{\max}\right) = \frac{1}{\varepsilon\alpha}\int_f^{F_{\max}} \frac{dF'}{C(F')} = \frac{2}{\varepsilon\nu}\sinh\theta\left(\cosh\theta + \rho\sinh\theta\right) \tag{B.16a}$$

with

$$\theta = \tfrac{1}{2}\varepsilon\nu \cdot (4 \text{ to } 6)\sqrt{\tau_{ex}}, \tag{B.16b}$$

|as shown in Appendix D. Although the boundary condition is irrelevent if $F_{\max}$ is chosen large enough, we find it cleanest to treat the boundary at $F_{\max}$ the same as the boundary at $F_{\min}$: We allow a delta function at $F_{\max}$,

$$Q(T,F) = Q^R(T)\delta(F - F_{\max}) \qquad \text{at } F = F_{\max}, \tag{B.17a}$$

where

$$\frac{dQ^R}{dT} = -\lim_{F\to F^-_{\max}} \tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q\right]_F . \tag{B.17b}$$

This ensures that probability is conserved exactly, and by examining the size of $Q^R(T)$, we can determine whether the boundary $F_{\max}$ needs to be increased. We also use absorbing boundary conditions,

$$D^2(F)Q(T,F) \to 0 \qquad \text{as } F \to F^-_{\max}, \tag{B.17c}$$

which ensures that $\tilde{F}(T)$ is exactly a Martingale.

To summarize, let us write the density as

$$Q(T,F) = \begin{cases} Q^L(T)\delta(F-F_{\min}) & \text{at } F = F_{\min} \\ Q^c(T,F) & \text{for } F_{\min} < F < F_{\max} \\ Q^R(T)\delta(F-F_{\max}) & \text{at } F = F_{\max} \end{cases} . \tag{B.18}$$

where the superscript ^{c} is being used to denote the continuous part of the density. Then $Q^c(t,F)$ satisfies the boundary value problem

$$Q^c_T = \tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q^c\right]_{FF} \qquad \text{for } F_{\min} < F < F_{\max}, \tag{B.19a}$$

with the boundary conditions

$$D^2(F)Q^c \to 0 \qquad \text{as } F \to F^+_{\min}, \tag{B.19b}$$

$$D^2(F)Q^c \to 0 \qquad \text{as } F \to F^-_{\max}, \tag{B.19c}$$

for $t < T < \tau_{ex}$, and the initial condition

$$Q^c(T,F) \to \delta(F-f) \qquad \text{as } T \to t^+ \tag{B.19d}$$

The probability at the boundaries is

$$\frac{dQ^L}{dT} = \lim_{F\to F^+_{\min}} \tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q^c\right]_F \tag{B.19e}$$

$$\frac{dQ^R}{dT} = -\lim_{F\to F^-_{\max}} \tfrac{1}{2}\varepsilon^2\alpha^2 \left[D^2(F)Q^c\right]_F \tag{B.19f}$$

and the initial conditions are

$$Q^L(0) = 0, \qquad Q^R(0) \to 0 \qquad \text{as } T \to t^+. \tag{B.19g}$$

B.1. Boundary layer analysis. We believe that the delta function $Q^L(T)$ arises because when the forward $\tilde{F}(T)$ is near enough to the boundary, other mechanisms come into play. After all, there must be some reason that $\tilde{F}(T)$ doesn't cross the boundary. To show how this could come about, consider the effective forward equation

$$Q_T = \tfrac{1}{2}\varepsilon^2\alpha^2 \left[\tilde{D}^2(F)Q\right]_{FF} \qquad \text{for } 0 < F < \infty, \tag{B.20a}$$

$$Q = \delta(F-f) \qquad \text{at } T \to 0, \tag{B.20b}$$

where $D(F)$ has been modified near the boundary:

$$\tilde{D}(F) = \begin{cases} D(\eta)F/\eta & \text{for } 0 < F < \eta \\ D(F) & \text{for } F > \eta \end{cases} . \tag{B.20c}$$

In the limit $\eta \to 0$, we will recover the absorbing boundary condition and the delta function $Q^L(T)$.

In the region $0 < F < \eta$, the new effective forward equation is

$$Q_T = \tfrac{1}{2}\left(\frac{\varepsilon\alpha}{\eta}D(\eta)\right)^2 \left[F^2 Q\right]_{FF} \qquad \text{for } 0 < F < \eta, \tag{B.21a}$$

$$Q = 0 \qquad \text{at } T \to 0. \tag{B.21b}$$

The boundary at $F = 0$ is an inaccessible (natural) boundary, and requires no boundary condition. At $F = \eta$,

$$Q(T, \eta^-) = Q(T, \eta^+) \tag{B.22a}$$

$$\left(\frac{\varepsilon\alpha}{\eta}D(\eta)\right)^2 \lim_{F\to\eta^-} \left[F^2 Q(T,F)\right]_F = \lim_{F\to\eta^+} \left[D^2(F) Q(T,F)\right]_F \tag{B.22b}$$

since the probability Q and the flux must be continuous.

Define $Q^L(T)$ as the total amount of probablity in $0 < F < \eta$,

$$Q^L(T) = \int_0^\eta Q(T,F)dF. \tag{B.23}$$

We note that

$$\frac{dQ^L(T)}{dT} = \int_0^\eta Q_T(T,F)dF = \left(\frac{\varepsilon\alpha}{\eta}D(\eta)\right)^2 \int_0^\eta \left[F^2 Q\right]_{FF} dF = \left(\frac{\varepsilon\alpha}{\eta}D(\eta)\right)^2 \lim_{F\to\eta-} \left[F^2 Q\right]_F, \tag{B.24}$$

so

$$\frac{dQ^L(T)}{dT} = \lim_{F\to\eta^+} \left[D^2(F) Q(T,F)\right]_F .$$

In the limit $\eta \to 0$, we have a finite probability $Q^L(T)$ in an infinitely thin region; i.e., a delta function:

$$Q(T,F) = Q^L(T)\delta(F), \tag{B.25a}$$

with

$$\frac{dQ^L(T)}{dT} = \lim_{F\to 0^+} \left[D^2(F) Q(T,F)\right]_F . \tag{B.25b}$$

To investigage the effective boundary condition, define the Fourier transform of Q,

$$\tilde{Q}(\lambda, F) = \int_0^\infty Q(T,F)e^{-\lambda T}dT. \tag{B.26}$$

Then

$$\frac{A}{\eta^2}\left[F^2\tilde{Q}\right]_{FF} - \lambda\tilde{Q} = 0 \qquad \text{for } 0 < F < \eta, \tag{B.27a}$$

where the constant A is

$$A = \tfrac{1}{2}\varepsilon^2\alpha^2 D^2(\eta) > 0. \tag{B.27b}$$

The general solution to **??** is

$$\tilde{Q}(\lambda, F) = C_1(\lambda)\,(F/\eta)^{\gamma_1} + C_2(\lambda)\,(F/\eta)^{\gamma_2}\,, \tag{B.28a}$$

where

$$\gamma_{1,2} = \frac{-3 \pm \sqrt{1 + 4\lambda\eta^2/A}}{2}. \tag{B.28b}$$

The integral

$$\int_0^\eta (F/\eta)^{\gamma_2}\, dF \tag{B.29}$$

is infinite when $\mathrm{Re}\,\{\lambda\} \geq 0$. (It suffices to consider the region $\mathrm{Re}\,\{\lambda\} \geq Z$ for any constant Z; and then using analytic continuation. See [CKP].) Therefore $C_2(\lambda) = 0$ as $\tilde{Q}(\lambda, F)$ must be integrable, and thus

$$\tilde{Q}(\lambda, F) = C_1(\lambda)\,(F/\eta)^{\gamma_1} \qquad \text{for } 0 < F < \eta. \tag{B.30}$$

Since

$$Q(T, \eta^+) = \tilde{Q}(\lambda, \eta^-) = C_1(\lambda), \tag{B.31}$$

and

$$\begin{aligned} \lim_{F\to\eta^+} \left[D^2(F)Q(T,F)\right]_F &= (\varepsilon\alpha D(\eta))^2\, C_1(\lambda) \left[(F/\eta)^{2+\gamma_1}\right]_F \\ &= (2+\gamma_1)\,(\varepsilon\alpha D(\eta))^2\, C_1(\lambda)/\eta, \end{aligned} \tag{B.32}$$

we have

$$\tfrac{1}{2}\,(\varepsilon\alpha D(\eta))^2\, Q(T, \eta^+) = \frac{2\eta}{2+\gamma_1} \lim_{F\to\eta^+} \left[D^2(F)Q(T,F)\right]_F. \tag{B.33}$$

In the limit that $\eta \to 0$, the right hand side goes to zero, and we obtain the absorbing boundary condition:

$$\lim_{\eta\to 0} \tfrac{1}{2}\,(\varepsilon\alpha D(\eta))^2\, Q(T, \eta) = 0. \tag{B.34}$$

Appendix C. Moment preserving finite difference schemes.

We require our numerical scheme to conserve probability and the first moment exactly, so that

$$Q^L(T) + \int_{F_{\min}}^{F_{\max}} Q^c(T, F)dF + Q^R(T) = 1, \tag{C.1a}$$

$$F_{\min}Q^L(T) + \int_{F_{\min}}^{F_{\max}} FQ^c(T, F)dF + F_{\max}Q^R(T) = f. \tag{C.1b}$$

This ensures call-put parity, and provided that $Q^c(T, F) \geq 0$ for all F, it also ensures that the numerical solution itself represents an exactly arbitrage free model.

To simplify notation, define

$$M(T,F) = \tfrac{1}{2}\varepsilon^2\alpha^2 D^2(F) = \tfrac{1}{2}\varepsilon^2\alpha^2\left(1 + 2\varepsilon\rho\nu z + \varepsilon^2\nu^2 z^2\right) e^{\varepsilon^2\rho\nu\alpha\Gamma T} C^2(F), \tag{C.2a}$$

where

$$z(F) = \frac{1}{\varepsilon\alpha}\int_f^F \frac{df'}{C(f')}, \qquad \Gamma = \frac{C(F) - C(f)}{F - f}. \tag{C.2b}$$

We also set $t = 0$ without loss of generality.

C.1. Grid generation. The integration domain is

$$F_{\min} < F < F_{\max}. \tag{C.3}$$

We discretize F so that

$$F_{\max} = F_{\min} + Jh. \tag{C.4a}$$

and define

$$F_j \equiv F_{\min} + \left(j - \tfrac{1}{2}\right) h \qquad \text{for } j = 0, 1, ..., J+1, \tag{C.4ba}$$

to be the midpoints of the intervals from $(j-1)\,h$ to jh. In this process we adjust h slightly so that f occurs exactly at the midpoint of it's interval:

$$f \equiv F_{j_0} = F_{\min} + \left(j_0 - \tfrac{1}{2}\right) h. \tag{C.4c}$$

This allows us to implement the initial conditions in an exactly bias-free manner.

Let $Q_j^n = Q^c(n\delta, F_j)$ be the probability density at $T = n\delta, F = F_j$. Specifically, define

$$Q_j^n = \frac{1}{h}\int_{F_{\min}+(j-1)h}^{F_{\min}+jh} Q^c(n\delta, F')dF' \qquad \text{for } j = 1, ..., J \tag{C.5}$$

so that hQ_j^n is the total probability in the j^{th} grid cell at time step n. We usually use around 200 to 500 points for our grid, and divide $0 < T < \tau_{ex}$ into 30 to 100 timesteps.

C.2. Finite difference scheme. We integrate the effective forward equation using a Crank-Nicholson scheme, which averages explicit and implicit centerred difference equations[8, NumRecInC]. Not only is this scheme unconditionally stable, it is second order accurate in time.

To advance from timestep n to $n+1$, we need to solve

$$\begin{aligned} Q_j^{n+1} - Q_j^n = {} & \frac{dt}{2h^2}\left\{M_{j+1}^{n+1}Q_{j+1}^{n+1} - 2M_j^{n+1}Q_j^{n+1} + M_{j-1}^{n+1}Q_{j-1}^{n+1}\right\} \\ & \frac{dt}{2h^2}\left\{M_{j+1}^{n}Q_{j+1}^{n} - 2M_j^{n}Q_j^{n} + M_{j-1}^{n}Q_{j-1}^{n}\right\} \qquad \text{for } j = 1, 2, ..., J, \end{aligned} \tag{C.6}$$

which we re-write as

$$\begin{aligned} Q_j^{n+1} - \frac{dt}{2h^2}\left\{M_{j+1}^{n+1}Q_{j+1}^{n+1} - 2M_j^{n+1}Q_j^{n+1} + M_{j-1}^{n+1}Q_{j-1}^{n+1}\right\} = {} & \\ Q_j^n + \frac{dt}{2h^2}\left\{M_{j+1}^{n}Q_{j+1}^{n} - 2M_j^{n}Q_j^{n} + M_{j-1}^{n}Q_{j-1}^{n}\right\} \qquad \text{for } j = 1, 2, ..., J. & \end{aligned} \tag{C.7a}$$

The absorbing boundary conditions yield

$$M_0^{n+1}Q_0^{n+1} = -M_1^{n+1}Q_1^{n+1} \qquad \text{at } j=0, \tag{C.7b}$$

$$M_{J+1}^{n+1}Q_{J+1}^{n+1} = -M_J^{n+1}Q_J^{n+1} \qquad \text{at } j=J+1. \tag{C.7c}$$

Note that the points $j=0$ and $j=J+1$ fall outside the domain; these shadow points simply enable us to obtain the correct boundary condition at the "true" boundaries $F_{\min}=F_{1/2}$ and $F_{\max}=F_{J+1/2}$. Note that this scheme only requires the solution of a tridiagonal system, so the computational work scales linearly with J and the number of timesteps N.

We also need to solve for the probabilities $Q^L(t)$ and $Q^R(t)$ at the left and right boundaries. Let

$$Q_L^n = Q^L(n\delta), \qquad Q_R^n = Q^R(n\delta), \tag{C.8}$$

be the boundary probabilities at timestep n. At each time step, after solving for $Q_0^{n+1}, Q_1^{n+1}, \ldots, Q_J^{n+1}, Q_{J+1}^{n+1}$, we update the values Q_L^{n+1} and Q_R^{n+1}

$$Q_L^{n+1} - Q_L^n = \frac{dt}{2h}\left\{M_1^{n+1}Q_1^{n+1} - M_0^{n+1}Q_0^{n+1} + M_1^nQ_1^n - M_0^nQ_0^n\right\}, \tag{C.9a}$$

$$Q_R^{n+1} - Q_R^n = -\frac{dt}{2h}\left\{M_{J+1}^{n+1}Q_{J+1}^{n+1} - M_J^{n+1}Q_J^{n+1} + M_{J+1}^nQ_{J+1}^n - M_J^nQ_J^n\right\}. \tag{C.9b}$$

Note that this is also second order accurte in time.

To set the initial condition, recall that we adjusted h so that

$$f = F_{j_0} \equiv F_{\min} + \left(j_0 - \tfrac{1}{2}\right)h. \tag{C.10}$$

for some integer j_0 when we set up the grid. The initial condition is simply

$$Q_L^0 = 0, \qquad Q_j^0 = \begin{cases} 0 & \text{for } j \neq j_0 \\ 1/h & \text{for } j = j_0 \end{cases}, \qquad Q_R^0 = 0. \tag{C.11}$$

C.3. Moments. It is easily seen that this scheme conserves probability. The total probabilty is

$$Q_L^n + \int_{F_{\min}}^{F_{\max}} Q^c(t_n, F')dF' + Q_R^n \equiv Q_L^n + \sum_{j=1}^{J} hQ_j^n + Q_R^n. \tag{C.12}$$

Summing equation C.7a over j,and using equations C.9a, C.9b, yields

$$\begin{aligned}\sum_{j=1}^{J} h\left(Q_j^{n+1} - Q_j^n\right) &= \frac{dt}{2h}\left\{M_{J+1}^{n+1}Q_{J+1}^{n+1} - M_J^{n+1}Q_J^{n+1} - M_1^{n+1}Q_1^{n+1} + M_0^{n+1}Q_0^{n+1}\right\} \\ &\quad + \frac{dt}{2h}\left\{M_{J+1}^nQ_{J+1}^n - M_J^nQ_J^n - M_1^nQ_1^n + M_0^nQ_0^n\right\} \\ &= -\left(Q_L^{n+1} - Q_L^n\right) - \left(Q_R^{n+1} - Q_R^n\right)\end{aligned} \tag{C.13}$$

for each timestep n. Thus

$$Q_L^{n+1} + \sum_{j=1}^{J} hQ_j^{n+1} + Q_R^{n+1} = Q_L^n + \sum_{j=1}^{J} hQ_j^n + Q_R^n, \tag{C.14a}$$

so the total probability is the same for all timesteps n. Since the total probability starts at 1 at $n = 0$,

$$Q_L^n + \sum_{j=1}^{J} hQ_j^n + Q_R^n = 1 \qquad \text{for all } n. \tag{C.15}$$

Similarliy, the first moment is

$$F_{\min}Q_L^n + \int_{F_{\min}}^{F_{\max}} F'Q^c(t_n, F')dF' + F_{\max}Q_R^n \equiv F_{\min}Q_L^n + \sum_{j=1}^{J} hF_jQ_j^n + F_{\max}Q_R^n. \tag{C.16}$$

We multiply eq. C.7a by F_j and sum over j. Since $F_{j+1} - 2F_j + F_{j-1} = 0$, this yields

$$\sum_{j=1}^{J} hF_j \left(Q_j^{n+1} - Q_j^n\right) \tag{C.17}$$

$$= \frac{dt}{2h}\left\{F_J M_{J+1}^{n+1}Q_{J+1}^{n+1} - F_0 M_1^{n+1}Q_1^{n+1} - F_{J+1}M_J^{n+1}Q_J^{n+1} + F_1 M_0^{n+1}Q_0^{n+1}\right\} \tag{C.18}$$

$$\frac{dt}{2h}\left\{F_J M_{J+1}^{n}Q_{J+1}^{n} - F_0 M_1^{n}Q_1^{n} - F_{J+1}M_J^{n}Q_J^{n} + F_1 M_0^{n}Q_0^{n}\right\}.$$

Using C.9a, C.9b, this becomes

$$F_{\min}\left(Q_L^{n+1} - Q_L^n\right) + \sum_{j=1}^{J} hF_j\left(Q_j^{n+1} - Q_j^n\right) + F_{\max}\left(Q_R^{n+1} - Q_R^n\right) \tag{C.19}$$

$$= \frac{dt}{4}\left(M_1^{n+1}Q_1^{n+1} + M_0^{n+1}Q_0^{n+1}\right) - \frac{dt}{4}\left(M_{J+1}^{n+1}Q_{J+1}^{n+1} + M_J^{n+1}Q_J^{n+1}\right)$$

$$+\frac{dt}{4}\left(M_1^nQ_1^n + M_0^nQ_0^n\right) - \frac{dt}{4}\left(M_{J+1}^nQ_{J+1}^n + M_J^nQ_J^n\right).$$

The absorbing boundary conditions C.7b, C.7c ensure that the right hand side is zero, so the first moment is conserved for each time step n:

$$F_{\min}Q_L^{n+1} + \sum_{j=1}^{J} hF_jQ_j^{n+1} + F_{\max}Q_R^{n+1} = F_{\min}Q_L^n + \sum_{j=1}^{J} hF_jQ_j^n + F_{\max}Q_R^n. \tag{C.20}$$

Hence,

$$F_{\min}Q_L^n + \sum_{j=1}^{J} hF_jQ_j^n + F_{\max}Q_R^n = hF_{j_0}Q_{j_0}^0 = f. \tag{C.21}$$

Thus the expected value of $\tilde{F}(T)$ remains exactly f for the numerical solution. This is a key reason we decided to use a uniform mesh; if we used a non-uniform mesh, a moment preserving finite difference scheme would yield a linear problem with a matrix of at least five non-zero diagonals instead of three.

C.4. Option pricing. We integrate the option prices assuming that the probability hQ_j^N is spread uniformly in each cell j. This yields

$$V_{call}(\tau_{ex},K) = f - K \qquad \text{for } K < F_{\min} \tag{C.22a}$$

$$V_{call}(\tau_{ex},K) = \tfrac{1}{2}\left(F_{\min}+kh-K\right)^2 Q_k^N + \sum_{j=k+1}^{J}\left[F_{\min}+\left(j-\tfrac{1}{2}\right)h-K\right]hQ_j^N + \left(F_{\max}-K\right)Q_R^N \qquad \text{for } F_{\min}+(k-1)h < K < F_{\min}+kh \tag{C.22b}$$

$$V_{call}(\tau_{ex},K) = 0 \qquad \text{for } K > F_{\max} \tag{C.22c}$$

for the call prices, and

$$V_{put}(\tau_{ex},K) = 0 \qquad \text{for } K < F_{\min} \tag{C.23a}$$

$$V_{put}(\tau_{ex},K) = \left(K-F_{\min}\right)Q_L^N + \sum_{j=1}^{k-1}\left[K-F_{\min}-\left(j-\tfrac{1}{2}\right)h\right]hQ_j^N + \tfrac{1}{2}\left[K-F_{\min}-(k-1)\,h\right]^2 Q_k^N \qquad \text{for } F_{\min}+(k-1)h < K < F_{\min}+kh \tag{C.23b}$$

$$V_{put}(\tau_{ex},K) = K - f \qquad \text{for } K > F_{\max} \tag{C.23c}$$

for the puts.

Appendix D. Dispersion.

We would like to be able to measure F, at least crudely, in terms of standard deviations from today's forward f. The effective forward equation can be written as

$$Q_T = \tfrac{1}{2}\varepsilon^2\alpha^2\left[D^2(F)Q\right]_{FF} \qquad \text{for } T > t, \tag{D.1a}$$

$$Q \to \delta(F-f) \qquad \text{as } T \to t, \tag{D.1b}$$

with

$$D^2(F) = \left(1+2\varepsilon\rho\nu z+\varepsilon^2\nu^2z^2\right)e^{\varepsilon^2\rho\nu a\Gamma(F)(T-t)}C^2(F). \tag{D.1c}$$

To leading order, the solution is a Gaussian density

$$Q\left(T,F\right) \approx \frac{1}{\sqrt{2\pi\left(T-t\right)}}e^{-y^2/2(T-t)}\frac{dy}{dF} \tag{D.2a}$$

with

$$y(F) = \frac{1}{\varepsilon\alpha}\int_f^F \frac{dF'}{D(F')}. \tag{D.2b}$$

Since we are working only to leading order, we can neglect the exponential factor $e^{\varepsilon^2\rho\nu a\Gamma(F)(T-t)}$ in $D(F)$. Integrating then yields

$$y(F) = \frac{1}{\varepsilon\nu}\log\left(\frac{\sqrt{1+2\varepsilon\rho\nu z(F)+\varepsilon^2\nu^2z^2(F)}+\rho+\varepsilon\nu z(F)}{1+\rho}\right), \tag{D.3a}$$

where

$$z(F) = \frac{1}{\varepsilon\alpha}\int_f^F \frac{dF'}{C(F')}. \tag{D.3b}$$

For F to be roughly N standard deviations above f on the exercise date τ_{ex}, we need y to be $+N\sqrt{\tau_{ex}-t}$. This occurs at the F where

$$z\left(F\right)=\frac{1}{\varepsilon\alpha}\int_{f}^{F}\frac{dF'}{C(F')}=\frac{2}{\varepsilon\nu}\sinh\theta\left(\cosh\theta+\rho\sinh\theta\right) \tag{D.4a}$$

with

$$\theta=\frac{\varepsilon\nu}{2}N\sqrt{\tau_{ex}-t}. \tag{D.4b}$$

Similarly, for F to be roughly N standard deviations below f on the exercise date, we need y to be $-N\sqrt{\tau_{ex}-t}$. This occurs where

$$-z\left(F\right)=\frac{1}{\varepsilon\alpha}\int_{F}^{f}\frac{dF'}{C(F')}=\frac{2}{\varepsilon\nu}\sinh\theta\left(\cosh\theta-\rho\sinh\theta\right) \tag{D.5a}$$

with

$$\theta=\frac{\varepsilon\nu}{2}N\sqrt{\tau_{ex}-t}. \tag{D.5b}$$

Appendix E. Shifted SABR model.

It is now commonly accepted that interest rates need not be strictly positive. Surely, though, if interest rates were to become too negative, then increasing amounts of money would be withdrawn from the banking system, putting a squeeze on deposits. So there should be some barrier to how negative interest rates can become, but this barrier is probably below zero. Thus, it may make more sense to use a shifted SABR model,

$$C(F)=\left(F+b\right)^{\beta}, \tag{E.1}$$

for some $b>0$.

For this model, the explicit implied vol formula is

$$\sigma_N(K)=\frac{\varepsilon\alpha\left(1-\beta\right)\left(f-K\right)}{\left(f+b\right)^{1-\beta}-\left(K+b\right)^{1-\beta}}\cdot\left(\frac{\zeta}{x(\zeta)}\right)\cdot \tag{E.2a}$$
$$\left\{1+\left[-\tfrac{1}{24}\frac{\beta\left(2-\beta\right)\left(1-\beta\right)^{2}\alpha^{2}\log^{2}\frac{f+b}{K+b}}{\left[\left(f+b\right)^{1-\beta}-\left(K+b\right)^{1-\beta}\right]^{2}}+\tfrac{1}{4}\rho\nu\alpha\frac{\left(f+b\right)^{\beta}-\left(K+b\right)^{\beta}}{f-K}+\frac{2-3\rho^{2}}{24}\nu^{2}\right]\varepsilon^{2}\tau_{ex}\right\},$$

where

$$\zeta=\frac{\nu}{\alpha}\frac{\left(f+b\right)^{1-\beta}-\left(K+b\right)^{1-\beta}}{1-\beta},\qquad x(\zeta)=\log\left(\frac{\sqrt{1-2\rho\zeta+\zeta^{2}}-\rho+\zeta}{1-\rho}\right). \tag{E.2b}$$

Our effective forward equation is

$$Q_T=\tfrac{1}{2}\varepsilon^{2}\alpha^{2}\left[\left(1+2\varepsilon\rho\nu z+\varepsilon^{2}\nu^{2}z^{2}\right)e^{\varepsilon^{2}\rho\nu a\Gamma(T-t)}C^{2}(F)Q\right]_{FF}\qquad\text{for } T>t. \tag{E.3a}$$

where

$$z(F)=\frac{\left(F+b\right)^{1-\beta}-\left(f+b\right)^{1-\beta}}{\varepsilon\alpha\left(1-\beta\right)},\qquad\Gamma=\frac{\left(F+b\right)^{\beta}-\left(f+b\right)^{\beta}}{F-f}. \tag{E.3b}$$

E.1. Stochastic normal model. As $\beta \to 0$, the shifted SABR model simplifies to the stochastic normal model,

$$C(F) = 1. \tag{E.4}$$

For this model, the explicit implied vol formula is

$$\sigma_N(K) = \varepsilon\alpha \cdot \left(\frac{\zeta}{x(\zeta)}\right) \cdot \left\{1 + \left[\tfrac{1}{4}\rho\nu\alpha + \tfrac{1}{24}\left(2 - 3\rho^2\right)\nu^2\right]\varepsilon^2\tau_{ex} + \cdots\right\}, \tag{E.5a}$$

where

$$\zeta = \frac{\nu}{\alpha}\left(f - K\right), \qquad x(\zeta) = \log\left(\frac{\sqrt{1 - 2\rho\zeta + \zeta^2} - \rho + \zeta}{1 - \rho}\right). \tag{E.5b}$$

The effective forward equation is

$$Q_T = \tfrac{1}{2}\varepsilon^2\left[\left(\alpha^2 + 2\rho\nu\alpha\left[F - f\right] + \nu^2\left[F - f\right]^2\right)Q\right]_{FF} \qquad \text{for } T > t. \tag{E.6a}$$

$$Q(T, F) \to \delta(F - f) \qquad \text{as } T \to t^+. \tag{E.6b}$$

Here there is no barrier, and the placement of $F_{\min}$ is an independent modeling decision.

REFERENCES

[1] D.T. Breeden and R. H. Litzenberger, *Prices of state-contingent claims implicit in option prices*, J. Business, 51 (1994), pp. 621-651.
[2] B. Dupire, *Pricing with a smile*, Risk, Jan. 1994, pp. 18–20.
[3] B. Dupire, *Pricing and hedging with smiles*, in Mathematics of Derivative Securities, M.A. H. Dempster and S. R. Pliska, eds., Cambridge University Press, Cambridge, 1997, pp. 103–111
[4] E. DermaN and I. Kani, *Riding on a smile*, Risk, Feb. 1994, pp. 32–39.
[5] E. Derman and I. Kani, *Stochastic implied trees: Arbitrage pricing with stochastic term and strike structure of volatility*, Int J. Theor Appl Finance, 1 (1998), pp. 61–110.
[6] J.M. Harrison and S. Pliska, *Martingales and stochastic integrals in the theory of continuous trading*, Stoch. Proc. Appl, 11 (1981), pp. 215-260.
[7] J.M. Harrison and D. Krebs, *Martingales and arbitrage in multiperiod securities markets*, J. Econ. Theory, 20 (1979), pp. 381-408.
[8] I. Karatzas, J.P. Lehoczky, S.E. Shreve, and G.L. Xus, *Martingale and duality methods for utility maximization in an incomplete market*, SIAM J. Control Optim, 29 (1991), pp. 702–730.
[9] J. Michael Steele, *Stochastic Calculus and Financial Applications*, Springer, 2001
[10] F. Jamshidean, *Libor and swap market models and measures*, Fin. Stoch. 1 (1997), pp. 293-330
[11] F. Black, *The pricing of commodity contracts*, Jour. Pol. Ec., 81 (1976), pp. 167-179.
[12] John C. Hull, *Options, Futures, and Other Derivative Securities*, Prentice Hall, 1997.
[13] P. Wilmott, *Paul Wilmott on Quantitative Finance*, John Wiley & Sons, 2000.
[14] Patrick S. Hagan and Diana E. Woodward, *Equivalent Black volatilities*, App. Math. Finance, **6** (1999), pp. 147–157.
[15] P. S. Hagan, A. Lesniewski and D. E. Woodward, *Geometric optics in finance*, in preparation.
[16] F. Wan, *Mathematical Models and Their Analysis,* Harper-Row, 1989.
[17] J. Hull and A. White, The pricing of options on assets with stochastic volatilities, J. of Finance, **42** (1987), pp. 281-300.
[18] S.L. Heston, *A closed-form solution for options with stochastic volatility with applications to bond and currency options*, Rev of Fin Studies, 6 (1993), pp. 327-343.
[19] A. Lewis, *Option Valuation Under Stochastic Volatility*, Financial Press, 2000.
[20] J.P. Fouque, G. Papanicolaou, K.R. Sirclair, *Derivatives in Financial Markets with Stochastic Volatility*, Cambridge Univ Press, 2000.
[21] N. A. Berner, *Hedging vanna & volga*, DKW, private communicatons
[22] J.D. Cole, *Perturbation Methods in Applied Mathematics*, Ginn-Blaisdell, 1968.

[23] J. Kevorkian and J.D. Cole, Perturbation Methods in Applied Mathematics, Springer-Verlag, 1985.
[24] J.F. Clouets, *Diffusion Approximation of a Transport Process in Random Media*, SIAM J Appl Math, **58** (1998), pp. 1604–1621.
[25] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*, Springer, 1988.
[26] B. Okdendal, *Stochastic Differential Equations*, Springer, 1998.
[27] M. Musiela and M. Rutkowski, *Martingale Methods in Financial Modelling*, Springer, 1998.
[28] G. B. Whitham, *Linear and Nonlinear Waves*, Wiley, 1974.
[29] J.C. Neu, *Thesis*, California Institute of Technology, 1978