On the Condition of Piecewise Polynomial Finite Element Bases

by

J. M. Varah
Computer Science Department
University of British Columbia
Vancouver, B. C. Canada

1. Introduction

There has been a great deal of interest lately in the approximate solution of various problems by so-called global methods: that is to find a solution of the form

$$\sum_{1}^{n} c_i \phi_i(x)$$

where the $\phi_i(x)$ are given functions. Data fitting problems and boundary value problems have both been successfully treated in this way, particularly when the basis funcitons $\{\phi_i(x)\}$ are piecewise polynomials with support over a small region in x. We refer to Schultz [2] and Strang and Fix [3] for a general discussion of such methods.

Of crucial importance in computing with these bases is their condition, or "amount of linear dependence". This can be measured by the condition number of the Gram matrix (mass matrix in [3]):

$$G_{ij} = \int \phi_i(x)\phi_j(x)dx.$$

Since G is a positive definite symmetric matrix, the condition number in the

l_2 norm is $\kappa(G)=\lambda max(G)/\lambda min(G)$. We shall use this throughout the paper. Of course, if the basis functions are orthogonal over the whole x-domain, $\kappa=1$; however when we demand that the support of each basis function be restricted to a small region, we no longer have orthogonality and the questions of the condition of various bases becomes interesting.

In this paper we consider the condition of the most common piecewise polynomial bases: cubic splines and piecewise cubic Hermite polynomials. For the former, this is merely a matter of direct computation; however for the latter there is still some choice to be made, and we investigate the problem of minimizing the condition.

2. The General Method

For a given general set of abscissas in the x-domain, the piecewise polynomial basis functions over these abscissas will vary throughout the interval and the Gram matrix, although banded, will have elements varying in size depending on the spacing of the abscissas. In order to isolate the effect of which basis is chosen, we shall consider only equally spaced abscissas. Then, as in [3,pg 209ff], the Gram matrix is Toeplitz, or block Toeplitz, and its condition is more readily discernable. Indeed, since the mesh spacing h only appears as a common factor, it does not influence the condition, so the important consideration is the condition of the doubly infinite Toeplitz (or block Toeplitz) Gram matrix. This corresponds to either a doubly infinite x-domain with fixed h, or a finite x-domain with $h\to 0$.

Now let us concentrate on a spline basis; these are piecewise polynomials of degree 2n-1, and continuity 2n-2 at the abscissas, with support over 2n intervals (see for example deBoor [1]). Since there is one basis function for each abscissa, each basis function $\phi_i(x)$ is a translated scaled copy of one basic function $S_n(x)$, centred at 0 with support (-n,n). Thus the Gram matrix G_n has the form

$$H_n = \frac{G_n}{h} = \begin{pmatrix} a_0 & a_1 & \cdots & a_{2n-1} & & 0 \\ a_1 & a_0 & \ddots & \ddots & \ddots & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \\ a_{2n-1} & & \ddots & \ddots & \ddots & \\ & \ddots & \ddots & \ddots & \ddots & a_{2n-1} \\ 0 & & & \ddots & a_1 & a_0 \end{pmatrix} \qquad (2.1)$$

where $a_i = \int S_n(x) S_n(x-i)\,dx$, i=0,...2n-1. The spectrum of the doubly infinite version of this Toeplitz matrix is well-known:

$$Sp(H_n) = \{p(\theta) = a_0 + 2 \sum_{1}^{2n-1} a_j \cos j\theta\}.$$

Thus $\kappa(H_n) = \dfrac{\max\limits_{\theta} p(\theta)}{\min\limits_{\theta} p(\theta)}$ and $0 < \kappa(H_n) < \infty$ since H_n is positive definite so $p(\theta) > 0$.

For n=1 (piecewise linear functions), $S_n(x)$ is the familiar roof function; as in [3,pg.211] we have $a_0=4$, $a_1=1$, and

$$\kappa(H_1) = \frac{\max\limits_{\theta} (4+2\cos\theta)}{\min\limits_{\theta} (4+2\cos\theta)} = 3.$$

For n=2 (cubic splines), $S_2(x)$ is the basic cubic spline (see Schultz [2,pg 73])

$$S_2(x) = \begin{cases} (2-x)^3, & 1 \le x \le 2 \\ 3x^3 - 6x^2 + 4, & 0 \le x \le 1 \\ S(-x), & x < 0 \end{cases}$$

A short computation gives

$$a_0 = 17\tfrac{9}{35},\ a_1 = 7\tfrac{23}{28},\ a_2 = \tfrac{6}{7},\ a_3 = \tfrac{1}{140}\ .$$

The corresponding $p(\theta)$ has its maximum at $\theta=0$: $p(0) = 34\frac{22}{35}$, and its minimum at $\theta=\pi$: $p(\pi) = 3\frac{11}{35}$. Thus

$$\kappa(H_2) = \frac{303}{29} \cong 10.45.$$

Estimates of condition numbers for higher dimensional spline bases can be found in de Boor [1].

3. The Cubic Hermite Polynomials

The general piecewise Hermite polynomials have degree 2n-1 and continuity n-1 at the abscissas; because of this lower continuity, there are more basis

functions: n associated with each abscissa. The so-called natural Hermite basis, obtained by Hermite interpolation of the delta function at successive abscissas, gives the smallest support possible, namely two subintervals. For n=2, for equally spaced abscissas, the basis functions are translated scaled copies of two functions defined over [-1,1] (see Schultz [2,pg.27]):

$$H^{(0)}(x) = \begin{cases} (2x+1)(x-1)^2, & 0 \le x \le 1 \\ H^{(0)}(-x), & -1 \le x \le 0 \end{cases}, \quad H^{(1)}(x) = \begin{cases} x(x-1)^2, & 0 \le x \le 1 \\ -H^{(1)}(-x), & -1 \le x \le 0 \end{cases} \tag{3.1}$$

Of course, this is not the only basis with minimal support; we could use translated scaled copies of any linear combination of these functions, say

$$B^{(0)} = H^{(0)} + \alpha H^{(1)}, \quad B^{(1)} = (\beta H^{(0)} + H^{(1)})s, \tag{3.2}$$

where we have included a scaling factor s as well.
One choice used in practice is $\alpha=-3$, $\beta=1/3$, $s=3$, which gives the B-spline basis:

$$B^{(0)}(x) = (1-x)^3, \quad 0 \le x \le 1, \qquad B^{(1)}(x) = (x-1)^2(5x+1), \quad 0 \le x \le 1$$
$$= B^{(1)}(-x), \quad -1 \le x \le 0 \qquad = B^{(0)}(-x), \quad -1 \le x \le 0.$$

These have the property of being positive throughout the interval of support.

In what follows we shall discuss the problem of choosing α,β,s to minimize the condition number of the Gram matrix. Because of the two basic basis functions and the fact that the support is two subintervals, the Gram matrix has the block-Toeplitz form

$$G = h \begin{pmatrix} A & C & & \\ C^T & A & C & \\ & \ddots & \ddots & \ddots \end{pmatrix}$$

where A and C are 2x2 blocks. (For the general Hermite case, A and C are nxn.) The spectrum of the doubly infinite version of this is given by the set of eigenvalues of the 2x2 positive definite Hermitian matrix

$$P(\theta) = A + C\, e^{i\theta} + C^T e^{-i\theta}.$$

$$\text{Thus} \quad \kappa(\alpha,\beta,s) = \frac{\max_{\theta} \lambda_1 (P(\theta))}{\min_{\theta} \lambda_2 (P(\theta))} . \tag{3.3}$$

First we consider $\alpha=\beta=0$; that is, using the natural Hermite basis with some scaling of the second function $H^{(1)}(x)$. The basic Gram matrix (with s=1) has, after a simple computation with (3.1),

$$A=\frac{1}{420}\begin{pmatrix} 312 & 0 \\ 0 & 8 \end{pmatrix}, \qquad C=\frac{1}{420}\begin{pmatrix} 54 & -13 \\ +13 & -3 \end{pmatrix}$$

This gives, except for a constant,

$$P(\theta)\equiv P_0(\theta)=\begin{pmatrix} 156 + 54\cos\theta & -13i\sin\theta \\ 13i\sin\theta & 4-3\cos\theta \end{pmatrix}$$

Since we are scaling the second basis function (see(3.2)), and this function affects the second row and column of A and C, we obtain

$$A(s)=DAD^T \ , \ C(s)=DCD^T, \ P(s;\theta)=DP_0(\theta)D^T$$

where $D=\begin{pmatrix} 1 & 0 \\ 0 & s \end{pmatrix}$.

Theorem 3.1: The scaled natural cubic Hermite basis (3.1) has condition $\kappa(s)\geq 7$. This minimum is achieved for $s_1^2< s^2<s_2^2$, where $s_1^2\approx 32$, $s_2^2\approx 96$.

Proof:

From the above, we have

$$P(\theta;s) =\begin{pmatrix} 156 + 54\cos\theta & (-13i\sin\theta)s \\ (13i\sin\theta)s & (4-3\cos\theta)s^2 \end{pmatrix} \tag{3.4}$$

Since $P(\theta;s)$ is positive definite, we can denote the larger by $\lambda_1(\theta;s)$, the smaller by $\lambda_2(\theta;s)$ and both are always positive. Since $\lambda_1(\theta;s)$ is larger than any diagonal element, and $\lambda_2(\theta;s)$ is correspondingly smaller,

$$\lambda_2(\theta;s) \leq (4-3\cos\theta)s^2 \leq \lambda_1(\theta;s).$$

Thus

$$\kappa(s) \frac{\max\limits_{\theta} \lambda_1(\theta;s)}{\min\limits_{\theta} \lambda_2(\theta;s)} \geq \frac{\max\limits_{\theta} (4-3\cos\theta)s^2}{\min\limits_{\theta} (4-3\cos\theta)s^2} = 7.$$

To show the range where this is actually achieved is more difficult. We have explicitly

$$2\lambda_1(\theta;s)=F+\sqrt{D}\ ,\quad 2\lambda_2(\theta;s)=F-\sqrt{D}\ ,$$

$$F=4(39+s^2)+3\cos\theta(18-s^2)$$

$$D=[4(39-s^2)+3\cos(18+s^2)]^2+676s^2\sin^2\theta.$$

First we compute these eigenvalues at the endpoints $\theta=0$ and π. (Since they are functions of $\cos\theta$, we need only consider the range $0\leq\theta\leq\pi$.) We have

$$2\lambda_1(0)=210+s^2+|210-s^2|\ ,\quad 2\lambda_2(0)=210+s^2-|210-s^2|$$

$$2\lambda_1(\pi)=102+7s^2+|102-7s^2|,\quad 2\lambda_2(\pi)=102+7s^2-|102-7s^2|.$$

Thus

$$2\lambda_1(0) = \max(420,2s^2),\quad 2\lambda_2(0) = \min(420,2s^2)$$

$$2\lambda_1(\pi) = \max(204,14s^2),\quad 2\lambda_2(\pi) = \min(204,14s^2).$$

Hence

$$\max(2\lambda_1(0),2\lambda_1(\pi)) = 420 (=2\lambda_1(0)) \text{ for } s^2\leq 30$$
$$= 14s^2 (=2\lambda_1(\pi)) \text{ for } s^2\geq 30$$

$$\min(2\lambda_2(0),2\lambda_2(\pi)) = 2s^2 (=2\lambda_2(0)) \text{ for } s^2\leq 102$$
$$= 204 (=2\lambda_2(\pi)) \text{ for } s^2\geq 102$$

So we see immediately that for $s^2<30$ and $s^2>102$,

$$\kappa(s) = \frac{\max\limits_{\theta} \lambda_1(\theta;s)}{\min\limits_{\theta} \lambda_2(\theta;s)} \geq \frac{\max(\lambda_1(0),\lambda_1(\pi))}{\min(\lambda_2(0),\lambda_2(\pi))} > 7.$$

However the right-hand ratio is exactly 7 for $30\leq s^2\leq 102$, and this will equal $\kappa(s)$ if the max and min are achieved at the endpoints.

To examine this, rewrite D as a function of $\cos\theta$:

$$D = a\cos^2\theta + 2b\cos\theta + c,$$

with $a = 16(39-s^2)^2+26^2s^2$, $b=12(18+s^2)(39-s^2)$, $c=9(18+s^2)^2 -26^2s^2$.
So as functions of $\cos\theta$, for fixed s,

$$\lambda_1' = F' + \tfrac{1}{2}\frac{D'}{\sqrt{D}}, \quad \lambda_2' = F' - \tfrac{1}{2}\frac{D'}{\sqrt{D}}$$

$$\lambda_{1,2}'' = \pm D^{-3/2}\left(D\frac{D''}{2} - (D'/2)^2\right).$$

Computing this term in parentheses, we find

$$\lambda_{1,2}''(\cos\theta) = \pm D^{-3/2}(ac-b^2).$$

So the signs of λ_1'' , λ_2'' are fixed for all θ, and they are opposite. In other words, for each fixed s, one eigenvalue is a convex function of $\cos\theta$ and the other is concave. And a brief computation gives

$$ac-b^2 = -7\cdot 26^2 s^2(s^4-128s^2+36\cdot 85).$$

Thus when the quadratic in s^2 is negative (which occurs for $s_1^2 \le s^2 \le s_2^2$, $s_1^2 = 64-2\sqrt{259} \cong 32$, $s_2^2 = 64+2\sqrt{259} \cong 96$), $\lambda_1(\cos\theta)$ is convex and $\lambda_2(\cos\theta)$ is concave; thus in this region the max of λ_1 and min of λ_2 must occur and the endpoints ($\theta=0$ or π).

Thus for $s_1^2 < s^2 < s_2^2$,

$$\kappa(s) = \frac{\max(\lambda_1(0),\lambda_1(\pi))}{\min(\lambda_2(0),\lambda_2(\pi))} = \frac{\lambda_1(\pi)}{\lambda_2(0)} = 7. \qquad \text{QED}$$

Now we consider the general basis (3.2). Because of the way this is formed, the corresponding P-matrix

$$P(\alpha,\beta,s;\theta) = S\,P_0(\theta)\,S^T, \quad S = \begin{pmatrix} 1 & \alpha \\ \beta s & s \end{pmatrix},$$

and the problem now is to find

$$\min_{\alpha,\beta,s}\kappa(P) = \min_{\alpha,\beta,s}\left[\frac{\max_\theta \lambda_1(SP_0(\theta)S^T)}{\min_\theta \lambda_2(SP_0(\theta)S^T)}\right] \qquad (3.5)$$

Theorem 3.2:

For the general cubic Hermite basis (3.2), the condition number $\kappa(P) \geq 7$.

Proof: First we decompose S into QR factors (Q orthogonal, R upper triangular), so that $P = QRP_oR^TQ^T$, and we can reduce the problem to

$$\min_{\alpha,\beta,s} \kappa(Q^TPQ) = \min_{\alpha,\beta,s} \kappa(RP_oR^T)$$

since this matrix has the same eigenvalues as P. In face, we have

$$R = \begin{pmatrix} p & r \\ o & q \end{pmatrix} = \frac{1}{\sqrt{1+\beta^2 s^2}} \begin{pmatrix} 1+\beta^2 s^2 & \alpha+\beta s^2 \\ 0 & s(\alpha\beta-1) \end{pmatrix}.$$

Clearly there is a one to one relationship between the triples (α,β,s) and (p,q,r), providing we keep $p \geq 1$. So we can reformulate our problem (3.5) as

$$\min_{R} \left[\frac{\max_{\theta} \lambda_1(RP_oR^T)}{\min_{\theta} \lambda_2(RP_oR^T)} \right], \tag{3.6}$$

Now we decompose $P_o(\theta)$ into triangular factors $P_o(\theta) = UU^*$.

From the definition of $P_o(\theta)$, this gives

$$U = \begin{pmatrix} x & iy \\ 0 & z \end{pmatrix} = \frac{1}{\sqrt{8-6\cos\theta}} \begin{pmatrix} \sqrt{7(65-36\cos\theta+\cos^2\theta)} - 13i\sin\theta & \\ 0 & 4-3\cos\theta \end{pmatrix}$$

Thus $RP_oR^T = (RU)(RU)^*$, with

$$RU = \begin{pmatrix} px & rz+ipy \\ 0 & qz \end{pmatrix}.$$

Now $\max_{\theta} \lambda_1(RP_oR^T) = \max_{\theta} ||RU||_2^2 \geq \max_{\theta} (qz(\theta))^2$

and $\min_{\theta} \lambda_2(RP_oR^T) = \min_{\theta} ||(RU)^{-1}||_2^{-2} \leq \min_{\theta} (qz(\theta))^2$

Thus $\min_R \kappa(P) \geq \dfrac{\max_\theta (z(\theta))^2}{\min_\theta (z(\theta))^2} = \dfrac{\max_\theta (4-3\cos\theta)}{\min_\theta (4-3\cos\theta)} = 7.$ QED

This minimum condition of 7 is in fact attained for other choices of α,β, and s. Suppose α and β are of opposite sign and we choose $s^2=s_o^2=-\alpha/\beta(>0)$. Then the matrix R used in the proof of Theorem 3.2 is

$$R = \sqrt{1-\alpha\beta}\begin{pmatrix} 1 & 0 \\ 0 & -s_o \end{pmatrix}$$

and the corresponding condition number κ is the same as that for the diagonally scaled matrix

$$\begin{pmatrix} 1 & 0 \\ 0 & s_o \end{pmatrix} P_o(\theta) \begin{pmatrix} 1 & 0 \\ 0 & s_o \end{pmatrix},$$

which is $P(\theta;s_o)$ in (3.4). Thus from Theorem 3.1, if $s_1^2<s_0^2=-\alpha/\beta<s_2^2$, this scaling produces a matrix with $\kappa=7$.

Finally, we return to the B-spline basis ($\alpha=-3,\beta=1/3,s=3$). It is fairly easy to check that this value of s does give the minimal condition for this choice of α and β (the two basic basis functions $B^{(o)}(x)$, $B^{(o)}(x)$ are then mirror images). And since $s^2=-\alpha/\beta$, this condition number is the same as the diagonally scaled matrix $P(\theta;3)$ in (3.4). From the proof of Theorem 3.1, the eigenvalues of this are

$$2\lambda_1(\theta;3)=F+\sqrt{D}\ , \quad 2\lambda_2(\theta;3)=F-\sqrt{D},$$

where

$$F=192+27\cos\theta$$

$$D=(120+81\cos\theta)^2 + 78^2\sin^2\theta$$

A brief computation shows that $\lambda_1(\theta;3)$ is maximized at $\theta=0$ and $\lambda_2(\theta;3)$ is minimized at $\theta=0$, giving

$$\kappa(P(\theta;3))=\frac{70}{3}\ .$$

Thus the condition of the B-spline basis is higher than the scaled natural cubic Hermite basis.

References

1. Carl de Boor, On calculating with B-splines. J. Approx. Thy. 6(1972), 50-62.
2. M.H. Schultz, Spline Analysis. Prentice-Hall, New York, 1973.
3. G. Strang and G. Fix, An Analysis of the Finite Element Method. Prentice-Hall, New York, 1973.