ABSTRACT

BLENDING-FUNCTION TECHNIQUES WITH APPLICATIONS TO DISCRETE LEAST SQUARES

By

Dale Russel Doty

The theory of blending function spaces (bivariate interpolation) is developed in the general setting of interpolation spaces. In this setting it is shown that blending function spaces have the desirable quality of doubling the order of accuracy with less computation when compared to standard tensor product spaces.

The dimensionality of discretized blending function spaces is derived, and several bases are explicitly constructed. The special example of Hermite spline blended piecewise polynomials is developed, showing that these spaces have bases with small support which are easy to calculate. These spaces offer maximum order of convergence for a minimum number of basis elements. For example, linearly blended piecewise cubic polynomials offer a fourth order approximation scheme, and cubic Hermite spline blended piecewise polynomials offer an eighth order scheme.

Next, using the exponential decay of the natural cubic cardinal splines and natural cubic spline blending, a derivative-free approximation scheme is developed, which is eighth order in the interior of the domain.

Algorithms with corresponding error estimates are given for solving the discrete least squares problem with unstructured data. For the univariate case, algorithms are developed using the space of cubic splines. The resulting error analysis indicates the necessary restrictions to be placed on the number and distribution of the data points to insure that the discrete least squares fit will be $O(h^m)$ to a function $f \in C^m[a, b]$ from which the data arises, where h is the mesh size and $1 \leq m \leq 4$. An example is given to illustrate that the discrete least squares fit need not be close to f if these conditions are not realized. For the bivariate case, algorithms and error analyses are given for the spaces of bicubic splines and discretized blending function spaces. It is shown that the discrete least squares fit to a bivariate function f is of the same order accuracy as the corresponding interpolation accuracy.

Discrete least squares is considered on general domains which have curved boundaries and are possibly multiply connected. This general domain is subdivided into "standard" subdomains, and explicit mappings from the unit square to these standard subdomains are constructed which are one-one, onto, and have easily calculated

inverses. Thus, discrete least squares over general domains reduces to the cases previously considered.

Finally, an extensive computational error analysis is given for a constrained least squares algorithm.

BLENDING-FUNCTION TECHNIQUES WITH APPLICATIONS TO DISCRETE LEAST SQUARES

By

Dale Russel Doty

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Mathematics

1975

ACKNOWLEDGMENTS

I wish to express my appreciation to Professor Gerald D. Taylor for his encouragement and guidance during the preparation of this thesis.

TABLE OF CONTENTS

Page

LIST OF FIGURES

CHAPTER 1

COMPUTATIONAL ERROR ANALYSIS FOR DISCRETE LEAST SQUARES

The following chapters will deal with solving the constrained least squares problems which arise from applications of Blending Function spaces.

Due to computational considerations, it is necessary to use methods of orthogonal factorization and pseudo-inverses to obtain solutions to these problems. It is therefore necessary to perform a perturbational and computational analysis of these methods to show how the numerical solutions are affected. We would hope that the condition numbers, χ, associated with our least squares and constraint equations, appear only to the first power in the error analysis, because our solutions contain only inverses to the first power. This, however, has been shown not to be the case, quoting van der Sluis [31]:

> "It caused something of a shock, therefore, when in 1966 Golub and Wilkinson asserted that already the multiplications QA and QB may produce errors in the solution containing a factor $\chi^2(A)$."

We will prove that the condition numbers associated with the constrained least squares equations appear only linearly in the error analysis except for the coefficient of the residual.

Section 1. Least Squares With Constraints

We are given the two matrices

(1.1) $A \in \mathbb{R}^{m \times n}$, and

(1.2) $C \in \mathbb{R}^{r \times n}$,

where $r < n$ and $m \geqslant n$.

Then we seek a solution $x \in \mathbb{R}^n$ which minimizes the norm of R

(1.3) $R = Ax - f$, where $f \in \mathbb{R}^m$,

and satisfies

(1.4) $Cx = g$, where $g \in \mathbb{R}^r$.

We can think of (1.3) as a least squares problem subject to the constraint equations given in (1.4).

We will develop here the method of Halliday and Hayes [21] for finding a solution to (1.3) and (1.4). But first we need to state some of the pertinent theories of factorization by orthogonal transformations. Householder [22] developed the theory of factorization into orthogonal transformations. Precisely, he has shown that if at step i the matrix C_i has the form

(1.5) $$C_i = \left[\begin{array}{c|c} L_i & 0 \\ \hline T_i & \end{array}\right],$$

where $L_i \in \mathbb{R}^{(i-1)\times(i-1)}$ is a lower-triangular matrix, $T_i \in \mathbb{R}^{(r-(i-1))\times n}$ is a rectangular matrix, and $C_1 = C$, then the orthogonal matrix $P_i \in \mathbb{R}^{n \times n}$ which introduces zeros into row i, from i+1 to n of the

matrix C_{i+1}, and leaves the first i-1 columns of C_i unaltered is given by

$$P_i = I - v_i v_i^T / H_i , \tag{1.6}$$

where

$$C_{i+1} = C_i P_i . \tag{1.7}$$

If we represent the entries of C_i by c^i_{jk} for $1 \leqslant j \leqslant r$, $1 \leqslant k \leqslant n$, then the vector $v_i \in \mathbb{R}^n$ is given by

$$v_i^T = (0, \cdots, 0, c^i_{ii} + \operatorname{sgn}(c^i_{ii}) S_i, c^i_{i,i+1}, \cdots, c^i_{i,n}) . \tag{1.8}$$

where sgn $(\cdot)$ is a function of a real variable defined by

$$\operatorname{sgn}(x) = \begin{cases} +1 & \text{if } x \geqslant 0 \\ -1 & \text{if } x < 0 \end{cases} . \tag{1.9}$$

S_i is defined by

$$S_i = \left[\sum_{j=1}^{n} (c^i_{ij})^2 \right]^{\frac{1}{2}} . \tag{1.10}$$

Finally H_i is defined by

$$H_i = S_i^2 + \left| c^i_{ii} \right| S_i . \tag{1.11}$$

Also, he has shown that

$$\left| c^{i+1}_{i,i} \right| = S_i , \tag{1.12}$$

or that the new diagonal element has the Euclidean length of the old row i, from i to n. This is a property of orthogonal

transformations, that the Euclidean length of any transformed row remains invariant.

If for some i, $S_i = 0$, then row i of C_i is zero from i to n, and we take $P_i = I$. If this is not the case, then both S_i and H_i are strictly positive. In either case, the factorization proceeds until the completion of step r, and we have a matrix, C_{r+1}, which is lower triangular

$$[L|0] = C_{r+1}, \text{ where } L = L_{r+1} . \tag{1.13}$$

Define Q by

$$Q = P_1 \cdot P_2 \cdot \cdots \cdot P_r . \tag{1.14}$$

Then Q is also an orthogonal matrix, i.e., $Q^{-1} = Q^T$, and it follows from (1.7) that

$$[L|0] = CQ, \text{ or} \tag{1.15a}$$

$$L = C Q_1 , \tag{1.15b}$$

where Q_1 is the first r columns of Q,

$$Q_1 = Q \begin{bmatrix} I_r \\ \hline 0 \end{bmatrix} . \tag{1.15c}$$

Then we may write (1.4) as

$$[L|0] \; Q^T x = g. \tag{1.16}$$

Section 1.1. Pivoting

Row pivoting can be included in the above algorithm for factoring C. This is done in the following manner. At step i of the factorization, row k, where $i \leqslant k \leqslant r$, is chosen as the next pivotal row by some pivotal strategy. Then we premultiply by a matrix $E_i \in \mathbb{R}^{rxr}$, which interchanges the rows i and k, see Deskins [7, p. 551]. We now factor P_i^{-1} out of the matrix $E_i C_i$ and define

$$C_{i+1} = (E_i C_i) P_i \,, \tag{1.17}$$

where P_i is obtained from row i of $E_i C_i$. We can conclude from (1.4), (1.14) and (1.17) that

$$\begin{aligned} [L|0]\, Q^T x &= E_r \cdot \cdots \cdot E_1 g \\ &= E_r \cdot \cdots \cdot E_1 C x \,. \end{aligned} \tag{1.18}$$

Therefore, for theoretical purposes, we will assume that the matrix C and vector g with which we are working, have had their rows or elements permuted beforehand. So that, when we apply our pivotal strategy, the pivotal rows will be chosen sequentially from 1 to r.

The type of pivoting that is usually used in practice is called "maximal row pivoting". At step i, this type of pivoting chooses the next pivot by the criterion that S_i is maximized. If two or more rows give the same value, then the first of these is chosen as the

pivotal row. For this type of pivoting, Golub [11] has shown that

(1.19) $$S_1 \geqslant S_2 \geqslant \cdots \geqslant S_r \geqslant 0 .$$

If for some i, $S_i = 0$, then from (1.12), L will have a zero at position i on its diagonal, which means that the rank of L will be less than r. But from (1.15a) we have $C = [L|0]\, Q^T$ which implies that the rank of C will also be less than r, see Deskins [7, p. 550]. Therefore, if we assume that C is of full rank r, then for $1 \leqslant i \leqslant r$ we must have

(1.20) $$S_i > 0 ,$$

and because the S_i for $1 \leqslant i \leqslant r$ are the absolute value of the diagonal entries of L we have

(1.21) L^{-1} exists.

It should be noted that if $r \geqslant n$, then C^T can be factored in the following way

$$C^T P_1 \cdot \cdots \cdot P_n = [L|0],$$

where by taking transposes and using $P_i^T = P_i \in R^{rxr}$

$$P_n \cdot \cdots \cdot P_1 C = \left[\frac{W}{0}\right] ,$$

and $W = L^T$ is an upper-triangular matrix. We will use both types of factorizations in what follows.

Section 1.2. A Constrained Least Squares Algorithm

Next, we shall describe an algorithm for the solution of constrained least squares which was developed by Hayes and Halliday [21]. This will be presented in detail, because we need their formulas for later reference.

Algorithm 1.1. (Hayes and Halliday [21]). If we have a least squares problem with constraints, as defined in (1.1)-(1.4), where C is of full rank, then the solution x is obtained as follows.

Step 1: Let C be of full rank, then for $1 \leq i \leq r$, we have from (1.20) that $S_i > 0$ and L is nonsingular. Thus there exists a Q given by (1.14) such that in (1.16)

$$[L \mid 0]\ Q^T x = g . \tag{1.16}$$

Step 2: Define the vectors $d_1 \in \mathbb{R}^r$ and $d_2 \in \mathbb{R}^{n-r}$ such that

$$\begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = Q^T x . \tag{1.23}$$

Then from (1.16) and (1.23) we have

$$d_1 = L^{-1} g . \tag{1.24}$$

Step 3: Solve for d_2, which upon inverting (1.23) will yield a solution x

$$x = Q \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} \overset{8}{.} \tag{1.25}$$

Toward this end, using (1.3) and (1.23) we have

$$\begin{aligned} A x &= A Q Q^T x \\ &= [B_1 \mid B_2] Q^T x \\ &= B_1 d_1 + B_2 d_2 \\ &= f + R , \end{aligned} \tag{1.26}$$

where B_1 is the first r columns, and B_2 the remaining n-r columns of the matrix A Q

$$[B_1 \mid B_2] = A Q . \tag{1.27}$$

From (1.26) we have

$$B_2 d_2 = f - B_1 d_1 + R . \tag{1.28}$$

Let B_2 be of full rank, then the factorization of B_2, using orthogonal transformations, proceeds in the following way. Define

$$B_{2,1} = B_2 \tag{1.29}$$

and for $1 \leq i \leq n-r$

$$B_{2,i+1} = U_i B_{2,i} , \tag{1.30}$$

where at step i, the orthogonal transformation $U_i \in \mathbb{R}^{m \times m}$ is defined

by (1.6), where $B_{2,i}^T$ plays the role of C_i in (1.8)-(1.12), and we are factoring from the left. Define

$$V = U_{n-r} \cdot \ldots \cdot U_1 \tag{1.31}$$

and

$$\left[\frac{W}{0}\right] = V B_2 = B_{2,m+1} \tag{1.32a}$$

or

$$W = V_1 B_2 , \tag{1.32b}$$

where from (1.20) and (1.21), $W \in \mathbb{R}^{(n-r)x(n-r)}$ is an upper triangular nonsingular matrix. Then the solution of (1.28) which minimizes R, [see 21], is given by

$$d_2 = W^{-1} V_1 \{f - B_1 d_1\} , \tag{1.33}$$

where V_1 is the first n-r rows of V

$$V_1 = [I_{n-r} | 0] V . \tag{1.34}$$

It is shown by Peters and Wilkinson [28] that d_2 as defined (1.33) is unique. Then by using (1.24), (1.33) and (1.25) we have the solution

$$x = Q \begin{bmatrix} L^{-1} g \\ \\ W^{-1} V_1 (f - B_1 L^{-1} g) \end{bmatrix} . \tag{1.35}$$

Section 2. Defining the Perturbation Problem

In this section, we shall derive an upper bound on the computational errors of the solution defined by Algorithm 1.1. Toward this end, we will develop a perturbational analysis of the problem defined in (1.1) to (1.4). Due to calculation and rounding errors, the problem which is stored in the computer is not the exact matrices and vectors as defined in (1.3) and (1.4). Instead, the original quantities have been perturbed to give the new problem:

(2.1) We are given the two matrices $\hat{A} \in \mathbb{R}^{m \times n}$ and $\hat{C} \in \mathbb{R}^{r \times n}$ where

(2.2) $r < n$ and $n \leqslant m$.

Then the solution $\hat{x}$ of the perturbed problem is the vector which minimizes the norm of

(2.3) $\hat{R} = \hat{A}\hat{x} - \hat{f}$, where $\hat{f} \in \mathbb{R}^{m}$,

and we require that $\hat{x}$ satisfy the constraint

(2.4) $\hat{C}\hat{x} = \hat{g}$, where $\hat{g} \in \mathbb{R}^{r}$.

We will denote with a "hat" all of the perturbed quantities, and their meaning will be the same as in Section 1. If $\hat{C}$ and $\hat{B}_2$ are of full rank, then Algorithm 1.1 gives the solution $\hat{x}$

$$\hat{x} = \hat{Q} \begin{bmatrix} \hat{L}^{-1}\hat{g} \\ \hat{W}^{-1}\hat{V}_1 \{\hat{f} - \hat{B}_1 \hat{L}^{-1}\hat{g}\} \end{bmatrix}. \tag{2.5}$$

We now define the perturbational quantities δA, δf, δC, δg, δP_i, δL etc. as the difference between a perturbed and unperturbed quantity. For example, $\delta A = \hat{A} - A$, $\delta f = \hat{f} - f$, etc.

Before we can estimate $||\hat{x} - x||$, we need some well known facts from the theory of matrices. Due to the lack of references for the following norm relations for non-square matrices, we will include in the next section a somewhat detailed development for completeness.

Section 3. Matrix Norm Inequalities

Definition 3.1: Let $A \in \mathbb{R}^{n \times n}$, then the complex number $\lambda_i(A)$ is an eigenvalue of the matrix A if and only if there exists a non-zero eigenvector $x_i \in \mathbb{C}^n$ such that $A x_i = \lambda_i(A) x_i$. The spectral radius $\rho(\cdot)$ of A is defined to be $\rho(A) = \max |\lambda_i(A)|$, see Varga [32, p. 9], and let $\mu(\cdot)$ of A be defined by $\mu(A) = \min_i \{ |\lambda_i(A)| \mid \lambda_i(A) \neq 0 \}$.

Definition 3.2: The Euclidean norm $||\cdot||$ of a vector $x \in \mathbb{R}^m$ is defined to be $||x|| = (\sum_{i=1}^{m} (x_i)^2)^{\frac{1}{2}}$. The Euclidean (Schur) norm $||\cdot||_E$ of a rectangular matrix $A \in \mathbb{R}^{n \times m}$ is defined to be $||A||_E = \left[\sum_{i=1}^{n} \sum_{j=1}^{m} (a_{ij})^2 \right]^{\frac{1}{2}}$, see [36, p. 81].

Definition 3.3: The spectral matrix norm $||\cdot||$ of a rectangular matrix $A \in \mathbb{R}^{n \times m}$, induced by the Euclidean vector norm, is defined to be $||A|| = \sup_{\substack{||x||=1 \\ x \in \mathbb{R}^m}} ||Ax||$, see Varga [32, p. 9].

Remark: For a vector $x \in \mathbb{R}^m$, the Euclidean vector norm, Euclidean matrix norm and spectral matrix norm are the same, see the following lemma.

Lemma 3.1. Given the rectangular matrix $A \in \mathbb{R}^{n \times m}$ and vector $x \in \mathbb{R}^m$ then

$$||A|| = \left[\rho(A^T A) \right]^{\frac{1}{2}} . \tag{3.1}$$

(3.2a) $\rho(A^T A) = \rho(AA^T)$.

(3.2b) $\mu(A^T A) = \mu(A A^T)$.

(3.3) If $A A^T \in \mathbb{R}^{nxn}$ is nonsingular, then

$||(A A^T)^{-1}|| = 1/\mu(A A^T)$.

(3.4) $||A||_E / \sqrt{n} \leq ||A|| \leq ||A||_E$.

(3.5) $||A^T|| = ||A||$.

(3.6) $||A^T||_E = ||A||_E$.

(3.7) If $Q \in \mathbb{R}^{nxn}$ is orthogonal (i.e. $Q^T = Q^{-1}$), then $||Q|| = 1$.

(3.8) If $Q \in \mathbb{R}^{nxn}$ is orthogonal, then $||QA|| = ||A||$.

(3.9) If $Q \in \mathbb{R}^{mxm}$ is orthogonal, then $||AQ|| = ||A||$.

(3.10) $||Ax|| \leq ||A||\ ||x||$.

(3.11) If $B \in \mathbb{R}^{mxr}$, then $||AB|| \leq ||A||\ ||B||$.

(3.12) If $B \in \mathbb{R}^{nxm}$, then $||A+B|| \leq ||A|| + ||B||$.

(3.13) If $A_1 \in \mathbb{R}^{nx(m-r)}$ is a matrix obtained by deleting the first or last r columns of A, then $||A_1|| \leq ||A||$.

(3.14) If $A_1 \in \mathbb{R}^{(n-r)xm}$ is a matrix obtained by deleting the first or last r rows of A, then $||A_1|| \leq ||A||$.

(3.15) If $Q \in \mathbb{R}^{mxm}$ is an orthogonal matrix and Q_1 is the first r rows or columns of Q, then $||Q_1|| \leq 1$.

(3.16) If $B \in \mathbb{R}^{nxn}$ and $||B|| < 1$, then the following inverses exist and

$||(I + B)^{-1}|| \leq 1/(1 - ||B||)$,

$||(I + B)^{-1} - I|| \leq ||B||/(1 - ||B||)$.

(3.17) If B, $\hat{B}$, $\delta B \in \mathbb{R}^{nxn}$, both B^{-1} and $\hat{B}^{-1}$ exist and $\delta B = \hat{B} - B$, then $\hat{B}^{-1} = B^{-1} - \hat{B}^{-1}\,\delta B\, B^{-1}$.

(3.18) If the vector v is any row or column of A, then $||v|| \leq ||A||$.

Proof (3.1): From Varga [32, p. 11] we have that because $A^T A$ is a non-negative definite (i.e. $x^T A^T A x \geq 0$) symmetric square matrix, the eigenvalues are all real and non-negative. Since $||Ax||^2/||x||^2 = x^T A^T A x / x^T x$, we have $||Ax||^2/||x||^2 \leq \rho(A^T A)$, where equality is taken on for an eigenvector corresponding to $\rho(A^T A)$.

(3.2a), (3.2b): We will show that the set of non-zero eigenvalues of $A^T A \in \mathbb{R}^{mxm}$ is the same as that of $A A^T \in \mathbb{R}^{nxn}$. Let λ be a non-zero eigenvalue of $A^T A$, then λ is real and there exists $x \neq 0$ such that $A^T A x = \lambda x$. Premultiply by A and we have $(A A^T)A x = \lambda A x$. The vector $A x$ is non-zero, because if it were zero we would have $\lambda x = A^T(A x) = 0$, and because $\lambda \neq 0$, this would imply that $x = 0$, which is a contradiction. Therefore, λ is an eigenvalue of $A A^T$. The reverse inclusion is proved in an identical manner.

(3.3): $A A^T$ is positive definite because $x^T A A^T x = ||A^T x||^2 \geq 0$, and for $x \neq 0$ we have, by using an argument similar to the one given in (3.2a) that $||A^T x|| > 0$. Because $A A^T$ is a

square matrix, we can use property G of Wilkinson [34 , p. 290] and conclude our result.

(3.4): If $||x|| = 1$, then $||A\,x||^2 = \sum_{i=1}^{n}\left[\sum_{j=1}^{m} a_{ij}\,x_j\right]^2$

$$\leqslant \sum_{i=1}^{n}\left[\sum_{j=1}^{m} a_{ij}^2 \cdot \sum_{j=1}^{m} x_j^2\right] = ||A||_E^2,$$

where we have used the Schwarz inequality on each term in brackets. Therefore, by using the definition of the spectral norm, we have $||A|| \leqslant ||A||_E$.

To prove the left hand side of the inequality we use (3.1) and (3.2) to obtain $||A||^2 = \rho(A\,A^T)$. Because $A\,A^T \in \mathbb{R}^{n\times n}$, we have from Marcus and Ming [26, p. 23] that $\sum_{i=1}^{n}(A\,A^T)_{ii} = \sum_{i=1}^{n}\lambda_i(A\,A^T)$. Because $x^T A\,A^T x = ||A^T x||^2 \geqslant 0$, and using [26 , p. 69] , we have that $A\,A^T$ is non-negative definite and $\lambda_i(A\,A^T) \geqslant 0$ for $1 \leqslant i \leqslant n$. Therefore $n\cdot\rho(A\,A^T) \geqslant \sum_{i=1}^{n}\lambda_i(A\,A^T) = \sum_{i=1}^{n}(A\,A^T)_{ii} = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2 = ||A||_E^2$, and we have proved (3.4).

(3.5): Use (3.1) and (3.2a).

(3.6): Use the definition of $||A||_E$.

(3.7): From (3.1) we have $||Q||^2 = \rho(Q^T Q) = \rho(I) = 1$, because all eigenvalues of the identity are 1.

(3.8): $||Q\,A||^2 = \rho(Q\,A)^T Q\,A) = \rho(A^T A) = ||A||^2$.

(3.9): Because Q^T is orthogonal, use (3.5) and (3.8).

(3.10): Follows from the definition of $||A||$.

(3.11): Using Varga [32, p. 11], the norm of AB is obtained for an unit eigenvector x corresponding to $\rho((AB)^T(AB))$. Then $||AB|| = ||A(Bx)|| \leqslant ||A||\ ||Bx|| \leqslant ||A||\ ||B||$, where we have used (3.10).

(3.12): Similar to (3.11).

(3.13): If A_1 is obtained by deleting the last r columns of A, then $A_1 = A\left[\frac{I}{0}\right]$, where $I \in \mathbb{R}^{(m-r)x(m-r)}$ is the identity matrix. From (3.11) and (3.1), $||A_1||^2 \leqslant ||A||^2\ \rho([I|0]\left[\frac{I}{0}\right]) = ||A||^2$. The other case is done in the same way.

(3.14): Similar to (3.13).

(3.15): Use (3.7), (3.13) and (3.14).

(3.16): Follows from [9, pp. 112-114].

(3.17): Because $I = \hat{B}^{-1} \cdot \hat{B} = \hat{B}^{-1} \cdot (B + \delta B)$

$= \hat{B}^{-1} \cdot B + \hat{B} \cdot \delta B$, we have $\hat{B}^{-1} \cdot B = I - \hat{B}^{-1} \cdot \delta B$,

from which our result follows.

(3.18): We will assume that $v \in \mathbb{R}^n$ is row ℓ of A, where $1 \leqslant \ell \leqslant n$. Define the vector $e_\ell \in \mathbb{R}^n$ to have zeros in all entries except for a unit in entry ℓ. It follows that $v^T = (e_\ell)^T A$ and $||e_\ell{}^T||^2 = ||e_\ell||^2 = \rho(e_\ell{}^T e_\ell) = \rho(1) = 1$.

■

Section 4. Condition Numbers of Non-square Matrices

Definition 4.1: If $A \in \mathbb{R}^{mxn}$ is any matrix, then the condition number $\chi(\cdot)$ of A is defined to be

$$\chi(A) = \left[\rho(A^T A)/\mu(A^T A)\right]^{\frac{1}{2}} \geq 1, \tag{4.1}$$

if A is of full rank, otherwise $\chi(A)$ is undefined. It should be recalled from Definition 3.1 that $\mu(A^T A)$ is the smallest non-zero eigenvalue of $A^T A$ in absolute value.

There are many definitions for condition numbers, depending upon the type of matrix. We will list some of them here and show that they are equivalent to (4.1).

If $A \in \mathbb{R}^{nxn}$ is a nonsingular matrix, then the condition number $\chi(\cdot)$ of A is defined to be

$$\chi(A) = ||A^{-1}|| \cdot ||A|| \geq 1. \tag{4.2}$$

If A is singular, then $\chi(A)$ is undefined, (see [23, p. 81]).

We would like to show that (4.1) and (4.2) are equivalent for $m = n$. This is clear, because if A is not of full rank, then $\chi(A)$ is undefined in both (4.1) and (4.2). If A is of full rank and square, then A^{-1} exists, and using property (I) of Wilkinson [34, p. 290] and (3.3) we have $||A^{-1}||^2 = ||A^{-1}A^{-T}|| = ||(A^T A)^{-1}|| = 1/\mu(A^T A)$. Thus we have that (4.1) is equivalent to (4.2).

If $A \in \mathbb{R}^{mxn}$ is a rectangular matrix, where $m \geq n$, then the condition number $\chi(\cdot)$ of A is defined to be

$$(4.3) \qquad \chi(A) = \underset{\substack{||x||=1 \\ x \in \mathbb{R}^n}}{\text{Sup}} \; ||A\,x|| \Big/ \underset{\substack{||x||=1 \\ x \in \mathbb{R}^n}}{\text{Inf}} \; ||A\,x|| \geqslant 1,$$

if A is of full rank n, otherwise $\chi(A)$ is undefined, (see Bjorck [3]).

To see that (4.1) and (4.3) are equivalent for $m \geqslant n$, we have from (3.1) and Definition 3.3 that $\underset{||x||=1}{\text{Sup}} \; ||A\,x|| = \rho^{\frac{1}{2}}(A^T A)$. Also, we have that $A^T A \in \mathbb{R}^{nxn}$ is of full rank, because if it were not, there would exist a $x \in \mathbb{R}^n$ such that $x \neq 0$ and $A^T A\,x = 0$. But this implies that $x^T A^T A\,x = ||A\,x||^2 = 0$, and in turn that $A\,x = 0$. This cannot happen if A is of full rank n and $x \neq 0$. Therefore, we have that the smallest eigenvalue of $A^T A$ is non-zero and $\min_{1 \leqslant i \leqslant n} \lambda_i(A^T A) = \mu(A^T A)$. Varga [32, p. 11], has shown that $\min_{1 \leqslant i \leqslant n} \lambda_i(A^T A) \leqslant ||A\,x||^2$ for $||x|| = 1$, where equality is taken on for the unit eigenvector x corresponding to $\mu(A^T A)$. Therefore, $\underset{||x||=1}{\text{Inf}} \; ||A\,x|| = \mu^{\frac{1}{2}}(A^T A)$, and we see that (4.3) is equivalent to (4.1).

Remark: The importance of defining $\mu(A^T A)$ to be the smalles non-zero eigenvalue in absolute value comes from the case where $m < n$. Here we have that $A^T A$ has a zero eigenvalue even if A is of full rank, however, $A\,A^T$ has not. Using (3.2b), we have that $\mu(A^T A) = \mu(A\,A^T) \neq 0$, and our condition number is finite.

Remark: The typical situation in error analysis is that we must bound the following

$$(4.4) \qquad ||A x||\ ||A^{-1} y|| \leqslant \chi(A)\ ||x||\ ||y||,$$

where A is nonsingular and square. Note that the left hand side of (4.4) will be much smaller than the right hand side unless both the vectors x and y are eigenvectors corresponding to $\rho(A^T A)$ and $\mu(A^T A)$ respectively. Therefore, in a practical problem, if A is ill-conditioned, we would expect (4.4) to be a gross overestimate of $||A x||\ ||A^{-1} y||$.

Remark: If Q is orthogonal, then $\chi(Q) = 1$.

Section 5. An Upper Bound for $||\delta Q||$

We will develop here a bound for the norm of the perturbational quantity $\delta Q = \hat{Q} - Q$, where Q is defined by (1.14). It should be noted that Q and $\hat{Q}$ are not the unique orthogonal matrices which reduce C and $\hat{C}$ respectively if $r < n$. For example, if we consider Q, then recalling (1.15a) we have $\left[L|0\right] = C\,Q$. If we post multiply by the orthogonal matrix $P = I - 2uu^T \in \mathbb{R}^{nxn}$, where $u \in \mathbb{R}^n$ is any unit vector which has its first r entries set to zero, then $C(Q\,P) = \left[L|0\right]$. This is true because P will leave unaltered the first r columns of $C\,Q$. Therefore, we cannot expect to obtain a bound for $||\delta Q||$ by manipulating the formula (1.15a). Instead, because Q is uniquely defined by (1.14), we will use the definition of the P_i given in (1.6) to obtain our estimate.

In doing this, we shall use the following equality

$$(5.1) \qquad (1 + K)^s = 1 + \sum_{j=1}^{s} K\,(1 + K)^{j-1},$$

where K is a non-negative real number and s is any natural number.

Also, for the following lemma, we recall that c^k_{kk} and $\hat{c}^k_{kk}$ are the k^{th} diagonal elements of the matrices C_k and $\hat{C}_k$ respectively. The matrix C_k and correspondingly $\hat{C}_k$, as defined in (1.7), represent the factorization at step k.

Lemma 5.1. If we use the factorization method of Section 1 by orthogonal transformation with maximal row pivoting, assume after pivoting that both of the matrices C and $\hat{C}$ have the same ordering of rows, also at step k of the factorization $\operatorname{sgn}(c_{kk}^{k}) = \operatorname{sgn}(\hat{c}_{kk}^{k})$ for $1 \leqslant k \leqslant r$, and finally that there exists a real non-negative number $\triangle$ such that

$$||\delta C||/||C|| \leqslant \triangle \Big/ \left[(9 + 4\sqrt{2} + 8\triangle)^{r-1} \chi(C)\right] , \tag{5.2a}$$

then it follows that

$$||\delta Q|| < K(r)\, \chi(C)\, ||\delta C||/||C||, \tag{5.2b}$$

where $K(r) = (1 + K)^{r} - 1$ and $K = 4(2 + \sqrt{2} + 2\triangle)$.

Remark: In (5.2a) we will use the fact that $9 + 4\sqrt{2} + 8\triangle = 1 + K$.

Proof: To estimate $||\delta Q||$, we express it as the following telescoping series

$$\begin{aligned} ||\delta Q|| &= ||\hat{P}_1 \cdot \hat{P}_2 \cdot \cdots \cdot \hat{P}_r - P_1 \cdot P_2 \cdot \cdots \cdot P_r|| \\ &= ||\delta P_1 \cdot \hat{P}_2 \cdot \cdots \cdot \hat{P}_r + P_1 \cdot \delta P_2 \cdot \hat{P}_3 \cdot \cdots \cdot \hat{P}_r \\ &\quad + \cdots + P_1 \cdot P_2 \cdot \cdots \cdot P_{r-1} \cdot \delta P_r|| \\ &\leqslant \sum_{k=1}^{r} ||\delta P_k||, \end{aligned} \tag{5.3}$$

where we have used (3.8) and (3.9) to obtain the last inequality.

To estimate δP_k we recall the following definitions from sections one and two, which we tabulate here for convenience, $1 \leqslant k \leqslant r$

$$(5.4)\quad C_1 = C,\ C_{k+1} = C_k P_k, \qquad \hat{C}_1 = \hat{C},\ \hat{C}_{k+1} = \hat{C}_k \hat{P}_k,$$

$$(5.5)\quad \left[L \mid 0\right] = C Q, \qquad \left[\hat{L} \mid 0\right] = \hat{C} \hat{Q},$$

$$(5.6)\quad P_k = I - v_k v_k^T / H_k, \qquad \hat{P}_k = I - \hat{v}_k \hat{v}_k^T / \hat{H}_k,$$

$$(5.7)\quad S_k = \left[\sum_{\ell=k}^{n} (c_{k\ell}^k)^2\right]^{\frac{1}{2}}, \qquad \hat{S}_k = \left[\sum_{\ell=k}^{n} (\hat{c}_{k\ell}^k)^2\right]^{\frac{1}{2}},$$

$$(5.8)\quad H_k = S_k^2 + \left| c_{kk}^k \right| S_k, \qquad \hat{H}_k = \hat{S}_k^2 + \left| \hat{c}_{kk}^k \right| \hat{S}_k \ .$$

The $v_k, \hat{v}_k \in \mathbb{R}^n$ are defined as

$$(5.9)\quad v_k^T = (0, \cdots, 0, c_{kk}^k + \operatorname{sgn}(c_{kk}^k) S_k, c_{k,k+1}^k, \cdots, c_{k,n}^k),$$

$$(5.10)\quad \hat{v}_k^T = (0, \cdots, 0, \hat{c}_{kk}^k + \operatorname{sgn}(\hat{c}_{kk}^k) \hat{S}_k, \hat{c}_{k,k+1}^k, \cdots, \hat{c}_{k,n}^k),$$

where the first $k-1$ entries are zero. Finally, we have from (5.9) (5.10), (5.7) and (5.8) the relations

$$(5.11)\quad ||v_k||^2 = 2(S_k^2 + \left| c_{kk}^k \right| S_k) = 2 H_k$$

$$(5.12)\quad ||\hat{v}_k||^2 = 2(\hat{S}_k^2 + \left| \hat{c}_{kk}^k \right| \hat{S}_k) = 2 \hat{H}_k \ .$$

We will now relate the perturbation δP_k to the perturbation at step k in $\delta C_k = \hat{C}_k - C_k$. From (5.6) we have

$$(5.13)\quad \begin{aligned} \delta P_k &= (I - \hat{v}_k \hat{v}_k^T / \hat{H}_k) - (I - v_k v_k^T / H_k) \\ &= (v_k v_k^T - \hat{v}_k \hat{v}_k^T) / H_k + \hat{v}_k \hat{v}_k^T (\hat{H}_k - H_k) / (\hat{H}_k \cdot H_k) \ . \end{aligned}$$

We define $\delta v_k \in \mathbb{R}^n$ by

$$(5.14)\quad \delta v_k = \hat{v}_k - v_k$$

$$= (0, \cdots, 0, \delta c^k_{kk} + \operatorname{sgn}(c^k_{kk})(\hat{S}_k - S_k), \delta c^k_{k,k+1}, \cdots, \delta c^k_{k,n})^T ,$$

where we have made use of our assumption that $\operatorname{sgn}(c^k_{kk}) = \operatorname{sgn}(\hat{c}^k_{kk})$. This assumption means that the initial perturbation of δC in C is not large enough to cause a sign change in c^k_{kk} at step k of the reduction. If the signs were different, then we can conclude that both

$$(5.15)\qquad |\hat{c}^k_{kk}|, \quad |c^k_{kk}| \leqslant |\delta c^k_{kk}| .$$

Later in our proof we will obtain an estimate for δC_k which will enable us to test if this can happen. If the signs do differ, however, we cannot guarantee that the matrix δP_k will have a small norm, which implies that this type of error analysis does not apply.

If we make use of (5.14), and substitute for v_k in (5.13), we obtain the following expression for δP_k

$$(5.16)\quad \delta P_k = -(v_k \cdot \delta v_k^T + \delta v_k \cdot v_k^T + \delta v_k \cdot \delta v_k^T)/H_k + \hat{v}_k \cdot \hat{v}_k^T (\hat{H}_k - H_k)/(\hat{H}_k \cdot H_k).$$

We will now introduce the following notation which will help in our making norm estimates. Define the column vectors C^k_i, $\hat{C}^k_i$, $\delta C^k_i \in \mathbb{R}^n$ to be the transpose of row i of the matrices C_k, $\hat{C}_k$ and δC_k respectively. Also, define the vectors $\tilde{C}^k_i$, $\hat{\tilde{C}}^k_i$, $\delta\tilde{C}^k_i \in R^n$ to be obtained by setting the first $k-1$ coordinates equal to zero of the vectors C^k_i, $\hat{C}^k_i$ and δC^k_i, respectively. If we apply the above

definitions to (5.7) we have

$$(5.17) \qquad ||\tilde{C}_k^k|| = S_k \quad \text{and} \quad ||\hat{\tilde{C}}_k^k|| = \hat{S}_k \ .$$

We can now use (5.17) and the above definitions to obtain a convenient expression for the norm of (5.14)

$$(5.18) \qquad ||\delta v_k||^2 = ||\delta\hat{C}_k^k||^2 + 2\delta c_{kk}^k \operatorname{sgn}(c_{kk}^k)(||\hat{\tilde{C}}_k^k|| - ||\tilde{C}_k^k||)$$

$$+ (||\hat{\tilde{C}}_k^k|| - ||\tilde{C}_k^k||)^2 \ .$$

By the triangle inequality we have

$$(5.19) \qquad \left| ||\hat{\tilde{C}}_k^k|| - ||\tilde{C}_k^k|| \right| \leq ||\hat{\tilde{C}}_k^k - \tilde{C}_k^k||$$

$$\leq ||\delta\tilde{C}_k^k|| \ .$$

Using (5.18) and (5.19) we find

$$(5.20) \qquad ||\delta v_k||^2 \leq 2||\delta\tilde{C}_k^k||^2 + 2|\delta c_{kk}^k| \cdot ||\delta\tilde{C}_k^k||$$

$$\leq 4||\delta\tilde{C}_k^k||^2 ,$$

where in the last inequality we have made use of the fact that $|\delta c_{kk}^k| \leq ||\delta C_k^k||$. This gives a bound for $||\delta v_k||$

$$(5.21) \qquad ||\delta v_k|| \leq 2||\delta\tilde{C}_{kk}^k|| \ .$$

We use (5.8), (5.17) and (5.19) to obtain a bound for $|\hat{H}_k - H_k|$

$$(5.22) \qquad |\hat{H}_k - H_k| \leq |\hat{S}_k^2 + |\hat{c}_{kk}^k|\hat{S}_k - (S_k^2 + |c_{kk}^k|S_k)|$$

$$\leq (\hat{S}_k + S_k) \cdot |\hat{S}_k - S_k| + |\hat{c}_{kk}^k| \cdot |\hat{S}_k - S_k| + |\ |\hat{c}_{kk}^k| - |c_{kk}^k|\ | \cdot S_k$$

$$\leqslant 2(||\hat{\tilde{C}}_k^k|| + ||\tilde{C}_k^k||) \cdot ||\delta\tilde{C}_k^k||$$

$$\leqslant 2(2S_k + ||\delta\tilde{C}_k^k||)\,||\delta\tilde{C}_k^k|| .$$

If we substitute (5.21) and (5.22) into (5.16) and use (3.5) and (3.11) we obtain the following after taking norms

$$||\delta P_k|| \leqslant ||\delta v_k||(2||v_k|| + ||\delta v_k||)/H_k \tag{5.23}$$

$$+ (||\hat{v}_k||^2/\hat{H}_k)\cdot |\hat{H}_k - H_k|/H_k$$

$$\leqslant 4||\delta\tilde{C}_k^k||\left[(||v_k|| + ||\delta\tilde{C}_k^k||)/H_k + (2S_k + ||\delta\tilde{C}_k^k||)/H_k\right],$$

where we have made use of (5.12) in the last inequality. From (5.8) we have

$$H_k \geqslant S_k^2 , \tag{5.24}$$

and using this with (5.11) leads to the estimate

$$(||v_k|| + ||\delta\tilde{C}_k^k||)/H_k \leqslant \sqrt{2}/\sqrt{H_k} + ||\delta\tilde{C}_k^k||/H_k \tag{5.25}$$

$$\leqslant (\sqrt{2} + ||\delta\tilde{C}_k^k||/S_k)/S_k .$$

If we combine (5.11), (5.23) and (5.25), we have the following bound for $||\delta P_k||$, using $||\delta\tilde{C}_k^k|| \leqslant ||\delta C_k^k||$

$$||\delta P_k|| \leqslant 4||\delta C_k^k||\cdot\left[(2+\sqrt{2}) + 2||\delta C_k^k||/S_k\right]/S_k . \tag{5.26}$$

A condition will be given later, as to when we have a bound for $||\delta C_k^k||/S_k$.

We now relate the perturbation in step k, which is δC_k, to the initial perturbation δC. We will do this by tracing the error backwards step by step. Toward this end we calculate the following estimates. From (5.4) we have for $1 \leqslant i \leqslant r$

$$\delta C_{i+1} = \delta C_i \hat{P}_i + C_i \delta P_i \quad . \tag{5.27}$$

For $1 \leqslant \ell \leqslant r$, row ℓ of the matrix δC_{i+1} is given by

$$(\delta C_\ell^{i+1})^T = (\delta C_\ell^i)^T \hat{P}_i + (C_\ell^i)^T \delta P_i \quad . \tag{5.28}$$

We want to replace the last term in (5.28) by something more convenient for our purposes. It will be shown that

$$(C_\ell^i)^T \delta P_i = (\tilde{C}_\ell^i)^T \delta P_i \quad . \tag{5.29}$$

To accomplish this, we recall (5.13), which gives

$$\delta P_i = v_i v_i^T / H_i - \hat{v}_i \hat{v}_i^T / \hat{H}_i \quad . \tag{5.30}$$

From (5.9) and (5.10), the first i-1 entries of v_i and $\hat{v}_i$ are zero, which, when we combine this with (5.30), implies that the first i-1 rows of δP_i are zero. This implies that irrespective of how the first i-1 entries of C_ℓ^i are changed, the resulting product will be the same, as long as the remaining entries of C_ℓ^i remain unchanged. Using the fact that $\tilde{C}_\ell^i$ agrees with C_ℓ^i in all but the first i-1 entries, we have

$$(\delta C_\ell^{i+1})^T = (\delta C_\ell^i)^T \hat{P}_i + (\tilde{C}_\ell^i)^T \delta P_i \quad . \tag{5.31}$$

Define the following non-negative numbers for $1 \leqslant i \leqslant r$ and $i \leqslant \ell \leqslant r$ by

$$\mu_{\ell i} = ||\tilde{C}_\ell^i||/||\tilde{C}_\ell^i|| = ||\tilde{C}_\ell^i||/S_i \ , \tag{5.32}$$

which, because of (1.20), are all finite.

Taking norms in (5.31) and using (5.32) and (3.9) we have

$$||\delta C_\ell^{i+1}|| \leqslant ||\delta C_\ell^i|| + \mu_{\ell i} S_i ||\delta P_i|| \ . \tag{5.33}$$

We will now prove by induction on i, for $1 \leqslant i \leqslant r$ and $1 \leqslant \ell \leqslant r$, that the following is true

$$||\delta C_\ell^{i+1}|| \leqslant \sum_{j=1}^{i} \mu_{\ell j} S_j ||\delta P_j|| + ||\delta C_\ell^1|| \ . \tag{5.34}$$

For $i = 1$, where $1 \leqslant \ell \leqslant r$, inequality (5.34) is just inequality (5.33). Assume that (5.34) is true for some i, such that $1 \leqslant i \leqslant r-1$, then we will prove that (5.34) is also true for i+1. Using (5.33) for i+1, and our induction hypothesis, we obtain for $1 \leqslant \ell \leqslant r$

$$\begin{aligned} ||\delta C_\ell^{i+2}|| &\leqslant ||\delta C_\ell^{i+1}|| + \mu_{\ell, i+1} S_{i+1} ||\delta P_{i+1}|| \\ &\leqslant \sum_{j=1}^{i} \mu_{\ell j} S_j ||\delta P_j|| + ||\delta C_\ell^1|| \\ &\qquad + \mu_{\ell, i+1} S_{i+1} ||\delta P_{i+1}|| \\ &\leqslant \sum_{j=1}^{i+1} \mu_{\ell j} S_j ||\delta P_j|| + ||\delta C_\ell^1|| \ . \end{aligned} \tag{5.35}$$

At this point, we will make use of our assumption of "maximal row pivoting". This assumption implies that all of the $\mu_{\ell j} \leqslant 1$ for $\ell \geqslant j$, because of the manner in which the next pivotal row is chosen (see Section 1.1). We have retained the $\mu_{\ell j}$ until this time, because we will refer to (5.34) later when we discuss pivotal strategies. Using this assumption, (5.34) now becomes for $\ell = i+1$

$$||\delta C_{i+1}^{i+1}|| \leqslant \sum_{j=1}^{i} S_j ||\delta P_j|| + ||\delta C_{i+1}^{1}|| . \tag{5.36}$$

At this point in our proof, we would like to show that (5.36) and our assumption (5.2a) leads to

$$||\delta C_k^k|| / S_k \leqslant \triangle, \text{ for } 1 \leqslant k \leqslant r, \tag{5.37}$$

and

$$||\delta C_k^k|| \leqslant (K+1)^{k-1} ||\delta C|| \quad \text{for } 1 \leqslant k \leqslant r. \tag{5.38}$$

Where (5.37) would yield the following simplification of (5.26)

$$||\delta P_k|| \leqslant K ||\delta C_k^k|| / S_k , \tag{5.39}$$

where $K = 4(2 + \sqrt{2} + 2\triangle)$.

Toward this end we will prove the following for $1 \leqslant k \leqslant r$

$$||C|| / S_k \leqslant \chi(C) . \tag{5.40}$$

This is true because of our assumptions of "maximal row pivoting", and that C is of full rank, and because we have from (1.19) and (1.20) that

$$(5.41) \qquad ||C||/S_k \leqslant ||C||/S_r ,$$

where $S_r > 0$. Because S_r is the absolute value of element r on the diagonal of L, which is a lower triangular matrix, we have $|\lambda_r(L)| = S_r$, where $\lambda_r(L)$ is an eigenvalue of L. Therefore, there is a unit vector $x \in \mathbb{R}^r$, such that $Lx = \lambda_r(L)x$. From (1.21) L^{-1} exists, so we have $L^{-1}x = (1/\lambda_r(L))x$. Taking norms, we find

$$(5.42) \qquad \begin{aligned} 1/S_r &= ||(1/\lambda_r(L))x|| \\ &= ||L^{-1}x|| \\ &\leqslant ||L^{-1}|| . \end{aligned}$$

If we use (3.1) we have $||L^{-1}||^2 = \rho(L^{-T}L^{-1}) = \rho((L\,L^T)^{-1})$. Using Wilkinson [34, p. 290] property (F), because $(L\,L^T)^{-1}$ is a symmetric matrix we have $\rho((L\,L^T)^{-1}) = ||(L\,L^T)^{-1}||$. Using (3.3), we have $||(L\,L^T)^{-1}|| = 1/\mu(L\,L^T)$, and from (5.5) and (3.2b) we have $\mu(L\,L^T) = \mu(C\,C^T) = \mu(C^T\,C)$. This proves that (5.40) is true, using Definition 4.1 for the condition number $\chi(C)$.

We shall prove (5.37), (5.38) and (5.39) simultaneously by induction on k, for $1 \leqslant k \leqslant r$, by using (5.2a) and (5.36). For $k=1$, using (5.40), we have that (5.37) is valid and hence also (5.39) because

$$(5.43) \qquad \begin{aligned} ||\delta C_1^1||/S_1 &\leqslant (||\delta C||/||C||)(||C||/S_1) \\ &\leqslant [\triangle/(\chi(C)(1+K)^{r-1})]\ \chi(C) \\ &\leqslant \triangle/(1+K)^{r-1} \\ &\leqslant \triangle . \end{aligned}$$

Also, (5.38) is trivial for $k=1$. Assume that (5.37), (5.38) and (5.39) are valid for all ℓ such that $1 \leqslant \ell \leqslant k$ where $1 \leqslant k < r$, then we will prove that they are also valid for $k+1$. Using (5.36) and our induction hypothesis, (5.38) is valid because

$$(5.44) \qquad ||\delta C_{k+1}^{k+1}|| \leqslant \sum_{j=1}^{k} S_j \, ||\delta P_j|| + ||\delta C_{k+1}^{1}||$$

$$\leqslant \sum_{j=1}^{k} K \, ||\delta C_j^j|| + ||\delta C_{k+1}^{1}||$$

$$\leqslant \sum_{j=1}^{k} K \, (K+1)^{j-1} \, ||\delta C|| + ||\delta C||$$

$$\leqslant (K+1)^k \, ||\delta C|| \ ,$$

where we have used (5.1). To see that (5.37) is valid for $k+1$

$$(5.45) \qquad ||\delta C_{k+1}^{k+1}||/S_{k+1} \leqslant \left[(K+1)^k \, ||\delta C||/||C||\right] \cdot \left[||C||/S_{k+1}\right]$$

$$\leqslant \triangle (K+1)^k / (K+1)^{r-1}$$

$$\leqslant \triangle \ ,$$

where we have used (5.44), (5.2a) and (5.40). Finally (5.39) follows from (5.37) and (5.26), and our induction is complete.

What we have accomplished is to express the perturbation in row k at step k of our factorization in terms of the initial perturbation in C. We have also shown in (5.39) that the perturbation in the elementary orthogonal transformation at step k is bounded in terms of the perturbation in row k of C_k. Therefore, if we combine

(5.38), (5.39) and (5.40), we obtain the following estimate for $||\delta P_k||$

$$(5.46) \qquad ||\delta P_k|| \leq K(1+K)^{k-1} ||\delta C||/S_k$$

$$\leq K(1+K)^{k-1} \chi(C) ||\delta C||/||C|| .$$

We are now in a position to obtain a bound for $||\delta Q||$. Using (5.3), (5.46) and (5.1), it is clear that

$$(5.47) \qquad ||\delta Q|| \leq \left[\sum_{k=1}^{r} K(1+K)^{k-1} \right] \chi(C) ||\delta C||/||C||$$

$$\leq K(r) \chi(C) ||\delta C||/||C|| ,$$

where

$$(5.48) \qquad K(r) = (1+K)^r - 1 .$$

■

Remark: In reference to the assumption that $\operatorname{sgn}(c^k_{kk}) = \operatorname{sgn}(\hat{c}^k_{kk})$: if it is not the case that we have equality for some k, then recalling (5.15)

$$(5.15) \qquad |\hat{c}^k_{kk}|, \ |c^k_{kk}| \leq |\delta c^k_{kk}| ,$$

let k be the initial natural number such that we do not have equality, then (5.15) is valid, and we have from (5.38)

$$(5.49) \qquad |\hat{c}^k_{kk}|, \ |c^k_{kk}| \leq (1 + K)^{k-1} ||\delta C|| .$$

This is a necessary condition for the signs to be different. The typical situation is as follows. We have a numerical bound for $||\delta C||$, and we are performing the factorization on the perturbed

matrix $\hat{C}$. We check at each step, k, of the factorization to see if (5.49) cannot hold for any k, then we can conclude that the factorization of the matrix C would proceed in the same way, and $\operatorname{sgn}(\hat{c}^{k}_{kk}) = \operatorname{sgn}(c^{k}_{kk})$ for $1 \leqslant k \leqslant r$.

Remark: On pivoting. The usual pivotal strategy used is "maximal row pivoting", see Golub [11]. Other strategies of various types have been tried with various degrees of computational success, see Jennings and Osborn [25]. Usually, no theoretical justification is given as to how the computational errors are affected by the choice of pivotal strategy.

We are able to give here a justification of sorts as to the desirability of "maximal row pivoting". Assume (5.39) holds, then we have from (5.34)

$$||\delta C^{i+1}_{i+1}|| \leqslant \sum_{j=1}^{i} (K\mu_{i+1,j})\,||\delta C^{j}_{j}|| + ||\delta C^{1}_{i+j}||\,. \tag{5.50}$$

Intuitively, it is seen by examining (5.50) that, by repeated back substitution, we obtain an expression which is the sum of products, each of which is made up of repeated factors of the type $(K\mu_{st})$. Because "maximal row pivoting" gives $\mu_{st} \leqslant 1$, it helps reduce the effect of the power of k in the error analysing giving, in general, a better bound then a strategy which allows the μ_{st} to be larger than unity.

Section 6. Basic Relations in the Perturbational Problem

Using the notation of Sections 1 through 5, we will establish the following relations, which will be used repeatedly.

Lemma 6.1. (6.1) $C\,C^T = L\,L^T$, $\hat{C}\,\hat{C}^T = \hat{L}\,\hat{L}^T$, $B_2^T B_2 = W^T W$ and $\hat{B}_2^T \hat{B}_2 = \hat{W}^T \hat{W}$.

(6.2) $L^{-1} C = Q_1$ and $B_2 W^{-1} = V_1$.

(6.3) $\chi(L) = \chi(C)$.

(6.4) $\chi(W) = \chi(B_2)$.

If $||L^{-1}\,\delta L|| \leqslant \Delta < 1$, then

(6.5a) $||\hat{L}|| \leqslant e_1 ||L||$, where $e_1 = (1 + \Delta)$,

(6.5b) $||\hat{L}^{-1}|| \leqslant e_2 ||L^{-1}||$, where $e_2 = 1/(1 - \Delta)$,

(6.5c) $\chi(\hat{L}) \leqslant e_3\, \chi(C)$, where $e_3 = (1 + \Delta)/(1 - \Delta)$.

If $||\delta W W^{-1}|| \leqslant \Delta < 1$, then

(6.6a) $||\hat{W}|| \leqslant e_1 ||W||$, where $e_1 = (1 + \Delta)$,

(6.6b) $||\hat{W}^{-1}|| \leqslant e_2 ||W^{-1}||$, where $e_2 = 1/(1 - \Delta)$,

(6.6c) $\chi(\hat{W}) \leqslant e_3\, \chi(B_2)$, where $e_3 = (1 + \Delta)/(1 - \Delta)$.

If $\sqrt{r/2}\ ||L^{-1}\delta L|| \leqslant \Delta < 1$,

then

(6.7a) $||L^{-1}\delta L|| \leqslant K_1(r)\, \chi(C)\, ||\delta C||/||C||$,

(6.7b) $||\delta Q_1|| \leqslant K_2(r)\, \chi(C)\, ||\delta C||/||C||$,

(6.7c) $\quad ||\delta L|| \leq K_3(r)\, \chi(C)\, ||\delta C||$,

where $K_1(r) = \sqrt{r}\,(2+\Delta)/(\sqrt{2}(1-\Delta))$,

$K_2(r) = 1 + K_1(r)$, and $K_3(r) = K_2(r) + 1$.

If $\sqrt{(n-r)/2}\ ||\delta W\, W^{-1}|| \leq \Delta < 1$,

then

(6.8a) $\quad ||\delta W\, W^{-1}|| \leq K_1(n-r)\, \chi(B_2)\, ||\delta B_2||/||B_2||$,

(6.8b) $\quad ||\delta V_1|| \leq K_2(n-r)\, \chi(B_2)\, ||\delta B_2||/||B_2||$,

(6.8c) $\quad ||\delta W|| \leq K_3(n-r)\, \chi(B_2)\, ||\delta B_2||$,

where $K_1(n-r)$, $K_2(n-r)$ and $K_3(n-r)$ are defined in (6.7).

Under the hypothesis of Lemma 5.1 we have

(6.9a) $\quad ||\delta B_1||,\ ||\delta B_2||,\ ||[\delta B_1 | \delta B_2]|| \leq (||\delta A||/||A||$

$$+ K(r)\, \chi(C)\, ||\delta C||/||C||)||A|| ,$$

(6.9b) $\quad ||\delta B_1|| \leq (||\delta A||/||A|| + ||\delta Q_1)\, ||A||$.

If $x = Q \cdot \begin{bmatrix} d_1 \\ d_2 \end{bmatrix}$, where $x \in \mathbb{R}^n$, $Q \in \mathbb{R}^{n \times n}$ is orthogonal, $d_1 \in \mathbb{R}^{n_1}$, $d_2 \in \mathbb{R}^{n_2}$ and $n_1 + n_2 = n$, then

(6.10) $\quad ||d_1||,\ ||d_2|| \leq ||x||$.

Proof (6.1): From (1.15a) we have $C C^T = [L|0]\, Q^t Q \begin{bmatrix} L^T \\ 0 \end{bmatrix} = L\, L^T$,

and from (1.32a) we have

$$B_2^T B_2 = [W^T|0]\, V\, V^T \begin{bmatrix} W \\ 0 \end{bmatrix} = W^T W .$$

(6.2): From (1.15a) we have $L^{-1} C = L^{-1} [L|0] Q^T = Q_1^T$, where the proof for W is similar.

(6.3): If we use (6.1), (3.2a) and (3.2b), then (6.3) follows from Definition 4.1.

(6.4): Same as (6.3).

(6.5a): Because $\hat{L} = L (I + L^{-1} \delta L)$ is valid, the result follows by taking norms.

(6.5b): Because $\hat{L}^{-1} = (I + L^{-1} \delta L)^{-1} L^{-1}$ is valid, the result follows by taking norms and using (3.16).

(6.5c): This follows from (6.5a), (6.5b) and (6.3).

(6.6a)-(6.6c): These are proved in a similar way to (6.5a) through (6.5c).

(6.7a): This is a result of Jennings and Osborne [25, p. 327] inequality (2.6) of their paper.

(6.7b): From (1.15a) we have $\delta C = \hat{L} \hat{Q}_1 - L Q_1^T = \delta L \hat{Q}_1^T + L \delta Q_1^T$ which gives $L \delta Q_1^T = \delta C - \delta L \hat{Q}_1^T$. Premultiplying by L^{-1} we have $\delta Q_1^T = L^{-1} \delta C - L^{-1} \delta L \hat{Q}_1^T$. Taking norms and using (3.5), (3.11), (3.13) and (3.7) we obtain $||\delta Q_1|| \leqslant ||L^{-1}|| \; ||\delta C|| + ||L^{-1} \delta L||$. After using (6.1), (6.3) and (6.7a) our result follows.

(6.7c): Using (1.15b) and taking norms we have $||\delta L|| \leqslant ||\delta C|| + ||C|| \; ||\delta Q_1||$, where we have used (3.11), (3.13) and (3.7). If we use (6.7b) and the fact that $\chi(C) \geqslant 1$ we have our result.

(6.8a)-(6.8c): These are proved in almost an identical way as (6.7a) through (6.7c).

(6.9a): From (1.27) we have $\left[\delta B_1 \mid \delta B_2\right] = \delta A\, \hat{Q} + A\, \delta Q$.

If we apply (3.13), (3.9) and Lemma 5.1 we have the result.

(6.9b): $\delta B_1 = \delta A\, \hat{Q}_1 + A\, \delta Q_1$.

(6.10): Use (3.8) and the definition of the Euclidean vector norm on $\begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = Q^T x$.

Section 7. Effects of Perturbational Errors

In this section we will show the effect of perturbational errors on a solution given by the method of Algorithm 1.1. In order to obtain an expression which is compact, it will be necessary to merge all terms of higher powers into those of the first power. These terms represent relative errors, which, if the perturbations and condition numbers are reasonable, should be very much smaller than unity. Therefore, it is not unreasonable to make the assumption that they are bounded away from unity. This is the reason for making assumptions (7.1a) through (7.1e). It should be clear, that even terms (7.1a) and (7.1b) represent relative errors, because $||L^{-1}\delta L|| \leqslant \chi(L)\,||\delta L||/||L||$. Also, (7.1c) should be reasonable, because in the unconstrained case it is always true that $||R||/||f|| \leqslant 1$, and we would hope that in a practical problem, f would be very near the space spanned by the columns of A, which would yield a very small relative error for the residual.

Theorem 7.1. The notation of Sections 1 through 6 is used. If there exists a real non-negative $\Delta < 1$, such that

(7.1a) $h_1||L^{-1}\delta L|| \leqslant \Delta$, where $h_1 = \max\{1, \sqrt{r/2}\}$,

(7.1b) $h_2||\delta W W^{-1}|| \leqslant \Delta$, where $h_2 = \max\{1, \sqrt{(n-r)/2}\}$,

(7.1c) $||R||/||f|| \leqslant \Delta$, $f \neq 0$

(7.1d) $||\delta Q_1|| \leqslant K_2(r)\,\chi(C)\,||\delta C||/||C|| \leqslant \Delta$,

(7.1e) $||\delta A||/||A|| \leqslant \Delta$,

where C, $\hat{C}$, B_2, $\hat{B}_2$ are of full rank and the hypothesis of Lemma 5.1 is satisfied then

$$
\begin{aligned}
(7.2)\qquad \|\hat{x}-x\|/\|x\| \leq\ & (e_2\|\delta g\|/\|g\| + K_5(r)\|\delta C\|/\|C\|)\chi(C) \\
& + (e_2^2\|\delta f\|/\|f\| + 2e_2\|\delta A\|/\|A\|) \\
& \cdot(\|A\|/\|B_2\|)\chi(B_2) + (e_2^2 e_4\|\delta g\|/\|g\| \\
& + e_2 K_4(r)\|\delta C\|/\|C\|)(\|A\|/\|B_2\|)\chi(C)\chi(B_2) \\
& + e_2^2 K_2(n-r)(\|A\|/\|B_2\|)^2(\|\delta A\|/\|A\| \\
& + K(r)\chi(C)\|\delta C\|/\|C\|)\chi^2(B_2)\|R\|/\|f\|,
\end{aligned}
$$

where the constants e_2, e_4, $K(r)$, $K_2(n-r)$, $K_4(r)$ and $K_5(r)$ are given by (6.5b), (7.21), (5.48), (6.7), (7.23) and (7.26) respectively.

Proof: We have from Algorithm 1.1 that solutions x and $\hat{x}$ exist to both problems, where x and $\hat{x}$ are given by

$$
(7.3a)\qquad x = Q\begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = Q\begin{bmatrix} L^{-1} g \\ W^{-1} V_1 \{f - B_1 L^{-1} g\} \end{bmatrix}
$$

$$
(7.3b)\qquad \hat{x} = \hat{Q}\begin{bmatrix} \hat{d}_1 \\ \hat{d}_2 \end{bmatrix} = \hat{Q}\begin{bmatrix} \hat{L}^{-1} \hat{g} \\ \hat{W}^{-1} \hat{V}_1 \{\hat{f} - \hat{B}_1 \hat{L}^{-1} \hat{g}\} \end{bmatrix}
$$

respectively. Subtracting x from $\hat{x}$ we obtain

$$(7.4) \qquad \hat{x} - x = \hat{Q} \begin{bmatrix} \delta d_1 \\ \delta d_2 \end{bmatrix} + \delta Q \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} ,$$

where $\delta d_1 = \hat{d}_1 - d_1$ and $\delta d_2 = \hat{d}_2 - d_2$. Using (6.10), (7.3a) and (7.4), we find after taking norms

$$(7.5) \qquad ||\hat{x}-x|| \leqslant ||\delta d_1|| + ||\delta d_2|| + ||\delta Q||\ ||x|| .$$

Lemma 5.1 gives an upper bound for $||\delta Q||$

$$(7.6) \qquad ||\delta Q|| < K(r)\ \chi(C)\ ||\delta C||/||C|| .$$

Therefore, all we have to find are upper bounds for $||\delta d_1||$ and $||\delta d_2||$. Using (7.3a) the expression for δd_1 is

$$(7.7) \qquad \begin{aligned} \delta d_1 &= \hat{L}^{-1}\hat{g} - L^{-1}g \\ &= \hat{L}^{-1}(\delta g - \delta L d_1) , \end{aligned}$$

where we have made use of (3.17) and the definition of d_1 to obtain our simplification. Because $\hat{L}^{-1} = (I + L^{-1}\delta L)^{-1} L^{-1}$, we obtain the following upon taking norms

$$(7.8) \qquad ||\delta d_1|| \leqslant e_2 \left[\chi(C)||\delta g||/(||C||\ ||x||) + ||L^{-1}\delta L||\right]\ ||x|| ,$$

where we have made use of (3.16), (6.3) and (6.10). From (1.4) we have that $||C||\ ||x|| \geqslant ||g||$, and substituting this and (6.7a) into (7.8) we have

$$(7.9) \qquad ||\delta d_1|| \leqslant e_2 \left[||\delta g||/||g|| + K_1(r)||\delta C||/||C||\right] \chi(C)||x|| .$$

Before estimating $||\delta d_2||$, define the vector $\mathcal{E} \in \mathbb{R}^m$ by

$$\mathcal{E} = \{\hat{f} - \hat{B}_1 \hat{d}_1\} - \{f - B_1 d_1\} \tag{7.10}$$
$$= \delta f - \hat{B}_1 \delta d_1 - \delta B_1 d_1 .$$

If we make use of the equalities (1.33), (3.17) and $\hat{V}_1 = V_1 + \delta V_1$, we obtain the following simplified expression for δd_2

$$\delta d_2 = \hat{W}^{-1} \hat{V}_1 \{\hat{f} - \hat{B}_1 \hat{d}_1\} - W^{-1} V_1 \{f - B_1 d_1\} \tag{7.11}$$
$$= \left[\hat{W}^{-1} \hat{V}_1 - W^{-1} V_1\right]\{f - B_1 d_1\} + \hat{W}^{-1} \hat{V}_1 \mathcal{E}$$
$$= \hat{W}^{-1}\left[\delta V_1 \{f - B_1 d_1\} - \delta W d_2 + \hat{V}_1 \mathcal{E}\right].$$

In the last expression we have made a simplification by using the definition of d_2 given in (1.33). From (1.28) we have $B_2 d_2 - R = f - B_1 d_1$. Substituting this into (7.11) we find

$$\delta d_2 = \hat{W}^{-1}\left[\delta V_1 B_2 d_2 - \delta V_1 R - \delta W d_2 + \hat{V}_1 \mathcal{E}\right] \tag{7.12}$$
$$= \hat{W}^{-1}\left[(\delta V_1 B_2 - \delta W) d_2 - \delta V_1 R + \hat{V}_1 \mathcal{E}\right].$$

Using (1.32b) we see that

$$\delta W = \hat{V}_1 \hat{B}_2 - V_1 B_2 \tag{7.13}$$
$$= \hat{V}_1 \delta B_2 + \delta V_1 B_2 .$$

If (7.13) is substituted into (7.12), we find after the cancellation of $\delta V_1 B_2$

$$\delta d_2 = \hat{W}^{-1}\left[-\hat{V}_1 \delta B_2 d_2 - \delta V_1 R + \hat{V}_1 \mathcal{E}\right]. \tag{7.14}$$

If substitution of (7.10) into (7.14) is made, we have after regrouping

$$(7.15)\qquad \delta d_2 = \hat{W}^{-1}\hat{V}_1\left[\delta f - \hat{B}_1\,\delta d_1 - \delta B_1\, d_1 - \delta B_2\, d_2\right] - \hat{W}^{-1}\delta V_1 R.$$

Taking norms in (7.15), and using (6.6b) and (6.10) we have

$$\begin{aligned}(7.16)\qquad ||\delta d_2|| \leqslant e_2\Big[&||\delta f||/(||A||\,||x||) + (||B_1||/||A|| \\ &+ ||\delta B_1||/||A||)\,||\delta d_1||/||x|| + ||\delta B_1||/||A|| \\ &+ ||\delta B_2||/||A||\Big]\,||A||\,||W^{-1}||\,||x|| \\ &+ e_2\,||W^{-1}||\,||\delta V_1||\,||R||\,.\end{aligned}$$

From (1.27), by using (3.9) and (3.13), we have that

$$(7.17)\qquad ||B_1||,\ ||B_2|| \leqslant ||A||\,.$$

From (1.3) and our hypothesis (7.1c) it follows

$$\begin{aligned}(7.18)\qquad ||A||\,||x|| &\geqslant ||f+R|| \\ &\geqslant ||f|| - ||R|| \\ &\geqslant ||f||/e_2 > 0\,,\end{aligned}$$

where $e_2 = 1/(1-\triangle)$.

Using inequalities (7.17) and (7.18) to simplify (7.16) we obtain

$$\begin{aligned}(7.19)\qquad ||\delta d_2|| \leqslant e_2\Big[&e_2||\delta f||/||f|| + (1 + ||\delta B_1||/||A||) \\ &\cdot ||\delta d_1||/||x|| + ||\delta B_1||/||A|| \\ &+ ||\delta B_2||/||A||\Big](||A||/||B_2||)\chi(B_2)||x|| \\ &+ e_2\,||W^{-1}||\,||\delta V_1||\,||R||\,,\end{aligned}$$

where (6.4) has been used.

From (6.9b), and our hypotheses (7.1d) and (7.1e) it follows

$$(7.20) \qquad ||\delta B_1|| \leqslant 2\triangle ||A|| ,$$

thereby simplifying the coefficient of $||\delta d_1||/||x||$ to be

$$(7.21) \qquad e_4 = 1 + 2\triangle .$$

Using (6.9a) as an upper bound for $||\delta B_1||$ and $||\delta B_2||$, and substituting (7.9) for $||\delta d_1||$, we obtain after regrouping

$$\begin{aligned}(7.22)\ ||\delta d_2|| \leqslant e_2 \big[& e_2 ||\delta f||/||f|| + 2||\delta A||/||A|| \\ & + (e_2 e_4 ||\delta g||/||g|| + K_4(r) ||\delta C||/||C||)\chi(C)\big] \\ & \cdot (||A||/||B_2||\chi(B_2)||x|| + e_2 ||W^{-1}|| \, ||\delta V_1|| \, ||R||,\end{aligned}$$

where

$$(7.23) \qquad K_4(r) = e_2 e_4 K_1(r) + 2K(r) ,$$

and $$K_1(r) = \sqrt{r}\,(2 + \triangle)/(\sqrt{2}(1 - \triangle)) .$$

We will now obtain an upper bound for the coefficient of $||R||$, by applying the estimates given in (6.4), (6.8b), (6.9a) and (7.18)

$$\begin{aligned}(7.24) \qquad & e_2 ||W^{-1}|| \, ||\delta V_1|| \, ||R|| \\ & \leqslant e_2 K_2(n-r)\chi^2(B_2)||R|| \, ||\delta B_2||/||B_2||^2 \\ & \leqslant e_2^2 K_2(n-r)\big[||\delta A||/||A|| + K(r)\chi(C)||\delta C||/||C||\big] \\ & \cdot \chi^2(B_2)(||A||/||B_2||)^2(||R||/||f||) \, ||x|| .\end{aligned}$$

If (7.6), (7.9), (7.22) and (7.24) are substituted into (7.5) and

regrouped according to condition numbers we obtain

$$
\begin{aligned}
(7.25)\qquad \|\hat{x}-x\|/\|x\| \leq\ & (e_2\|\delta g\|/\|g\| \\
& + K_5(r)\|\delta C\|/\|C\|)\,\chi(C) + (e_2^2\|\delta f\|/\|f\| \\
& + 2e_2\|\delta A\|/\|A\|)(\|A\|/\|B_2\|)\,\chi(B_2) \\
& + (e_2^2 e_4\,\|\delta g\|/\|g\| \\
& + e_2 K_4(r)\|\delta C\|/\|C\|)(\|A\|/\|B_2\|)\,\chi(C)\,\chi(B_2) \\
& + e_2^2 K_2(n-r)(\|A\|/\|B_2\|)^2(\|\delta A\|/\|A\| \\
& + K(r)\chi(C)\|\delta C\|/\|C\|)\chi^2(B_2)\,\|R\|/\|f\|\,,
\end{aligned}
$$

where

$$
(7.26)\qquad K_5(r) = e_2 K_1(r) + K(r)\,.
$$

We have shown that the condition numbers appear only linearly in (7.25), except for the coefficient of $\|R\|/\|f\|$. The question could be asked, if whether the term $\chi^2(B_2)$ is reasonable, or if it might be the fault of the error analysis. Van der Sluis [31] has shown, using a geometrical argument, that squaring of the condition number in the unconstrained case can indeed be realized. If $\chi(B_2)\,\|R\|/\|f\|$ is not too large, however, the effect of $\chi^2(B_2)$ can be minimized, and we would expect the condition numbers to appear only linearly in the error.

Also, the full effect of a condition number is seldom or possibly never realized, see Wilkinson [34].

Finally, because A and B_2 are calculated during our solution of (1.1) to (1.4), we have the bound

$$||A||/||B_2|| \leqslant \sqrt{n-r}\ ||A||_E/||B_2||_E\ ,$$

which is easy to calculate.

Section 8. Rounding Errors in Computation

We will state here for later reference some known results on computational errors as developed by Wilkinson [36], [35]. For our purposes, we will limit ourselves strictly to floating point computation.

To develop our results, we introduce the very general notation $fl(\cdot op \cdot)$, which is used by Wilkinson [36] and others. The "fl" notation means, that if we have two quantities A and B, which could be numbers, vectors, matrices, etc., and a corresponding operator "op" such as scalar addition, inner product, matrix multiplication, etc., then fl (A op B) is that quantity which would result in the computer by performing that operation by some computer program using floating point arithmetic.

A floating point binary number x consists of two parts, the binary exponent b and the binary mantissa a, where $-\frac{1}{2} \geqslant a > -1$ or $\frac{1}{2} \leqslant a < 1$. Then x is expressed as $x = a \cdot 2^b$. The computer memory allocated for x is limited to a certain number of digits, which are divided up in some way between a and b. We will denote by t the number of binary digits allocated to the mantissa. Also, we will assume that the number of digits allotted to the exponent is large enough to accommodate all of our calculations, and will not concern ourselves here with the problem of exponent under- or over-flow. Finally, we will assume that the computer with which we are working

has a double-precision accumulator. This means when single precision floating point numbers are retrieved from the memory to perform certain arithmetic operations, they are allocated a double precision mantissa before the operation is performed. The operation is carried out in double-precision, and the result is not rounded to single-precision until it is sent back to the memory.

Under these assumptions Wilkinson [36] has shown the following results.

Section 8.1. Vector Addition

Given two real numbers a and b, there exists a real ε, such that

$$(8.1) \qquad fl\,(a + b) = (a + b)(1 + \varepsilon)\,,$$

where

$$(8.2) \qquad |\varepsilon| \leqslant 2^{-t}\,.$$

Therefore, if we have two vectors $x, y \in R^n$,

$$(8.3) \qquad ||fl(x + y) - (x + y)|| \leqslant 2^{-t}\,||x+y||\,.$$

Section 8.2. Matrix Multiplication

Given a matrix $A \in \mathbb{R}^{mxn}$ and a vector $b \in \mathbb{R}^n$, let E denote the computational error in calculating $A \cdot b$

$$(8.4) \qquad E = fl\,(A \cdot b) - A \cdot b\,,$$

where the entries of E are represented by e_i for $1 \leqslant i \leqslant m$.

Wilkinson [36], p. 83 has shown that under the assumption $n\,2^{-t} < .1$

$$(8.5)\qquad |e_i| \leq 1.06 \cdot 2^{-t}\Big[n|a_{i1}||b_1| + \sum_{k=2}^{n} (n-(k-2))|a_{ik}||b_k|\Big]$$

$$\leq 1.06 \cdot 2^{-t}\, n \sum_{k=1}^{n} |a_{ik}||b_k|$$

$$\leq 1.06 \cdot 2^{-t}\, n \Big(\sum_{k=1}^{n} a_{ik}^2\Big)^{\frac{1}{2}} \Big(\sum_{k=1}^{n} b_k^2\Big)^{\frac{1}{2}} ,$$

where the last inequality is obtained by using Schwarz's inequality. Therefore, it follows from the definition of $||\cdot||_E$ and (3.4) that

$$(8.6)\qquad ||E|| \leq 1.06 \cdot 2^{-t}\, n\, ||A||_E\, ||b||$$

$$\leq 2^{-t}\, \overline{K_1}(n)\, ||A||\, ||b|| ,$$

where

$$(8.7)\qquad \overline{K_1}(n) = 1.06\, n^{3/2}$$

Section 8.3. Solution of a Triangular Set of Equations

Let $L \in \mathbb{R}^{n\times n}$ be an upper- or lower-triangular non-singular matrix. We now wish to know the computational error introduced by solving the linear equation

$$(8.8)\qquad Lx = b, \text{ where } x, b \in \mathbb{R}^n ,$$

for the vector x. Without loss of generality we will assume here that L is lower triangular. Wilkinson [36] has shown that the entries x_i of x are calculated sequentially from i=1 to i=n

by using

$$(8.9) \qquad x_r = fl((-l_{r1}\,x_1 - l_{r2}\,x_2 - \cdots - l_{r,r-1}\,x_{r-1} + b_r)/l_{rr}),$$

for $1 \leqslant r \leqslant n$. Note that the vector x which we have constructed in (8.9) is in general not the solution of (8.8), but rather, it is an exact solution of the perturbed problem

$$(8.10) \qquad (L + \delta L)\,x = b,$$

where the lower triangular matrix δL is bounded by [36, p. 103]

$$(8.11) \qquad \begin{aligned} ||\delta L|| &\leqslant ||\delta L||_E \\ &\leqslant 1.06\, 2^{-t}(1 + 1.06 \cdot 2^{-t} \cdot 3(n+2)/2)\, ||L||_E \\ &\leqslant 2^{-t}\, \overline{K_2}(n)\, ||L||, \end{aligned}$$

with

$$(8.12) \qquad \overline{K_2}(n) = 1.06\, \sqrt{n}\,(1 + 1.06\, 2^{-t} 3(n+2)/2),$$

and the value of n restricted sufficiently to make the term $1.06\, 2^{-t} 3(n+2)/2 \ll 1$. Also, if L^{-1} exists, then Wilkinson [36] has shown that $(L + \delta L)^{-1}$ also exists.

Section 8.4. Orthogonal Transformation.

An elementary orthogonal transformation $P \in \mathbb{R}^{n \times n}$ has the form $P = I - 2ww^T$, where $w \in \mathbb{R}^n$ and $||w|| = 1$. Given the vector w, we will denote the matrix "P", calculated using floating point arithmetic, by the symbol $\overleftarrow{P}$, (which may no longer be orthogonal), i.e.

(8.13) $\quad \overline{P} = fl(P)$,

where it is assumed that $n2^{-t} < .1$. Wilkinson [35] has shown that if $A \in \mathbb{R}^{nxm}$ is any matrix, premultiplication by P using floating point arithmetic gives

(8.14) $\quad ||fl(\overline{P}A) - PA|| \leqslant 2^{-t}\beta \sqrt{m}\ ||A||$,

where

(8.15) $\quad \beta = 12.36$.

Also if A is premultiplied sequentially by orthogonal transformations $P_1, P_2, \cdots, P_m$, and for $1 \leqslant i \leqslant m$ we define

(8.16a) $\quad \overline{P_i} = fl\,(P_i)$

(8.16b) $\quad \overline{A}_{i+1} = fl\,(\overline{P}_i \cdot \overline{A}_i)$, where $\overline{A}_1 = A$.

Then there exists a $\delta A \in \mathbb{R}^{nxm}$ such that

(8.17) $\quad \overline{A}_{m+1} = P_m \cdot P_{m-1} \cdot \cdots \cdot P_1\,(A + \delta A)$,

with

(8.18) $\quad ||\delta A|| \leqslant ||\delta A||_E$

$$\leqslant m2^{-t}\beta\,(1 + 2^{-t}\beta)^{m-1}\ ||A||_E$$

$$\leqslant 2^{-t}\overline{K}_3(m)\ ||A|| ,$$

where β is given in (8.15) and $\overline{K}_3(m)$ is defined by

(8.19) $\quad \overline{K}_3(m) = \beta\, m^{3/2}\,(1 + 2^{-t}\beta)^{m-1}$.

The results of (8.14) through (8.19) hold also for post multiplication by orthogonal transformations. This is true because

$$B\,P = (P\,B^T)^T, \tag{8.20}$$

where $B \in \mathbb{R}^{m\times n}$ and $P^T = P$, and our norms are invariant under transposition.

Section 9. Rounding Error Analysis

We now wish to analyze the effect of computational errors on our least squares solution with constraints. We will show that most of the error can be accounted for by perturbational induced errors in the initial problem.

Algorithm 1.1 gives the exact solution (1.1) to (1.4) to be

$$x = Q \begin{bmatrix} L^{-1} g \\ W^{-1} V_1 \{f - B_1 L^{-1} g\} \end{bmatrix} . \tag{9.1}$$

We will now define stepwise the order of computation defined in (9.1). This will determine the effect of the errors.

Step 1: Reduce C by orthogonal transformations to obtain L.

Define for $1 \leq i \leq r$

$$\overline{C}_{i+1} = fl\,(\overline{C}_i \overline{P}_i), \text{ where } \overline{C}_1 = C \text{ and } [\overline{L}|0] = \overline{C}_{r+1}, \tag{9.2}$$

and

$$\overline{P}_i = fl\,(\hat{P}_i), \tag{9.3}$$

where $\overline{P}_i \in \mathbb{R}^{nxn}$ is that elementary orthogonal matrix which exactly reduces row i of $\overline{C}_i$. Define the orthogonal matrix $\hat{Q}$ by

$$\hat{Q} = \hat{P}_1 \cdot \hat{P}_2 \cdot \cdots \cdot \hat{P}_r . \tag{9.4}$$

From (8.17) and (8.18) there exists a perturbation δC in C such that

$$[\overline{L}|0] = (C + \delta C)\,\hat{Q}, \tag{9.5}$$

where

(9.6a) $\quad ||\delta C|| \leqslant 2^{-t} \overline{K}_3(r)\, ||C||$, or

(9.6b) $\quad ||\delta C|| = O(2^{-t}\, ||C||)$.

Step 2: Calculate $\overline{L}^{-1} g$. From (8.10) and (8.11) there exists a perturbation δL^* in $\overline{L}$ such that

(9.7) $\quad fl\,(\overline{L}^{-1} g) = (\overline{L} + \delta L^*)^{-1} g,$

where $(\overline{L} + \delta L^*)^{-1}$ exists because $\overline{L}^{-1}$ does and

(9.8) $\quad ||\delta L^*|| \leqslant 2^{-t} \overline{K}_2(r)\, ||\overline{L}||$.

Step 3: Calculate $\overline{B}_1$ and $\overline{B}_2$. Define for $1 \leqslant i \leqslant r$

(9.9) $\quad \overline{A}_{i+1} = fl\,(\overline{A}_i\, \overline{P}_i)$ where $\overline{A}_1 = A$,

(9.10) $\quad \left[\overline{B}_1 | \overline{B}_2\right] = \overline{A}_{r+1}$,

where $\overline{P}_i$ is defined by (9.3). From (8.17) and (8.18) there exists a matrix $\delta A^* \in \mathbb{R}^{mxn}$ such that

(9.11) $\quad \left[\overline{B}_1 | \overline{B}_2\right] = (A + \delta A^*)\, \hat{Q}$,

where

(9.12a) $\quad ||\delta A^*|| \leqslant 2^{-t} \overline{K}_3(n)\, ||A||$, or

(9.12b) $\quad ||\delta A^*|| = O(2^{-t}\, ||A||)$.

Step 4: Calculate $\bar{B}_1(\bar{L} + \delta L^*)^{-1} g$.

Define $\varepsilon_1 \in \mathbb{R}^m$ by

$$(9.13) \qquad fl(\bar{B}_1 \cdot (\bar{L} + \delta L^*)^{-1} g) = \bar{B}_1(\bar{L} + \delta L^*)^{-1} g + \varepsilon_1 .$$

A bound for ε_1 is given in (8.6)

$$(9.14) \qquad ||\varepsilon_1|| \leqslant 2^{-t} \bar{K}_1(r) \, ||\bar{B}_1|| \, ||(\bar{L} + \delta L^*)^{-1} g|| .$$

Step 5: Calculate $fl(f - [\bar{B}_1(\bar{L} + \delta L^*)^{-1} g + \varepsilon_1])$.

Define $\varepsilon_2 \in \mathbb{R}^m$ by

$$(9.15) \qquad fl(f - [\bar{B}_1(\bar{L} + \delta L^*)^{-1} g + \varepsilon_1])$$

$$= f - (\bar{B}_1(\bar{L} + \delta L^*)^{-1} g + \varepsilon_1) + \varepsilon_2 .$$

A bound for ε_2 is given by (8.3)

$$(9.16) \qquad ||\varepsilon_2|| \leqslant 2^{-t} ||f - (\bar{B}_1(\bar{L} + \delta L^*)^{-1} g + \varepsilon_1)||$$

Step 6: Calculate $\bar{W}$ by reducing $\bar{B}_2$ with elementary orthogonal transformations.

Define for $1 \leqslant i \leqslant n-r$

$$(9.17) \qquad \bar{B}_{2,i+r} = fl(\bar{U}_i \bar{B}_{2,i}) \text{ and } \bar{B}_{2,1} = B_2 ,$$

where

$$(9.18) \qquad \bar{U}_i = fl(\hat{U}_i) .$$

$\hat{U}_i \in \mathbb{R}^{mxm}$ is that elementary orthogonal transformation which exactly reduces $\bar{B}_{2,i}$. Define the orthogonal matrix $\hat{V}$ by

$$\hat{V} = \hat{U}_{n-r} \cdot \cdots \cdot \hat{U}_1 \quad . \tag{9.19}$$

From (8.17) and (8.18) there exists a matrix $\delta B_2^* \in \mathbb{R}^{mx(n-r)}$ such that

$$\begin{bmatrix} \overline{W} \\ \overline{0} \end{bmatrix} = \overline{B}_{2,((n-r)+1)} = \hat{V}(\overline{B}_2 + \delta B_2^*) \tag{9.20}$$

and

$$||\delta B_2^*|| \leqslant 2^{-t} \overline{K}_3(n-r) \, ||\overline{B}_2|| \, , \tag{9.21}$$

where $\overline{W} \in \mathbb{R}^{(n-r)x(n-r)}$ is an upper triangular matrix.

Step 7: Premultiplication by $\hat{V}$.

Define

$$\overline{z}_1 = \{ f - (\overline{B}_1(\overline{L} + \delta L^*)^{-1} g + \mathcal{E}_1) + \mathcal{E}_2 \} \tag{9.22}$$

and for $1 \leqslant i \leqslant n-r$

$$\overline{z}_{i+1} = fl(\overline{U}_i \, \overline{z}_i) \, . \tag{9.23}$$

From (8.17) and (8.18) there exists a vector $\delta z \in \mathbb{R}^m$ such that

$$\overline{z}_{(n-r)+1} = \hat{V}(\overline{z}_1 + \delta z) \, , \tag{9.24}$$

where

$$||\delta z|| \leqslant 2^{-t} \overline{K}_3(n-r) \, ||\overline{z}_1|| \, . \tag{9.25}$$

We now define $y \in \mathbb{R}^{n-r}$ by

$$y = \left[I_{n-r} | 0 \right] \overline{z}_{r+1} = \hat{V}_1 (\overline{z}_1 + \delta z) \, , \tag{9.26}$$

where

(9.27) $\quad \hat{V}_1 = \left[I_{n-r} \mid 0 \right] \hat{V}$.

It is clear that $\hat{V}_1$ is the first n-r rows of $\hat{V}$.

Step 8: Calculate $\overline{W}^{-1} y$.

We have from (8.10) and (8.11), that there exists a perturbation δW^* such that

(9.28) $\quad fl(\overline{W}^{-1} y) = (\overline{W} + \delta W^*)^{-1} y$,

where $(\overline{W} + \delta W^*)^{-1}$ exists because $\overline{W}^{-1}$ exists and

(9.29) $\quad ||\delta W^*|| \leqslant 2^{-t} \overline{K}_2(n-r) \, ||\overline{W}||$.

Step 9: Premultiply by Q.

Define $\overline{\ell}_1 \in \mathbb{R}^n$ from (9.7), (9.28), (9.26) and (9.22) by

(9.30) $$\overline{\ell}_1 = \begin{bmatrix} (\overline{L} + \delta L^*)^{-1} g \\ \\ (\overline{W} + \delta W^*)^{-1} \hat{V}_1 \{\delta z + f - \overline{B}_1 (\overline{L} + \delta L^*)^{-1} g - \varepsilon_1 + \varepsilon_2\} \end{bmatrix}$$

and for $1 \leqslant i \leqslant r$

(9.31) $\quad \overline{\ell}_{i+1} = fl(\overline{P}_{(r+1-i)} \, \overline{\ell}_i)$.

Therefore our calculated solution is given by

(9.32) $\quad \overline{x} = \overline{\ell}_{r+1}$,

and from (8.17) and (8.18) there exists a vector $\delta \ell^* \in \mathbb{R}^n$ such that

(9.33) $\quad \overline{x} = \hat{Q} (\overline{\ell}_1 + \delta \ell^*)$,

where

$$(9.34) \qquad ||\delta \ell^*|| \leq 2^{-t} \overline{K}_3(1)\, ||\overline{\ell}_1|| \,.$$

From (9.33) and (9.30) we have an exact representation of our calculated least squares vector $\overline{x}$ using floating point arithmetic. We now proceed to estimate how far $\overline{x}$ is from the true solution x given in (9.1). This will be accomplished in two parts. First we will estimate how far $\overline{x}$ is from an intermediate vector $\hat{x}$ which is the exact solution of a perturbed problem, and then estimate how far $\hat{x}$ is from the exact solution x of the original problem.

In order to obtain a compact expression, we will make the assumption that the following quantities are moderately small. For a real Δ satisfying $0 < \Delta < 1$ assume that

$$(9.35a) \qquad h_1 ||L^{-1} L|| \leq \Delta \text{ where } h_1 = \max\{1, \sqrt{r/2}\} \,,$$

$$(9.35b) \qquad 2^{-t} \overline{K}_2(r)\, \chi(C) \leq (\Delta - \Delta^2)/(1+\Delta) \,,$$

$$(9.35c) \qquad h_2 ||\delta W W^{-1}|| \leq \Delta \text{ where } h_2 = \max\{1, \sqrt{(n-r)/2}\} \,,$$

$$(9.35d) \qquad 2^{-t} \overline{K}_2(n-r)\, \chi(B_2) \leq (\Delta - \Delta^2)/(1+\Delta) \,,$$

$$(9.35e) \qquad ||\hat{R}||/||f|| \leq \Delta \,,$$

$$(9.35f) \qquad ||\hat{x} - x||/||x|| \leq \Delta \,,$$

where $\hat{x}$ and $\hat{R}$ will be defined in (9.37) and (9.38). Also, assume that the conditions needed to satisfy the hypotheses of Lemma 5.1 and Theorem 7.1 are met.

It should be clear from (9.8), (6.5c), (9.35a) and (9.35b) that

(9.35g) $$||\bar{L}^{-1}||\ ||\delta L^{*}|| \leq \Delta ,$$

where $\hat{L} = \bar{L} = \delta L + L$. Likewise from (9.29), (6.6c), (9.35c) and (9.35d) we have

(9.35h) $$||\delta W^{*}||\ ||\bar{W}^{-1}|| \leq \Delta ,$$

where $\hat{W} = \bar{W} = \delta W + W$.

We will now account for most of our errors by interpreting them as errors induced by a perturbation of the initial problem (9.1), and then apply Theorem 7.1 to bound this portion of the error. Consider the following constrained least squares problem

(9.36a) $$\hat{C} = C + \delta C,\ \hat{g} = g,\ \text{i.e. } \delta g = 0 ,$$

where δC is defined in (9.5),

(9.36b) $$\hat{A} = A + \delta A,\ \hat{f} = f,\ \text{i.e. } \delta f = 0 ,$$

where

(9.36c) $$\delta A = \left[0 \mid \delta B_2^{*}\right] \hat{Q}^{T} + \delta A^{*} ,$$

δB_2^{*} and δA^{*} are defined in (9.20) and (9.11) respectively, with solution $\hat{x}$ which satisfies

(9.37) $$\hat{C}\hat{x} = g ,$$

minimizing the norm of

(9.38) $$\hat{R} = \hat{A}\hat{x} - f .$$

From (9.5) it is clear that the orthogonal matrix $\hat{Q}$ reduces $\hat{C}$ exactly to $[\bar{L}|0]$, where (9.6b) gives a bound for $||\delta C||$. Also, from (9.36b) and (9.11) we have

$$\hat{A}\,\hat{Q} = [\bar{B}_1 | \bar{B}_2 + \delta B_2^*], \tag{9.39}$$

and from (9.20) the orthogonal matrix $\hat{V}$ reduces $\bar{B}_2 + \delta B_2^*$ to $\begin{bmatrix} \bar{W} \\ \hline 0 \end{bmatrix}$. This is the reason we have $\hat{L} = \bar{L}$ and $\hat{W} = \bar{W}$.

Therefore, using Algorithm 1.1 we have

$$\hat{x} = \hat{Q} \begin{bmatrix} \bar{L}^{-1} g \\ \\ \bar{W}^{-1} \hat{V}_1 \{f - \bar{B}_1 \bar{L}^{-1} g\} \end{bmatrix} = \hat{Q} \begin{bmatrix} \hat{d}_1 \\ \hat{d}_2 \end{bmatrix}, \tag{9.40}$$

where

$$(\bar{B}_2 + \delta B_2^*)\,\hat{d}_2 = f - \bar{B}_1 \hat{d}_1 + \hat{R}. \tag{9.41}$$

In order to obtain a bound for $||\bar{B}_1||$, $||\bar{B}_2||$ we use (9.11) and apply (3.13), (3.9) and (9.12b) to obtain

$$||\bar{B}_1||,\ ||\bar{B}_2|| = O(||A||). \tag{9.42}$$

Using (9.42) on (9.21) we have

$$||\delta B_2^*|| = O(2^{-t}\,||A||). \tag{9.43}$$

Therefore, from (9.36c), (9.43), (3.13) and (9.12b) it follows that

$$||\delta A|| = O(2^{-t}\,||A||), \tag{9.44}$$

where $\delta A = \delta B_2^* [0|I]\,\hat{Q}^T + \delta A^*$.

If we use Theorem 7.1, (9.44) and (9.6b) it follows that

$$(9.45)\qquad ||\hat{x}-x||/||x|| = O(2^{-t}\chi(C))$$

$$+ O(2^{-t}\chi(B_2)\,||A||/||B_2||)$$

$$+ O(2^{-t}\chi(C)\,\chi(B_2)\,||A||/||B_2||)$$

$$+ O(2^{-t}\chi(C)\,\chi^2(B_2)(||R||/||f||)(||A||^2/$$

$$||B_2||^2)),$$

where in the last expression we have merged terms by using the fact that $\chi(C) \geq 1$.

We now want to estimate the error which we cannot account for by perturbation. Toward this end use (9.33), (9.30) and (9.40) to define $\Delta_1 \in \mathbb{R}^r$ and $\Delta_2 \in \mathbb{R}^{n-r}$

$$(9.46)\qquad \bar{x} - \hat{x} = \hat{Q}\,\delta\ell^* + \hat{Q}\begin{bmatrix}\Delta_1\\ \Delta_2\end{bmatrix}.$$

Taking norms, it follows from (9.46) and (6.10) that

$$(9.47)\qquad ||\bar{x}-\hat{x}|| \leq ||\delta\ell^*|| + ||\Delta_1|| + ||\Delta_2||.$$

We will consider first the vector Δ_1, and use (9.30), (9.40) and (3.17) to obtain

$$(9.48)\qquad \Delta_1 = (\bar{L} + \delta L^*)^{-1} g - \bar{L}^{-1} g$$

$$= -(I + \bar{L}^{-1}\,\delta L^*)^{-1}\,\bar{L}^{-1}\,\delta L^*\,\hat{d}_1 .$$

Taking norms and using (9.35g) and (3.16) bounds the first inverse, using (9.8), (9.35a), (6.5c) and (6.10) yields

$$(9.49) \qquad ||\Delta_1|| = O(2^{-t}\,\chi(C)\,||\hat{x}||).$$

Before evaluating $||\Delta_2||$, we will bound the vectors $\mathcal{E}_1$ and $\mathcal{E}_2$. To accomplish this, we show the following result from (9.39) by using (3.9) and (3.13)

$$(9.50) \qquad ||\bar{B}_1||,\ ||\bar{B}_2 + \delta B_2^*|| \leq ||\hat{A}||\ .$$

Also, it follows from the fact that $\bar{L}^{-1} g = \hat{d}_1$

$$(9.51) \qquad (\bar{L} + \delta L^*)^{-1} g = (I + \bar{L}^{-1}\delta L^*)^{-1}\,\hat{d}_1\ .$$

Therefore, from (9.14), (9.50), and (9.51) we have using (9.35g), (3.16) and (6.10) that

$$(9.52) \qquad ||\mathcal{E}_1|| = O(2^{-t}\,||\hat{A}||\ ||\hat{x}||)\ .$$

From (9.16), a bound for $||\mathcal{E}_2||$ follows from the simplification of the expression

$$\begin{aligned}(9.53) \qquad f - B_1(\bar{L} + \delta L^*)^{-1} g - \mathcal{E}_1 &= f - \bar{B}_1\,\bar{L}^{-1} g + \bar{B}_1\,(\bar{L} \\ &\quad + \delta L^*)^{-1}\,\delta L^*\,\bar{L}^{-1} g - \mathcal{E}_1 \\ &= (\bar{B}_2 + \delta B_2^*)\,\hat{d}_2 - \hat{R} + \bar{B}_1(I + \bar{L}^{-1}\delta L^*)^{-1}\bar{L}^{-1}\delta L^*\,\hat{d}_1 - \mathcal{E}_1,\end{aligned}$$

where we have used (3.17), (9.40) and (9.41). In (9.35g) it is assumed that $||\bar{L}^{-1}||\ ||\delta L^*|| \leq \Delta < 1$, upon applying this <u>twice</u> and using (9.50), (3.16), (6.10) and (9.52) we have that the norm of

(9.53) is $O(||\hat{A}||\ ||\hat{x}||) + O(||\hat{R}||)$, hence from (9.16)

$$(9.54) \qquad ||\mathcal{E}_2|| = O(2^{-2}\ ||\hat{A}||\ ||\hat{x}||) + O(2^{-t} \cdot [||\hat{R}||/(||\hat{A}||\ ||\hat{x}||)]||\hat{A}||\ ||\hat{x}||) .$$

Using (9.38) we find $||\hat{A}||\ ||\hat{x}|| \geq ||f|| - ||\hat{R}||$, which, when combined with (9.54) and the assumption (9.35e) that $||\hat{R}||/||f|| \leq \Delta < 1$, yields

$$(9.55) \qquad ||\mathcal{E}_2|| = O(2^{-t}\ ||\hat{A}||\ ||\hat{x}||) .$$

We now define the vector $\mathcal{E}_3 \in \mathbb{R}^m$ by

$$(9.56) \qquad \mathcal{E}_3 = \{f - \bar{B}_1(\bar{L} + \delta L^*)^{-1} g - \mathcal{E}_1 + \mathcal{E}_2\} - \{f - \bar{B}_1 \bar{L}^{-1} g\}$$
$$= \bar{B}_1 (I + \bar{L}^{-1} \delta L^*)^{-1} \bar{L}^{-1} \delta L^* \hat{d}_1 - \mathcal{E}_1 + \mathcal{E}_2 .$$

A bound for $||\mathcal{E}_3||$ is obtained in similar fashion as that employed for $||\mathcal{E}_2||$, therefore we have from (9.50), (9.35g), (3.16), (9.8), (9.35a), (6.5c), (9.52) and (9.55) that

$$(9.57) \qquad ||\mathcal{E}_3|| = O(2^{-t} \chi(C)\ ||\hat{A}||\ ||\hat{x}||) .$$

Finally, using (9.25) and (9.22) we have

$$(9.58) \qquad ||\delta z|| \leq 2^{-t} \bar{K}_3(n-r)\ ||f - \bar{B}_1(\bar{L} + \delta L^*)^{-1} g - \mathcal{E}_1 + \mathcal{E}_2|| .$$

Upon examination of (9.16) and (9.58), we see that they differ only by $\mathcal{E}_2$ and the factor $\bar{K}_3(n-r)$, therefore using (9.55) it follows that

$$(9.59) \qquad ||\delta z|| = O(2^{-t}\ ||\hat{A}||\ ||\hat{x}||) .$$

The definition of Δ_2 comes from (9.46), (9.40), (9.33) and (9.30),

which yields the following simplification using (9.56), (3.17) and (9.40)

$$\Delta_2 = (\bar{W} + \delta W^*)^{-1} \hat{V}_1 \left[\delta z + \{f - \bar{B}_1 (\bar{L} + \delta L^*)^{-1} g - \mathcal{E}_1 + \mathcal{E}_2\}\right] - \bar{W}^{-1} \hat{V}_1 \left[f - \bar{B}_1 \bar{L}^{-1} g\right]$$
$$= \bar{W}^{-1} (I + \delta W^* \bar{W}^{-1})^{-1} \left[-\delta W^* \hat{d}_2 + \hat{V}_1 (\mathcal{E}_3 + \delta z)\right]. \tag{9.60}$$

Using (9.35h), (3.16), (9.29), (6.10), (3.14), (3.7), (9.57), (9.59), (6.6b) and the fact that $\hat{W} = \bar{W}$, it follows upon taking norms

$$||\Delta_2|| = O(2^{-t} \chi(C) ||W^{-1}|| ||\hat{A}|| ||\hat{x}||), \tag{9.61}$$

where we have used the fact that $||\bar{W}|| = ||\bar{B}_2 + \delta B_2^*|| \leq ||\hat{A}||$, and $\chi(C) \geq 1$.

Finally, from (9.44) it follows that $||\hat{A}|| = O(||A||)$, hence

$$||\Delta_2|| = O(2^{-t} \chi(C) \chi(B_2) ||\hat{x}|| ||A||/||B_2||), \tag{9.62}$$

where we have used the fact that $||W|| = ||B_2||$ and (6.4).

We are now in a position to bound $||\delta \ell^*||$. From (9.34) we have

$$||\delta \ell^*|| = O(2^{-t} ||\bar{\ell}_1||), \tag{9.63}$$

and substitution of (9.33) into (9.46) yields after rearrangement

$$\bar{\ell}_1 = \begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix} + \hat{Q}^T \hat{x}. \tag{9.64}$$

Therefore, a bound for $||\delta \ell^*||$ follows from (9.63), (9.64), (9.49) and (9.62)

$$(9.65)\qquad ||\delta \ell^{*}|| = O(2^{-2t}\,\chi(C)\,||\hat{x}||) + O(2^{-2t}\,\chi(C)\,\chi(B_2)\,||\hat{x}||\,||A||/||B_2||) + O(2^{-t}\,||\hat{x}||).$$

We now have our bound for $||\bar{x}-\hat{x}||$ from (9.47), (9.49), (9.62), (9.65) and the fact that $\chi(C) \geq 1$

$$(9.66)\qquad ||\bar{x}-\hat{x}|| = O(2^{-t}\,\chi(C)\,||\hat{x}||) + O(2^{-t}\,\chi(C)\,\chi(B_2)\,||\hat{x}||\,||A||/||B_2||).$$

From (9.45), we have a bound for $||\hat{x}-x||/||x||$, which can be tested to see if our assumption that $||\hat{x}-x||/||x|| < \Delta$ is reasonable. If it is, then from (9.66) we have

$$(9.67)\qquad ||\bar{x}-\hat{x}|| = O(2^{-t}\,\chi(C)\,||x||) + O(2^{-t}\,\chi(C)\,\chi(B_2)\,||x||\,||A||/||B_2||).$$

From (9.45) and (9.67) follows our final result

$$(9.68)\qquad ||\bar{x}-x||/||x|| = O(2^{-t}\,\chi(C)) + O(2^{-t}\,\chi(B_2)\,||A||/||B_2||) + O(2^{-t}\,\chi(C)\,\chi(B_2)\,||A||/||B_2||) + O(2^{-t}\,\chi(C)\,\chi^{2}(B_2)(||R||/||f||)(||A||/||B_2||)^{2}).$$

(9.68) shows the errors introduced by solving (9.1) using floating point arithmetic. Here we have suppressed all quantities which are constant (depend only on Δ, r, n, and m) to illustrate the

dependence on the machine accuracy, 2^{-t}, and condition of the matrices C and B_2. As hoped for, the influence of the condition of the constraint equations C appears only as a linear factor in (9.68). However, the term of major influence is the last one, in which the condition of B_2 is squared. The effect of this term will be minimized if the quantity $\chi(B_2)\,||R||/||f||$ is not too large, i.e. the residual R is small compared to f.

CHAPTER 2

BLENDING FUNCTION THEORY

In Section 1, an introduction to blending function theory as developed by Gordon and Hall [12, 13, 14, 15, 16, 17] is given. The presentation will be a generalization of their results to the more general setting of interpolation spaces. Most of the proofs of Section 1 will follow from their own. However, the more general setting of Section 1 will enable the application of blending function techniques in the following chapter to be accomplished with greater ease.

In Section 2 the dimension of discretized blending function spaces will be shown and several bases will be explicitly developed in terms of the cardinal bases for the corresponding interpolation spaces.

In Sections 3 and 4, natural cubic spline blending will be developed with error bounds given for "approximate" natural cubic spline blending.

Finally, for $f \in C[a,b]$ and $g \in C([a,b]x[c,d])$ we define

$$||f|| = \sup_{a \leq x \leq b} |f(x)| \quad \text{and} \quad ||g|| = \sup_{\substack{a \leq x \leq b \\ c \leq y \leq d}} |g(x,y)|.$$

Section 1. Spline and Blending Function Interpolants

Given a mesh $\pi_x: a = x_1 < x_2 < \cdots < x_M = b$, the space of cubic splines on π_x is defined to be

(1.1) $$S^2(\pi_x) = \{ s \in C^2[a,b] \mid s(x) \text{ is a cubic polynomial on } [x_i, x_{i+1}] \text{ for } 1 \leqslant i \leqslant M-1 \},$$

see Schultz [29].

Let $f \in C^{(1)}[a,b]$, then the type 1 cubic spline interpolant, s_f, associated with f is defined to be the unique spline which satisfies

(1.2a) $$s_f(x_i) = f(x_i), \quad 1 \leqslant i \leqslant M$$

(1.2b) $$s_f'(x_i) = f'(x_i), \quad i = 1, M.$$

The following theorem of Carlson and Hall [5] gives error bounds for type 1 spline interpolation.

Theorem 1.1 (Carlson and Hall [5]). Let $f \in C^{(m)}[a,b]$ and π_x be a mesh on $[a,b]$. Then for $1 \leqslant m \leqslant 4$, $0 \leqslant r \leqslant \min\{m, 3\}$

(1.3) $$||(s_f - f)^{(r)}|| \leqslant \mathcal{E}_{mr} ||f^{(m)}|| \, h_x^{m-r},$$

where the mesh size $h_x = \max_{1 \leqslant i \leqslant M-1} \Delta x_i$; $\Delta x_i = x_{i+1} - x_i$, the mesh ratio $M_\pi = \max_{1 \leqslant i \leqslant M-1} \Delta x_i / \min_{1 \leqslant i \leqslant M-1} \Delta x_i$ and $\mathcal{E}_{mr}$ is given in Table 1.

TABLE 1

$\mathcal{E}_{mr}$	r = 0	r = 1	r = 2	r = 3
m = 1	15/4	14	--	--
m = 2	9/8	4	10	--
m = 3	71/216	31/27	5	$(63 + 8\ M_{\pi}^{2})/9$
m = 4	5/384	$(9+\sqrt{3})/216$	5	$(2+M_{\pi}^{2})/4$

Bivariate Functions:

We will present here an introduction to blending functions as developed by Gordon and Hall [12],[13],[14], [16] . To accomplish this, it will be desirable to introduce the following general notation, which will allow us to develop at one time many of their results. Although, it should be pointed out that the methods of proof used here are just slight generalizations of their own.

In what follows, we use the notation

$$(1.4) \qquad f^{(k,\ell)} = D^{(k,\ell)}[f] = \partial^{k+\ell} f/\partial x^{k} \partial y^{\ell} .$$

Definition 1.1: Let $C^{(m,n)}([a,b] \times [c,d]) = C^{(m,n)}$ be the space of real valued functions with domain $[a,b] \times [c,d]$ such that if $f \in C^{(m,n)}$ then $f^{(k,\ell)}$ exists and is continuous for $0 \leqslant k \leqslant m$ and $0 \leqslant \ell \leqslant n$.

Definition 1.2: A projection operator (projector) P is an idempotent linear operator from a function space onto a subspace of the function space.

We will consider here only projectors which separate the x and y variables, where the dependence on one of the variables can be represented by an element from a finite dimensional vector space. This space is to satisfy an interpolation property for certain values of the function and its derivatives at specified points. Therefore, we introduce the following notation:

We are given the mesh $\pi_x : a \leq x_1 \leq x_2 \leq \cdots \leq x_M \leq b$ where the points are not necessarily distinct. Also, the non-negative integer m, the interpolation function $\alpha : I_M \longrightarrow J_m$, where $I_M = \{i \text{ is an integer } | 1 \leq i \leq M\}$ and $J_m = \{j \text{ is an integer } | 0 \leq j \leq m\}$, satisfying the following restraint: if there exists distinct natural numbers s and t such that $x_s = x_t$ then $\alpha(s) \neq \alpha(t)$.

Definition 1.3: $V(\pi_x, M, m) \subseteq C^{(m)}[a, b]$ is a finite dimensional interpolation vector space with respect to α on the mesh π_x if and only if there exists an algebraic basis $\{\varphi_i\}_{i=1}^{M}$ satisfying the following cardinality conditions for $1 \leq i, j \leq M$.

$$(1.5) \qquad (D^{(\alpha(i))}(\phi_i))(x_j) = \delta_{ij} = \begin{cases} 1 \text{ if } i=j \\ 0 \text{ if } i \neq j \end{cases} .$$

Such a basis for $V(\pi_x, M, m)$ is called a cardinal basis.

In terms of the cardinal basis, if $f \in C^{(m)}[a,b]$, the the function

$$(1.6) \qquad p(x) = \sum_{i=1}^{M} f^{(\alpha(i))}(x_i)\cdot \phi_i(x)$$

satisfies the interpolation conditions

$$(1.7) \qquad (D^{(\alpha(i))}p)(x_i) = f^{(\alpha(i))}(x_i), \text{ for } 1 \leqslant i \leqslant M.$$

Example 1.1: Let $V(\pi_x, M, 0)$ be the space of Lagrange interpolation polynomials of degree $M-1$ on $[a,b]$ interpolating on the mesh $\pi_x: a \leqslant x_1 < x_2 < \cdots < x_M \leqslant b$, then for $1 \leqslant i \leqslant M$, $\alpha(i) = 0$ and the cardinal basis is $\{\ell_i\}_{i=1}^{M}$ where

$$(1.8) \qquad \ell_i(x) = \prod_{\substack{j=1 \\ j\neq i}}^{M} (x-x_j) \Big/ \prod_{\substack{j=1 \\ j\neq i}}^{M} (x_i-x_j).$$

Given $f \in C^{(0)}[a,b]$, then the interpolation polynomial is given by

$$(1.9) \qquad p(x) = \sum_{i=1}^{M} f^{(\alpha(i))}(x_i)\,\ell_i(x),$$

where $f^{(\alpha(i))}(x_i) = f(x_i)$.

Example 1.2: Type 1 cubic splines on $M-2 \geqslant 2$ knots.

Then $V(\pi_x, M, 1) = S^2(\pi_x)$ on the mesh $\pi_x: a = x_1 = x_2 < x_3 < \cdots < x_{M-2} < x_{M-1} = x_M = b$, where $\alpha(1) = \alpha(M) = 1$ and for $2 \leqslant i \leqslant M-1$, $\alpha(i) = 0$. Also, the interpolation spline satisfies conditions (1.2a) and (1.2b), and the cardinal basis $\{C_i(x)\}_{i=1}^{M}$ satisfies for $1 \leqslant j \leqslant M$

(1.10a) $C_i(x_j) = \delta_{ij}$, $C_i'(x_1) = C_i(x_M) = 0$ for $2 \leqslant i \leqslant M-1$

(1.10b) $C_1'(x_1) = C_M'(x_M) = 1$, $C_i(x_j) = 0$ for $i = 1, M$.

Example 1.3: Cubic Hermite splines on M knots.

Then $V(\pi_x, 2M, 1) = H^1(\pi_x)$ on the mesh $\pi_x: a = x_1 = x_2 < x_3 = x_4 < \cdots < x_{2M-1} = x_{2M} = b$, where $\alpha(i) = (i-1) \bmod 2$ for $1 \leqslant i \leqslant 2M$. Given $f \in C^{(1)}[a, b]$, the Hermite cubic spline interpolant $p(x)$ satisfies $p^{(\alpha(i))}(x_i) = f^{(\alpha(i))}(x_i)$ for $1 \leqslant i \leqslant 2M$. The cardinal basis $\{H_i\}_{i=1}^{2M}$ satisfies for $1 \leqslant j \leqslant M$ and $1 \leqslant i \leqslant M$

(1.11a) $H_{2i-1}(x_{2j-1}) = \delta_{ij}$, $H_{2i-1}'(x_{2j}) = 0$,

(1.11b) $H_{2i}(x_{2j-1}) = 0$, $H_{2i}'(x_{2j}) = \delta_{ij}$.

Remark: Even though our notation $V(\pi_x, M, m)$ does not include a reference to α and $\{\phi_i\}_{i=1}^{M}$, it is understood that these two quantities are always associated with $V(\pi_x, M, m)$.

We are now in a position to define the projection operator P_x of a bivariate function. Given the interpolation vector space $V(\pi_x, M, m)$, the corresponding interpolation function α, basis $\{\phi_i\}_{i=1}^{M}$ and $f \in C^{(m, 0)}$, we define $P_x[f]$ pointwise to be

(1.12a) $$(P_x[f])(x, y) = \sum_{i=1}^{M} f^{(\alpha(i), 0)}(x_i, y)\, \phi_i(x),$$

where it is clear that

(1.12b) $\quad (D^{(\alpha(i),\,0)}\, P_x[f])\,(x_i,\, y) = f^{(\alpha(i),\,0)}\,(x_i,\, y)$

for $1 \leqslant i \leqslant M$. Examination of (1.12a) and (1.12b) shows that P_x is indeed a projection operator on $C^{(m,\,0)}$.

We define the univariate interpolation vector space $V(\pi_y,\, N,\, n) \subseteq C^{(n)}[c,\, d]$ in the variable y with interpolation function $\beta : I_N \longrightarrow J_N$ and cardinal basis $\{\psi_j\}_{j=1}^{N}$ in an identical manner as above. The corresponding projection operator $P_y[f]$ for $f \in C^{(0,\,n)}$ is defined by

$$(1.13) \qquad (P_y[f])\,(x, y) = \sum_{j=1}^{N} f^{(0,\,\beta(j))}\,(x,\, y_j)\ \psi_j(y)\,,$$

where for $1 \leqslant j \leqslant N$

$$(1.14) \qquad (D^{(0,\,\beta(j))}\, P_y[f])\,(x,\, y_j) = f^{(0,\,\beta(j))}\,(x,\, y_j)\,.$$

In the usual manner, we define the sum or difference of two operators to be the pointwise sum or difference, and the product as composition. Direct calculation shows that the following statements are valid on $C^{(m_1,\,n_1)}$ where $m_1 \geqslant m$ and $n_1 \geqslant n$, and P_x and P_y are defined as above.

$$(1.15a) \qquad P_x\, P_y = P_y\, P_x\,,$$

$$(1.15b) \qquad D^{(0,\,\ell)}\, P_x = P_x\, D^{(0,\,\ell)} \quad \text{for} \quad 0 \leqslant \ell \leqslant n_1, \text{ and}$$

$$(1.15c) \qquad D^{(k,\,0)}\, P_y = P_y\, D^{(k,\,0)} \quad \text{for} \quad 0 \leqslant k \leqslant m_1\,.$$

We define the projection operators

(1.16a) $R_x = I - P_x$,

(1.16b) $R_y = I - P_y$,

where I is the identity operator. The following relations are a direct consequence of (1.15a) to (1.16b). On $C^{(m_1, n_1)}$ where $m_1 \geqslant m$ and $n_1 \geqslant n$

(1.17a) $R_x R_y = R_y R_x$

(1.17b) $D^{(0, \ell)} R_x = R_x D^{(0, \ell)}$ for $0 \leqslant \ell \leqslant n_1$, and

(1.17c) $D^{(k, 0)} R_y = R_y D^{(k, 0)}$ for $0 \leqslant k \leqslant m_1$.

The first observation that we make is that given $f \in C^{(m, n)}$, then $P_x P_y[f]$ is the tensor product interpolant to f, where

$$(1.18) \qquad (P_x P_y[f])(x, y) = \sum_{i=1}^{M} \sum_{j=1}^{N} f^{(\alpha(i), \beta(j))}(x_i, y_j)\, \phi_i(x)\, \psi_j(y) ,$$

see Gordon and Hall [14]. It follows from (1.18) that for $1 \leqslant i \leqslant M$ and $1 \leqslant j \leqslant N$

$$(1.19) \qquad (D^{(\alpha(i), \beta(j))} P_x P_y[f])(x_i, y_j) = f^{(\alpha(i), \beta(j))}(x_i, y_j) .$$

The error of the tensor product interpolant to f is given by

$$(1.20) \qquad (I - P_x P_y)[f] = R_x[f] + R_y[f] - R_x R_y[f] .$$

Gordon and Hall [17] (see Theorem 1.2 below) have shown that there is an interpolation scheme using the projection operators P_x and P_y which will eliminate the terms $R_x[f]$ and $R_y[f]$ in (1.20).

Definition: The operator on $C^{(m,n)}$ defined by

$$(1.21) \qquad P_x \oplus P_y = P_x + P_y - P_x P_y$$

is the blending function operator and $P_x \oplus P_y[f]$ is the blending function interpolant to the function $f \in C^{(m,n)}$.

Theorem 1.2. (Gordon and Hall [17]). If $f \in C^{(m,n)}$ then

$$(1.22) \qquad (I - P_x \oplus P_y)[f] = R_x R_y [f] .$$

Proof: $$\begin{aligned} I - (P_x + P_y - P_x P_y) &= I - P_x + (I - P_x) P_y \\ &= (I - P_x)(I - P_y) \\ &= R_x R_y . \end{aligned}$$

■

The set of functional values and derivatives on which $P_x \oplus P_y$ interpolates f is given in the following theorem.

Theorem 1.3. (Gordon and Hall [17]). If $f \in C^{(m,n)}$, then for $1 \leq i \leq M$ and $y \in [c,d]$

$$(1.23a) \qquad (D^{(\alpha(i),0)} P_x \oplus P_y[f])(x_i, y) = f^{(\alpha(i),0)}(x_i, y) ,$$

also, for $1 \leqslant j \leqslant N$ and $x \in [a, b]$

$$(1.23b) \qquad (D^{(0, \beta(j))} P_x \oplus P_y[f])(x, y_j) = f^{(0, \beta(j))}(x, y_j) .$$

Proof: By direct calculation. ∎

What we have accomplished is the approximation of a bivariate function with a sum of univariate functions. Therefore, it should be clear that the image of $C^{(m, n)}$ under the operator $P_x \oplus P_y$ is not a finite dimensional space. In comparison, from (1.18), it is clear that the image of $C^{(m, n)}$ under the tensor product operator $P_x P_y$ is a finite dimensional space. Therefore, to implement the blending function method, it is usually necessary to make second level approximations to $f^{(\alpha(i), 0)}(x_i, \cdot)$ and $f^{(0, \beta(j))}(\cdot, y_j)$ which have an accuracy compatible with $P_x \oplus P_y[f]$. This is the sacrifice which must be made in order to achieve this gain in accuracy. Next, we want to show how much more accuracy is gained by blending function techniques, to see if it will justify the extra work of implementation. Toward this end, we note that the accuracy of blending function methods depends upon the interpolation accuracy of the two univariate spaces $V(\pi_x, M, m)$ and $V(\pi_y, N, n)$. Therefore, we introduce the following general notation for an univariate error bound.

Definition 1.4: Let $V(\pi_x, M, m_1)$ be an interpolation vector space, with interpolation function α, and cardinal basis $\{\phi_i\}_{i=1}^{M}$. We say $V(\pi_x, M, m_1)$ has an error bound for the k^{th} derivative if and only if there exists a function $g_x(k) = g_x(h_x, m_1, m_2, k)$ such that if $\hat{x} \in [a, b]$ is any point where $\phi_i^{(k)}(\hat{x})$ exists for $1 \leq i \leq M$, and $f \in C^{(m_2)}[a, b]$ where $m_2 \geq \max\{m_1, k\}$ then

$$(1.24) \qquad |f^{(k)}(\hat{x}) - \sum_{i=1}^{M} f^{(\alpha(i))}(x_i)\, \phi_i^{(k)}(\hat{x})| \leq g_x(k) || f^{(m_2)} ||,$$

where h_x is a parameter of the mesh π_x. It is clear, if $V(\pi_x, M, m_1)$ has an error bound for the k^{th} derivative, $g \in C^{(m_2, n)}$, $y \in [c, d]$, P_x is the corresponding projection operator and x is any point satisfying Definition 1.4, then

$$(1.25a) \qquad (D^{(k,0)} P_x[f])(\hat{x}, y) \text{ exists, and}$$

$$(1.25b) \qquad (D^{(k,0)} P_x[f])(\hat{x}, \cdot) \in C^{(n)}([c, d]).$$

Therefore, it follows from (1.15b) that for $0 \leq \ell \leq n$

$$(1.26a) \qquad (D^{(k,\ell)} P_x[f])(\hat{x}, y) = (D^{(k,0)} P_x[f^{(0,\ell)}])(\hat{x}, y),$$

correspondingly

$$(1.26b) \qquad (D^{(k,\ell)} R_x[f])(\hat{x}, y) = (D^{(k,0)} R_x[f^{(0,\ell)}])(\hat{x}, y)$$

and finally

$$(1.26c) \qquad |(D^{(k,\ell)} R_x[f])(\hat{x}, y)| \leq g_x(k) \sup_{a \leq t \leq b} |f^{(m_2,\ell)}(t, y)|.$$

For the interpolation space $V(\pi_y, N, n_1)$, interpolation function $\beta: I_N \longrightarrow J_n$ and cardinal basis $\{\psi_j\}_{j=1}^{N}$, we use an analogous definition for an error bound for the ℓ^{th} derivative, and we have conclusions similar to those given by (1.25a) to (1.26c).

We are now in a position to prove the following theorem, which is a generalization of the error analysis given by Gordon and Hall [17].

Theorem 1.4. Let $V(\pi_x, M, m_1)$ be an interpolation vector space which has an error bound for the k^{th} derivative for $0 \leqslant k \leqslant m_3$, where m_3 is an integer such that $0 \leqslant m_3 \leqslant m_2$. Also, let $V(\pi_y, N, n_1)$ be an interpolation vector space which has an error bound for the ℓ^{th} derivative for $0 \leqslant \ell \leqslant n_3$, where n_3 is an integer such that $0 \leqslant n_3 \leqslant n_2$. If $f \in C^{(m_2, n_2)}$, then

$$|((I - P_x \oplus P_y)[f])^{(k,\ell)}(\hat{x}, \hat{y})| \leqslant g_x(k)\, g_y(\ell)\, ||f^{(m_2, n_2)}||, \tag{1.27}$$

and

$$|((I - P_x\, P_y)[f])^{(k,\ell)}(\hat{x}, \hat{y})| \leqslant g_x(k)||f^{(m_2, \ell)}|| + g_y(\ell)||f^{(k, n_2)}|| + g_x(k)\, g_y(\ell)||f^{(m_2, n_2)}||, \tag{1.28}$$

where $\hat{x}$, $\hat{y}$, m_2, n_2, $g_x(k)$ and $g_y(\ell)$ are given in Definition 1.4.

Proof: For $0 \leqslant q \leqslant n_2$ and for each $y \in [c, d]$ it follows from (1.26c)

$$(1.29)\qquad |(D^{(k,q)} R_x[f])(\hat{x}, y)| \leqslant g_x(k) \sup_{a \leqslant x \leqslant b} |f^{(m_2, q)}(\cdot, y)| .$$

Because $D^{(k,0)} R_x[f](\hat{x}, \cdot) \in C^{(n_2)}[c, d]$ and $\psi_j^{(\ell)}(\hat{y})$ exists for $1 \leqslant j \leqslant N$ we have that $(D^{(k,\ell)} R_y R_x[f])(\hat{x}, \hat{y})$ exists and from (1.29) and Theorem 1.2

$$(1.30)\qquad |(D^{(k,\ell)} R_y R_x[f])(\hat{x}, \hat{y})| = |(D^{(0,\ell)} R_y D^{(k,0)} R_x[f])(\hat{x}, \hat{y})|$$

$$\leqslant g_y(\ell) \sup_{c \leqslant y \leqslant d} |D^{(k, n_2)} R_x[f](\hat{x}, \cdot)|$$

$$\leqslant g_x(k)\, g_y(\ell)\, ||f^{(m_2, n_2)}|| .$$

The proof of (1.28) follows in a similar way from (1.20), (1.29) and (1.30)

■

Corollary 1.1. (Gordon and Hall [17]). If $f \in C^{(m,n)}([o, h] \times [o, h])$, $h < 1$, and the spaces $V(\pi_x, m, 0)$ and $V(\pi_y, n, 0)$ are polynomial spaces as defined in Example 1.1, then

$$(1.31)\qquad ||((I - P_x \oplus P_y)[f])^{(k,\ell)}|| \leqslant \bar{\mathcal{E}}_{mk}\, \bar{\mathcal{E}}_{n\ell}\, ||f^{(m,n)}||\, h^{m+n-(k+\ell)}$$

and

$$(1.32)\qquad ||((I - P_x\, P_y)[f])^{(k,\ell)}|| \leqslant \bar{\mathcal{E}}_{mk} ||f^{(m,\ell)}||\, h^{m-k}$$

$$+ \bar{\mathcal{E}}_{n\ell} ||f^{(k,n)}||\, h^{n-\ell} + \bar{\mathcal{E}}_{mk}\, \bar{\mathcal{E}}_{n\ell} ||f^{(m,n)}||\, h^{m+n-(k+\ell)} ,$$

where $0 \leqslant k < m$, $0 \leqslant \ell < n$, $\bar{\mathcal{E}}_{mk} = 1/(m-k)!$ and $\bar{\mathcal{E}}_{n\ell} = 1/(n-\ell)!$.

Proof: From [24, p. 289], if $f \in C^{(m)}[0, h]$, then

$$(1.33) \qquad |f^{(k)}(x) - \sum_{i=1}^{m} f^{(\alpha(i))}(x_i)\, \ell_i^{(k)}(x)| \leq \bar{\mathcal{E}}_{mk} ||f^{(m)}|| h^{m-k},$$

hence $g_x(h, 0, m, k) = \bar{\mathcal{E}}_{mk} h^{m-k}$. An application of Theorem 1.4 completes the proof. ∎

Corollary 1.2. (Carlson and Hall [5]). If $f \in C^{(m,n)}([a, b]x[c, d])$, $1 \leq m, n \leq 4$ and the spaces $V(\pi_x, M, 1)$ and $V(\pi_y, N, 1)$ are cubic spline spaces given in Example 1.2, then

$$(1.34) \qquad ||((I - P_x \oplus P_y)[f])^{(k,\ell)}|| \leq \mathcal{E}_{mk} \mathcal{E}_{n\ell} ||f^{(m,n)}|| h_x^{m-k} h_y^{n-\ell},$$

and

$$(1.35) \qquad ||((I - P_x P_y)[f])^{(k,\ell)}|| \leq \mathcal{E}_{mk} ||f^{(m,\ell)}|| h_x^{m-k}$$

$$+ \mathcal{E}_{n\ell} ||f^{(k,n)}|| h_y^{n-\ell} + \mathcal{E}_{mk} \mathcal{E}_{n\ell} ||f^{(m,n)}|| h_x^{m-k} h_y^{n-\ell},$$

where $0 \leq k \leq \min\{m, 3\}$, $0 \leq \ell \leq \min\{n, 3\}$, also $\mathcal{E}_{mk}$ and $\mathcal{E}_{n\ell}$ are given in Table 1, $h_x = \max_{2 \leq i \leq M-2} (x_{i+1} - x_i)$ and

$h_y = \max_{2 \leq j \leq N-2} (y_{j+1} - y_j)$.

Proof: We have an error bound for the derivatives given by Theorem 1.1, where $g_x(h_x, 1, m, k) = \mathcal{E}_{mk} h_x^{m-k}$. ∎

Carlson and Hall in the same paper [5] have given other error bounds for $P_x P_y[f]$ under weaker continuity requirements for f.

Corollary 1.3. If $f \in C^{(m,n)}([a,b] \times [c,d])$, $1 \leq m, n \leq 4$ and the spaces $V(\pi_x, 2M, 1)$ and $V(\pi_y, 2N, 1)$ are Hermite cubic spline spaces defined in Example 1.3 then

$$(1.36) \qquad ||((I - P_x \oplus P_y)[f])^{(h,\ell)}|| \leq \hat{\varepsilon}_{mk} \hat{\varepsilon}_{n\ell} ||f^{(m,n)}|| h_x^{m-k} h_y^{n-\ell},$$

and

$$(1.37) \qquad ||((I - P_x P_y)[f])^{(k,\ell)}|| \leq \hat{\varepsilon}_{mk} ||f^{(m,\ell)}|| h_x^{m-k} + \hat{\varepsilon}_{n\ell} ||f^{(k,n)}|| h_y^{n-\ell} + \hat{\varepsilon}_{mk} \hat{\varepsilon}_{n\ell} ||f^{(m,n)}|| h_x^{m-k} h_y^{n-\ell},$$

where $0 \leq k \leq \min\{m, 3\}$, $0 \leq \ell \leq \min\{n, 3\}$ also $\hat{\varepsilon}_{mk}$ and $\hat{\varepsilon}_{n\ell}$ are given in Table 2, $h_x = \max_{1 \leq i \leq M-1} (x_{2i+1} - x_{2i})$,

$h_y = \max_{1 \leq j \leq N-1} (y_{2j+1} - y_{2j})$.

TABLE 2

$\hat{\varepsilon}_{mk}$	k = 0	k = 1	k = 2	k = 3
m = 1	5/4	4	---	---
m = 2	3/4	5/2	12	---
m = 3	7/24	1	5	7
m = 4	1/384	$\sqrt{3}/216$	1/12	1/2

Proof: From Carlson and Hall [5], if $f \in C^{(m)}[a,b]$, then

$$(1.38) \qquad \| f^{(k)} - \sum_{i=1}^{2M} f^{(\alpha(i))}(x_i) H_i \| \leq \hat{\mathcal{E}}_{mk} \| f^{(m)} \| h_x^{m-k},$$

implying that $g_x(h_x, 1, m, k) = \hat{\mathcal{E}}_{mk} h_x^{m-k}$.

■

We are now in a position to describe the second level decomposition of the blending function interpolant, see [17]. The purpose of this is to create a finite dimensional interpolation scheme while preserving the accuracy of our blending techniques.

Define the

Meshes: $\bar{\pi}_x$: $a \leq \bar{x}_1 \leq \bar{x}_2 \leq \cdots \leq \bar{x}_{\bar{M}} \leq b$

$\bar{\pi}_y$: $c \leq \bar{y}_1 \leq \bar{y}_2 \leq \cdots \leq \bar{y}_{\bar{N}} \leq d$

Interpolation functions: $\bar{\alpha}: I_{\bar{M}} \longrightarrow J_{\bar{m}}$

$\bar{\beta}: I_{\bar{N}} \longrightarrow J_{\bar{n}}$

Cardinal bases: $\{\bar{\phi}_i\}_{i=1}^{\bar{M}}$

$\{\bar{\psi}_j\}_{j=1}^{\bar{N}}$

Projection operators: $\bar{P}_x$ and $\bar{R}_x = I - \bar{P}_x$

$\bar{P}_y$ and $\bar{R}_y = I - \bar{P}_y$.

From this define the interpolation spaces $\bar{V}(\bar{\pi}_x, \bar{M}, \bar{m}) \subseteq C^{(\bar{m})}([a,b])$ and $\bar{V}(\bar{\pi}_y, \bar{N}, \bar{n}) \subseteq C^{(\bar{n})}([c,d])$.

For $f \in C^{(m^*, n^*)}$, $m^* = \max\{m, \overline{m}\}$ and $n^* = \max\{n, \overline{n}\}$ define the discrete blending approximation to $P_x \oplus P_y[f]$ as

$$\overline{P_x \oplus P_y}[f] = \overline{P}_x P_y[f] + \overline{P}_y P_x[f] - P_x P_y[f], \tag{1.39}$$

where, for example

$$\overline{P}_x P_y[f](x, y) = \sum_{i=1}^{\overline{M}} \sum_{j=1}^{N} f^{(\overline{\alpha}(i), \beta(j))}(\overline{x}_i, y_j)\, \overline{\phi}_i(x)\, \psi_j(y). \tag{1.40}$$

In general, the discrete blending approximation $\overline{P_x \oplus P_y}[f]$ does not interpolate values of f and its derivatives. However, with the following restriction on the interpolation spaces, we will prove that $\overline{P_x \oplus P_y}[f]$ does indeed interpolate.

Definition 1.5: Let $V(\pi_x, M, m)$ and $\overline{V}(\overline{\pi}_x, \overline{M}, \overline{m})$ be interpolation vector spaces with interpolation functions α and $\overline{\alpha}$ respectively, then $V(\pi_x, M, m)$ is subordinate to $\overline{V}(\overline{\pi}_x, \overline{M}, \overline{m})$ if and only if $m \leqslant \overline{m}$ and for each i where $1 \leqslant i \leqslant M$ there exists an $\overline{i}$ such that $1 \leqslant \overline{i} \leqslant \overline{M}$, $x_i = \overline{x}_{\overline{i}}$ and $\alpha(i) = \overline{\alpha}(\overline{i})$.

We have the following generalization of a theorem due to Gordon and Hall [17].

Theorem 1.5. Given the interpolation spaces $V(\pi_x, M, m)$, $\overline{V}(\overline{\pi}_x, \overline{M}, m)$, $V(\pi_y, N, n)$ and $\overline{V}(\overline{\pi}_y, \overline{N}, n)$ where $V(\pi_x, M, m)$ is subordinate to $\overline{V}(\overline{\pi}_x, \overline{M}, m)$ and $V(\pi_y, N, n)$ is subordinate to $\overline{V}(\overline{\pi}_y, \overline{N}, n)$, if $f \in C^{(m, n)}$ then for $1 \leqslant i \leqslant M$, $1 \leqslant j \leqslant N$, $1 \leqslant \overline{i} \leqslant \overline{M}$

and $1 \leqslant \bar{j} \leqslant \bar{N}$

$$(1.41) \qquad (D^{(\bar{\alpha}(\bar{i}),\beta(j))}\overline{P_x \oplus P_y}[f])(\bar{x}_{\bar{i}}, y_j) = f^{(\bar{\alpha}(\bar{i}),\beta(j))}(\bar{x}_{\bar{i}}, y_j)$$

and

$$(1.42) \qquad (D^{(\alpha(i),\bar{\beta}(\bar{j}))}\overline{P_x \oplus P_y}[f])(x_i, \bar{y}_{\bar{j}}) = f^{(\alpha(i),\bar{\beta}(\bar{j}))}(x_i, \bar{y}_{\bar{j}}) .$$

Proof: By direct calculation. ∎

Even if our blending spaces are not subordinate to those used in the second level decomposition, $\overline{P_x \oplus P_y}[f]$ is still an approximation to $P_x \oplus P_y[f]$ and hence to f also. This is the conclusion of the following theorem.

Theorem 1.6. (Gordon and Hall [17]). Given the interpolation spaces $V(\pi_x, M, m)$, $\bar{V}(\bar{\pi}_x, \bar{M}, \bar{m})$, $V(\pi_y, N, n)$ and $\bar{V}(\bar{\pi}_y, \bar{N}, \bar{n})$, if $f \in C^{(m^*, n^*)}$ where $m^* = \max\{m, \bar{m}\}$ and $n^* = \max\{n, \bar{n}\}$ then

$$(1.43) \qquad (I - \overline{P_x \oplus P_y})[f] = \bar{R}_x[f] + \bar{R}_y[f] + R_x R_y[f] - \bar{R}_x R_y[f] - \bar{R}_y R_x[f] .$$

Proof:
$$I - \overline{P_x \oplus P_y} = I - \bar{P}_x P_y - \bar{P}_y P_x + P_x P_y$$

$$= I - P_x \oplus P_y + \bar{R}_x P_y + \bar{R}_y P_x$$

$$= R_x R_y + \bar{R}_x - \bar{R}_x R_y + \bar{R}_y - \bar{R}_y R_x .$$

∎

An examination of Theorem 1.6 and Theorem 1.4 yields the following generalization of a theorem by Gordon and Hall [17].

Theorem 1.7. Given the nonnegative integers $m_3 \leqslant \min\{m_2, \overline{m}_2\}$ and $n_3 \leqslant \min\{n_2, \overline{n}_2\}$, the interpolation spaces $V(\pi_x, M, m_1)$ and $\overline{V}(\overline{\pi}_x, \overline{M}, \overline{m}_1)$ which have error bounds for k, where $0 \leqslant k \leqslant m_3$, also the interpolation spaces $V(\pi_y, N, n_1)$ and $\overline{V}(\overline{\pi}_y, \overline{N}, \overline{n}_1)$ which have error bounds for ℓ where $0 \leqslant \ell \leqslant n_3$, and if $f \in C^{(m_2^*, n_2^*)}$ where $m_2^* = \max\{m_2, \overline{m}_2\}$ and $n_2^* = \max\{n_2, \overline{n}_2\}$ then

$$
\begin{aligned}
(1.44) \qquad |((I - \overline{P_x \oplus P_y})[f])^{(k,\ell)}(\hat{x}, \hat{y})| &\leqslant \overline{g}_x(k)\,||f^{(\overline{m}_2, \ell)}|| \\
&+ \overline{g}_y(\ell)\,||f^{(k, \overline{n}_2)}|| + g_x(k)\, g_y(\ell)\,||f^{(m_2, n_2)}|| \\
&+ \overline{g}_x(k)\, g_y(\ell)\,||f^{(\overline{m}_2, n_2)}|| + g_x(k)\, \overline{g}_y(\ell)\,||f^{(m_2, \overline{n}_2)}||,
\end{aligned}
$$

where $\hat{x}$ and $\hat{y}$ satisfy the conditions of Definition 1.4.

Proof: Similar to the proof of Theorem 1.4. ∎

If it is possible, by some procedure, to increase the accuracy of the spaces $\overline{V}(\overline{\pi}_x, \overline{M}, \overline{m}_1)$ and $\overline{V}(\overline{\pi}_y, \overline{N}, \overline{n}_1)$, we see from Theorem 1.7 that the total increase in accuracy for $\overline{P_x \oplus P_y}$ is limited by the term $g_x(k)\, g_y(\ell)\,||f^{(m_2, n_2)}||$. Therefore, it is not possible to increase the accuracy beyond that of the original blending approximation.

Example 1.4: If we take for each of our interpolation spaces the space of cubic splines defined in Example 1.2, then $V(\pi_x, M, 1) = S^2(\pi_x)$, $\overline{V}(\overline{\pi}_x, \overline{M}, 1) = S^2(\overline{\pi}_x)$, $V(\pi_y, N, 1) = S^2(\pi_y)$, $\overline{V}(\overline{\pi}_y, \overline{M}, 1) = S^2(\overline{\pi}_y)$, $m_2 = \overline{m}_2$, $n_2 = \overline{n}_2$, $m_3 = \min\{m_2, 3\}$, $n_3 = \min\{n_2, 3\}$ $1 \leq m_2, n_2 \leq 4$ and if $f \in C^{(m_2, n_2)}$ then

$$
\begin{aligned}
(1.45) \qquad ||((I - \overline{P_x \oplus P_y}([f]))^{(k,\ell)}|| \leq\ & \mathcal{E}_{m_2,k}||f^{(m_2,\ell)}||\overline{h}_x^{\,m_2-k} \\
& + \mathcal{E}_{n_2,\ell}||f^{(k,n_2)}||\overline{h}_y^{\,n_2-\ell} \\
& + \mathcal{E}_{m_2,k}\mathcal{E}_{n_2,\ell}||f^{(m_2,n_2)}||h_x^{\,m_2-k}h_y^{\,n_2-\ell} \\
& + \mathcal{E}_{m_2,k}\mathcal{E}_{n_2,\ell}||f^{(m_2,n_2)}||\overline{h}_x^{\,m_2-k}h_y^{\,n_2-\ell} \\
& + \mathcal{E}_{m_2,k}\mathcal{E}_{n_2,\ell}||f^{(m_2,n_2)}||h_x^{\,m_2-k}\overline{h}_y^{\,n_2-\ell} \quad .
\end{aligned}
$$

For ease of comparison, we will let $h = \max\{h_x, h_y\}$, $\overline{h} = \max\{\overline{h}_x, \overline{h}_y\}$, $m_2 = n_2 = 4$ and $k = \ell = 0$. Then the accuracy in (1.45) is limited to $O(h^8)$, and therefore we take $\overline{h} = h^2$ to preserve the overall accuracy of the scheme. This tells us how much the meshes $\overline{\pi}_x$ and $\overline{\pi}_y$ must be refined in (1.45) to obtain a scheme which is $O(h^8)$.

In comparison, for bicubic splines (tensor product (1.35)) the accuracy is $O(h^4)$. In the next section we will compute the dimension of discretized blending function spaces, and therefore will be able to show that the increase in accuracy is worth the extra labor of implementation.

Remark: In our notation $V(\pi_x, M, m)$, m can be any integer which satisfies $m^* \leq m \leq m^{**}$, where m^* is the smallest integer which allows enough continuity to perform our interpolation and m^{**} is the largest integer (if it exists) for which $V(\pi_x, M, m) \subseteq C^{(m^{**})}[a, b]$. In Example 1.2, for type 1 cubic spline interpolation, m can be equal to 1 or 2. In Example 1.3, for cubic Hermite splines, $m = 1$ is the only choice. Note that by increasing m, we restrict the choice of f which will satisfy our theorems. However, sometimes it is necessary to use a $m > m^*$, if, for example, we wish to satisfy Definition 1.5 or Theorem 1.5.

Section 2. Dimension of Discretized Blending Function Spaces

In applications other than interpolation, such as discrete least squares, Ritz-Galerkin methods and collocation, it is necessary to have a finite dimensional space on which to do the computation. Therefore, we introduce the following definition.

Definition 2.1: Let $V(\pi_x, M, m)$, $\overline{V}(\overline{\pi}_x, \overline{M}, \overline{m})$, $V(\pi_y, N, n)$ and $\overline{V}(\overline{\pi}_y, \overline{N}, \overline{n})$ be interpolation spaces (see Definition 1.3), then define the discretized blending function space $DBF(V(\pi_x, M, m), V(\pi_y, N, n); \overline{V}(\overline{\pi}_x, \overline{M}, \overline{m}), \overline{V}(\overline{\pi}_y, \overline{N}, \overline{n}))$ to be the image of $C^{(m^*, n^*)}([a,b]x[c,d])$ under the linear discretized blending function operator $\overline{P}_x P_y + \overline{P}_y P_x - P_x P_y$, (denoted by $\overline{P_x \oplus P_y}$), where $m^* = \max\{m, \overline{m}\}$ and $n^* = \max\{n, \overline{n}\}$.

From this notation, it is understood that the blending spaces are $V(\pi_x, M, m)$ and $V(\pi_y, N, n)$, and the spaces which give the second level approximations are $\overline{V}(\overline{\pi}_x, \overline{M}, \overline{m})$ and $\overline{V}(\overline{\pi}_y, \overline{N}, \overline{n})$.

When the interpolation spaces are understood, we will drop them from our notation and denote the space of discretized blending functions as DBF.

It is clear that DBF is indeed a vector space.

Finally, for this section, $\alpha, \overline{\alpha}, \beta$ and $\overline{\beta}$ will be the interpolation functions and $\{\phi_i\}_{i=1}^{M}$, $\{\overline{\phi}_i\}_{i=1}^{\overline{M}}$, $\{\psi_j\}_{j=1}^{N}$ and $\{\overline{\psi}_j\}_{j=1}^{\overline{N}}$ will be

the cardinal bases for $V(\pi_x, M, m)$, $\overline{V}(\overline{\pi}_x, \overline{M}, \overline{m})$, $V(\pi_y, N, n)$ and $\overline{V}(\overline{\pi}_y, \overline{N}, \overline{n})$, respectively.

We will need the following information about product space. If $V(\pi_x, M, m)$ and $V(\pi_y, N, n)$ are interpolation spaces, then their product is defined to be

$$(2.1) \qquad V(\pi_x, M, m) \otimes V(\pi_y, N, n) = \text{span}\,\{fg \mid f \in V(\pi_x, M, m) \text{ and } g \in V(\pi_y, N, n)\}.$$

An element $f \otimes g \in V(\pi_x, M, m) \otimes V(\pi_y, N, n)$ will be abbreviated by $fg = f \otimes g$, where $f \in V(\pi_x, M, m)$, $g \in V(\pi_y, N, n)$ and $fg(x, y) = f(x)\, g(y)$ for $(x, y) \in [a, b] \times [c, d]$.

If $\{u_i\}_{i=1}^{M}$ is any basis for $V(\pi_x, M, m)$ and $\{w_j\}_{j=1}^{N}$ is any basis for $V(\pi_y, N, n)$ then $\{u_i\, w_j\}_{\substack{1 \le i \le M \\ 1 \le j \le N}}$ is a basis for $V(\pi_x, M, m) \otimes V(\pi_y, N, n)$. Therefore,

$$(2.2) \qquad \begin{aligned} &\text{Dimension}\,(V(\pi_x, M, m) \otimes V(\pi_y, N, n)) \\ &\quad = \text{Dim}\,(V(\pi_x, M, m)) \cdot \text{Dim}\,(V(\pi_y, N, n)) \\ &\quad = M \cdot N\ . \end{aligned}$$

Given the interpolation spaces $V(\pi_x, M, m)$, $\overline{V}(\overline{\pi}_x, \overline{M}, \overline{m})$, $V(\pi_y, N, n)$ and $\overline{V}(\overline{\pi}_y, \overline{N}, \overline{n})$, and, for example, if $f \in V(\pi_x, M, m) \otimes V(\pi_y, N, n)$ then

$$(2.3) \qquad f = \sum_{i=1}^{M} \sum_{j=1}^{N} f^{(\alpha(i), \beta(j))}(x_i, y_j)\ \phi_i\, \psi_j\ ,$$

by the uniqueness of representation. Also

$$(2.4)\qquad (V(\pi_x, M, m) \cap \bar{V}(\bar{\pi}_x, \bar{M}, \bar{m})) \circledast V(\pi_y, N, n)$$

$$= (V(\pi_x, M, m) \circledast V(\pi_y, N, n))$$

$$\cap (\bar{V}(\bar{\pi}_x, \bar{M}, \bar{m}) \circledast V(\pi_y, N, n)),$$

and

$$(2.5)\qquad V(\pi_x, M, m) \circledast (V(\pi_y, N, n) \cap \bar{V}(\bar{\pi}_y, \bar{N}, \bar{n}))$$

$$= (V(\pi_x, M, m) \circledast V(\pi_y, N, n))$$

$$\cap (V(\pi_x, M, m) \circledast \bar{V}(\bar{\pi}_y, \bar{N}, \bar{n})).$$

We see that (2.4) is true, because the left hand side is clearly contained in the right hand side. Also, if f is an element of the right hand side of (2.4), then

$$(2.6)\qquad f = \sum_{j=1}^{N} \left\{ \sum_{i=1}^{M} f^{(\alpha(i),\beta(j))}(x_i, y_j)\, \phi_i \right\} \psi_j$$

$$= \sum_{j=1}^{N} \left\{ \sum_{i=1}^{\bar{M}} f^{(\bar{\alpha}(i),\beta(j))}(\bar{x}_i, y_j)\, \bar{\phi}_i \right\} \psi_j .$$

We will have proved our result if the bracketed expressions in (2.6) are contained in $V(\pi_x, M, m) \cap \bar{V}(\bar{\pi}_x, \bar{M}, \bar{m})$ for $1 \leqslant j \leqslant N$. Using the cardinality conditions, we have for each j satisfying $1 \leqslant j \leqslant N$

$$(2.7)\qquad f^{(0,\beta(j))}(\cdot, y_j) = \sum_{i=1}^{M} f^{(\alpha(i),\beta(j))}(x_i, y_j)\, \phi_i \in V(\pi_x, M, m)$$

$$= \sum_{i=1}^{\bar{M}} f^{(\bar{\alpha}(i),\beta(j))}(\bar{x}_i, y_j)\, \bar{\phi}_i \in \bar{V}(\bar{\pi}_x, \bar{M}, \bar{m}),$$

which proves (2.4). We note that (2.5) follows in a similar manner.

We now show that the following statement is valid for $m \leqslant \overline{m}$ or $n \leqslant \overline{n}$

$$(2.8)\qquad (\overline{V}(\overline{\pi}_x, \overline{M}, \overline{m}) \circledast V(\pi_y, N, n)) \bigcap (V(\pi_x, M, m) \circledast \overline{V}(\overline{\pi}_y, \overline{N}, \overline{n})) \subseteq V(\pi_x, M, m) \circledast V(\pi_y, N, n) .$$

To see that (2.8) is true, let f be an element of the left hand side. Then

$$(2.9)\qquad f = \sum_{j=1}^{N} \left\{ \sum_{i=1}^{\overline{M}} f^{(\overline{\alpha}(i), \beta(j))}(\overline{x}_i, y_j)\, \overline{\phi}_i \right\} \psi_j$$

$$= \sum_{i=1}^{M} \sum_{j=1}^{\overline{N}} f^{(\alpha(i), \overline{\beta}(j))}(x_i, \overline{y}_j)\, \phi_i \quad \overline{\psi}_j .$$

We will be done if it can be shown that the bracketed expression in (2.9) is in $V(\pi_x, M, m)$ for $1 \leqslant j \leqslant N$. Using $n \leqslant \overline{n}$ and the cardinality conditions for $1 \leqslant \hat{j} \leqslant N$, then

$$(2.10)\qquad f^{(0, \beta(\hat{j}))}(\cdot, y_{\hat{j}}) = \sum_{i=1}^{\overline{M}} f^{(\overline{\alpha}(i), \beta(\hat{j}))}(\overline{x}_i, y_{\hat{j}})\ \overline{\phi}_i$$

$$= \sum_{i=1}^{M} \left[\sum_{j=1}^{\overline{N}} f^{(\alpha(i), \overline{\beta}(j))}(x_i, \overline{y}_j)\, \overline{\psi}_j(\beta(\hat{j}))(y_{\hat{j}}) \right] \phi_i$$

$$\in V(\pi_x, M, m) .$$

Finally, from (1.39) and (1.40) it is clear that

$$(2.11)\qquad DBF \subseteq \operatorname{span}\{(\overline{V}(\overline{\pi}_x, \overline{M}, \overline{m}) \circledast V(\pi_y, N, n)) \cup (V\pi_x, M, m) \circledast \overline{V}(\overline{\pi}_y, \overline{N}, \overline{n})) \cup (V(\pi_x, M, m) \circledast V(\pi_y, N, n))\}.$$

Usually, however, DBF is a proper subset. The above yields the following theorem.

Theorem 2.1. Given the interpolation spaces $V(\pi_x, M, m)$, $\overline{V}(\overline{\pi}_x, \overline{M}, \overline{m})$, $V(\pi_y, N, n)$ and $\overline{V}(\overline{\pi}_y, \overline{N}, \overline{n})$, then

$$\begin{aligned}(2.12)\qquad \text{Dim (DBF)} \leqslant\ & \overline{M}N + M\overline{N} + MN \\ & - N \cdot \text{Dim}\,(\overline{V}(\overline{\pi}_x, \overline{M}, \overline{m}) \cap V(\pi_x, M, m)) \\ & - M \cdot \text{Dim}\,(\overline{V}(\overline{\pi}_y, \overline{N}, \overline{n}) \cap V(\pi_y, N, n)) .\end{aligned}$$

Proof: From [7, p. 468], we have for finite dimensional vector spaces V_1 and V_2

$$(2.13)\qquad \text{Dim (span}\{V_1 \cup V_2\}) = \text{Dim}\,(V_1) + \text{Dim}\,(V_2) - \text{Dim}(V_1 \cap V_2).$$

Therefore, from (2.11) we have

$$\begin{aligned}(2.14)\qquad \text{Dim (DBF)} \leqslant\ & \text{Dim}\,(\overline{V}(\overline{\pi}_x, \overline{M}, \overline{m}) \circledast V(\pi_y, N, n)) \\ & + \text{Dim}\,(V(\pi_x, M, m) \circledast \overline{V}(\overline{\pi}_y, \overline{N}, \overline{n})) \\ & + \text{Dim}\,(V(\pi_x, M, m) \circledast V(\pi_y, N, n)) \\ & - \text{Dim}((\overline{V}(\overline{\pi}_x, \overline{M}, \overline{m}) \circledast V(\pi_y, N, n)) \cap (V(\pi_x, M, m) \circledast \overline{V}(\overline{\pi}_y, \overline{N}, \overline{n}))) \\ & - \text{Dim}((\overline{V}(\overline{\pi}_x, \overline{M}, \overline{m}) \circledast V(\pi_y, N, n)) \cap (V(\pi_x, M, m) \circledast V(\pi_y, N, n))) \\ & - \text{Dim}((V(\pi_x, M, m) \circledast \overline{V}(\overline{\pi}_y, \overline{N}, \overline{n})) \cap (V(\pi_x, M, m) \circledast V(\pi_y, N, n))) \\ & + \text{Dim}((\overline{V}(\overline{\pi}_x, \overline{M}, \overline{m}) \circledast V(\pi_y, N, n)) \cap (V(\pi_x, M, m) \circledast \overline{V}(\overline{\pi}_y, \overline{N}, \overline{n})) \\ & \qquad \cap (V(\pi_x, M, m) \circledast V(\pi_y, N, n))) .\end{aligned}$$

The fourth and last terms together are non-positive, hence, can be dropped (if $m \leqslant \overline{m}$ or $n \leqslant \overline{n}$ then 2.8 would show that they are equal and the bound is best possible). Using (2.2), (2.4) and (2.5) we have our result.

■

For the proof of the next theorem, we introduce the following notation.

If $V(\pi_x, M, m)$ is subordinate to $\overline{V}(\overline{\pi}_x, \overline{M}, \overline{m})$, then from Definition 1.5, for each i satisfying $1 \leqslant i \leqslant M$, there corresponds a unique $\overline{i}$ such that $1 \leqslant \overline{i} \leqslant \overline{M}$, $\overline{x}_{\overline{i}} = x_i$ and $\alpha(i) = \overline{\alpha}(\overline{i})$. Thus, we define the index set IM to be the set which contains each of the $\overline{i}$ satisfying the above conditions. Finally, we define the index set $\overline{IM}$ to be the set of the remaining $\overline{M} - M$ integers.

Correspondingly, if $V(\pi_y, N, n)$ is subordinate to $(\overline{V}(\overline{\pi}_y, \overline{N}, \overline{n})$, then we define the index sets JN and $\overline{JN}$ in an analogous way. We will now prove a lemma which gives a lower bound on Dim (DBF), and eventually a basis for DBF.

Lemma 2.1. Given the interpolation spaces $V(\pi_x, M, m)$ $V(\pi_y, N, n)$, $\overline{V}(\overline{\pi}_x, \overline{M}, m)$ and $\overline{V}(\overline{\pi}_y, \overline{N}, n)$ such that $V(\pi_x, M, m)$ is subordinate to $\overline{V}(\overline{\pi}_x, \overline{M}, \overline{m})$ and $V(\pi_y, N, n)$ is subordinate to $\overline{V}(\overline{\pi}_y, \overline{N}, n)$, and if

$$Q_1 = \{\overline{\phi}_i\, \psi_j\}_{i \in \overline{IM},\ 1 \leqslant j \leqslant N'} \qquad Q_2 = \{\phi_i\, \overline{\psi}_j\}_{j \in \overline{JN},\ 1 \leqslant i \leqslant M'}$$

$$Q_3 = \{\overline{\phi}_i\, \psi_j + P_x\, \overline{P}_y[\overline{\phi}_i\, \psi_j] - P_x[\overline{\phi}_i\, \psi_j]\}_{j \in IM,\ 1 \leqslant j \leqslant N'}$$

$$Q_4 = \{\phi_i\, \overline{\psi}_j + \overline{P}_x\, P_y[\phi_i \overline{\psi}_j] - P_y[\phi_i\, \overline{\psi}_j]\}_{j \in JN,\ 1 < i < M'}$$

$$Q_5 = \{P_y[\overline{\phi}_i\, \overline{\psi}_j] + P_x[\overline{\phi}_i\, \overline{\psi}_j] - P_x\, P_y[\overline{\phi}_i\, \overline{\psi}_j]\}_{i \in IM,\ j \in JN'}$$

$$Q_6 = \{\overline{P}_x[\phi_i\, \psi_j] + \overline{P}_y[\phi_i\, \psi_j] - \phi_i\, \psi_j\}_{1 \leqslant i \leqslant M,\ 1 \leqslant j \leqslant N'}$$

$$T_1 = Q_1 \cup Q_2 \cup Q_3\,, \qquad T_2 = Q_1 \cup Q_2 \cup Q_4\,,$$

$$T_3 = Q_1 \cup Q_2 \cup Q_5\,, \qquad T_4 = Q_1 \cup Q_2 \cup Q_6\,,$$

then each of the sets T_1, T_2, T_3 and T_4 is a linearly independent subset of DBF.

<u>Proof</u>: The sets Q_3 through Q_6 are clearly contained in DBF, thus we need only show that Q_1 and Q_2 are subsets of DBF. Because $\overline{\phi}_i\, \psi_j \in C^{(m,n)}$, then for $\overline{i} \in \overline{IM}$ and $1 \leqslant j \leqslant N$ we have

$$\overline{P_x \oplus P_y}[\overline{\phi}_{\overline{i}}\, \psi_j] = \overline{\phi}_{\overline{i}}\, \psi_j + \overline{P}_y\, P_x[\overline{\phi}_{\overline{i}}\, \psi_j] - P_x[\overline{\phi}_{\overline{i}}\, \psi_j]\,. \tag{2.15}$$

We make the observation that for $1 \leqslant i \leqslant M$ and $\overline{i} \in \overline{IM}$

$$\overline{\phi}_{\overline{i}}^{(\alpha(i))}(x_i) = 0\,, \tag{2.16}$$

which implies that $P_x[\overline{\phi}_{\overline{i}}\, \psi_j] = 0$. Hence Q_1, and in a similar manner also Q_2 are contained in DBF. It then follows that T_1 through T_4 are subsets of DBF.

We now show the linear independence of the sets T_1 through T_4. For notational convenience and also to save space we will only prove the linear independence of T_3. If T_3 is not linearly independent, then there exists real numbers a_{ij} for $i \in \overline{IM}$, $1 \leqslant j \leqslant N$. also b_{ij} for $1 \leqslant i \leqslant M$, $j \in \overline{JN}$ and c_{ij} for $i \in IM$ and $j \in JM$ not all zero such that

$$(2.17) \qquad \sum_{i \in \overline{IM}} \sum_{1 \leqslant j \leqslant N} a_{ij} \bar{\phi}_i \psi_j + \sum_{1 \leqslant i \leqslant M} \sum_{j \in \overline{JN}} b_{ij} \phi_i \bar{\psi}_j$$

$$+ \sum_{i \in IM} \sum_{j \in JN} c_{ij} (P_y[\bar{\phi}_i \bar{\psi}_j] + P_x[\bar{\phi}_i \bar{\psi}_j] - P_x P_y[\bar{\phi}_i \bar{\psi}_j]) = 0 .$$

We will show that all c_{ij} are equal to zero. To accomplish this, apply the cardinality conditions for each $\hat{i} \in IM$ and $\hat{j} \in JN$ to (2.17). This means the application of the differential operator $D^{(\bar{\alpha}(\hat{i}), \bar{\beta}(\hat{j}))}$ to (2.17) and evaluating at the point $(\bar{x}_{\hat{i}}, \bar{y}_{\hat{j}})$. Because of the definition of IM and JN, we have for $i \in \overline{IM}$ and $j \in \overline{JN}$

$$(2.18) \qquad \bar{\phi}_i^{(\bar{\alpha}(\hat{i}))}(\bar{x}_{\hat{i}}) = 0 \quad \text{and} \quad \bar{\psi}_j^{(\bar{\beta}(\hat{j}))}(\bar{y}_{\hat{i}}) = 0 ,$$

which implies that there is no contribution from the first two sums in (2.17). Also, it can be shown that

$$(2.19) \qquad (D^{(\bar{\alpha}(\hat{i}), \bar{\beta}(\hat{j}))}(\sum_{i \in IM} \sum_{j \in JN} c_{ij} P_y[\bar{\phi}_i \bar{\psi}_j]))(\bar{x}_{\hat{i}}, \bar{y}_{\hat{j}}) = c_{\hat{i}\hat{j}} ,$$

$$(2.20) \qquad (D^{(\bar{\alpha}(\hat{i}), \bar{\beta}(\hat{j}))}(\sum_{i \in IM} \sum_{j \in JN} c_{ij} P_x[\bar{\phi}_i \bar{\psi}_j]))(\bar{x}_{\hat{i}}, \bar{y}_{\hat{j}}) = c_{\hat{i}\hat{j}}$$

and

$$(2.21)\qquad (D^{(\overline{\alpha}(\hat{i}),\,\overline{\beta}(\hat{j}))}\Big(\sum_{i\in IM}\sum_{j\in JN} c_{ij}\,P_x\,P_y\big[\overline{\phi}_i\,\overline{\psi}_j\big]\Big))(\overline{x}_{\hat{j}},\overline{y}_{\hat{j}}) = c_{\hat{i}\hat{j}}\,,$$

from which it follows, using (2.17), that $c_{\hat{i}\hat{j}} = 0$ for $\hat{i}\in IM$ and $\hat{j}\in JN$. To conserve space, we will only show that (2.21) is valid, and merely note that (2.19) and (2.20) are similar.

Because our blending spaces are subordinate, there corresponds an i_o where $1\leqslant i_o\leqslant M$, and j_o where $1\leqslant j_o\leqslant N$ such that $\overline{\alpha}(\hat{i}) = \alpha(i_o)$, $\overline{x}_{\hat{i}} = x_{i_o}$, $\overline{\beta}(\hat{j}) = \beta(j_o)$ and $\overline{y}_{\hat{j}} = y_{j_o}$. Then

$$(2.22)\qquad (D^{(\overline{\alpha}(\hat{i}),\,\overline{\beta}(\hat{j}))}\Big(\sum_{i\in IM}\sum_{j\in JN} c_{ij}\,P_x\,P_y\big[\overline{\phi}_i\,\overline{\psi}_j\big]\Big))\,(\overline{x}_{\hat{i}},\,\overline{y}_{\hat{j}})$$

$$= \sum_{i\in IM}\sum_{j\in JN} c_{ij}\sum_{s=1}^{M}\sum_{t=1}^{N}\overline{\phi}_i^{(\alpha(s))}(x_s)\,\overline{\psi}_j^{(\beta(t))}(y_t)\,\phi_s^{(\overline{\alpha}(\hat{i}))}(\overline{x}_{\hat{i}})\,\psi_t^{(\overline{\beta}(\hat{j}))}(\overline{y}_{\hat{j}})$$

$$= c_{\hat{i}\hat{j}}\,,$$

where we have used the cardinal interpolation conditions and the correspondence given above to obtain our result.

With all $c_{ij} = 0$, it is now a simple matter, using the appropriate cardinality conditions, to prove that all $a_{ij} = 0$ for $i\in\overline{IM}$, $1\leqslant j\leqslant N$, and all $b_{ij} = 0$ for $1\leqslant i\leqslant M$ and $j\in\overline{JN}$. Therefore, T_3 cannot be linearly dependent, and we have completed our proof.

■

For discretized blending function spaces which have subordinate interpolation spaces, we now have a lower bound for Dim (DBF).

Theorem 2.2. Given the interpolation spaces $V(\pi_x, M, m)$, $\overline{V}(\overline{\pi}_x, \overline{M}, m)$, $V(\pi_y, N, n)$ and $\overline{V}(\overline{\pi}_y, \overline{N}, n)$ such that $V(\pi_x, M, m)$ is subordinate to $\overline{V}(\overline{\pi}_x, \overline{M}, m)$ and $V(\pi_y, N, n)$ is subordinate to $\overline{V}(\overline{\pi}_y, \overline{N}, n)$, then

$$\text{(2.23a)} \qquad \text{Dim (DBF)} \geq \overline{M} \cdot N + M \cdot \overline{N} - M \cdot N$$

$$\text{(2.23b)} \qquad \text{Dim (DBF)} \leq \overline{M} N + M \overline{N} + M \cdot N$$
$$- N \cdot \text{Dim}\,(\overline{V}(\overline{\pi}_x, \overline{M}, m) \cap V(\pi_x, M, m))$$
$$- M \cdot \text{Dim}\,(\overline{V}(\overline{\pi}_y, \overline{N}, n) \cap V(\pi_y, N, n)) .$$

Proof: The upper bound follows from Theorem 2.1 and the lower bound from Lemma 2.1.

■

We would like to know the exact dimension of DBF, and have a basis for the space. Examination of Theorem 2.2 shows the assumption needed to obtain our result.

Theorem 2.3. Given the interpolation spaces $V(\pi_x, M, m)$, $\overline{V}(\overline{\pi}_x, \overline{M}, m)$, $V(\pi_y, N, n)$ and $\overline{V}(\overline{\pi}_y, \overline{N}, n)$ such that $V(\pi_x, M, m)$ is subordinate to $\overline{V}(\overline{\pi}_x, \overline{M}, m)$, $V(\pi_y, N, n)$ is subordinate to $\overline{V}(\overline{\pi}_y, \overline{N}, n)$, $V(\pi_x, M, m) \subseteq \overline{V}(\overline{\pi}_x, \overline{M}, m)$, $V(\pi_y, N, n) \subseteq \overline{V}(\overline{\pi}_y, \overline{N}, n)$, and if

$$S_1 = \{\overline{\phi}_i \psi_j\}_{i \in \overline{IM},\ 1 \leq j \leq N'} \qquad S_2 = \{\phi_i \overline{\psi}_j\}_{j \in \overline{JN},\ 1 \leq i \leq M},$$

$$S_3 = \{\overline{\phi}_i\, \psi_j\}_{i\in IM,\ 1\leqslant j\leqslant N}\,, \quad S_4 = \{\phi_i\, \overline{\psi}_j\}_{j\in JN,\ 1\leqslant i\leqslant M}\,,$$

$$S_5 = \{P_y[\overline{\phi}_i\, \overline{\psi}_j] + P_x[\overline{\phi}_i\, \overline{\psi}_j] - P_x\, P_y[\phi_i\, \psi_j]\}_{i\in IM,\ j\in JN}\,,$$

$$S_6 = \{\phi_i\, \psi_j\}_{1\leqslant i\leqslant M,\ 1\leqslant j\leqslant N}\,,$$

$$T_1 = S_1 \cup S_2 \cup S_3\,, \qquad T_2 = S_1 \cup S_2 \cup S_4\,,$$

$$T_3 = S_1 \cup S_2 \cup S_5\,, \qquad T_4 = S_1 \cup S_2 \cup S_6\,,$$

then

$$\text{Dim (DBF)} = \overline{M}\cdot N + M\cdot \overline{N} - M\cdot N \tag{2.24}$$

and each of the sets T_1, T_2, T_3 and T_4 forms a basis for DBF.

Proof: From Theorem 2.2, we have that (2.24) is true because of the containment of our blending spaces.

Also, because of the uniqueness of representation, we make the observation that

$$\overline{P}_y[\overline{\phi}_i\, \psi_j] = \overline{\phi}_i\, \psi_j \quad \text{for} \quad 1\leqslant i\leqslant \overline{M},\ 1\leqslant j\leqslant N\,, \tag{2.25a}$$

$$\overline{P}_x[\phi_i\, \overline{\psi}_j] = \phi_i\, \overline{\psi}_j \quad \text{for} \quad 1\leqslant i\leqslant M,\ 1\leqslant j\leqslant \overline{N}\,, \tag{2.25b}$$

and

$$\overline{P}_x[\phi_i\, \psi_j] = \overline{P}_y[\phi_i\, \psi_j] = \phi_i\, \psi_j \quad \text{for} \quad 1\leqslant i\leqslant M,\ 1\leqslant j\leqslant N. \tag{2.25c}$$

Using (2.25a), (2.25b) and (2.25c), then in Lemma 2.1, Q_3 through Q_6 reduce to S_3 through S_6 respectively. Thus from Lemma 2.1, it follows that T_1 through T_4 each forms a basis for DBF, and we have proved our result.

■

Remark: S_5 is not as complicated as it appears. Because our blending interpolation spaces are subordinate, for each $i \in IM$ and $j \in JN$ there exists an i_o where $1 \leqslant i_o \leqslant M$ and j_o where $1 \leqslant j_o \leqslant N$ such that

$$S_5 = \{\bar{\phi}_i \psi_{j_o} + \phi_{i_o} \bar{\psi}_j - \phi_{i_o} \psi_{j_o}\} . \tag{2.26}$$

Often in practice, it is not convenient to work with the cardinal basis for DBF. If one is willing to work with a spanning set rather than a basis, the following theorem is useful.

Theorem 2.4. Given the interpolation spaces $V(\pi_x, M, m)$, $\bar{V}(\bar{\pi}_x, \bar{M}, m)$, $V(\pi_y, N, n)$ and $\bar{V}(\bar{\pi}_y, \bar{N}, n)$ such that $V(\pi_x, M, m)$ is subordinate to $\bar{V}(\bar{\pi}_x, \bar{M}, m)$, $V(\pi_y, N, n)$ is subordinate to $\bar{V}(\bar{\pi}_y, \bar{N}, n)$, $V(\pi_x, M, m) \subseteq \bar{V}(\bar{\pi}_x, \bar{M}, m)$, $V(\pi_y, N, n) \subseteq \bar{V}(\bar{\pi}_y, \bar{N}, n)$ and if $\{A_i\}_{i=1}^{M}$, $\{\bar{A}_i\}_{i=1}^{\bar{M}}$, $\{B_j\}_{j=1}^{N}$ and $\{\bar{B}\}_{j=1}^{\bar{N}}$ are any bases for $V(\pi_x, M, m)$, $\bar{V}(\bar{\pi}_x, \bar{M}, m)$, $V(\pi_y, N, n)$ and $\bar{V}(\bar{\pi}_y, \bar{N}, n)$, respectively, then

$$T_5 = \{A_i \bar{B}_j\}_{1 \leqslant i \leqslant M,\ 1 \leqslant j \leqslant \bar{N}} \cup \{\bar{A}_i B_j\}_{1 \leqslant i \leqslant \bar{M},\ 1 \leqslant j \leqslant N} \tag{2.27}$$

is a spanning set for DBF.

Proof: We first show that $T_5 \subseteq$ DBF. For $1 \leqslant i \leqslant M$ and $1 \leqslant j \leqslant \bar{N}$ we have $A_i \bar{B}_j \in C^{(m,n)}$ and

$$(2.28)\qquad \overline{P_x \oplus P_y}\left[A_i\ \overline{B}_j\right] = (\overline{P}_x P_y + \overline{P}_y P_x - P_x P_y)\left[A_i\ \overline{B}_j\right]$$

$$= \overline{P}_x P_y\left[A_i\ \overline{B}_j\right] + A_i\ \overline{B}_j - P_y\left[A_i\ \overline{B}_j\right]$$

$$= A_i\ \overline{B}_j\ ,$$

since $\overline{P}_x\left[A_i\ \overline{B}_j\right] = A_i\ \overline{B}_j$, which implies $A_i\ \overline{B}_j \in$ DBF. In a similar way, for $1 \leqslant i \leqslant \overline{M}$ and $1 \leqslant j \leqslant N$ we have $\overline{A}_i\ B_j \in$ DBF, which implies that $T_5 \subseteq$ DBF.

Because $V(\pi_x, M, m) \subseteq \overline{V}(\overline{\pi}_x, \overline{M}, m)$, it is clear from (2.1), which is the definition of a product space, that

$$(2.29)\qquad V(\pi_x, M, m) \otimes V(\pi_y, N, n) \subseteq \overline{V}(\overline{\pi}_x, \overline{M}, m) \otimes V(\pi_y, N, n)\ .$$

It follows from (2.11), (2.29) and the statement following (2.1) that T_5 is indeed a spanning set for DBF.

Remark: For certain spaces, if $V(\pi_x, M, m) \subseteq \overline{V}(\overline{\pi}_x, \overline{M}, m)$, then this implies that $V(\pi_x, M, m)$ is subordinate to $\overline{V}(\overline{\pi}_x, \overline{M}, m)$. For example, consider the cubic spline spaces of Example 1.2, if $S^2(\pi_x) \subseteq S^2(\overline{\pi}_x)$, then $S^2(\pi_x)$ is subordinate to $S^2(\overline{\pi}_x)$. This is clear, because if $x_i \neq a$ and $\neq b$ is a knot of the mesh π_x, then there exists a cubic spline in $S^2(\pi_x)$ which has a jump discontinuity in the third derivative at x_i. However, all cubic splines in $S^2(\overline{\pi}_x)$ are cubic polynomials between the knots of $\overline{\pi}_x$, from which it follows that x_i must be a knot of $\overline{\pi}_x$.

If we consider polynomial spaces of Example 1.1, then even if $V(\pi_x, M, 0) \subseteq \overline{V}(\overline{\pi}_x, \overline{M}, 0)$, it does not necessarily follow that

$V(\pi_x, M, 0)$ is subordinate to $\overline{V}(\overline{\pi}_x, \overline{M}, 0)$, because π_x and $\overline{\pi}_x$ could be different.

Finally, let $V(\pi_x, M, 0)$ be the space of piecewise linear functions and $\overline{V}(\pi_x, M, 0)$ be the space of polynomials of degree M-1, then even though $V(\pi_x, M, 0)$ is subordinate to $\overline{V}(\pi_x, M, 0)$, it is not the case that $V(\pi_x, M, 0) \subseteq \overline{V}(\pi_x, M, 0)$.

Example: In Theorem 2.1 and 2.2, even if the intersection of our interpolation spaces contains only the zero vector, we will show by the following example that the bounds on the dimension are still sharp. Let

(2.30a) $\pi_x = x_1 \in [1, 2]$, $\overline{\pi}_x = \overline{x}_1 \in [1, 2]$,

(2.30b) $V(\pi_x, 1, 0) = \text{span}\{1\}$ with cardinal basis $\{1\}$,

(2.30c) $\overline{V}(\overline{\pi}_x, 1, 0) = \text{span}\{x\}$ with cardinal basis $\{x/\overline{x}_1\}$,

(2.30d) $\pi_x = y_1 \in [1, 2]$, $\overline{\pi}_y = \overline{y}_1 \in [1, 2]$,

(2.30e) $V(\pi_y, 1, 0) = \text{span}\{1\}$ with cardinal basis $\{1\}$,

(2.30f) $\overline{V}(\overline{\pi}_y, 1, 0) = \text{span}\{y\}$ with cardinal basis $\{y/\overline{y}_i\}$,

and if $x_1 \neq \overline{x}_1$, $y_1 \neq \overline{y}_1$ and $f \in C([1, 2] \times [1, 2])$

then

$$\overline{P_x \oplus P_y}[f](x, y) = f(\overline{x}_1, y_1) \cdot (x/\overline{x}_1) \cdot 1 + f(x_1, \overline{y}_1) \cdot 1 \cdot (y/\overline{y}_1) - f(x_1, y_1) \cdot 1 \cdot 1 \tag{2.31}$$

and

(2.32) $\quad DBF = \{a\cdot x + b\cdot y + c\cdot 1 \mid a, b \text{ and } c \text{ are real}\}$,

therefore

(2.33) $\quad Dim\,(DBF) = 3$,

which is the upper bound of Theorem 2.1 and Theorem 2.2.

Using (2.30a) through (2.30f), and if $x_1 \neq x_1$ but $y_1 = y_1$, then from (2.31) we have

(2.34) $\quad DBF = \{a\cdot x + b\cdot (y/y_1 - 1) \mid a \text{ and } b \text{ are real}\}$,

therefore

(2.35) $\quad Dim\,(DBF) = 2$.

Finally, if (2.30a) through (2.30f) hold, where $V(\pi_x, 1, 0)$ and $V(\pi_y, 1, 0)$ are subordinate to $\overline{V}(\overline{\pi}_x, 1, 0)$ and $\overline{V}(\overline{\pi}_y, 1, 0)$, respectively, (i.e., $\overline{x}_1 = x_1$ and $y_1 = \overline{y}_1$), then

(2.36) $\quad DBF = \{a(x/x_1 + y/y_1 - 1) \mid a \text{ is real}\}$,

and

(2.37) $\quad Dim\,(DBF) = 1$,

which is the lower bound of Theorem 2.2.

<u>Remark:</u> If $V(\pi_x, M, m)$, $\overline{V}(\overline{\pi}_x, \overline{M}, m)$, $V(\pi_y, N, n)$ and $\overline{V}(\overline{\pi}_y, \overline{N}, n)$ are the spaces of cubic splines $S^2(\pi_x)$, $S^2(\overline{\pi}_x)$, $S^2(\pi_y)$ and $S^2(\overline{\pi}_y)$, respectively, then a convenient basis with which to work is the basis of "B-splines" (see [29, p. 73]), which have

support on at most four consecutive intervals and are non-negative. The B-spline basis is also relatively easy to construct when compared to the cardinal basis. In this case it is often advantageous to work with a spanning set of bicubic B-splines defined in Theorem 2.4, and use methods for overdetermined systems to solve the resulting equations.

Section 3. Natural Cubic Blending

Often, information about the normal derivatives around the boundary does not exist, or it is not known to sufficient accuracy to be compatible with an eighth order method. In the case where some deterioration of accuracy is acceptable near the boundary of our domain, we can use natural cubic blending to obtain a method which is $0(h^8)$ in the interior.

The natural cubic spline basis $\{A_i\}_{i=0}^{M+1}$ for $S^2(\pi_x) \subseteq C^{(2)}[a,b]$ on the mesh $\pi_x: a = x_1 < x_2 < \cdots < x_M = b$ satisfies for $1 \leqslant i, j \leqslant M$ the conditions

$$A_i(x_j) = \delta_{ij}, \quad A_i''(x_1) = A_i''(x_M) = 0 , \tag{3.1a}$$

$$A_0(x_j) = 0 , \quad A_0''(x_1) = 1, \quad A_0''(x_M) = 0 \tag{3.1b}$$

$$A_{M+1}(x_j) = 0, \quad A_{M+1}''(x_1) = 0, \quad A_{M+1}''(x_M) = 1 . \tag{3.1c}$$

The natural cubic spline interpolant to a function $f \in C[a,b]$ is defined to be

$$s(x) = \sum_{i=1}^{M} f(x_i)\, A_i(x) , \tag{3.2}$$

where it is clear that $s''(x_1) = s''(x_M) = 0$.

The following theorem, which is crucial to the proof of Theorem 3.2, shows the behavior of the natural cubic spline interpolant of one variable, see Hall [19].

Theorem 3.1. (Hall [19]). Let s be the natural cubic spline interpolant to $f \in C^m[a,b]$, $m = 2$, 3 or 4, then for $x \in [x_i, x_{i+1}]$, $1 \leq i \leq M-1$

$$(3.3) \qquad |(s-f)^{(k)}(x)| \leq ||f^{(m)}|| \{ \mathcal{E}_{mk} h_x^{m-k} + K_m h_x^{m-1} \alpha_k \Delta_i \} + 1/2 \, R \, h_x \alpha_k \Delta_i ,$$

$$(3.4) \qquad R = \max \{ |f''(a)|, |f''(b)| \}$$

for $0 \leq k \leq 2$, where the mesh size h_x is defined to be

$$(3.5) \qquad h_x = \max_{1 \leq i \leq M-1} \Delta x_i, \ \Delta x_i = x_{i+1} - x_i, \ \Delta_i = \{ 2^{1-i} + 2^{1-M+i} \},$$

and the constants $\mathcal{E}_{m,k}$, K_m and α_k are given in the following tables.

Table 3

$\mathcal{E}_{mk}$	k = 0	k = 1	k = 2
m = 2	9/8	4	10
m = 3	71/216	31/27	5
m = 4	5/384	$(9+\sqrt{3})/216$	5

Table 4

α_k	k = 0	k = 1	k = 2
	$\Delta x_i/4$	1	$6/\Delta x_i$

Table 5

K_m	m = 4	m = 3	m = 2
	7/24	1	5/2

Table 6

β_ℓ	$\ell = 0$	$\ell = 1$	$\ell = 2$
	$\Delta y_j/4$	1	$6/\Delta y_j$

We now generalize this theorem to two variables. For a mesh on the y - axis $\pi_y: c = y_1 < y_2 < \cdots < y_N = d$, let $S^2(\pi_y) \subseteq C^{(2)}[c, d]$ be the space of cubic splines in the variable y. The natural cubic spline basis $\{B_j\}_{j=0}^{N+1}$ for $S^2(\pi_y)$ satisfies conditions similar to (3.1a) - (3.1c). Define the projection operators P_x and P_y by

$$(3.6) \qquad (P_x[f])(x, y) = \sum_{i=1}^{M} f(x_i, y)\, A_i(x)$$

and

$$(3.7) \qquad (P_y[f])(x, y) = \sum_{j=1}^{N} f(x, y_j)\, B_j(y) \ ,$$

for each $f \in C([a, b] \times [c, d])$. The natural cubic spline blended interpolant to f is defined to be

$$(3.8) \qquad N B[f] = P_x[f] + P_y[f] - P_x P_y[f] \ .$$

Define $h_y = \max_{1 \leq j \leq N-1} \Delta y_j$, where $\Delta y_j = y_{j+1} - y_j$.

We have the following theorem.

Theorem 3.2: Let NB be the natural cubic spline blended interpolant to $f \in C^{(m,n)}([a,b]x[c,d])$. If $x \in [x_i, x_{i+1}]$, $y \in [y_j, y_{j+1}]$, $1 \leq i \leq M-1$, $1 \leq j \leq N-1$, $2 \leq m, n \leq 4$ and $0 \leq k, \ell \leq 2$, then

$$(3.9) \quad |(((I-NB)[f])^{(k,\ell)})(x,y)|$$

$$\leq ||f^{(m,n)}|| \{\mathcal{E}_{mk} h_x^{m-k} + K_m h_x^{m-1} \alpha_k \Delta_i\} \{\mathcal{E}_{n\ell} h_y^{n-\ell}$$

$$+ K_n h_y^{n-1} \beta_\ell \bar{\Delta}_j\} + 1/2 ||f^{(m,2)}|| \{\mathcal{E}_{mk} h_x^{m-k}$$

$$+ K_m h_x^{m-1} \alpha_k \Delta_i\} \bar{\Delta}_j \beta_\ell h_y + 1/2 ||f^{(2,n)}|| \{\mathcal{E}_{n\ell} h_y^{n-\ell}$$

$$+ K_n h_y^{n-1} \beta_\ell \bar{\Delta}_j\} \Delta_i \alpha_k h_x + 1/4 ||f^{(2,2)}|| \Delta_i \bar{\Delta}_j \alpha_k \beta_\ell h_x h_y,$$

where $\Delta_i = \{2^{1-i} + 2^{1-M+i}\}$, $\bar{\Delta}_j = \{2^{1-j} + 2^{1-N+j}\}$,

$\mathcal{E}_{mk}$ and $\mathcal{E}_{n\ell}$ are given in Table 3, K_m and K_n are given in Table 5, α_k is given in Table 4, β_ℓ is given in Table 6 and h_x and h_y are the mesh sizes of π_x and π_y, respectively.

Proof: From Section 1 we have $I - NB = R_x R_y$, where $I - P_x = R_x$ and $I - P_y = R_y$. Also from (1.17b) and (1.17c) we have $D^{(0,\ell)} R_x = R_x D^{(0,\ell)}$ and $D^{(k,0)} R_y = R_y D^{(k,0)}$. Define $g = D^{(0,\ell)} R_y [f]$. Then, for fixed x and y,

$$(3.10)\qquad |(((I - NB)[f])^{(k,\ell)})(x,y)| = |((R_x R_y[f])^{(k,\ell)})(x,y)|$$

$$= |(D^{(k,0)} R_x D^{(0,\ell)} R_y[f])(x,y)|$$

$$= |(D^{(k,0)} R_x g)(x,y)|$$

$$= |E(x,y)| .$$

For each fixed y, we consider $P_x[g](\cdot, y)$, which is the natural cubic spline interpolant to the univariable function $g(\cdot, y)$, where $g(\cdot, y) \in C^m[a,b]$. From Theorem 3.1 we have

$$(3.11)\qquad |E(x,y)| \leqslant ||g^{(m,0)}(\cdot, y)|| \{\mathcal{E}_{mk} h_x^{m-k} + K_m h_x^{m-1} \alpha_k \Delta_i\}$$

$$+ 1/2\, R_1(y)\, \Delta_i\, \alpha_k\, h_x ,$$

where

$$(3.12)\qquad R_1(y) = \max \{ |g^{(2,0)}(a,y)|,\ |g^{(2,0)}(b,y)| \} .$$

From the definition of g, we have for $0 \leqslant t \leqslant m$,

$$(3.13)\qquad g^{(t,0)} = D^{(t,0)}(D^{(0,\ell)} R_y[f])$$

$$= D^{(0,\ell)} R_y[f^{(t,0)}] .$$

Therefore, applying Theorem 3.1 again, for any fixed ξ such that $a \leqslant \xi \leqslant b$, we have

$$(3.14)\qquad |g^{(t,0)}(\xi, y)| \leqslant ||f^{(t,n)}(\xi, \cdot)|| \{\mathcal{E}_{n\ell} h_y^{n-\ell} + K_n h_y^{n-1} \beta_\ell \bar{\Delta}_j\}$$

$$+ 1/2\, R_2(\xi)\, \bar{\Delta}_j\, \beta_\ell\, h_y ,$$

where

$$(3.15)\qquad R_2(\xi) = \max\{|f^{(t,2)}(\xi, c)|, |f^{(t,2)}(\xi, d)|\} \; .$$

$$\leqslant ||f^{(t,2)}|| \; .$$

Taking the supremum over all ξ on the right of inequality (3.14) we have

$$(3.16)\qquad |g^{(t,0)}(\xi, y)| \leqslant ||f^{(t,n)}||\{\mathcal{E}_{n\ell}\, h_y^{n-\ell} + K_n\, h_y^{n-1}\, \beta_\ell\, \bar{\Delta}_j$$

$$+ 1/2\; ||f^{(t,2)}||\, \bar{\Delta}_j\, \beta_\ell\, h_y \; .$$

Substituting (3.16), for $t = m$ and $t = 2$, into (3.11) and (3.12) respectively, our conclusion follows after regrouping.

■

We will now make a series of remarks about Theorem 3.2.

Remark 3.1: The error still goes to zero even if only one of our meshes, π_x or π_y, is refined.

Remark 3.2: Because of the exponential decay of the terms Δ_i and $\bar{\Delta}_j$, if $a < a_1 < b_1 < b$ and $c < c_1 < d_1 < d$, then there exists a mesh fine enough so that the convergence on the subrectangle $[a_1, b_1] \times [c_1, d_1]$ is $O(h^{m+n-\ell-k})$, where $h = \max\{h_x, h_y\}$.

Remark 3.3: Moreover, because the exponential decay of the terms Δ_i and $\bar{\Delta}_j$ tend toward zero faster than the term $1/h^2 = (N-1)^2$

tends toward infinity, we have the area of higher order convergence increasing as we refine our meshes π_x and π_y.

Remark 3.4: Explicitly, the rate of increase is given by

$$a_1 - a = b - b_1 = c_1 - c = d - d_1 = (K \log (N-1))/(N-1) \longrightarrow 0$$

as $N \longrightarrow \infty$, where K is chosen with consideration of the bound on the derivatives of f.

Remark 3.5: For purposes of illustration, we will consider the special case of $k = \ell = 0$, $a = c = 0$, $b = d = 1$, $M = N$ and $m = n = 4$. Therefore, both α_0 and β_0 have a factor of h, and because of the exponential decay of the terms Δ_i and $\overline{\Delta}_j$ as we move away from the boundary, we have the following situation for our error, illustrated in Figure 1.

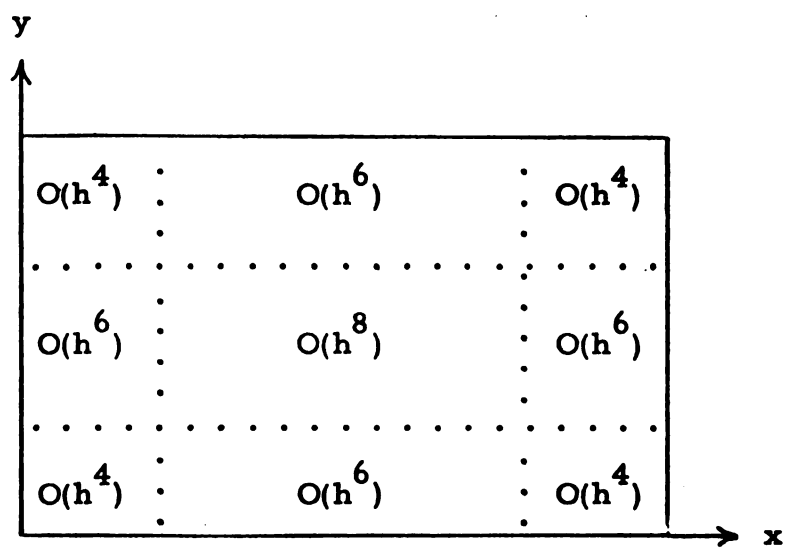

Figure 1

Section 4. Exponential Decay of Natural Cubic Cardinal Splines

Given a mesh $\pi_x: a = x_1 < x_2 < \cdots < x_M = b$ of M knots, the natural cardinal splines $A_i(x) \in S^2(\pi_x)$ are uniquely determined by the $M + 2$ conditions given in (3.1a) to (3.1c).

We will now prove the exponential decay of the natural cubic cardinal splines by a series of lemmas and theorems. Much of what follows parallels the results of Birkhoff and De Boor [2], for cardinal splines with first derivative end conditions, and we will quote theorems from their paper which are also valid for natural splines. Finally, we will use the abbreviation M. V. T. for the Mean Value Theorem and I. V. T. for the Intermediate Value Theorem.

Lemma 4.1. (Birkhoff and De Boor [2]). If $p(x)$ is a cubic polynomial which vanishes at 0 and $h \neq 0$ then

$$p'(h) = -2p'(0) - h\,p''(0)/2 \tag{4.1a}$$

$$p''(h)/2 = -3\,p'(0)/h - p''(0)\,. \tag{4.1b}$$

Corollary 4.1. For $1 \leqslant i \leqslant M$, the natural cubic cardinal splines satisfy

$$A_i'(x_{j+1}) = -2A_i'(x_j) - \Delta x_j\, A_i''(x_j)/2 \tag{4.2a}$$

$$A_i''(x_{j+1})/2 = -3A_i'(x_j)/\Delta x_j - A_i''(x_j)\,, \tag{4.2b}$$

where $1 \leqslant j < i-1$, and

$$A_i'(x_{j-1}) = -2\,A_i'(x_j) + \Delta x_{j-1}\,A_i''(x_j)/2 \tag{4.3a}$$

$$A_i''(x_{j-1})/2 = 3\,A_i'(x_j)/\Delta x_{j-1} - A_i''(x_j)\ , \tag{4.3b}$$

where $i+1 < j \leqslant M$.

Define the function $\operatorname{sgn}(\cdot)$ of a real variable by

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0\ . \end{cases} \tag{4.4}$$

Lemma 4.2. (Birkhoff and De Boor [2]). Let $s(x)$ be any cubic spline function with knots at the x_j, which satisfy for some i, $s(x_{i-1}) = s(x_{i+1}) = 0$, $s(x_i) > 0$, $s'(x_{i-1})\,s''(x_{i-1}) \geqslant 0$, $s'(x_{i+1})\,s''(x_{i+1}) \leqslant 0$. Then $s'(x_{i-1}) \geqslant 0$, $s''(x_{i-1}) \geqslant 0$, $s'(x_{i+1}) \leqslant 0$, $s''(x_{i+1}) \geqslant 0$, $s''(x_i) \leqslant 0$ and $s(x) \geqslant 0$ on $[x_{i-1}, x_{i+1}]$.

Lemma 4.3. For $2 \leqslant i \leqslant M-1$, $A_i(x)$ satisfies Lemma 4.2 on the interval $[x_{i-1}, x_{i+1}]$.

Proof: From Corollary 4.1 and the fact that $A_i''(x_1) = A_i''(x_M) = 0$. ∎

Lemma 4.4. For $1 \leqslant i \leqslant M$ the natural cubic cardinal splines satisfy

$$\operatorname{sgn}(A_i'(x_j)) = \operatorname{sgn}(A_i''(x_j)) \neq 0 \tag{4.5a}$$

(4.5b) $\operatorname{sgn}(A_i'(x_j)) = -\operatorname{sgn}(A_i'(x_{j-1})) \neq 0$,

where $2 \leqslant j \leqslant i-1$, and for $i+1 \leqslant j \leqslant M-1$

(4.6a) $\operatorname{sgn}(A_i'(x_j)) = -\operatorname{sgn}(A_i''(x_j)) \neq 0$

(4.6b) $\operatorname{sgn}(A_i'(x_j)) = -\operatorname{sgn}(A_i'(x_{j+1})) \neq 0$.

Proof: Case 1: $2 \leqslant i \leqslant M-1$.

We will first prove that $A_i'(x_1)$ and $A_i'(x_N)$ are both non-zero. If this is not the case, say $A_i'(x_1) = 0$, then inductively $A_i(x) \equiv 0$ on $[x_1, x_{i-1}]$. Therefore, on $[x_{i-1}, x_i]$, $A(x_{i-1}) = A'(x_{i-1}) = A''(x_{i-1}) = 0$, $A_i(x_i) = 1$ and from Lemmas 4.2 and 4.3 we also have $A_i''(x_i) < 0$. Several applications of the M. V. T. gives $\xi \in (x_{i-1}, x_i)$ such that $A_i''(\xi) > 0$ and the I. V. T. gives another distinct zero for A_i''. This implies that A_i is linear on $[x_{i-1}, x_i]$ because $A_i'' \equiv 0$ on $[x_{i-1}, x_i]$. But this is a contradiction to the fact that A_i must satisfy the conditions $A_i(x_{i-1}) = A_i'(x_{i-1}) = 0$ and $A_i(x_i) = 1$. An identical argument shows that $A_i'(x_M) \neq 0$. The result now follows inductively from Corollary 4.1.

Case 2: $i = 1, M$.

For $i = M$, if $A_M'(x_1) = 0$ then $A_M(x) \equiv 0$ on $[x_1, x_{M-1}]$. We then know that A_M'' has zeros at x_{M-1} and x_M, showing that A_M cannot satisfy all of the required conditions, therefore $A_M'(x_1) \neq 0$. An identical argument shows that $A_1'(x_M) \neq 0$.

The result now follows inductively from Corollary 4.1.

■

Lemma 4.5. If $1 \leqslant j \leqslant i-2$ and $x \in [x_j, x_{j+1}]$, then $A_i'(x_{j+1})\, A_i(x) \leqslant 0$ and if $i+1 \leqslant j \leqslant M-1$ for $x \in [x_j, x_{j+1}]$, then $A_i'(x_j)\, A_i(x) \geqslant 0$.

Proof: From Lemma 4.4, $\operatorname{sgn}(A_i'(x_j)) = -\operatorname{sgn}(A_i'(x_{j+1})) \neq 0$. It is now clear that if the conclusion of the lemma is not true, then A_i will have four distinct zeros in $[x_j, x_{j+1}]$, which is false. The other case is proved in a similar way.

■

Lemma 4.6. For $1 \leqslant i \leqslant M$ we have $A_i'(x_{i-1}) > 0$ and $A_i'(x_{i+1}) < 0$.

Proof: Case 1: $2 \leqslant i \leqslant M-1$.

The result follows from Lemmas 4.2 through 4.4.

Case 2: $i = 1, M$.

For $i = M$, $A_M(x_M) = 1$, $A_M''(x_M) = 0$. If $A_M'(x_{M-1}) < 0$, we have from Lemma 4.4 that $A_M''(x_{M-1}) < 0$. Two applications of the M. V. T. and one of the I. V. T. yields another zero for A_M'' which cannot happen if A_M is to satisfy the required conditions at x_{M-1} and x_M. The case where $i = 1$ follows in the same manner.

■

Lemma 4.7. On $[x_1, x_2]$, $\max |A_1(x)| = 1$, and on $[x_{M-1}, x_M]$ $\max |A_M(x)| = 1$.

Proof: On $[x_{M-1}, x_M]$, direct calculation yields a real number a such that $0 > a > -1/(2\Delta x_{M-1}^3)$ and

$$A_M(x) = a(x-x_M)^3 + (1-a(\Delta x_{M-1})^3)(x-x_M)/\Delta x_{M-1} + 1 . \tag{4.7}$$

Examination of (4.7) shows that A_M has no interior relative maximums or minimums on $[x_{M-1}, x_M]$. The bound for A_1 follows in the same way. ∎

We will now give a proof similar to that of Birkhoff and De Boor [2] for natural cardinal cubic splines.

Lemma 4.8. For $1 \leqslant j \leqslant i-2$, if $x_j \leqslant x \leqslant x_{j+1}$, then $|A_i(x)| \leqslant \Delta x_j |A_i'(x_{j+1})|$ and if $i+1 \leqslant j \leqslant M-1$ for $x_j \leqslant x \leqslant x_{j+1}$, then $|A_i(x)| \leqslant \Delta x_j |A_i'(x_j)|$.

Proof: Case 1: $1 \leqslant j \leqslant i-2$.

Without loss of generality assume $A_i'(x_{j+1}) < 0$. From Lemma 4.5 $A_i(x) \geqslant 0$ on $[x_j, x_{j+1}]$, and from Lemma 4.4 $A_i''(x_{j+1}) < 0$, $A_i'(x_j) > 0$, $A_i''(x_j) \geqslant 0$. Our proof is by contradiction. Let ξ_1 be the point where A_i obtains its absolute maximum on $[x_j, x_{j+1}]$, then $A_i'(\xi_1) = 0$. Assume $A_i(\xi_1) > -A_i'(x_{j+1})\Delta x_j > 0$,

then because $A_i(x_{j+1}) = 0$, we have from the M. V. T. a ξ_2 satisfying $\xi_1 < \xi_2 < x_{j+1}$ such that $A_i'(\xi_2) = -A_i(\xi_1)/\Delta x_j < A_i'(x_{j+1}) < 0$. Applying the M. V. T. again we have the existence of ξ_3 satisfying $\xi_2 < \xi_3 < x_{j+1}$ such that $A_i''(\xi_3) > 0$. Because $A_i''(x_{j+1}) < 0$, the I. V. T. gives a ξ_4 satisfying $\xi_3 < \xi_4 < x_{j+1}$ such that $A_i''(\xi_4) = 0$. Because $A_i'(x_j) > 0$, $A_i'(\xi_1) = 0$ and $A_i''(x_j) \geq 0$ ($A_i''(x_1) = 0$), we have a second zero of A_i'', which implies that A_i is linear on $[x_j, x_{j+1}]$. This contradicts the fact that A_i must satisfy $A_i(x_j) = A_i(x_{j+1}) = 0$ and $A_i'(x_j) > 0$.

The proof of the other case is identical.

■

Paralleling the results of Birkhoff and De Boor [2], we have the following two lemmas for natural cubic cardinal splines.

Lemma 4.9: For $1 \leq j \leq i-2$

$$|A_i'(x_j)| \leq 1/2 \ |A_i'(x_{j+1})| ,$$

and for $i+1 \leq j \leq M-1$

$$|A_i'(x_{j+1})| \leq 1/2 \ |A_i'(x_j)| .$$

Proof: From Corollary 4.1.

■

Lemma 4.10. For $1 \leqslant j \leqslant i-2$, on $[x_j, x_{j+1}]$ we have

$$|A_i(x)| \leqslant 2^{j-i+2} |A_i'(x_{i-1})| \Delta x_j ,$$

and for $i+2 \leqslant j \leqslant M$, on $[x_{j-1}, x_j]$ we have

$$|A_i(x)| \leqslant 2^{i+2-j} |A_i'(x_{i+1})| \Delta x_{j-1} .$$

Proof: Follows inductively from Lemmas 4.8 and 4.9.

■

We have now shown through a sequence of Lemmas that the natural cubic cardinal splines behave in the same manner as the splines which have zero derivatives as their boundary conditions. Therefore, the proof of the following Lemma is now identical to that given by Birkhoff and De Boor [2], and we will only state their conclusion.

Lemma 4.11. (Birkhoff and De Boor [2]). For $2 \leqslant i \leqslant M-1$ on $[x_{i-1}, x_{i+1}]$ we have $0 \leqslant A_i(x) \leqslant L$ and $|A_i'(x_{i-1})| \leqslant L/\Delta x_{i-1}$, $|A'(x_{i+1})| \leqslant L/\Delta x_i$, where $L = 3 M_{\pi_x} (M_{\pi_x}+1)^2/(3+4M_{\pi_x})$, the mesh ratio is $M_{\pi_x} = \max \Delta x_i / \min \Delta x_i$.

We will now show similar bounds for $|A_1'(x_2)|$ and $|A_N'(x_{M-1})|$ by constructing a majorant, $U_i(x)$, for the end splines. Consider first the spline A_M, then define the unique cubic spline U_M on $[x_{M-2}, x_M]$ satisfying $U_M(x_{M-2}) = U(x_{M-1}) = U_M''(x_{M-2}) =$

$U''_M(x_M) = 0$ and $U_M(x_M) = 1$. Define the cubic spline T on $[x_{M-2}, x_M]$ by $T(x) = (U_M - A_M)(x)$, then we will prove $T'(x_{M-1}) \geqslant 0$. If $M-2 = 1$, then $T''(x_{M-2}) = 0$ and we are done. If $M-2 > 1$, then $T''(x_{M-2}) \neq 0$, and we will show that $T'(x_{M-1}) > 0$. To do this, we use the fact that $T''(x_{M-2}) = 0 - A''_M(x_{M-2}) > 0$, which follows from Lemmas 4.4 and 4.6, and that T satisfies $T(x_M) = T''(x_M) = T(x_{M-1}) = T(x_{M-2}) = 0$. Also, T satisfies the conditions of Lemma 4.1 and therefore (4.3a) and (4.3b), for x_{M-2}, x_{M-1}, x_M. Because $T''(x_M) = 0$ and $T''(x_{M-2}) > 0$, we have that $T'(x_M) \neq 0$. From (4.3a) and (4.3b) we have that $\operatorname{sgn}(T'(x_{j-1})) = -\operatorname{sgn}(T'(x_j)) \neq 0$ for $j = M, M-1$ and $\operatorname{sgn}(T''(x_j)) = -\operatorname{sgn}(T'(x_j)) \neq 0$ for $j = M-2, M-1$. Therefore, because $T''(x_{M-2}) > 0$ it follows that $T'(x_{M-1}) = U'_M(x_{M-1}) - A'_M(x_{M-1}) > 0$. Using Lemma 4.6, and our above results, it follows that $U'_M(x_{M-1}) \geqslant A'_M(x_{M-1}) > 0$.

$U_M(x)$ is given explicitly by $p_1(x)$ on $[x_{M-2}, x_{M-1}]$ and $p_2(x)$ on $[x_{M-1}, x_M]$, where

$$p_1(x) = b(x-x_{M-1})(x-x_{M-2})(x-x_{M-1}+2\Delta x_{M-2})$$

$$p_2(x) = (x-x_{M-1})(-b\Delta x_{M-2}[(x-x_M)^2 - \Delta x_{M-1}(x-x_M)/\Delta x_{M-1} + 1]/\Delta x_{M-1})$$

where $b = 1/[2\Delta x_{M-2}\Delta x_{M-1}(\Delta x_{M-1} + \Delta x_{M-2})]$

and $U'_M(x_{M-1}) = \Delta x_{M-2}/(\Delta x_{M-1}[\Delta x_{M-1} + \Delta x_{M-2}])$.

Therefore,

$$(4.8) \qquad 0 < A'_M(x_{M-1}) \leqslant M_{\pi_x} / (\Delta x_{M-1} [1 + M_{\pi_x}]) .$$

A bound for $A'_1(x_2)$ follows in an identical way.

The above yields the following theorem.

Theorem 4.1. If $1 \leqslant i \leqslant M$, $1 \leqslant j \leqslant i-2$ and $x_j \leqslant x \leqslant x_{j+1}$, then

$$|A_i(x)| \leqslant 2^{2+j-i} L M_{\pi_x} .$$

If $i+2 \leqslant j \leqslant M$ and $x_{j-1} \leqslant x \leqslant x_j$, then

$$|A_i(x)| \leqslant 2^{2+i-j} L M_{\pi_x} .$$

For $2 \leqslant i \leqslant M-1$, $|A_i(x)| \leqslant L$ on $[x_{i-1}, x_{i+1}]$, $|A_1(x)| \leqslant L$ on $[x_1, x_2]$ and $|A_M(x)| \leqslant L$ on $[x_{M-1}, x_M]$, where $L = 3 M_{\pi_x} (M_{\pi_x} + 1)^2 / (3 + 4 M_{\pi_x})$ and M_{π_x} is the mesh ratio.

Proof: If $2 \leqslant i \leqslant M-1$, the result follows from Lemmas 4.10 and 4.11. For $i = 1, M$ our result follows from Lemmas 4.7 and 4.10 and inequality (4.8), where we have used the liberal estimate that $M_{\pi_x} / (1 + M_{\pi_x}) \leqslant L$ and $1 \leqslant L$. ∎

We are now in a position to develop the error estimate for approximate natural spline blending. Usually, it is not possible to use all of the values of f along the mesh lines. Therefore, it is

reasonable to replace the $f(x_i, y)$ and $f(x, y_j)$ with appropriate approximations $p_i(y)$ and $q_j(x)$. The type of approximation is arbitrary as long as the overall accuracy of the scheme is preserved. For example, typical choices of approximation could be univariate spline or polynomial interpolation. We then blend these functions with natural splines to obtain an approximation to f. We would expect this approximation to be close to the original function f. Exactly how close is shown in the following theorem.

We are given the meshes $\pi_x: a = x_1 < x_2 < \cdots < x_M = b$ and $\pi_y: c = y_1 < y_2 < \cdots < y_N = d$ and the natural cubic cardinal splines $\{A_i\}_{i=1}^{M} \subseteq S^2(\pi_x)$ and $\{B_j\}_{j=1}^{N} \subseteq S^2(\pi_y)$.

Theorem 4.2. Let $f \in C^{(m,n)}([a,b]x[c,d])$, where $2 \leqslant m, n \leqslant 4$, and if the approximations to the mesh functions satisfy $||f(x_i, \cdot) - p_i(\cdot)|| \leqslant \mathcal{E}_x$ and $||f(\cdot, y_j) - q_j(\cdot)|| \leqslant \mathcal{E}_y$ for $1 \leqslant i \leqslant M$ and $1 \leqslant j \leqslant N$, then the natural blending function approximation defined by

$$(4.9) \qquad \bar{s}(x,y) = \sum_{i=1}^{M} p_i(y)\, A_i(x) + \sum_{j=1}^{N} q_j(x)\, B_i(y) - \sum_{i=1}^{M}\sum_{j=1}^{N} (1/2\, p_i(y_j) + 1/2\, q_j(x_i)) A_i(x)\, B_j(y)$$

satisfies

$$(4.10) \qquad |(f-\bar{s})(x,y)| \leqslant |(I-NB)[f](x,y)| + \mathcal{E}_x K(M_{\pi_x}) + \mathcal{E}_y K(M_{\pi_y}) + 1/2(\mathcal{E}_x + \mathcal{E}_y) K(M_{\pi_x}) K(M_{\pi_y}),$$

where a bound for the first term on the right of inequality (4.10) is given by Theorem 3.2, M_{π_x} and M_{π_y} are the mesh ratios, and $K(\xi) = 6\xi(2\xi + 1)(\xi + 1)^2/(3 + 4\xi)$.

Proof: Estimate the term

$$|(NB[f] - \bar{s})(x, y)| \leq \mathcal{E}_x \sum_{i=1}^{M} |A_i(x)| + \mathcal{E}_y \sum_{j=1}^{N} |B_j(y)|$$

$$+ 1/2(\mathcal{E}_x + \mathcal{E}_y)\left(\sum_{i=1}^{M} |A_i(x)|\right)\left(\sum_{j=1}^{N} |B_j(y)|\right)$$

$$\leq \mathcal{E}_x K(M_{\pi_x}) + \mathcal{E}_y K(M_{\pi_y}) + 1/2(\mathcal{E}_x + \mathcal{E}_y)$$

$$\cdot K(M_{\pi_x}) K(M_{\pi_y}) .$$

The proof is completed by applying the triangle inequality, and carefully observing the bound for each term in each interval given by Theorem 4.1 and using the fact that $\sum_{i=0}^{\infty} 2^i = 2$.

■

Remark 4.1: It should be noted that the result of Theorem 3.2 is independent of the mesh ratios M_{π_x} and M_{π_y}, while the exponential decay of the basis functions is dependent on both M_{π_x} and M_{π_y}.

Remark 4.2: To see the size of $K(M_{\pi_x})$, consider the special case of equal spacing, where $M_{\pi_x} = 1$, then $K(1) = 72/7$.

Remark 4.3: If $m = n = 4$ and $\mathcal{E} = O(h^8)$, then the error satisfies Figure 1 from Remark 3.5 of Theorem 3.2.

CHAPTER 3

DISCRETE LEAST SQUARES

This chapter is devoted to the development of discrete least squares algorithms on unstructured data sets, and to showing the accuracy that can be expected depending upon the distribution of the data points and the smoothness of the function f from which the data originates.

Sections 1 and 2 develop preliminary estimates for use in the following section. In Section 3, an example is given to show the necessity of having a sufficient number of data points reasonably distributed to guarantee that the discrete least squares fit will be close the function f. The remaining portion of this section gives error estimates for several univariate discrete least squares algorithms.

Sections 4 and 5 extend the error estimates of Section 3 to bivariate functions, and algorithms are developed using discretized blending function spaces.

Finally, in Section 6, general domains with curved boundarys are considered, and this case is reduced to the methods of Section 5.

Section 1. Uniform Error Estimates

We will develop here an error analysis in the uniform norm for univariate cubic splines in terms of interpolation errors at the knots.

Let $s \in S^2(\pi_x)$ be a cubic spline on the mesh $\pi_x: a = x_1 < x_2 < \cdots < x_M = b$, and set

$$(1.1) \qquad s_j = s(x_j),\ s'_j = s^{(1)}(x_j) \text{ and } s''_j = s^{(2)}(s_j) \text{ for } 1 \leqslant j \leqslant M.$$

The mesh π_x is uniform if $x_{j+1} - x_j = h = (b-a)/(M-1)$ for $1 \leqslant j \leqslant M-1$. Also, throughout this chapter, if $f \in C[a,b]$, then

$$(1.2) \qquad ||f|| = \sup_{a \leqslant x \leqslant b} |f(x)| ,$$

and if $g \in C([a,b]x[c,d])$ then

$$(1.3) \qquad ||g|| = \sup_{\substack{a \leqslant x \leqslant b \\ c \leqslant y \leqslant d}} |g(x,y)| .$$

Finally, for $x \in \mathbb{R}^N$ where $x = (x_1, x_2, \ldots, x_N)^T$ define the vector norm $||x||_\infty = \max_{1 \leqslant i \leqslant N} |x_i|$. The matrix norm $||\cdot||_\infty$ of a matrix $A = [a_{ij}] \in \mathbb{R}^{NxN}$ subordinate to this vector norm is defined to be

$$||A||_\infty = \max_{1 \leqslant i \leqslant N} \sum_{j=1}^{N} |a_{ij}| , \text{ (see [9 , p. 108])} .$$

Lemma 1.1. Let $s \in S^2(\pi_x)$ be a cubic spline on the uniform mesh π_x such that $|s_j| \leqslant \xi$ and $|s'_j| \leqslant \eta$ for $1 \leqslant j \leqslant M$, then

$$(1.4) \qquad ||s|| \leqslant \xi + h\eta/4.$$

Proof: From [1, p. 12], if s is a cubic polynomial on $[x_{j-1}, x_i]$ for $1 < j \leqslant M$ then

$$(1.5) \qquad s(x) = s'_{j-1}(x_j - x)^2(x - x_{j-1})/h^2 - s'_j(x - x_{j-1})^2(x_j - x)/h^2 + s_{j-1}(x_j - x)^2[2(x - x_{j-1}) + h]/h^3 + s_j(x - x_{j-1})^2[2(x_j - x) + h]/h^3,$$

for $x \in [x_{j-1}, x_j]$. Taking absolute values in (1.5) and observing that all of the polynomial factors of s'_{j-1}, s'_j, s_{j-1} and s_j are positive, we have after a little algebra,

$$(1.6) \qquad |s(x)| \leqslant \eta[(x_j - x)^2(x - x_{j-1}) + (x - x_{j-1})^2(x_j - x)]/h^2 + \xi .$$

The maximum of the right hand side of (1.6) occurs for $x = x_{j-1} + h/2$, and our result follows. ∎

Lemma 1.2. Let $s \in \mathring{S}(\pi_x)$ be a cubic spline on the uniform mesh π_x such that $|s_j| \leqslant \xi$ for $1 \leqslant j \leqslant M$ and $|s'_k| \leqslant \eta$ for $k = 1, M$, then

$$(1.7) \qquad |s'_j| \leqslant 3\xi/h + \eta/2 \quad \text{for} \quad 2 \leqslant j \leqslant M-1 .$$

Proof: From [1, p. 12], if s is a cubic spline then for $2 \leqslant j \leqslant M-1$

(1.8a) $$s'_{j-1} + 4s'_j + s'_{j+1} = 3(s_{j+1} - s_{j-1})/h ,$$

(1.8b) $$4s'_2 + s'_3 = 3(s_3 - s_1)/h - s'_1 ,$$

and

(1.8c) $$s'_{M-2} + 4s'_{M-1} = 3(s_M - s_{M-2})/h - s'_M .$$

Writing (1.8a) - (1.8c) in matrix notation we obtain

(1.9)
$$\begin{bmatrix} 4 & 1 & & & & \\ 1 & 4 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & 1 & 4 & 1 & \\ & & & 1 & 4 & \end{bmatrix} \begin{bmatrix} s'_2 \\ s'_3 \\ \vdots \\ s'_{M-1} \end{bmatrix} = \begin{bmatrix} 3(s_3 - s_1)/h - s'_1 \\ 3(s_4 - s_2)/h \\ \vdots \\ 3(s_{j+1} - s_{j-1})/h \\ \vdots \\ 3(s_{M-1} - s_{M-3})/h \\ 3(s_M - s_{M-2})/h - s'_M \end{bmatrix} ,$$

or $AX = B$, where A, X, and B correspond to the quantities in (1.9).

Premultiply the tridiagonal matrix $A \in \mathbb{R}^{(M-2)\times(M-2)}$ by the matrix $D = (1/4)I$, where I is the identity matrix. We obtain $DA = I + C$ where

(1.10)
$$C = 1/4 \begin{bmatrix} 0 & 1 & & & & \\ 1 & 0 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & 1 & 0 & 1 & \\ & & & 1 & 0 & \end{bmatrix} .$$

Examination of (1.10) yields $||C||_\infty = 1/2$, and using (3.16) of Chapter 1 we have the existence of $(DA)^{-1}$ and

$$(1.11) \qquad ||(DA)^{-1}||_\infty \leqslant 1/(1 - ||C||_\infty) \leqslant 2 .$$

From (1.11) we can obtain a bound on the derivatives of s at the knots

$$(1.12) \qquad ||X||_\infty = ||(DA)^{-1} DB||_\infty \leqslant ||(DA)^{-1}||_\infty ||D||_\infty ||B||_\infty \leqslant 1/2 \, ||B||_\infty .$$

To obtain a bound on $||B||_\infty$ we use (1.9) which yields

$$(1.13) \qquad ||B||_\infty = \max \{ \max_{3 \leqslant j \leqslant M-2} |3(s_{j+1} - s_{j-1})/h| , |3(s_3 - s_1)/h - s'_1|, |3(s_M - s_{M-2})/h - s'_M| \} ,$$

where

$$(1.14) \qquad \max_{3 \leqslant j \leqslant M-2} |3(s_{j+1} - s_{j-1})/h| \leqslant 6\xi/h ,$$

$$(1.15) \qquad |3(s_3 - s_1)/h - s'_1| \leqslant 6\xi/h + \eta ,$$

and

$$(1.16) \qquad |3(s_M - s_{M-2})/h - s'_M| \leqslant 6\xi/h + \eta .$$

Therefore $||B||_\infty \leqslant 6\xi/h + \eta$, completing our proof of (1.7). ∎

Combining Lemmas 1.1 and 1.2 gives a uniform norm estimate for s in terms of its values at the knots.

Lemma 1.3. Let $s \in S^2(\pi_x)$ be a cubic spline on the uniform mesh π_x such that $|s_j| \leqslant \xi$ for $1 \leqslant j \leqslant M$ and $|s'_k| \leqslant \eta$ for $k = 1, M$, then

$$||s|| \leqslant (7/4)\,\xi + (1/4)\,h\eta . \tag{1.17}$$

Proof: Using Lemma 1.2 and our hypothesis we have for $1 \leqslant j \leqslant M$

$$\begin{aligned} |s'_j| &\leqslant \max\{\eta,\ 3\xi/h + \eta/2\} \\ &\leqslant 3\,\xi/h + \eta . \end{aligned} \tag{1.18}$$

Combining (1.18) and Lemma 1.1 completes our proof.

■

Lemma 1.3 is really a stability result for errors in interpolation. To see this, let s_1 be the cubic spline which interpolates the following values of $f \in C^{(m)}[a,b]$, $1 \leqslant m \leqslant 4$

$$s_1(x_i) = f(x_i) , \quad \text{for } 1 \leqslant i \leqslant M, \tag{1.19}$$

and

$$s'_1(x_k) = f'(x_k), \quad \text{for } k = 1, M. \tag{1.20}$$

Also, let s_2 be the cubic spline which interpolates the following functional values of f with errors

$$(1.21) \qquad s_2(x_i) = f(x_i) + \xi_i \text{ , for } 1 \leqslant i \leqslant M,$$

and

$$(1.22) \qquad s_2'(x_k) = f'(x_k) + \eta_k \text{ , for } k = 1, M,$$

where $|\xi_i| \leqslant \xi$ and $|\eta_k| \leqslant \eta$. Using Theorem 1.1 of Chapter 2 and Lemma 1.3 we have an error bound for the approximate interpolation spline s_2

$$(1.23) \qquad ||f-s_2|| \leqslant ||f-s_1|| + ||s_1-s_2||$$

$$\leqslant \mathcal{E}_{m,0} ||f^{(m)}|| \, h^m + (7/4)\xi + (1/4)\, h \cdot \eta ,$$

where $\mathcal{E}_{m,0}$ are given in Table 1 of Chapter 2.

Remark 1.1: Often in performing spline interpolation, an approximate method must be employed to estimate the derivatives at the end points. It is clear from (1.23) that the values $f'(x_1)$ and $f'(x_M)$ must be approximated to at least $0(h^{m-1})$ to preserve the accuracy of the interpolation scheme. Specifically, Lagrange interpolation polynomials can be used to estimate the derivatives. Let p_1 and p_M be the Lagrange interpolation polynomials of degree m-1 on m points in the intervals $[x_1, x_2]$ and $[x_{M-1}, x_M]$ respectively. From [24, p. 289] we have for k = 1, M

$$(1.24) \qquad |f'(x_k) - p_k'(x_k)| \leq h^{m-1} ||f^{(m)}||/(m-1)! \ .$$

Therefore, if we define s_3 to be that cubic spline which interpolates $s_3(x_i) = f(x_i) + \xi_i$ for $1 \leq i \leq M$, $s_3'(x_1) = p_1'(x_1)$ and $s_3'(x_M) = p_M'(x_M)$, then

$$(1.25) \qquad ||f-s_3|| < (\mathcal{E}_{m,0} + 1/[4(m-1)!])||f^{(m)}||h^m + (7/4)\xi \ .$$

Examination of (1.25) indicates that the preservation of the order of the method depends upon limiting the interpolation errors, ξ_i, to be $0(h^m)$. We will use (1.24) later to obtain an error bound for univariate least squares.

Section 2. A Uniform Bound

For this section we assume that we are given the set $X \subseteq [a, b]$ of $\hat{M} \geqslant 3(M-1)$ data points and the uniform mesh $\pi_x : a = x_1 < x_2 < \cdots < x_M = b$, where $h = (b-a)/(M-1)$. For each j where $2 \leqslant j \leqslant M-1$ we assume that there exists six fixed data points which are designated by $\{\bar{x}_j^i\}_{i=1}^3 \cup \{x_j^i\}_{i=1}^3 \subseteq X$ such that

$$(2.1a) \qquad x_{j-1} \leqslant \bar{x}_j^3 < \bar{x}_j^2 < \bar{x}_j^1 < x_j < x_j^1 < x_j^2 < x_j^3 \leqslant x_{j+1},$$

and also the existence of $\{x_1^i\}_{i=1}^3 \subseteq X$ and $\{\bar{x}_M^i\} \subseteq X$ such that

$$(2.1b) \qquad x_1 < x_1^1 < x_1^2 < x_1^3 \leqslant x_2 \text{ and } x_{M-1} \leqslant \bar{x}_M^3 < \bar{x}_M^2 < \bar{x}_M^1 < x_M.$$

Remark: It need not be the case that $x_{j-1}^1 = \bar{x}_j^3$ or $x_{j-1}^2 = \bar{x}_j^2$ or $x_{j-1}^3 = \bar{x}_j^1$, although equality in any of the above is acceptable as long as they satisfy the conditions of Lemma 2.1.

Also, for notational convenience, we define $x_j = \bar{x}_j^0 = x_j^0$, and powers of x_j^i will be written as $(x_j^i)^n$ to avoid confusion.

We introduce the following notation for the k^{th} divided difference of the cubic spline $s \in \hat{S}(\pi_x)$

$$(2.2a) \qquad \Delta_j^k = s\left[x_j^0, x_j^1, \cdots, x_j^k\right] = \sum_{i=0}^{k} s(x_j^i)\,\mu(k, j, i), \text{ and}$$

$$(2.2b) \qquad \bar{\Delta}_j^k = s\left[\bar{x}_j^0, \bar{x}_j^1, \cdots, \bar{x}_j^k\right] = \sum_{i=0}^{k} s(\bar{x}_j^i)\,\mu(k, j, i),$$

where, for $0 \leqslant i \leqslant k$

$$(2.3a)\qquad \mu(k,j,i) = 1 \Big/ \prod_{\substack{\ell=0 \\ \ell\neq i}}^{k} (x_j^i - x_j^\ell), \quad \text{and}$$

$$(2.3b)\qquad \bar{\mu}(k,j,i) = 1 \Big/ \prod_{\substack{\ell=0 \\ \ell\neq i}}^{k} (\bar{x}_j^i - \bar{x}_j^\ell),$$

(see [24, p. 247]).

Lemma 2.1. Let $s \in S^2(\pi_x)$ be a cubic spline such that $|s'(x_1)|$, $|s'(x_M)| \leq \eta$ and $|s(\bar{x}_j^i)|$, $|s(x_j^i)| \leq \xi$ where the data points $\bar{x}_j^i$ and x_j^i satisfy (2.1a) and (2.1b). Also, define the real numbers γ, α and $\hat{\alpha}$ by

$$(2.4)\qquad \gamma h = \min\Big\{ \min_{1\leq j\leq M-1} \Big\{ -h - \sum_{i=1}^{3} (x_j - x_j^i) \Big\}, \min_{2\leq j\leq M} \Big\{ -h + \sum_{i=1}^{3} (x_j - \bar{x}_j^i) \Big\} \Big\},$$

$$(2.5a)\qquad \alpha h = \min\Big\{ \min_{1\leq j\leq M-1} \min_{\substack{0\leq i,\ell\leq 3 \\ i\neq\ell}} \{ |x_j^i - x_j^\ell| \}, \min_{2\leq j\leq M} \min_{\substack{0\leq i,\ell\leq 3 \\ i\neq\ell}} \{ |\bar{x}_j^i - \bar{x}_j^\ell| \} \Big\} > 0,$$

and

$$(2.5b)\qquad \hat{\alpha} h^3 = \min\Big\{ \min_{1\leq j\leq M-1} \min_{1\leq i\leq 3} |1/\mu(3,j,i)|, \min_{2\leq j\leq M} \min_{1\leq i\leq 3} |1/\bar{\mu}(3,j,i)| \Big\} > 0.$$

If $\gamma > 0$, then

(2.6) $$||s|| \leqslant 7(1-\alpha)(1-2\alpha)\left[(3-6\alpha)\,\xi \,/\, \hat{\alpha} + h\eta\right]/(4\gamma) + h\eta/4.$$

Remark: From (2.5a) and (2.5b) it follows that $1/3 \geqslant \hat{\alpha} > 0$ and $1 > \alpha > 0$.

Proof From [1, p. 10], $s \in S_3(\pi_x)$ is a cubic spline if and only if for $2 \leqslant j \leqslant M-1$

(2.6a) $$(h^2/6)(s''_{j-1} + 4s''_j + s''_{j+1}) = s_{j-1} - 2s_j + s_{j+1},$$

(2.6b) $$(h^2/6)(2s''_1 + s''_2) = s_2 - s_1 - h s'_1,$$

and

(2.6c) $$(h^2/6)(s''_{M-1} + 2s''_M) = h s'_M - s_M + s_{M-1}.$$

Because $s \in C^{(2)}[a, b]$ is a cubic polynomial on $[x_j, x_{j+1}]$ for $1 \leqslant j \leqslant M-1$ we have $s''_{j+1} = s''_j + s'''(x_j)$, where $s'''(x_j)$ is evaluated from the right. Also, $s'''(x) = \text{constant} = 6\Delta^3_j$ on $[x_j, x_{j+1}]$, (see [24, p. 249]. Therefore,

(2.7) $$s''_{j+1} = s''_j + 6h\Delta^3_j, \text{ for } 1 \leqslant j \leqslant M-1,$$

and correspondingly

(2.8) $$s''_{j-1} = s''_j - 6h\overline{\Delta}^3_j, \text{ for } 2 \leqslant j \leqslant M.$$

Because s is cubic on $[x_j, x_{j+1}]$ for $1 \leqslant j \leqslant M-1$, we represent s on that interval as

(2.9) $$s_j = \Delta^0_j + (x-x^0_j)\Delta^1_j + (x-x^0_j)(x-x^1_j)\Delta^2_j$$

$$+ (x-x_j^0)(x-x_j^1)(x-x_j^2)\,\Delta_j^3 ,$$

(see [24, p. 248]). Differentiating (2.9) twice and evaluating at x_j we obtain

$$(2.10) \qquad s_j'' = 2\,\Delta_j^2 + 2(2x_j - x_j^1 - s_j^2)\,\Delta_j^3 .$$

Correspondingly, on $[x_{j-1}, x_j]$ for $2 \leqslant j \leqslant M$ we have

$$(2.11) \qquad s_j'' = 2\,\bar{\Delta}_j^2 + 2(2x_j - \bar{x}_j^1 - \bar{x}_j^2)\,\bar{\Delta}_j^3 .$$

Substituting (2.7), (2.8), (2.10) and (2.11) into (2.6a) through (2.6c) it follows that for $2 \leqslant j \leqslant M-1$

$$(2.12a) \qquad h^2\big[(2x_j - \bar{x}_j^1 - \bar{x}_j^2 - h)\,\bar{\Delta}_j^3 + \bar{\Delta}_j^2 + \Delta_j^2 + (2x_j - x_j^1 - x_j^2 + h)\,\Delta_j^3\big] = s_{j-1} - 2s_j + s_{j+1} ,$$

$$(2.12b) \qquad h^2\big[(2x_1 - x_1^1 - x_1^2 + h)\,\Delta_1^3 + \Delta_1^2\big] = s_2 - s_1 - h s' ,$$

and

$$(2.12c) \qquad h^2\big[(2x_M - \bar{x}_M^1 - \bar{x}_M^2 - h)\,\bar{\Delta}_M^3 + \bar{\Delta}_M^2\big] = h s'_M - s_M + s_{M-1} .$$

Substituting (2.2a) and (2.2b) for our divided differences into (2.12a) through (2.12c), we obtain the following after regrouping

$$(2.13a) \qquad h^2\big[(2x_j - \bar{x}_j^1 - \bar{x}_j^2 - h)\,\bar{\mu}(3,j,0) + \bar{\mu}(2,j,0) + \mu(2,j,0) + (2x_j - x_j^1 - x_j^2 + h)\,\mu(3,j,0)\big]\, s_j + 2s_j - s_{j-1} - s_{j+1}$$

$$= -h^2\Big[(2x_j - \bar{x}_j^1 - \bar{x}_j^2 - h) \sum_{i=1}^{3} s(\bar{x}_j^i)\,\bar{\mu}(3,j,i)$$

$$+ \sum_{i=1}^{2} s(\bar{x}_j^i)\,\bar{\mu}(2,j,i) + \sum_{i=1}^{2} s(x_j^i)\,\mu(2,j,i)$$

$$+ (2x_j - x_j^1 - x_j^2 + h) \sum_{i=1}^{3} s(x_j^i)\,\mu(3,j,i)\Big] ,$$

(2.13b) $$h^2\Big[(2x_1 - x_1^1 - x_1^2 + h)\,\mu(3,1,0) + \mu(2,1,0)\Big]\, s_1 + s_1 - s_2$$

$$= -h^2\Big[(2x_1 - x_1^1 - x_1^2 + h) \sum_{i=1}^{3} s(x_1^i)\,\mu(3,1,i)$$

$$+ \sum_{i=1}^{2} s(x_1^i)\,\mu(2,1,i)\Big] - h s_1' ,$$

and

(2.13c) $$h^2\Big[(2x_M - \bar{x}_M^1 - \bar{x}_M^2 - h)\,\bar{\mu}(3,M,0) + \bar{\mu}(2,M,0)\Big]\, s_M$$

$$+ s_M - s_{M-1}$$

$$= -h^2\Big[(2s_M - \bar{x}_M^1 - \bar{x}_M^2 - h) \sum_{i=1}^{3} s(\bar{x}_M^i)\,\bar{\mu}(3,M,0)$$

$$+ \sum_{i=1}^{2} s(\bar{x}_M^i)\,\bar{\mu}(2,M,0)\Big] + h s_M' .$$

Using (2.3a) and (2.3b), we simplify (2.13a) through (2.13b) to obtain

(2.14a) $$-s_{j-1} + (\bar{\varepsilon}_j + 2 + \varepsilon_j)\, s_j - s_{j+1} = \bar{\delta}_j + \delta_j ,$$

(2.14b) $$(1 + \varepsilon_1)\, s_1 - s_2 = \delta_1 - h s_1' ,$$

and

(2.14c) $$(\bar{\varepsilon}_M + 1)\, s_M - s_{M-1} = \bar{\delta}_M + h s_M' ,$$

where for $1 \leqslant j \leqslant M-1$

$$\text{(2.15a)} \qquad \mathcal{E}_j = h^2\Big[h + \sum_{i=1}^{3} (x_j - x_j^i)\Big]\, \mu(3,j,0) > 0 \ ,$$

$$\text{(2.15b)} \qquad \delta_j = -h^2 \sum_{i=1}^{3} \Big[h + \sum_{\substack{\ell=1 \\ \ell \neq i}}^{3} (x_j - x_j^{\ell})\Big] s(x_j^i)\, \mu(3,j,i) \ ,$$

and for $2 \leqslant j \leqslant M$

$$\text{(2.15c)} \qquad \bar{\mathcal{E}}_j = h^2\Big[-h + \sum_{i=1}^{3} (x_j - \bar{x}_j^i)\Big]\, \bar{\mu}(3,j,0) > 0 \ ,$$

$$\text{(2.15d)} \qquad \bar{\delta}_j = -h^2 \sum_{i=1}^{3} \Big[-h + \sum_{\substack{\ell=1 \\ \ell \neq i}}^{3} (x_j - \bar{x}_j^{\ell})\Big] s(\bar{x}_j^i)\, \bar{\mu}(3,j,i) \ .$$

In (2.15a) and (2.15c) we have used (2.4) and the fact that $\gamma > 0$, $\mu(3,j,0) < 0$ and $\bar{\mu}(3,j,0) > 0$ to show that they are positive. In matrix form, equations (2.14a) through (2.14c) become

$$\text{(2.16)} \qquad \begin{bmatrix} 1+\mathcal{E}_1 & -1 & & & \\ -1 & \bar{\mathcal{E}}_2+2+\mathcal{E}_2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & \mathcal{E}_{M-1}+2+\mathcal{E}_{M-1} & -1 \\ & & & -1 & \mathcal{E}_M+1 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_M \end{bmatrix}$$

$$= \begin{bmatrix} \bar{\delta}_1 - h s'_1 \\ \bar{\delta}_2 + \delta_2 \\ \vdots \\ \bar{\delta}_{M-1} + \delta_{M-1} \\ \bar{\delta}_M + h s'_M \end{bmatrix},$$

or $A S = E$, where A, S and E correspond to the quantities in (2.16). In order to estimate the norm of the vector S we define the diagonal matrix D

(2.17) $$D = \begin{bmatrix} 1 + \varepsilon_1 & & & & \\ 0 & \bar{\varepsilon}_2 + 2 + \varepsilon_2 & & & \\ & \ddots & \ddots & & \\ & & \bar{\varepsilon}_{M-1} + 2 + \varepsilon_{M-1} & 0 & \\ & & & \bar{\varepsilon}_M + 1 & \end{bmatrix}.$$

From (2.15a) and (2.15c) we see that all of the diagonal entries of the matrix D are positive, which implies that D^{-1} exists. Using this fact, and premultiplying A by D^{-1} we obtain

$$D^{-1}A = I + B \,, \tag{2.18}$$

where the tridiagonal matrix B is given by

$$B=(-1)\cdot\begin{bmatrix} 0 & 1/(1+\varepsilon_1) & & & \\ 1/(\bar{\varepsilon}_2+2+\varepsilon_2) & 0 & 1/(\bar{\varepsilon}_2+2+\varepsilon_2) & & \\ & \ddots & \ddots & \ddots & \\ & & 1/(\bar{\varepsilon}_{M-1}+2+\varepsilon_{M-1}) & 0 & 1/(\bar{\varepsilon}_{M-1}+2+\varepsilon_{M-1}) \\ & & & 1/(\varepsilon_M+1) & 0 \end{bmatrix} \tag{2.19}$$

If $||B||_\infty < 1$, then, from (3.16) of Chapter 1, we have that $(D^{-1}A)^{-1}$ exists and

$$\begin{aligned} ||S||_\infty &\leq ||(D^{-1}A)^{-1}||_\infty ||D^{-1}E||_\infty \\ &\leq \left[1/(1-||B||_\infty)\right] ||D^{-1}E||_\infty . \end{aligned} \tag{2.20}$$

We now proceed to calculate bounds for $||B||_\infty$ and $||D^{-1}E||_\infty$.

Examination of (2.19), (2.15a) and (2.15c) shows that

$$(2.21)\qquad ||B||_\infty = \max\Big\{ \max_{2\leqslant j\leqslant M-1} \{2/(\bar{\varepsilon}_j+2+\varepsilon_j),\ 1/(1+\varepsilon_1),\ 1/(\bar{\varepsilon}_M+1)\}\Big\},$$

which implies that $\bar{\varepsilon}_j$ and ε_j should be bounded from below to bound $||B||_\infty$ from above.

Using (2.15a), (2.4) and (2.5a) it is clear that for $1 \leqslant j \leqslant M-1$

$$\begin{aligned}(2.22)\qquad \varepsilon_j &\geqslant h^3 \gamma |\mu(3,j,0)| \\ &\geqslant h^3 \gamma / |(x_j - x_j^1)(x_j - x_j^2)(x_j - x_j^3)| \\ &\geqslant h^3 \gamma / \left[h\cdot(1-2\alpha)\cdot h(1-\alpha)\cdot h\right] \\ &\geqslant \gamma / \left[(1-\alpha)(1-2\alpha)\right] .\end{aligned}$$

Correspondingly, from (2.15c), (2.4) and (2.5a) we have for $2 \leqslant j \leqslant M$ that

$$(2.23)\qquad \bar{\varepsilon}_j \geqslant \gamma / \left[(1-\alpha)(1-2\alpha)\right] .$$

Combining (2.21), (2.22) and (2.23) yields

$$\begin{aligned}(2.24)\qquad ||B||_\infty &\leqslant (1-\alpha)(1-2\alpha) / \left[(1-\alpha)(1-2\alpha) + \gamma\right] \\ &< 1 .\end{aligned}$$

In order to obtain a bound on $||D^{-1}E||_\infty$ we use (2.16) and (2.17) which gives

$$(2.25)\quad D^{-1}E = \begin{bmatrix} (\delta_1 - hs'_1)/(1+\varepsilon_1) \\ (\bar{\delta}_2 + \delta_2)/(\bar{\varepsilon}_2 + 2 + \varepsilon_2) \\ \vdots \\ (\bar{\delta}_{M-1} + \delta_{M-1})/(\bar{\varepsilon}_{M-1} + 2 + \varepsilon_{M-1}) \\ (\bar{\delta}_M + hs'_M)/(\bar{\varepsilon}_M + 1) \end{bmatrix} .$$

Which implies using (2.22) and (2.23) that

$$(2.26)\qquad ||D^{-1}E||_\infty \leqslant \left[(1-\alpha)(1-2\alpha)/((1-\alpha)(1-2\alpha)+\gamma)\right] ||\hat{E}||_\infty ,$$

where

$$(2.27)\quad \hat{E} = \begin{bmatrix} \delta_1 - hs'_1 \\ (\bar{\delta}_2 + \delta_2)/2 \\ \vdots \\ (\bar{\delta}_{M-1} + \delta_{M-1})/2 \\ \bar{\delta}_M + hs'_M \end{bmatrix} .$$

An upper bound on $||\hat{E}||_\infty$ is obtained by bounding $\bar{\delta}_j$ and δ_j from above. Using (2.5a) and carefully observing internal cancellations we have for $1 \leqslant i \leqslant 3$

$$(2.28)\qquad \Big|h + \sum_{\substack{\ell=1\\ \ell\neq i}}^{3} (x_j - x_j^{\ell})\Big| \leqslant h\,[1 - i\cdot\alpha]$$

and

$$(2.29)\qquad \Big| -h + \sum_{\substack{\ell=1\\ \ell\neq i}}^{3} (x_j - \bar{x}_j^{\ell})\Big| \leqslant h\,[1 - i\cdot\alpha]\ .$$

Combining (2.15b), (2.15d), (2.5b), (2.28) and (2.29) it follows that for $1 \leqslant j \leqslant M-1$

$$(2.30)\qquad |\delta_j| \leqslant (3 - 6\alpha)\xi\,/\hat{\alpha}\ ,$$

and for $2 \leqslant j \leqslant M$

$$(2.31)\qquad |\bar{\delta}_j| \leqslant (3 - 6\alpha)\,\xi\,/\hat{\alpha}\ .$$

Therefore, from (2.27), (2.30) and (2.31) follows an upper bound for $||\hat{E}||_\infty$

$$(2.32)\qquad ||\hat{E}||_\infty = \max\Big\{ \max_{2\leqslant j\leqslant M-1} \{\,|\bar{\delta}_j+\delta_j|/2\,\},\ |\delta_1 - h_1 s'_1|\ ,\ |\bar{\delta}_M + h s'_M|\Big\}$$

$$\leqslant (3 - 6\alpha)\xi\,/\hat{\alpha} + h\eta\ .$$

Combining (2.20), (2.24), (2.26) and (2.32) yields

$$(2.33)\qquad ||S||_\infty \leqslant (1-\alpha)(1-2\alpha)\Big[(3-6\alpha)\xi/\hat{\alpha} + h\eta\Big]\,/\,\gamma\ .$$

Our result follows from (2.33) and Lemma 1.3.

■

Section 3. Univariate Discrete Least Squares

Assume that we are given an unstructured set $X = \{\hat{x}_i\}_{i=1}^{\hat{M}} \subseteq [a,b]$ of $\hat{M}$ data points, and the mesh $\pi_x: a = x_1 < x_2 < \cdots < x_M = b$, where $h_x = \max_{1 \leq j \leq M-1} (x_{j+1} - x_j)$. If $f \in C^{(m)}[a,b]$, where $1 \leq m \leq 4$, then a cubic spline $s_{LS} \in S^2(\pi_x)$ which minimizes the Euclidean norm of the residual vector $R \in \mathbb{R}^{\hat{M}}$, where component i of R is

$$R_i = f(\hat{x}_i) - s_{LS}(\hat{x}_i), \tag{3.1}$$

i.e.

$$||R|| = \left(\sum_{i=1}^{\hat{M}} (f(\hat{x}_i) - s_{LS}(\hat{x}_i))^2 \right)^{1/2} = \min_{s \in S^2(\pi_x)} \left(\sum_{i=1}^{\hat{M}} (f(\hat{x}_i) - s(\hat{x}_i))^2 \right)^{1/2}, \tag{3.2}$$

is a discrete least squares approximation to f on the unstructured data set X. It would be desirable to have an estimate as to how close s_{LS} is to f in the uniform norm similar to the estimate we have for the cubic interpolation spline $s_f \in S^2(\pi_x)$, (Theorem 1.1 of Chapter 2). We would hope that if f is smooth and the residual vector R is small, then s_{LS} is close to f. This, however, is not the case, even if the cubic interpolation spline s_f is close to f. Therefore, as a prelude to the next theorem, we will give the following example which illustrates the importance of having sufficient data points which are reasonably distributed on $[a,b]$ with respect to the mesh π_x.

Example 3.1: We will construct the discrete least squares approximation $s_{LS} \in S^2(\pi_x)$ on the uniform mesh $\pi_x: a = x_1 < x_2 < \cdots < x_M = M$, where $x_i = i$ for $1 \leq i \leq M$, $h_x = 1$ and $M > 3$ to the function $f \in C^{(4)}[1, M]$

$$f(x) = \begin{cases} \mathcal{E}(4(x-5/4))^5 & \text{if } 1 \leq x \leq 5/4 \\ \\ 0 & \text{if } 5/4 < x \leq M \end{cases} . \tag{3.3}$$

Direct calculation using (3.3) yields $||f|| = \mathcal{E}$ and $||f^{(4)}|| = 30720\,\mathcal{E}$. If $s_f \in S^2(\pi_x)$ is the cubic interpolation spline which interpolates f on the mesh π_x, then from Theorem 1.1 of Chapter 2

$$||f - s_f|| \leq 400\,\mathcal{E} . \tag{3.4}$$

We will explicitly construct the discrete least squares spline approximation s_{LS} to f on the following $M+2$ data points

$$X = \{1 + i/4\}_{i=0}^{3} \cup \{j + \mathcal{E}\}_{j=2}^{M-1} \subseteq [1, M] , \tag{3.5}$$

where $f(1) = -\mathcal{E}$, $f(1 + i/4) = 0$ for $1 \leq i \leq 3$, $f(j+\mathcal{E}) = 0$ for $2 \leq j \leq M-1$, and we assume $0 < \mathcal{E} \leq 1\ 2$. Our construction will show that the residual vector $R \in \mathbb{R}^{M+2}$ is zero, and that s_{LS} is the unique discrete least squares solution to f on the data set X.

If R is zero, then s_{LS} interpolates f at the points $\{1 + i/4\}_{i=0}^{3}$, which implies that on $[1, 2]$, s_{LS} is uniquely given by

$$s_{LS}(x) = \mathcal{E} \cdot 32 \cdot (x - 5/4)(x-3/2)(x-7/4)/3 , \tag{3.6}$$

where

(3.7a) $s_{LS}(2) = \mathcal{E} > 0$,

(3.7b) $s'_{LS}(2) > 0$,

and

(3.7c) $s''_{LS}(2) > 0$.

Because $s_{LS} \in C^{(2)}[a, b]$ and $s_{LS}(j+\mathcal{E}) = f(j+\mathcal{E}) = 0$, it follows that on the interval $[j, j+1]$, s_{LS} has the unique representation

(3.8) $$s_{LS}(x) = -(s''_{LS}(j)/(2\mathcal{E}) + s'_{LS}(j)/\mathcal{E}^2 + s_{LS}(j)/\mathcal{E}^3)(x-j)^3 + (s''_{LS}(j)/2)(x-j)^2 + s'_{LS}(j)(x-j) + s_{LS}(j) \ .$$

If s_{LS} has been uniquely constructed on the interval $[1, j]$, then (3.8) gives a unique extension to $[1, j+1]$, hence inductively to $[1, M]$. From (3.8), if $s_{LS}(j)$, $s'_{LS}(j)$ and $s''_{LS}(j)$ are all non-zero and of the same sign, then $s_{LS}(j+1)$, $s'_{LS}(j+1)$ and $s''_{LS}(j+1)$ are all non-zero with the opposite sign and

(3.9) $$|s_{LS}(j+1)| > |s_{LS}(j)|/\mathcal{E}^2 \ ,$$

where we have used our assumption that $0 < \mathcal{E} \leqslant 1/2$ for the conservative lower bound in (3.9). Combining (3.7a), (3.7b), (3.7c) and (3.9) we conclude that for $3 \leqslant j \leqslant M$

(3.10) $$|s_{LS}(j)| > (1/\mathcal{E})^{2j-5} \ ,$$

(3.11) $$||f - s_{LS}|| > (1/\mathcal{E})^{2M-5} \ ,$$

and

$$(3.12) \qquad ||s_f - s_{LS}|| > (1/\varepsilon)^{2M-5} ,$$

where we have used the fact that $f(M) = s_f(M) = 0$. Therefore, by decreasing ε, we can cause f and its first four derivatives to be as small in norm as desired. However, the norm of the error in (3.11) can be made as large as desired, even though the residual vector R is zero.

From this example, it is clear that an acceptable upper bound on $||f-s_{LS}||$ can be obtained only if we place restrictions on the number and distribution of the data points with respect to the mesh π_x. Finally, observe that even if s_f is close to f, s_{LS} need not be close to either s_f or f.

Remark 3.1: The observation should be made that the mesh size $h = 1$ was chosen only for convenience, and its size is not crucial to the above example.

The above example motivates the hypothesis of the following theorem.

Theorem 3.1. Let $s_{LS} \in S^2(\pi_x)$ be a discrete least squares approximation to $f \in C^{(m)}[a,b]$, $1 \leqslant m \leqslant 4$, on the set X of $\hat{M}$ unstructured data points, where π_x is a uniform mesh. Assume for each j, where $2 \leqslant j \leqslant M-1$, that there exists six data points which we designate by $\{\bar{x}_j^i\}_{i=1}^3 \cup \{x_j^i\}_{i=1}^3 \subseteq X$ satisfying (2.1a), and also

$\{x_j^i\}_{i=1}^3 \subseteq X$ and $\{\bar{x}_M^i\}_{i=1}^3 \subseteq X$ satisfying (2.1b). Let the real numbers γ, α and $\hat{\alpha}$ be defined as in (2.4), (2.5a) and (2.5b), respectively. If $\gamma > 0$, $|f'(x_k) - s'_{LS}(s_k)| \leq \eta$, for $k = 1, M$, and at all of the data points x_j^i and $\bar{x}_j^i$ defined above $|f(x_j^i) - s_{LS}(x_j^i)| \leq \xi$ and $|f(\bar{x}_j^i) - s_{LS}(\bar{x}_j^i)| \leq \xi$, then

$$(3.13) \quad ||f-s_{LX}|| \leq \mathcal{E}_{m,0}(21(1-\alpha)(1-2\alpha)^2/(4\hat{\alpha}\gamma)+1)||f^{(m)}||\, h_x^m + 21(1-\alpha)(1-2\alpha)^2\xi/(4\hat{\alpha}\gamma) + (7(1-\alpha)(1-2\alpha)/(4\gamma)+1/4)\, h_x\, \eta ,$$

where $\mathcal{E}_{m,0}$ is given in Table 1 of Chapter 2 and $h_x = (b-a)/(M-1)$.

Proof: From Theorem 1.1 of Chapter 2, if $s_f \in S^2(\pi_x)$ is the cubic interpolation spline to f, then

$$(3.14) \quad ||f-s_{LS}|| \leq ||f-s_f|| + ||s_f-s_{LS}|| \leq \mathcal{E}_{m,0}||f^{(m)}||\, h_x^m + ||s_f-s_{LS}|| .$$

We now obtain a bound for $||s_f-s_{LS}||$ from Lemma 2.1. First observe that $s_f-s_{LS} \in S^2(\pi_x)$ and

$$(3.15) \quad |s'_f(x_k)-s'_{LS}(x_k)| = |f'(x_k) - s'_{LS}(x_k)| \leq \eta ,$$

for $k = 1, M$. Furthermore, for our specified data points we have

$$(3.16)\qquad |s_f(x_j^i) - s_{LS}(x_j^i)| \leqslant |s_f(x_j^i) - f(x_j^i)| + |f(x_j^i) - s_{LS}(x_j^i)|$$

$$\leqslant ||s_f - f|| + \xi$$

$$\leqslant \mathcal{E}_{m,0}||f^{(m)}||h_x^m + \xi ,$$

and correspondingly

$$(3.17)\qquad |s_f(\bar{x}_j^i) - s_{LS}(\bar{x}_j^i)| \leqslant \mathcal{E}_{m,0}||f^{(m)}||h_x^m + \xi .$$

Therefore, from Lemma 2.1

$$(3.18)\qquad ||s_f - s_{LS}|| \leqslant 7(1-\alpha)(1-2\alpha)(3-6\alpha)(\mathcal{E}_{m,0}||f^{(m)}||h_x^m$$

$$+ \xi)/4\hat{\alpha}\gamma + (7(1-\alpha)(1-2\alpha)/(4\gamma) + 1/4)\, h_x \eta .$$

Combining (3.14) and (3.18) yields (3.13) and the proof is complete.

■

Remark 3.2: In Theorem 3.1 it is permissible to use either of the following estimates for ξ ,

$$(3.19)\qquad \xi = ||R||_\infty ,$$

or

$$(3.20)\qquad \xi = ||R|| ,$$

where $||R||_\infty \leqslant ||R||$. In practice, however, $||R||$ is usually a poor choice when compared to $||R||_\infty$, especially for large $\hat{M}$.

Remark 3.3: It should be noted that during the numerical solution of a discrete least squares problem, the residual vector R is

calculated, (see [4 , 11, 21 , 28]). Therefore, R is available for use in the estimates (3.19) and (3.20).

Remark 3.4: Examination of (2.3a) through (2.5b) yields the following lower bound for γ and $\hat{\alpha}$

(3.21) $\qquad \gamma \geq 6\alpha - 1$

and

(3.22) $\qquad \hat{\alpha} \geq 2\alpha^3$.

However, 6α-1 and $2\alpha^3$ are usually much smaller than γ and $\hat{\alpha}$, respectively, and their use in Theorem 3.1 deteriorates the bound (3.13).

Example: If we assume that the data points X are distributed in such a way as to give $\alpha \geq 1/4$, then from (3.21) and (3.22) $\gamma \geq 1/2$ and $\hat{\alpha} \geq 1/32$. This gives the following bound from (3.13)

(3.23) $\qquad ||f-s_{LS}|| \leq \mathcal{E}_{m,0} 127\, ||f^{(m)}||h_x^m + 126\,\xi + (85/4)\, h_x \cdot \eta$.

Examination of Theorem 3.1 shows the need of insuring the smallness of η . We want to approximate $f'(x_k)$ for $k = 1, M$ in some manner which is compatible with the h^m. One possibility is as follows. If each of the intervals $[x_1, x_2]$ and $[x_{M-1}, x_M]$ contains at least m data points, respectively (the hypotheses of Theorem 3.1 guarantees at least 3), then using Lagrange interpolation polynomials of degree m-1, we can approximate $f'(x_k)$ for

$k = 1, M$ with the derivative of these polynomials at the end points of our interval x_1 and x_M (see Remark 1.1). If a constrained least squares algorithm is used to insure that the least squares spline s_{LS} has these values for its end derivatives, then inequality (1.24) gives a bound for M. Therefore, the following Corollary is an immediate consequence of Theorem 3.1 and the above construction.

Corollary 3.1. If the hypotheses of Theorem 3.1 are satisfied and there exists m data points in each of the intervals $[x_1, x_2]$ and $[x_{M-1}, x_M]$ on which the Lagrange interpolation polynomials of degree m-1 are constructed to approximate $f'(x_k)$, $k = 1, M$, and $s'_{LS}(x_k)$ is equal to those approximations, then

$$(3.24) \qquad ||f-s_{LS}|| \leq \left[\mathcal{E}_{m,0}(21(1-\alpha)(1-2\alpha)^2/(4\hat{\alpha}\gamma) + 1)\right.$$

$$\left. + (7(1-\alpha)(1-2\alpha)/4\gamma + 1/4)/(m-1)!\right] ||f^{(m)}||h_x^m$$

$$+ 21(1-\alpha)(1-2\alpha)^2\xi/(4\hat{\alpha}\gamma) \ .$$

It would also be desirable to have an a priori bound on $||f-s_{LS}||$, knowing only that $f \in C^{(m)}[a,b]$ and the distribution of the data points X. Toward this end we prove the following Corollary.

Corollary 3.2. If the hypotheses of both Theorem 3.1 and Corollary 3.1 are satisfied, then

$$(3.25)\qquad ||f-s_{LS}|| \leq \left[\mathcal{E}_{m,0}(21(1-\alpha)(1-2\alpha)^2(1+\hat{M}^{1/2})/(4\hat{\alpha}\gamma)+1)\right.$$

$$\left. + (7(1-\alpha)(1-2\alpha)/(4\gamma) + 1/4)/(m-1)!\right] ||f^{(m)}||\, h_x^m \ ,$$

where $\mathcal{E}_{m,0}$ is given in Table 1 of Chapter 2.

Proof: Examination of (3.24) shows that all that remains is to bound ξ. This will be accomplished by obtaining a bound for $||R||$ and using (3.20). The cubic interpolation spline s_f is a candidate for the discrete least squares approximation to f, hence the norm of its residual vector must be greater than or equal to $||R||$, i.e.,

$$(3.26)\qquad ||R|| \leq \Big(\sum_{i=1}^{\hat{M}} (f(\hat{x}_i) - s_f(\hat{x}_i))^2\Big)^{1/2}$$

$$\leq \Big(\sum_{i=1}^{\hat{M}} ||f - s_f||^2\Big)^{1/2}$$

$$\leq \mathcal{E}_{m,0}||f^{(m)}||\, h_x^m\, \hat{M}^{1/2} \ ,$$

where an application of Theorem 1.1 of Chapter 2 has been made. Combining (3.24), (3.20) and (3.26) completes the proof.

■

Because $\hat{M}^{1/2} > (3(M-1))^{1/2} > h_x^{-1/2}$, the power of h_x is no longer m, as (3.26) might indicate, but rather no larger than m-1/2, depending upon the number of data points. In practice, of course, we would expect (3.26) to be a rather crude estimate and

would hope that ξ is much closer to $\mathcal{E}_{m,0}||f^{(m)}||h_x^m$ instead of the bound given in (3.26).

We now give a more stringent condition on the data points, X, which will insure that the discrete least squares spline s_{LS} will be uniformly close to f. In preparation for the following theorem, the real non-negative number $||\cdot||_X$ of a function f is defined to be

$$(3.27) \qquad ||f||_X = \max_{\substack{1 \leqslant i \leqslant \hat{M} \\ \hat{x}_i \in X}} |f(\hat{x}_i)| ,$$

which is used in the following Lemma (see [6]).

Lemma 3.1. [6 , p. 91]. Let P be an algebraic polynomial of degree $\leqslant n$ on the interval $[a,b]$, then

$$(3.28) \qquad ||P||(1 - n^2 \beta_1(X)) \leqslant ||P||_X ,$$

where

$$(3.29) \qquad \beta_1(X) = \max_{x \in [a,b]} \min_{1 \leqslant i \leqslant \hat{M}} \{2|x-\hat{x}_i|/(b-a)\} ,$$

and

$$(3.30) \qquad ||P||(1 - (n\hat{\beta}_1(X))^2/2) \leqslant ||P||_X ,$$

where

$$(3.31)\qquad \hat{\beta}_1(X) = \max_{x\in[a,b]} \min_{1\leqslant i\leqslant \hat{M}} \left|\cos^{-1}((2x-(a+b))/(b-a)) - \cos^{-1}((2\hat{x}_i-(a+b))/(b-a))\right| .$$

The above Lemma yields the following theorem.

<u>Theorem 3.2.</u> Let $X = \{\hat{x}_i\}_{i=1}^{\hat{M}} \subseteq [a,b]$ be a set of $\hat{M}$ unstructured data points and $\pi_x: a = x_1 < x_2 < \cdots < x_M = b$ be a mesh on $[a,b]$. Let $s_{LS} \in S^2(\pi_x)$ be a discrete least squares approximation to $f \in C^{(m)}[a,b]$, $1 \leqslant m \leqslant 4$, on the data set X with residual vector $R \in \mathbb{R}^{\hat{M}}$. Define $\beta(X)$ and $\hat{\beta}(X)$ by

$$(3.32)\qquad \beta(X) = \max_{1\leqslant j\leqslant M-1} \max_{x\in[x_j,x_{j+1}]} \min_{\substack{1\leqslant i\leqslant \hat{M}\\ \hat{x}_i\in[x_j,x_{j+1}]}} \{2|x-\hat{x}_i|/(x_{j+1}-x_j)\} ,$$

$$(3.33)\qquad \hat{\beta}(X) = \max_{1\leqslant j\leqslant M-1} \max_{x\in[x_j,x_{j+1}]} \min_{\substack{1\leqslant i\leqslant \hat{M}\\ \hat{x}_i\in[x_j,x_{j+1}]}} \{ \left|\cos^{-1}((2x-(x_j+x_{j+1}))/(x_{j+1}-x_j)) - \cos^{-1}((2\hat{x}_i-(x_j+x_{j+1}))/(x_{j+1}-x_j))\right| \} .$$

If $\beta(X) < 1/9$, then

$$(3.34)\qquad ||f-s_{LS}|| \leqslant \left[(2-9\beta(X))\mathcal{E}_{m,0}||f^{(m)}||h_x^m + ||R||_\infty\right]/(1-9\beta(X)) ,$$

or if $\hat{\beta}(X) < \sqrt{2}/3$, then

$$(3.35) \qquad ||f-s_{LS}|| \leq \left[(2-(3\hat{\beta}(X))^2/2)\mathcal{E}_{m,0}||f^{(m)}||h_x^m + ||R||_\infty\right]/(1-(3\hat{\beta}(X))^2/2),$$

where $\mathcal{E}_{m,0}$ are given in Table 1 of Chapter 2 and $h_x = \max_{1\leq j\leq M-1}(x_{j+1}-x_j)$.

Proof: Let $s_f \in S^2(\pi_x)$ be the interpolation spline to $f \in C^{(m)}[a,b]$. It follows from Theorem 1.1 of Chapter 2 that

$$(3.36) \qquad ||f-s_{LS}|| \leq ||f-s_f|| + ||s_f - s_{LS}|| \leq \mathcal{E}_{m,0}||f^{(m)}||h_x^m + ||s_f - s_{LS}||.$$

To prove (3.34), we first note that $s_f - s_{LS}$ is a cubic polynomial on $[x_j, x_{j+1}]$ for $1 \leq j \leq M-1$. Because $\beta(X) < 1/9$, it follows by applying Lemma 3.1 to each interval $[x_j, x_{j+1}]$ that

$$(3.37) \qquad ||s_f - s_{LS}|| \leq ||s_f - s_{LS}||_X/(1 - 9\beta(X)).$$

Also, using (3.27) and the triangle inequality

$$(3.38) \qquad ||s_f - s_{LS}||_X \leq \max_{1\leq i\leq \hat{M}} \{|(f-s_f)(\hat{x}_i)| + |(f-s_{LS})(\hat{x}_i)|\} \leq ||f - s_f|| + ||R||_\infty \leq \mathcal{E}_{m,0}||f^{(m)}||\, h_x^m + ||R||_\infty .$$

Combining (3.36), (3.37) and (3.38) yields (3.34). The proof of (3.35) is identical to that of (3.34) and is omitted, thus completing our proof. ∎

Remark 3.5: Because of the high density required of the data points, X, it is not necessary to specify an estimate for $f'(x_k)$, $k = 1, M$, and therefore also unnecessary to employ a constrained least squares algorithm to solve for s_{LS}.

Remark 3.6: As was done in (3.26) we have

$$(3.39) \qquad ||R||_\infty \leq ||R|| \leq \sqrt{\hat{M}}\ \varepsilon_{m,0}\ ||f^{(m)}||\ h_x^m \ .$$

The bound (3.39) can be used in Theorem 3.2 to obtain an a priori bound for $||f-s_{LS}||$, where we only need to know X and $f \in C^{(m)}[a,b]$. However, because $\sqrt{\hat{M}} > \sqrt{M-1} \geq h_x^{-1/2}$, the best a priori bound obtainable using this method can not have the exponent of h exceed $m-1/2$.

Remark 3.7: The observation should be made that a uniform mesh is required for an application of Theorem 3.1. However, the conclusion of Theorem 3.2 is independent of the mesh spacing.

This section is concluded by recording a few observations on the distribution of the data points required by each of the above theorems. The hypotheses of Theorem 3.1 are satisfied if there exists at least three data points in each interval $[x_j, x_{j+1}]$ such that the spacing satisfies (2.1a), (2.1b) and in (2.4) we have $\gamma > 0$. The observation should be made that this is a reasonably weak condition to be placed on X. For example, in each interval, all of the three data points could lie in just half the interval, say $[x_j, x_j + h_x/2]$ and $\gamma > 0$ could still be realized. The data points $x_j^1 = h_x/3 + x_j$, $x_j^2 = 5h_x/12 + x_j$ and $x_j^3 = h_x/2 + x_j$, all of which lie in the half interval, are surely acceptable, and give the value $\gamma = 1/4 > 0$. If X contains an excess of data points beyond that required to fulfill the requirements (2.1a) and (2.1b), then a judicious selection of the three data points required for each interval can maximize the quantities γ, α and $\hat{\alpha}$ and therefore minimize the upper bound given in (3.13) for $||f - s_{LS}||$.

For $\beta(X) < 1/9$ in Theorem 3.2, the number of data points in each interval $[x_j, x_{j+1}]$ must exceed 9 and be distributed in a reasonably uniform manner throughout each interval. Usually for unstructured data, this will require the number of data points in each interval to be far in excess of 9. In Theorem 3.2, the number of data points in each interval must exceed $3\pi/(2\sqrt{2}) \cong 3.3$ to insure that $\hat{\beta}(X) < \sqrt{2}/3$. However, uniform distribution of the data points

is not what is required. Examination of (3.33) discloses that more data points are required near the end points of the interval rather than near the center.

Section 4. Bivariate Least Squares With Data on Mesh Lines

Let $\pi_x : a = x_1 < x_2 < \cdots < x_M = b$ and $\pi_y : c = y_1 < y_2 < \cdots < y_N = d$ be univariate meshes on $[a, b]$ and $[c, d]$ respectively. On $[a, b] \times [c, d]$ define the bivariate mesh $\pi_x \oplus \pi_y$ which consists of the M vertical mesh lines $x = x_i$ and N horizontal mesh lines $y = y_j$.

If unstructured data is given only on the mesh lines, $\pi_y \oplus \pi_y$, then it is possible to avoid solving a matrix problem of high dimension by solving a discrete least squares problem on each vertical and horizontal mesh line. The $M + N$ discrete least squares solutions are then blended with natural splines to obtain a bivariate approximation. Toward this end, we note that sufficient theory has been developed to yield several algorithms and corresponding error bounds.

For the first algorithm, the $M + N$ univariate discrete least squares cubic splines are constructed on each of the mesh lines of $\pi_x \oplus \pi_y$ from the cubic spline spaces $S^2(\pi_x)$ and $S^2(\pi_y)$. If we have sufficient data points on each of the $M + N$ mesh lines, then Theorem 3.1 or Theorem 3.2 would yield an error estimate for each of these splines. If $f \in C^{(m, n)}([a, b] \times [c, d])$, $1 \leqslant m, n \leqslant 4$, and the residual vector for each of the discrete least squares splines is small enough, then each would also be an approximation of order m or n, where $1 \leqslant m, n \leqslant 4$. Hence, we could blend these splines with

linear functions, which would give a bivariate approximation to f which is of order $\min\{m, 2\} + \min\{n, 2\}$.

Therefore, we introduce linear blending, see Gordon [12]. In the notation of Chapter 2, define $V(\pi_x, M, 0)$ to be the interpolation vector space of piecewise linear continuous functions, such that the basis functions $\{\phi_i\}_{i=1}^{M}$ are defined by

$$(4.1) \qquad \phi_i(x) = \begin{cases} (x-x_{i-1})/(x_i-x_{i-1}) & \text{if } x_{i-1} \leqslant x \leqslant x_i \\ (x_{i+1}-x)/(x_{i+1}-x_i) & \text{if } x_i \leqslant x \leqslant x_{i+1} \\ 0 \quad \text{otherwise} \end{cases}$$

and the interpolation function α is defined by $\alpha(i) = 0$ for $1 \leqslant i \leqslant M$. Define the corresponding space $V(\pi_y, N, 0)$ with basis $\{\psi_j\}_{j=1}^{N}$ and interpolation function β, of piecewise linear continuous functions on the mesh π_y in an identical fashion. $V(\pi_x, M, 0)$ (and correspondingly $V(\pi_y, N, 0)$) has the following error analysis: if $g \in C^{(m)}[a, b]$ then

$$(4.2) \qquad \Big|\Big| g - \sum_{i=1}^{M} g(x_i)\,\phi_i \Big|\Big| \leqslant K_m^* \,||f^{(m)}||\, h_x^{\,m} ,$$

where $1 \leqslant m \leqslant 2$, $K_1^* = 2$ and $K_2^* = 1/8$. When $m = 1$, the proof of (4.2) follows from [24, pp. 248-249], where, on the interval $[x_i, x_{i+1}]$

$$(4.3)\qquad \left|g(x) - \sum_{i=1}^{M} g(x_i)\,\phi_i(x)\right| = \left|(x-x_i)(x-x_{i+1})\,g\left[x_i, x_{i+1}, x\right]\right|$$

$$= \left|(x-x_{i+1})\left\{g\left[x_{i+1}, x\right] - g\left[x_i, x_{i+1}\right]\right\}\right|$$

$$\leqslant 2h_x\left|\left|g^{(1)}\right|\right| .$$

The proof of the case where $m = 2$ follows from [24, p. 248]. The error in linear blending interpolation to a function $f \in C^{(m,n)}$ $([a,b]\times[c,d])$, $1 \leqslant m, n \leqslant 2$ is given by Theorem 1.4 of Chapter 2 as

$$(4.4)\qquad \left|\left|f - P_x \oplus P_y\, f\right|\right| \leqslant K_m^* K_n^* \left|\left|f^{(m,n)}\right|\right| h_x^m h_y^n ,$$

where $K_1^* = 2$, $K_2^* = 1/8$, $h_x = \max_{1\leqslant i\leqslant M-1} (x_{i+1}-x_i)$ and

$h_y = \max_{1\leqslant j\leqslant N-1} (y_{j+1}-y_j)$.

If $h = \max(h_x, h_y)$ and $m = n = 2$, then in terms of h, this gives a fourth order approximation to f.

We now define stepwise an algorithm, which was described above. This algorithm will be designated as Algorithm 4.1.

Algorithm 4.1: Linear Blended Discrete Least Squares with Approximation to Boundary Derivaties.

Let $\pi_x: a = x_1 < x_2 < \cdots < x_M = b$ and $\pi_y: c = y_1 < y_2 < \cdots < y_N = d$ be uniform univariate meshes, and let $\pi_x \oplus \pi_y$ be the bivariate mesh of mesh lines defined above. For $1 \leqslant j \leqslant N$, let

$X_j = \{\hat{x}_{ji}\}_{i=1}^{\hat{M}_j} \subseteq [a, b]$ be an unstructured discrete data set of $\hat{M}_j$ points, and for $1 \leq i \leq M$ let $Y_i = \{\hat{y}_{ij}\}_{j=1}^{\hat{N}_i} \subseteq [c, d]$ be an unstructured discrete data set of $\hat{N}_i$ points.

Step 1: Let $f \in C^{(m,n)}([a, b] \times [c, d])$ where $1 \leq m, n \leq 4$. For $1 \leq j \leq N$ construct $s^x_{LS,j} \in S^2(\pi_x)$, which is an univariate discrete least squares cubic spline which minimizes the Euclidean norm of the residual vector $R^x_j \in \mathbb{R}^{\hat{M}_j}$, where component i of R^x_j is given by

$$(R^x_j)_i = f(\hat{x}_{ji}, y_j) - s_{LS,j}(\hat{x}_{ji}) \quad , \tag{4.5}$$

and $\hat{x}_{ji} \in X_j$. Also for each k, k=1, M, $(s^x_{LS,j})'(x_k)$ is constrained to equal the approximation to $f^{(1,0)}(x_k, y_j)$ given by the Lagrange interpolation polynomial of degree m-1 constructed on the m data points $\{\hat{\hat{x}}_{ji}\}_{i=1}^{m} \subseteq X_j$, where $\{\hat{\hat{x}}_{ji}\}_{i=1}^{m} \subseteq [x_1, x_2]$ for k=1 or $\subseteq [x_{M-1}, x_M]$ for k=2, which interpolates the values of $f(\hat{\hat{x}}_{ji}, y_j)$ (see Remark 1.1 for procedure). Correspondingly, for $1 \leq i \leq M$, construct $s^y_{LS,i} \in S^2(\pi_y)$, which is a univariate discrete least squares cubic spline which minimizes the Euclidean norm of the residual vector $R^y_i \in \mathbb{R}^{\hat{N}_i}$, where component j of R^y_i is given by

$$(R^y_i)_j = f(x_i, \hat{y}_{ij}) - s^y_{LS,i}(\hat{y}_{ij}) \quad , \tag{4.6}$$

and $\hat{y}_{ij} \in Y_i$. Also, for each k, k=1, N, $(s^y_{LS,i})'(y_k)$ is constrained to equal the approximation to $f^{(0,1)}(x_i, y_k)$ given by the Lagrange

interpolation polynomial of degree n-1 constructed on the n data points $\{\hat{\hat{y}}_{ij}\}_{j=1}^{n} \subseteq Y_i$, where $\{\hat{\hat{y}}_{ij}\}_{j=1}^{n} \subseteq [y_1, y_2]$ for k=1, or $\subseteq [y_{N-1}, y_N]$ for k=N, which interpolates the values $f(x_i, \hat{\hat{y}}_{ij})$.

Step 2: Linearly blend the M + N discrete least squares cubic splines to obtain the following approximation to f

$$(4.7) \qquad s(x,y) = \sum_{i=1}^{M} s^{y}_{LS,i}(y)\,\phi_i(x) + \sum_{j=1}^{N} s^{x}_{LS,j}(x)\,\psi_j(y)$$

$$- \sum_{i=1}^{M}\sum_{j=1}^{N} (1/2)(s^{y}_{LS,i}(y_j) + s^{x}_{LS,j}(x_i))\,\phi_i(x)\,\psi_j(y) \quad .$$

In order to obtain an error estimate, define the following parameters of the data sets X_j and Y_i. For each j, $1 \leq j \leq N$, assume that there are m distinct data points of X_j in each of the intervals $[x_1, x_2]$ and $[x_{M-1}, x_M]$, and for $1 < i < M$, assume that there exists fixed data points in X_j satisfying (2.1a) and also fixed data points satisfying (2.1b) with respect to the mesh π_x. For these fixed data points in X_j, define the real numbers γ_j^x, α_j^x and $\hat{\alpha}_j^x$ with respect to the mesh π_x by (2.4), (2.5a) and (2.5b), respectively (recall the notation that $x_i = \bar{x}_i^0 = x_i$ is a knot of the mesh π_x). Finally, define the real numbers

$$(4.8) \qquad \gamma^x = \min_{1 \leq j \leq N} \gamma_j^x \quad ,$$

$$(4.9) \qquad \alpha^x = \min_{1 \leq j \leq N} \alpha_j^x > 0 \quad ,$$

and

$$\hat{\alpha}^x = \min_{1 \leqslant j \leqslant N} \hat{\alpha}_j^x > 0 . \tag{4.10}$$

Correspondingly, for each i, $1 \leqslant i \leqslant M$, identical assumptions on the data sets Y_i, with respect to the mesh π_y, are made. Also, the real numbers γ^y, $\alpha^y > 0$ and $\hat{\alpha}^y > 0$ are defined in a manner analogous to (4.8), (4.9) and (4.10).

For Algorithm 4.1, the following error estimate is valid.

Theorem 4.1. Let s, defined in (4.7), be constructed by Algorithm 4.1. If $f \in C^{(m,n)}([a,b]\times[c,d])$, $\gamma^x > 0$ and $\gamma^y > 0$, then

$$\begin{aligned} \|f-s\| \leqslant\ & K^*_{m^*} K^*_{n^*} \|f^{(m^*,n^*)}\| h_x^{m^*} h_y^{n^*} \\ & + (3/2)\Big[\mathcal{E}_{m,0}(21(1-\alpha^x)(1-2\alpha^x)^2/(4\hat{\alpha}^x\gamma^x) + 1) \\ & + (7(1-\alpha^x)(1-2\alpha^x)/(4\gamma^x) + 1/4)/(m-1)!\Big] \|f^{(m,0)}\| h_x^m \\ & + (63/2)(1-\alpha^x)(1-2\alpha^x)^2 \max_{1\leqslant j\leqslant N} \|R_j^x\|_\infty / (4\hat{\alpha}^x\gamma^x) \\ & + (3/2)\Big[\mathcal{E}_{n,0}(21(1-\alpha^y)(1-2\alpha^y)^2/(4\hat{\alpha}^y\gamma^y) + 1) \\ & + (7(1-\alpha^y)(1-2\alpha^y)/(4\gamma^y) + 1/4)/(n-1)!\Big] \|f^{(0,n)}\| h_y^n \\ & + (63/2)(1-\alpha^y)(1-2\alpha^y)^2 \max_{1\leqslant i\leqslant M} \|R_i^y\|_\infty / (4\hat{\alpha}^y\gamma^y), \end{aligned} \tag{4.11}$$

where $m^* = \min\{m, 2\}$, $n^* = \min\{n, 2\}$, $\mathcal{E}_{m,0}$ and $\mathcal{E}_{n,0}$ are

given in Table 1 of Chapter 2, $K_1^* = 2$, $K_2^* = 1/8$, $h_x = (b-a)/(M-1)$ and $h_y = (d-c)/(N-1)$.

Proof: From (4.4), we have

$$(4.12)\qquad ||f-s|| \leqslant ||f - P_x \oplus P_y[f]|| + ||P_x \oplus P_y[f]-s||$$

$$\leqslant K_{m^*}^* K_{n^*}^* ||f^{(m^*, n^*)}|| h_x^{m^*} h_y^{n^*} + ||P_x \oplus P_y[f]-s||.$$

Writing out $P_x \oplus P_y[f] - s$ and using (4.7), gives

$$(4.13)\qquad P_x \oplus P_y[f] - s = \sum_{i=1}^{M} (f(x_i, \cdot) - s^y_{LS,i})\phi_i$$

$$+ \sum_{j=1}^{N} (f(\cdot, y_j) - s^x_{LS,j})\,\psi_j$$

$$- \sum_{i=1}^{M}\sum_{j=1}^{N} (1/2)(f(x_i, y_j) - s^y_{LS,i}(y_j) + f(x_i, y_j) - s^x_{LS,j}(x_i))\,\phi_i\,\psi_j .$$

Taking absolute values, observing that both ϕ_i and ψ_j are non-negative, $\sum_{i=1}^{M} \phi_i = 1$ and $\sum_{j=1}^{N} \psi_j = 1$, then (4.13) reduces to

$$(4.14)\qquad |(P_x \oplus P_y[f]-s)(x, y)| \leqslant (3/2) \max_{1\leqslant i\leqslant M} ||f(x_i, \cdot)$$

$$- s^y_{LS,i}|| + (3/2) \max_{1\leqslant j\leqslant N} ||f(\cdot, y_j)-s^x_{LS,j}|| .$$

Corollary 3.1 yields

$$(4.15)\qquad \max_{1\leq i\leq M} \|f(x_i,\cdot) - s^y_{LS,i}\|$$

$$\leq \Big[\mathcal{E}_{n,0}(21(1-\alpha^y)(1-2\alpha^y)^2/(4\hat{\alpha}^y\gamma^y)+1)$$

$$+ (7(1-\alpha^y)(1-2\alpha^y)/(4\gamma^y)+1/4)/(n-1)!\Big]\,\|f^{(0,n)}\|\,h_y^n$$

$$+ 21\,(1-\alpha^y)(1-2\alpha^y)^2 \max_{1\leq i\leq M} \|R_i^y\|_\infty/(4\hat{\alpha}^y\gamma^y)\,,$$

with a corresponding bound for $\max_{1\leq j\leq N} \|f(\cdot,y_j) - s^x_{LS,j}\|$. Combining (4.12), (4.14) and (4.15) completes the proof of the Theorem.

■

We first observe that if T_1 is the total number of data points, then T_1 must satisfy

$$(4.16)\qquad T_1 = \sum_{i=1}^{M} \hat{N}_i + \sum_{j=1}^{N} \hat{M}_j \geq 6\,MN - (M+N)\,,$$

for sufficient data points to be available to apply Theorem 4.1. Next, because only univariate least squares is being performed, the size (total number of entries) of the largest matrix that must be stored in the computer at one time is given by

$$(4.17)\qquad T_2 = \max\{(M+2)\max_{1\leq j\leq N} \hat{M}_j,\ (N+2)\max_{1\leq i\leq M} \hat{N}_i\}\,.$$

Instead, if a bivariate least squares solution was obtained using the discretized blending function space of linearly blended cubic splines, the matrix problem to be solved would be significantly larger than T_2, (usually of the order of $(T_2)^2$).

Example 4.1: Consider the special case when $m = n = 4$ and $h = \max(h_x, h_y)$. Hence $m^* = n^* = 2$ and (4.11) gives a bound in which the power of h is four. Also, an a priori bound can be obtained using Theorem 4.1 and the reasoning of (3.26) on each residual vector R_j^x and R_i^y. However, the highest realizable power of h is 3 1/2.

If each of the data sets X_j and Y_i has sufficient data points, the following algorithm can be applied, where it is not necessary to approximate the normal derivatives or to use uniform meshes.

Algorithm 4.2: Let $\pi_x: a = x_1 < x_2 < \cdots < x_M = b$ and $\pi_y: c = y_1 < y_2 < \cdots < y_N = d$ be univariate meshes and let $\pi_x \oplus \pi_y$ be the corresponding bivariate mesh. For $1 \leq j \leq N$, let $X_j = \{\hat{x}_{ji}\}_{i=1}^{\hat{M}_j} \subseteq [a, b]$ be an unstructured discrete data set of $\hat{M}_j$ points and for $1 \leq i \leq M$ let $Y_i = \{\hat{y}_{ij}\}_{j=1}^{\hat{N}_i} \subseteq [c, d]$ be discrete unstructured data sets of $\hat{N}_i$ points.

Step 1: Let $f \in C^{(m,n)}([a, b]x[c, d])$, where $1 \leq m, n \leq 4$. For $1 \leq j \leq N$, construct $s^x_{LS,j} \in S^2(\pi_x)$ which is a univariate discrete

least squares cubic spline which minimizes the Euclidean norm of the residual vector $R_y^x \in \mathbb{R}^{\hat{M}_j}$ given by (4.5). Correspondingly, for $1 \leqslant i \leqslant M$, construct $s^y_{LS,i} \in S^2(\pi_y)$ which minimizes the norm of the residual vector $R_i^y \in \mathbb{R}^{\hat{N}_i}$ defined in (4.6).

Step 2: Linearly blend the discrete least squares splines to obtain the bivariate approximation s defined in (4.7).

In order to obtain an error bound, calculate for each j, $1 \leqslant j \leqslant N$, the real numbers $\beta(X_j)$ and $\hat{\beta}(X_j)$ with respect to the mesh π_x from (3.32) and (3.33), respectively. Define the real numbers β_x and $\hat{\beta}_x$ by

$$\beta_x = \max_{1 \leqslant j \leqslant N} \beta(X_j) \quad , \tag{4.18}$$

$$\hat{\beta}_x = \max_{1 \leqslant j \leqslant N} \hat{\beta}(X_j) \quad . \tag{4.19}$$

Correspondingly, for each i, $1 \leqslant i \leqslant M$ calculate the numbers $\beta(Y_i)$ and $\hat{\beta}(Y_i)$ with respect to the mesh π_y from (3.32) and (3.33), respectively, and define the real numbers β_y and $\hat{\beta}_y$ by

$$\beta_y = \max_{1 \leqslant i \leqslant M} \beta(Y_i) \quad , \tag{4.20}$$

$$\hat{\beta}_y = \max_{1 \leqslant i \leqslant M} \hat{\beta}(Y_i) \quad . \tag{4.21}$$

The following error estimate is a consequence of Theorem 3.2.

Theorem 4.2. Let s, defined by (4.7), be constructed by Algorithm 4.2 and $f \in C^{(m,n)}([a,b]\times[c,d])$. If $\beta_x < 1/9$ and $\beta_y < 1/9$, then

$$(4.22)\qquad ||f-s|| \leq K^*_{m^*} K^*_{n^*} ||f^{(m^*,n^*)}|| h_x^{m^*} h_y^{n^*}$$

$$+(3/2)\left[(2-9\beta_x)\,\mathcal{E}_{m,0}||f^{(m,0)}||h_x^m + \max_{1\leq j\leq N}||R_j^x||_\infty\right]/(1-9\beta_x)$$

$$+(3/2)\left[(2-9\beta_y)\,\mathcal{E}_{n,0}||f^{(0,n)}||h_y^n + \max_{1\leq i\leq M}||R_i^y||_\infty\right]/(1-9\beta_y),$$

or if $\hat{\beta}_x < \sqrt{2}/3$ and $\hat{\beta}_y < \sqrt{2}/3$, then

$$(4.23)\qquad ||f-s|| \leq K_{m^*} K_{n^*} ||f^{(m^*,n^*)}|| h_x^{m^*} h_y^{n^*}$$

$$+(3/2)\left[(2-(3\hat{\beta}_x)^2/2)\,\mathcal{E}_{m,0}||f^{(m,0)}||h_x^m + \max_{1\leq j\leq N}||R_j^x||_\infty\right]$$

$$/(1-(3\hat{\beta}_x)^2/2)$$

$$+(3/2)\left[(2-(3\hat{\beta}_y)^2/2)\,\mathcal{E}_{n,0}||f^{(0,n)}||h_y^n\right.$$

$$\left. + \max_{1\leq i\leq M}||R_i^y||_\infty\right]/(1-(3\hat{\beta}_y)^2/2),$$

where $m^* = \min\{m,2\}$, $n^* = \min\{n,2\}$, $\mathcal{E}_{m,0}$ and $\mathcal{E}_{n,0}$ are given in Table 1 of Chapter 2, $K^*_1 = 2$, $K^*_2 = 1/8$, $h_x = \max_{1\leq i\leq M-1}(x_{i+1}-x_i)$ and $h_y = \max_{1\leq j\leq N-1}(y_{j+1}-y_j)$.

Proof: The proof of Theorem 4.2 is essentially identical to the proof of Theorem 4.1. We merely note that combining (4.12), (4.14) and the bounds of Theorem 3.2 yields our result.

■

Algorithms 4.1 and 4.2 are hybrid algorithms, which combine piecewise cubic and piecewise linear splines. If it is desired that cubic splines be used throughout, then the following algorithm, outlined below, may be useful.

We blend our discrete least squares cubic splines with natural cubic splines (see Section 3 of Chapter 2). If $f \in C^{(m,n)}([a,b]x[c,d])$, $2 \leq m,n \leq 4$, then this method has a potential accuracy of order $m+n$ in the "interior of the region", (see Sections 3 and 4 of Chapter 2). Examination of Theorem 4.2 of Chapter 2 indicates that preservation of this accuracy requires the discrete least squares cubic spline approximations to be of order $m+n$, when their residual vectors are small. To accomplish this, on each mesh line of $\pi_x \oplus \pi_y$, we replace the meshes π_x and π_y by uniform univariate meshes $\bar{\pi}_x: a = \bar{x}_1 < \bar{x}_2 < \cdots < \bar{x}_{\bar{M}} = b$ and $\bar{\pi}_y: c = \bar{y}_1 < \bar{y}_2 < \cdots < \bar{y}_{\bar{N}} = d$, respectively. From Theorem 1.1 of Chapter 2, the interpolation accuracy of the univariate cubic spline spaces $S^2(\bar{\pi}_x)$ and $S^2(\bar{\pi}_y)$ is $O(\bar{h}_x^m)$ and $O(\bar{h}_y^n)$ respectively, where $\bar{h}_x = (b-a)/(\bar{M}-1)$ and $\bar{h}_y = (d-c)/(\bar{N}-1)$. Therefore, for the preservation of our

accuracy, it is necessary to refine the meshes $\bar{\pi}_x$ and $\bar{\pi}_y$ sufficiently to have $\bar{h}_x^m \leq h_x^m\, h_y^n$ and $\bar{h}_y^n \leq h_y^m\, h_y^n$.

Remark: For the special case where $m = n = 4$ and $h_x = h_y = h$, this reduces to $\bar{h}_x \leq h^2$ and $\bar{h}_y \leq h^2$.

The above observations lead to the following algorithm.

Algorithm 4.3: Let $\pi_x: a = x_1 < x_2 < \cdots < x_M = b$, $\pi_y: c = y_1 < y_2 < \cdots < y_N = d$, $\bar{\pi}_x: a = \bar{x}_1 < \bar{x}_2 < \cdots < \bar{x}_{\bar{M}} = b$ and $\bar{\pi}_y: c = \bar{y}_1 < \bar{y}_2 < \cdots < \bar{y}_{\bar{N}} = d$ be univariate meshes such that $\bar{\pi}_x$ and $\bar{\pi}_y$ are uniform, $\bar{h}_x^m \leq h_x^m\, h_y^n$ and $\bar{h}_y^n \leq h_x^m\, h_y^n$ (these last two conditions are not necessary for Theorem 4.3 to remain valid, however, to preserve the desirability of Algorithm 4.3, attempts should be made to insure that these two conditions hold, or at least nearly hold). Let $\pi_x \oplus \pi_y$ be the bivariate mesh of $M+N$ mesh lines. The data points will only be specified on the mesh $\pi_x \oplus \pi_y$. For $1 \leq j \leq N$, let $X_j = \{\hat{x}_{ji}\}_{i=1}^{\hat{M}_j} \subseteq [a, b]$ be a discrete unstructured data set of $\hat{M}_j$ points, and for $1 \leq i \leq M$, let $Y_i = \{\hat{y}_{ij}\}_{j=1}^{\hat{N}_i} \subseteq [c, d]$ be a discrete unstructured data set of $\hat{N}_i$ points.

Step 1: Let $f \in C^{(m,n)}([a, b] \times [c, d])$ where $2 \leq m, n \leq 4$. For $1 \leq j \leq N$ construct $s^x_{LS,\,j} \in S^2(\bar{\pi}_x)$, which is a univariate discrete least squares cubic spline which minimizes the Euclidean norm of

the residual vector $R_j^x \in \mathbb{R}^{\hat{M}_j}$ defined by (4.5). Also, for $k = 1, \bar{M}$, $(s^x_{LS,j})'(x_k)$ is constrained to equal the approximation to $f^{(1,0)}(\bar{x}_k, y_j)$ given by the Lagrange interpolation polynomial of degree m-1 constructed on the m data points $\{\hat{\hat{x}}_{ji}\}_{i=1}^m \subseteq X_j$, where $\{\hat{\hat{x}}_{ji}\}_{i=1}^m \subseteq [\bar{x}_1, \bar{x}_2]$ for $k = 1$ or $\subseteq [\bar{x}_{\bar{M}-1}, \bar{x}_{\bar{M}}]$ for $k = \bar{M}$, which interpolates the values $f(\hat{\hat{x}}_{ji}, y_j)$. Correspondingly, for $1 \leqslant i \leqslant M$, construct $s^y_{LS,i} \in S^2(\bar{\pi}_y)$, which is a univariate discrete least squares cubic spline which minimizes the Euclidean norm of the residual vector $R_i^y \in \mathbb{R}^{\hat{N}_i}$ defined by (4.6). Also, for $k=1, \bar{N}$, $(s^y_{LS,i})'(\bar{y}_k)$ is constrained to equal the approximation to $f^{(0,1)}(x_i, \bar{y}_k)$ given by the Lagrange interpolation polynomial of degree n-1 constructed on the n data points $\{\hat{\hat{y}}_{ij}\}_{j=1}^n \subseteq Y_i$, where $\{\hat{\hat{y}}_{ij}\}_{j=1}^n \subseteq [\bar{y}_1, \bar{y}_2]$ for $k = 1$ or $\subseteq [\bar{y}_{\bar{N}-1}, \bar{y}_{\bar{N}}]$ for $k = \bar{N}$ which interpolates the values $f(x_i, \hat{\hat{y}}_{ij})$.

Step 2: Blend the above univariate discrete least squares cubic splines with natural cubic splines (see Section 3 of Chapter 2) to obtain the following bivariate approximation to f, where $\{A_i\}_{i=1}^M \subseteq S^2(\pi_x)$ and $\{B_j\}_{j=1}^N \subseteq S^2(\pi_y)$ are natural cardinal basis functions,

$$(4.24) \qquad s = \sum_{i=1}^{M} s^y_{LS,i} A_i + \sum_{j=1}^{N} s^x_{LS,j} B_j - \sum_{i=1}^{M}\sum_{j=1}^{N} (1/2)(s^y_{LS,i}(y_j) + s^x_{LS,j}(x_i)) A_i B_j .$$

The following assumptions are made on the distribution of the data points so that an error analysis can be performed. For each j, $1 \leq j \leq N$, assume that there exists m distinct data points of X_j in each interval $[\bar{x}_1, \bar{x}_2]$ and $[\bar{x}_{\bar{M}-1}, \bar{x}_{\bar{M}}]$, and for $1 \leq i \leq \bar{M}$, assume that there exists fixed data points in X_j satisfying (2.1a) and also fixed data points in X_j satisfying (2.1b) with respect to the mesh $\bar{\pi}_x$. For these fixed data points in X_j, define the real numbers $\bar{\gamma}_j^x$, $\bar{\alpha}_j^x$ and $\overset{\Delta x}{\alpha}_j$ with respect to the mesh $\bar{\pi}_x$ by (2.4), (2.5a) and (2.5b) respectively. Define the real numbers $\bar{\gamma}^x$, $\bar{\alpha}^x$ and $\overset{\Delta x}{\alpha}$ to be the minimum of $\bar{\gamma}_j^x$, $\bar{\alpha}_j^x$ and $\overset{\Delta x}{\alpha}_j$, respectively, for $1 \leq j \leq N$. Correspondingly, for each i, $1 \leq i \leq M$, identical assumptions on the data sets Y_j, with respect to the mesh $\bar{\pi}_y$, are made. The real numbers $\bar{\gamma}^y$, $\bar{\alpha}^y$ and $\overset{\Delta y}{\alpha}$ are defined in an analogous manner.

Algorithm 4.3, along with the assumptions on the distribution of the data points, gives the following theorem.

Theorem 4.3. Let s, defined in (4.24), be constructed by Algorithm 4.3. If $f \in C^{(m,n)}([a,b]\times[c,d])$, $\bar{\gamma}^x > 0$ and $\bar{\gamma}^y > 0$, then for $x \in [x_i, x_{i+1}]$ and $y \in [y_j, y_{j+1}]$

$$|f(x,y) - s(x,y)| \leq ||f^{(m,n)}||\{\mathcal{E}_{m,0}\, h_x^m + K_m(h_x^{m-1}\Delta x_i)\Delta_i/4\}\{\mathcal{E}_{n,0}\, h_y^n + K_n(h_y^{n-1}\Delta y_j)\bar{\Delta}_j/4\} \tag{4.25}$$

$$+ (1/8)||f^{(m,2)}||\{\mathcal{E}_{m,0}\, h_x^m + K_m(h_x^{m-1}\Delta x_i)\Delta_i/4\}(h_y\Delta y_j)\bar{\Delta}_j$$

$$+ (1/8)||f^{(2,n)}||\{\mathcal{E}_{n,0}\, h_y^n + K_n(h_y^{n-1}\Delta y_j)\bar{\Delta}_j/4\}(h_x\Delta x_i)\Delta_i$$

$$+ (1/64)||f^{(2,2)}||\{(h_x\Delta x_i)(h_y\Delta y_j)\}\Delta_i\bar{\Delta}_j$$

$$+ K(M_{\pi_y})(1+K(M_{\pi_x})/2)\{\left[\mathcal{E}_{m,0}(21(1-\bar{\alpha}^x)(1-2\bar{\alpha}^x)^2/(4\bar{\alpha}^{\Delta x}\bar{\gamma}^x)+1)\right.$$

$$\left.+ (7(1-\bar{\alpha}^x)(1-2\bar{\alpha}^x)/(4\bar{\gamma}^x)+1/4)/(m-1)!\right]||f^{(m,0)}||\,\bar{h}_x^m$$

$$+ 21(1-\bar{\alpha}^x)(1-2\bar{\alpha}^x)^2 \max_{1\le j\le N}||R_j^x||_\infty/(4\bar{\alpha}^{\Delta x}\bar{\gamma}^x)\}$$

$$+ K(M_{\pi_x})(1+K(M_{\pi_y})/2)\{\left[\mathcal{E}_{n,0}(21(1-\bar{\alpha}^y)(1-2\bar{\alpha}^y)^2/(4\bar{\alpha}^{\Delta y}\bar{\gamma}^y)+1)\right.$$

$$\left.+ (7(1-\bar{\alpha}^y)(1-2\bar{\alpha}^y)/(4\bar{\gamma}^y)+1/4)/(n-1)!\right]||f^{(0,n)}||\,\bar{h}_y^n$$

$$+ 21(1-\bar{\alpha}^y)(1-2\bar{\alpha}^y)^2 \max_{1\le i\le M}||R_i^y||_\infty/(4\bar{\alpha}^{\Delta y}\bar{\gamma}^y)\}\,,$$

where $\mathcal{E}_{m,0}$ and $\mathcal{E}_{n,0}$ are given in Table 3 of Chapter 2, K_m and K_n are given in Table 5 of Chapter 2, $\Delta x_i = x_{i+1}-x_i$, $\Delta y_j = y_{j+1}-y_j$, $h_x = \max_{1\le i\le M-1}\Delta x_i$, $h_y = \max_{1\le j\le N-1}\Delta y_j$, $\Delta_i = \{2^{1-i}+2^{1-M+i}\}$, $\bar{\Delta}_j = \{2^{1-j}+2^{1-N+j}\}$, $M_{\pi_x} = \max_{1\le i\le M-1}\Delta x_i \,/\, \min_{1\le i\le M-1}\Delta x_i$, $M_{\pi_y} = \max_{1\le j\le N-1}\Delta y_j \,/\, \min_{1\le j\le N-1}\Delta y_j$, $K(\xi) = 6\xi(2\xi+1)(\xi+1)^2/(3+4\xi)$, $\bar{h}_x = (b-a)/(\bar{M}-1)$ and $\bar{h}_y = (d-c)/(\bar{N}-1)$.

Proof: This is a direct consequence of Corollary 3.1 and Theorem 4.2 of Chapter 2.

■

Remark: The observation should be made that the order of accuracy depends upon which rectangle, $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$, the point (x, y) belongs to, because of the exponential decay of the terms Δ_i and $\overline{\Delta}_j$ toward the center of the region. If the norm of the residual vectors are small enough, $m = n = 4$, $h = \max(h_x, h_y)$, $\bar{h}_x \leq h^2$ and $\bar{h}_y \leq h^2$, then the error satisfies Figure 1 of Chapter 2, i.e., eighth order near the center and less near the boundary (see Section 3 of Chapter 2).

Remark: The continuity required by Theorem 4.1 and Theorem 4.2 is $C^{(m,0)} \cap C^{(0,n)} \cap C^{(m^*,n^*)} \supseteq C^{(m,n)}$, where $m^* = \min\{m, 2\}$ and $n^* = \min\{n, 2\}$. However, Theorem 4.3 requires continuity of $C^{(m,0)} \cap C^{(0,n)} \cap C^{(m,n)} = C^{(m,n)}$, hence, using cubic splines throughout the algorithm makes more efficient use of the available continuity.

Section 5. Bivariate Least Squares with Unstructured Data

Let $X = \{(\hat{x}_i, \hat{y}_i)\}_{i=1}^{\hat{M}} \subseteq [a,b]x[c,d]$ be a set of $\hat{M}$ unstructured data points. We will describe here several bivariate finite dimensional vector spaces from which discrete least squares fits can be calculated on the data set X. The first two spaces, bicubic splines and cubically blended cubic splines are known (see [29], p. 49 and [13], respectively). The Hermite blended piecewise polynomials are new and have some interesting properties. For this chapter, define the real number $||\cdot||_X$ of a bivariate function $f \in C^{(0,0)}([a,b]x[c,d])$ to be

$$(5.1) \qquad ||f||_X = \max_{1 \leqslant i \leqslant \hat{M}} |f(\hat{x}_i, \hat{y}_i)| .$$

Section 5.1. Bicubic Splines

Let $\pi_x: a = x_1 < x_2 < \cdots < x_M = b$ and $\pi_y: c = y_1 < y_2 < \cdots < y_N = d$ be univariate meshes, then the space of bicubic splines is defined to be $S^2(\pi_x) \circledast S^2(\pi_y)$, where $S^2(\pi_x)$ and $S^2(\pi_y)$ are cubic spline spaces of Example 1.2 of Chapter 2 (see [29, p. 49]). From the remark following (2.1) of Chapter 2, if $\{u_i\}_{i=1}^{M+2}$ and $\{w_j\}_{j=1}^{N+2}$ are any bases for $S^2(\pi_x)$ and $S^2(\pi_y)$, respectively, then $\{u_i w_j\}_{1 \leqslant i \leqslant M+2,\ 1 \leqslant j \leqslant N+2}$ is a basis for $S^2(\pi_x) \circledast S^2(\pi_y)$, and the dimension of this space is $(M+2)(N+2)$.

Let $f \in C^{(m,n)}([a,b]x[c,d])$, $1 \leqslant m, n \leqslant 4$, then $s \in S^2(\pi_x) \circledast S^2(\pi_y)$ is a discrete least squares cubic spline fit (for f on X) if it

minimizes the Euclidean norm of the residual vector $R \in \mathbb{R}^{\hat{M}}$, where component i is given by

$$R_i = f(\hat{x}_i, \hat{y}_i) - s(\hat{x}_i, \hat{y}_i), \tag{5.2}$$

for $(\hat{x}_i, \hat{y}_i) \in X$ and $1 \leqslant i \leqslant \hat{M}$.

Before deriving an error analysis, the bivariate generalization of Lemma 3.1 is given (see [6, p. 91]).

Lemma 5.1. Let Q be a bivariate polynomial of degree $\leqslant n$ in both variables, then

$$||Q|| \, (1 - n^2 \, \bar{\beta}(X)) \leqslant ||Q||_X, \text{ where} \tag{5.3}$$

$$\bar{\beta}(X) = \max_{\substack{a \leqslant x \leqslant b \\ c \leqslant y \leqslant d}} \min_{(\hat{x}_i, \hat{y}_i) \in X} \{ 2|x - \hat{x}_i|/(b-a) + 2|y - \hat{y}_i|/(c-d) \}, \tag{5.4}$$

and

$$||Q|| \, (1 - \tfrac{1}{2}(n \, \hat{\beta}(X))^2) \leqslant ||Q||_X, \text{ where} \tag{5.5}$$

$$\hat{\beta}(X) = \max_{\substack{a \leqslant x \leqslant b \\ c \leqslant y \leqslant d}} \min_{(\hat{x}_i, \hat{y}_i) \in X} \{ |\cos^{-1}((2x-(a+b))/(b-a)) - \cos^{-1}((2\hat{x}_i-(a+b))/(b-a))| + |\cos^{-1}((2y-(c+d))/(d-c)) - \cos^{-1}((2\hat{y}_i-(c+d))/(c-d))| \}. \tag{5.6}$$

Proof: The proof of (5.3) is along similar lines as the proof of (5.5) and is omitted. The proof of the univariate case of (5.5) is given in [6, pp. 91-92].

Let $(\bar{x}, \bar{y}) \in [a, b] \times [c, d]$ be such that $|Q(\bar{x}, \bar{y})| = ||Q||$, then from (5.6), there exists $(\hat{x}, \hat{y}) \in X$ such that if $\bar{\alpha} = \cos^{-1}((2\bar{x}-(a+b))/(b-a))$, $\bar{\theta} = \cos^{-1}((2\bar{y}-(c+d))/(d-c))$, $\hat{\alpha} = \cos^{-1}((2\hat{x}-(a+b))/(b-a))$ and $\hat{\theta} = \cos^{-1}((2\hat{y}-(c+d))/(d-c))$, then

$$\hat{\bar{\beta}}(X) \geq |\bar{\alpha} - \hat{\alpha}| + |\bar{\theta} - \hat{\theta}| . \tag{5.7}$$

Define the bivariate trigonometric polynomial of degree n to be $R(\alpha, \theta) = Q(((b-a)\cos\alpha + a + b)/2, ((d-c)\cos\theta + c + d)/2)$. Because $||Q|| = |R(\bar{\alpha}, \bar{\theta})| = ||R||_{[-\pi, 0] \times [-\pi, 0]} = ||R||_{[-\infty, \infty] \times [-\infty, \infty]}$ then

$$R^{(1,0)}(\bar{\alpha}, \bar{\theta}) = R^{(0,1)}(\bar{\alpha}, \bar{\theta}) = 0 , \tag{5.8}$$

and we have

$$R(\bar{\alpha}, \bar{\theta}) = \int_C R^{(1,0)}(\alpha, \theta)d\alpha + R^{(0,1)}(\alpha, \theta)d\theta + R(\hat{\alpha}, \hat{\theta}) , \tag{5.9}$$

where C is a straight line from $(\hat{\alpha}, \hat{\theta})$ to $(\bar{\alpha}, \bar{\theta})$. If C is parameterized by $t \in [0, 1]$, then

$$R(\alpha, \theta) = \int_0^1 \left[\frac{R^{(1,0)}(t(\bar{\alpha}-\hat{\alpha})+\hat{\alpha}, t(\bar{\theta}-\hat{\theta})+\hat{\theta}) - R^{(1,0)}(\bar{\alpha}, \bar{\theta})}{(t-1)} (t-1) \right. \tag{5.10}$$

$$\left. + \frac{R^{(0,1)}(t(\bar{\alpha}-\hat{\alpha})+\hat{\alpha}, t(\bar{\theta}-\hat{\theta})+\hat{\theta}) - R^{(0,1)}(\bar{\alpha}, \bar{\theta})}{(t-1)} (t-1) \right] dt$$

$$+ R(\hat{\alpha}, \hat{\theta}) .$$

By the Mean Value Theorem for each $t \in [0, 1)$, there exists a $\tilde{t} \in (t, 1)$, such that

$$(5.11) \qquad (R^{(1,0)}(t(\bar{\alpha}-\hat{\alpha})+\hat{\alpha},\ t(\bar{\theta}-\hat{\theta})+\hat{\theta}) - R^{(1,0)}(\bar{\alpha}, \bar{\theta}))/(t-1)$$

$$= R^{(2,0)}(\tilde{t}(\bar{\alpha}-\hat{\alpha})+\hat{\alpha},\ \tilde{t}(\bar{\theta}-\hat{\theta})+\hat{\theta})\ (\bar{\alpha}-\hat{\alpha})$$

$$+ R^{(1,1)}(\tilde{t}(\bar{\alpha}-\hat{\alpha})+\hat{\alpha},\ \tilde{t}(\bar{\theta}-\hat{\theta})+\hat{\theta})\ (\bar{\theta}-\hat{\theta}),$$

with a similar expression for the second term in brackets on the right hand side of (5.10). Taking absolute values and substituting (5.11) into (5.10) we have, after two applications of Bernstein's inequality (see [6, p. 91]),

$$(5.12) \qquad ||Q|| = |R(\bar{\alpha}, \bar{\theta})|$$

$$\leq ||Q|| n^2 (|\bar{\alpha}-\hat{\alpha}|+|\bar{\theta}-\hat{\theta}|)^2 \int_0^1 |t-1|dt + |R(\hat{\alpha}, \hat{\theta})|$$

$$\leq ||Q|| n^2 (|\bar{\alpha}-\hat{\alpha}|+|\bar{\theta}-\hat{\theta}|)^2/2 + ||Q||_X$$

$$\leq ||Q|| (n\hat{\bar{\beta}}(X))^2/2 + ||Q||_X,$$

where (5.7) has been used and the fact that $|R(\hat{\alpha}, \hat{\theta})| = |Q(\hat{x}, \hat{y})| \leq ||Q||_X$.

■

Define the real numbers $\bar{\beta}(X, \pi_x, \pi_y)$ and $\hat{\bar{\beta}}(X, \pi_x, \pi_y)$ with respect to the univariate meshes π_x and π_y to be

$$(5.13)\qquad \bar{\beta}(X, \pi_x, \pi_y) = \max_{\substack{1\leq i\leq M-1\\ 1\leq j\leq N-1}} \max_{\substack{x\in[x_i, x_{i+1}]\\ y\in[y_j, y_{j+1}]}} \min_{\substack{(\hat{x}_k, \hat{y}_k)\in X\\ \hat{x}_k\in[x_i, x_{i+1}]\\ \hat{y}_k\in[y_j, y_{j+1}]}}$$

$$\{2|\hat{x}_k - x|/(x_{i+1} - x_i) + 2|\hat{y}_k - y|/(y_{j+1} - y_j)\},$$

and

$$(5.14)\qquad \hat{\bar{\beta}}(X, \pi_x, \pi_y) = \max_{\substack{1\leq i\leq M-1\\ 1\leq j\leq N-1}} \max_{\substack{x\in[x_i, x_{i+1}]\\ y\in[y_j, y_{j+1}]}} \min_{\substack{(\hat{x}_k, \hat{y}_k)\in X\\ \hat{x}_k\in[x_i, x_{i+1}]\\ \hat{y}_k\in[y_j, y_{j+1}]}}$$

$$\{|\cos^{-1}((2\hat{x}_k - (x_i + x_{i+1}))/(x_{i+1} - x_i)) - \cos^{-1}((2x - (x_i + x_{i+1}))/(x_{i+1} - x_i))| + |\cos^{-1}((2\hat{y}_k - (y_{j+1} + y_j))/(y_{j+1} - y_j)) - \cos^{-1}((2y - (y_{j+1} + y_j))/(y_{j+1} - y_j))|\}.$$

<u>Theorem 5.1</u>. Let $s\in S^2(\pi_x) \circledast S^2(\pi_y)$ be a discrete least squares bicubic spline on the data set X to $f\in C^{(m,n)}([a, b]\times[c, d])$ for $1 \leq m, n \leq 4$. If $\bar{\beta}(x) < 1/9$, then

$$(5.15)\qquad ||f - s|| \leq \{\mathcal{E}_{m,0}||f^{(m,0)}||h_x^m + \mathcal{E}_{n,0}||f^{(0,n)}||h_y^n + \mathcal{E}_{m,0}\mathcal{E}_{n,0}||f^{(m,n)}||h_x^m h_y^n\}\{1 + 1/(1 - 9\bar{\beta}(X, \pi_x, \pi_y))\} + ||R||_\infty/(1 - 9\bar{\beta}(X, \pi_x, \pi_y)).$$

If $\hat{\bar{\beta}}(X, \pi_x, \pi_y) < \sqrt{2}/3$, then

$$(5.16)\qquad ||f-s|| \leqslant \{\mathcal{E}_{m,0}||f^{(m,0)}||h_x^m + \mathcal{E}_{n,0}||f^{(0,n)}||h_y^n$$

$$+ \mathcal{E}_{m,0}\,\mathcal{E}_{n,0}||f^{(m,n)}||h_x^m\,h_y^n\}\{1 + 1/(1-(3\hat{\bar{\beta}}(X, \pi_x, \pi_y))^2/2)\}$$

$$+ ||R||_\infty/(1-(3\hat{\bar{\beta}}(X, \pi_x, \pi_y))^2/2),$$

where $\mathcal{E}_{m,0}$ and $\mathcal{E}_{n,0}$ are given in Table 1 of Chapter 2 and h_x and h_y are the mesh sizes of π_x and π_y, respectively.

Proof: From Corollary 1.2 of Chapter 2

$$(5.17)\qquad ||f-s|| \leqslant ||f-P_x P_y[f]|| + ||P_x P_y[f]-s||$$

$$\leqslant \mathcal{E}_{m,0}||f^{(m,0)}||h_x^m + \mathcal{E}_{n,0}||f^{(0,n)}||h_y^n$$

$$+ \mathcal{E}_{m,0}\,\mathcal{E}_{n,0}||f^{(m,n)}||h_x^m\,h_y^n + ||P_x P_y[f]-s||,$$

where $P_x P_y[f]$ is the bicubic interpolant to f defined in (1.18) of Chapter 2. Note that $P_x P_y[f]-s$ is a bivariate cubic polynomial on each rectangle $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$ for $1 \leqslant i \leqslant M-1$ and $1 \leqslant j \leqslant N-1$, hence from Lemma 5.1

$$(5.18)\qquad |(P_x P_y[f]-s)(x, y)| \leqslant ||P_x P_y[f]-s||_X/(1-9\bar{\beta}(X, \pi_x, \pi_y)),$$

because $\bar{\beta}(X, \pi_x, \pi_y) < 1/9$. It follows that

$$(5.19)\qquad ||P_x P_y[f]-s||_X \leqslant \max_{1\leqslant i\leqslant \hat{M}} |(P_x P_y[f]-f)(\hat{x}_i, \hat{y}_i)|$$

$$+ \max_{1\leqslant i\leqslant \hat{M}} |(f-s)(\hat{x}_i, \hat{y}_i)|$$

$$\leq ||P_x P_y[f]-f|| + ||R||_\infty \ .$$

Using Corollary 1.2 of Chapter 2 and combining (5.17), (5.18) and (5.19) completes the proof.

■

Remark: Note that we have the choice of using any basis of $S^2(\pi_x)$ and $S^2(\pi_y)$. Usually, bicubic B-splines are chosen because they have the desirable property that they have support on at most sixteen adjacent rectangles, are non-negative, and easy to calculate, see [21]. This property minimizes the number of evaluations of the basis functions which must be made to construct the least squares matrix.

Section 5.2 Cubically Blended Cubic Splines.

Let $\pi_x: a = x_1 < x_2 < \cdots < x_M = b$, $\overline{\pi}_x: a = \overline{x}_1 < \overline{x}_2 < \cdots < \overline{x}_{\overline{M}} = b$, $\pi_y: c = y_1 < y_2 < \cdots < y_N = d$ and $\overline{\pi}_y: c = \overline{y}_1 < \overline{y}_2 < \cdots < \overline{y}_{\overline{N}} = d$ be univariate meshes with the corresponding cubic spline spaces $S^2(\pi_x)$, $S^2(\overline{\pi}_x)$, $S^2(\pi_y)$ and $S^2(\overline{\pi}_y)$.

From Definition 2.1 of Chapter 2, the discretized blending function space of cubically blended cubic splines, DBF_1, is the image of $C^{(1,1)}$ under the map $\overline{P_x \oplus P_y} = \overline{P}_x P_y + P_x \overline{P}_y - P_x P_y$.

$\overline{\pi}_x$ is said to be a refinement of the mesh π_x if and only if the knots of the mesh π_x are knots of the mesh $\overline{\pi}_x$. For cubic spline spaces, if $\overline{\pi}_x$ is a refinement of π_x, then $S^2(\pi_x)$ is

subordinate to $S^2(\bar{\pi}_x)$ and $S^2(\pi_x) \subseteq S^2(\bar{\pi}_x)$ (see remark following (2.29) of Chapter 2).

If the meshes $\bar{\pi}_x$ and $\bar{\pi}_y$ are refinements of π_x and π_y, respectively, then from Theorem 2.3 of Chapter 2

$$\text{Dim}(DBF_1) = (M+2)(\bar{N}+2) + (\bar{M}+2)(N+2) - (M+2)(N+2), \tag{5.20}$$

and T_1 through T_4, given in that theorem, each form a basis for DBF_1.

Finally, a bound for $||f - \overline{P_x \oplus P_y}[f]||$ is given in Example 1.4 of Chapter 2.

Let $f \in C^{(m,n)}([a,b]x[c,d])$, $1 \leq m, n \leq 4$, then $s \in DBF_1$ is a discrete least squares solution on the data set X if the Euclidean norm of the residual vector R defined in (5.2) is minimized. Combining the above observations and Lemma 5.1 yields the following theorem.

Theorem 5.2. Let $s \in DBF_1$ be a discrete least squares solution on the data set X to $f \in C^{(m,n)}([a,b]x[c,d])$ for $1 \leq m, n \leq 4$. If $\bar{\beta}(X, \bar{\pi}_x, \bar{\pi}_y) < 1/9$, then

$$||f-s|| \leq \{ \mathcal{E}_{m,0} ||f^{(m,0)}|| \bar{h}_x + \mathcal{E}_{n,0} ||f^{(0,n)}|| \bar{h}_y \tag{5.21}$$

$$+ \mathcal{E}_{m,0}\mathcal{E}_{n,0} ||f^{(m,n)}|| h_x^m h_y^n + \mathcal{E}_{m,0}\mathcal{E}_{n,0} ||f^{(m,n)}|| \bar{h}_x^m h_y^n$$

$$+ \mathcal{E}_{m,0}\,\mathcal{E}_{n,0}||f^{(m,n)}||h_x^m \bar{h}_y^n\}(1+1/(1-9\bar{\beta}(X,\bar{\pi}_x,\bar{\pi}_y)))$$

$$+ ||R||_\infty/(1-9\bar{\beta}(X,\bar{\pi}_x,\bar{\pi}_y)) .$$

If $\hat{\bar{\beta}}(X,\bar{\pi}_x,\bar{\pi}_y) < \sqrt{2}/3$, then

$$||f-s|| \leq \{\mathcal{E}_{m,0}||f^{(m,0)}||\bar{h}_x + \mathcal{E}_{n,0}||f^{(0,n)}||\bar{h}_y \tag{5.22}$$

$$+ \mathcal{E}_{m,0}\,\mathcal{E}_{n,0}||f^{(m,n)}||h_x^m h_y^n + \mathcal{E}_{m,0}\,\mathcal{E}_{n,0}||f^{(m,n)}||\bar{h}_x^m h_y^n$$

$$+ \mathcal{E}_{m,0}\,\mathcal{E}_{n,0}||f^{(m,n)}||h_x^m \bar{h}_y^n\}(1+1/(1-(3\hat{\bar{\beta}}(X,\bar{\pi}_x,\bar{\pi}_y))^2/2))$$

$$+ ||R||_\infty/(1-(3\hat{\bar{\beta}}(X,\bar{\pi}_x,\bar{\pi}_y))^2/2) ,$$

where $\mathcal{E}_{m,0}$ and $\mathcal{E}_{n,0}$ are given in Table 1 of Chapter 2, h_x, $\bar{h}_x$, h_y and $\bar{h}_y$ are the mesh sizes of π_x, $\bar{\pi}_x$, π_y and $\bar{\pi}_y$, respectively.

Proof: Identical to Theorem 5.1, hence omitted.

■

Section 5.3. Hermite Blended Piecewise Polynomials.

Let $\pi_x: a = x_1 = x_2 < x_3 = x_4 < \cdots < x_{2M-1} = x_{2M} = b$ and $\pi_y: c = y_1 = y_2 < y_3 = y_4 < \cdots < y_{2N-1} = y_{2N} = d$ be univariate meshes and $V(\pi_x, 2M, 1)$ and $V(\pi_y, 2N, 1)$ be the interpolation spaces of cubic Hermite splines of Example 1.3 of Chapter 2. Let $\{\phi_i\}_{i=1}^{2M}$ and $\{\psi_j\}_{j=1}^{2N}$ be the cardinal bases (see [29, pp. 25-27] for explicit representation), α and β be the interpolation functions

for the interpolation spaces $V(\pi_x, 2M, 1)$, and $V(\pi_y, 2N, 1)$, respectively, given in Example 1.3 of Chapter 2.

If $g \in C^{(4)}[a, b]$, then

$$(5.23) \qquad \left|\left|g - \sum_{i=1}^{2M} g^{(\alpha(i))}(x_i)\, \phi_i\right|\right| \leqslant (1/384)\left|\left|g^{(4)}\right|\right| h_x^4 ,$$

with a corresponding error bound for the interpolation space $V(\pi_y, 2N, 1)$, see Carlson and Hall [5].

Refine the mesh π_x by adding $k-2$ additional points between each of the knots x_{2i} and x_{2i+1}, for $1 \leqslant i \leqslant M-1$, to obtain $\bar{\pi}_x: a = \bar{x}_1 = \bar{x}_2 < \bar{x}_3 < \cdots < \bar{x}_k < \bar{x}_{k+1} = \bar{x}_{k+2} < \bar{x}_{k+3} < \cdots < \bar{x}_{(M-1)k} < \bar{x}_{(M-1)k+1} = \bar{x}_{(M-1)k+2} = b$, where we have the correspondence for $1 \leqslant i \leqslant M$

$$(5.24) \qquad x_{2i-1} = \bar{x}_{(i-1)k+1} \quad \text{and} \quad x_{2i} = \bar{x}_{(i-1)k+2} .$$

Define the interpolation function

$$(5.25) \qquad \bar{\alpha}(i) = \begin{cases} 1 & \text{if } (i-2) \bmod k = 0 \\ 0 & \text{otherwise} \end{cases} .$$

$\bar{V}(\bar{\pi}_x, (M-1)k+2, 1)$ is the interpolation space of continuously differentiable functions such that on the interval $[x_{ik+1}, x_{(i+1)k+2})]$ each element is a polynomial of degree $k+1$ for $0 \leqslant i \leqslant M-2$. The cardinal basis for this space is represented by $\{\bar{\phi}_i\}_{i=1}^{(M-1)k+2}$.

By the manner in which the new points have been added we have that $V(\pi_x, 2M, 1)$ is subordinate to $\bar{V}(\bar{\pi}_x, (M-1)k+2, 1)$ and $V(\pi_x, 2M, 1) \subseteq \bar{V}(\bar{\pi}_x, (M-1)k+2, 1)$.

Correspondingly, refine the mesh π_y by adding ℓ-2 points between each of the points y_{2j} and y_{2j+1} of π_y for $1 \leqslant j \leqslant N-1$ to obtain the mesh $\bar{\pi}_y: c=\bar{y}_1 = \bar{y}_2 < \bar{y}_3 < \cdots < \bar{y}_\ell < \bar{y}_{\ell+1} = \bar{y}_{\ell+2} < \cdots < \bar{y}_{(N-1)\ell} < \bar{y}_{(N-1)\ell+1} = \bar{y}_{(N-1)\ell+2} = d$. Construct the interpolation space of piecewise polynomials of degree $\ell+1$, $\bar{V}(\bar{\pi}_y, (N-1)\ell+2, 1)$, with interpolation function $\bar{\beta}$ and cardinal basis $\{\bar{\psi}\}_{j=1}^{(N-1)\ell+2}$ in an analogous manner. Then $V(\pi_y, 2N, 1)$ is subordinate to $\bar{V}(\bar{\pi}_y, (N-1)\ell+2, 1)$ and $V(\pi_y, 2N, 1) \subseteq \bar{V}(\bar{\pi}_y, (N-1)\ell+2, 1)$.

Construct the discretized blending function space DBF_2; then, from Theorem 2.3 of Chapter 2,

$$\text{(5.26)} \qquad \text{Dim}(DBF_2) = 2M\left[(N-1)\ell+2\right] + \left[2N\ (M-1)k+2\right] - 4MN,$$

and each of the sets T_1 through T_4 forms a basis for DBF_2.

We will examine the basis T_4 in detail. If $IM = \{i \mid i=(s-1)k+1$ or $i = (s-1)k+2,\ 1 \leqslant s \leqslant M\}$, $\overline{IM} = \{i \mid 1 \leqslant i \leqslant (M-1)k+2$ and $i \notin IM\}$, $JN = \{j \mid j=(s-1)\ell+1$ or $j = (s-1)\ell+2,\ 1 \leqslant s \leqslant N\}$ and $\overline{JN} = \{j \mid 1 \leqslant j \leqslant (N-1)\ell+2$ and $j \notin JN\}$, then $T_4 = S_1 \cup S_2 \cup S_6$, where $S_1 = \{\bar{\phi}_i \psi_j\}_{i \in \overline{IM},\ 1 \leqslant j \leqslant 2N}$, $S_2 = \{\phi_i \bar{\psi}_j\}_{1 \leqslant i \leqslant 2M,\ j \in \overline{JN}}$ and $S_6 = \{\phi_i \psi_j\}_{1 \leqslant i \leqslant 2M,\ 1 \leqslant j \leqslant 2N}$.

For $\bar{\phi}_i \psi_j \in S_1$, let $\hat{i} = (2((i-1)-(i-1) \bmod k)/k)+2$ and $\hat{j} = j-(j+1) \bmod 2$, then $\bar{\phi}_i \psi_j$ has support on the rectangle $[x_{\hat{i}}, x_{\hat{i}+1}] \times [y_{\hat{j}-1}, y_{\hat{j}+2}]$, where we restrict ourselves to the domain $[a, b] \times [c, d]$ (recall that $y_{\hat{j}} = y_{\hat{j}+1}$). Correspondingly for $\phi_i \bar{\psi}_j \in S_2$, if

$\hat{j} = (2((j-1)-(j-1) \bmod \ell)/\ell) + 2$ and $\hat{i} = i-(i+1) \bmod 2$, then $\phi_i \bar{\psi}_j$ has support on $[x_{\hat{i}-1}, x_{\hat{i}+2}] \times [y_{\hat{j}}, y_{\hat{j}+1}]$, where $x_{\hat{i}} = x_{\hat{i}+1}$ and we restrict ourselves to $[a, b]\times[c, d]$. Finally, for $\phi_i \psi_j \in S_6$, if $\hat{i} = i-(i+1) \bmod 2$ and $\hat{j} = j-(j+1) \bmod 2$, then $\phi_i\psi_j$ has support on $[x_{\hat{i}-1}, x_{\hat{i}+2}]\times[y_{\hat{j}-1}, y_{\hat{j}+2}]$. Paraphrasing the above, $\bar{\phi}_i\psi_j$ and $\phi_i\bar{\psi}_j$ have support on at most two adjacent rectangles for the meshes π_x and π_y, while the support of $\phi_i\psi_j$ is over at most four adjacent rectangles.

The cubic Hermite bases $\{\phi_i\}_{i=1}^{2M}$ and $\{\psi_j\}_{j=1}^{2N}$ are easy to calculate and can be stored in the computer in polynomial form. Actually, only two basis functions need be stored, as the others can be derived from these two (for explicit representation see [29, pp. 25-27]). Correspondingly, the basis of piecewise polynomials $\{\bar{\phi}_i\}_{i=1}^{(M-1)k+2}$ of degree k+1, and the basis of piecewise polynomials $\{\bar{\psi}_j\}_{j=1}^{(N-1)\ell+2}$ of degree $\ell+1$ are easy to calculate and can be stored in the computer in polynomial form. Again, for similar reasons, only k+2 or $\ell+2$ polynomials need be stored as the others can be easily generated from these.

Note that the above basis circumvents many of the difficulties associated with the implementation of discretized blending function spaces, hence the main reason for its introduction.

It is a simple matter to calculate the interpolation accuracy of the interpolation spaces $\bar{V}(\bar{\pi}_x, (M-1)k+2, 1)$ (correspondingly, for

$\overline{V}(\overline{\pi}_y, (N-1)\ell+2, 1))$. For $x\epsilon[\overline{x}_{sk+1}, \overline{x}_{(s+1)k+2}]$ and $g \epsilon C^{(k+2)}[a,b]$

$$(5.27) \qquad g(x) - \sum_{i=1}^{(M-1)k+2} g^{(\overline{\alpha}(i))}(\overline{x}_i)\ \overline{\phi}_i(x)$$

$$= \prod_{i=1}^{k+2} (x-\overline{x}_{sk+1})\ g^{(k+2)}\ \xi(x))/(k+2)!\ ,$$

where $\xi(x)\epsilon [\overline{x}_{sk+1}, \overline{x}_{(s+1)k+2}]$, see [33, pp. 1-5].

Hence

$$(5.28) \qquad ||g - \sum_{i=1}^{(M-1)k+2} g^{(\overline{\alpha}(i))}(\overline{x}_i)\ \overline{\phi}_i||$$

$$\leqslant \prod_{i=1}^{k+2} |x-\overline{x}_{sk+i}|\ ||g^{(k+2)}||/(k+2)!$$

$$\leqslant ||g^{(k+2)}||h_x^{k+2}/(k+2)!\ .$$

Remark: Note that $\max_{0\leqslant s\leqslant M-2} (\overline{x}_{(s+1)k+2} - \overline{x}_{sk+1})$

$= \max_{0\leqslant s\leqslant M-2} (x_{2(s+1)+1} - x_{2s+1}) = h_x$, the mesh size of π_x.

Remark: If k is fixed and the spacing of the points x_{sk+1} for $3 \leqslant i \leqslant k$ is also fixed, then a more refined estimate can be given for this term. For example, if $k = 6$, and $\overline{x}_{sk+i}$ are equally spaced for $3 \leqslant i \leqslant k$, then $\prod_{i=1}^{8} |x-\overline{x}_{6s+i}| < h_x^8\ 24/5^6$.

Combining the above and using Theorem 1.7 of Chapter 2, we have the following error estimate for $f \in C^{(k+2, \ell+2)}([a, b]x[c, d])$ (note that k and $\ell \geq 2$),

$$(5.29) \quad ||f-\overline{P_x \oplus P_y}[f]|| \leq (1/(k+2)!)||f^{(k+2, 0)}||h_x^{k+2}$$

$$+(1/(\ell+2)!)||f^{(0, \ell+2)}||h_y^{\ell+2}+(1/384)^2||f^{(4, 4)}||h_x^4 h_y^4$$

$$+(1/(384\cdot(k+2)!))||f^{(k+2, 4)}||h_x^{k+2} h_y^4+(1/(384(\ell+2)!))$$

$$\cdot ||f^{(4, \ell+2)}||h_x^4 h_y^{\ell+2} \quad .$$

The limiting accuracy of the above interpolation scheme is $h_x^4 h_y^4$, thus we choose ℓ and k sufficiently large to insure that h_x^{k+2}, $h_y^{\ell+2} \leq h_x^4 h_y^4$. For example, if $h = \max(h_x, h_y)$ then this would imply that in terms of h, the choice of $\ell = k = 6$ would suffice.

Remark: If $h = \max(h_x, h_y)$, and we assume sufficient continuity of f, then both of the discretized blending function spaces, cubically blended cubic splines = DBF_1, and cubic Hermite blended piecewise polynomials = DBF_2, give eighth order approximations to f. For large M and N, examination of (5.20) and (5.26) shows that Dim (DBF_1) is much larger than Dim (DBF_2). Thus, there will be a corresponding savings of computer memory needed to store the least squares matrix using DBF_2. However, in order to obtain this savings, the continuity required to implement DBF_2 must be larger than DBF_1.

Let $X = \{(\hat{x}_i, \hat{y}_i)\}_{i=1}^{\hat{M}} \subseteq [a, b]x[c, d]$ be a set of $\hat{M}$ unstructured data points. Let $f \in C^{(k+2, \ell+2)}([a, b]x[c, d])$, then a discrete least squares solution $s \in DBF_2$ is a function which minimizes the Euclidean norm of the residual vector $R \in \mathbb{R}^{\hat{M}}$ defined in (5.2).

Theorem 5.3. Let $s \in DBF_2$ be a discrete least squares solution on the data set X to $f \in C^{(k+2, \ell+2)}([a, b]x[c, d])$. If $\bar{\beta}(X, \pi_x, \pi_y) < 1/t^2$, then

$$(5.30)\quad ||f-s|| \leq \{(1/(k+2)!)||f^{(k+2, 0)}||h_x^{k+2}$$
$$+ (1/(\ell+2)!)||f^{(0, \ell+2)}||h_y^{\ell+2} + (1/384)^2||f^{(4, 4)}|| h_x^4 h_y^4$$
$$+ (1/(384\cdot(k+2)!))||f^{(k+2, 4)}||h_x^{k+2} h_y^4$$
$$+ (1/(384\cdot(\ell+2)!))||f^{(4, \ell+2)}||h_x^4 h_y^{\ell+2}\}\{1-1/(1-$$
$$t^2\bar{\beta}(X, \pi_x, \pi_y))\} + ||R||_\infty/(1-t^2\bar{\beta}(X, \pi_x, \pi_y)) .$$

If $\hat{\bar{\beta}}(X, \pi_x, \pi_y) < \sqrt{2}/t$, then

$$(5.31)\quad ||f-s|| \leq \{(1/(k+2)!)||f^{(k+2, 0)}|| h_x^{k+2}$$
$$+ (1/(\ell+2)!)||f^{(0, \ell+2)}||h_y^{\ell+2} + (1/384)^2||f^{(4, 4)}|| h_x^4 h_y^4$$
$$+ (1/(384\cdot(k+2)!))||f^{(k+2, 4)}||h_x^{k+2} h_y^4$$
$$+ (1/(384\cdot(\ell+2)!))||f^{(4, \ell+2)}||h_x^4 h_y^{\ell+2}\}\{1-1/(1-(t\hat{\bar{\beta}}(X, \pi_x, \pi_y))^2/2)\}$$
$$+ ||R||_\infty/(1-(t\hat{\bar{\beta}}(X, \pi_x, \pi_y))^2/2) ,$$

where $k, \ell \geq 2$ and $t = \max\{k+1, \ell+1\}$.

Proof: Similar to Theorem 5.1, hence omitted.

■

Section 5.4. Linear Blending

As a final example, we shall consider the discretized blending function space of linearly blended piecewise cubic polynomials.

Let $\pi_x: a = x_1 < x_2 < \cdots < x_M = b$ and $\pi_y: c = y_1 < y_2 < \cdots < y_N = d$ be univariate meshes. The interpolation spaces of piecewise linear continuous functions $V(\pi_x, M, 0)$ and $V(\pi_y, N, 0)$ were defined in Section 4, where the basis functions $\{\phi_i\}_{i=1}^{M}$ and $\{\psi_j\}_{j=1}^{N}$ were given by (4.1) and the error estimate by (4.2).

Refine the meshes π_x and π_y by adding two knots between each of the knots of π_x and π_y to obtain $\bar{\pi}_x: a = \bar{x}_1 < \bar{x}_2 < \bar{x}_3 < \cdots < \bar{x}_{3M-2} = b$ and $\bar{\pi}_y: c = \bar{y}_1 < \bar{y}_2 < \bar{y}_3 < \cdots < \bar{y}_{3N-2} = d$, respectively, where

$$(5.32) \quad x_i = \bar{x}_{3(i-1)+1} \text{ for } 1 \leqslant i \leqslant M, \text{ and } y_j = y_{3(j-1)+1} \text{ for } 1 \leqslant j \leqslant N.$$

Define the interpolation space of piecewise continuous cubic polynomials $\bar{V}(\bar{\pi}_x, 3M-2, 0)$ (respectively, $\bar{V}(\bar{\pi}_y, 3N-2, 0)$) with interpolation function $\bar{\alpha}(i) = 0$ for $1 \leqslant i \leqslant 3M-2$ ($\bar{\beta}(j) = 0$ for $1 \leqslant j \leqslant 3N-2$) and cardinal basis of piecewise continuous cubic polynomials $\{\bar{\phi}_i\}_{i=1}^{3M-2}$ ($\{\bar{\psi}_j\}_{j=1}^{3N-2}$). For a fixed i, such that $1 \leqslant i \leqslant 3M-2$, let $s = ((i-1)-(i-1) \bmod 3)/3$ and $k = i - 3s$, if $s \leqslant M-2$, then $\bar{\phi}_i$ is the cubic

Lagrange interpolation polynomial on the four points $\{\bar{x}_{3s+t}\}_{t=1}^{4}$ such that $\bar{\phi}_i(\bar{x}_{3s+t}) = \delta_{kt}$. If $k = 2, 3$, then $\bar{\phi}_i$ is identically zero off the interval $[\bar{x}_{3s+1}, \bar{x}_{3(s+1)+1}] = [x_{s+1}, x_{s+2}]$. If $k = 1$ and $i > 1$, or $s = M-1$, then $\bar{\phi}_i$ is also the cubic Lagrange interpolation polynomial on the four points $\{\bar{x}_{3(s-1)+t}\}_{t=1}^{4}$ such that $\bar{\phi}_i(\bar{x}_{3(s-1)+t}) = \delta_{t,4}$ and identically zero off the interval $[x_s, x_{s+2}]$. Thus, the support of $\{\phi_i\}_{i=1}^{M}$ and $\{\bar{\phi}\}_{i=1}^{3M-2}$ ($\{\psi_j\}_{j=1}^{N}$ and $\{\bar{\psi}_j\}_{j=1}^{3N-2}$ consists of at most two adjacent intervals of the mesh π_x (π_y).

Note that $V(\pi_x, M, 0)$ is subordinate to $\bar{V}(\bar{\pi}_x, 3M-2, 0)$ and $V(\pi_x, M, 0) \subseteq \bar{V}(\bar{\pi}_x, 3M-2, 0)$ with corresponding relations holding for $V(\pi_y, N, 0)$ and $\bar{V}(\bar{\pi}_y, 3N-2, 0)$. Theorem 2.3 of Chapter 2 gives the dimension of the discretized blending function space of linearly blended piecewise cubic polynomials = DBF_3 as

$$\text{Dim}\,(DBF_3) = 5MN - 2(M+N), \tag{5.33}$$

where T_1 through T_4 each forms a basis for DBF_3. Note that the cardinal basis elements ϕ_i, $\bar{\phi}_i$, ψ_j and $\bar{\psi}_j$ are particularly simple, easy to store and compute, with support over at most two adjacent intervals of the meshes π_x or π_y. Hence, the basis elements for DBF_3 will share these desirable properties of being very simple, easy to store and compute, with support over at most four adjacent rectangles.

On the interval $[x_i, x_{i+1}]$, the interpolation error of $\bar{V}(\bar{\pi}_x, 3M-2, 0)$ is just the difference of the cubic Lagrange

interpolation polynomial and a function $g \in C^{(4)}[a, b]$. Thus, on this interval, we have from [24, p. 249]

$$(5.34) \quad |g(x) - \sum_{k=1}^{3M-2} g(\bar{x}_k)\, \bar{\phi}_k(x)| \leq \prod_{s=1}^{4} |x - \bar{x}_{3(i-1)+s}|\, ||g^{(4)}||/4! \leq K h_x^4\, ||g^{(4)}||/4!,$$

where K is a constant less than one.

Remark: If structure is given to the four data points $\{x_{3(i-1)+s}\}_{s=1}^{4}$, then various estimates for K can be given. For example, if the points are equally spaced in each interval $[x_i, x_{i+1}]$, then direct calculation shows that $K = 1/81$. If the points are the zeros of the cubic Chebyshev polynomial (see [24, pp. 228] for definition, explanation and bound), then $K = 1/128$, which is the best possible for any distribution of the four points in each interval.

From Theorem 1.7 of Chapter 2, for $f \in C^{(4,4)}([a, b] \times [c, d])$

$$(5.35) \quad ||f - \overline{P_x \oplus P_y}[f]|| \leq (K/24)||f^{(4,0)}||\, h_x^4 + (K/24)||f^{(0,4)}||h_y^4 + (1/64)||f^{(2,2)}||\, h_x^2 h_y^2 + (K/192)||f^{(4,2)}||h_x^4 h_y^2 + (K/192)||f^{(2,4)}||h_x^2 h_y^4,$$

where $K < 1$ is some constant, which is defined above.

Let $f \in C^{(4,4)}([a, b] \times [c, d])$, then a discrete least squares solution $s \in DBF_3$ is a function which minimizes the Euclidean norm of the residual vector $R \in \mathbb{R}^{\hat{M}}$ defined in (5.2) on the data set X.

Theorem 5.4. Let $s \in DBF_3$ be a discrete least squares solution to $f \in C^{(4,4)}([a,b]x[c,d])$ on the data set X. If $\bar{\beta}(X, \pi_x, \pi_y) < 1/9$, then

$$(5.36)\quad ||f-s|| \leq \{(K/24)||f^{(4,0)}||h_x^4 + (K/24)||f^{(0,4)}||h_y^4$$
$$+ (1/64)||f^{(2,2)}||h_x^2 h_y^2 + (K/192)||f^{(4,2)}|| h_x^4 h_y^2$$
$$+ (K/192)||f^{(2,4)}|| h_x^2 h_y^4\}\{1-1/(1-9\bar{\beta}(X, \pi_x, \pi_y))\}$$
$$+ ||R||_\infty /(1-9\bar{\beta}(X, \pi_x, \pi_y)) .$$

If $\hat{\bar{\beta}}(X, \pi_x, \pi_y) < \sqrt{2}/3$, then

$$(5.37)\quad ||f-s|| \leq \{(K/24)||f^{(4,0)}||h_x^4 + (K/24)||f^{(0,4)}||h_y^4$$
$$+ (1/64)||f^{(2,2)}||h_x^2 h_y^2 + (K/192)||f^{(4,2)}||h_x^4 h_y^2$$
$$+ (K/192)||f^{(2,4)}||h_x^2 h_y^4\}\{1-1/(1-(3\hat{\bar{\beta}}(X, \pi_x, \pi_y))^2/2)\}$$
$$+ ||R||_\infty /(1-(3\hat{\bar{\beta}}(X, \pi_x, \pi_y))^2/2) ,$$

where $K < 1$ is some constant, depending upon the meshes $\bar{\pi}_x$ and $\bar{\pi}_y$, and h_x and h_y are the mesh sizes of π_x and π_y .

Proof: Similar to the proof of Theorem 5.1.

■

A procedure is now presented which minimizes the computer storage needed to compute a discrete least squares solution. Note

that for the discretized blending function spaces considered, even for relatively small M and N, a large matrix must be stored in the computer, with a correspondingly large number of computational operations needed to obtain the numerical solution. The procedure given here is to solve a separate discrete least squares problem on each rectangle $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$ for $1 \leqslant i \leqslant M-1$ and $1 \leqslant j \leqslant N-1$. Restricted to this rectangle, the discretized blending function space of linearly blended piecewise continuous cubic polynomials will be denoted by DBF_{ij}, where it is clear from the previous notation what the meshes, interpolation spaces, and their cardinal basis functions will be. From (5.33), it follows that $\text{Dim}(DBF_{ij}) = 12$, with a basis given by T_1 through T_4 of Theorem 2.3 of Chapter 2. If $X_{ij} = X \cap ([x_i, x_{i+1}] \times [y_j, y_{j+1}])$ for $1 \leqslant i \leqslant M-1$ and $1 \leqslant j \leqslant N-1$ represents the $\hat{M}_{ij}$ data points which are in each of the rectangular regions (note that a data point could be in more than one X_{ij} if it lies on a mesh line of the bivariate mesh $\pi_x \oplus \pi_y$), then the matrix problem to be stored at any one time has only $12\hat{M}_{ij}$ elements. This is significantly less storage than that required by any of the previous methods. If $\hat{M}_{ij}$ is not too large, then most computers, even those having a very small memory capacity, can store the full least squares matrix in central memory.

Let $f \in C^{(4,4)}([x_i, x_{i+1}] \times [y_j, y_{j+1}])$ and let $s_{ij} \in DBF_{ij}$ be a discrete least squares solution which minimizes the residual vector $R_{ij} \in \mathbb{R}^{\hat{M}_{ij}}$, where each of the $\hat{M}_{ij}$ elements of R_{ij} is the difference of f and s_{ij} evaluated at one of the $\hat{M}_{ij}$ data points of X_{ij}. Theorem 5.4 gives an error estimate for $||f-s_{ij}||$ on $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$. The set $\{s_{ij}\}_{1 \leq i \leq M-1,\ 1 \leq j \leq N-1}$ forms a "patch network" of discrete least squares functions over the full domain $[a, b] \times [c, d]$. However, this "patch network" is not necessarily continuous across the mesh lines of $\pi_x \oplus \pi_y$. In order to remedy this situation and obtain an approximation to f which has global continuity on $[a, b] \times [c, d]$, the following scheme is introduced which produces an approximation $\tilde{s} \in DBF_3$ to f.

If T_4 of Theorem 2.3 of Chapter 2 is chosen as a basis for DBF_{ij} on $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$, then

$$(5.38)\quad T_4 = \{\bar{\phi}_s^{ij}\ \psi_t^{ij}\}_{2 \leq s \leq 3,\ 1 \leq t \leq 2} \cup \{\phi_s^{ij}\bar{\psi}_t^{ij}\}_{1 \leq s \leq 2,\ 2 \leq t \leq 3} \cup \{\phi_s^{ij}\ \psi_t^{ij}\}_{1 \leq s, t \leq 2},$$

and

$$(5.39)\quad s_{ij} = \sum_{s=2}^{3}\sum_{t=1}^{2} a_{st}^{ij}\,\bar{\phi}_s^{ij}\,\psi_t^{ij} + \sum_{s=1}^{2}\sum_{t=2}^{3} b_{st}^{ij}\,\phi_s^{ij}\,\bar{\psi}_t^{ij} + \sum_{s=1}^{2}\sum_{t=1}^{2} c_{st}^{ij}\,\phi_s^{ij}\,\psi_t^{ij}\ .$$

For each fixed i and j, let $0 \leqslant u, v \leqslant 1$, and define the average of the "a", "b" and "c" coefficients to be

$$(5.40)\quad \tilde{a}^{ij}_{2+u,\,1+v} = \sum_{t=0}^{1} a^{i,\,j+v-t}_{2+u,\,1+t} \Big/ DA(i, j, u, v)\,,$$

$$(5.41)\quad \tilde{b}^{ij}_{1+u,\,2+v} = \sum_{s=0}^{1} b^{i+u-s,\,j}_{1+s,\,2+v} \Big/ DB(i, j, u, v)\,,$$

and

$$(5.42)\quad \tilde{c}^{ij}_{1+u,\,1+v} = \sum_{s=0}^{1}\sum_{t=0}^{1} c^{i+u-s,\,j+v-t}_{1+s,\,1+t} \Big/ DC(i, j, u, v)\,.$$

where the indices are restricted to insure that $1 \leqslant i, i+u-s \leqslant M-1$ and $1 \leqslant j, j+v-t \leqslant N-1$. The $DA(i, j, u, v) = 1$ or 2, $DB(i, j, u, v) = 1$ or 2 and $DC(i, j, u, v) = 1$, 2 or 4 are just the total number of terms which have been summed in (5.40), (5.41) and (5.42), respectively. For example, $DC(1, 1, 0, 0) = 1$, and if $M, N > 2$, then $DC(1, 1, 1, 0) = 2$ and $DC(1, 1, 1, 1) = 4$, etc.

Define

$$(5.43)\quad \tilde{s}_{ij} = \sum_{s=2}^{3}\sum_{t=1}^{2} \tilde{a}^{ij}_{st}\, \bar{\phi}^{ij}_{st}\, \psi^{ij}_{t} + \sum_{s=1}^{2}\sum_{t=2}^{3} \tilde{b}^{ij}_{st}\, \phi^{ij}_{s}\, \bar{\psi}^{ij}_{t} + \sum_{s=1}^{2}\sum_{t=1}^{2} \tilde{c}^{ij}_{st}\, \phi^{ij}_{s}\, \psi^{ij}_{t}\,,$$

to be that element in DBF_{ij} which has as its coefficients the average "a's", "b's" and "c's". Correspondingly, define $\tilde{s} \in DBF_3$ by

$$(5.44)\quad \tilde{s} = \sum_{i=1}^{M-1}\sum_{s=2}^{3}\sum_{j=1}^{N} \tilde{a}_{s1}^{ij}\ \bar{\phi}_{3(i-1)+s}\ \psi_j$$

$$+ \sum_{i=1}^{M}\sum_{j=1}^{N-1}\sum_{t=2}^{3} \tilde{b}_{1t}^{ij}\ \phi_i\ \bar{\psi}_{3(j-1)+t} + \sum_{i=1}^{M}\sum_{j=1}^{N} \tilde{c}_{1,1}^{ij}\ \phi_i\ \psi_j\ ,$$

where we define $\tilde{a}_{s,1}^{i,N} = \tilde{a}_{s,2}^{i,N-1}$, $\tilde{b}_{1,t}^{M,j} = \tilde{b}_{2,t}^{M-1,j}$,

$\tilde{c}_{1,1}^{i,N} = \tilde{c}_{1,2}^{i,N-1}$, $\tilde{c}_{1,1}^{M,j} = \tilde{c}_{2,1}^{M-1,j}$ and $\tilde{c}_{1,1}^{M,N} = \tilde{c}_{2,2}^{M-1,N-1}$.

Let the real number $\mathcal{E}>0$ be such that

$$(5.45)\quad \max_{\substack{1\leqslant i\leqslant M-1\\ 1\leqslant j\leqslant N-1}} ||f-s_{ij}||_{[x_i, x_{i+1}]\times[y_j, y_{j+1}]} \leqslant \mathcal{E}\ ,$$

then we have the following theorem.

Theorem 5.5. Let $\tilde{s} \in DBF_3$, defined in (5.44), be constructed as above, then

$$(5.46)\quad ||f-\tilde{s}|| \leqslant \{5/2 + 2\{ \max_{1\leqslant i\leqslant M-1} \sum_{s=2}^{3} ||\bar{\phi}_{3(i-1)+s}||$$

$$+ \max_{1\leqslant j\leqslant N-1} \sum_{t=2}^{3} ||\bar{\psi}_{3(j-1)+t}||\}\ \mathcal{E},$$

where $\mathcal{E}$ satisfies (5.45).

Proof: It will be shown that for $1 \leqslant i \leqslant M-1$, $1 \leqslant j \leqslant N-1$, and $1 \leqslant s, t \leqslant 2$ that

$$(5.47)\quad |\tilde{c}_{st}^{ij} - c_{st}^{ij}| \leqslant (3/2)\ \mathcal{E}\ .$$

Because of the cardinality conditions, (5.39) and (5.45), we have for $1 \leq i \leq M-1$, $1 \leq j \leq N-1$, $1 \leq s, t \leq 2$

$$(5.48) \quad |c^{ij}_{st} - f(x_{i+s-1}, y_{j+t-1})| \leq \mathcal{E}.$$

Using (5.48), (5.42), the triangle inequality, and observing the internal cancellation of one of the terms yields (5.47) (notice that if $DC(i, j, s, t) = k$ for $k = 1, 2$ then $|\tilde{c}^{ij}_{st} - c^{ij}_{st}| \leq (k-1)\mathcal{E}$). Also, for $2 \leq s \leq 3$ and $1 \leq t \leq 2$ we have

$$(5.49) \quad |\tilde{a}^{ij}_{st} - a^{ij}_{st}| \leq 2\mathcal{E} \text{ and } |\tilde{b}^{ij}_{ts} - b^{ij}_{ts}| \leq 2\mathcal{E}.$$

The proof is given only for the first inequality, as the other is nearly identical. From the cardinality conditions, (5.39) and (5.45), we have

$$(5.50) \quad |a^{ij}_{st} + \sum_{k=1}^{2} c^{ij}_{kt}\, \phi^{ij}_{k}(\bar{x}_{3(i-1)+s}) - f(\bar{x}_{3(i-1)+s}, y_{j+t-1})| \leq \mathcal{E},$$

for all $1 \leq i \leq M-1$, $1 \leq j \leq N-1$, $2 \leq s \leq 3$ and $1 \leq t \leq 2$.

Consider the case where $DA(i, j, u, v) = 2$ (as the case where $DA(i, j, u, v) = 1$ is trivial), and without loss of generality let $t = 1$. Then from (5.40), (5.49) and the triangle inequality we have the following after some algebra and cancellation

$$(5.51) \quad |\tilde{a}^{ij}_{s1} - a^{ij}_{s1}| = \tfrac{1}{2}|a^{i,j-1}_{s,2} - a^{i,j}_{s,1}|$$

$$\leq \tfrac{1}{2}\left[\mathcal{E} + \mathcal{E} + \sum_{k=1}^{2} |c^{i,j-1}_{k,2} - c^{i,j}_{k,1}|\, \phi^{i,j}_{k}(\bar{x}_{3(i-1)+s})\right]$$

$$\leqslant \varepsilon + \tfrac{1}{2} \max_{k=1,2} \left| c_{k,2}^{i,j-1} - c_{k,1}^{i,j} \right|$$

$$\leqslant 2\,\varepsilon,$$

where the second expression was obtained using the fact that $\phi_k^{ij} = \phi_k^{i,j-1} \geqslant 0$, the third from the fact that $\sum_{k=1}^{2} \phi_k^{ij} \equiv 1$, and the last from the triangle inequality and (5.48).

From (5.39), (5.43), (5.45), (5.49), (5.47) and $(x, y) \in [x_i, x_{i+1}] \times [y_j, y_{j+1}]$

$$(5.52) \quad \left|(f-\tilde{s}_{ij})(x, y)\right| \leqslant \left|(f-s_{ij})(x, y)\right| + \left|(s_{ij} - \tilde{s}_{ij})(x, y)\right|$$

$$\leqslant \varepsilon + \sum_{s=2}^{3} \sum_{t=1}^{2} 2\varepsilon \left|\bar{\phi}_s^{ij}(x)\right| \psi_t^{ij}(y)$$

$$+ \sum_{s=1}^{2} \sum_{t=2}^{3} 2\varepsilon \left|\bar{\psi}_t^{ij}(y)\right| \phi_s^{ij}(x) + \sum_{s=1}^{2} \sum_{t=1}^{2} (3\varepsilon/2)\, \phi_s^{ij}(x)\, \psi_t^{ij}(y)$$

$$\leqslant (5/2)\varepsilon + 2\varepsilon \sum_{s=2}^{3} \left(\left|\bar{\phi}_s^{ij}(x)\right| + \left|\bar{\psi}_s^{ij}(y)\right|\right)$$

$$\leqslant (5/2)\varepsilon + 2\varepsilon \Big\{ \max_{1 \leqslant i \leqslant M-1} \left[\sum_{s=2}^{3} \left|\left|\bar{\phi}_{3(i-1)+s}\right|\right| \right]$$

$$+ \max_{1 \leqslant j \leqslant N-1} \left[\sum_{t=2}^{3} \left|\left|\bar{\psi}_{3(j-1)+t}\right|\right| \right] \Big\},$$

where we have used the fact that $\phi_s^{ij} \geqslant 0$, $\psi_t^{ij} \geqslant 0$, $\sum_{s=1}^{2} \phi_s^{ij} \equiv 1$,

$\sum_{s=1}^{2} \psi_s^{ij} \equiv 1$, $\bar{\phi}_s^{ij} = \bar{\phi}_{3(i-1)+s}$ on $[x_i, x_{i+1}]$ and $\bar{\psi}_s^{ij} = \bar{\psi}_{3(j-1)+s}$ on $[y_j, y_{j+1}]$.

We now show that $\tilde{s}_{ij} = \tilde{s}$ on $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$. This is accomplished by making the following observations. For $1 \leqslant s, t \leqslant 2$, $1 \leqslant i \leqslant M-1$, $1 \leqslant j \leqslant N-1$ we have $\bar{\phi}_s^{ij} = \bar{\phi}_{3(i-1)+s}$, $\bar{\psi}_t^{ij} = \bar{\psi}_{3(j-1)+t}$, $\phi_s^{ij} = \phi_{i+s-1}$ and $\psi_t^{ij} = \psi_{j+t-1}$. The following three sets are sets of equal elements

$$\{\tilde{a}^{i,j+v-t}_{2+u,1+t}\}_{0\leqslant t\leqslant 1}, \quad \{\tilde{b}^{i+u-s,j}_{1+s,2+v}\}_{0\leqslant s\leqslant 1} \text{ and } \{\tilde{c}^{i+u-s,j+v-t}_{1+s,\ 1+t}\}_{0\leqslant s,\, t\leqslant 1},$$

for $0 \leqslant u, v \leqslant 1$, which follows from (5.40), (5.41) and (5.42), respectively. Direct substitution of the above sets and relations into (5.44) shows that $\tilde{s}_{ij} = \tilde{s}$.

■

Remark: Once the meshes $\bar{\pi}_x$ and $\bar{\pi}_y$ have been specified, then direct calculation gives a bound for

$$\max_{1\leqslant i\leqslant M-1} \left[\sum_{s=2}^{3} \left|\left| \bar{\phi}_{3(i-1)+s} \right|\right| \right] + \max_{1\leqslant j\leqslant N-1} \left[\sum_{t=2}^{3} \left|\left| \bar{\psi}_{3(j-1)+t} \right|\right| \right].$$

For example, if on each of the intervals $[x_i, x_{i+1}]$ $([y_j, y_{j+1}])$ we assume that the four points $\{\bar{x}_{3(i-1)+s}\}_{s=1}^{4}$ $(\{\bar{y}_{3(j-1)+t}\}_{t=1}^{4})$ are equally spaced, then direct calculation shows that for Theorem 5.5

(5.53) $$||f-\tilde{s}|| \leq (5/2 + 4(10 + 7\sqrt{7})/27)\,\mathcal{E}\,.$$

Remark: If Theorem 5.4 is applied to each of the spaces DBF_{ij} to obtain an upper bound for $||f-s_{ij}||_{[x_i, x_{i+1}]\times[y_j, y_{j+1}]}$, then the maximum of these can be used for $\mathcal{E}$ in Theorem 5.5 .

Section 6. Domain Transformations

In the previous sections, the domain on which the data was given was always the rectangle $[a, b]x[c, d]$. In this section, a method will be presented which will remove this restriction and allow singly and multiply connected domains with curved boundarys. This will be accomplished by using vector valued blending techniques (see Gordon and Hall [16], [17]) for domain transformations. Their procedure is to divide the boundary of a bounded region $\Omega \subset \mathbb{R}^2$ into four parameterized boundary curves. Blending these four curves yields a mapping $\vec{U}: [0, 1]x[0, 1] \longrightarrow \mathbb{R}^2$ which maps the boundary of $[0, 1]x[0, 1]$ onto the boundary of Ω. Of major concern is, under what conditions is the map $\vec{U}$ univalent? Some conditions have been given in [16] to insure univalency, however, the major responsibility of insuring univalency usually rests with the ability of the person implementing the scheme. For application to least squares, it is imperative to know the point $\vec{U}^{-1}(x, y)$, where $(x, y) \in X \subset \Omega$ is a data point. Even though an inverse for $\vec{U}$ may exist, it is not explicitly known. Thus, the following procedure is introduced which allows the inverse of $\vec{U}$ to be calculated at any point $(x, y) \in \Omega$. This will be accomplished by considering several special domains and then subdividing the domain Ω into a finite collection of these special domains.

Section 6.1 Type 1 Domain Transformations

Consider the closed region $\Omega \subset R^2$ which has the following form

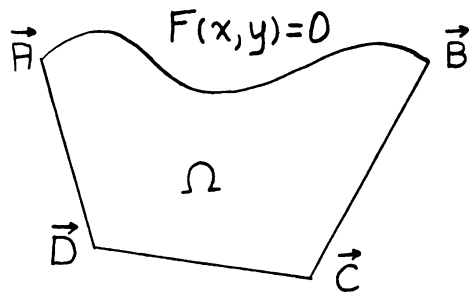

Figure 2. The Region Ω.

where the distinct points $\vec{A} = (a_1, a_2)$, $\vec{B} = (b_1, b_2)$, $\vec{C} = (c_1, c_2)$ and $\vec{D} = (d_1, d_2)$ are ordered as in Figure 2 to form a quadrilateral. We desire a map $\vec{U}: I^2 \longrightarrow \Omega$, where $I^2 = [0, 1] \times [0, 1]$, whose range is exactly Ω and is univalent.

Let $F \in C^{(1,1)}(\Omega^*)$, where Ω^* is some open region containing the curve $F(x, y) = 0$ from $\vec{A}$ to $\vec{B}$ where

$$(6.1) \qquad (F^{(1,0)}(x, y))^2 + (F^{(0,1)}(x, y))^2 > 0$$

for all points (x, y) on the curve $F(x, y) = 0$ from $\vec{A}$ to $\vec{B}$.

For each $s \in [0, 1]$, construct the straight line $\vec{L}(r; s)$ that contains the two points $(1-s)\vec{A} + s\vec{B}$ and $(1-s)\vec{D} + s\vec{C}$ (which are on the straight lines $\overline{AB}$ and $\overline{DC}$, respectively)

$$(6.2) \qquad \vec{L}(r; s) = r\{(1-s)\vec{A} + s\vec{B} - ((1-s)\vec{D} + s\vec{C})\} + (1-s)\vec{D} + s\vec{C},$$

where $r \in R$ and s is a parameter. The x and y components of

$\vec{L}(r;s)$ are expressed as $L^x(r;s)$ and $L^y(r;s)$, respectively. The line $\vec{L}(r;s)$ can also be expressed in the form

$$(6.3) \qquad y = \frac{(1-s)(a_2-d_2) + s(b_2-c_2)}{(1-s)(a_1-d_1) + s(b_1-c_1)} (x-((1-s)a_1+sb_1))+(1-s)a_2+sb_2$$

or

$$(6.4) \qquad y = \eta(s)\, x + \xi(s) .$$

Remark 6.1: If for some $s \in [0,1]$, the denominator of $\eta(s)$ is zero, i.e., $(1-s)(a_1-d_1) + s(b_1-c_1) = 0$, or if $|\eta(s)| >> 1$, then express the line $\vec{L}(r;s)$ as $x = \tilde{\eta}(s)\, y + \tilde{\xi}(s)$ where $\tilde{\eta}(s) = 1/\eta(s)$ and, in the following discussion, interchange the rolls of the x and y variables. Without loss of generality, throughout this section we will assume that $\eta(s)$ is finite for the value of s considered.

The following assumptions about the lines $\vec{L}(r;s)$ $(y = \eta(s)\, x + \xi(s))$ and the curve $F(x,y) = 0$ are made. For each $s \in [0,1]$, the straight line $\vec{L}(r;s)$ $(y = \eta(s)\, x + \xi(s))$ intersects the curve $F(x,y) = 0$ once and only once. Also, each point of $F(x,y) = 0$ from $\vec{A}$ to $\vec{B}$ lies on a line $\vec{L}(r;s)$ $(y = \eta(s)\, x + \xi(s))$ for some $s \in [0,1]$ (these conditions are usually not too restrictive, since a domain often can be subdivided until it holds). The line $\vec{L}(r;s)$ $(y = \eta(s)\, x + \xi(s))$ does not intersect the curve $F(x,y) = 0$ tangentially. Also, for $s \in [0,1]$, assume that for distinct values of s the lines $\vec{L}(r;s)$ $(y = \eta(s)\, x + \xi(s))$ do not have a point of intersection in the

region Ω and that the curve $F(x, y) = 0$ does not intersect the interior of the line $\overline{DC}$. Finally, for $s \in [0, 1]$, if $\vec{P}$ is that point of intersection of the straight line $\vec{L}(r; s)$ and curve $F(x, y) = 0$, then $\{ t(\vec{P} - ((1-s)\vec{D} + s\vec{C})) + (1-s)\vec{D} + s\vec{C} \mid \; 0 \leqslant t \leqslant 1 \} \subseteq \Omega$.

Remark 6.2: This final assumption can be proved from the others, but its proof leads us away from the desired results of this section.

Parameterization of $F(x, y) = 0$

We wish to locate the point designated by $(x(s), y(s)) = \vec{F}(s) \in \Omega^*$ where the straight line $\vec{L}(r; s)$ $(y = \eta(s)\, x + \xi(x))$ intersects the curve $F(x, y) = 0$. This can be accomplished by either one of the following two procedures which we develop simultaneously.

For a fixed $s \in [0, 1]$, let

(6.5a) $\qquad z(r) = F(\vec{L}(r; s))$

and

(6.5b) $\qquad w(x) = F(x, \; \eta(s)\, x + \xi(s))$.

The root $\hat{r}$ where $z(\hat{r}) = 0$ or $x(s)$ where $w(x(s)) = 0$ is the desired solution. One of the many root finding techniques for univariate functions can be employed to locate the root in (6.5a) and (6.5b); for example, Newton's method or the method of false position (see [24, pp. 97-100]). If Newton's method is chosen, then the derivatives $z'(r)$ and $w'(x)$ are calculated as follows.

If $\vec{L}(r; s)$ and $(x, \eta(s)\, x + \xi(s)) \in \Omega^*$, then because $F \in C^{(1, 1)}(\Omega^*)$, we have

$$(6.6a) \qquad z'(r) = F^{(1, 0)}(\vec{L}(r; s))(L^x(r; s))' + F^{(0, 1)}(\vec{L}(r; s))(L^y(r; s))'$$

$$= F^{(1, 0)}(\vec{L}(r; s))\{(1-s)(a_1 - d_1) + s(b_1 - c_1)\}$$

$$+ F^{(0, 1)}(\vec{L}(r; s))\{(1-s)(a_2 - d_2) + s(b_2 - c_2)\}$$

$$(6.6b) \qquad w'(x) = F^{(1, 0)}(x, \eta(s)\, x + \xi(s)) + F^{(0, 1)}(x, \eta(s)\, x + \xi(s))\, \eta(s).$$

Let $\hat{r}$ and $x(s)$ correspond to the roots of (6.6a) and (6.6b), respectively. It will be shown that both $z'(\hat{r})$ and $w'(x(s))$ are nonzero.

If $z'(\hat{r}) = 0$, then without loss of generality, from (6.1) we assume that $F^{(0, 1)}(\vec{L}(\hat{r}; s)) \neq 0$. Thus from (6.6a)

$$(6.7) \qquad (L^x(\hat{r}; s))' \{-F^{(1, 0)}(\vec{L}(\hat{r}; s)) \,/\, F^{(0, 1)}(\vec{L}(\hat{r}; s))\} =$$

$$(L^y(\hat{r}; s))'.$$

If $(L^x(\hat{r}; s))' = 0$, then $(L^y(\hat{r}; s))' = 0$. Using (6.2) this implies that the straight lines $\overline{AB}$ and $\overline{DC}$ intersect, which cannot happen, implying that $(L^x(\hat{r}; s)) \neq 0$. The Implicit Function Theorem (see [10, p. 256]) implies that the slope of the curve $F(x, y) = 0$ at $\vec{L}(\hat{r}; s)$ is given by the term in brackets of (6.7) which would equal $(L^y(\hat{r}; s))' / (L^x(\hat{r}; s))' = \eta(s)$ which is the slope of the straight line $\vec{L}(r; s)$. Our nontangential intersection assumption guarantees that this cannot happen, implying that $z'(\hat{r}) \neq 0$.

The argument that $w'(x(s)) \neq 0$, is similar and is omitted.

For sufficiently close initial estimates r_1 and x_1, for $i = 1, 2, \cdots$

$$r_{i+1} = r_i - z(r_i)/z'(r_i) \tag{6.8a}$$

and

$$x_{i+1} = x_i - w(x_i)/w'(x_i), \tag{6.8b}$$

where the r_i converge to $\hat{r}$ and the x_i converge to $x(s)$.

Remark 6.3: If $F \in C^{(2,2)}(\Omega^*)$, then $z(r)$ and $w(x)$ are twice differentiable in a neighborhood of $\hat{r}$ and $x(s)$, respectively. Under these conditions, the convergence of Newton's method is quadratic (see [24, p. 98]).

Remark 6.4: Finding an initial estimate for $\hat{r}$ is often easier than for $x(s)$, because $r = 1$ corresponds to a point on the line $\overline{AB}$ and the curve $F(x, y) = 0$ is usually "somewhat near" $\overline{AB}$.

Having calculated $\hat{r}$ and $x(s)$, then

$$\vec{L}(\hat{r}; s) \tag{6.9a}$$

or

$$(x(s), \eta(s)\, x(s) + \xi(s)) = (x(s), y(s)) \tag{6.9b}$$

is the point of intersection of $F(x, y) = 0$ and $\vec{L}(r; s)$ $(y = \eta(s)\, x + \xi(s))$. This completes the parameterization of the curve $F(x, y) = 0$.

Remark 6.5: If the curve is given as $y = f(x)$ or $x = g(y)$, then the extension of the above to this case is obvious.

Calculation of $\vec{U}(s, t)$

The straight lines $\overline{DA}$, $\overline{CB}$ and $\overline{DC}$ are parameterized linearly as follows for $s \in [0, 1]$ and $t \in [0, 1]$

$$t(\vec{A} - \vec{D}) + \vec{D} = (1 - t)\vec{D} + t\vec{A} \tag{6.10a}$$

$$t(\vec{B} - \vec{C}) + \vec{C} = (1 - t)\vec{C} + t\vec{B} \tag{6.10b}$$

and

$$s(\vec{C} - \vec{D}) + \vec{D} = (1 - s)\vec{D} + s\vec{C}\,. \tag{6.10c}$$

For notational convenience we define the vector $\vec{F}(s) = (x(s), y(s))$, where there is no confusion between the vector $\vec{F}(s)$ and the curve $F(x, y) = 0$ of which $(x(s), y(s))$ is a point.

Linearly blend the four curves $\overline{DA}$, $\overline{CB}$, $\overline{DC}$ and $F(x, y) = 0$ (see [16] for procedure) to obtain the vector valued map $\vec{U}: I^2 \longrightarrow \mathbb{R}^2$

$$\begin{aligned}\vec{U}(s, t) = {} & (1-s)\{(1-t)\vec{D} + t\vec{A}\} + s\{(1-t)\vec{C} + t\vec{B}\} \\ & + (1-t)\{(1-s)\vec{D} + s\vec{C}\} + t\vec{F}(s) \\ & - (1-s)(1-t)\vec{D} - (1-s)t\vec{A} - s(1-t)\vec{C} - st\vec{B} \\ = {} & (1-t)\{(1-s)\vec{D} + s\vec{C}\} + t\vec{F}(s)\,.\end{aligned} \tag{6.11}$$

Remark 6.6: Observe that $\vec{U}(s, 0) = (1-s)\vec{D} + s\vec{C}$, $\vec{U}(s, 1) = \vec{F}(s)$, $\vec{U}(0, t) = (1-t)\vec{D} + t\vec{F}(0) = (1-t)\vec{D} + t\vec{A}$ and $\vec{U}(1, t) = (1-t)\vec{C} + t\vec{F}(1) = (1-t)\vec{C} + t\vec{B}$. Thus the mapping $\vec{U}$ carries the boundary of I^2 onto the boundary of Ω.

$\vec{U}$ is univalent. Let $(s_1, t_1), (s_2, t_2) \in I^2$ be two points such that $\vec{U}(s_1, t_1) = \vec{U}(s_2, t_2)$. Observe that the points $\vec{F}(s_1)$ and $(1-s_1)\vec{D} + s_1\vec{C}$ are on the line $\vec{L}(r; s_1)$ by construction. From this and (6.11) it is clear that $\vec{U}(s_1, t_1)$ is on the same line $\vec{L}(r; s_1)$ because

$$\vec{U}(s_1, t) = t\{\vec{F}(s_1) - [(1-s_1)\vec{D} + s_1\vec{C}]\} + (1-s_1)\vec{D} + s_1\vec{C} \tag{6.12}$$

is a straight line passing through these two points. Making the corresponding observation that $\vec{U}(s_2, t_2)$ is on the line $\vec{L}(r; s_2)$, then we have $s_1 = s_2$ from the assumption that the lines $\vec{L}(r; s)$ have no points of intersection in Ω for distinct values of s and for $t \in [0, 1]$.

From (6.12) and our assumption that $\vec{F}(s)$ does not intersect $\overline{DC}$, it is clear that $t_1 = t_2$ because

$$\vec{U}(s_1, t_2) - \vec{U}(s_1, t_1) = (t_2 - t_1)(\vec{F}(s_1) - [(1-s_1)\vec{D} + s_1\vec{C}]), \tag{6.13}$$

hence, $\vec{U}$ is a univalent mapping.

Remark 6.7: For linear blending, the map $\vec{U}$ will not always be univalent. However, with the special parameterization given, this problem is circumvented.

Continuity of $\vec{U}$. If $F \in C^{(1,1)}(\Omega^*)$, then we will prove that $\vec{U} \in C^{(1,1)}(I^2)$. Examination of (6.11) shows that the continuity of $\vec{U}$ is limited only by the continuity of $\vec{F}(s) = (x(s), y(s))$, hence it will be shown that $x(s), y(s) \in C^{(1)}[0,1]$.

For a fixed $\hat{s} \in [0,1]$, the denominator of $\eta(\hat{s})$ is assumed non-zero, i.e. $|(1-\hat{s})(a_1-d_1) + \hat{s}(b_1-c_1)| = \mathcal{E} > 0$. Hence, for $s \in (\hat{s}-\delta, \hat{s}+\delta) \cap [0,1]$, it follows that $|(1-s)(a_1-d_1) + s(b_1-c_1)| \geqslant \mathcal{E}/2 > 0$ where $\delta = \mathcal{E}/2$ if $|-(a_1-d_1) + (b_1-c_1)| \leqslant 1$ or $\delta = \mathcal{E}/(2|-(a_1-d_1) + (b_1-c_1)|)$ otherwise. Direct calculation shows that both $\eta(s)$ and $\xi(s)$ are continuously differentiable in the interval $(\hat{s}-\delta, \hat{s}+\delta) \cap [0,1]$. Also observe that $F(x, \eta(s)\,x + \xi(s))$ is continuously differentiable as a bivariate function of x and s for $s \in (\hat{s}-\delta, \hat{s}+\delta) \cap [0,1]$ and x such that $(x, \eta(s)\,x + \xi(s)) \in \Omega^*$ because $F \in C^{(1,1)}(\Omega^*)$. The implicit function theorem (see [10, p. 257]) implies the existence of $x(s)$ which is a unique continuously differentiable function in some neighborhood N of $\hat{s}$ if $F^{(1,0)}(x(\hat{s}), \eta(\hat{s})\,x(\hat{s}) + \xi(\hat{s})) + F^{(0,1)}(x(\hat{s}), \eta(\hat{s})\,x(\hat{s}) + \xi(\hat{s}))\,\eta(\hat{s}) \neq 0$ (note that our assumptions guarantee that $x(s)$ and $y(s)$ are both univalent, and we only need prove continuity). It has already been shown that this cannot be zero (see the discussion following (6.6b)). Thus

$$x'(s) = -(\eta'(s)\,x(s) + \xi'(s))F^{(0,1)} \Big/ (F^{(1,0)} + \eta(s)\,F^{(0,1)}), \tag{6.14}$$

where $F^{(0,1)}$ and $F^{(1,0)}$ are evaluated at the point $(x(s), \eta(s)$ $x(s) + \xi(s))$ for $s \in N$.

From (6.4), the first derivative of $y(s)$ also exists for $s \in N$.

Remark 6.8: If higher continuity of F is assumed, then higher derivatives of $x(s)$ and $y(s)$ follow by repeated differentiation of (6.14) and (6.4) (observe that the denominator of (6.14) cannot be zero).

$\vec{U}([0,1]x[0,1]) = \Omega$. Because $\vec{U}$ is continuous, univalent, and maps the boundary of I^2 onto the boundary of Ω (see Remark 6.6), we have from Theorem 13.1 of [27, p. 121] that $\vec{U}([0,1]x[0,1]) = \Omega$, i.e., the range of $\vec{U}$ is precisely Ω.

Remark 6.9: Linear blending will not always yield a map whose range is Ω. It is possible for the mapping to "spill over" the boundary of Ω into the complement of Ω, (see [16]).

Summarizing, to construct a univalent and onto bivariate map $\vec{U}$, parameterize $F(x, y) = 0$ with respect to s by the above method and use linear blending. The observation should be made, given the point $(s, t) \in I^2$, that $F(x, y) = 0$ need only be parameterized for that single value of s.

Calculating $\vec{U}^{-1}(x, y)$

A procedure is now given to calculate $\vec{U}^{-1}(\hat{x}, \hat{y})$ for $(\hat{x}, \hat{y}) \in \Omega$. It was shown above that $\vec{U}$ is a univalent onto map; thus a unique point $(\hat{s}, \hat{t}) \in I^2$ exists such that $\vec{U}(\hat{s}, \hat{t}) = (\hat{x}, \hat{y})$.

Coordinate $\hat{s}$ is calculated first. From (6.11), observe that $(\hat{x}, \hat{y})$ lies on the straight line $\vec{L}(r; \hat{s})$ $(y = \eta(\hat{s})\, x + \xi(\hat{s}))$. The two points $(1-\hat{s})\,\vec{A} + \hat{s}\vec{B}$ and $(1-\hat{s})\,\vec{D} + \hat{s}\vec{C}$ are also on the same line. Thus, we desire the value $\hat{s}$ which causes the vector $(v_1, v_2) = (1-\hat{s})\,\vec{A} + \hat{s}\vec{B} - [(1-\hat{s})\,\vec{D} + \hat{s}\vec{C}]$ to be parallel to the vector $(w_1, w_2) = (\hat{x}, \hat{y}) - [(1-\hat{s})\,\vec{D} + \hat{s}\vec{C}]$, i.e. (w_1, w_2) to be a scalar multiple of (v_1, v_2). If $(v_1, v_2) \neq (0, 0)$, which is the case because $\vec{A} \neq \vec{D}$ and $\vec{B} \neq \vec{C}$, then the above is equivalent to

$$\det \begin{pmatrix} w_1 & v_1 \\ w_2 & v_2 \end{pmatrix} = 0 . \tag{6.15}$$

Direct calculation shows that s must satisfy

$$0 = K_1 \hat{s}^2 + K_2 \hat{s} + K_3 , \tag{6.16}$$

where

$$K_1 = (d_1 - c_1)(b_2 - a_2) - (d_2 - c_2)(b_1 - a_1) \tag{6.17}$$

$$K_2 = (d_1 - c_1)(a_2 - d_2) + (\hat{x} - d_1)(b_2 - c_2 + d_2 - a_2) \tag{6.18}$$

$$-(d_2 - c_2)(a_1 - d_1) - (\hat{y} - d_2)(b_1 - c_1 + d_1 - a_1)$$

and

$$(6.19) \qquad K_3 = (\hat{x}-d_1)(a_2-d_2) - (\hat{y}-d_2)(a_1-d_1) .$$

If $K_1 = 0$, then this is equivalent to

$$(6.20) \qquad \det(\vec{B} - \vec{A},\ \vec{C} - \vec{D}) = 0 ,$$

or that $\overline{AB}$ be parallel to $\overline{DC}$. Hence, equation (6.16) reduces to a linear equation. In either case, because of our construction, $\hat{s}$ is unique in the interval $[0, 1]$.

The coordinate $\hat{t}$ is computed as follows. First, calculate the point of intersection $\vec{F}(\hat{s})$, of the line $\vec{L}(r; \hat{s})$ $(y = \eta(\hat{s})\, x + \xi(\hat{s}))$ and the curve $F(x, y) = 0$ (the procedure is identical to that given above). $\hat{t}$ is given by

$$(6.21) \qquad \hat{t} = \left|\left|(\hat{x}, \hat{y}) - \{(1-\hat{s})\vec{D} + \hat{s}\vec{C}\}\right|\right| \Big/ \left|\left|\vec{F}(\hat{s}) - \{(1-\hat{s})\vec{D} + \hat{s}\vec{C}\}\right|\right| .$$

The denominator of (6.21) is non-zero because it is assumed that the curve $F(x, y) = 0$ does not intersect the line $\overline{DC}$.

Using the fact that $(1-\hat{s})\vec{A} + \hat{s}\vec{B}$, $(1-\hat{s})\vec{D} + \hat{s}\vec{C}$, $\vec{F}(\hat{s})$ and $(\hat{x}, \hat{y})$ are all on the line $\vec{L}(r; \hat{s})$ $(y = \eta(\hat{s})\, x + \xi(\hat{s}))$ by construction, then direct calculation shows that $\vec{U}(\hat{s}, \hat{t}) = (\hat{x}, \hat{y})$.

To reiterate, $\vec{U}^{-1}(\hat{x}, \hat{y})$ is calculated by first solving for $\hat{s}$ from the quadratic (6.16), parameterizing $F(x, y) = 0$ to obtain $\vec{F}(\hat{s})$, and then calculating $\hat{t}$ from (6.21).

Section 6.2. Type 2 Transformations

The above procedure is modified by letting two of the points in Figure 2 coincide, either $\vec{A} = \vec{D}$, $\vec{B} = \vec{C}$, $\vec{A} = \vec{B}$ or $\vec{D} = \vec{C}$ and the resulting procedure will be denoted as a type 2 domain transformation.

Type 2a Transformations. The case where $\vec{A} = \vec{D}$ will be examined first, observing that the case where $\vec{B} = \vec{C}$ is nearly identical. Nothing is changed in parameterizing $F(x, y) = 0$ to obtain $\vec{F}(s)$ (using the fact that if $s = 0$, then $\vec{F}(0) = \vec{A} = \vec{D}$), and $\vec{U}(s, t)$ is given by equation (6.11). $\vec{U}$ is no longer univalent, because the points $(0, t)$ for $t \in [0, 1]$ are all mapped to the single point $\vec{A} = \vec{D}$ (see Remark 6.6). However, the argument given above proves that $\vec{U}$ is univalent on $(0, 1] \times [0, 1]$. Nothing is changed in the argument on the continuity of $\vec{U}$, taking $\eta(s) = (b_2 - c_2)/(b_1 - c_1) = \eta$ for $s \in [0, 1]$. However, the proof that $\vec{U}(I^2) = \Omega$ is no longer valid for this case because $\vec{U}$ is no longer univalent. To circumvent this problem, we first show that $\vec{U}$ cannot map to a point outside of Ω. Let $(\hat{s}, \hat{t}) \in (0, 1] \times [0, 1]$, then from the above construction, $\vec{F}(\hat{s})$ and $\vec{U}(\hat{s}, \hat{t})$ are both on the straight line $\vec{L}(r; \hat{s})$ which can be expressed as $\vec{U}(\hat{s}, t) = t\{\vec{F}(\hat{s}) - [(1 - \hat{s})\vec{D} + \hat{s}\vec{C}]\} + (1 - \hat{s})\vec{D} + \hat{s}\vec{C}$. Because $\hat{t} \in [0, 1]$, our assumptions guarantee that $\vec{U}(\hat{s}, \hat{t}) = (\hat{x}, \hat{y}) \in \Omega$, proving that $\vec{U}$ maps into Ω. To prove that $\vec{U}$ maps onto Ω, let $(\hat{x}, \hat{y}) \in \Omega$. If $(\hat{x}, \hat{y}) = \vec{A} = \vec{D}$ or $(\hat{x}, \hat{y})$ is a point on the line $\overline{BC}$, then take $\hat{s} = 0$ or

= 1, respectively. If $(\hat{x}, \hat{y})$ is a point on the curve $F(x, y) = 0$ or the straight line $\overline{AC} = \overline{DC}$, then take $\hat{s}$ such that $(\hat{x}, \hat{y}) = \vec{F}(\hat{s})$ or $=(1-\hat{s})\vec{D}+\hat{s}\vec{C}$, respectively. If $(\hat{x}, \hat{y})$ is not on the boundary of Ω, then $(\hat{x}, \hat{y})$ is a point on the straight line $y = \eta(x-\hat{x})+\hat{y}$, which is parallel to the line $\overline{BC}$ (recall $\eta(s) = \eta$ = constant for $s \in [0, 1]$). This line must intersect the boundary of Ω at least twice and our assumptions imply that this line must also be one of the lines $\vec{L}(r; \hat{s})$ for some $\hat{s} \in (0, 1)$. To see this, note that $y = \eta(x-\hat{x})+\hat{y}$ cannot intersect the point $\vec{A} = \vec{D}$ or the line $\overline{BC}$ which is parallel to it. If it intersects the curve $F(x, y) = 0$ at some point, say $\vec{F}(\hat{s})$, then the lines $y = \eta(x-\hat{x}) + \hat{y}$ and $\vec{L}(r; \hat{s})$ are identical because they have the same slope and a point in common. A similar argument holds if $y = \eta(x-\hat{x}) + \hat{y}$ intersects the curve $\overline{DC}$ at some point $(1-\hat{s})\vec{D}+\hat{s}\vec{C}$, in either case take $s = \hat{s}$. Also note that this value of s is unique because $y = \eta(x-\hat{x}) + \hat{y}$ is a parameter line $\vec{L}(r; \hat{s})$.

If $\hat{s} \neq 0$, then compute $\hat{t}$ from (6.21), and direct calculation shows that $\vec{U}(\hat{s}, \hat{t}) = (\hat{x}, \hat{y})$ (if $\hat{s} = 0$, then any $\hat{t} \in [0, 1]$ is acceptable).

Type 2b Transformations. Consider the case where $\vec{D} = \vec{C}$ (note that the case where $\vec{A} = \vec{B}$ is similar and less complex). All of the lines $\vec{L}(r; s)$ $(y = \eta(s)\, x + \xi(s))$ intersect at the point $\vec{D} = \vec{C}$, but no other point of Ω. $F(x, y) = 0$ is parameterized as above and the map $\vec{U}$ reduces to

(6.22) $\vec{U}(s, t) = (1-t)\vec{D} + t\vec{F}(s)$.

$\vec{U}$ is not univalent because the points $(s, 0)$ for $s \in [0, 1]$ are mapped to the single point $\vec{D} = \vec{C}$. However, the above proof for the univalency of $\vec{U}$ holds for $(s, t) \in [0, 1]x(0, 1]$. The differentiability of $\vec{U}$ follows from the above argument. Finally, the proof that $\vec{U}(I^2) = \Omega$ is similar to that of type 2a transformations where we work with the straight line $t\{(x, y) - \vec{D}\} + \vec{D}$.

Section 6.3. General Domains. It should be understood that for the region Ω, the side on which the curve $F(x, y) = 0$ is located was specified as it was above for purposes of illustration. If it is desired, then the curve $F(x, y) = 0$ can be any one of the four sides of the quadrilateral region, where minor modifications of the above presentation will yield the appropriate procedure and formulas for the construction of $\vec{U}$.

We now consider closed regions Γ which can be subdivided into N subregions Ω_i such that

(6.23) $$\Gamma = \bigcup_{i=1} \Omega_i ,$$

and for $1 \leqslant i, j \leqslant N$, $i \neq j$

(6.24) $$\text{Int}\,(\Omega_i) \cap \text{Int}\,(\Omega_j) = \phi ,$$

where each Ω_j is a type 1 or type 2 region satisfying the assumptions of Section 6.1 or 6.2, respectively. For each subregion

construct the map $\vec{U}_i : I^2 \longrightarrow \Omega_i$. Several examples will be given to illustrate the procedure (where the numbers in the following illustrations have the obvious correspondence).

Figure 3

In Figure 3, Ω_1, Ω_3, Ω_4 and Ω_6 are type 2b regions (type 2a could be used with minor modifications) and Ω_2 and Ω_5 are type 1 regions. Note that the left hand side of I_1^2 and I_4^2 (I_3^2 and I_6^2) are identified in the obvious manner and the adjacent side of I_1^2 and I_4^2 (I_3^2 and I_6^2) are mapped to the single point $\vec{P}_1$ ($\vec{P}_2$).

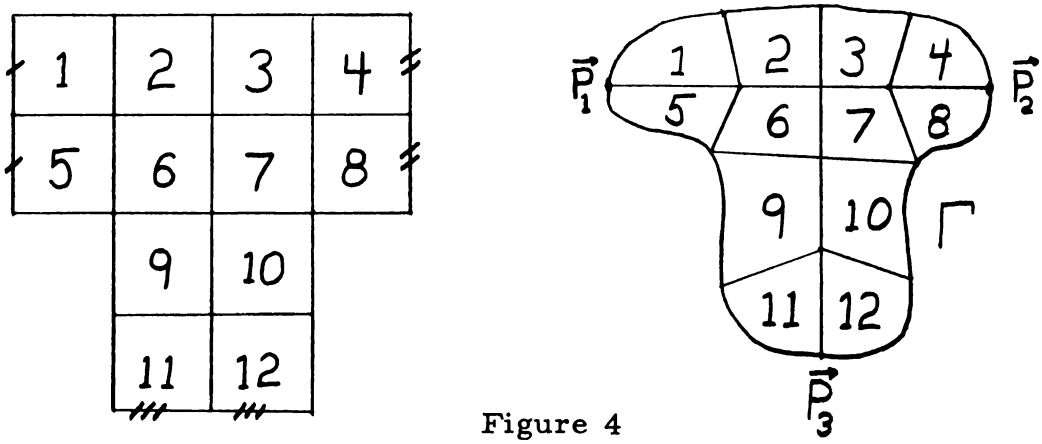

Figure 4

In Figure 4 Ω_1, Ω_4, Ω_5, Ω_8, Ω_{11} and Ω_{12} are type 2a regions and the remaining regions are type 1, where the left hand sides of I_1^2 and I_5^2 are mapped to the point $\vec{P}_1$, etc.

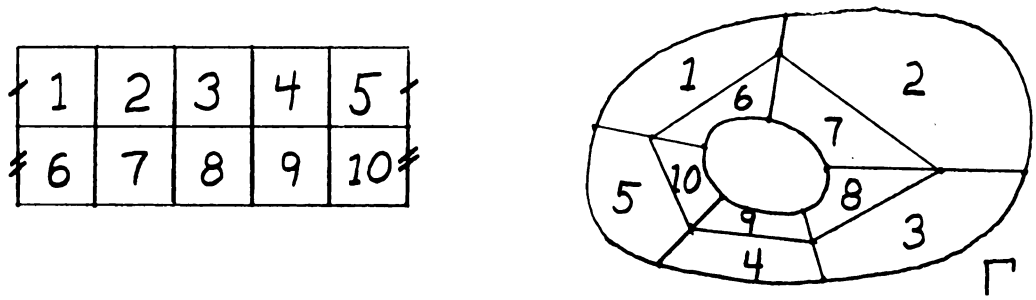

Figure 5

In Figure 5, the regions Ω_i for $1 \leqslant i \leqslant 10$ are all type 1, where the left hand side of I_1^2 and I_6^2 are identified with the right hand side of I_5^2 and I_{10}^2, respectively.

The procedure of subdividing Γ should be lucid from the above examples.

Section 6.4. Discrete Least Squares Over Γ. Let $f \in C^{(0,0)}(\Gamma)$, where Γ satisfies (6.23) and (6.24) and let $X = \{(\hat{x}_k, \hat{y}_k)\}_{k=1}^{\hat{M}} \subseteq \Gamma$ be a set of $\hat{M}$ unstructured data points. A discrete least squares fit $\tilde{g}_i : \Omega_i \longrightarrow \mathbb{R}$ to f is constructed over each region Ω_i in the following manner. Choose a finite dimensional bivariate real valued function space $S(I^2)$ over I^2 (for example, any of those of Section 5) and construct the map $\vec{U}_i : I^2 \longrightarrow \Omega_i$.

If Ω_i is a type 2 domain, then constrain $S(I^2)$ to insure that each element of $S(I^2)$ is constant on that single boundary line of I^2 which is mapped to the single point $\vec{P}$ (see Section 6.2) (for the spaces given in Section 5, this can easily be accomplished by modifying the basis elements). With this restriction there is no ambiguity for the element $s \circ \vec{U}_i^{-1}$, where $s \in S(I^2)$, because s is single valued on the inverse image (under $\vec{U}_i^{-1}$) of each point of the type 2 region Ω_i. Thus we define $\vec{U}_i^{-1}(\vec{P})$ to be (for example) the midpoint of the boundary line of I^2 whose image under $\vec{U}_i$ is the point $\vec{P}$.

Define $X_i = \{(\hat{x}_{ij}, \hat{y}_{ij})\}_{j=1}^{\hat{M}_i} = X \cap \Omega_i$ to be the set of $\hat{M}_i$ data points which are in Ω_i and define $ST_i = \{\vec{U}_i^{-1}(\hat{x}_{ij}, \hat{y}_{ij})\}_{j=1}^{\hat{M}_i}$, where for $1 \leqslant j \leqslant \hat{M}_i$

$$(\hat{s}_{ij}, \hat{t}_{ij}) = \vec{U}_i^{-1}(\hat{x}_{ij}, \hat{y}_{ij}) . \tag{6.25}$$

Finally, define the finite dimensional function space $V(\Omega_i) = \{s \circ \vec{U}_i^{-1} \mid s \in S(I^2)\}$.

We say $\tilde{g}_i \in V(\Omega_i)$ $(\tilde{s}_i \in S(I^2))$ is a discrete least squares fit to f $(f \circ \vec{U}_i)$ on the data set $X_i \subseteq \Omega_i$ $(ST_i \subseteq I^2)$ if the Euclidean norm of the residual vector $R_i \in \mathbb{R}^{\hat{M}_i}$ $(\tilde{R}_i \in \mathbb{R}^{\hat{M}_i})$ is minimized, where component j of R_i is given by

$$(R_i)_j = f(\hat{x}_{ij}, \hat{y}_{ij}) - \tilde{g}_i(\hat{x}_{ij}, \hat{y}_{ij}) \tag{6.26}$$

and component j of $\tilde{R}_i$ is given by

$$(6.27) \qquad (\tilde{R}_i)_j = (f \circ \vec{U}_i)(\hat{s}_{ij}, \hat{t}_{ij}) - \tilde{s}_i(\hat{s}_{ij}, \hat{t}_{ij}) .$$

It is clear from (6.25), (6.26) and (6.27) that if $\tilde{g}_i$ ($\tilde{s}_i$) is a discrete least squares fit, then there exists $s^* \in S(I^2)$ ($g^* = \tilde{s}_i \circ \vec{U}_i^{-1} \in V(\Omega_i)$) such that $s^* \circ \vec{U}_i^{-1} = \tilde{g}_i$ ($g^* \circ \vec{U}_i = \tilde{s}_i$) where

$$(6.28) \qquad ||R_i|| = \Big(\sum_{j=1}^{\hat{M}_i} \Big[(f \circ \vec{U}_i)(\hat{s}_{ij}, \hat{t}_{ij}) - s^*(\hat{s}_{ij}, \hat{t}_{ij}) \Big]^2 \Big)^{1/2} ,$$

and

$$(6.29) \qquad ||\tilde{R}_i|| = \Big(\sum_{j=1}^{\hat{M}_i} \Big[f(\hat{x}_{ij}, \hat{y}_{ij}) - g^*(\hat{x}_{ij}, \hat{y}_{ij}) \Big]^2 \Big)^{1/2} ,$$

showing that $||R_i|| = ||\tilde{R}_i||$ because $||R_i|| \geqslant ||\tilde{R}_i||$ and $||R_i|| \leqslant ||\tilde{R}_i||$. Therefore, a discrete least squares fit $\tilde{g}_i \in V(\Omega_i)$ can be calculated from a discrete least squares fit $\tilde{s}_i \in S(I^2)$ by taking

$$(6.30) \qquad \tilde{g}_i = \tilde{s}_j \circ \vec{U}_i^{-1} .$$

Observing that

$$(6.31) \qquad ||f - \tilde{g}_i||_{\Omega_i} = ||f \circ \vec{U}_i - \tilde{s}_i||_{I^2} ,$$

an error analysis for $f - \tilde{g}_i$ over Ω_i can be given in terms of $f \circ \vec{U}_i - \tilde{s}_i$ over I^2.

Thus, the case of least squares over a general domain Γ is reduced to least squares over rectangular domains and the error analysis reduces to bounding $||f \circ \vec{U}_i - \tilde{s}_i||$ where, for example, $f \circ \vec{U}_i$ replaces f in the formulas (5.15), (5.16), (5.21), (5.22), (5.30), (5.31), (5.36) and (5.37).

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Ahlberg, Nilson, and Walsh, The Theory of Splines and Their Applications, Academic Press: New York, 1967.

2. Birkhoff, G., and C. DeBoor, "Error Bounds for Spline Interpolation," Journal of Mathematics and Mechanics, Vol. 13, No. 5 (1964) pp. 827-835.

3. Björck, Å., "Solving Linear Least Squares Problems by Gram-Schmidt Orthogonalization," Nordisk Tidskrift for Informations - Behandling, Vol. 7 (1967).

4. Businger, P., and G. H. Golub, "Linear Least Squares Solutions by Householder Transformations," Numerische Mathematik, Vol. 7 (1965) pp. 269-276.

5. Carlson, R. E., and C. A. Hall, "Error Bounds for Bicubic Spline Interpolation," Journal of Approximation Theory, Vol. 7, No. 1, Jan. 1973, pp. 41-47.

6. Cheney, E. W., Introduction to Approximation Theory, McGraw-Hall: New York, 1966.

7. Deskins, W. E., Abstract Algebra, MacMillan Co.: New York, 1964.

8. Dugundji, J., Topology, Allyn and Bacon, Inc.: Boston, 1966.

9. Faddeev, D. K., and V. N. Faddeeva, Computational Methods of Linear Algebra, W. H. Freedman and Company: San Francisco, 1963.

10. Fulks, W., Advanced Calculus - An Introduction to Analysis, John Wiley and Sons, Inc.: New York, 1961.

11. Golub, G., "Numerical Methods for Solving Linear Least Squares Problems," Numerische Mathematik, Vol. 7 (1965) pp. 206-216.

12. Gordon, W. J. "'Blending-Function' Methods of Bivariate and Multivariate Interpolation and Approximation," General Motors Research Publication, GMR-834, Warren, Michigan, October, 1968.

13. Gordon, W. J., "Spline-Blended Surface Interpolation Through Curve Networks," Journal of Mathematics and Mechanics, Vol. 18, No. 10, April, 1969, pp. 931-952.

14. Gordon, W. J., and C. A. Hall, "Discretization Error Bounds for Transfinite Elements," General Motors Research Publication, GMR-1196, Warren, Michigan, May, 1972.

15. Gordon, W. J., and C. A. Hall, "Geometric Aspects of the Finite Element Method," in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, Academic Press, Inc.: New York, 1972.

16. Gordon, W. J., and C. A. Hall, "Geometric Aspects of the Finite Element Method: Construction of Curvilinear Coordinate Systems and Their Application to Mesh Generation," General Motors Research Publication, GMR-1286, Warren, Michigan.

17. Gordon, W. J., and C. A. Hall, "Transfinite Element Methods: Blending-Function Interpolation over Arbitrary Curved Element Domains," Numerische Mathematik, Vol. 21, 1973, pp. 109-129.

18. Gordon, W. J., and J. A. Wixom, "Pseudo-Harmonic Interpolation on Convex Domains," General Motors Research Publication, GMR-1248, Warren, Michigan, August 1972.

19. Hall, C. A., "Natural Cubic and Bicubic Spline Interpolation," SIAM J. Numerical Analysis, Vol. 10, No. 6, Dec, 1973, pp. 1055-1060.

20. Hall, C. A., "On Error Bounds for Spline Interpolation," Journal of Approximation Theory, Vol. 1, 1968, pp. 209-218.

21. Halliday, J., and J. G. Hayes, "The Least-Squares Fitting of Cubic Spline Surfaces to General Data Sets," National Physical Laboratory, NPL Report NAC 22, December, 1972.

22. Householder, A. S., "Unitary Triangularization of a Non-symmetric Matrix," Assoc. Comput. Mach., Vol. 5, 1958, pp. 339-342.

23. Householder, A. S., The Theory of Matrices in Numerical Analysis, Blardell Publishing Co.: 1964.

24. Isaacson, E., and H. B. Keller, Analysis of Numerical Methods, Wiley and Sons, Inc.: New York, 1966.

25. Jennings, L. S., and M. R. Osborne, "A Direct Error Analysis for Least Squares," Numerische Mathematik, Vol. 22, 1974, pp. 325-332.

26. Marcus, M., and H. Ming, A Survey of Matrix Theory and Matrix Inequalities, Allyn and Bacon Inc.: Boston, 1964.

27. Newman, M. H. A., Topology of Plane Sets, Cambridge University Press: 1964.

28. Peters, G. and J. H Wilkinson, "The Least Squares Problem and Pseudo-Inverses," The Computer Journal, Vol. 13, No. 3, August, 1970, pp. 309-316.

29. Schultz, M H., Spline Analysis, Prentice-Hall, Inc.: Englewood Cliffs, N. J., 1973.

30. Soble, A. B., "Majorants of Polynomial Derivatives," American Math Monthly, Vol. 64, 1957, pp. 639-643.

31. van der Sluis, A., "Stability of the Solution of Linear Least Squares Problems," Numerische Mathematik, Vol. 23, 1975, pp. 241-254.

32. Varga, R. S. Matrix Iterative Analysis, Prentice-Hall, Inc.: Englewood Cliffs, N. J., 1962.

33. Wendroff, B., Theoretical Numerical Analysis, Academic Press: New York, 1966.

34. Wilkinson, J. H., "Error Analysis of Direct Methods of Matrix Inversion," Association for Computing Machinery, Vol. 8, No. 1, Jan, 1961.

35. Wilkinson, J. H., "Error Analysis of Transformations Based on the Use of Matrices of the Form $I-2ww^H$," in Error in Digital Computation, Vol. 2, Edited by Louis B. Rall, Wiley and Sons, Inc.: 1965.

36. Wilkinson, J. H., Rounding Errors in Algebraic Processes, Prentice-Hall, Inc.: Englewood Cliffs, N.J., 1962.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) BLENDING-FUNCTION TECHNIQUES WITH APPLICATIONS TO DISCRETE LEAST SQUARES		5. TYPE OF REPORT & PERIOD COVERED Interim 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Dale Russel Doty		8. CONTRACT OR GRANT NUMBER(s) AFOSR-72-2271
9. PERFORMING ORGANIZATION NAME AND ADDRESS Michigan State University Department of Mathematics East Lansing, MI 48824		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 9749-03
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research (NM) 1400 Wilson Blvd. Arlington, VA 22209		12. REPORT DATE August, 1975 13. NUMBER OF PAGES 230
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

blending-functions, least squares, splines,error analysis, interpolation, natural splines, domain transformations

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The theory of blending function spaces (bivariate interpolation) is developed in the general setting of interpolation spaces. In this setting it is shown that blending function spaces have the desirable quality of doubling the order of accuracy with less computation when compared to standard tensor product spaces.

The dimensionality of discretized blending function spaces is derived, and

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

several bases are explicitly constructed. The special example of Hermite spline blended piecewise polynomials is developed, showing that these spaces have bases with small support which are easy to calculate. These spaces offer maximum order of convergence for a minimum number of basis elements. For example, linearly blended piecewise cubic polynomials offer a fourth order approximation scheme, and cubic Hermite spline blended piecewise polynomials offer an eighth order scheme.

Next, using the exponential decay of the natural cubic cardinal splines and natural cubic spline blending, a derivative-free approximation scheme is developed, which is eighth order in the interior of the domain.

Algorithms with corresponding error estimates are given for solving the discrete least squares problem with unstructured data. For the univariate case, algorithms are developed using the space of cubic splines. The resulting error analysis indicates the necessary restrictions to be placed on the number and distribution of the data points to insure that the discrete least squares fit will be $O(h^m)$ to a function $f \in C^m[a, b]$ from which the data arises, where h is the mesh size and $1 \leq m \leq 4$. An example is given to illustrate that the discrete least squares fit need not be close to f if these conditions are not realized. For the bivariate case, algorithms and error analyses are given for the spaces of bicubic splines and discretized blending function spaces. It is shown that the discrete least squares fit to a bivariate function f is of the same order accuracy as the corresponding interpolation accuracy.

Discrete least squares is considered on general domains which have curved boundaries and are possibly multiply connected. This general domain is subdivided into "standard" subdomains, and explicit mappings from the unit square to these standard subdomains are constructed which are one-one, onto, and have easily calculated inverses. Thus, discrete least squares over general domains reduces to the cases previously considered.

Finally, an extensive computational error analysis is given for a constrained least squares algorithm.