Necessary and Sufficient No-Arbitrage Conditions for the SSVI/S3 Volatility Curve

Timothy R. Klassen, PFT Analytics

Abstract

The "SSVI" (aka "S3") implied volatility curve is the simplest curve that has three parameters to describe the at-the-money behavior of implied volatilities for a given term, while also having a sensible functional form in the call and put wings. We describe the necessary and sufficient conditions on its three parameters to avoid butterfly arbitrage. By considering dimensionless parameters in normalized strike space, a simple picture of the no-arbitrage region emerges. For fixed (normalized) volatility it is compact, simply connected, and surprisingly large for realistic volatilities. For very large (normalized) volatility the asymptotic wing constraints of Lee are not just necessary but also sufficient for the absence of arbitrage everywhere.

First Version: January 31, 2016

1 Introduction

Implied volatility surfaces are a basic building block in the infrastructure to price, trade and risk-manage derivatives. They are the primary ingredient used by market makers to price vanilla options (listed or OTC) and many light exotics. Implied volatility surfaces, with sensible wing extrapolation and preferably arbitrage-free, are also an important input for the calibration of more sophisticated "SLVJ" models required for more exotic derivatives/structured products. The existence of an arbitrage-free implied volatility surface is equivalent to the existence of a well-defined local volatility surface.

The last 15 years have seen a large amount of attention being devoted to the design of implied (and local) volatility curves and surfaces. The subject has slowly moved from being a dark art practiced by small, isolated cabals of quants and traders at most major trading firm, to the light of public scrutiny in academic publications, conference talks, etc. The move from ad-hoc volatility curves with obvious flaws towards more first-principle based approaches was, in the public domain, largely prompted by the work of Roger Lee, on the allowed asymptotic wing behavior [1], and Jim Gatheral [2, 3, 4], especially his work on a particular 5-parameter volatility curve known as "SVI". We refer to [4, 5], and references therein, for background and details on recent progress in the design of sensible volatility curves and surfaces.

Here we concentrate on a particular volatility curve that has received significant attention recently, the "SSVI" [4] aka "S3"[1] 3-parameter curve. It is a special case of the SVI curve.

[1] Independently of [4], it was designed by the present author in 2003 as the simplest curve that has 3-parameters to describe the ATM behavior while also having sensible wing behavior. It was used for the arbitrage-free calibration of implied and local volatility surfaces of many underliers at Wachovia Securities from that time on, and subsequently spread to several other trading firms.

We say "curve" and not "surface" since we will only be concerned with slices of the volatility surface for a given maturity T. In other words, we are here only concerned with the conditions for the absence of butterfly arbitrage. Working with suitable dimensionless parameters, we answer this question by mapping out the region of parameter space where there is no arbitrage. We do so by a combination of analytical and numerical work. For the subsets of parameters where either the slope or curvature of the SSVI/S3 curve vanishes, we derive analytical expressions of the no-arbitrage boundary for the remaining two parameters. This allows us to give an analytical expression that is, for practical purposes, a reasonably accurate presentation of the exact no-arbitrage region.

We add a few words of general motivation for the SSVI/S3 curve (for more details see [5]). It has the minimum number of parameters, three, required to have any hope of describing realistic volatility curves, via some kind of ATM (at-the-money) volatility, slope and curvature (aka "skew" and "convexity", etc). Naive designs using three parameters, e.g. using a parabola in log-moneyness for volatility (or volatility-squared) violate the by now well-known constraints [1] on the asymptotic wing behavior.[2] Similar problem arise in 3-parameter curves that are approximate solutions to a "microscopic" model, but ignore the wing asymptotics (e.g. the SABR curve).

We consider the SSVI/S3 curve a sort of "null hypothesis" of volatility curve fitting. And empirically, the vast majority of e.g. the US equity options universe (on stock, ETFs, indices, and futures) can be fit extremely well within bid-asks. It is only the top 100 or 200 or so underliers, by liquidity, out of several thousands, that can not be fit well enough with this curve [5]. Similar remarks apply to other asset classes. The SSVI/S3 curve, as input to a suitable implied volatility surface calibration algorithm, makes it relatively easy to produce arbitrage-free implied and local volatility surfaces in practice.

2 The SSVI/S3 Volatility Curve

We refer to [4, 5] and references therein for background and details on volatility curve and surface modeling. We assume that the implied volatility we will be talking about is that of some "pure" stock that has been cleansed of the effect of cash dividends (continuous or discrete proportional dividends are fine); see e.g. [6] for details on different pure stock models used for the pricing of vanilla (and other) options.

Vanilla options can then sensibly be priced in terms of the Black formula. In other words, the price of un-discounted calls and puts will be written in terms of forward F[3] and strike K as

$$\hat{C} = +F\,N(d_+) \;-\; K\,N(d_-) \tag{1}$$

$$\hat{P} = -F\,N(-d_+) + K\,N(-d_-) \tag{2}$$

Here $N(x)$ is the standard cumulative normal distribution, and in terms of log-moneyness y

$$y \;:=\; \ln(K/F) \tag{3}$$

we have

$$d_\pm \;=\; \frac{-y}{\hat{\sigma}} \pm \tfrac{1}{2}\hat{\sigma} \quad , \qquad \hat{\sigma} := \sigma\sqrt{T}\ , \tag{4}$$

[2] The most sensible parabola in log-moneyness would be for volatility to the fourth power, which has a chance of sensible wing behavior. But that implies the same asymptotics in the call and put wing, which is unnecessarily restrictive, while also violating empirical fact and common sense.

[3] The forward F here may or may not be that of the "observed" stock price, depending on the dividend model [6].

where $\sigma = \sigma(T, K)$ is the implied volatility of the option, and T the maturity (in a suitable time convention). Note that once (un-discounted) prices are normalized by the forward, the Black formula is a function of just two dimensionless variables, y and $\hat{\sigma}$. The region around $K = F$, or equivalently, $y = 0$, will be referred to as the "at-the-forward" or "ATF" region.

In practice we prefer to think of a volatility curve parametrization in terms of a dimensionless *shape* curve written in *normalized-strike* space, and one (dimensionfull) overall volatility parameter. Specifically, for the SSVI/S3 curve we write the square of the implied volatility at a given term T as

$$\sigma^2(z) \;=\; \sigma_0^2 \, \left(\tfrac{1}{2}(1 + s_2 z) + \sqrt{\tfrac{1}{4}(1 + s_2 z)^2 + \tfrac{1}{2} c_2 z^2} \right) \;, \tag{5}$$

where

$$z \;:=\; \ln(K/F)/\hat{\sigma}_0 \;=\; y/\hat{\sigma}_0 \tag{6}$$

will be referred to as "normalized strike", defined in terms of the normalized ATF volatility

$$\hat{\sigma}_0 := \sigma_0 \sqrt{T} \; .$$

Normalized strike (NS) provides a convenient and intuitive way of thinking about moneyness, as it makes moneyness (more) comparable across maturities and underliers (without some of the practical complications that using e.g. d_- entails). Besides the ATF volatility σ_0, the other two parameters of the SSVI/S3 curve, s_2, c_2, are the slope and curvature of the shape curve. In other words, the Taylor series of the above for small normalized strike is

$$\sigma^2(z) \;=\; \sigma_0^2 \, \left(1 + s_2 z + \tfrac{1}{2} c_2 z^2 + \mathcal{O}(z^3)\right) \;,$$

as is easily checked. We use the subscript '2' in s_2, c_2 as this is the expansion of σ^2.[4] The a priori range of parameters is $\sigma_0 \geq 0$, $-\infty < s_2 < \infty$, $c_2 \geq 0$. There are good theoretical and practical reasons for parametrizing implied volatility curves in terms of σ_0, s_2, c_2. Some will be seen here, for others we refer to [5].

Note that $c_2 \geq 0$ guarantees that the implied variance never becomes negative (and the converse holds too). In fact, the minimum variance is given by

$$\sigma^2_{\min} \;=\; \frac{\sigma_0^2}{1 + \frac{1}{2} s_2^2 / c_2} \; .$$

For future reference we define the *total implied variance*

$$w := \sigma^2 \, T \tag{7}$$

It is a product of the dimensionless shape curve and the total ATF variance parameter $\hat{\sigma}_0^2$, obviously.

Note that for large $|z|$ we have

$$\sigma^2(z) \;\to\; \sigma_0^2 \left(\sqrt{\tfrac{1}{4} s_2^2 + \tfrac{1}{2} c_2} \;\pm\; \tfrac{1}{2} s_2 \right) |z| \;+\; \tfrac{1}{2} \pm \frac{\frac{1}{4} s_2}{\sqrt{\frac{1}{4} s_2^2 + \frac{1}{2} c_2}} \;+\; \mathcal{O}(|z|^{-1}) \tag{8}$$

consistent with the at-worst linear behavior required by the Lee [1] bounds. One might argue that (5) is the simplest 3-parameter curve with a generic Taylor series around the origin (almost,

[4] Occasionally it is also useful to think about the expansion of σ itself, $\sigma(z) = \sigma_0 \, (1 + s_1 z + \frac{1}{2} c_1 z^2 + \ldots)$. Its first two coefficients are related to s_2, c_2 as follows: $s_1 = \frac{1}{2} s_2$, $c_1 = \frac{1}{2} c_2 - \frac{1}{4} s_2^2$.

variance curvature has to be positive), consistent with this behavior. For future reference let us give a name to the asymptotic coefficients in the above,

$$C_\pm \;:=\; \sqrt{\tfrac{1}{4}s_2^2 + \tfrac{1}{2}c_2} \;\pm\; \tfrac{1}{2}s_2 \;, \tag{9}$$

relevant for the call and put wing, respectively. The Lee bounds, $w(y) < 2|y|$, guaranteeing no arbitrage for asymptotically large $|y|$, become $\hat\sigma_0 C_\pm < 2$.

Note that instead of using s_2, c_2 we can equivalently parametrize SSVI/S3 in terms of C_+ and C_-. The relationship is $s_2 = C_+ - C_-$, $c_2 = 2\,C_+\,C_-$.[5]

Since we will always be working at fixed term T, the relation between y and z is a trivial rescaling by the constant $\hat\sigma_0$. For some questions it is more convenient to work with y, for others with z. By a slight abuse of notation we will mostly use the same names for various quantities, whether we consider them functions of y or z.

3 Butterfly Arbitrage

3.1 General considerations

For completeness we quickly review the results we will need in the following. For details and references to proofs within a rigorous mathematical setup see [4]. Under standard assumptions, vanilla options can be priced by integrating an *implied density* [7] against the pay-off, with the probability density ρ_K (in strike- or really future spot-space) given by the second-derivative of European vanilla prices as $\rho_K(S) = \partial_K^2 \hat C|_{K=S} = \partial_K^2 \hat P|_{K=S}$. A simple Jacobian relates this density $\rho_K(S)$ to that in terms of y, $\rho_y(y)$, (or z, or any other variable), and one can show that

$$\rho_y(y) \;=\; \frac{g(y)}{\hat\sigma(y)}\; n(d_-(y)) \;, \tag{10}$$

where $n(x) = N'(x)$ is the normal density, and the *density factor* function $g(y)$ is given by

$$g(y) \;=\; \left(1 - \frac{y\,w'(y)}{2w(y)}\right)^2 \;-\; \frac{1}{4}\left(\frac{1}{w(y)} + \frac{1}{4}\right) w'(y)^2 \;+\; \frac{1}{2}\,w''(y) \tag{11}$$

Here we think of the total variance w as a function of y. In the Black-Scholes case $w'(y) \equiv 0 \equiv w''(y)$, so that $g(y) \equiv 1$.

In deriving (10) is is convenient to first show that the *cumulative density* $c(y)$ is given by

$$c(y) \;=\; 1 \;-\; N(d_-(y)) \;+\; n(d_-(y))\,\hat\sigma'(y) \;. \tag{12}$$

This in turn is clear, since the cumulative density is just $\int_0^K dS_T\,\rho_K(S_T) \;\equiv\; \partial_K \hat P$. Let us for reference give the expressions for the strike-derivatives of calls and puts, namely,

$$\partial_K \hat C \;=\; n(d_-)\,\hat\sigma'(y) - N(d_-) \tag{13}$$

$$\partial_K \hat P \;=\; n(d_-)\,\hat\sigma'(y) + N(-d_-) \tag{14}$$

Using $N(x) + N(-x) = 1$ implies (12). After some calculation, (10), (11) follow from $\rho_y(y) = \partial_y c(y)$.

[5]Also, the relationship to the parameters θ, ρ, φ in section 4 of [4] is: $\hat\sigma_0 = \sqrt{\theta}$, $s_2 = \hat\sigma_0 \rho\varphi$, $c_2 = \frac{1}{2}\hat\sigma_0^2(1-\rho^2)\varphi^2$.

Definition (Butterfly Arbitrage): There is no butterfly arbitrage if and only if $g(y) \geq 0$ for all $y \in \mathbb{R}$ and $\lim_{y\to\infty} d_+(y) = -\infty$.

Remark: The first requirement is obvious, since all other factors in the implied density (10) are strictly positive. The second condition follows from the requirement that call prices vanish as strike $K \to \infty$; the equivalence is clear from (1). For clarity of we will consider these two conditions separately (and sometimes even just refer to the former as the absence of butterfly arbitrage). Normalizability of the density (10) holds, as seen from (12), under very mild smoothness assumptions on $y \mapsto \hat{\sigma}(y)$, guaranteeing that $n(d_-(y))\,\hat{\sigma}'(y) \to 0$ as $y \to \infty$, that hold for all volatility curves one might consider in practice.

Note that equations (10), (11) provide a somewhat magical recipe: One can plug in "any reasonable" volatility curve and out comes a (normalizable, non-negative) probability density. Of course, there is no magic: The requirement that $g(y) \geq 0$ for all y is in general very non-trivial to satisfy.

We now want to rewrite the density factor $g(y)$ for the case where the implied variance is written in terms of a shape function $f(z)$ of normalized strike and the normalized volatility, i.e. for the case that

$$w(y) = \hat{\sigma}_0^2 \ f(z) \ . \tag{15}$$

Then $w'(y) = \hat{\sigma}_0 f'(z)$ and $w''(y) = f''(z)$, so that

$$g(y) = \left(1 - \frac{z\,f'(z)}{2f(z)}\right)^2 \ - \ \frac{1}{4}\,\frac{f'(z)^2}{f(z)} \ - \ \frac{\hat{\sigma}_0^2}{16}\,f'(z)^2 \ + \ \frac{1}{2}\,f''(z) \ =: \ g(z) \ . \tag{16}$$

As indicated, by a slight abuse of notation we can now consider the density factor to be a function of z rather than y. Note that the dependence on $\hat{\sigma}_0$ is in exactly one term in (16), that is strictly negative and decreasing in $\hat{\sigma}_0$ for the non-flat case (except for z with $f'(z) = 0$). This simple monotonic dependence on $\hat{\sigma}_0$ only holds if we consider shape functions of normalized strike. It is important for the emergence of the simple picture of the no-arbitrage region we will describe later.

For the specific case of the SSVI/S3 curve, the question is for which values of $\hat{\sigma}_0, s_2, c_2$ does $g(z)$ stay non-negative for all values of z. To get a better understanding of this problem, we will first consider several limiting cases; some are applicable to any volatility curve ($|z| \to 0$, $|z| \to \infty$), some are specific to the SSVI/S3 curve (vanishing c_2 or s_2), before considering the general case.

3.2 Behavior of $d_+(y)$ as $y \to \infty$

Thinking of d_+ as a function of z, a simple calculation shows that for an arbitrary volatility curve whose shape function has the asymptotic behavior $f(z) = C_+ z + \mathcal{O}(1)$ as $z \to \infty$

$$d_+ \ = \ \left(\tfrac{1}{2}\hat{\sigma}_0\sqrt{C_+} - \frac{1}{\sqrt{C_+}}\right)\sqrt{z} \ + \ \mathcal{O}(1/\sqrt{z}) \ ,$$

which implies

$$\lim_{z\to\infty} d_+ \ = \ \begin{cases} -\infty & \text{for} \quad \hat{\sigma}_0\, C_+ < 2 \\ \ \ 0 & \text{for} \quad \hat{\sigma}_0\, C_+ = 2 \\ +\infty & \text{for} \quad \hat{\sigma}_0\, C_+ > 2 \end{cases} \tag{17}$$

So, for call prices to vanish as strike goes to infinity requires $\hat{\sigma}_0\, C_+ < 2$.[6]

[6] And by the same calculation this conclusion even holds if $f(z) = C_+ z + \mathcal{O}(\sqrt{z})$.

3.3 No-arbitrage at small z

The ATF region does not, usually, provide very strong arbitrage constraints on on the SSVI/S3 curve. But it is still interesting to consider, and there is one result that is generic to any volatility curve that we can mention. Namely, assume the fixed term variance $w(z)$ is twice differentiable at $z=0$, i.e. has an expansion of the form

$$w(z) \;=\; \hat{\sigma}_0^2\,(1+s_2 z+\tfrac{1}{2}c_2 z^2+\ldots)\ ,$$

where the higher order terms do not contribute to the second derivative $w''(0)$ at $z=0$ (so they do not necessarily have to be $\mathcal{O}(z^3)$ terms). All volatility curves one should reasonably consider in practice have this property, we suspect. Substituting this into (16) then gives

$$g(z\to 0) \;=\; 1+\tfrac{1}{2}c_2-\tfrac{1}{4}s_2^2\,(1+\tfrac{1}{4}\hat{\sigma}_0^2)\ . \tag{18}$$

Non-negativity of $g(0)$ gives an upper bound on the ATF slope given $\hat{\sigma}_0$ and c_2:

$$s_2^2 \;\le\; \frac{4+2c_2}{1+\frac{1}{4}\hat{\sigma}_0^2}\ . \tag{19}$$

Note that this bound on the slope *increases* with curvature c_2, and decreases with volatility.

A slightly more interesting result is the bound it provides on the curvature. Namely, whereas for SSVI/S3 the curvature c_2 has to be non-negative, we know that empirically this is not the case for the implied volatility curves of stocks for maturities right after e.g. earnings dates, which can have a pronounced W-shape (especially for technology names like AAPL, GOOG, NFLX, etc). These volatility curves have to be parameterized in more general ways to allow $c_2<0$, which in turn allows for bi-modal implied densities (see [5] for details).

The constraint from $g(0)\ge 0$ can then most illuminatingly be written in terms of the ATF curvature $c_1=\frac{1}{2}c_2-\frac{1}{4}s_2^2$ of the volatility shape (see our earlier footnote) as

$$c_1 \;\ge\; -1 \;+\; \tfrac{1}{16}\,s_2^2\,\hat{\sigma}_0^2\ . \tag{20}$$

The last term tends to be very small for liquid terms, so this is essentially saying that even though the ATF curvature can go negative, there is a simple lower bound $c_1>-1$ to avoid butterfly arbitrage in the ATF region.

3.4 No-arbitrage at large $|z|$

The first thing to note about this limit is that it is not uniform in $|z|$ and the parameters $\hat{\sigma}_0, s_2, c_2$. Namely, if $s_2=0=c_2$, the Black-Scholes case, then $g(z)\equiv 1$ for all z. But if $s_2\ne 0$ or $c_2>0$, then

$$g(z) \;=\; \tfrac{1}{4}\,(1-\tfrac{1}{4}\hat{\sigma}_0^2C_\pm^2) \;-\; \frac{1}{4|z|}\left(C_\pm-\frac{1}{\sqrt{\frac{1}{4}s_2^2+\frac{1}{2}c_2}}\right) \;+\; \mathcal{O}(z^{-2}) \quad \text{as} \quad z\to\pm\infty \tag{21}$$

If after taking $|z|\to\infty$ we take $s_2\to 0, c_2\to 0$, in any order and along any path, or $\hat{\sigma}\to 0$, we have $g(\pm\infty)=\frac{1}{4}$.

From the above we see that if $\hat{\sigma}_0C_\pm<2$ holds, there will be no arbitrage at sufficiently large $|z|$ (the sign of the $\mathcal{O}(1/z)$ term can be positive or negative, so we need a strict inequality in general). There will be, if $\hat{\sigma}_0C_\pm>2$. In other words, we recover the Lee bounds, as expected.

3.5 SSVI/S3 no-arbitrage conditions for $c_2 = 0$

Theorem 1: When $c_2 = 0$, the SSVI/S3 curve has no butterfly arbitrage if and only if $|s_2| \leq s_2^\star(\hat\sigma_0)$, where

$$s_2^\star(\hat\sigma_0)^2 \;:=\; \begin{cases} 4 - \hat\sigma_0^2 & \text{for} \quad \hat\sigma_0^2 \leq 2 \\ 4/\hat\sigma_0^2 & \text{for} \quad \hat\sigma_0^2 \geq 2 \end{cases}$$

Proof: If $c_2 = 0$ the variance has a hockey-stick shape and

$$\begin{aligned} f(z) &= (1 + s_2 z)_+ && (22)\\ f'(z) &= s_2\, \theta(1 + s_2 z) && (23)\\ f''(z) &= s_2^2\, \delta(1 + s_2 z) && (24) \end{aligned}$$

in terms of the Heaviside step function θ and the Dirac delta function δ. Without loss of generality we can assume $s_2 > 0$. Technically, $g(z)$ is only well-defined for $c_2 = 0$ when $1 + s_2 z > 0$. But it is clear what happens for $1 + s_2 z \leq 0$, in the flat region of the hockey stick: The volatility and implied density $\rho_y(y)$ vanishes. Any kind of sensible, regularized definition of $g(z)$ via a suitable $c_2 \to 0$ limit will lead to vanishing density $\rho_y(y) = 0$ when $1 + s_2 z \leq 0$, due to the rapid exponential decay of $n(d_-(y))$ in (10). Note that this is despite the fact that $g(z)$ diverges quadratically when $z \to -1/s_2$ from above (see below).

When looking for potential arbitrage we can therefore assume $1 + s_2 z > 0$ from now on. We know there that $g(z) \equiv 1$ if $s_2 = 0$, and we established the conditions under which $g(z) > 0$ in some region close to $z = 0$. We will now find the smallest $s_2 = s_2^\star(\hat\sigma_0) > 0$ where $g'(z)$ has a zero $z^\star$ with $1 + s_2 z^\star > 0$. We then know that $g(z) \geq 0$ for all relevant z and $s_2 \leq s_2^\star$. We will also show that $g(z) < 0$ somewhere if $s_2 > s_2^\star$.

Under our assumptions we have

$$g(z) \;=\; \left(1 - \frac{\frac{1}{2} s_2 z}{1 + s_2 z}\right)^2 \;-\; \tfrac{1}{4}\,\frac{s_2^2}{1 + s_2 z} \;-\; \tfrac{1}{16}\,\hat\sigma_0^2\, s_2^2\;. \tag{25}$$

We see that to avoid $g(z) < 0$ for any sufficiently large $|z|$ requires $\hat\sigma_0 |s_2| < 2$, which of course is just a special case of $\hat\sigma_0\, C_\pm < 2$ from section 3.4. Next, we have

$$g'(z) \;=\; \tfrac{1}{4}\, s_2\, \frac{s_2^2 - 4 - s_2 z\,(2 - s_2^2)}{(1 + s_2 z)^3}\;.$$

It has the unique zero

$$z^\star = -\frac{4 - s_2^2}{s_2\,(2 - s_2^2)}\;,$$

for which $1 + s_2 z^\star = -2/(2 - s_2^2)$, and

$$g(z^\star) \;=\; \tfrac{1}{16} s_2^2\, \left(4 - (s_2^2 + \hat\sigma_0^2)\right)\;.$$

So, for $s_2^2 \leq 2$ the above zero is outside the relevant domain, and only $\hat\sigma_0 |s_2| < 2$ is necessary to avoid arbitrage. For $2 < s_2^2 \leq 4$, on the other hand, we have a relevant zero of $g'(z)$, and $g(z^\star) < 0$ if $s_2^2 > 4 - \hat\sigma_0^2$. Conversely, $g(z^\star) \geq 0$ if $s_2^2 \leq 4 - \hat\sigma_0^2$. Recall from section 3.3 that if $s_2^2 > 4$ there is arbitrage, since then $g(0) < 0$. $\square$

3.6 SSVI/S3 no-arbitrage conditions for $s_2 = 0$

Theorem 2: When $s_2 = 0$, the SSVI/S3 curve has no butterfly arbitrage if and only if $c_2 \le c_2^\star(\hat\sigma_0)$, where

$$c_2^\star(\hat\sigma_0) \;:=\; \begin{cases} \frac{5-\frac{1}{8}\hat\sigma_0^2}{\left(1-\frac{1}{8}\hat\sigma_0^2\right)^2+\hat\sigma_0^2} \;+\; \sqrt{\left(\frac{5-\frac{1}{8}\hat\sigma_0^2}{\left(1-\frac{1}{8}\hat\sigma_0^2\right)^2+\hat\sigma_0^2}\right)^2 \;-\; \frac{1}{\left(1-\frac{1}{8}\hat\sigma_0^2\right)^2+\hat\sigma_0^2}} & \text{for} \quad \hat\sigma_0^2 \le 4 \\[2ex] 8/\hat\sigma_0^2 & \text{for} \quad \hat\sigma_0^2 \ge 4 \end{cases}$$

Proof: When $s_2 = 0$ we have a symmetric volatility curve in NS-space with

$$f(z) \;=\; \tfrac{1}{2}\,(1+\sqrt{1+2c_2\,z^2}) \tag{26}$$

$$f'(z) \;=\; \frac{c_2\,z}{\sqrt{1+2c_2\,z^2}} \tag{27}$$

$$f''(z) \;=\; \frac{c_2}{\sqrt{1+2c_2\,z^2}^{\,3}} \tag{28}$$

We now want to find the conditions under which $g(z)$ does or does not become strictly negative. Because of the $z \to -z$ symmetry we can restrict ourselves to positive, say, z. We can also assume $c_2 > 0$. It is then convenient to think of $g(z)$ in terms of the new variable $q := \sqrt{1+2c_2z^2}$, which is strictly monotonically related to z, and satisfies $q \ge 1$. In other words, we are replacing

$$z^2 \to z(q)^2 \;=\; \frac{q^2-1}{2c_2}\;.$$

Substituting all this into (16) gives, after some algebra

$$g(z(q)) \;=\; \frac{q+1}{4\,q^3}\left[q^2\,(1-\alpha) \;+\; q\,(1+\alpha-c_2) \;+\; 2c_2\right], \tag{29}$$

where for convenience we have introduced

$$\alpha \;:=\; \alpha(c_2,\hat\sigma_0^2) \;=\; \tfrac{1}{8}\,c_2\,\hat\sigma_0^2\;.$$

We now have to analyze the quadratic form in q in the square brackets in (29); the factor multiplying it always positive. Let us give it a name

$$Q(q) \;:=\; q^2\,(1-\alpha) \;+\; q\,(1+\alpha-c_2) \;+\; 2c_2\;.$$

To start, it is clear: $Q(q)$ will go negative (for sufficiently large q) when $\alpha > 1$, and also for $\alpha = 1$ if and only if $c_2 > 2$. Note that the condition for call prices to vanish for $z \to \infty$ reduces to $\alpha < 1$ here.

Now assume $\alpha < 1$. Clearly $Q(q) > 0$ for small c_2, α for any q. To see at what $c_2^\star = c_2^\star(\hat\sigma_0) > 0$ arbitrage develops we need to find the zeros of $Q(q)$ with $q \ge 1$. Its roots are

$$q_\pm^\star \;=\; -\tfrac{1}{2}\,\frac{1+\alpha-c_2}{1-\alpha} \;\pm\; \tfrac{1}{2}\sqrt{\left(\frac{1+\alpha-c_2}{1-\alpha}\right)^2 - \frac{8c_2}{1-\alpha}} \tag{30}$$

For small $c_2 > 0$ the both roots $q_\pm^\star$ are negative, since the first term in $q_\pm^\star$ is negative, and the term under the square root is positive (and less in magnitude than the square of the first). As c_2

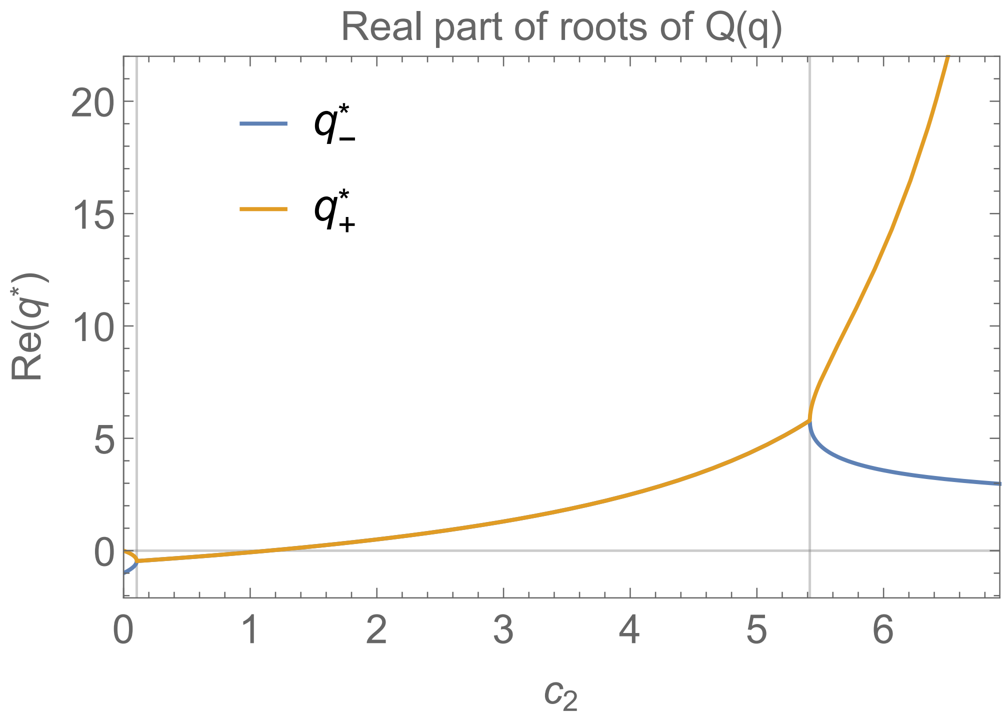

Figure 1: The real part of the roots of the quadratic form $Q(q)$ determining whether there is arbitrage for $s_2 = 0$, as a function of c_2, for $\hat{\sigma}_0 = 1$. The vertical lines bound the region where the roots $q^\star_\pm$ are complex conjugates. Beyond the upper vertical line, $c_2 > c_2^\star(\hat{\sigma}_0)$, where the roots are real, butterfly arbitrage exists.

gets larger the first term will at some point go through 0, and $q^\star_\pm$ will be complex conjugates, since the term under the square root is negative. Eventually c_2 will become so large that the first term is positive *and* the term under the square root is 0. (The behavior of the roots $q^\star_\pm$ as a function of c_2 is illustrated in figure 1.) At this point we have a degenerate, second-order positive root, which we will call $q^\star$. This happens at the first $c_2 = c_2^\star$ for which a root $q^\star \geq 1$ can develop, and it is clear that for any larger c_2 we will have $Q(q) < 0$ for any q between the now distinct, real $q^\star_\pm$ (as long as α stays less than 1, see below). Inserting the definition of α, we see that the critical c_2 is determined by

$$c_2^2\,((1-\tfrac{1}{8}\hat{\sigma}_0^2)^2 + \hat{\sigma}_0^2) \;-\; c_2(8 + 2(1-\tfrac{1}{8}\hat{\sigma}_0^2)) \;+\; 1 \;=\; 0$$

with the relevant solution being

$$c_2^\star(\hat{\sigma}_0) \;=\; \frac{5-\frac{1}{8}\hat{\sigma}_0^2}{\left(1-\frac{1}{8}\hat{\sigma}_0^2\right)^2 + \hat{\sigma}_0^2} \;+\; \sqrt{\left(\frac{5-\frac{1}{8}\hat{\sigma}_0^2}{\left(1-\frac{1}{8}\hat{\sigma}_0^2\right)^2 + \hat{\sigma}_0^2}\right)^2 \;-\; \frac{1}{\left(1-\frac{1}{8}\hat{\sigma}_0^2\right)^2 + \hat{\sigma}_0^2}} \tag{31}$$

(The negative sign corresponds to the smallest c_2 above which $q^\star_\pm$ become complex conjugates, and where $q^\star_\pm$ are still negative, i.e. not relevant.) It remains to check whether for this $c_2^\star$ we have $q^\star \geq 1$ and $\alpha < 1$. The answer is yes, as long as $\hat{\sigma}_0^2 < 4$. As $\hat{\sigma}_0^2 \to 4$, we have $\alpha(c_2^\star(\hat{\sigma}_0), \hat{\sigma}_0) \to 1$, i.e. $c_2^\star(\hat{\sigma}_0) \to 2$, and $q^\star \to \infty$. For $\hat{\sigma}_0^2 > 4$ we have $q^\star < 0$.

So, for $\hat{\sigma}_0^2 \geq 4$ only the bound $\alpha \leq 1$ or $c_2 \leq 8/\hat{\sigma}_0^2$ is relevant (strict $<$ if we want call prices to vanish for large strikes), but for smaller $\hat{\sigma}_0^2$ the above $c_2^\star(\hat{\sigma}_0)$ provides the boundary of the no-arbitrage region. $\square$

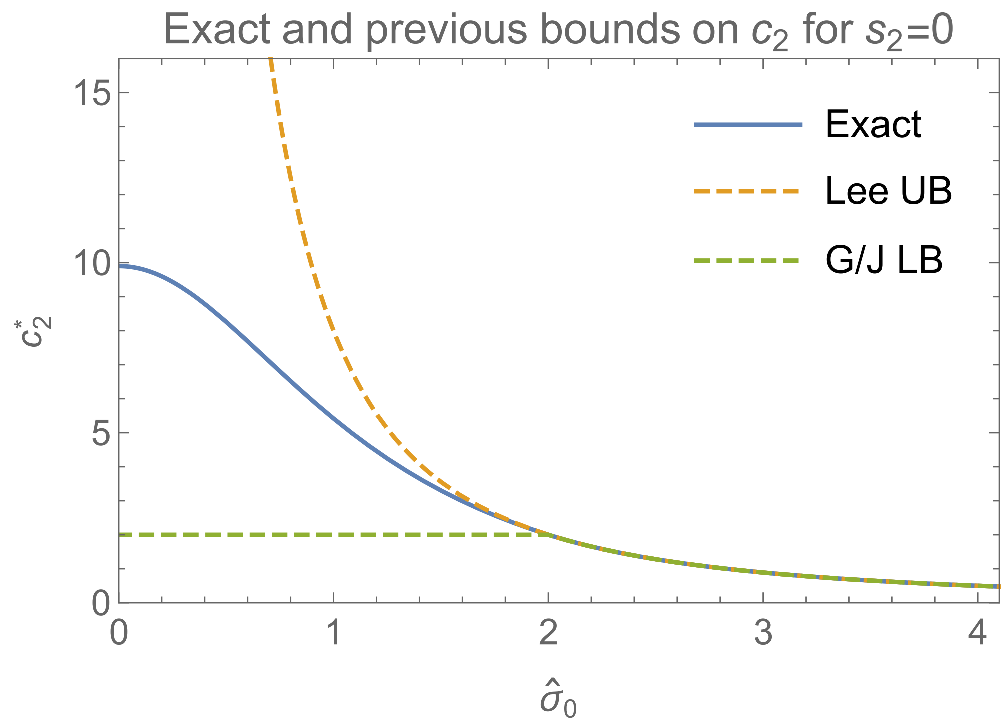

Figure 2: The exact no-arbitrage boundary $c_2(\hat{\sigma}_0)$ for $s_2 = 0$ from Theorem 2 compared with previous upper (UB) and lower (LB) bounds from [1], respectively [4].

Remark: The behavior of the exact $c_2^\star(\hat{\sigma}_0)$ and the upper bound (necessary, not sufficient) based on the wing asympotics [1], and the lower bound (sufficient, not necessary) of Gatheral and Jacquier [4], are illustrated in figure 2.

3.7 General Case: $s_2 \neq 0$, $c_2 > 0$

In the general case we have to analyze a function (proportional to) $g(z)$ which is a combination of linear, quadratic and a square root of quadratic terms in z, s_2 and c_2, for given $\hat{\sigma}_0$, to find when it develops zeros. Just possibly, a combination of clever and tedious calculations might lead to some clear (semi-)analytic results. We have not found a way to do this (yet). What is obvious from (16) is that for fixed z, s_2, c_2, the function $\hat{\sigma}_0 \mapsto g(z|\hat{\sigma}_0, s_2, c_2)$ is strictly decreasing in $\hat{\sigma}_0$, as long as the curve is not constant (i.e. the non-Black-Scholes case).[7] This implies that the no-arbitrage region will strictly shrink with increasing $\hat{\sigma}_0$.

Otherwise, we here just note that it is straightforward to analyze the no-arbitrage region numerically. For given $\hat{\sigma}_0$, loop over $c_2 \in (\,0, c_2^\star(\hat{\sigma}_0)\,)$ and find for each c_2 the "first" value (lim inf) of $s_2 > 0$ where $g(z)$ goes negative for some z. Numerically it is clear that arbitrage then develops for any $|s_2|$ above this critical value. An analytical proof is left for future research.[8]

[7]Except for the one value of z where the curve has its minimum. But the point(s) z where arbitrage occurs can never have $f'(z) = 0$, since the latter implies $g(z) = 1 + \frac{1}{2} f''(z) \geq 1$, from (16).

[8]This does not seem to follow from some generic monotonicity property of $g(z)$ for the SSVI/S3 case. For example $\partial_{s_2} g(z, \hat{\sigma}_0, s_2, c_2) < 0$ does not hold for all $z, \hat{\sigma}_0, c_2$. Nor is this true for all z when $\hat{\sigma}_0, s_2, c_2$ are on the arbitrage boundary. It seems to generically only hold on the arbitrage boundary for z around the point where $g(z)$ goes negative.

This leads to the no-arbitrage boundaries illustrated in figure 3. For comparison we have also included earlier results. First, the sufficient but not necessary condition from [4] (condition 2 of Theorem 4.2) reads, in our notation $(C_+ + C_-)\, C_\pm \le 2$. Expressed as explicit maximal values of one shape parameter in terms of another, they read

$$s_2^\star(c_2) \;=\; (2-c_2)\Big/\sqrt{2-\tfrac{1}{2}c_2} \qquad \text{with} \quad 0 \le c_2 \le 2 \tag{32}$$

$$c_2^\star(s_2) \;=\; 2-\tfrac{1}{4}s_2^2 - |s_2|\sqrt{1+\tfrac{1}{16}s_2^2} \qquad \text{with} \quad 0 \le |s_2| \le \sqrt{2} \tag{33}$$

Note that they do not depend on volatility.

Second, the necessary but generally not sufficient constraints from the asymptotic wing behavior [1], $\hat\sigma_0\, C_\pm \le 2$, correspond to the following straight line bounds on the parameters

$$\text{Asymptotic Wing Constraints:} \qquad c_2^\star \;=\; \frac{8}{\hat\sigma_0^2} - \frac{4|s_2|}{\hat\sigma_0} \qquad \text{or} \qquad |s_2^\star| \;=\; \frac{2}{\hat\sigma_0} - \frac{c_2\hat\sigma_0}{4} \tag{34}$$

We conclude this section with a couple of remarks:

- There is a curious feature in the exact no-arbitrage boundary in figure 3, namely the "bulge" on the rhs or small c_2 and $\hat\sigma_0$. In this situation the arbitrage beyond the boundary will occur at small $|z|$, in which case the results from (19) tell us that the maximal allowed value of $|s_2|$ increases with curvature.

- As already illustrated in the figure, but perhaps worthwhile pointing out explicitly, the no-arbitrage boundary has a well-defined $\hat\sigma_0 \to 0$ limit, as is clear from (16). In this limit $s_2^\star(0) = 2$ and $c_2^\star(0) = 5+\sqrt{24} \approx 9.898979$. In practice, most liquid options have $\hat\sigma_0 \ll 1$, so one is often not far from this limit. Is is perhaps surprising how large the no-arbitrage region is in this situation. In practice, with a proper fitting framework, one virtually never sees cases where one is close to the SSVI/S3 arbitrage boundary.[9]

- As mentioned earlier, for call prices to vanish at infinite strike we need the strict $|s_2| < s_2^\star(\hat\sigma_0)$ for $c_2\!=\!0$, $\hat\sigma_0^2 \ge 2$, and $c_2 < c_2^\star(\hat\sigma_0)$ for $s_2\!=\!0$, $\hat\sigma_0^2 \ge 4$.

- Perhaps most importantly, the exact results in Theorem 1 and 2 combined with the numerical results summarized in this figure lead to a simple sufficient (and close to necessary) no-arbitrage constraint for given $\hat\sigma_0$, obtained by connecting the limiting points on the $s_2 = 0$ and $c_2 = 0$ axes with a straight line, namely:

  $$\text{No arbitrage for:} \qquad \frac{|s_2|}{s_2^\star(\hat\sigma_0)} + \frac{c_2}{c_2^\star(\hat\sigma_0)} \;\le\; 1 \tag{35}$$

  For $\hat\sigma_0 \ge 2$ the straight line defined by (35) actually becomes the exact no-arbitrage boundary, as can be checked numerically to high precision (we do not have an analytical proof at present). Using the explicit results for $s_2^\star(\hat\sigma_0)$ and $c_2^\star(\hat\sigma_0)$ for this case, it is clear that this line is identical to (34). In other words, for $\hat\sigma_0 \ge 2$ the asymptotic wings constraints are not just necessary but also sufficient for no-arbitrage everywhere.

[9] The present author has fitted thousands of names in real-time with this curve over many years, and we are not aware of encountering a single (butterfly) arbitrage situation (we checked for this in various ways, in production and in overnight replays on captured real-time option prices).

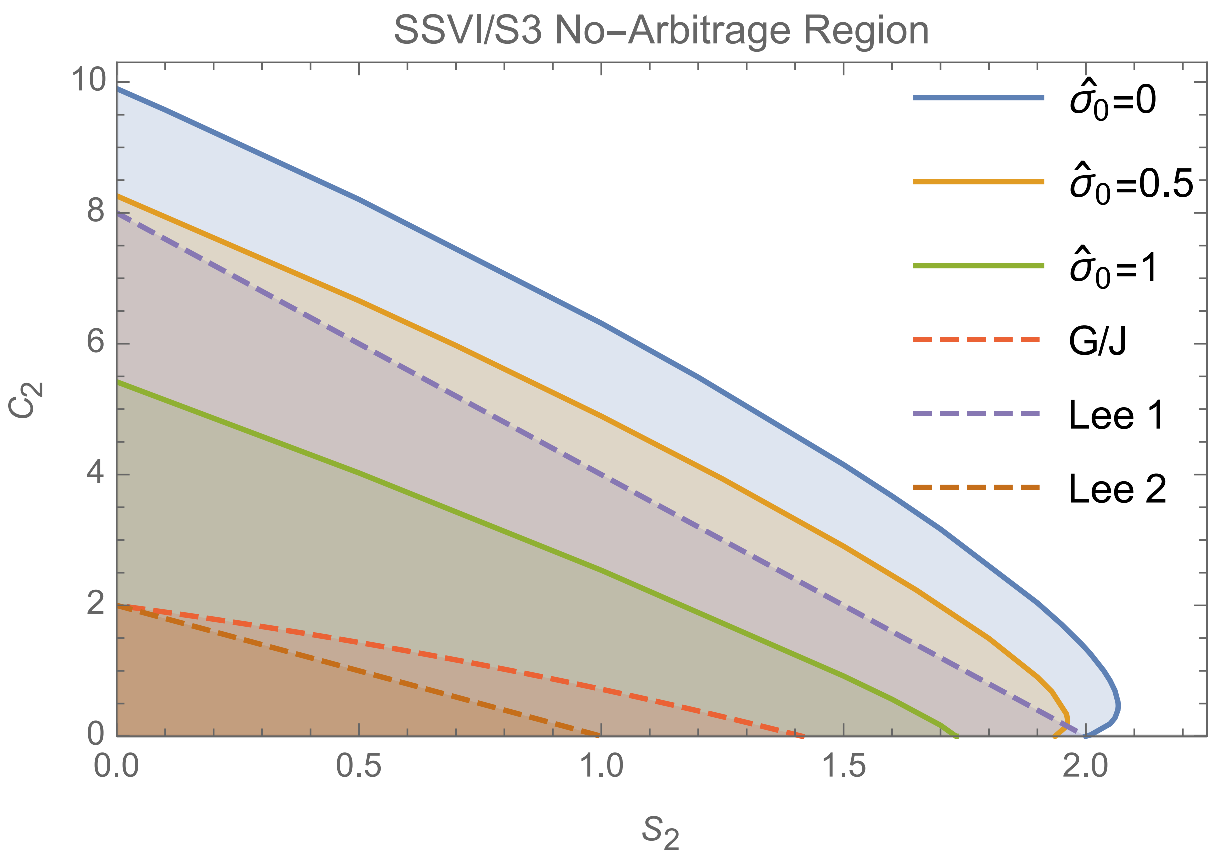

Figure 3: The exact no-butterfly-arbitrage regions of the SSVI/S3 curve in the $(|s_2|, c_2)$-plane for various $\hat{\sigma}_0$, as well as condition 2 from Theorem 4.2 of [4], and the regions in which there is no arbitrage in the asymptotic wings for $\hat{\sigma}_0 = 1, 2$ (labelled "Lee 1, 2"). For $\hat{\sigma}_0 \geq 2$ the exact no-arbitrage boundary becomes identical to the straight line asymptotic wing constraint (i.e. "Lee 2" for $\hat{\sigma}_0 = 2$).

4 Conclusion and Outlook

We have described the exact no-butterfly-arbitrage boundary of the 3-parameter SSVI/S3 volatility curve, sharpening the upper and lower bounds established in [1, 4]. When working with normalized strike space parameters, a simple picture emerges, summarized in Theorems 1, 2, and figures 2 and 3. The no-arbitrage region is surprisingly large for most realistic volatilities, explaining the excellent no-arbitrage behavior of this curve observed in practice in an (equity) implied volatility fitting context. This should encourage further use of this curve in practice[10] – we like to think of it as the "null hypothesis" of implied volatility curve fitting. This is true for all asset classes, as the defining characteristics of SSVI/S3 are motivated by simplicity and avoiding arbitrage, which should hold for all underliers.

However, we should point out that the volatility curves for options on more liquid underliers, at least in the equity domain, can have significantly more structure and require more degrees of freedom to be consistently fit within bid-ask spreads [5]. For example, the curvature c_2 can become significantly negative for maturities right after important events (earnings for technology names, FOMC for the SPY ETF, etc). This is a qualitative feature not allowed by the SSVI/S3 or SVI curves.

One general lesson from our results, we think, is that it is useful to work in normalized strike space when designing and handling volatility curves and surfaces. This will be discussed in more detail in [5].

On the mathematical side, we should mention that we have not been able to prove all our results analytically, leaving room for further research.

Acknowledgements: I would like to thank Peter Carr, Jim Gatheral, and Jiri Hoogland for discussions and comments on the manuscript.

References

[1] R. Lee, *The Moment Formula for Implied Volatilities at Extreme Strikes*, Mathematical Finance, **14**, 469–480, 2004.

[2] J. Gatheral, *The Volatility Surface*, Wiley, 2006.

[3] J. Gatheral, *A parsimonious arbitrage-free implied volatility parameterization with application to the valuation of volatility derivatives*, Presentation at Global Derivatives, 2004.

[4] J. Gatheral, A. Jacquier, *Arbitrage-Free SVI Volatility Surface*, Quantitative Finance, **14(1)**, 59–71, 2014.

[5] T.R. Klassen, *A General and Practical Volatility Curve Framework*, in preparation, 2016.

[6] T.R. Klassen, *Pricing Vanilla Options with Cash Dividends*, available at ssrn, 2015.

[7] D.T. Breeden, R.H Litzenberger, *Prices of State-Contingent Claims Implicit in Option Prices*, Journal of Business **51(4)**, 621–651, 1978.

[10] For example, any other 3-parameter volatility curve with bad wing behavior can be "fixed" by mapping it to SSVI/S3, e.g. by matching their second-order Taylor series.