1990A&A...239..443S

Astron. Astrophys. 239, 443–450 (1990)

ASTRONOMY
AND
ASTROPHYSICS

A simple method for monotonic interpolation in one dimension

M. Steffen

Institut für Theoretische Physik und Sternwarte der Universität, Olshausenstr. 40, D-2300 Kiel 1, Federal Republic of Germany

Received April 25; accepted August 15, 1990

Abstract. A simple method is proposed for a 1-dimensional interpolation on a given set of data points (x_i, y_i). In each interval (x_i, x_{i+1}) the interpolation function is assumed to be a third-order polynomial passing through the data points. The slope at each grid point is determined in such a way as to guarantee a monotonic behavior of the interpolating function. The result is a smooth curve with continuous first-order derivatives that passes through any given set of data points without spurious oscillations. Local extrema can occur only at grid points where they are given by the data, but not in between two adjacent grid points. The method gives exact results if the data points correspond to a second-order polynomial.

Key words: numerical methods – interpolation – hydrodynamics

1. Introduction

Interpolation on a discrete set of data points is a problem encountered in many fields of science and engineering. Consequently, a great number of methods exists (e.g. Späth, 1990). The method presented in this paper was developed for application in numerical hydrodynamics based on the method of characteristics, where interpolation of the physical variables in between the grid points is one of the basic steps in advancing the solution in time. For a description of various characteristics codes used for astrophysical applications see e.g. Ulmschneider et al. (1977), Stefanik et al. (1984) or Schmitz (1986).

If smoothness of the interpolation curve is of major importance, cubic splines are often the method of choice. Here the interpolation is based on third-order polynomials in each interval (x_i, x_{i+1}). The slope at each x_i is determined from the condition that the second-order derivative is to be continuous at the junction points, which is a global problem resulting in a non-local dependence of the interpolation function on the data points. Cubic splines produce very smooth curves, first- and second-order derivatives changing continuously across the data set (for details see e.g. Ahlberg et al., 1967, Lancaster and Salkauskas, 1986, or Späth, 1990).

The price to pay for smoothness is high, however, as is well known from interpolation by higher order polynomials. To achieve smoothness, the curve has to perform spurious oscillations in many cases, often leading to completely unrealistic interpolation results, especially with non-equidistant grids. Moreover, due to the non-local character of this interpolation scheme, the results of the interpolation in a given interval will depend on the data points far away from this interval, at least in principle.

These properties can be of major concern in the above mentioned hydrodynamics calculations, for example. In the presence of strong gradients in the data (shocks), the usual spline interpolation generally leads to severe oscillations of the interpolated variables, resulting in physically meaningless data, in extreme cases producing negative values of pressure or temperature. This is demonstrated schematically in Fig. 1 (for the astrophysical background of this example see section 3). Furthermore, the long distance effect of the splines propagates disturbances instantaneously across the computational domain, whereas physically the propagation velocity is given by the speed of sound. The non-local behavior of the spline interpolation is obvious in the examples displayed in Figs. 1 and 2.

The long distance effect can be eliminated at the expense of smoothness if the determination of the slope at the junction

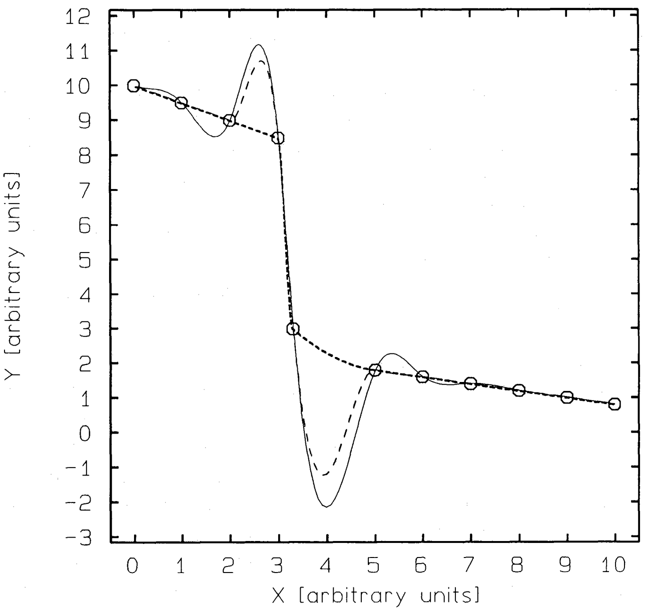

Fig. 1. Comparison of different interpolation methods discussed in the text for the example data points indicated by circles. Full-drawn: splines, long-dashed: Bessel's method, short-dashed: Akima's method.

Send offprint requests to: M. Steffen

© European Southern Observatory • Provided by the NASA Astrophysics Data System

1990A&A...239..443S

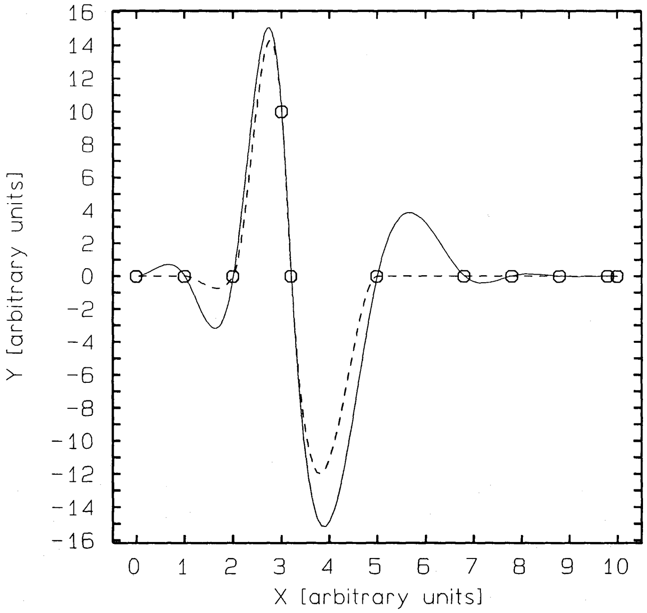

Fig. 2. Spline interpolation (full-drawn) and Bessel's interpolation method (dashed) for the example data points indicated by circles. The disturbance introduced by the 4th point from the left is noticeable over many intervals in the case of the spline interpolation, whereas it is confined to 4 intervals in the case of Bessel's method. In both cases, the maximum in the data is not mapped onto an equivalent maximum in the interpolation curve.

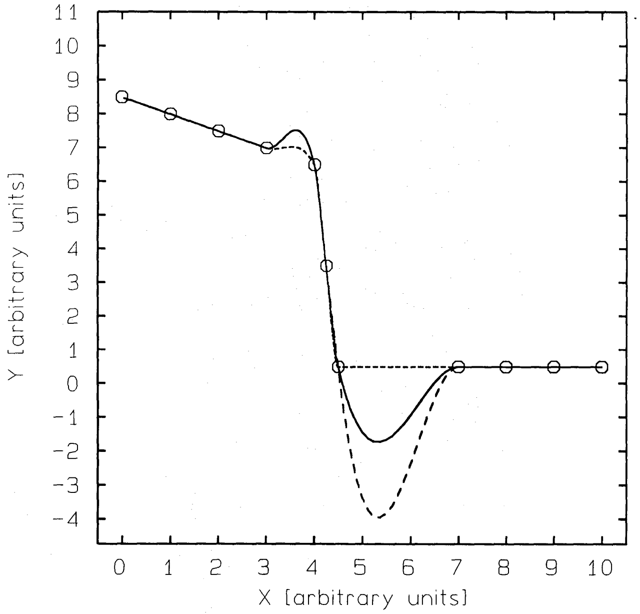

Fig. 3. Akima's interpolation for a data set where this method shows unwanted oscillations. The interpolation function jumps from the full-drawn to the long-dashed curve if the 4th data point from the right is shifted by an 'infinitesimal' amount (from (7.0000,0.5000) to (7.0000, 0.5001), and to the short-dashed curve when the 5th data point from the left is moved by an 'infinitesimal' amount (from (4.0000, 6.5000) to (4.0000, 6.5001).

points is based on local considerations. In this case, continuity of the second-order derivative is lost in general. However, for many purposes this may be more acceptable than a non-local interpolation behavior. An example for such a scheme is interpolation by piecewise cubic functions with locally determined slopes. A natural choice for the slopes at an internal grid point x_i is to compute it from the unique parabola passing through $(x_{i-1}, y_{i-1};\ x_i, y_i;\ x_{i+1}, y_{i+1})$ (Bessel's methods). Given the value and the slope of the variable to be interpolated at each grid point (Hermite-type interpolation), the piecewise cubic function is uniquely determined. Consequently, for Bessel's method interpolation results depend only on local conditions. As may be seen from Figs. 1 and 2, oscillations are somewhat reduced relative to those of the splines (at least in these examples) and are confined to just the critical intervals. However, in the presence of sharply changing slopes in the given data, the oscillations will still be unacceptably large for hydrodynamics applications.

A simple way to rule out unwanted oscillations is to use a piecewise linear interpolation function. But since in this case not even the first-order derivative changes continuously across the junction points, the resulting interpolation curve can no longer be considered as smooth. For hydrodynamics applications this is not appropriate because the physical variables as well as their first-order derivatives have to be interpolated from the same interpolation function. For this reason, the interpolation schemes discussed in this paper are restricted to those resulting in functions with at least continuous first-order derivatives.

In an effort to suppress unnatural wiggles in the interpolation curve, Akima (1970) proposed a new method to determine the slopes at the grid points from local conditions. Again, the interpolation function is composed of piecewise third-order polynomials. This method completely eliminates unwanted oscillations in the example given in Fig. 1. However, Akima's method (implemented in IMSL-Library as subroutine CSAKM) does not guarantee for every situation that the resulting interpolation curve connects two adjacent data points in a monotonic way. It is easy to construct examples where this method leads to unrealistic oscillations, too (see Fig. 3). Another serious drawback of Akima's interpolation is the property that an 'infinitesimal' change in one of the data points can result in large changes in the interpolation curve. This feature, which is also demonstrated in Fig. 3, is unacceptable for hydrodynamics applications.

The interpolation method proposed in this paper, also assuming a representation by piecewise cubic functions, was developed along lines of reasoning very similar to those of Akima. In many cases the resulting interpolation curves are indeed very similar. However, the method described here guarantees a monotonic behavior for every situation, in contrast to Akima's method. This means that local extrema in the data will result in corresponding extrema at the same position and of the same amplitude in the interpolation function, in contrast to the other methods discussed so far (see also Fig. 2). Equally important, the resulting monotonic interpolation curve depends on the given data points in a continuous way, i.e. small changes in the data will lead to small changes in the interpolating function. Like Bessel's method the proposed scheme gives exact results if the data points are derived from a second-order polynomial, whereas the same is true for Akima's method only if the grid points are equidistant.

In the next section we give the details of our monotonic interpolation scheme, while a few example results are presented

© European Southern Observatory • Provided by the NASA Astrophysics Data System

1990A&A...239..443S

in section 3. Section 4 contains some concluding remarks and summarizes the main points.

2. Monotonic interpolation by a piecewise cubic function

2.1. The method

The aim is to construct a piecewise cubic interpolation function that passes through N given data points in a monotonic way with the slope of the curve changing continuously across the junction points. The method described in the following is conceptually simple and easy to code.

In interval (x_i, x_{i+1}) the interpolation function is written as

$$f_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i, \; i = 1, N - 1. \quad (1)$$

If we know at each grid point i the value y_i and the slope y_i' of the variable to be interpolated, the coefficients a_i, b_i, c_i, d_i are uniquely determined:

$$a_i = (y_i' + y_{i+1}' - 2s_i)/h_i^2 \quad , \qquad (2)$$

$$b_i = (3s_i - 2y_i' - y_{i+1}')/h_i \quad , \qquad (3)$$

$$c_i = y_i' \quad , \qquad (4)$$

$$d_i = y_i \quad , \qquad (5)$$

where

$$h_i = (x_{i+1} - x_i) \qquad (6)$$

is the size of the interval, and

$$s_i = (y_{i+1} - y_i)/(x_{i+1} - x_i) \qquad (7)$$

is the slope of the secant through points $(x_i, y_i; \, x_{i+1}, y_{i+1})$. The piecewise cubic function constructed from given y_i and y_i' (Hermite-type interpolation) automatically has a continuous first-order derivative over the whole data set.

The remaining problem is to determine the first-order derivatives at the grid points from the given data points in such a way from local considerations that the resulting interpolation curve behaves monotonically.

The idea of our procedure is simple. Basically we determine the slope at an internal point i from the unique parabola (parabola (A) in Fig. 4) passing through points $(x_{i-1}, y_{i-1}; \, x_i, y_i; \, x_{i+1}, y_{i+1})$, as in Bessel's method (see above), that is

$$p_i = (s_{i-1}h_i + s_i h_{i-1})/(h_{i-1} + h_i) \quad . \qquad (8)$$

Note that for an equidistant x-grid ($h_{i-1} = h_i$) this is the same as the slope of the secant through points $(x_{i-1}, y_{i-1}; \, x_{i+1}, y_{i+1})$. As long as parabola (A) is monotonic in both intervals (x_{i-1}, x_i) and (x_i, x_{i+1}), its slope at center point i is considered a reasonable estimate for y_i' and we set $y_i' = p_i$. Otherwise (the case illustrated in Fig. 4), p_i is not accepted as y_i'. Rather, the absolute value of the slope at point i is reduced to p_i^*, the absolute value of which is to be just small enough to assure that both parabola (B) and parabola (C) (see Fig. 4) starting with this reduced slope at (x_i, y_i) and passing through (x_{i-1}, y_{i-1}) and (x_{i+1}, y_{i+1}), respectively, will be monotonic in their respective intervals. Now the position of the minimum/maximum of parabola (B) and parabola (C), respectively, is given by:

$$x_B^* = x_i - \frac{h_{i-1}\, p_i^*}{2(p_i^* - s_{i-1})} \quad , \qquad (9a)$$

$$x_C^* = x_i + \frac{h_i\, p_i^*}{2(p_i^* - s_i)} \quad . \qquad (9b)$$

From expressions (9a) and (9b) it may be seen that x_B^* will not lie inside the interval (x_{i-1}, x_i) if

$$0 \le \frac{p_i^*}{s_{i-1}} \le 2 \quad , \qquad (10a)$$

and similarly x_C^* will not lie inside the interval (x_i, x_{i+1}) if

$$0 \le \frac{p_i^*}{s_i} \le 2 \quad . \qquad (10b)$$

If s_{i-1} and s_i have different sign or at least one of these slopes is zero, the only way to satisfy (10a) and (10b) simultaneously is by $p_i^* = 0$. In this case the extremum is located at point i.

In view of this result we determine the final value of y_i' such that (10a) and (10b) will be satisfied simultaneously by $p_i^* = y_i'$, i.e.

$$y_i' = \begin{cases} 0 & \text{if } s_{i-1}\, s_i \le 0 \\ 2a \min(|s_{i-1}|, |s_i|) & \text{if } \begin{array}{l} |p_i| > 2|s_{i-1}| \text{ or} \\ |p_i| > 2|s_i| \end{array} \\ a = \text{sign}(s_{i-1}) = \text{sign}(s_i) & \\ p_i & \text{otherwise.} \end{cases} \qquad (11)$$

In brief, the slope at i must not be greater in absolute value than twice the minimum of the absolute value of the slopes given by the left- and right-handed finite differences. If these have different sign, y_i' must be zero. These rules may be condensed into a single FORTRAN statement:

```
Y1(i)=(SIGN(1.0,S(i-1))+SIGN(1.0,S(i)))*
MIN(ABS(S(i-1)),ABS(S(i)),0.5*ABS(P(i))),
```

where `Y1(i)` is to be identified with y_i' and the other symbols have the same meaning as above. At the end points ($i = 1$; $i = N$) appropriate boundary conditions have to be applied (see below).

Note that according to (11) y_i' is a continuous function of data points $i-1, i, i+1$. Assume s_{i-1} and s_i have the same sign. If their absolute values are not too different, $y_i' = p_i$, which is clearly a continuous function. If $|s_i|$ is increased (while s_{i-1} is fixed) $|y_i'|$ will increase until $|s_i| \ge |s_{i-1}|(2h_{i-1} + h_i)/h_{i-1}$. From this point on $y_i' = 2s_{i-1}$, independent of s_i. On the other hand, decreasing $|s_i|$ will result in a decreasing $|y_i'|$. Beyond the point where $|s_i| \le |s_{i-1}|h_i/(2h_i + h_{i-1})$, $y_i' = 2s_i$, which goes to zero as s_i goes to zero. If s_i changes sign y_i' stays at zero irrespective of the particular value of s_i. So it is clear that y_i' does not suffer any jumps as the data points are changed continuously. Hence, the whole interpolation function, depending only on y_i and y_i', will respond to changes in the data in a continuous way.

The main point still has to be proved: given the y_i' at each grid point determined from the y_i according to (11), the resulting piecewise *cubic* interpolation function is monotonic in each interval. Due to the local character of the problem we need to consider only one interval (x_i, x_{i+1}) and have to show that the first-order derivative of the interpolation function $f_i'(x)$ does not change sign inside this interval. We may assume that $s_i \neq 0$, because otherwise y_i' and $y_{i+1}' = 0$ according to (11) and the interpolation function is just a horizontal straight line. Now

$$f_i'(x) = 3a_i(x - x_i)^2 + 2b_i(x - x_i) + c_i \qquad (12)$$

© European Southern Observatory • Provided by the NASA Astrophysics Data System

1990A&A...239..443S

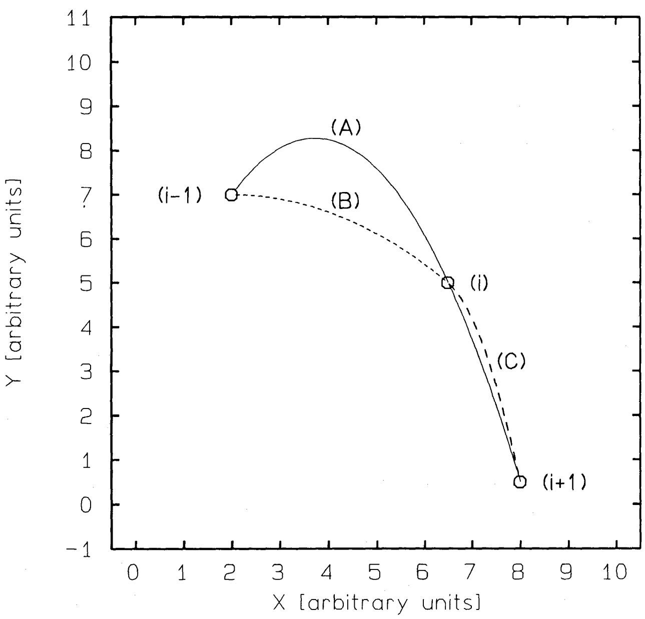

Fig. 4. Visualization of parabola (A) (full-drawn), parabola (B) (short-dashed) and parabola (C) (long-dashed) referred to in the text.

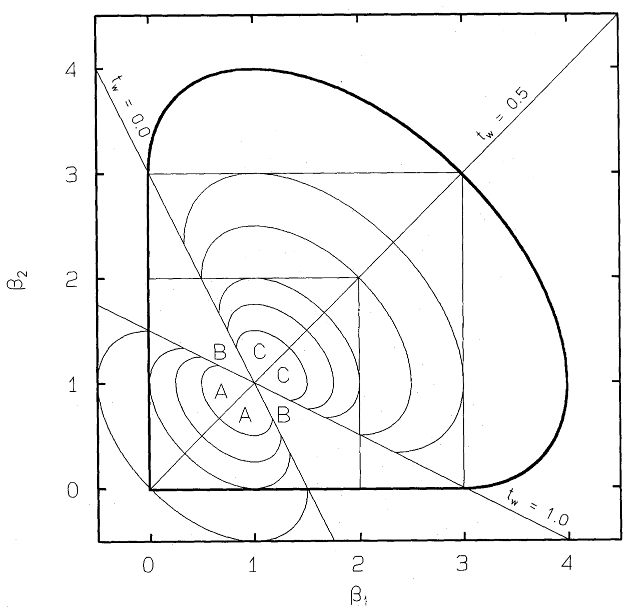

Fig. 5. Representation of the β_1/β_2-plane. β_1/β_2 combinations falling inside the area bounded by the heavy line result in a monotonic behavior of the cubic interpolation function. On the straight lines labelled $t_w = 0.0$, $t_w = 0.5$, and $t_w = 1.0$, the inflection point is located at the left side, in the middle, and at the right side of the interval, respectively. In sectors "B", the inflection point falls outside the interval. The slope of the interpolation function at the inflection points is constant on elliptical arcs. In sectors "A", $\beta_{\max} = \frac{7}{6}$, $\frac{5}{4}$, $\frac{4}{3}$, and $\frac{3}{2}$ proceeding from smaller to larger elliptical arcs. In sectors "C", $\beta_{\min} = \frac{5}{6}$, $\frac{3}{4}$, $\frac{2}{3}$, $\frac{1}{2}$, $\frac{1}{3}$, and 0 on progressively larger elliptical arcs.

or, introducing (2) to (4),

$$f_i'(t) = 3(y_i' + y_{i+1}' - 2s_i)t^2 + 2(3s_i - 2y_i' - y_{i+1}')t + y_i' \quad , \tag{13}$$

where $t = (x - x_i)/(x_{i+1} - x_i)$ varies between 0 and 1 in the considered interval. Defining

$$\beta(t) = f_i'(t)/s_i \quad , \tag{14}$$

$$\beta_1 = y_i'/s_i \quad (\geq 0) \quad , \tag{15}$$

$$\beta_2 = y_{i+1}'/s_i \quad (\geq 0) \quad , \tag{16}$$

equation (13) reads

$$\beta(t) = 3(\beta_1 + \beta_2 - 2)t^2 + 2(3 - 2\beta_1 - \beta_2)t + \beta_1 \quad . \tag{17}$$

Since $\beta_1, \beta_2 \geq 0$, this relation may be rewritten as

$$\begin{aligned}\beta(t) &= (\sqrt{\beta_1}(1-t) - \sqrt{\beta_2}\,t)^2 \\ &\quad + 2t(1-t)(3 + \sqrt{\beta_1\beta_2} - \beta_1 - \beta_2) \\ &\equiv U + VW.\end{aligned} \tag{18}$$

We note that the first term on the right-hand side (U), being a squared quantity, cannot become negative and that $V = t(1-t) \geq 0$ inside the interval (x_i, x_{i+1}). Then β will be greater (or equal) zero inside the interval if $W \geq 0$. So we must have

$$\beta_1 + \beta_2 - \sqrt{\beta_1\beta_2} \leq 3, \quad \beta_1, \beta_2 \geq 0 \quad . \tag{19}$$

This condition is not only sufficient to assure $\beta(t) \geq 0$ for $0 \leq t \leq 1$, but also necessary, because U always becomes zero somewhere inside the interval.

Based on relations (19), Fig. 5 shows the region in the β_1/β_2-plane where the cubic interpolation function is strictly monotonic inside the relevant interval (x_i, x_{i+1}). Inside this monotonicity region in the β_1/β_2- plane, the interpolation function either has no extrema at all or these are located somewhere outside the relevant interval. Although unimportant in the context of our interpolation scheme, it may be of interest to know where, for given β_1 and β_2, the extrema of the corresponding cubic polynomial are located, if they exist at all. In the Appendix we give a method how to find this information for any point in the β_1/β_2-plane from a simple geometrical construction.

Realizing that (11) implies

$$0 \leq \beta_1 \leq 2 \quad \text{and} \quad 0 \leq \beta_2 \leq 2 \quad , \tag{20}$$

we can see immediately from Fig. 5 that this condition easily assures monotonicity of the piecewise cubic interpolation function. Moreover, it is clear that we could also have used the somewhat less restrictive condition

$$0 \leq \beta_1 \leq 3 \quad \text{and} \quad 0 \leq \beta_2 \leq 3 \quad , \tag{21}$$

instead. In principle, some β_1 or β_2 could even be allowed to lie between 3 and 4 without losing monotonicity, if the respective β_i at the other side of the considered interval is in the right range. However, the required relations between the β_i would make the problem potentially non-local. Using the simpler condition (21) would assure locality of the interpolation scheme but is incompatible with the basic determination of the y_i' by parabolas: even the slopes derived from non-monotonic parabolas would be accepted as y_i', which we consider a doubtful procedure. The problem is avoided if we adopt condition (20), which is also preferred because inside the square $0 \leq \beta_1 \leq 2$; $0 \leq \beta_2 \leq 2$ we have a perfectly symmetric situation.

© European Southern Observatory • Provided by the NASA Astrophysics Data System

1990A&A...239..443S

To see this, we investigate the different areas in the β_1/β_2-plane in some more detail. The second-order derivative of the cubic interpolation function across interval (x_i, x_{i+1}) is given by

$$f_i''(x) = 6a_i(x - x_i) + 2b_i = 6a_i h_i t + 2b_i \quad . \tag{22}$$

The inflection point where $f_i'' = 0$ is located at

$$t_w = -\frac{b_i}{3a_i h_i} = \frac{(2\beta_1 + \beta_2 - 3)}{3(\beta_1 + \beta_2 - 2)} \quad . \tag{23}$$

In the β_1/β_2-plane, t_w is constant on straight lines through $(\beta_1, \beta_2) = (1, 1)$. In Fig. 5 we have indicated those lines where $t_w = 0$ $(2\beta_1 + \beta_2 = 3)$, $t_w = 0.5$ $(\beta_1 = \beta_2)$, and $t_w = 1$ $(\beta_1 + 2\beta_2 = 3)$. The lines $t_w = 0$ and $t_w = 1$ define the boundaries of different sectors "A", "B", and "C" in the plane. In sectors "A", the inflection point of the interpolating polynomial is inside the respective interval and the slope reaches a local *maximum* at this point. On the other hand, β_1/β_2 combinations located in sectors "B" result in an interpolation curve which is not only monotonic but also has monotonically changing slope across the interval of interest. Finally, in sectors "C" the inflection point lies again inside the interval, and the slope attains its *minimum* value at this point. The value of the slope at the inflection points is constant on the elliptical curves also indicated in Fig. 5. From these we see that in sectors "A" the slope of the cubic interpolation function is at most a factor 1.5 higher than the slope of the secant. In sectors "C" the slope at the inflection point can go down to zero at the outer boundary line. However, if we restrict β_1 and β_2 to be smaller or equal 2, (eqn. (11), eqn. (20)), the slope is nowhere less than 0.5 times the slope of the secant of the respective interval. The areas of sectors "A" and "C" become equal in this case and it is obvious from a look at Fig. 5 that conditions are perfectly symmetric inside the square given by $0 \leq \beta_1 \leq 2$; $0 \leq \beta_2 \leq 2$.

2.2. *Boundary conditions*

The boundary points ($i = 1$; $i = N$) have to be treated somewhat differently than the inner points.

For some problems, the value of the slope at the end points may be known independently. In this case these values can be used directly for y_1' and y_N', respectively. Monotonicity in the boundary intervals is guaranteed independently of the slopes at the inner points if y_1' has the same sign as s_1 (y_N' has the same sign as s_{N-1}) and $|y_1'| \leq 3|s_1|$ ($|y_N'| \leq 3|s_{N-1}|$). For example, $y_1' = y_N' = 0$ may be a natural boundary condition for some problems, clearly resulting in monotonic interpolation in the boundary intervals.

If the slopes at the end points are not known independently, there are several different ways to calculate them from the data points.

The simplest possibility is to define the unknown slopes by the one-sided finite differences, that is to set $y_1' = s_1$ and $y_N' = s_{N-1}$. In this case monotonicity is automatically guaranteed. The same is true if the slopes at the end points are calculated from an exponential through $(x_1, y_1; x_2, y_2)$ and $(x_{N-1}, y_{N-1}; x_N, y_N)$, respectively.

Somewhat more sophisticated, a natural modification of condition (11) is to calculate p_1 from the parabola through the three boundary points $(x_1, y_1; x_2, y_2; x_2, y_3)$ as

$$p_1 = s_1 \left(1 + \frac{h_1}{h_1 + h_2}\right) - s_2 \left(\frac{h_1}{h_1 + h_2}\right) \quad . \tag{24}$$

Similarly,

$$p_N = s_{N-1} \left(1 + \frac{h_{N-1}}{h_{N-1} + h_{N-2}}\right) - s_{N-2} \frac{h_{N-1}}{h_{N-1} + h_{N-2}} . \tag{25}$$

In analogy with (11) we set

$$y_1' = \begin{cases} 0 & \text{if } p_1 s_1 \leq 0 \\ 2s_1 & \text{if } |p_1| > 2|s_1| \\ p_1 & \text{otherwise} \end{cases} \tag{26}$$

and

$$y_N' = \begin{cases} 0 & \text{if } p_N s_{N-1} \leq 0 \\ 2s_{N-1} & \text{if } |p_N| > 2|s_{N-1}| \\ p_N & \text{otherwise.} \end{cases} \tag{27}$$

Finally, it may be desirable to require the second-order derivative to vanish at the end points ($f_1'' = f_N'' = 0$). From (23) we find for $t_w = 0$

$$\beta_1 = \frac{3 - \beta_2}{2} \quad \text{or} \quad y_1' = \frac{3}{2}s_1 - \frac{1}{2}y_2' \quad . \tag{28}$$

Since $0 \leq \beta_2 \leq 2$ we have $0.5 \leq \beta_1 \leq 1.5$ and monotonicity is guaranteed. At the right end point we obtain for $t_w = 1$

$$\beta_2 = \frac{3 - \beta_1}{2} \quad \text{or} \quad y_N' = \frac{3}{2}s_{N-1} - \frac{1}{2}y_{N-2}' \quad . \tag{29}$$

Again, $0.5 \leq \beta_2 \leq 1.5$, resulting in monotonic interpolation in the last interval.

3. Example results

At this point we want to show just a few results by applying our monotonic interpolation method to the examples of Figs. 1, 2, and 3. The corresponding plots are shown in Figs. 6, 7, and

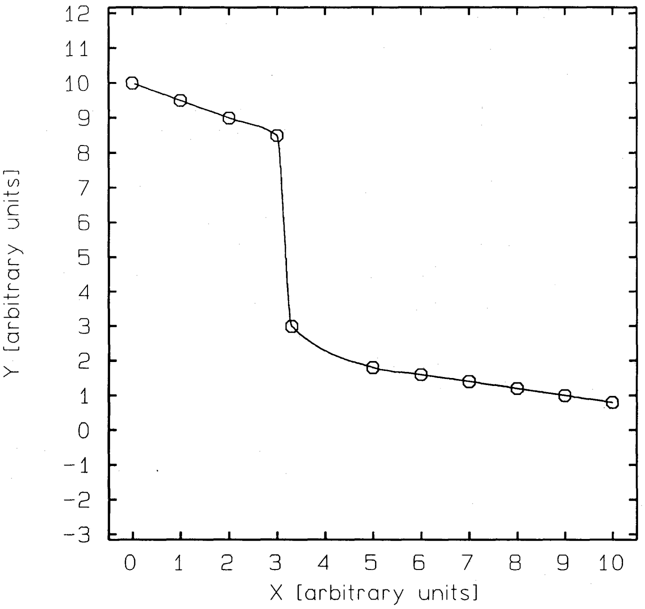

Fig. 6. Same as Fig. 1, but using the monotonic interpolation scheme described in this paper.

© European Southern Observatory • Provided by the NASA Astrophysics Data System

1990A&A...239..443S

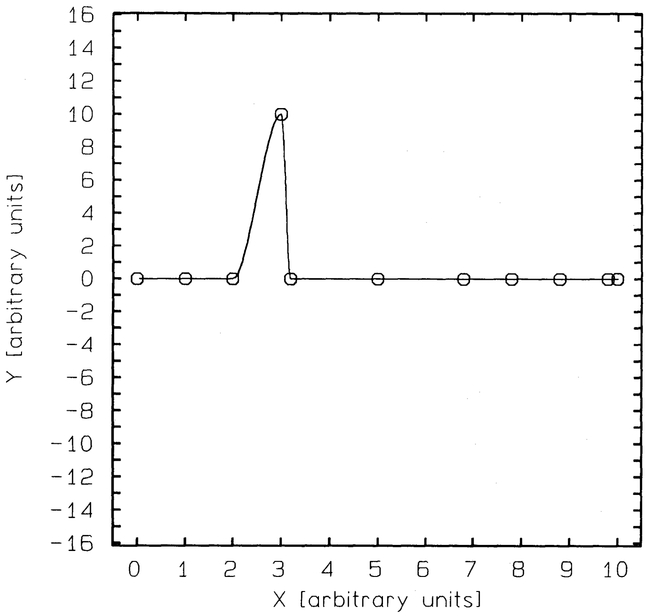

Fig. 7. Same as Fig. 2, but using the monotonic interpolation scheme described in this paper.

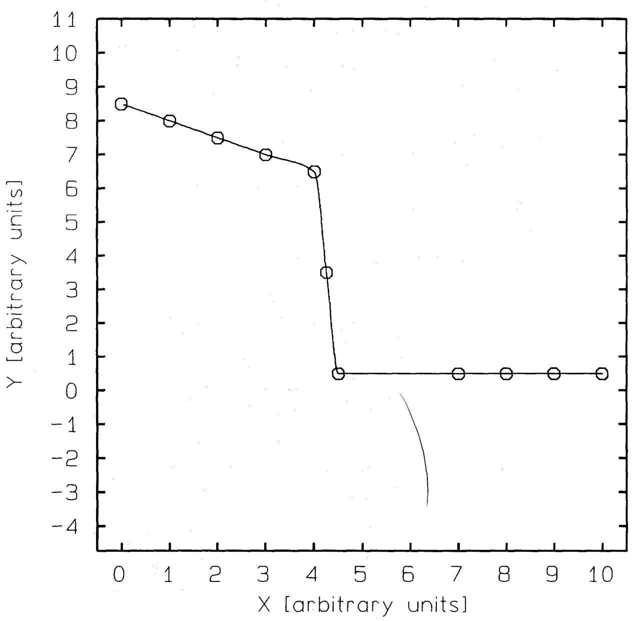

Fig. 8. Same as Fig. 3, but using the monotonic interpolation scheme described in this paper.

8. Boundary conditions are identical and uncritical in all these cases.

The first example (Figs. 1/6) was chosen to represent schematically the temperature structure in the hot rising parts of the solar photosphere as obtained from hydrodynamical calculations (Steffen et al., 1989). It is obvious that calculating the advection of internal energy by interpolation of the temperature in between the grid points becomes an unrealistic procedure if interpolation is based on normal splines or on Bessel's method. However, a physically consistent description of the hydrodynamical problem is possible when a monotonic interpolation is used. Note that our monotonic scheme and Akima's method result in almost identical curves in this example.

The second example (Figs. 2/7) shows the interpolation curves for a situation where all except one data point are on a constant level. The reaction of the spline is heavy oscillations in the vicinity of the disturbance. While Bessel's interpolation shows oscillations of similar amplitude, the effect of the disturbance is confined to 4 intervals. Finally, the monotonic interpolation undergoes no oscillations at all and the interpolation curve deviates from a horizontal straight line in just 2 intervals, clearly demonstrating the completely local behavior of our method.

The third example (Figs. 3/8) demonstrates the shortcomings of Akima's (1970) method and the superiority of our scheme. For the situation investigated in this example, Akima's interpolation is essentially undefined, resulting in greatly different curves if certain data points are displaced by an 'infinitesimal' amount. Moreover, another independent drawback of this method is that in some situations it produces unwanted wiggles (Fig. 3). In contrast, our method connects every set of data points by piecewise monotonic interpolation curves, their shape depending on the given data points in a continuous way.

4. Concluding remarks

We have described a simple method for 1-dimensional monotonic interpolation based on piecewise cubic functions. It is applicable to arbitrary sets of data points, not just to data sets with monotonically decreasing or increasing values. By monotonic interpolation we mean that the interpolation curves behave monotonically in each interval, not permitting minima/maxima to occur in between adjacent data points. The resulting curve smoothly passes through the given data points without unwanted oscillations, the first-order derivative being continuous across the data set.

The point is to find a reasonable way to prescribe the slopes at the knot points from the given data. Our method is nonlinear in general. This means, if we have two data sets (1, 2), with identical abscissae (x_i) but different ordinates ($y_{1,i}$, $y_{2,i}$) and generate a third data set (3), (x_i, $y_{3,i}$), where $y_{3,i} = y_{1,i} + y_{2,i}$, the interpolation curve through these data points can be different from the sum of the interpolation curves through (x_i, $y_{1,i}$) and (x_i, $y_{2,i}$), respectively. Only if the slopes are not too different in adjacent intervals such that $y'_i = p_i$ for all i in data sets 1 and 2, the interpolation curve for data set 3 will be the superposition of the interpolation curves through data sets 1 and 2.

As mentioned before, our method gives exact results if the data points are derived from a second-degree polynomial, provided the extremum (if in the data set) falls on a knot point, non-equidistant x_i causing no problems.

Our method requires only straightforward procedures and is easy to implement as a computer subroutine. Being a local scheme it requires less computer time than usual spline interpolation because there is no need to solve a system of equations. In many respects it compares favourably with similar methods found in the literature. In the preceding section we have demonstrated the superiority over Akima's (1970) interpolation method.

When this manuscript was almost finished, another method for monotonic interpolation, also based on piecewise cubic functions, came to our attention (Fritsch and Carlson, 1980). This method has some serious drawbacks as noticed by Fritsch and Butland (1984): (1) setting up the derivatives at the knot points requires 2 passes over the data, (2) the results are dependent on the order in which the data are processed, and

1990A&A...239..443S

(3) the scheme is potentially non-local. Note that none of these defects is present in our method.

A reasonable alternative to the method proposed in this paper to determine the y_i' may be to calculate these slopes from a *weighted harmonic mean* as proposed by Brodlie (1980), also investigated by Fritsch and Butland (1984). Here the y_i' are calculated as

$$\frac{1}{y_i'} = \frac{h_{i-1}+2h_i}{3(h_{i-1}+h_i)}\,\frac{1}{s_{i-1}} + \frac{2h_{i-1}+h_i}{3(h_{i-1}+h_i)}\,\frac{1}{s_i} \tag{30}$$

if $s_{i-1}\,s_i > 0$. Otherwise, $y_i' = 0$, as in our method. Note that the slope calculated according to (30) is restriced to $|y_i'| \le 2|s_{i-1}|$, $|y_i'| \le 2|s_i|$ for an equidistant grid ($h_{i-1} = h_i$). In other words, and in agreement with our method, $\beta_1, \beta_2 \le 2$. Being calculated from a harmonic mean, the y_i' derived from (30) are always lower than the y_i' calculated according to (11). The difference can be up to 25 % and it is clear that application of Brodlie's formula will not yield exact results for a second-order polynomial.

As a final remark let us note that monotonic interpolation in 2 dimensions is a more complicated issue. There seems to be no simple way to extend the ideas presented in this paper to bi-cubic interpolation in 2 dimensions. The only method to achieve 2-dimensional monotonicity known to the author is bi-linear interpolation.

Acknowledgements. The considerations presented in the Appendix are largely due to H.-G. Ludwig. This work was supported by the Deutsche Forschungsgemeinschaft under grant Ho 596/21-2.

References

Ahlberg, J.H., Nilson, E.N., Walsh, J.L.: 1967, *The Theory of Splines and their Applications*, Academic Press, New York

Akima, H.: 1970, *J. ACM* **17**, 589

Brodlie, K.W.: 1980, in *Mathematical Methods in Computer Graphics and Design*, ed. K.W. Brodlie, Academic Press, New York, pp. 1 - 37

Fritsch, F.N., Carlson, R.E.: 1980, *SIAM J. Numer. Anal.* **17**, 238

Fritsch, F.N., Butland, J.: 1984, *SIAM J. Sci. Stat. Comput.* **5**, 300

Lancaster, P., Salkauskas, K.: 1986, *Curve and Surface Fitting*, Academic Press, New York

Schmitz, F.: 1986, *Astron. Astrophys.* **166**, 368

Späth, H.: 1990, *Eindimensionale Spline-Interpolations-Algorithmen*, Oldenbourg, München

Stefanik, R.P., Ulmschneider, P., Hammer, R., Durrant, C.J.: 1984, *Astron. Astrophys.* **134**, 77

Steffen, M., Ludwig, H.-G., Krüß, A.: 1989, *Astron. Astrophys.* **213**, 371

Ulmschneider, P., Kalkofen, W., Nowak, T., Bohn, U.: 1977, *Astron. Astrophys.* **54**, 61

Appendix : A geometrical method for locating extrema and inflection point of a third-order polynomial

In this Appendix we investigate the question how the position of the extrema and of the inflection point of a third-order polynomial can be determined if the corresponding curve is known to pass through points (x_1, y_1) and (x_2, y_2) with slopes y_1' and y_2', respectively. This information uniquely determines the cubic function which in general has two extrema (one minimum and one maximum) and one inflection point. Here we determine the location of these characteristic points from geometrical considerations in the β_1/β_2-plane, permitting a quick survey of all possible cases.

Define

$$h = x_2 - x_1 \quad , \tag{A1}$$

$$t = x/h \quad , \tag{A2}$$

$$s = (y_2 - y_1)/h \quad , \tag{A3}$$

$$\beta_1 = y_1'/s \quad , \tag{A4}$$

$$\beta_2 = y_2'/s \quad , \tag{A5}$$

where we have assumed $h \neq 0$ and $s \neq 0$. From (17) it may be seen that on a straight line in the β_1/β_2 -plane (see Figs. 5 and 9) defined by

$$3(\beta_1 + \beta_2 - 2)t_e^2 + 2(3 - 2\beta_1 - \beta_2)t_e + \beta_1 = 0 \tag{A6}$$

one of the two extrema is located at $t = t_e$. To each value of the parameter t_e there corresponds a different straight line. For example, we have $t_e = 0$ along the β_2-axis ($\beta_1 = 0$) and $t_e = 1$ on the β_1-axis ($\beta_2 = 0$). In Fig. 9 we have also indicated the straight lines where $t_e = -1, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, 2$, and $\pm\infty$.

It can be shown that the complete set of straight lines generated by $-\infty \le t_e \le +\infty$ has a common envelope which turns out to be the ellipse displayed in Fig. 9 (partly visible also in Fig. 5). Its center lies at $(\beta_1, \beta_2) = (2, 2)$, its semi-minor axis is along the line $\beta_1 = \beta_2$ and has the length $\sqrt{2}$, while the length of the semi-major axis is $\sqrt{6}$. This ellipse may be represented with t_e as a parameter by

$$\beta_1 = \frac{t_e^2}{t_e^2 - t_e + 1/3}; \quad \beta_2 = \frac{(t_e - 1)^2}{t_e^2 - t_e + 1/3} \quad , \tag{A7}$$

where $-\infty \le t_e \le +\infty$. Note that the denominator has no real root. Each point on the perimeter of the ellipse is characterized by a different value of t_e, which corresponds to the straight line tangent to the ellipse at this point. We have indicated some of these t_e values by tick marks along the perimeter of the ellipse in Fig. 9.

After these preparatory considerations it is easy to find for any arbitrary point P_0 in the β_1/β_2-plane the position of the two extrema and of the inflection point. For this purpose we just have to draw the two possible straight lines passing through P_0 and being tangent to the ellipse. There we can read the corresponding t_e values marked at the perimeter of the ellipse which directly give the location of the extrema. The inflection point always lies half way between the two extrema, if these exist. It is clear from (23) that its position is constant on straight lines through $(\beta_1, \beta_2) = (1, 1)$. Furthermore, from (23) and (A7) we find that $t_w = t_e$ on the perimeter of the ellipse. Thus, in any case the location of the inflection point can be found geometrically by drawing a straight line from P_0 through $(\beta_1, \beta_2) = (1, 1)$. The t_e value where this line intersects the ellipse then is to be identified with t_w, the position of the inflection point.

Now we can quickly discuss all the possible cases. In the following we refer to the interval ($0 \le t \le 1$) as I.

1) Quadrant I: $0 \le \beta_1; 0 \le \beta_2$: Here we have 4 different cases.

1990A&A...239..443S

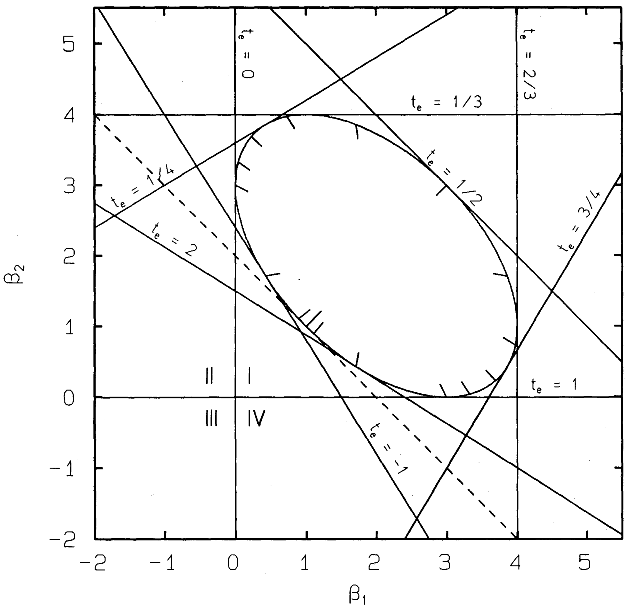

Fig. 9. On these straight lines drawn in the β_1/β_2-plane the position of one extremum, t_e, of the corresponding cubic polynomial is constant. The complete set of straight lines generated by $-\infty \le t_e \le +\infty$ has the indicated ellipse as a common envelope. Starting at the long tick mark at $(\beta_1, \beta_2) = (1, 1)$ and proceeding clockwise, the straight lines being tangent to the ellipse at the respective tick mark positions correspond to t_e values of $\pm\infty, -9, -1, 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2,$ and 10. On the dashed line $\beta_1 + \beta_2 = 2$ we have $t_e = \pm\infty$.

a) P_0 outside the ellipse and β_1 or $\beta_2 > 3$: there are two real extrema and both are located inside interval I, the inflection point, lying half way between the extrema, also is to be found inside I.

b) P_0 inside the ellipse: no tangent lines can be drawn which means that no real extrema exist, the cubic function increases or decreases monotonically over the whole range $-\infty \le t \le +\infty$. However, a real inflection point always exists. In most of the ellipse it lies inside interval I, but in the small areas of the ellipse belonging to sectors "B" in Fig. 5, t_w can be anywhere outside I.

c) P_0 on the perimeter of the ellipse: the two real extrema and the inflection point all fall on one point (given by the corresponding t_e), forming a saddle point.

d) P_0 outside the ellipse and $\beta_1, \beta_2 < 3$: the two real extrema are both located outside interval I, while the inflection point can be inside or outside this interval. It is outside I if P_0 lies in the areas belonging to sectors "B" in Fig. 5. The line $\beta_1 + \beta_2 = 2$ is a special case. On this line at least one of the two extrema as well as the inflection point are at $\pm\infty$.

2) Quadrant II: $0 \ge \beta_1$; $0 \le \beta_2$: both extrema are real, one is inside, the other outside interval I. The inflection point can be anywhere between $-\infty$ and $+\infty$, except between $\frac{1}{3}$ and $\frac{1}{2}$.

3) Quadrant III: $0 \ge \beta_1$; $0 \ge \beta_2$: both extrema are real and lie inside interval I, but not inside the interval $\frac{1}{3} \le t \le \frac{2}{3}$. The inflection point, however, is located between $\frac{1}{3} \le t_w \le \frac{2}{3}$.

4) Quadrant IV: $0 \le \beta_1$; $0 \ge \beta_2$: this quadrant is a mirror image of quadrant II and we find essentially the same conditions. Here the inflection point can be found anywhere except between $\frac{1}{2}$ and $\frac{2}{3}$.

Let us finally note that this geometrical method for the determination of the extrema and the inflection point is also applicable (with a small modification) if $s = 0$. In this case first draw a "direction line" through $(\beta_1, \beta_2) = (0, 0)$ with slope $\beta_2/\beta_1 = y_2'/y_1'$. The two possible straight lines parallel to this line and tangent to the ellipse will give the two values for t_e, and the intersection of the straight line through $(\beta_1, \beta_2) = (1, 1)$ and parallel to the "direction line" with the ellipse gives the position of the inflection point t_w. The detailed discussion of the case $s = 0$ is left to the reader.

This article was processed by the author using Springer-Verlag TEX AA macro package 1989.

© European Southern Observatory • Provided by the NASA Astrophysics Data System