Discrete Comput Geom 1:123–140 (1986)

Discrete & Computational Geometry
© 1986 Springer-Verlag New York Inc.

# Smooth, Easy to Compute Interpolating Splines*

John D. Hobby

Computer Science Department, Stanford University, Stanford, CA 94305

**Abstract.** We present a system of interpolating splines with first-order and approximate second-order geometric continuity. The curves are easily computed in linear time by solving a diagonally dominant, tridiagonal system of linear equations. Emphasis is placed on the need to find aesthetically pleasing curves in a wide range of circumstances; favorable results are obtained even when the knots are very unequally spaced or widely separated. The curves are invariant under translation, rotation, and scaling, and the effects of a local change fall off exponentially as one moves away from the disturbed knot.

Approximate second-order continuity is achieved by using a linear "mock curvature" function in place of the actual endpoint curvature for each spline segment and choosing tangent directions at knots so as to equalize these. This avoids extraneous solutions and other forms of undesirable behavior without seriously compromising the quality of the results.

The actual spline segments can come from any family of curves whose endpoint curvatures can be suitably approximated, but we propose a specific family of parametric cubics. There is freedom to allow tangent directions and "tension" parameters to be specified at knots, and special "curl" parameters may be given for additional control near the endpoints of open curves.

## 1. Introduction

The problem of fitting a smooth curve through a set of points on the plane has many important applications in computer graphics, computer-aided design, and typesetting. Often there is no preexisting curve to approximate except possibly a freehand drawing, and the only requirement is to find an aesthetically pleasing

*This research was supported in part by the National Science Foundation under grants IST-820-1926 and MCS-83-00984 and by the Systems Development Foundation.

curve that the computer can easily manipulate. For some interactive applications the curves can be controlled by manipulating points that do not lie on the curve, but many applications require the control points to lie on the curve. For example, the control points may be obtained by digitizing key points on a drawing, or there may be a priori knowledge that the curve must pass through certain points.

Suppose the curve must pass through points $z_0, z_1, \ldots, z_n$; and either $z_0$ and $z_n$ are to be the endpoints of the curve or $z_0 = z_n$ and the curve is to be a closed loop. Optionally, there may be direction vectors $w_i$ specifying the slope of the curve at some $z_i$. For example, some of the $z_i$ may have been selected as vertical extrema so that the curve must pass through them horizontally. It is desirable for the curve to be invariant under translation, rotation, and scaling in the sense that if $T$ is such a transformation, then applying $T$ to the computed curve should yield the same result as computing a new curve through $Tz_i$ for $0 \le i \le n$ with direction vectors $Tw_i - T(0,0)$.

The curve should have at least approximate continuity of slope and curvature at knots where no directions are given, and it would also be desirable to have some notion of *extensibility* and *locality*. A system of splines is extensible if the curve generated from knots $z_0, z_1, \ldots, z_n$ is identical to that generated from knots $z'_0, z'_1, \ldots, z'_{n+1}$ where $z'_i = z_i$ for $i < k$, $z'_i = z_{i-1}$ for $i > k$, and $z'_k$ is on the curve segment joining $z_{k-1}$ and $z_k$. In other words, adding a new knot already on the curve must not change it. In practice it is extremely difficult to achieve exact extensibility. The only well-known extensible spline family is the "curve of least energy" that minimizes the integral of squared curvature with respect to arc length [5, 9], but this curve is difficult to work with. It is interesting to note that when the knots are nearly collinear, the curve of least energy approaches the simple nonparametric cubic spline passing through the given knots with continuous second derivative. The splines that we deal with here will share this property.

The concept of locality is that if one of the knots or direction vectors is perturbed, the changes should be confined to a few surrounding spline segments. Specifically, there should be some constant $k$ such that changes to $z_i$ or $w_i$ effect only the curve segments between $z_{i-k}$ and $z_{i+k}$. As the example of Fig. 1 shows, it is difficult to have both locality and continuity of curvature even when the knots are collinear. If the knots $z_0$, $z_1$, $z_2$, and $z_3$ are as shown in the figure but $w_0$ is in the direction of $z_1 - z_0$, then the desired curve is obviously a straight line. if $w_0$ is changed, then locality demands that the spline segments between $z_k$ and $n$ must remain straight, yet there is no way a cubic curve can join such a straight line with continuous curvature.

Most local interpolating splines of moderate complexity do not have continuous curvature. Perhaps the best known splines of this type are the cardinal splines given by a cubic function $z(t)$, where the velocity vector $z'(i)$ at knot $z_i$ is $\frac{1}{2}(z_{i+1} - z_{i-1})$. Figure 2 shows an example of cardinal splines in comparison to the splines that will be developed in this paper. The price paid for locality is that Fig. 2(b) has wild discontinuities in curvature while Fig. 2(a) does not.

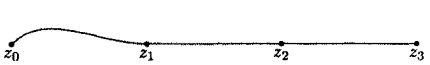


**Fig. 1.** The effect of changing $w_0$ while preserving exact locality.

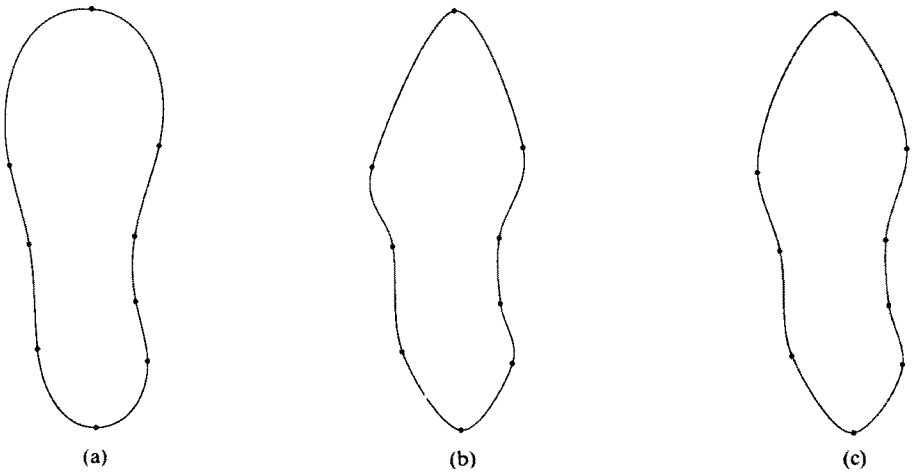


**Fig. 2.** (a) Cubics with mock curvature constraints. (b) Cardinal splines. (c) Barsky–DeRose splines with $\beta_i = \|z_{i+1} - z_i\| / \|z_i - z_{i-1}\|$.

Other well-known local cubic interpolating splines can be viewed as generalizations of cardinal splines. To see this, it is convenient to parameterize the cubic spline in terms of the incoming and outgoing velocity vectors $D_i^-$ and $D_i^+$ at each knot so that the spline segment between $z_i$ and $z_{i+1}$ has Bézier control points $z_i$, $z_i + \frac{1}{3}D_i^+$, $z_{i+1} - \frac{1}{3}D_{i+1}^-$, and $z_{i+1}$. For cardinal splines as described above, $D_i^- = D_i^+ = \frac{1}{2}(z_{i+1} - z_{i-1})$; in the most general form

$$D_i^- = a_i(z_i - z_{i-1}) + b_i(z_{i+1} - z_i) \quad \text{and}$$

$$D_i^+ = c_i(z_i - z_{i-1}) + d_i(z_{i+1} - z_i), \tag{1}$$

where $a_i$, $b_i$, $c_i$, and $d_i$ are parameters that may be used to control the shape of the curve. For instance, Kochanek and Bartels [7] give a scheme where the user can give continuity, tension, and bias parameters for each knot, and these determine $a_i$, $b_i$, $c_i$, and $d_i$. The default settings of these parameters result in $a_i = b_i = c_i = d_i = \frac{1}{2}$, giving simple cardinal splines as shown in Fig. 2(b).

Another approach to local interpolating splines can be found in [3], where DeRose and Barsky give an infinite family of spline curves based on two integer parameters: One parameter gives the desired order of continuity, and the other determines whether the splines interpolate the knots or just pass near them. DeRose and Barsky give explicit equations for cubic interpolating splines of this form with first-order continuity, but their approach requires polynomials of degree at least 5 in order to obtain interpolating splines with continuous curvature. In the cubic interpolating splines derived explicitly in [3], each knot $z_i$ has a single shape parameter $\beta_i$ associated with it. The splines are exactly those

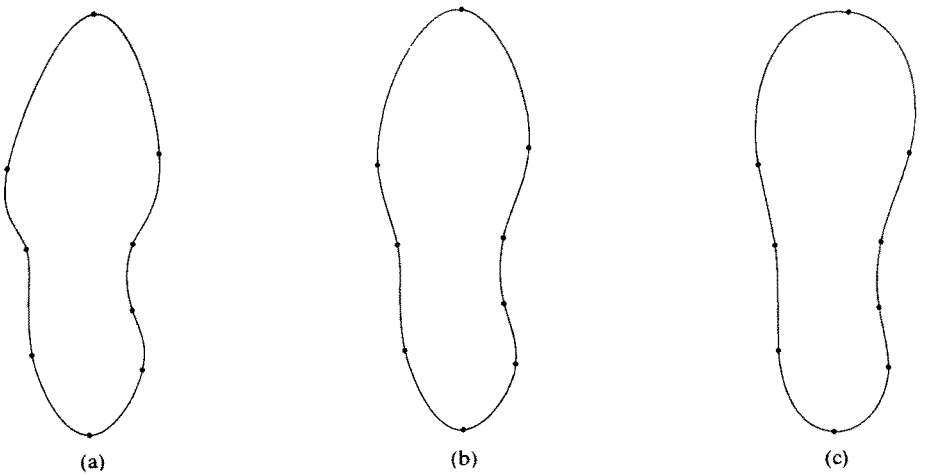


**Fig. 3.** (a) Natural cubics. (b) Cubics with chordal parameterization. (c) Cubics with mock curvature constraints.

produced by (1) where

$$\alpha_i = \frac{\beta_i}{\beta_i+1}, \qquad b_i = \frac{1}{\beta_i(\beta_i+1)}, \qquad c_i = \frac{\beta_i^2}{\beta_i+1}, \quad \text{and} \quad d_i = \frac{1}{\beta_i+1}.$$

The default suggested in [3] is to set all $\beta_i = 1$ so as to obtain cardinal splines as in Fig. 2(b), but somewhat better results can be obtained by setting $\beta_i$ equal to the ratio of Euclidean lengths $\|z_{i+1} - z_i\|/\|z_i - z_{i-1}\|$ as shown in Fig. 2(c).

In this article we shall see how to obtain smother curves by sacrificing locality and settling for an exponential decline in influence of local changes. The best known interpolating splines that behave this way are the $C^2$ continuous parametric cubic splines, sometimes called "natural cubic splines." If no directions are given, there is a unique piecewise parametric cubic, closed curve that is $C^2$ continuous with respect to the parameter and passes through $n$ given points in order. This curve can easily be computed by solving linear equations.

As Epstein shows in [4], natural cubic splines do not perform well for unequally spaced knots because the spacing of parameter values at knots does not reflect the spacing of the knots. Better results can be obtained by setting the parameter at each knot $z_i$ to a value $t_i$, where $t_j - t_{j-1} = \|z_j - z_{j-1}\|$ for $1 \le j \le n$, and requiring second-order continuity with respect to the parameter as shown in Fig. 3(b). This chordal parameterization improves on the uniform parameterization of Fig. 3(a), but the splines that we shall develop still have more gentle curvature in this case as shown in Fig. 3(c).

Figure 3 points out the difference between *geometric* and *parametric* continuity. Natural cubic splines have second-order parametric continuity because they are described by a function $z(t)$ that is $C^2$ continuous. Geometric continuity is a

relaxation of parametric continuity: A spline function $z(t)$ has $k$th-order geometric continuity if there is a continuous, monotonic increasing function $f$ such that $z(f(t))$ is $C^k$ continuous. In other words, geometric continuity is independent of the parameterization.

Requiring first- and second-order continuity with respect to the parameter uses up four degrees of freedom per knot, enough to completely determine a parametric cubic spline as shown in Fig. 3(a). In Fig. 3(b), one of these degrees of freedom is reclaimed and put to better use by altering the parameter spacing, but another degree of freedom can be made available by requiring only continuity of slope and curvature.

Other results on geometric continuity in cubic splines can be found in [1] and [2] where Barsky and Beatty show how two extra degrees of freedom can be obtained for B-splines by requiring only geometric continuity. We need to obtain similar degrees of freedom for interpolating splines, but we shall first concentrate on finding good defaults to work from. The new parameters will be of little value if it is difficult to set them so as to obtain reasonable results. We cannot afford to assume arbitrarily that the best defaults are those that yield some form of parametric $C^2$ continuity.

In [8], Manning takes an interesting approach to this problem. He defines a specific family of curves so that there is a unique one for each pair of initial and final points and directions. He then selects spline directions at each knot so as to achieve geometric continuity. His approach provides a certain degree of locality in that effects of local perturbations do not propagate past knots where a direction is given.

With Manning's approach, both degrees of freedom are available to control the shape of the curve, and defaults can be selected so as to obtain the most pleasing curves. Section 2 explains how to select the defaults by choosing two functions and using them to determine the magnitudes of the velocities at each knot in such a way as to guarantee that the curves generated will be independent of scaling, rotation, and reflection. We can then provide two "tension" parameters for each knot by simply dividing them into these functions. Essentially the same approach would work for other kinds of curves, although there may be more parameters to choose. We select parametric cubics here because they are essentially the simplest curves that can pass through two arbitrary points in two arbitrary directions.

On apparent disadvantage to this approach is the difficulty in solving for the directions that provide continuity of curvature. Manning proposes an iterative approximation scheme that seems to work well in practice, but he admits that there is not always a unique solution and there is no guarantee that the iteration always converges to the desired solution. Cubic splines often have very low curvature at their endpoints when they have very sharp bends internally, and this can introduce extraneous solutions as shown in Fig. 4. The three curves shown are all curvature continuous open curves that have given directions at $z_0$ and $z_2$, but regardless of the initial conditions, Manning's iteration always converges to one of the asymmetrical ones with sharp bends. If $z_0$ is raised and $z_2$ lowered until the angle $z_0z_1z_2$ is about 122°, the asymmetrical solutions merge with the symmetrical ones and the rate of convergence for Manning's iteration approaches zero.

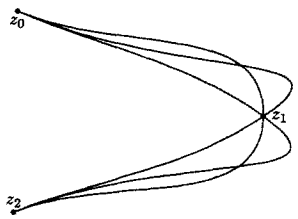


**Fig. 4.** Three splines of the type proposed in [8].

While these kinds of problems do not seem to occur when the angles involved are not so large, much additional testing would be necessary in order to verify this. In Section 3, we show how all such problems can be avoided by setting up a system of linear equations that are easy to solve and guarantee *approximate* continuity of curvature. We derive the specific equations appropriate for the families of curves discussed in Section 2, but similar equations could be derived for many different classes of curves.

## 2. Choosing a Family of Parametric Cubics

The subproblem to be solved in this section can be stated as follows: Given two points $z_i = (x_i, y_i)$ and $z_{i+1} = (x_{i+1}, y_{i+1})$, and two unit vectors $w_i$ and $w_{i+1}$, find an aesthetically pleasing parametric cubic $z(t)$ so that $z(0) = z_i$, $z(1) = z_{i+1}$, $z'(0) = \alpha w_i$, and $z'(1) = \beta w_{i+1}$, where $\alpha$ and $\beta$ are positive real numbers and $z'(t)$ is the componentwise derivative $(x'(t), y'(t))$ of $z(t) = (x(t), y(t))$. We also wish to introduce two "tension" parameters $\tau_i$ and $\bar{\tau}_{i+1}$ such that pleasing curves will be obtained when $\tau_i = \bar{\tau}_{i+1} = 1$, and as the tensions approach $\infty$, the curves will approach the line segment joining $z_i$ to $z_{i+1}$.

In order to guarantee that the results are independent of translation, rotation, and scaling, we shall begin by finding a function $\hat{z}(t)$ such that

$$\hat{z}(0) = (0,0), \qquad \hat{z}(1) = (1,0), \qquad \hat{z}'(0) = \frac{1}{\tau_i}\rho(\theta,\phi)(\cos\theta, \sin\theta), \quad \text{and}$$

$$\hat{z}'(1) = \frac{1}{\bar{\tau}_{i+1}}\sigma(\theta,\phi)(\cos\phi, -\sin\phi), \tag{2}$$

where $\rho$ and $\sigma$ are positive real functions to be determined later, $\theta = \arg w_i - \arg(z_{i+1} - z_i)$, and $\phi = \arg(z_{i+1} - z_i) - \arg w_{i+1}$. We refer to $\rho$ and $\sigma$ as *velocity parameter functions* because they determine the magnitude of the velocity at $t = 0$ and at $t = 1$. (Here $\arg(x, y)$ is the angle $\omega$ such that $(x, y)$ is a positive multiple of $(\cos\omega, \sin\omega)$.) We then set

$$z(t) = \left(\begin{pmatrix} x_i \\ y_i \end{pmatrix} + \begin{pmatrix} x_{i+1} - x_i & y_i - y_{i+1} \\ y_{i+1} - y_i & x_{i+1} - x_i \end{pmatrix} \hat{z}(t)^T\right)^T. \tag{3}$$

It is not hard to see that the parametric cubic satisfying (2) has Bézier control points $(0,0)$, $(\rho/3\tau_i)(\cos\theta, \sin\theta)$, $(1 - (\sigma/3\bar{\tau}_{i+1})\cos\phi, (\sigma/3\bar{\tau}_{i+1})\sin\phi)$, and $(1,0)$,

so that

$$\hat{z}(t) = \frac{\rho}{\tau_i} t(1-t)^2(\cos\theta, \sin\theta) + t^2(1-t)\left(3 - \frac{\sigma\sin\phi}{\bar{\tau}_{i+1}}, \frac{\sigma\sin\phi}{\bar{\tau}_{i+1}}\right) + t^3(1,0). \quad (4)$$

It only remains to choose positive functions $\rho(\theta, \phi)$ and $\sigma(\theta, \phi)$ so that $\rho(\theta, \phi) = \sigma(\phi, \theta) = \rho(-\theta, -\phi)$.

In [8] Manning chooses

$$\rho(\theta, \phi) = \frac{2}{1 + (1-c)\cos\theta + c\cos\phi} \quad \text{and}$$

$$\sigma(\theta, \phi) = \frac{2}{1 + c\cos\phi + (1-c)\cos\theta} \quad (5)$$

and then empirically selects $c = \frac{2}{3}$ to obtain the most pleasing family of curves. Here we shall attempt to do a systematic analysis of the vast range of possible functions to determine whether slightly more complicated velocity parameter functions will yield better results. These functions will have to be evaluated only once for each segment of the spline curve, and they have a strong influence on the final shape.

### 2.1. *Mathematical Measures of Smoothness*

One common way of evaluating the smoothness of curves is to integrate the square of curvature with respect to arc length. For $\hat{z} = (\hat{x}, \hat{y})$

$$\int k^2\, ds = \int_0^1 \frac{(\hat{y}''\hat{x}' - \hat{x}''\hat{y}')^2}{(\hat{x}'^2 + \hat{y}'^2)^{5/2}}\, dt. \quad (6)$$

This can be simplified somewhat but it still proves to be analytically intractable. This is not surprising considering the complex behavior of numerical solutions.

Equation (6) is exactly the energy function that the curve of least energy minimizes, but if we restrict $\hat{z}$ to be the cubic spline (4), we can investigate the velocity parameter functions that minimize (6). Actually we should consider the smallest local minimum since (6) approaches 0 as $\rho$ and $\sigma$ approach $\infty$ for fixed $\theta$ and $\phi$.

Unfortunately, numerical integration of $k^2\, ds$ proves to be slow and imprecise, and it would have to be repeated a large number of times in order to get a good idea what the velocity parameter functions $\rho(\theta, \phi)$ and $\sigma(\theta, \phi)$ should be. Instead we shall introduce two other measures of smoothness that behave similarly:

$$\max_{0 \le t \le 1} |k(t)| \quad (7)$$

and

$$\max_{0 \le t \le 1} \left|\frac{dk}{ds}\right|. \quad (8)$$

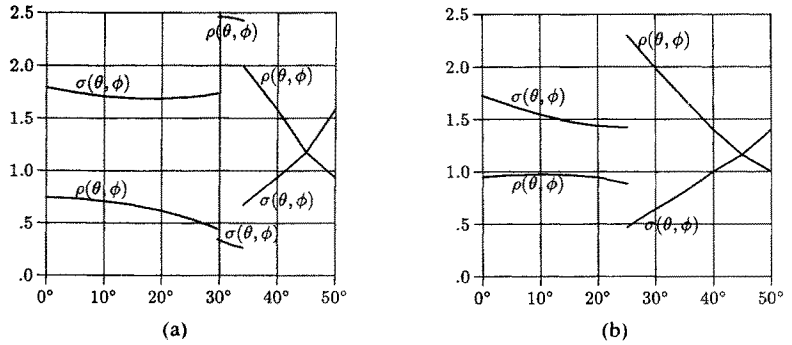


**Fig. 5.** The functions $\rho(\theta, \phi)$ and $\sigma(\theta, \phi)$ versus $\theta$ for $\phi = 90° - \theta$, minimizing (a) the magnitude of curvature and (b) the magnitude of curvature change.

Using (7) to measure smoothness corresponds to taking the $\infty$-norm of curvature instead of the 2-norm; using (8) gives roughly similar results but applies a greater penalty to short periods of relatively high curvature. The velocity parameter functions that minimize (8) turn out to be somewhat better behaved than those that minimize (7), but the overall character is the same for both measures of smoothness.

For fixed $\theta$ and $\phi$, the smoothness measures can have multiple local minima and the relative smoothness at the local minima can change as $\theta$ and $\phi$ change. Therefore, it should not be surprising that the "optimum" velocity parameter functions have large discontinuities where they catastrophically change from one local minimum to another. When this happens there tend to be $(\rho, \sigma)$ values between the two minima that also generate relatively "smooth" curves, so it is not really necessary to use discontinuous functions for $\rho$ and $\sigma$.

Figure 5 illustrates the most basic catastrophe. Near $\theta = \phi = 45°$, the "optimum" $\rho$ increases and the $\sigma$ decreases as $\theta$ decreases. This action tends to reduce the curvature where it is maximum near $t = 1$ without introducing other points of high curvature. When $0 \approx \theta \ll \phi$, the situation is entirely different. The high curvature near $t = 1$ is best controlled by making $\sigma$ large, and increasing $\rho$ beyond what is needed to control the curvature at $t = 0$ just makes the problem worse.

When $\rho$ and $\sigma$ are chosen to minimize (7) as shown in Fig. 5(a), there are actually two catastrophes, and the short segment between them is particularly interesting. Ordinarily, extremely small $\sigma$ values lead to high curvature near $t = 1$, but at $\theta = 34.1°$, $\phi = 55.9°$, the curvature is actually minimized by choosing $\sigma = 0.261$. The choice of $\rho = 2.423$ here is extremely critical. As shown in Fig. 6, this has the effect of making the last three Bézier control points almost collinear, so that the endpoint curvature is not too large in spite of the low velocity. (The bold dots in the figure are the Bézier control points of (4).)

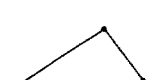

**Fig. 6.** The Bézier control polygon for (4) with $\theta = 34.1°$, $\phi = 55.9°$, where $\rho$ and $\sigma$ are chosen to minimize curvature.

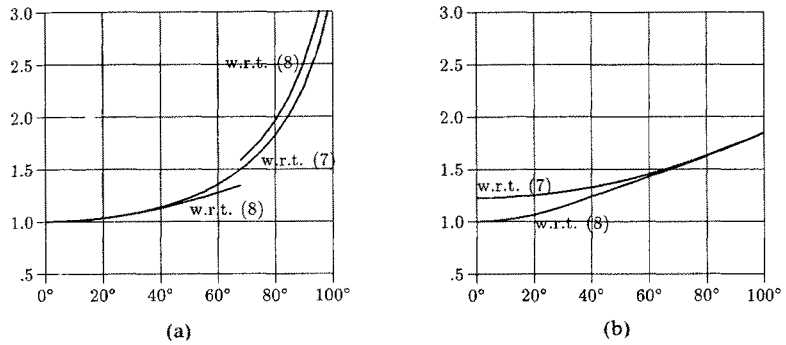


**Fig. 7.** (a) "Optimum" $\rho$ and $\sigma$ versus $\theta$ when (a) $\theta = \phi$ and (b) $\theta = -\phi$ for both smoothness measures.

The bizarre situation shown in Fig. 6 does not occur when minimizing the derivative of curvature, but there is still a catastrophe near 25°. When $\theta + \phi = 90°$ and $\theta < 25°$ the "optimum" cubic has a point of inflection, but when $\theta > 25°$ it has none.

Figure 7 shows how the "optimal" velocity parameter functions grow as $\theta$ and $\phi$ increase. Minimizing (8) produces a catastrophe at 68° where $\rho$ and $\sigma$ increase to about 1.58 and the cubic acquires a single point of maximum curvature at $t = 0.5$. For $\rho = \sigma \approx 1.34$, the maximum curvature occurs at $t \approx 0.08$ and $t \approx 0.92$. Intermediate values for $\rho$ and $\sigma$ produce relatively high change in curvature near $t = 0$ and $t = 1$.

Minimizing (7) instead of (8) avoids the catastrophe at $\theta = \phi = 68°$, but then $\rho$ and $\sigma$ do not approach unique limits as $(\theta, \phi) \to (0,0)$. Along $\theta = -\phi$ the limit is $\sqrt{6}/2$, while $\rho$ and $\sigma$ approach 1 as $\theta \to 0$ when $\theta = \phi$. Under the approximation $k \approx d^2y/dx^2$, when either (6) or (8) is used as the measure of smoothness, it can be shown that the optimum curves are cubics where $\rho \cos\theta = \sigma \cos\phi = 1$. Thus, it seems reasonable to let $(\rho, \sigma) \to (1,1)$ as $(\theta, \phi) \to (0,0)$.

### 2.2. *Velocity Parameter Functions*

We are now ready to choose velocity parameter functions $\rho(\theta, \phi)$ and $\sigma(\theta, \phi)$ so as to obtain an aesthetically pleasing family of curves from (4). In order to be useful as splines these curves should be fairly easy to compute, and they should respond predictably to changes in $\theta$ and $\phi$. The "optimum" velocity parameter functions illustrated in Figs. 5 and 7 fail on both counts, but they still provide useful guidelines. The catastrophes in the "optimum" velocity parameter functions are not features to be preserved, but rather points where many different possible values for $\rho$ and $\sigma$ are feasible. For instance when $\theta = 28°$ and $\phi = 62°$, good results are obtained for almost any values of $\rho$ and $\sigma$ as long as $\rho + \sigma \approx 2.5$.

The actual choice of velocity parameters is necessarily somewhat arbitrary, and there is a trade-off between "smoothness" and simplicity. Other properties such as approximate extensibility and predictable response to changes are also

important, but empirical studies indicate that these goals also tend to be served by maximizing "smoothness" and smoothing out the catastrophes.

We have already decided that $\rho(0,0) = \sigma(0,0) = 1$. Thus for small angles we can approximate the behavior of the curve of least energy and achieve very good approximate extensibility. For $\theta = \phi$, we should approximate the behavior of the functions shown in Fig. 7(a), but these functions increase too rapidly or large angles: They seem to approach $\infty$ well before $\theta$ and $\phi$ reach 180°. It is convenient to let

$$\rho = \sigma = \frac{2}{1+\cos\theta} \quad \text{for } \theta = \phi \tag{9}$$

so as to obtain good approximations to circles.

Because of the symmetry requirements $\rho(\theta,\phi) = \sigma(\phi,\theta) = \rho(-\theta,-\phi)$, it suffices to choose $\rho$ and $\sigma$ for $0 \le |\theta| \le \phi \le 180°$. Figure 8(a) shows the $\rho(\theta,\phi)$ and $\sigma(\theta,\phi)$ that minimize (8) for $\phi = 80°$. Figure 8(b) shows practical functions $\rho$ and $\sigma$ that smooth out the catastrophes and are consistent with (9). Similar plots for smaller $\phi$ would have $\rho$ and $\sigma$ closer to each other and closer to 1. (The slope discontinuities at 64.7° and 69.6° are due to changes in the relative sizes of extrema in $dk/ds$ at different parts of the cubic curve.)

If complexity is of no concern, we might want to choose $\rho(\theta,\phi)$ and $\sigma(\theta,\phi)$ as follows for $0 < |\theta| \le \phi < \pi$ with angles measured in radians:

$$\begin{aligned}
\rho &= f(\theta,\phi) + \gamma(\phi)\sin\left(\psi_\phi\frac{\theta}{\phi}\right),\\
\sigma &= f(\theta,\phi) - \gamma(\phi)\sin\left(\psi_\phi\frac{\theta}{\phi}\right),\\
f(\theta,\phi) &= \frac{a\alpha^2+\alpha+2c}{\alpha+c\cos\beta+c},\\
\alpha &= (\phi-\theta)\left(\frac{\phi-\theta}{2\phi}\right)^d, \qquad \beta = \frac{\theta+\phi}{2}\left(\frac{2\phi}{\theta+\phi}\right)^e,\\
\gamma(\phi) &= \frac{1.17}{\pi}\phi - 0.15\sin(2\phi),\\
\psi_\phi(x) &= \pi\left(x+(x^2-1)\left(\left(0.32-\frac{\phi}{2\pi}\right)x+0.5-\frac{\phi}{2\pi}\right)\right).
\end{aligned} \tag{10}$$

A least-squares fit of $f$ to $\frac{1}{2}(\rho+\sigma)$ with $\rho$ and $\sigma$ chosen to minimize (8) yielded $a = 0.2678306$, $c = 0.2638750$, $d = 1.402539$, and $e = 0.7539063$. A possible refinement is to require $\rho \le 1.5\sin\phi/\sin(\theta+\phi)$ when $\phi > \frac{1}{2}\pi$ and $-\phi < \theta < 0$ so as to avoid any possibility of generating a curve with a cusp in it. (This only affects the above functions when $\phi > 145°$.)

It is desirable to have a simpler approximation that does not use transcendental functions other than sines and cosines of $\theta$ and $\phi$. One such approximation is

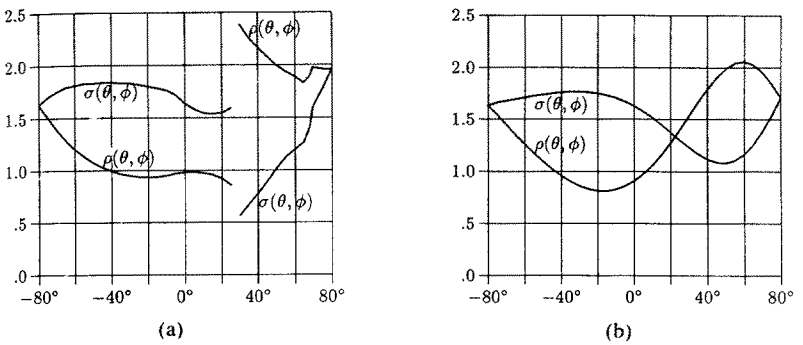


**Fig. 8.** (a) "Optimum" and (b) practical functions $\rho(\theta,\phi)$ and $\sigma(\theta,\phi)$ versus $\theta$ for $\phi = 80°$.

the following functions which were developed for the new METAFONT system [6]:

$$\rho = \frac{2+\alpha}{1+(1-c)\cos\theta + c\cos\phi},$$

$$\sigma = \frac{2-\alpha}{1+(1-c)\cos\phi + c\cos\theta},$$

where

$$\alpha = a(\sin\theta - b\sin\phi)(\sin\phi - b\sin\theta)(\cos\theta - \cos\phi). \tag{11}$$

The constants $a$, $b$, and $c$ were chosen to minimize an error function based on the value of (8) for curves generated by (11) for 116 different $(\theta,\phi)$ pairs. This suggested $a = 1.597$, $b = 0.0700$, and $c = 0.370$, but since empirical evidence indicated that large values for $|\rho - \sigma|$ were causing problems, METAFONT uses the slightly perturbed values $a = \sqrt{2}$, $b = \frac{1}{16}$, and $c = \frac{1}{2}(3-\sqrt{5})$.

Figures 9(a) and (b) show how velocity parameter functions compare in terms of the maximum magnitude of change in curvature. The vertical axis in each graph gives the ratio of the value of (8) for the indicated velocity parameter functions to that for the "optimum" velocity parameter functions. In other words for each value of $\theta$ and $\phi$, the graphs give the ratio of (8) to its minimum possible value. Since (10) and (11) were derived in order to minimize (8), it is not surprising that they perform better than (5) in Figs. 9(a) and (b). We would also expect the more complicated functions defined in (10) to perform better than the simpler versions of (11), and this is indeed the case except near $\theta = -20°$, $\phi = 80°$, where $f(\theta,\phi)$ is a little too small.

Figures 10(a) and (b) correspond to Figs. 9(a) and (b), except that the ratios are based on the maximum magnitude of curvature instead of the maximum magnitude of change in curvature. The graphs show that the curves generated by (10) and (11) do have less curvature than Manning's curves, especially when $\phi$ is fairly large and $\theta < 0$. The figures also show that (7) and (8) do behave similarly in that they both produce the same overall ranking: Curves from (10) are smoother than those from (11) which are smoother than Manning's curves.

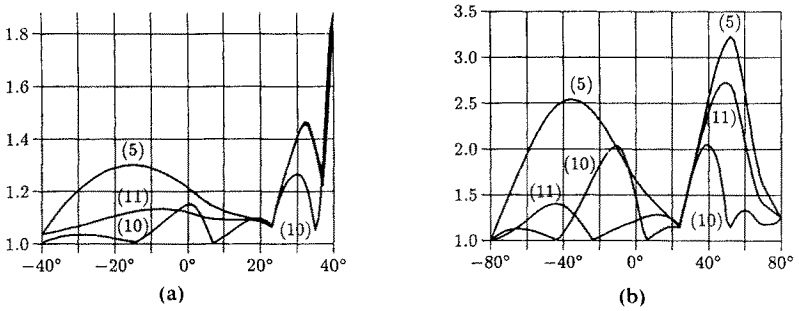


**Fig. 9.** (a), (b) Smoothness ratios based on (8), comparing optimum velocity parameter functions to those of (5), (10), and (11).

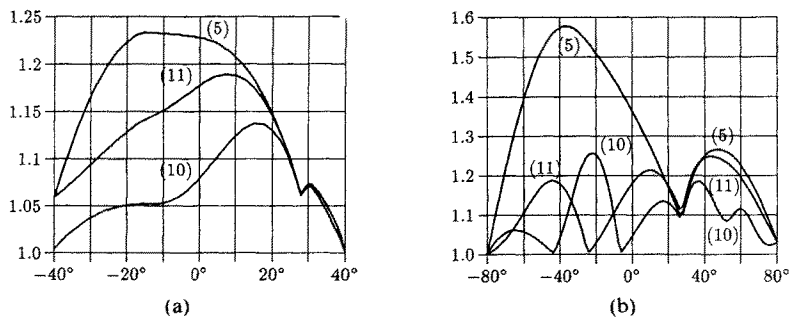


**Fig. 10.** (a), (b) Smoothness ratios based on (7), comparing optimum velocity parameter functions to those of (5), (10), and (11).

Figure 11 shows some of the curves generated by (10) and (11). They are similar for moderate angles, but the simpler equations set $\rho$ too small and $\sigma$ too large when $\phi = -90°$. Equation (11) does not perform well in such extreme cases because it does not allow $\rho - \sigma$ to be large enough when $0 \ll \theta/\phi < 1$ without making $\sigma - \rho$ too large when $-1 < \theta/\phi < 0$ or moving the crossover point where $\rho = \sigma$ too close to $\theta/\phi = 0$.

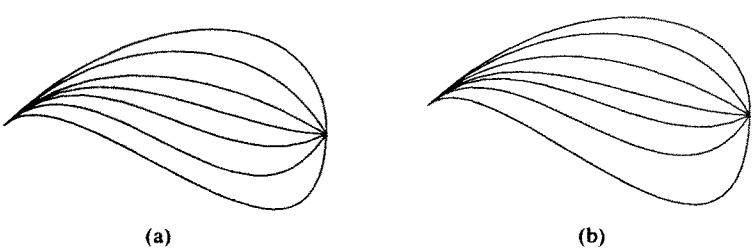


**Fig. 11.** (a) Curves from (10) and (b) curves from (11), both with $\theta = 45°$.

## 3. Mock Curvature Constraints

Here we need to extend the notation of Section 2 so that $\theta_j = \arg w_j - \arg(z_{j+1} - z_j)$ for $0 \le j < n$, $\phi_j = \arg(z_j - z_{j-1}) - \arg w_j$ for $0 < j \le n$, and $d_j$ is the Euclidean length of the vector $z_{j+1} - z_j$ for $0 \le j < n$. If the problem is to find a closed curve with no directions given, it will be convenient to sometimes use alternative names $z_{n+1}$, $\theta_n$, $\phi_{n+1}$, $\tau_n$, and $\bar{\tau}_{n+1}$ for $z_1$, $\theta_0$, $\phi_1$, $\tau_0$, and $\bar{\tau}_1$, respectively. We can then define $\psi_j = \arg(z_{j+1} - z_j) - \arg(z_j - z_{j-1})$ for $1 \le j \le n'$, where $n' = n$ for closed curve problems with no directions given and $n' = n - 1$ otherwise. Unless stated otherwise, all $\psi_j$ are at most 180° in absolute value.

Where $w_i$ has been given in advance, it simply determines $\phi_i$ and $\theta_i$; other $w_i$ need to be determined by solving for $\theta_i$ and $\phi_i$. Since the problem of finding direction angles can be broken into independent subproblems separated by knots where directions are given, we can assume that no directions other than $w_0$ and $w_n$ are given. For closed curve problems we can assume that no directions at all are given, otherwise the problem could be reduced to one or more open curve problems.

The requirement that the curvature be continuous at some knot $z_i$, $0 < i < n$, is

$$k_1(z_{i-1}, z_i, w_{i-1}, w_i, \tau_{i-1}, \bar{\tau}_i) = k_0(z_i, z_{i+1}, w_i, w_{i+1}, \tau_i, \bar{\tau}_{i+1}),$$

where $k_0$ and $k_1$ are functions that give the curvature at $t = 0$ and $t = 1$ in terms of the endpoints, terminal directions, and tension parameters for the family of curves being used. Because of the requirement for invariance under translation, rotation, and scaling, there exists a function $k$ such that

$$k_0(z_j, z_{j+1}, w_j, w_{j+1}, \tau_j, \bar{\tau}_{j+1}) = k(\theta_j, \phi_{j+1}, \tau_j, \bar{\tau}_{j+1})/d_j$$

and

$$k_1(z_j, z_{j+1}, w_j, w_{j+1}, \tau_j, \bar{\tau}_{j+1}) = k(\phi_{j+1}, \theta_j, \bar{\tau}_{j+1}, \tau_j)/d_j. \tag{12}$$

Any particular family of curves determines a specific function $k$ that satisfies (12). The corresponding *mock curvature* function $\hat{k}$ consists of the linear terms in the Taylor series for $k(\theta, \phi, \tau, \bar{\tau})$, expanded about $(\theta, \phi) = (0, 0)$. For the curves determined by (3) and (4) with $\rho$ and $\sigma$ determined by (10) or (11),

$$k(\theta, \phi, \tau, \bar{\tau}) = \frac{2\sigma(\theta, \phi)\sin(\theta + \phi)/\bar{\tau} - 6\sin\theta}{(\rho(\theta, \phi)/\tau)^2}$$

and

$$\hat{k}(\theta, \phi, \tau, \bar{\tau}) = \tau^2\left(\frac{2(\theta + \phi)}{\bar{\tau}} - 6\theta\right), \tag{13}$$

where the angles are measured in radians. Since the tension parameters are always known in advance, they are treated like constants in this expansion.

Instead of requiring continuity of curvature, we will obtain a system of linear equations by requiring the mock curvatures to match at each knot. When the angles $\theta_i$ and $\phi_i$ are small, this guarantees that the resulting splines will have approximate second-order continuity. For instance

$$\frac{|\hat{k}(\theta,\phi,1,1)-k(\theta,\phi,1,1)|}{\max(|k(\theta,\phi,1,1)|,|k(\phi,\theta,1,1)|)} < 0.11$$

when $|\theta|, |\phi| \le 22.5°$ and $\rho$ and $\sigma$ are determined by (11) with $a = \sqrt{2}$, $b = \frac{1}{16}$, and $c = \frac{1}{2}(3-\sqrt{5})$.

In many applications the angles $\theta_i$ and $\phi_i$ seldom get much larger than 22.5° but experience has shown that the splines usually appear very smooth even when some $\theta_i$ and $\phi_i$ are much larger than 22.5°. An iterative algorithm such as Manning's can be used to obtain true curvature continuity if desired, but this necessitates giving up the simplicity and guaranteed behavior of a system of linear equations.

Continuity of mock curvature requires

$$\frac{\hat{k}(\phi_i,\theta_{i-1},\bar{\tau}_i,\tau_{i-1})}{d_{i-1}} - \frac{\hat{k}(\theta_i,\phi_{i+1},\tau_i,\bar{\tau}_{i+1})}{d_i} = 0 \quad \text{for } 1 \le i \le n'. \tag{14}$$

Combining this with the first-order continuity equations

$$\theta_i + \phi_i = -\psi_i \quad \text{for } 1 \le i \le n' \tag{15}$$

gives enough equations to determine all $\theta_i$ and $\phi_i$ for closed curve problems. For open curve problems when directions $w_0$ and $w_n$ are given in advance, these provide the necessary additional equations by fixing $\theta_0$ and $\phi_n$, but otherwise additional constraints are needed.

The additional constraints are controlled by special "curl" parameters $\chi_0$ and $\chi_n$. There should be one such constraint for each endpoint where no direction is specified. They have the form

$$\hat{k}(\theta_0,\phi_1,\tau_0,\bar{\tau}_1) = \chi_0\hat{k}(\phi_1,\theta_0,\bar{\tau}_1,\tau_0) \tag{16a}$$

and

$$\hat{k}(\phi_n,\theta_{n-1},\bar{\tau}_n,\tau_{n-1}) = \chi_n\hat{k}(\theta_{n-1},\phi_n,\tau_{n-1},\bar{\tau}_n). \tag{16b}$$

The curl parameters give the ratio of the mock curvature at the endpoints to that at the adjacent knots. They should probably have default values of 1 so that the first and last spline segments will usually be good approximations to circular arcs as in Fig. 12(b).

We now have a system of equations consisting of (14), (15), and possibly (16a) and/or (16b). If $\theta_0$ or $\phi_n$ have been given in advance then they may be

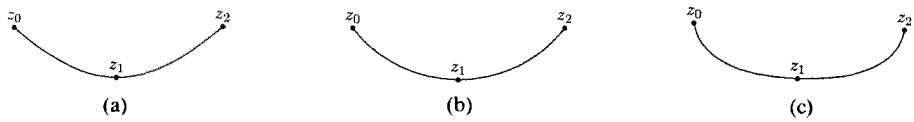


**Fig. 12.** (a) $\chi_0 = \chi_2 = 0$. (b) $\chi_0 = \chi_2 = 1$. (c) $\chi_0 = \chi_2 = \infty$.

regarded as constants. The first step is to rewrite (16a) and (16b) as

$$\theta_0 = \frac{\tau_0^3 + \chi_0 \bar{\tau}_1^3 (3\tau_0 - 1)}{\tau_0^3 (3\bar{\tau}_1 - 1) + \chi_0 \bar{\tau}_1^3} \phi_1 \quad \text{and} \quad \phi_n = \frac{\bar{\tau}_n^3 + \chi_n \tau_{n-1}^3 (3\bar{\tau}_n - 1)}{\bar{\tau}_n^3 (3\tau_{n-1} - 1) + \chi_n \tau_{n-1}^3} \theta_{n-1} \tag{17}$$

so that $\theta_0$ and $\phi_n$ can be eliminated. Then (15) may be used to eliminate all $\phi_i$ so that the remaining variables are $\theta_1, \theta_2, \ldots, \theta_{n'}$, and the remaining equations are those given by (14) with appropriate substitutions. This system has some important properties that may be summarized as follows.

**Theorem 1.** *If $n \geq 2$, if all tension parameters satisfy the bound $\tau_i, \bar{\tau}_i \geq \tau_{\min} > \frac{2}{3}$, and if any curl parameters satisfy $\chi_0, \chi_n \geq 0$, then after the aforementioned substitutions, all coefficients of $\theta_1, \theta_2, \ldots, \theta_{n'}$ in (14) are nonnegative, and for each $i$, the coefficient of $\theta_i$ is at least $3\tau_{\min} - 1$ times the sum of all the other coefficients in that equation.*

*Proof.* The bounds on the tension parameters guarantee that the coefficient of $\theta$ in (13) will be negative, the coefficient of $\phi$ will be positive, and the magnitude of the former will be at least $3\tau_{\min} - 1$ times the latter. When $1 < i < n'$ in (14), the only relevant substitutions are $\phi_j = -\psi_j - \phi_j$ for $|j - i| \leq 1$, so the coefficients of $\theta_{i-1}$, $\theta_i$, and $\theta_{i+1}$ clearly have the required properties. For closed curve problems, the same holds for $i = 1$ and $i = n'$, otherwise additional substitutions eliminate $\theta_0$ and $\phi_n$ so that $\hat{k}(\phi_1, \theta_0, \bar{\tau}_1, \tau_0)$ depends only on $\phi_1$ and $\hat{k}(\theta_{n-1}, \phi_n, \tau_{n-1}, \bar{\tau}_n)$ depends only on $\theta_{n-1}$. We need only show that both of these variables have nonpositive coefficients. This is clearly true for given directions, and it also holds when curl parameters apply since the coefficients in (17) are at most $3\tau_0 - 1$ and $3\bar{\tau}_n - 1$, respectively. □

Theorem 1 shows that subject to certain reasonable limitations on the tension and curl parameters, the system of equations is diagonally dominant, and hence it has a unique solution. Actually the solution is unique only up to the choice of the angles $\psi_i$. Ordinarily all $\psi_i$ should be chosen so that they are at most 180° in absolute value, but it is possible to add multiples of 360° to them. The effect of such a change is usually to add a loop to the curve as in Fig. 13.

Theorem 1 also shows that the splines have approximate locality in the sense that changes in direction angles fall off exponentially as one moves away from a disturbance. Specifically, suppose a given direction $\theta_0$ is displaced by an angle $\delta$ and let $A$ be the matrix of coefficients of $\theta_1, \theta_2, \ldots \theta_{n'}$ from (14) after the substitutions. The change in $\theta_1, \theta_2, \ldots, \theta_{n'}$ due to this displacement is given by the solution vector $\Theta$ to $A\Theta = \delta e_1$ where $e_1 = (1,0,0,\ldots,0)^T$.

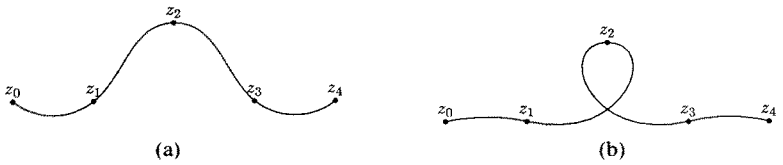


**Fig. 13.** A spline computed (a) with $\psi_2 = -90°$ and (b) with $\psi_2 = 270°$.

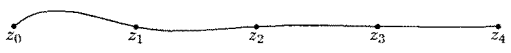


**Fig. 14.** Exponential decline in the effect of a 45° change in direction.

We know that $A$ is tridiagonal with nonnegative entries, and within each row the diagonal element dominates the sum of the other two elements by at least a factor of $3\tau_{\min} - 1$. It is not hard to see that for any two adjacent components of $\Theta$, either $\Theta_k = 0$ or $\Theta_{k-1}/\Theta_k \le 1 - 3\tau_{\min}$. This is trivial for $k = n'$, and it may be extended inductively to smaller $k$ using the fact that $A_{kk} \ge (3\tau_{\min} - 1)(A_{k,k-1} + A_{k,k+1})$. Thus, $j$ knots away from where a given direction is changed; the effect of the change is reduced by at least a factor of $(3\tau_{\min} - 1)^j$. In practice the reduction is often by a somewhat greater amount as in Fig. 14 where $\tau_{\min} = 1$ and $\theta_0/\theta_1 = -\frac{41}{11}$.

When a knot $z_i$ is displaced, three mock curvature constraints are directly affected due to changes in $d_{i-1}$, $d_i$, $\psi_{i-1}$, $\psi_i$, and $\psi_{j+1}$. The adjustment will cause some change in $\phi_{i-1}$ and $\theta_{i+1}$, and the effect on $\theta_{i+j}$ and $\theta_{i-j}$ for $j > 1$ is equivalent to what would happen if directions $w_{i+1}$ and $w_{i-1}$ were given in advance. The change in $\theta_{i+j}$ will be at most $1/(3\tau_{\min} - 1)^{j-1}$ as great as the change in $\theta_{i+1}$, and the change in $\theta_{i-j}$ will be at most $1/(3\tau_{\min} - 1)^{j-1}$ as great as the change in $\phi_{i-1}$. If the original problem were to find a closed curve with no directions given, then these two effects will add together so that the change in $\theta_{i+j}$ will be at most $1/(3\tau_{\min} - 1)^{j-1}$ of the change in $\theta_{i+1}$ plus $1/(3\tau_{\min} - 1)^{n-1-j}$ of the change in $\phi_{i-1}$.

## 4. Conclusion

We have developed a tridiagonal system of linear equations that can be solved in linear time to determine the spline direction at each knot so as to match mock curvatures. It is necessary to use arctangents to set up the system of equations, and to use sines and cosines to recover the resulting spline directions; but this work can be reduced to one arctangent, one sine, and one cosine per knot on the spline. When all unit direction vectors $w_i$ have been determined, the $i$th spline segment is the cubic whose Bézier control points are

$$z_i, \quad z_i + \frac{\rho(\theta_i, \phi_{i+1})}{3\tau_i}\|z_{i+1} - z_i\|w_i, \quad z_{i+1} - \frac{\sigma(\theta_i, \phi_{i+1})}{3\bar{\tau}_{i+1}}\|z_{i+1} - z_i\|w_{i+1},$$

and $z_{i+1}$.

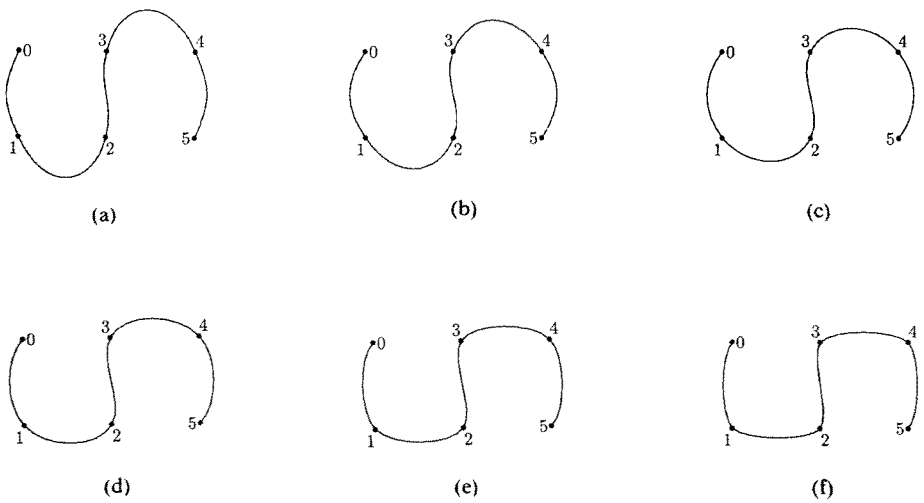


**Fig. 15.** (a) $\tau_i = \bar{\tau}_i = 0.75$. (b) $\tau_i = \bar{\tau}_i = 0.85$. (c) $\tau_i = \bar{\tau}_i = 1$. (d) $\tau_i = \bar{\tau}_i = 1.2$. (e) $\tau_i = \bar{\tau}_i = 1.5$. (f) $\tau_i = \bar{\tau}_i = 2$.

Thus, the total work required in order to find a computationally convenient representation of the desired spline is linear in the number of knots. The constant factor obviously depends on the implementation, but it is usually not much higher than the similar factor for natural cubic splines.

The splines do not have locality, but we have shown that changes in direction angles fall off exponentially. The rate of decline depends on how small the tension parameters are allowed to be, but at least a factor of 2 per knot is guaranteed for the default tensions, and decay rates as high as a factor of 4 per knot are typical. It should be noted that an exponential decline in angular change does not guarantee that curve displacements decline similarly because it is technically feasible for $d_j$ to be exponential in $j$.

The curve families discussed in Section 2 and defined by (10) and (11) are somewhat arbitrary, and the concept of mock curvature could be applied to other families of curves. It would be desirable to find $\rho$ and $\sigma$ functions simpler than (10) that perform better than (11), although even the simplified functions of (11) produce very good results for problems such as that shown in Figs. 2(a) and 3(c). In fact the splines produced by (11) have performed outstandingly in extensive tests with Knuth's new METAFONT system [6]. It is easy to compute the splines by solving the equations given in Section 3, but readers wishing a detailed description of one implementation can refer to [6].

The tension parameters have proved very useful in METAFONT, because it is often much easier to control a shape by specifying tensions than by specifying directions or giving more knots. Figure 15 shows the effect of adjusting all the tension parameters simultaneously, and Fig. 16 shows the effect of adjusting only one tension parameter and leaving the rest of the tension parameters at the default values ($\tau_i = \bar{\tau}_i = 1$). Note how the tension effects the direction chosen at knot 2 in Fig. 16.

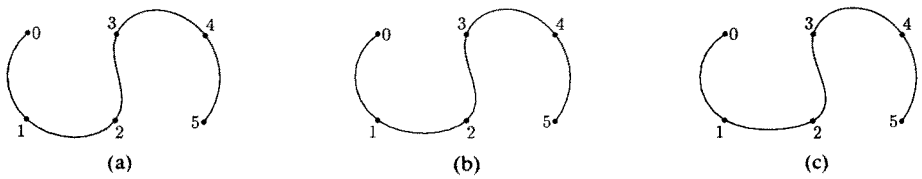


**Fig. 16.** (a) $\bar{\tau}_2 = 1.2$. (b) $\bar{\tau}_2 = 1.5$. (c) $\bar{\tau}_2 = 2$.

## Acknowledgments

The author is indebted to Professor Donald Knuth of Stanford University for doing the first complete implementation and helping to refine some of the ideas. Professor Kevin Karplus of Cornell University helped with the numerical investigation of the "optimal" $\rho$ and $\sigma$ functions.

## References

1. Brian A. Barsky, The Beta Spline: A Local Representation Based on Shape Parameters and Fundamental Geometric Measures, Ph.D. thesis, Univ. of Utah, December, 1981.
2. Brian A. Barsky, John C. Beatty, Varying the betas in beta-splines, Univ. of California, Berkeley TR CSD 82/112, December 1982.
3. Tony D. DeRose and Brian A. Barsky, Geometric continuity and shape parameters for Catmull-Rom Splines, Graphics Interface '84, 57–64.
4. M. P. Epstein, Parametric interpolation, SIAM Journal of Numerical Analysis 13 (1976), 261–268.
5. B. K. P. Horn, The curve of least energy, Massachusetts Institute of Technology A.I. Memo No. 612, January 1981.
6. Donald E. Knuth, Computers and Typesetting, Vol. 4: METAFONT the Program, Addison Wesley, to appear.
7. Doris H. U. Kochanek and Richard H. Bartels, Interpolating splines with local tension, continuity, and bias control, Computer Graphics 18 (1984), 33–41.
8. J. R. Manning, Continuity conditions for spline curves, Computer Journal 17 (1974), 181–186.
9. Even Mehlum, Curve and Surface Fitting Based on Variational Criteriæ for Smoothness, Oslo, 1969.

*Received January 31, 1985, and in revised form September 4, 1985.*