A STABLE HIGH ORDER INTERPOLATION SCHEME FOR SUPERCONVERGENT DATA

by

HONGSUNG JIN

A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Master of Science (Mathematical and Computer Sciences).

Golden, Colorado
Date Nov 9 /1993

Signed:
HONGSUNG JIN

Approved: Steven G. Pruess
Dr. Steven Pruess
Professor of Mathematical and Computer Sciences
Thesis Advisor

Golden, Colorado
Date Nov. 9, 1993

Dr. Ardel J. Boes
Professor and Head,
Department of Mathematical and Computer Sciences

ABSTRACT

A local collocation scheme is developed which yields stable high order accurate interpolants of discrete data arising from the numerical solution of a differential equation. The existence of uniformly superconvergent approximation is proved. Some choices for the local collocation points are suggested. Numerical examples are shown which illustrate the stability, even for the case of highly nonuniform meshes which have proven difficult in prior studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENT

I would like thank many persons who support and encourage my study. Thanks are due Dr. Willy Hereman for his kind and sincere helping with Mathematica, and for his encouragement on my committee. I would also like thank Dr. Van Vleck for his support on my committee.

Above all, my most thanks go to my advisor, Dr. Steven Pruess. Without his consistent guidance and encouragement, the work presented in this thesis could not have been done at all.

Thanks are also due my wife and sons for their patience and understanding Finally, I would like to dedicate this thesis to my parents in Korea.

Chapter 1

INTRODUCTION

The collocation method for approximating solutions of differential equations requires that the residuals, using the approximate solution, vanish at a finite number of so-called collocation points (for the exact solution the residual would vanish everywhere). Numerical approximations by the collocation method to solve two point boundary value problems are often computed only on a discrete set of mesh points and some type of interpolation is necessary if solutions are desired at points not in the mesh. Russell and Shampine (1972) tried the collocation method with piecewise polynomial functions. They suggested and showed that collocation with piecewise polynomials to approximate an mth order nonlinear differential equation had an $O(h^k)$ error bound when the number of collocation points between successive mesh points was k. de Boor and Swartz (1973) proved using the theory of projection methods that the collocation method with piecewise polynomials could be more accurate if a proper choice of collocation points was used. They used Gauss points for the set of collocation points. The error was optimal, being $O(h^{m+k})$ between the mesh points, which is the same as the error bounds $O(h^{m+k})$ from Galerkin's method and also the error of the best approximation by such polynomials (see de Boor, 1978). At the mesh points, the error was $O(h^{2k})$ which is called superconvergent. It was assumed that k was greater or equal to m.

Ascher, Pruess and Russell (1983) tested collocation with some different choices of bases such as B-spline basis, Hermite basis and monomial basis. They concluded that the monomial basis was recommended. Pruess (1986) attempted to boost the accuracy of the collocation solution at nonmesh points to the superconvergent accuracy at the mesh points. While some schemes of piecewise polynomial interpolation were adequate on many examples, all the schemes suffered when highly nonuniform meshes were used. The essential difficulty is the lack of sufficient superconvergent data to produce high order accurate uniform interpolants. To preserve the same order of accuracy, if data from neighboring intervals are used, this leads to inaccuracies when local mesh ratios are high (see Pruess, 1986).

In this thesis we discuss and illustrate a means of overcoming this difficulty; since it appears essential to use only local data, the missing information is supplied by a local collocation of the underlying differential equation. These local data are called secondary collocation points. To get the proper superconvergent accuracy between mesh points, three different secondary collocation point sets such as open uniform mesh, closed uniform mesh, and Gauss points are chosen and tested. Those local collocation points are analyzed using a related Hermite-Birkhoff interpolation scheme. In chapter 2 the Hermite-Birkhoff interpolation problem is studied using a monomial basis. To analyze the suitability of choices of secondary collocation points, the determinant of the coefficient matrices was calculated for some cases. We did not generalize this for all cases, but 10 cases for specific m and k are studied. In all cases the Gauss points make the Hermite-Birkhoff matrix singular. Hence the Gauss points are a poor choice for secondary collocation points. In chapter 3 the existence of the underlying projection is proved, and the uniformly superconvergent approximation is shown to have the error bound;

$$|u - \bar{u}| \leq Ch^{2k}.$$

In chapter 4 three sets of secondary collocation points are compared to the standard collocation solution. Both the open and closed uniform mesh choices are good approximations, superior to standard collocation. The open uniform mesh especially has an order of error comparable to superconvergence at the mesh points. Only a scalar differential equation is tested in this paper, the extension to the system case based on the same scheme is possible (see Pruess, 1993). The resulting algorithms are straightforward for linear differential equations; extensions to nonlinear problems by iteration on approximating linear problems should be self-evident.

Chapter 2

HERMITE-BIRKHOFF INTERPOLATION

2.1 What is Hermite-Birkhoff interpolation ?

A particular Hermite-Birkhoff interpolation scheme will be used to explain the behavior of the algorithms developed in later chapters. This problem is the following; Given integers m and k with $k > m$, some sufficiently smooth function $U(t)$, and a set of points $\{\sigma_i\}$ in [-1,1], find a polynomial $p(t)$ of order $2k$ satisfying

1. $(D^{j-1}p)(-1) = (D^{j-1}U)(-1)$ for j = 1,2,...,m;
2. $(D^{j-1}p)(1) = (D^{j-1}U)(1)$ for j = 1,2,...,m;
3. $(D^{m}p)(\sigma_i) = (D^{m}U)(\sigma_i)$ for i = 1,2,...,2k-2m.

(2.1)

This Hermite-Birkhoff interpolation problem has $2m$ conditions on the boundary and $2k - 2m$ conditions between boundaries. In the uniformly superconvergent approximation problem, we have $2m$ boundary conditions on each mesh interval and a set of $2k - 2m$ collocation points $\{\sigma_i\}$ which will be chosen by three different ways. A particular Hermite-Birkhoff interpolation problem can be used in a uniformly superconvergent approximation problem in a manner to be seen in the next chapter. The uniformly superconvergent piecewise polynomial is going to have order $2k$. In general there may be no solution or many solutions to Hermite-Birkhoff problems. For some $\{\sigma_i\}$ this problem has a unique solution, but for some $\{\sigma_i\}$ it does not. The existence and uniqueness of the Hermite-Birkhoff problem follows from the determinant of a coefficient matrix associated with $p(t)$ being nonzero. In this chapter we are going to use a monomial representation for $p(t)$ and the above $2k$ interpolation conditions to construct a linear system whose unknowns are the coefficients of $p(t)$. The choice of $\{\sigma_i\}$ is kept symmetric,

$$\sigma_1 > \sigma_2 > \cdots > \sigma_{k-m} > 0.$$

and for $i > k - m$

$$\sigma_i = -\sigma_{2k-2m+1-i}.$$

2.2 Well-posedness of the Hermite-Birkhoff problem using polynomial p(t) of order 2k

If we write the polynomial $p(t)$ as

$$p(t) = c_1 + c_2 t + c_3 t^2 + \ldots + c_{2k} t^{2k-1}. \tag{2.2}$$

then

$$(D^{j-1}p)(t) = c_1 D^{j-1}t^0 + c_2 D^{j-1}t^1 + \ldots + c_{2k} D^{j-1}t^{2k-1} = \sum_{i=1}^{2k} c_i D^{j-1}t^{i-1}. \tag{2.3}$$

for $j = 1, 2, \ldots, m$. When we substitute 1 for t in $(D^{j-1}p)(t)$, we can construct an m by $2k$ coefficient matrix, and when we use -1 instead of 1 we can construct another m by $2k$ coefficient matrix. Finally, if we use the condition that $(D^m p)(\sigma_i)$ is given for $i = 1, 2, \ldots, 2k - 2m$, we can get a $(2k - 2m)$ by $2k$ matrix. Therefore we can construct the entire $2k$ by $2k$ coefficient matrix M. In particular, for $t = 1$

$$\begin{pmatrix} 1 & 1 & 1 & \dots & \dots & 1 \\ 0 & 1 & 2 & \dots & \dots & 2k-1 \\ 0 & 0 & 2*1 & 3*2 & \dots & (2k-1)(2k-2) \\ 0 & 0 & 0 & 3*2*1 & \dots & (2k-1)(2k-2)(2k-3) \\ \vdots & \dots & \dots & \dots & \dots & \vdots \\ 0 & \dots & 0 & P^{m-1}_{m-1} & \dots & P^{m-1}_{2k-1} \end{pmatrix} \tag{2.4}$$

for $t = -1$

$$\begin{pmatrix} 1 & -1 & 1 & \dots & \dots & (-1)^{2k-1} * 1 \\ 0 & 1 & -2 & \dots & \dots & (-1)^{2k-2} * 2k-1 \\ 0 & 0 & 2*1 & -3*2 & \dots & (-1)^{2k-3} * (2k-1)(2k-2) \\ 0 & 0 & 0 & 3*2*1 & \dots & (-1)^{2k-4} * (2k-1)(2k-2)(2k-3) \\ \vdots & \dots & \dots & \dots & \dots & \vdots \\ 0 & \dots & 0 & P^{m-1}_{m-1} & \dots & (-1)^{2k-m} * P^{m-1}_{2k-1} \end{pmatrix} \tag{2.5}$$

for $t = \sigma_i$

$$\begin{pmatrix} 0 & \dots & P^m_m \sigma^0_1 & P^m_{m+1} \sigma^1_1 & \dots & P^m_{2k-1} \sigma_1^{2k-m-1} \\ 0 & \dots & P^m_m \sigma^0_2 & P^m_{m+1} \sigma^1_2 & \dots & P^m_{2k-1} \sigma_2^{2k-m-1} \\ 0 & \dots & P^m_m \sigma^0_3 & P^m_{m+1} \sigma^1_3 & \dots & P^m_{2k-1} \sigma_3^{2k-m-1} \\ \vdots & \dots & \dots & \dots & \dots & \vdots \\ \vdots & \dots & \dots & \dots & \dots & \vdots \\ 0 & \dots & P^m_m \sigma^0_{2k-2m} & P^m_{m+1} \sigma^1_{2k-2m} & \dots & P^m_{2k-1} \sigma_{2k-2m}^{2k-m-1} \end{pmatrix} . \tag{2.6}$$

Here we define $P_i^j = i!/(i-j)!$ for convenience. After combining equations (2.4),(2.5) and (2.6) we can get the coefficient matrix $M :=$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & \cdots & \cdots & \cdots & 1 \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ 0 & \cdots & 0 & P_{m-1}^{m-1} & P_{m}^{m-1} & \cdots & \cdots & P_{2k-2}^{m-1} & P_{2k-1}^{m-1} \\ 1 & -1 & 1 & -1 & 1 & \cdots & \cdots & \cdots & (-1)^{2k-1} * 1 \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ 0 & \cdots & 0 & P_{m-1}^{m-1} & -P_{m}^{m-1} & \cdots & \cdots & (-1)^{2k-m-1} P_{2k-2}^{m-1} & (-1)^{2k-m} P_{2k-1}^{m-1} \\ 0 & \cdots & 0 & P_{m}^{m} \sigma_1^0 & P_{m+1}^{m} \sigma_1^1 & \cdots & \cdots & P_{2k-2}^{m} \sigma_1^{2k-m} & P_{2k-1}^{m} \sigma_1^{2k-m-1} \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ 0 & \cdots & 0 & P_{m}^{m} \sigma_{2k-2m}^0 & P_{m+1}^{m} \sigma_{2k-2m}^1 & \cdots & \cdots & P_{2k-2}^{m} \sigma_{2k-2m}^{2k-m} & P_{2k-1}^{m} \sigma_{2k-2m}^{2k-m-1} \end{pmatrix} . \tag{2.7}$$

The determinant of this matrix for general m and k is not given in this paper. But some interesting patterns are found for individual cases using Mathematica (see Wolfram, 1991). As an example for $m = 2$ and $k = 4$ we can construct the coefficient matrix M.

$$M = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 0 & 0 & 2 & 6\sigma_1 & 12\sigma_1^2 & 20\sigma_1^3 & 30\sigma_1^4 & 42\sigma_1^5 \\ 0 & 0 & 2 & 6\sigma_2 & 12\sigma_2^2 & 20\sigma_2^3 & 30\sigma_2^4 & 42\sigma_2^5 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 0 & 1 & -2 & 3 & -4 & 5 & -6 & 7 \\ 0 & 0 & 2 & -6\sigma_1 & 12\sigma_1^2 & -20\sigma_1^3 & 30\sigma_1^4 & -42\sigma_1^5 \\ 0 & 0 & 2 & -6\sigma_2 & 12\sigma_2^2 & -20\sigma_2^3 & 30\sigma_2^4 & -42\sigma_2^5 \end{pmatrix} . \tag{2.8}$$

The determinant of the matrix M is evaluated as follows.

$$det\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 0 & 0 & 2 & 6\sigma_2 & 12\sigma_2^2 & 20\sigma_2^3 & 30\sigma_2^4 & 42\sigma_2^5 \\ 0 & 0 & 2 & 6\sigma_1 & 12\sigma_1^2 & 20\sigma_1^3 & 30\sigma_1^4 & 42\sigma_1^5 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 0 & 1 & -2 & 3 & -4 & 5 & -6 & 7 \\ 0 & 0 & 2 & -6\sigma_2 & 12\sigma_2^2 & -20\sigma_2^3 & 30\sigma_2^4 & -42\sigma_2^5 \\ 0 & 0 & 2 & -6\sigma_1 & 12\sigma_1^2 & -20\sigma_1^3 & 30\sigma_1^4 & -42\sigma_1^5 \end{pmatrix} \tag{2.9}$$

$$= C det\begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 3 & 0 & 5 & 0 & 7 \\ 0 & 0 & 2 & 0 & 4 & 0 & 6 & 0 \\ 0 & 0 & 2 & 0 & 12\sigma_1^2 & 0 & 30\sigma_1^4 & 0 \\ 0 & 0 & 0 & 6\sigma_1 & 0 & 20\sigma_1^3 & 0 & 42\sigma_1^5 \\ 0 & 0 & 2 & 0 & 12\sigma_2^2 & 0 & 30\sigma_2^4 & 0 \\ 0 & 0 & 0 & 6\sigma_2 & 0 & 20\sigma_2^3 & 0 & 42\sigma_2^5 \end{pmatrix} . \tag{2.10}$$

After appropriate row and column interchanges this becomes

$$Cdet\begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 4 & 6 & 0 & 0 & 0 & 0 \\ 0 & 2 & 12\sigma_1^2 & 30\sigma_1^4 & 0 & 0 & 0 & 0 \\ 0 & 2 & 12\sigma_2^2 & 30\sigma_2^4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 3 & 5 & 7 \\ 0 & 0 & 0 & 0 & 0 & 6\sigma_1 & 20\sigma_1^3 & 42\sigma_1^5 \\ 0 & 0 & 0 & 0 & 0 & 6\sigma_2 & 20\sigma_2^3 & 42\sigma_2^5 \end{pmatrix}. \tag{2.11}$$

$$= Cdet\begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 2 & 4 & 6 \\ 0 & 2 & 12\sigma_1^2 & 30\sigma_1^4 \\ 0 & 2 & 12\sigma_2^2 & 30\sigma_2^4 \end{pmatrix} * Cdet\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 3 & 5 & 7 \\ 0 & 6\sigma_1 & 20\sigma_1^3 & 42\sigma_1^5 \\ 0 & 6\sigma_2 & 20\sigma_2^3 & 42\sigma_2^5 \end{pmatrix} \tag{2.12}$$

$$= C(\sigma_2^2 - \sigma_1^2)(3 - 5\sigma_1^2 - 5\sigma_2^2 + 15\sigma_1^2\sigma_2^2) * \sigma_1\sigma_2(\sigma_2^2 -$$

$$\sigma_1^2)(15 - 21\sigma_1^2 - 21\sigma_2^2 + 35\sigma_1^2\sigma_2).$$

In general, we conjecture that

$$detM = CF_0F_1F_2 \tag{2.13}$$

where C is a generic constant. The three factors F_0, F_1 and F_2 are polynomials in the various σ_i. Because the Legendre polynomials are commonly occuring factors, we let $P_m(x)$ stand for the mth degree Legendre polynomial. The normalization constant will be absorbed into C. With proper column and row interchange the determinant of M can be written as;

$$detM = Cdet \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \tag{2.14}$$

where the matrices A and B are each k by k and C is a constant. So $detM = CdetA * detB$. When $\sigma_i = 0$, $detM$ is zero for any k and m and when $\sigma_i = \pm\sigma_j$ for $k \geq m + 2$, $detM$ is also zero. Factor F_0 accounts for these two properties; i.e.,

$$F_0 = \prod \sigma_j \cdot \prod_{i>j} (\sigma_i^2 - \sigma_j^2)^2.$$

F_1 comes from $detA$ after factoring out $\sigma_i^2 - \sigma_j^2$ and possibly σ_i. F_2 comes from $detB$ after factoring out $\sigma_i - \sigma_j$ and possibly σ_i. In general A has the form

$$\begin{pmatrix} 1 & 1 & 1 & \cdots & \cdots & \cdots & 1 & 1 \\ 0 & 2 & 4 & \cdots & \cdots & \cdots & 2k-4 & 2k-2 \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ 0 & \cdots & 0 & P_{2n-2}^{2n-2} & P_{2n}^{2n-2} & \cdots & \cdots & P_{2k-2}^{2n-2} \\ 0 & \cdots & 0 & 0 & P_{2n}^{2n-1} & \cdots & \cdots & P_{2k-2}^{2n-1} \\ 0 & \cdots & 0 & 0 & P_{2n}^{2n}\sigma_1^0 & P_{2n+2}^{2n}\sigma_1^2 & \cdots & P_{2k-2}^{2n}\sigma_1^{2k-2n-2} \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ 0 & \cdots & 0 & 0 & P_{2n}^{2n}\sigma_{k-m}^0 & P_{2n+2}^{2n}\sigma_{k-m}^2 & \cdots & P_{2k-2}^{2n}\sigma_{k-m}^{2k-2n-2} \end{pmatrix} \tag{2.15}$$

when $m = 2n$ for integer n

$$\begin{pmatrix} 1 & 1 & 1 & \dots & \dots & \dots & 1 & 1 \\ 0 & 2 & 4 & \dots & \dots & \dots & 2k-4 & 2k-2 \\ \vdots & \dots & \dots & \dots & \dots & \dots & \dots & \vdots \\ 0 & \dots & 0 & P^{2n-1}_{2n} & P^{2n-1}_{2n+2} & \dots & \dots & P^{2n-1}_{2k-2} \\ 0 & \dots & 0 & P^{2n}_{2n} & P^{2n}_{2n+2} & \dots & \dots & P^{2n}_{2k-2} \\ 0 & \dots & 0 & 0 & P^{2n+1}_{2n+2}\sigma^{1}_{1} & P^{2n+1}_{2n+4}\sigma^{3}_{1} & \dots & P^{2n+1}_{2k-2}\sigma^{2k-2n-3}_{1} \\ \vdots & \dots & \dots & \dots & \dots & \dots & \dots & \vdots \\ 0 & \dots & 0 & 0 & P^{2n+1}_{2n+2}\sigma^{1}_{k-m} & P^{2n+1}_{2n+2}\sigma^{3}_{k-m} & \dots & P^{2n+1}_{2k-2}\sigma^{2k-2n-3}_{k-m} \end{pmatrix} \tag{2.16}$$

when $m = 2n + 1$ for integer n.

For the matrix B;

$$\begin{pmatrix} 1 & 1 & 1 & \dots & \dots & \dots & 1 & 1 \\ 1 & 3 & 5 & \dots & \dots & \dots & 2k-3 & 2k-1 \\ \vdots & \dots & \dots & \dots & \dots & \dots & \dots & \vdots \\ 0 & \dots & 0 & P^{2n-2}_{2n-1} & P^{2n-2}_{2n+1} & \dots & \dots & P^{2n-2}_{2k-1} \\ 0 & \dots & 0 & P^{2n-1}_{2n-1} & P^{2n-1}_{2n+1} & \dots & \dots & P^{2n-1}_{2k-1} \\ 0 & \dots & 0 & 0 & P^{2n}_{2n+1}\sigma^{1}_{1} & P^{2n}_{2n+3}\sigma^{3}_{1} & \dots & P^{2n}_{2k-1}\sigma^{2k-2n-1}_{1} \\ \vdots & \dots & \dots & \dots & \dots & \dots & \dots & \vdots \\ 0 & \dots & 0 & 0 & P^{2n}_{2n+1}\sigma^{1}_{k-m} & P^{2n}_{2n+3}\sigma^{3}_{k-m} & \dots & P^{2n}_{2k-1}\sigma^{2k-2n-1}_{k-m} \end{pmatrix} \tag{2.17}$$

when $m = 2n$ for integer n

$$\begin{pmatrix} 1 & 1 & 1 & \dots & \dots & \dots & 1 & 1 \\ 1 & 3 & 5 & \dots & \dots & \dots & 2k-3 & 2k-1 \\ \vdots & \dots & \dots & \dots & \dots & \dots & \dots & \vdots \\ 0 & \dots & 0 & P^{2n-1}_{2n-1} & P^{2n-1}_{2n+1} & \dots & \dots & P^{2n-1}_{2k-1} \\ 0 & \dots & 0 & P^{2n}_{2n-1} & P^{2n}_{2n+1} & \dots & \dots & P^{2n}_{2k-1} \\ 0 & \dots & 0 & 0 & P^{2n+1}_{2n+1}\sigma_1^0 & P^{2n+1}_{2n+3}\sigma_1^2 & \dots & P^{2n+1}_{2k-1}\sigma_1^{2k-2n-2} \\ \vdots & \dots & \dots & \dots & \dots & \dots & \dots & \vdots \\ 0 & \dots & 0 & 0 & P^{2n+1}_{2n+1}\sigma^0_{k-m} & P^{2n+1}_{2n+3}\sigma^2_{k-m} & \dots & P^{2n+1}_{2k-1}\sigma^{2k-2n-2}_{k-m} \end{pmatrix} \tag{2.18}$$

when $m = 2n+1$ for integer n. While we were unable to find general formulas for $detA$ and $detB$, the following individual cases can be studied.

Case 1): $m = 1, k = 2$

$$detA = det \begin{pmatrix} 1 & 1 \\ 0 & \sigma_1 \end{pmatrix} = \sigma_1 \tag{2.19}$$

$$detB = det \begin{pmatrix} 1 & 1 \\ 1 & 3\sigma_1^2 \end{pmatrix} = -1 + 3\sigma_1^2. \tag{2.20}$$

Then in (2.13)

$$F_0(\sigma_1) = \sigma_1$$

$$F_1(\sigma_1) = 1$$

$$F_2(\sigma_1) = -1 + 3\sigma_1^2 = P_2(\sigma_1).$$

Case 2): $m = 1, k = 3$

$$detA = det \begin{pmatrix} 1 & 1 & 1 \\ 0 & 2\sigma_1 & 4\sigma_1^3 \\ 0 & 2\sigma_2 & 4\sigma_2^3 \end{pmatrix} = 8\sigma_1\sigma_2(\sigma_2^2 - \sigma_1^2) \tag{2.21}$$

$$detB = det \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3\sigma_1^2 & 5\sigma_1^4 \\ 1 & 3\sigma_2^2 & 5\sigma_2^4 \end{pmatrix} = (\sigma_2^2 - \sigma_1^2)(3 - 5\sigma_1^2 - 5\sigma_2^2 + 15\sigma_1^2\sigma_2^2). \tag{2.22}$$

Then in (2.13)

$$F_0(\sigma_1, \sigma_2) = \sigma_1\sigma_2(\sigma_2^2 - \sigma_1^2)^2$$

$$F_1(\sigma_1, \sigma_2) = 1$$

$$F_2(\sigma_1, \sigma_2) = 3 - 5\sigma_1^2 - 5\sigma_2^2 + 15\sigma_1^2\sigma_2^2.$$

Case 3): $m = 1, k = 4$

$$detA = det \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 2\sigma_1 & 4\sigma_1^3 & 6\sigma_1^5 \\ 0 & 2\sigma_2 & 4\sigma_2^3 & 6\sigma_2^5 \\ 0 & 2\sigma_3 & 4\sigma_3^3 & 6\sigma_3^5 \end{pmatrix} = 48\sigma_1\sigma_2\sigma_3(\sigma_2^2 - \sigma_1^2)(\sigma_3^2 - \sigma_1^2)(\sigma_3^2 - \sigma_2^2) \tag{2.23}$$

$$detB = det \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 3\sigma_1^2 & 5\sigma_1^4 & 7\sigma_1^6 \\ 1 & 3\sigma_2^2 & 5\sigma_2^4 & 7\sigma_2^6 \\ 1 & 3\sigma_3^2 & 5\sigma_3^4 & 7\sigma_3^6 \end{pmatrix} \tag{2.24}$$

$$= (\sigma_2^2 - \sigma_1^2)(\sigma_3^2 - \sigma_1^2)(\sigma_3^2 - \sigma_2^2)(-15 + 21\sigma_1^2 + 21\sigma_2^2 + 21\sigma_3^2 -$$

$$35\sigma_1^2\sigma_2^2 - 35\sigma_1^2\sigma_3^2 - 35\sigma_2^2\sigma_3^2 + 105\sigma_1^2\sigma_2^2\sigma_3^2).$$

Then in (2.13)

$$F_0(\sigma_1, \sigma_2, \sigma_3) = \sigma_1\sigma_2\sigma_3(\sigma_2^2 - \sigma_1^2)^2(\sigma_3^2 - \sigma_1^2)^2(\sigma_3^2 - \sigma_2^2)^2$$

$$F_1(\sigma_1, \sigma_2, \sigma_3) = 1$$

$$F_2(\sigma_1, \sigma_2, \sigma_3) = 15 - 21\sigma_1^2 - 21\sigma_2^2 - 21\sigma_3^2 + 35\sigma_1^2\sigma_2^2 + 35\sigma_1^2\sigma_3^2 + 35\sigma_2^2\sigma_3^2 - 105\sigma_1^2\sigma_2^2\sigma_3^2.$$

Case 4): $m = 2, k = 3$

$$detA = det\begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & 4 \\ 0 & 2 & 12\sigma_1^2 \end{pmatrix} = 8(-1 + 3\sigma_1^2) \tag{2.25}$$

$$detB = det\begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \\ 0 & 6\sigma_1 & 20\sigma_1^3 \end{pmatrix} = 8\sigma_1(-3 + 5\sigma_1^2). \tag{2.26}$$

Then in (2.13)

$$F_0(\sigma_1) = \sigma_1$$

$$F_1(\sigma_1) = P_2(\sigma_1)$$

$$F_2(\sigma_1) = 5\sigma_1^2 - 3.$$

Case 5): $m = 2, k = 4$

$$detA = det \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 2 & 4 & 6 \\ 0 & 2 & 12\sigma_1^2 & 30\sigma_1^4 \\ 0 & 2 & 12\sigma_2^2 & 30\sigma_2^4 \end{pmatrix} = 48(\sigma_2^2 - \sigma_1^2)(3 - 5\sigma_1^2 - 5\sigma_2^2 + 15\sigma_1^2\sigma_2^2) \quad (2.27)$$

$$detB = det \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 3 & 5 & 7 \\ 0 & 6\sigma_1 & 20\sigma_1^3 & 42\sigma_1^5 \\ 0 & 6\sigma_1 & 20\sigma_1^3 & 42\sigma_1^5 \end{pmatrix} \quad (2.28)$$

$= 48\sigma_1\sigma_2(\sigma_2^2 - \sigma_1^2)(15 - 21\sigma_1^2 - 21\sigma_2^2 + 35\sigma_1^2\sigma_2)$.

Then in (2.13)

$$F_0(\sigma_1, \sigma_2) = \sigma_1\sigma_2(\sigma_2^2 - \sigma_1^2)^2$$

$$F_1(\sigma_1, \sigma_2) = 3 - 5\sigma_1^2 - 5\sigma_2^2 + 15\sigma_1^2\sigma_2^2$$

$$F_2(\sigma_1, \sigma_2) = 15 - 21\sigma_1^2 - 21\sigma_2^2 + 35\sigma_1^2\sigma_2^2.$$

Case 6): $m = 2, k = 5$

$$detA = det \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 2 & 4 & 6 & 8 \\ 0 & 2 & 12\sigma_1^2 & 30\sigma_1^4 & 56\sigma_1^6 \\ 0 & 2 & 12\sigma_2^2 & 30\sigma_2^4 & 56\sigma_2^6 \\ 0 & 2 & 12\sigma_3^2 & 30\sigma_3^4 & 56\sigma_3^6 \end{pmatrix} \quad (2.29)$$

$$= 384(\sigma_2^2 - \sigma_1^2)(\sigma_3^2 - \sigma_1^2)(\sigma_3^2 - \sigma_2^2)(-15 + 21\sigma_1^2 + 21\sigma_2^2 + 21\sigma_3^2 -$$

$$35\sigma_1^2\sigma_2^2 - 35\sigma_1^2\sigma_3^2 - 35\sigma_2^2\sigma_3^2 + 105\sigma_1^2\sigma_2^2\sigma_3^2)$$

$$detB = det\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 5 & 7 & 9 \\ 0 & 6\sigma_1 & 20\sigma_1^3 & 42\sigma_1^5 & 72\sigma_1^7 \\ 0 & 6\sigma_2 & 20\sigma_2^3 & 42\sigma_2^5 & 72\sigma_2^7 \\ 0 & 6\sigma_3 & 20\sigma_3^3 & 42\sigma_3^5 & 72\sigma_3^7 \end{pmatrix} \tag{2.30}$$

$$= 1152\sigma_1\sigma_2\sigma_3(\sigma_2^2 - \sigma_1^2)(\sigma_3^2 - \sigma_1^2)(\sigma_3^2 - \sigma_2^2)(-35 + 45\sigma_1^2 + 45\sigma_2^2 + 45\sigma_3^2 -$$

$$63\sigma_1^2\sigma_2^2 - 63\sigma_1^2\sigma_3^2 - 63\sigma_2^2\sigma_3^2 + 105\sigma_1^2\sigma_2^2\sigma_3^2).$$

Then in (2.13)

$$F_0(\sigma_1, \sigma_2, \sigma_3) = \sigma_1\sigma_2\sigma_3(\sigma_2^2 - \sigma_1^2)^2(\sigma_3^2 - \sigma_1^2)^2(\sigma_3^2 - \sigma_2^2)^2$$

$$F_1(\sigma_1, \sigma_2, \sigma_3) = 15 - 21\sigma_1^2 - 21\sigma_2^2 - 21\sigma_3^2 + 35\sigma_1^2\sigma_2^2 + 35\sigma_1^2\sigma_3^2 + 35\sigma_2^2\sigma_3^2 - 105\sigma_1^2\sigma_2^2\sigma_3^2$$

$$F_2(\sigma_1, \sigma_2, \sigma_3) = 35 - 45\sigma_1^2 - 45\sigma_2^2 - 45\sigma_3^2 + 63\sigma_1^2\sigma_2^2 + 63\sigma_1^2\sigma_3^2 + 63\sigma_2^2\sigma_3^2 - 105\sigma_1^2\sigma_2^2\sigma_3^2.$$

Case 7): $m = 2, k = 6$

$$detA = det\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 2 & 4 & 6 & 8 & 10 \\ 0 & 2 & 12\sigma_1^2 & 30\sigma_1^4 & 56\sigma_1^6 & 90\sigma_1^8 \\ 0 & 2 & 12\sigma_2^2 & 30\sigma_2^4 & 56\sigma_2^6 & 90\sigma_2^8 \\ 0 & 2 & 12\sigma_3^2 & 30\sigma_3^4 & 56\sigma_3^6 & 90\sigma_3^8 \\ 0 & 2 & 12\sigma_4^2 & 30\sigma_4^4 & 56\sigma_4^6 & 90\sigma_4^8 \end{pmatrix} \tag{2.31}$$

$$= 11520(\sigma_2^2 - \sigma_1^2)(\sigma_3^2 - \sigma_1^2)(\sigma_3^2 - \sigma_2^2)(\sigma_4^2 - \sigma_1^2)(\sigma_4^2 - \sigma_2^2)(\sigma_4^2 - \sigma_3^2)$$

$$(35 - 45\sigma_1^2 - 45\sigma_2^2 - 45\sigma_3^2 - 45\sigma_4^2 + 63\sigma_1^2\sigma_2^2 + 63\sigma_1^2\sigma_3^2 + 63\sigma_1^2\sigma_4^2 + 63\sigma_2^2\sigma_3^2 +$$

$$63\sigma_3^2\sigma_4^2 - 105\sigma_1^2\sigma_2^2\sigma_3^2 - 105\sigma_1^2\sigma_2^2\sigma_4^2 - 105\sigma_1^2\sigma_3^2\sigma_4^2 - 105\sigma_2^2\sigma_3^2\sigma_4^2 + 315\sigma_1^2\sigma_2^2\sigma_3^2\sigma_4^2)$$

$$detB = det\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 5 & 7 & 9 & 11 \\ 0 & 6\sigma_1 & 20\sigma_1^3 & 42\sigma_1^5 & 72\sigma_1^7 & 110\sigma_1^9 \\ 0 & 6\sigma_2 & 20\sigma_2^3 & 42\sigma_2^5 & 72\sigma_2^7 & 110\sigma_2^9 \\ 0 & 6\sigma_3 & 20\sigma_3^3 & 42\sigma_3^5 & 72\sigma_3^7 & 110\sigma_3^9 \\ 0 & 6\sigma_4 & 20\sigma_4^3 & 42\sigma_4^5 & 72\sigma_4^7 & 110\sigma_4^9 \end{pmatrix} \qquad (2.32)$$

$$= 11520\sigma_1\sigma_2\sigma_3\sigma_3(\sigma_2^2 - \sigma_1^2)(\sigma_3^2 - \sigma_1^2)$$

$$(\sigma_4^2 - \sigma_1^2)(\sigma_3^2 - \sigma_2^2)(\sigma_4^2 - \sigma_2^2)(\sigma_4^2 - \sigma_3^2)$$

$$(315 - 385\sigma_1^2 - 385\sigma_2^2 - 385\sigma_3^2 - 385\sigma_4^2 + 495\sigma_1^2\sigma_2^2 + 495\sigma_1^2\sigma_3^2 +$$

$$495\sigma_1^2\sigma_4^2 + 495\sigma_2^2\sigma_3^2 + 495\sigma_3^2\sigma_4^2 - 693\sigma_1^2\sigma_2^2\sigma_3^2 -$$

$$693\sigma_1^2\sigma_2^2\sigma_4^2 - 693\sigma_1^2\sigma_3^2\sigma_4^2 - 693\sigma_2^2\sigma_3^2\sigma_4^2 + 1155\sigma_1^2\sigma_2^2\sigma_3^2\sigma_4^2).$$

Then in (2.13)

$$F_0(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = \sigma_1\sigma_2\sigma_3\sigma_4 \prod_{i>j}(\sigma_i^2 - \sigma_j^2)^2$$

$$F_1(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = 35 - 45(\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2) + 63(\sigma_1^2\sigma_2^2 + \sigma_1^2\sigma_3^2 + \sigma_1^2\sigma_4^2 + \sigma_2^2\sigma_3^2 + \sigma_2^2\sigma_4^2 + \sigma_3^2\sigma_4^2)$$

$$-105(\sigma_1^2\sigma_2^2\sigma_3^2 + \sigma_1^2\sigma_2^2\sigma_4^2 + \sigma_1^2\sigma_3^2\sigma_4^2 + \sigma_2^2\sigma_3^2\sigma_4^2) - 315\sigma_1^2\sigma_2^2\sigma_3^2\sigma_4^2$$

$$F_2(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = 315 - 385(\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2) + 495(\sigma_1^2\sigma_2^2 + \sigma_1^2\sigma_3^2 + \sigma_1^2\sigma_4^2 + \sigma_2^2\sigma_3^2 + \sigma_2^2\sigma_4^2 + \sigma_3^2\sigma_4^2)$$

$$-693(\sigma_1^2\sigma_2^2\sigma_3^2+\sigma_1^2\sigma_2^2\sigma_4^2+\sigma_1^2\sigma_3^2\sigma_4^2+\sigma_2^2\sigma_3^2\sigma_4^2)-1155\sigma_1^2\sigma_2^2\sigma_3^2\sigma_4^2.$$

Case 8): $m=3, k=4$

$$detA = det\begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 2 & 4 & 6 \\ 0 & 2 & 12 & 30 \\ 0 & 0 & 24\sigma_1 & 120\sigma_1^3 \end{pmatrix} = 384\sigma_1(-3+5\sigma_1^2) \tag{2.33}$$

$$detB = det\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 3 & 5 & 7 \\ 0 & 6 & 20 & 42 \\ 0 & 6 & 60\sigma_1^2 & 210\sigma_1^4 \end{pmatrix} = 96(3-30\sigma_1^2+35\sigma_1^4). \tag{2.34}$$

Then in (2.13)

$$F_0(\sigma_1)=\sigma_1$$

$$F_1(\sigma_1)=5\sigma_1^2-3$$

$$F_2(\sigma_1)=P_4(\sigma_1).$$

Case 9): $m=4, k=5$

$$detA = det\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 2 & 4 & 6 & 8 \\ 0 & 2 & 12 & 30 & 56 \\ 0 & 0 & 24 & 120 & 336 \\ 0 & 0 & 24 & 360\sigma_1^2 & 1680\sigma_1^4 \end{pmatrix} = 36864(3-30\sigma_1^2+35\sigma_1^4) \tag{2.35}$$

$$detB = det\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 5 & 7 & 9 \\ 0 & 6 & 20 & 42 & 72 \\ 0 & 6 & 60 & 210 & 504 \\ 0 & 0 & 120\sigma_1 & 840\sigma_1^3 & 3024\sigma_1^5 \end{pmatrix} = 36864\sigma_1(15 - 70\sigma_1^2 + 63\sigma_1^4). \quad (2.36)$$

Then in (2.13)

$$F_0(\sigma_1) = \sigma_1$$

$$F_1(\sigma_1) = P_4(\sigma_1)$$

$$F_2(\sigma_1) = 15 - 70\sigma_1^2 + 63\sigma_1^4.$$

Case 10): $m = 4, k = 6$

$$detA = det\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 2 & 4 & 6 & 8 & 10 \\ 0 & 2 & 12 & 30 & 56 & 90 \\ 0 & 0 & 24 & 120 & 336 & 720 \\ 0 & 0 & 24 & 360\sigma_1^2 & 1680\sigma_1^4 & 5040\sigma_1^6 \\ 0 & 0 & 24 & 360\sigma_2^2 & 1680\sigma_2^4 & 5040\sigma_2^6 \end{pmatrix} \quad (2.37)$$

$$= 8847360(\sigma_2^2 - \sigma_1^2)(15 - 70\sigma_1^2 - 70\sigma_2^2 +$$

$$63\sigma_1^4 + 63\sigma_2^4 + 588\sigma_1^2\sigma_2^2 - 630\sigma_1^4\sigma_2^2 - 630\sigma_1^2\sigma_2^4 + 735\sigma_1^4\sigma_2^4)$$

$$detB = det \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 5 & 7 & 9 & 11 \\ 0 & 6 & 20 & 42 & 72 & 110 \\ 0 & 6 & 60 & 210 & 504 & 990 \\ 0 & 0 & 120\sigma_1 & 840\sigma_1^3 & 3024\sigma_1^5 & 7920\sigma_1^7 \\ 0 & 0 & 120\sigma_2 & 840\sigma_2^3 & 3024\sigma_2^5 & 7920\sigma_2^7 \end{pmatrix} \qquad (2.38)$$

$$= 8847360\sigma_1\sigma_2(\sigma_2^2 - \sigma_1^2)(175 - 630\sigma_1^2 - 630\sigma_2^2 + 495\sigma_1^4 + 495\sigma_2^4 +$$

$$2700\sigma_1^2\sigma_2^2 - 2310\sigma_1^4\sigma_2^2 - 2310\sigma_1^2\sigma_2^4 + 2079\sigma_1^4\sigma_2^4).$$

Then in (2.13)

$$F_0(\sigma_1, \sigma_2) = \sigma_1\sigma_2(\sigma_2^2 - \sigma_1^2)^2$$

$$F_1(\sigma_1, \sigma_2) = 15 - 70\sigma_1^2 - 70\sigma_2^2 + 63\sigma_1^4 + 63\sigma_2^4 +$$

$$588\sigma_1^2\sigma_2^2 - 630\sigma_1^4\sigma_2^2 - 630\sigma_1^2\sigma_2^4 + 735\sigma_1^4\sigma_2^4$$

$$F_2(\sigma_1, \sigma_2) = 175 - 630\sigma_1^2 - 630\sigma_2^2 + 495\sigma_1^4 + 495\sigma_2^4 +$$

$$2700\sigma_1^2\sigma_2^2 - 2310\sigma_1^4\sigma_2^2 - 2310\sigma_1^2\sigma_2^4 + 2079\sigma_1^4\sigma_2^4.$$

In cases (1),(4),(8), and (9), $detA$ and $detB$ were Legendre polynomials. And when $detA$ is the mth degree Legendre polynomial, $detB$ turns out to be P_{m+1}. This suggests that when $k = m + 1$ the determinant of M is $CP_m(x)P_k(x)$. Moreover, it is verified by direct computation that F_2 in (2) and both F_1 and F_2 in (5) and (10) vanish for $\sigma_1 = \sqrt{(70 + \sqrt{480})/70}$ and $\sigma_2 = \sqrt{(70 - \sqrt{480})/70}$ the positive roots of $P_4(x)$. Similarly, F_2 in (3) and both F_1 and F_2 in (6) vanish when $\sigma_1, \sigma_2, \sigma_3$ are the positive roots of $P_6(x)$. Finally, both F_1 and F_2 in (7) vanish at the four positive roots of $P_8(x)$. So for all cases the matrix is singular when the roots of Legendre polynomials (Gauss points) are chosen as σ_i. Consequently, Gauss points produce an ill-posed Hermite-Birkhoff problem.

In contrast to choosing the Gauss points, it is easily checked that in all the above cases for k and m, a uniformly spaced choice of points either open (not including ±1) or closed (including ±1), has a unique solution to the corresponding Hermite-Birkhoff problem.

Chapter 3

FORMULA FOR A SINGLE EQUATION

All approximate solutions considered will be piecewise polynomials. The set of breakpoints for pieces is denoted by Δ. Hence for the finite interval $[a, b]$, Δ is a set with $\Delta = \{a = x_1 < x_2 < \cdots < x_{N+1} = b\}$; define $h_n = x_{n+1} - x_n$ and $h = \max h_n$. The symbol $P_{k,\Delta}$ denotes the space of piecewise polynomials of order k (degree $< k$) with breakpoints in Δ. When we produce an approximation $\hat{u}(x)$ in $P_{m+k,\Delta} \cap C^{m-1}[a, b]$ to the exact solution $u(x)$ using collocation at Gaussian points (see deBoor and Swartz, 1973), for sufficiently small h and for $k \geq m$ the error bound is

$$|D^{j-1}u - D^{j-1}\hat{u}| \leq Ch^{2k} \tag{3.1}$$

at the breakpoints. At general points x not in Δ,

$$|D^{j-1}u - D^{j-1}\hat{u}| \leq Ch^{m+k-j+1} \tag{3.2}$$

for some generic constant C. Here the order of the error bound for x in Δ is bigger than that for x not in Δ so higher order superconvergence occurs at the mesh points as long as $k > m$. In order to boost the order of the error bound to $2k$ for the points which are not in the breakpoints , one more approximation to $\hat{u}(x)$ using collocation is tried. For $\hat{u}(x)$ we needed the requirement that it satisfies the differential equation at k points in each interval and the m side conditions. Other side conditions were given from continuity conditions and the interpolation points in each interval were chosen to be the Gauss points: these collocation points are the zeros of the kth Legendre polynomial on each interval.

Therefore, $\hat{u}(x)$ is a piecewise polynomial of order $m + k$. The final uniformly superconvergent approximation $\bar{u}(x)$ is a piecewise polynomial of order $2k$. Hence we need $2k$ conditions for $\bar{u}(x)$. On each piece, we have $2m$ accurate (superconvergent) values, m at each end. Hence we have to pick $2(k - m)$ collocation points called secondary collocation points, for $\bar{u}(x)$. In this paper a particular Hermite-Birkhoff interpolation problem is used to motivate the choice of the $2(k-m)$ collocation points. As shown in chapter 2, the Gauss points are not good for the secondary collocation points. Some methods of choosing these secondary collocation points are mentioned, and tried numerically in chapter 4. Before we derive formulas for $\bar{u}(x)$, we have to prove the existence of $\bar{u}(x)$. Also, the stability of $\bar{u}(x)$ is needed. In order to accomplish these tasks we define some linear maps denoted by Q,Q_Δ,S_n. S_n is the one to one linear change of variable map from $C[x_n, x_{n+1}]$ to $C[-1, 1]$, Q is a linear projector which maps $C[-1, 1]$ to P_k(the space of polynomials of order k), and Q_Δ is the induced linear projector

$$Q_\Delta F := S_n^{-1} Q S_n F, \tag{3.3}$$

i.e.,

$$F|_{[x_n,x_{n+1}]} \rightarrow (S_n F)|_{[-1,1]} \rightarrow (QS_n F)|_{[-1,1]} \rightarrow (S_n^{-1} Q S_n F)|_{[x_n,x_{n+1}]}.$$

Theorem 1 *Assume that:*

1. *$\bar{u}(x)$ is a piecewise polynomial of order $2k$ with breakpoints in Δ.*
2. *$(D^{j-1}\bar{u})(x_n)$ is given for $1 \leq j \leq m$; $x_n \in \Delta$.*
3. *on each $[x_n, x_{n+1}]$*

$$D^m \bar{u} = \sum_{j=1}^{m} c_j D^{j-1} \bar{u} + f \tag{3.4}$$

evaluated at some σ_i.

4. *the associated Hermite-Birkhoff problem(2.1) is well-posed for this choice of* k, m *and* σ_i.

Then Q_Δ *exists and is bounded.*

Proof.

The proof consists of two steps; first we are going to find a coefficient matrix H for a problem on $C[-1,1]$, second we are going to get a coefficient matrix A for $C[x_n, x_{n+1}]$. And then these matrices H and A are going to be compared to prove that Q_Δ exists and is bounded.

Step 1) In $C[-1,1]$ let $p(t)$ be the polynomial of order $2k$ given by

$$p(t) = \sum_{j=1}^{k} (D^{j-1}p)(-1)\Phi_j(-t)(-1)^{j-1} + \sum_{j=1}^{k} (D^{j-1}p)(1)\Phi_j(t) \tag{3.5}$$

where $\Phi_j(t)$ satisfies

$$(D^{\nu-1}\Phi_j)(-1) = 0 \qquad 1 \le \nu \le k \tag{3.6}$$

$$(D^{\nu-1}\Phi_j)(1) = \delta_{\nu j} \qquad 1 \le \nu \le k \tag{3.7}$$

with D here denoting differentiation with respect to t. If we take $(j-1)$ derivatives on both sides in equation (3.5) then

$$(D^{j-1}p)(t) = \sum_{\nu=1}^{k} (D^{\nu-1}p)(-1)(D^{j-1}\Phi_\nu)(-t) + \sum_{\nu=1}^{k} (D^{\nu-1}p)(1)(D^{j-1}\Phi_\nu)(t). \tag{3.8}$$

To check whether this polynomial $p(t)$ can be related to Hermite-Birkhoff interpolation problem let's substitute -1 and 1 for t in equation (3.8).

For $t = -1$;

$$(D^{j-1}p)(-1) = \sum_{\nu=1}^{k}(D^{\nu-1}p)(-1)(D^{j-1}\Phi_\nu)(1) + \sum_{\nu=1}^{k}(D^{\nu-1}p)(1)(D^{j-1}\Phi_\nu)(-1) \quad (3.9)$$

the second term vanishes because of the character of the function Φ and the first term vanishes except when j equals ν. Therefore,

$$(D^{j-1}p)(-1) = (D^{j-1}p)(-1) \quad (3.10)$$

which means that the notation is consistent.

For $t = 1$

$$(D^{j-1}p)(1) = \sum_{\nu=1}^{k}(D^{\nu-1}p)(-1)(D^{j-1}\Phi_\nu)(-1) + \sum_{\nu=1}^{k}(D^{\nu-1}p)(1)(D^{j-1}\Phi_\nu)(1) \quad (3.11)$$

the first term vanishes because of the character of the function Φ and the second term vanishes except when j equals ν; again the notation is consistent. Now we can pick the $2k - 2m$ interpolation points between $-1 \leq t \leq 1$ to check the third condition of the Hermite-Birkhoff interpolation problem.

$$(D^m p)(\sigma_i) = \sum_{\nu=1}^{k}(D^{\nu-1}p)(-1)(D^m\Phi_\nu)(-\sigma_i) + \sum_{\nu=1}^{k}(D^{\nu-1}p)(1)(D^m\Phi_\nu)(\sigma_i) \quad (3.12)$$

where $(D^m p)(\sigma_i) = (D^m U)(\sigma_i)$ for $i = 1, 2, \ldots, 2k - 2m$. $(D^m U)(\sigma_i)$ is given and $D^{\nu-1}p(-1)$ is known for ν=1,2,...,m. The equation (3.12) is rearranged by knowns and unknowns.

After rearranging

$$\sum_{\nu=m+1}^{k}(D^{\nu-1}p)(-1)(D^m\Phi_\nu)(-\sigma_i)+\sum_{\nu=m+1}^{k}(D^{\nu-1}p)(1)(D^m\Phi_\nu)(\sigma_i)$$

$$=(D^m p)(\sigma_i)-\sum_{\nu=1}^{m}(D^{\nu-1}p)(-1)(D^m\Phi_\nu)(-\sigma_i)+\sum_{\nu=1}^{m}(D^{\nu-1}p)(1)(D^m\Phi_\nu)(\sigma_i). \tag{3.13}$$

This is a $2k-2m$ system, and we can express this system in matrix form as

$$Hx=b \tag{3.14}$$

where

$$b_i=(D^m p)(\sigma_i)-\sum_{\nu=1}^{m}(D^{\nu-1}p)(-1)(D^m\Phi_\nu)(-\sigma_i)+\sum_{\nu=1}^{m}(D^{\nu-1}p)(1)(D^m\Phi_\nu)(\sigma_i) \tag{3.15}$$

for $i=1,2,\ldots,2k-2m$.

$$x=\{(D^m p)(-1),(D^{m+1}p)(-1),\ldots,(D^{k-1}p)(-1),(D^m p)(1),\ldots,(D^{k-1}p)(1)\}^T$$

$$H=\begin{pmatrix}
D^m\Phi_{m+1}(-\sigma_1) & \ldots & D^m\Phi_k(-\sigma_1) & D^m\Phi_{m+1}(\sigma_1) & \ldots & D^m\Phi_k(\sigma_1)\\
D^m\Phi_{m+1}(-\sigma_2) & \ldots & D^m\Phi_k(-\sigma_2) & D^m\Phi_{m+2}(\sigma_2) & \ldots & D^m\Phi_k(\sigma_2)\\
\vdots & & & & & \vdots\\
\vdots & & & & & \vdots\\
\vdots & & & & & \vdots\\
D^m\Phi_{m+1}(-\sigma_{2k-2m}) & \ldots & D^m\Phi_k(-\sigma_{2k-2m}) & D^m\Phi_{m+1}(\sigma_{2k-2m}) & \ldots & D^m\Phi_k(\sigma_{2k-2m})
\end{pmatrix} \tag{3.16}$$

Step 2) For $C[x_n, x_{n+1}]$ a local Hermite representation is used for $\bar{u}$, i.e.,

$$\bar{u}(x) = \sum_{j=1}^{k}(D^{j-1}\bar{u})(x_n^+)\Phi_j((x_{n+1} + x_n - 2x)/h_n)((-h_n)/2)^{j-1}$$

$$+\sum_{j=1}^{k}(D^{j-1}\bar{u})(x_{n+1}^-)\Phi_j((2x - x_n - x_{n+1})/h_n)(h_n/2)^{j-1}$$

where $\Phi_j(t)$ is as before (3.6)-(3.7) and consequently $\bar{u}(x)$ is a piecewise polynomial of order $2k$ and is in $C^{m-1}[a,b]$ as long as

$$(D^{j-1}\bar{u})(x_n^-) = (D^{j-1}\bar{u})(x_n^+) = (D^{j-1}\bar{u})(x_n)$$

at the mesh points.

By assumption 2

$$(D^{\nu-1}\bar{u})(x_n)$$

$$= \sum_{j=1}^{k}(D^{j-1}\bar{u})(x_n^+)(D^{\nu-1}\Phi_j)((x_{n+1} + x_n - 2x)/h_n)((-h_n)/2)^{j-\nu}$$

$$+\sum_{j=1}^{k}(D^{j-1}\bar{u})(x_{n+1}^-)(D^{\nu-1}\Phi_j)((2x - x_{n+1} - x_n)/h_n)((h_n)/2)^{j-\nu} \tag{3.17}$$

for $1 \leq \nu \leq m$; $x_n \in \Delta$. If we substitute the above equation into equation (3.4)

$$\sum_{j=1}^{k}(D^{j-1}\bar{u})(x_n^+)(D^m\Phi_j)(-\sigma_i)(-h_n/2)^{j-1-m}$$

$$+\sum_{j=1}^{k}(D^{j-1}\bar{u})(x_{n+1}^-)(D^m\Phi_j)(\sigma_i)(h_n/2)^{j-1-m}$$

$$= \sum_{\nu=1}^{m} c_\nu(\sigma_{in})\{\sum_{j=1}^{k}(D^{j-1}\bar{u})(x_n^+)(D^{\nu-1}\Phi_j)(-\sigma_i)(-h_n/2)^{j-\nu}+$$

$$\sum_{j=1}^{k}(D^{j-1}\bar{u})(x_{n+1}^{-})(D^{\nu-1}\Phi_j)(\sigma_i)(h_n/2)^{j-\nu}\} + f(\sigma_{in}). \tag{3.18}$$

After rearranging this

$$\begin{aligned}
&\sum_{j=m+1}^{k}(D^{j-1}\bar{u})(x_n^{+})(-h_n/2)^{j-m-1}\{(D^m\Phi_j)(-\sigma_i)- \\
&\quad \sum_{\nu=1}^{m}c_\nu(\sigma_{in})(D^{\nu-1}\Phi_j)(-\sigma_i)(-h_n/2)^{m+1-\nu}\} \\
&+\sum_{j=m+1}^{k}(D^{j-1}\bar{u})(x_{n+1}^{-})(h_n/2)^{j-m-1}\{(D^m\Phi_j)(\sigma_i)- \\
&\quad \sum_{\nu=1}^{m}c_\nu(\sigma_{in})(D^{\nu-1}\Phi_j)(\sigma_i)(h_n/2)^{m+1-\nu}\} \\
&= f(\sigma_{in})-\sum_{j=1}^{m}(D^{j-1}\bar{u})(x_n^{+})(-h_n/2)^{j-m-1}\{(D^m\Phi_j)(-\sigma_i)- \\
&\quad \sum_{\nu=1}^{m}c_\nu(\sigma_{in})(D^{\nu-1}\Phi_j)(-\sigma_i)(-h_n/2)^{m+1-\nu}\} \\
&+\sum_{j=1}^{m}(D^{j-1}\bar{u})(x_{n+1}^{-})(h_n/2)^{j-m-1}\{(D^m\Phi_j)(\sigma_i)- \\
&\quad \sum_{\nu=1}^{m}c_\nu(\sigma_{in})(D^{\nu-1}\Phi_j)(\sigma_i)(h_n/2)^{m+1-\nu}\}.
\end{aligned}$$

This is another $2k-2m$ system for $i = 1,...,2k-2m$, and this can be written in matrix form as

$$Ay = c \tag{3.19}$$

where

$$c_i = f(\sigma_{in}) - \sum_{j=1}^{m}(D^{j-1}\hat{u})(x_n)\{(D^m\Phi_j)(-\sigma_i)(-h_n/2)^{j-m-1}$$

$$- \sum_{\nu=1}^{m} c_\nu(\sigma_{in})(D^{\nu-1}\Phi_j)(-\sigma_i)(-h_n)^{j-\nu}\}$$

$$- \sum_{j=1}^{m} (D^{j-1}\hat{u})(x_{n+1})\{(D^m\Phi_j)(\sigma_i)(h_n/2)^{j-m-1}$$

$$- \sum_{\nu=1}^{m} c_\nu(\sigma_{in})(D^{\nu-1}\Phi_j)(\sigma_i)(h_n/2)^{j-\nu}\}.$$

for $i = 1, ..., 2k-2m$,

$$y = \{(D^m\bar{u})(x_n^+), \dots, (D^{k-1}\bar{u})(x_n^+), (D^m\bar{u})(x_{n+1}^-), \dots, (D^{k-1}\bar{u})(x_{n+1}^-)\}^T.$$

The matrix A can be written as

$$A = \begin{pmatrix} D^m\Phi_{m+1}(-\sigma_1)(-h_n/2)^0 & \dots & \dots & D^m\Phi_k(\sigma_1)(h_n/2)^{k-m-1} \\ D^m\Phi_{m+1}(-\sigma_2)(-h_n/2)^0 & \dots & \dots & D^m\Phi_k(\sigma_2)(h_n/2)^{k-m-1} \\ \vdots & & & \vdots \\ D^m\Phi_{m+1}(-\sigma_{2k-2m})(-h_n/2)^0 & \dots & \dots & D^m\Phi_k(\sigma_{2k-2m})(h_n/2)^{k-m-1} \end{pmatrix} \tag{3.20}$$

$$- \begin{pmatrix} \sum_{\nu=1}^{m} c_\nu(\sigma_{in})(D^m\Phi_j)(-\sigma_1)(-h_n/2)^{m+1-\nu} & \dots \\ \vdots & \vdots \\ \sum_{\nu=1}^{m} c_\nu(\sigma_{in})(D^m\Phi_j)(-\sigma_{2k-2m})(-h_n/2)^{m+1-\nu} & \dots \end{pmatrix} \tag{3.21}$$

which we choose to write as

$$A = BD + \Delta A \tag{3.22}$$

where B equals

$$B = \begin{pmatrix} D^m\Phi_{m+1}(-\sigma_1) & \cdots & D^m\Phi_k(-\sigma_1) & D^m\Phi_{m+1}(\sigma_1) & \cdots & D^m\Phi_k(\sigma_1) \\ D^m\Phi_{m+1}(-\sigma_2) & \cdots & D^m\Phi_k(-\sigma_2) & D^m\Phi_{m+2}(\sigma_2) & \cdots & D^m\Phi_k(\sigma_2) \\ \vdots & & & & & \vdots \\ \vdots & & & & & \vdots \\ \vdots & & & & & \vdots \\ D^m\Phi_{m+1}(-\sigma_{2k-2m}) & \cdots & D^m\Phi_k(-\sigma_{2k-2m}) & D^m\Phi_{m+1}(\sigma_{2k-2m}) & \cdots & D^m\Phi_k(\sigma_{2k-2m}) \end{pmatrix} \tag{3.23}$$

which is the same as H in equation (3.16). The diagonal matrix D is

$$D = \begin{pmatrix} (-h_n/2)^0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & \cdots & \cdots & \vdots \\ \vdots & 0 & (-h_n/2)^{k-m-1} & 0 & \cdots & \vdots \\ \vdots & \cdots & 0 & (h_n/2)^0 & 0 & \vdots \\ \vdots & \cdots & \cdots & 0 & \cdots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & (h_n/2)^{k-m-1} \end{pmatrix}. \tag{3.24}$$

From hypothesis (4), the associated Hermite-Birkhoff problem is well-posed, i.e., H is invertible so that $(HD)^{-1}$ exists and $||A - HD|| = ||\Delta A||$. Finally $||A - HD||_\infty = ||\Delta A||_\infty =$

$$max_{(i)} \sum_{\nu=1}^{m} |c_\nu(\sigma_{in})| * |(D^m\Phi_j)(-\sigma_i)| * |(-h_n/2)^{m+1-\nu}| + \ldots$$

$$\ldots + \sum_{\nu=1}^{m} |c_\nu(\sigma_{in})| * |(D^m\Phi_j)(\sigma_i)| * |(h_n/2)^{k-\nu}|$$

$$\leq (2k-2m) \sum_{\nu=1}^{m} |c_p(\sigma_q)| * |(D^m\Phi_j)(-\sigma_r)| * |(-h_n/2)^{m+1-\nu}|$$

$$\leq (2k-2m)m|c_p(\sigma_q)| * |(D^m\Phi_j)(-\sigma_r)| * |(-h_n/2)|$$

where $|c_p(\sigma_q)|$ is maximum in $|c_\nu(\sigma_{in})|$ for $\nu = 1, 2, \ldots, m$ and for $i = 1, 2, \ldots, 2k-2m$, and $|(D^m\Phi_j)(-\sigma_r)|$ is maximum in $|(D^m\Phi_j)(-\sigma_i)|$ for $i = 1, 2, \ldots, 2k-2m$. Clearly, $||\Delta A|| \leq C * h$ for some constant C independent of the mesh; consequently for h sufficiently small A is invertible, i.e., Q_Δ exists and is bounded. □

Theorem 2 *If $u \in C^{2k}[a,b]$ and $\bar{u}$ is as in Theorem 1, then there exists a positive constant C such that for each $\Delta \in \Pi$,*

$$||D^j u - D^j \bar{u}|| \leq Ch^{2k-j+1} \textit{for } j = 1, 2, \ldots, 2k \tag{3.25}$$

where Π is the set of all partitions of $[a,b]$ with $\Delta = \{a = x_1 < \ldots < x_{N+1} = b\}$.

Proof : If Q is the given projector then for each Δ, on (x_n, x_{n+1}) let $F(t) = u(x(t))$, $\hat{F}(t) = \bar{u}(x(t))$, $x(t) = (h_n/2)t + (x_n + x_{n+1})/2$. Then $F \in C^{2k}[-1,1]$ and $\hat{F} = QF$. So

$$\begin{aligned}\sup_{(x_n,x_{n+1})} |D^{j-1}u(x) - D^{j-1}\bar{u}(x)| &= (2/h_n)^{j-1} \sup_{(-1,1)} |D^{j-1}(F - \bar{F})| \\ &= (2/h_n)^{j-1} \sup_{(-1,1)} |D^{j-1}[(1-Q)F]| \qquad (3.26) \\ &\text{for } j = 1, \ldots, 2k.\end{aligned}$$

Q induces on $C[-1,1]$ a map Q_j defined as follows

$$Q_j : C[-1,1] \to P_{2k-j+1} : F \to D^{j-1}(Q(D^{-j+1}F)).$$

So Q_j is a continuous linear projector from $C[-1,1]$ onto P_{2k-j+1}. After using

Lebesque's inequality

$$||(1 - Q_j)D^{j-1}F|| \leq ||1 - Q||\ dist(D^{j-1}F, P_{2k-j+1}) \qquad (3.27)$$

for $j = 1, \ldots, 2k$.

From Lorentz (1966) for $F \in C^{2k-j+1}[-1, 1]$,

$$dist(D^{j-1}F, P_{2k-j+1}) \leq \frac{||D^{2k}F||}{2^{2k-j}(2k-j+2)!}. \qquad (3.28)$$

Applying this to the inequality (3.27)

$$||(1 - Q_j)D^{j-1}F|| \leq ||1 - Q_j|| \frac{||D^{2k}F||}{2^{2k-j+1}(2k-j+2)!}. \qquad (3.29)$$

Substituting the above equation into equation (3.26) yields

$$\sup_{(x_n, x_{n+1})} |D^{j-1}u(x) - D^{j-1}\bar{u}(x)| \leq C_j(2/h_n)^{2k-j+1} \sup_{(x_n, x_{n+1})} |D^{2k}u(x)|, \qquad (3.30)$$

where $C_j = \frac{||1-Q_j||}{2^{2(2k-j+1)}(2k-j+2)!}$.

Therefore

$$||D^{j-1}u - D^{j-1}\bar{u}|| \leq Ch^{2k-j+1} \text{for } j = 1, 2, \ldots, 2k \qquad (3.31)$$

for some positive constant C. □

Chapter 4

EXAMPLES

In this chapter the uniformly superconvergent interpolation using local collocation is tested to verify the expected behavior. Many examples were tested to verify the algorithm but only four examples are presented. The maximum error of the uniformly superconvergent interpolants is compared with that of the collocation solution. At points not in the mesh, the collocation solution has an error of order $m+k$, while the uniformly superconvergent approximation is expected to have $2k$ order of error. So we could predict that the uniformly superconvergent approximation is as accurate as the collocation solution at the mesh points. But the accuracy depends on the way of choosing the secondary collocation points. Three different choices for the secondary collocation points were explored for each example. The first choice is open uniform :

$$\sigma_i = (k-m+1-i)/(2k-2m+1)$$

where the $2k-2m$ points are interior to $(-1,1)$ and equidistant. The second one is closed uniform :

$$\sigma_i = (2k-2m-i)/(2k-2m-1)$$

where 2 points are chosen at the ends and then the other $(2k-2m-2)$ points are interior and equidistant. The last choice is Gauss points which is the set of roots of Legendre polynomial of degree $2k-2m$. The choice of Gauss points seems a natural one for collocation because of its importance in generating the superconvergent $\bar{u}$. But when the secondary collocation points were chosen to be Gauss points, this made the coefficient matrix singular in chapter 2. Therefore the Gauss points are expected to be poor secondary collocation points. Also, for the case $k = 2m$ there is an easy argument which shows the problem. In this case, $2k - 2m = k$, so the set of $2k - 2m$ secondary collocation points for $\bar{u}$ coincide with the set of k Gauss collocation points for the superconvergent interpolant $\hat{u}$. Since they both are in $P_{2k,\Delta} \cap C^{m-1}[a, b]$, if $\bar{u}$ is uniquely determined, i.e., its associated coefficient matrix is invertible, then $\bar{u} = \hat{u}$. Consequently $\bar{u}$ would degenerate to a $(m + k)$th order piecewise polynomial, and it can not have an error bound like (3.25). The open set should not have an ill-posed underlying Hermite-Birkhoff problem, so its coefficient matrices are expected to be better behaved. Similarly, the closed set should not make the coefficient matrices singular. However in the closed case the interpolant uses the differential equation at mesh points which correspond to $\sigma_i = -1$ or 1. This choice can be poor on singularly perturbed problems because the errors can be magnified by the reciprocal of the perturbation parameter. Hence, we anticipate that, at least on singularly perturbed problems, the closed choice will be inferior to the open one.

The first example is a second order ordinary differential equation;

$$D^2u(x) = D^1u(x) + xu(x) + (-x^3 - 2x^2 + 2x + 1)e^{4x} \tag{4.1}$$

$$u(0) = 0 \qquad u(1) = 0.$$

The exact answer is

$$u(x) = x(x-1)\exp(4x).$$

Figure 4.1 shows the plot of the exact answer for example 1. Figure 4.2 shows the plot of the derivative of the exact answer for example 1.

Table 1 and 2 show the three types of errors. The first is at the mesh points. The

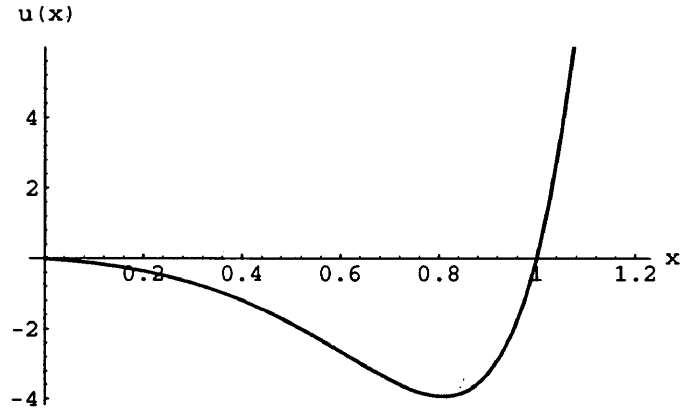

FIG. 4.1. Plot of $u(x)$ for Example 1.

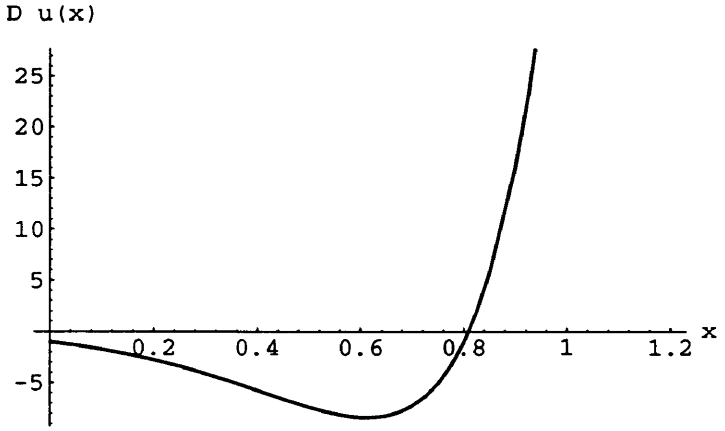

FIG. 4.2. Plot of $Du(x)$ for Example 1.

second is the error in the collocation solution $\hat{u}$ between the mesh points and the third is for the uniformly superconvergent interpolants based on three choice of secondary collocation points. Table 1 is the case of $k = 3$ with various N values (number of mesh intervals) using equally spaced mesh intervals. The error in the open set is very close to the order of the error at the mesh points. The closed set is less accurate than the open one, but still it is much superior to the collocation solution. The decrease in the derivative accuracy for all cases is reflecting the expected error bounds (3.25). With N getting bigger, we get more accurate values. Note that the choice of Gauss points produced a large error as expected from chapter 2-3.

Table 4.1. Absolute errors for Example 1.
$k = 3$ *with various* N

		At the mesh points	$\hat{u}(x)$	$\bar{u}(x)$		
				open	closed	Gauss
$N = 4$	u	0.995E-4	0.542E-2	0.273E-3	0.205E-2	0.539E+1
	Du	0.678E-3	0.137E+0	0.908E-2	0.280E-1	0.679E+2
$N = 8$	u	0.163E-5	0.223E-3	0.523E-5	0.435E-4	0.499E+1
	Du	0.112E-4	0.118E-1	0.388E-3	0.117E-2	0.126E+3
$N = 16$	u	0.259E-7	0.801E-5	0.879E-7	0.798E-6	0.479E+1
	Du	0.177E-6	0.873E-3	0.141E-4	0.425E-4	0.241E+3
$N = 32$	u	0.408E-9	0.263E-6	0.141E-8	0.135E-7	0.479E+1
	Du	0.277E-8	0.592E-4	0.477E-6	0.144E-5	0.484E+3
$N = 64$	u	0.639E-11	0.869E-8	0.223E-10	0.220E-9	0.479E+1
	Du	0.434E-10	0.386E-5	0.155E-7	0.467E-7	0.966E+3

Because we used equally spaced mesh points we can easily estimate the rate of convergence. This estimate comes from equation (3.25).

$$\frac{|\text{error for } h|}{|\text{error for } h/2|} \approx \frac{Ch^{2k-j+1}}{C(h/2)^{2k-j+1}} \tag{4.2}$$

$$= 2^{2k-j+1} \quad \text{for } j = 1, 2, \ldots, 2k.$$

We used $(2k-1)$st degree polynomial as interpolants for $u(j=1)$, so 2^{2k} factors are expected. For Du we used $2k-2$ degree polynomial $(j=2)$, so 2^{2k-1} factors are expected. Table 2 shows these rates of convergent for the open case with $k=3$. The factors are very close to 64 for u and 32 for Du as we expected.

Table 4.2. Rate of convergence.
open case, $k = 3$, *various* N

N		4 to 8	8 to 16	16 to 32	32 to 64
factor	u	52.2	59.5	62.3	63.2
	Du	23.4	27.5	29.6	30.8

Table 3 shows the case for $k=4$ and various N. It is easily seen that the uniformly superconvergent approximation is superior to the collocation approximation. The error of the open case is very close to the mesh point error as in Table 1; the open case is still better than the closed case. The error at the Gauss points is notable. It is almost equal to the errors in $\bar{u}$. Here $k=4$ and $m=2$, so $k=2m$, which is the case mentioned earlier in this chapter where $\hat{u}(x)=\bar{u}(x)$. For this special case the uniform superconvergence does not occur, so the Gauss points as secondary collocation points result in an error of order $m+k$ at most. The error in Gauss points degenerates as N gets bigger than 32 because of the ill-conditioning in the coefficient matrix. Therefore Gauss points are not good as secondary collocation points. After $N=32$ the accuracy of all cases does not improve because the mesh interval is too fine. The error is dominated by the double precision roundoff. So the case $N=64$ is not good for the estimation of rate of convergence as seen in Table 4.

Table 4 shows the rate of convergence for $k=4$. Hence 2^8(=256) is expected as

Table 4.3. Absolute errors for Example 1.
$k = 4$ with various N

		At the mesh points	$\hat{u}(x)$	$\bar{u}(x)$		
				open	closed	Gauss
$N = 4$	u	0.196E-6	0.267E-3	0.653E-6	0.614E-5	0.267E-3
	Du	0.156E-5	0.984E-2	0.179E-4	0.122E-3	0.984E-2
$N = 8$	u	0.808E-9	0.576E-5	0.311E-8	0.322E-7	0.578E-5
	Du	0.643E-8	0.416E-3	0.189E-6	0.127E-5	0.418E-3
$N = 16$	u	0.323E-11	0.106E-6	0.130E-10	0.146E-9	0.483E-6
	Du	0.255E-10	0.151E-4	0.171E-8	0.115E-7	0.684E-4
$N = 32$	u	0.111E-13	0.179E-8	0.535E-13	0.615E-12	0.716E-5
	Du	0.995E-13	0.510E-6	0.144E-10	0.966E-10	0.203E-2
$N = 64$	u	0.754E-14	0.291E-10	0.799E-14	0.888E-14	0.116E-3
	Du	0.213E-13	0.166E-7	0.462E-12	0.107E-11	0.657E-1

a convergence factor for u and 2^7(=128) for Du. The results in Table 4 are close to what we expected.

Table 4.4. Rate of convergence.
open case, $k = 4$, various N

N		4 to 8	8 to 16	16 to 32	32 to 64
factor	u	210	239	243	66
	Du	94.7	111	119	31.2

The second example is a second order differential equation with perturbation parameter ϵ.

$$D^2u(x) = u(x)/\epsilon - x/\epsilon \tag{4.3}$$

$$u(1) = 2 \qquad u(-1) = 2.$$

The exact answer is

$$u(x) = c_1 e^{x/\sqrt{\epsilon}} + c_2 e^{-x/\sqrt{\epsilon}} + x \tag{4.4}$$

where

$$c_1 = \frac{e^{1/\sqrt{\epsilon}} - 3e^{-1/\sqrt{\epsilon}}}{e^{2/\sqrt{\epsilon}} - e^{-2/\sqrt{\epsilon}}}$$

$$c_2 = \frac{3e^{1/\sqrt{\epsilon}} - e^{-1/\sqrt{\epsilon}}}{e^{2/\sqrt{\epsilon}} - e^{-2/\sqrt{\epsilon}}}.$$

Figure 4.3 shows the plot of the exact answer for example 2. Figure 4.4 shows the plot of the derivative of exact answer for example 2. The mesh intervals are chosen as equally spaced at first, then remeshed so that an approximation to $||u^{(2k)}||^{1/(2k)}$ was roughly equidistributed. For a given mesh the *maximum local mesh ratio* is defined by

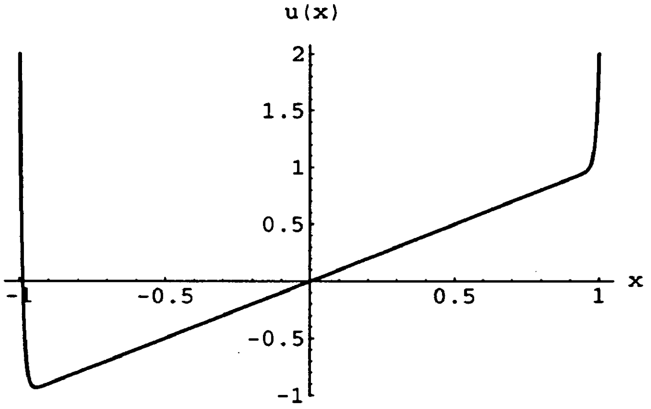

FIG. 4.3. Plot of $u(x)$ for Example 2.

$$\max_i\{\max[h_i/h_{i-1}, h_{i-1}/h_i]\};$$

the *global mesh ratio* is

$$\max h_i / \min h_i.$$

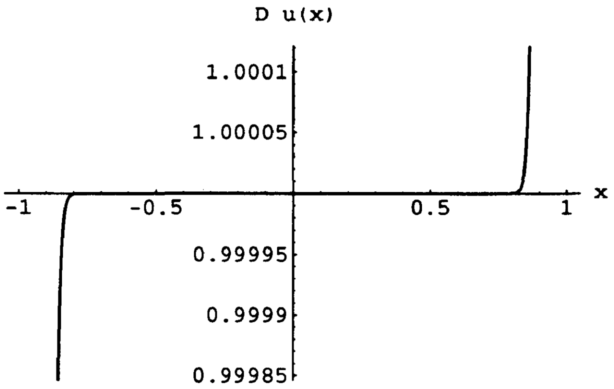

FIG. 4.4. Plot of $Du(x)$ for Example 2.

Table 5 shows the local and global mesh ratios from the final remeshed mesh points, for $k = 4$ and $\epsilon = 10^{-4}$ with various N. Since the mesh ratio varies from 3 to 100, the rate of convergence is meaningless.

Table 4.5. Local and global mesh ratio of Example 2.
$k = 4$, $\epsilon = 10^{-4}$, *various* N

	maximum local mesh ratio	global mesh ratio
$n = 16$	11.0	75.0
$n = 32$	4.2	48.7
$n = 48$	3.2	100.7
$n = 64$	3.6	100.2

Table 6 shows the absolute errors for the meshes in Table 5. The error in the closed case is very close to the mesh point error. This problem is singularly perturbed at both ends. The choice of secondary collocation points for the closed case includes both these ends. The meshes are more dense near the ends. Hence the result from this closed case is very accurate. For the open case the error is more accurate than the closed case. As shown in Table 6 the errors in $\bar{u}$ and $D\bar{u}$ for the open case are exactly the same as those at mesh points. In other words the accuracy for this open case did not degrade at all. The errors for Gauss points are very close to those of $\hat{u}(x)$. This is what we expect because this case is one where $k = 2m$.

Table 4.6. Absolute errors for Example 2.
$k = 4$, $\epsilon = 10^{-4}$, *various* N

		At the mesh points	$\hat{u}(x)$		$\bar{u}(x)$	
				open	closed	Gauss
$N = 16$	u	0.513E-6	0.817E-5	0.513E-6	0.281E-5	0.817E-5
	Du	0.475E-4	0.694E-2	0.475E-4	0.103E-3	0.694E-2
$N = 32$	u	0.112E-7	0.850E-6	0.850E-7	0.412E-7	0.850E-6
	Du	0.108E-5	0.107E-2	0.107E-5	0.427E-5	0.107E-2
$N = 48$	u	0.115E-8	0.699E-8	0.115E-8	0.220E-8	0.714E-8
	Du	0.101E-6	0.198E-4	0.101E-6	0.253E-6	0.202E-4
$N = 64$	u	0.213E-9	0.142E-8	0.213E-9	0.351E-9	0.172E-8
	Du	0.179E-7	0.460E-5	0.179E-7	0.470E-7	0.657E-5

Table 7 shows the results when we approximate example 2 with a higher degree polynomial than in Table 6. The closed case is as accurate as the case at mesh points. The accuracy for the open case did not deteriorate at all. But the Gauss points as secondary collocation points show huge errors. The decline in the derivative accuracy was roughly 10^{-2} for the uniformly superconvergent approximation compared to the function error. That is because of the factor $1/\sqrt{\epsilon}$.

Table 8 is the case for for $k = 4$ and $\epsilon = 10^{-5}$ with various N. As in Table 6 and Table 7 the uniformly superconvergent interpolant is better than standard collocation. The open case is still better than the closed case, but the accuracy deteriorated a little bit. That is because this case is more steeply perturbed than the one in Table 6.

Table 4.7. Absolute errors for Example 2.
$k = 6$, $\epsilon = 10^{-4}$, *various* N

		At the mesh points	$\hat{u}(x)$		$\bar{u}(x)$	
				open	closed	Gauss
$N = 16$	u	0.194E-9	0.261E-6	0.194E-9	0.205E-9	0.912E+2
	Du	0.289E-7	0.195E-3	0.289E-7	0.289E-7	0.179E+5
$N = 32$	u	0.910E-12	0.183E-8	0.910E-12	0.910E-12	0.338E+2
	Du	0.901E-10	0.264E-5	0.901E-10	0.923E-10	0.680E+4
$N = 48$	u	0.911E-13	0.193E-8	0.911E-13	0.911E-13	0.699E+2
	Du	0.130E-10	0.276E-5	0.130E-10	0.130E-10	0.133E+5
$N = 64$	u	0.203E-13	0.232E-9	0.203E-13	0.203E-13	0.591E+2
	Du	0.199E-11	0.438E-6	0.199E-11	0.223E-11	0.112E+5

Table 4.8. Absolute errors for Example 2.
$k = 4$, $\epsilon = 10^{-5}$, *various* N

		At the mesh points	$\hat{u}(x)$		$\bar{u}(x)$	
				open	closed	Gauss
$N = 16$	u	0.386E-4	0.251E-2	0.489E-4	0.304E-3	0.251E-2
	Du	0.209E+0	0.242E+1	0.209E+0	0.257E+0	0.242E+1
$N = 32$	u	0.253E-5	0.381E-3	0.298E-5	0.219E-4	0.381E-3
	Du	0.142E-1	0.523E+0	0.142E-1	0.229E-1	0.523E+0
$N = 48$	u	0.120E-7	0.496E-5	0.120E-7	0.569E-7	0.496E-5
	Du	0.309E-4	0.145E-1	0.309E-4	0.109E-3	0.145E-1
$N = 64$	u	0.118E-8	0.937E-6	0.119E-8	0.610E-8	0.937E-6
	Du	0.306E-5	0.365E-2	0.306E-5	0.149E-4	0.365E-2

The next example is a second order ordinary differential equation with perturbation parameter ϵ which can cause a spike at $x = 0$.

$$D^2u(x) = -xDu(x)/\epsilon - (\pi x sin^2 \pi x)/\epsilon - \pi^2 cos \pi x \tag{4.5}$$

$$u(-1) = -2 \qquad u(1) = 0.$$

The exact answer is

$$u(x) = \cos \pi x + \mathrm{erf}(x/\sqrt{2\epsilon})/\mathrm{erf}(1/\sqrt{2\epsilon}).$$

Figure 4.5 shows the plot of the exact answer for example 3. Figure 4.6 shows the plot of the derivative of exact answer for example 3. This problem is singularly perturbed near $x = 0$, so nonuniform meshes were used. The rate of convergence was meaningless as example 2. As in example 2, the mesh interval was chosen so that an approximation to $||u^{(2k)}||^{1/2k}$ was roughly equidistributed. Table 9 gives the absolute

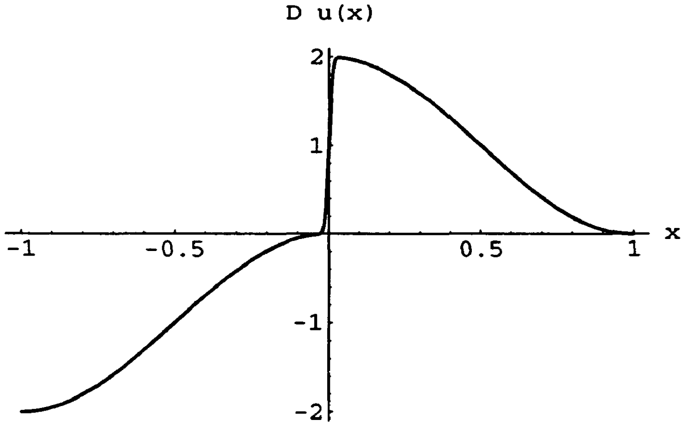

FIG. 4.5. Plot of $u(x)$ for Example 3.

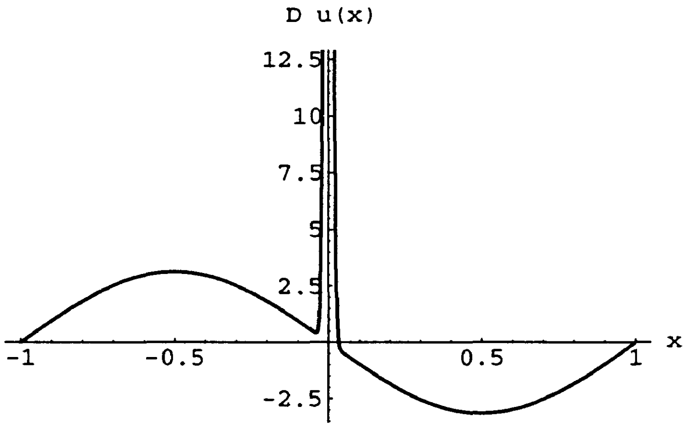

FIG. 4.6. Plot of $Du(x)$ for Example 3.

errors for $k = 4$ and $\epsilon = 10^{-2}$ with various N. The open case was better than the closed case. Even the closed case does not degrade at all in the order. Both cases were superior to the collocation solution. The errors for Gauss points were very close to those for standard collocation as expected.

Table 4.9. Absolute errors for Example 3.
$k = 4$, $\epsilon = 10^{-2}$, *various* N

		At the mesh points	$\hat{u}(x)$	$\bar{u}(x)$		
				open	closed	Gauss
$N = 36$	u	0.178E-8	0.182E-7	0.188E-8	0.497E-8	0.182E-7
	Du	0.165E-6	0.443E-5	0.165E-6	0.201E-6	0.442E-5
$N = 48$	u	0.246E-9	0.506E-8	0.326E-9	0.629E-9	0.506E-8
	Du	0.270E-7	0.139E-5	0.270E-7	0.301E-7	0.141E-5
$N = 64$	u	0.439E-10	0.105E-8	0.607E-10	0.997E-10	0.103E-8
	Du	0.527E-8	0.438E-6	0.527E-8	0.527E-8	0.452E-6

Table 10 shows the same case as Table 9 except $\epsilon = 10^{-4}$. This value of ϵ made a sharper spike near the origin. It made the closed case worse than standard collocation, because the reciprocal of ϵ made the error big. But still the closed case was not bad; furthermore it was better than the standard one when we approximated Du with 64 intervals. The open case is still accurate to the same order of the error of the mesh points. All results worsen compared to Table 9 because of the smaller parameter ϵ.

Table 4.10. Absolute errors for Example 3.
$k = 4$, $\epsilon = 10^{-4}$, *various* N

		At the mesh points	$\hat{u}(x)$		$\bar{u}(x)$	
				open	closed	Gauss
$N = 36$	u	0.140E-7	0.110E-6	0.327E-7	0.872E-5	0.110E-6
	Du	0.786E-5	0.213E-3	0.786E-5	0.276E-3	0.213E-3
$N = 48$	u	0.366E-8	0.680E-7	0.406E-8	0.339E-6	0.680E-7
	Du	0.158E-5	0.128E-3	0.158E-5	0.146E-4	0.128E-3
$N = 64$	u	0.767E-9	0.183E-7	0.846E-9	0.276E-7	0.183E-7
	Du	0.410E-6	0.433E-4	0.410E-6	0.185E-5	0.433E-4

Table 11 shows the case of $\epsilon = 10^{-6}$ with $k = 4$ as above. The errors were severely magnified by the reciprocal perturbation parameter. For any number of the mesh points shown here, the closed case was the worst one. Hence for the closed secondary collocation points, interpolants using the differential equation at mesh points were frequently poor on singularly perturbed problems. The accuracy improved by increasing the number of mesh points.

The final example is an oscillatory function.

$$D^2u(x) = -28\pi u(x)/(1 + 7x)^2 \tag{4.6}$$

Table 4.11. Absolute errors for Example 3.
$k = 4$, $\epsilon = 10^{-6}$, *various N*

		At the mesh points	$\hat{u}(x)$	$\bar{u}(x)$		
				open	closed	Gauss
$N = 36$	u	0.976E-2	0.165E-1	0.231E-1	0.207E+3	0.165E-1
	Du	0.469E+2	0.469E+2	0.469E+2	0.687E+4	0.446E+2
$N = 48$	u	0.151E-4	0.161E-3	0.193E-4	0.185E+0	0.161E-3
	Du	0.612E-1	0.915E+0	0.612E-1	0.145E+2	0.915E+0
$N = 64$	u	0.740E-9	0.147E-7	0.521E-8	0.183E-3	0.319E-7
	Du	0.475E-5	0.219E-3	0.475E-5	0.110E-1	0.220E-3

$$u(0) = 1 \qquad u(1) = 0.$$

The exact answer is

$$(1 + 7x)\cos[4\pi/(1 + 7x)].$$

Figure 4.7 shows the plot of the exact answer this example. Figure 4.8 shows the plot of the Derivative of exact answer for example 1. Table 12 shows the absolute errors for different degree piecewise polynomial approximation with $N = 24$. Here we used a variable mesh chosen the same way as the other examples. The closed case for $k = 4$ is inferior to the open case, but it is more accurate than standard collocation. When $k = 6$ both cases did not deteriorate at all for u and they are much superior to $\hat{u}$. The errors at Gauss points were big for $k = 6$, and equal to the errors of $\hat{u}$ for $k = 4$, as expected from chapter 2.

Table 13 shows the absolute errors for a different number of mesh points with $k = 6$. For the open and the closed case the accuracy did not degrade at all in the order of error. They were much superior to the collocation approximation.

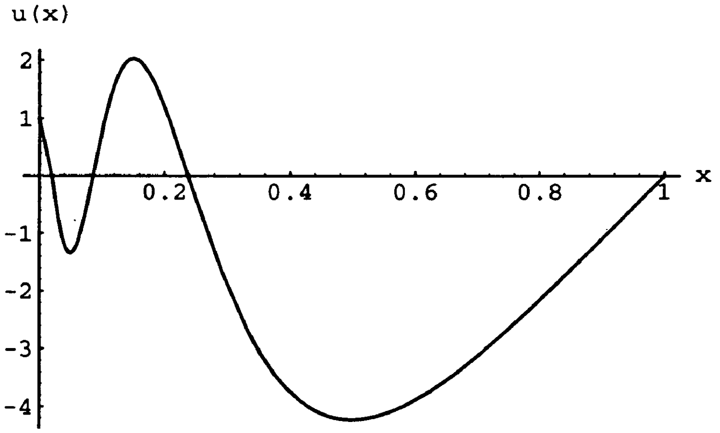

FIG. 4.7. Plot of $u(x)$ for Example 4.

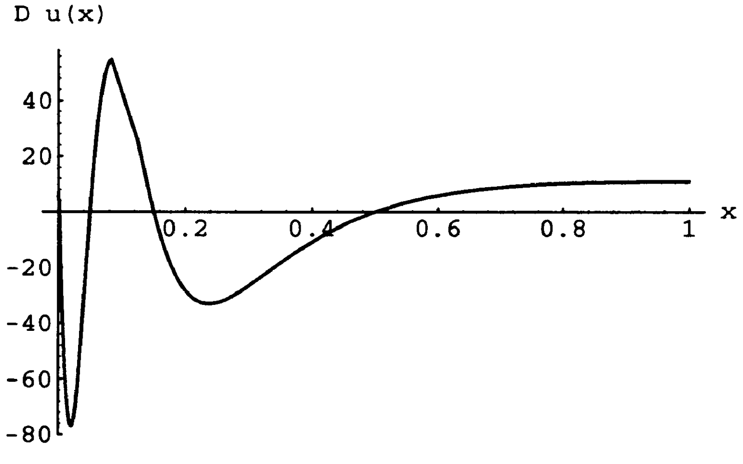

FIG. 4.8. Plot of $Du(x)$ for Example 4.

Table 4.12. Absolute errors for Example 4 with n=24.

		At the mesh points	$\hat{u}(x)$	$\bar{u}(x)$ open	closed	Gauss
$k = 4$	u	0.950E-8	0.240E-6	0.105E-7	0.227E-7	0.240E-6
	Du	0.419E-6	0.243E-3	0.719E-6	0.451E-5	0.243E-3
$k = 6$	u	0.799E-13	0.811E-9	0.799E-13	0.799E-13	0.847E+1
	Du	0.455E-11	0.715E-6	0.455E-11	0.482E-11	0.822E+3

Table 4.13. Absolute errors for Example 4 when k=6.

		At the mesh points	$\hat{u}(x)$	$\bar{u}(x)$ open	closed	Gauss
$N = 16$	u	0.630E-11	0.679E-8	0.644E-11	0.643E-11	0.819E+1
	Du	0.140E-9	0.521E-5	0.222E-9	0.201E-9	0.498E+3
$N = 24$	u	0.799E-13	0.811E-9	0.799E-13	0.799E-13	0.847E+1
	Du	0.455E-11	0.715E-6	0.455E-11	0.482E-11	0.822E+3

Chapter 5

CONCLUSION

Two point boundary value problems were solved using local collocation with piecewise polynomials. The superconvergent error bound at mesh points was $O(h^{2k})$. The error at nonmesh points could have the same order as at the mesh points, which means that uniform superconvergence can occur. In order to achieve this superconvergent accuracy of approximation at nonmesh points, three sets of secondary collocation were studied. In general Gauss points performed too badly to be useful, but when $k = 2m$, Gauss points turned out to be equivalent to standard collocation. The closed uniform set was better in approximation than standard collocation for many cases. But as in example 3, the closed case was worse than standard collocation for a singularly perturbed differential equation. The open uniform mesh points preserve the accuracy of superconvergence very well through all the examples. As shown in Table 4.2 and Table 4.4, the rate of convergence had an order $2k$ for u and $(2k-1)$ for Du, consistent with superconvergence. As seen in other examples, this method worked very nicely even on very irregular meshes and on singularly perturbed problems.

The analysis of the determinant of the Hermite-Birkhoff coefficient matrix can be generalized in future work. Improper row and column scaling sometimes made the condition number so big that it was difficult to analyze the condition of local collocation points. An alternative algorithm which can handle scaling problems and different representation is being studied. This algorithm needs to be rewritten for first order systems to compare with the one for variable order systems.

REFERENCES

[1] U. Ascher, J. Christiansen, and R. Russell. A collocation solver for mixed order systems of boundary value problems, Math. Comp., 33 (1979), 659–679.

[2] U. Ascher, S. Pruess, and R. Russell, On spline basis selection for solving differential equations, SIAM J. Numer. Anal., 20 (1983), 121–132.

[3] C. de Boor and B. Swartz, Collocation at Gaussian Points, SIAM J. Numer. Anal., 10 (1973), 582–606.

[4] R. D. Russell and L. F. Shampine, A collocation method for boundary value problems, Numer. Math., 19 (1972), 1–28.

[5] C. de Boor, A Practical Guide to Splines, Springer-Verlag, Madison, New York, 1978.

[6] E. Houstis, A collocation method for systems of nonlinear ordinary differential equation, J. Math. Anal. Appl., 62 (1978), 24–37.

[7] S. Pruess, Estimating the eigenvalues of Sturm-Liouville Problems by approxi -mating the coefficients, SIAM J. Numer. Anal., 10 (1973), 55–68

[8] S. Pruess, Interpolation schemes for collocation solutions of two point boundary value problems, SIAM J. Sci. and Stat. Comp., 7 (1986), 322–333.

[9] S. Pruess, Solving Linear Boundary Value Problems by Approximating the Coefficients, Math. Comp., 27 (1973), 551–561.

[10] G. Lorentz, Approximation of Functions, Holt, Rinehart and Winston, New York, 1966

[11] S. Pruess, Stability bounds for local Lagrangian interpolation, J. Approx. Theory, 53 (1988), 117–127.

[12] S. Wolfram, Mathematica, Addison–Wesley, Redwood City, California, 1991.

[13] S. Pruess and H. Jin A Stable High Order Interpolation Scheme for Superconvergent Data, Colorado School of Mines, Dept. of Math. and CS., technical report, MCS-93-11, to be submitted.