Rational Approximations to the Exponential Function for the Numerical Solution of the Heat Conduction Problem

by

Stewart Roy Trickett

Department of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
N2L 3G1

Research Report CS-84-09

April 1984

Abstract

A common approach for the numerical solution of the heat conduction problem is to reduce it to a set of ordinary differential equations through discretization of the space variables. Unfortunately the resulting system displays certain properties, particularly sparseness and stiffness, which make it unsuited to solution by many standard numerical ODE solvers. The methods examined here are an alternate formulation of the semi-implicit Runga-Kutta schemes, whereby the problem is further reduced to that of estimating an exponential function involving matrix arguments. Stability and implementation considerations lead us to examine rational approximations of the form

$$e^{-z} \approx (c_0 + c_1 z + \cdots + c_m z^m) / (1 + bz)^n, \; m \leq n,$$

where the value of b is chosen first, and the numerator coefficients are then calculated to give maximum order at zero.

The primary aim of this thesis is to determine what values of the parameter b most benefit the solution of the heat conduction problem. We find that by making the method exact for some critical eigenvalue of the complementary problem, performance during the transient phase can be greatly enhanced. The ability of the approximations to satisfy such a criterion is established by two existence theorems. Results concerning A_0-stability and the attenuation of high frequency components are also given. Finally, a physical problem involving heat conduction in a thermal print head is used to more fully demonstrate the behaviour of these methods.

Table of Contents

Chapter 1
Introduction

1.1 The Semi-Discrete Problem

We will investigate numerical solutions for the following class of parabolic initial boundary value problems, defined inside the finite region Ω, with boundary $\partial\Omega$ and boundary normal $\eta(x)$:

$$\begin{aligned}
&\frac{\partial u(x,t)}{\partial t} = \nabla\cdot(\alpha(x)\nabla u(x,t)) - \beta(x)u(x,t) + s(x,t) \\
&u(x,0) = f(x),\ \ x \in \Omega \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (1.1.1) \\
&\gamma u(x,t) + \delta u_\eta(x,t) = g(x,t),\ \ x \in \partial\Omega,\ \ t > 0.
\end{aligned}$$

It is assumed that $\alpha(x)$ is positive and suitably smooth, and that $\beta(x)$ is non-negative and at least piece-wise continuous inside the region Ω. The problem can be generalized to include mild non-linearities in u, time dependent coefficients α and β, and so on, and the results stated here will often apply. The physical problems which can be represented by (1.1.1) are numerous, and include heat conduction [2], fluid flow in a porous medium [28], and mass transfer [1].

The numerical approach to this problem will be that of semi-discretization, or “the method of lines“. We will illustrate this for Dirichlet boundary conditions and Ω a finite interval $[a,b]$ in $\mathbb{R}^1$. The extension to higher dimension and other boundary conditions is straightforward, and can be found in a number of sources (for example, Varga [39], chapter 6).

The region $[a,b]$ is first discretized into N subintervals $[x_i, x_{i+1}]$, $i=0,\ldots,N-1$, where $x_0=a$, $x_N=b$, and $x_i<x_{i+1}$. A common strategy is to select a uniform mesh where all intervals have an identical length h,

and we will normally assume this has been done. Define

$$\mathbf{u}(t) = (u_1(t), \ldots, u_{N-1}(t))^T,$$

where $u_i(t)$ is to be considered an approximation to $u(x_i,t)$. The system can now be written in the “semi-discrete“ form

$$\frac{d\mathbf{u}(t)}{dt} = -\mathbf{B}^{-1}\mathbf{A}\mathbf{u}(t) + \mathbf{B}^{-1}\mathbf{s}(t); \tag{1.1.2}$$

$$\mathbf{u}(0) = (f(x_1), \ldots, f(x_{N-1}))^T.$$

The length $N-1$ vector $\mathbf{s}(t)$ arises from the source term $s(x,t)$ and any inhomogeneous boundary conditions. The $N-1$ by $N-1$ matrices **A** and **B** result from the discrete approximation of the operator L given by

$$Lu = -\nabla\cdot(\alpha(x)\nabla u) + \beta(x)u,$$

with the appropriate boundary conditions.

For example, a standard method for estimating the operator L is through the central difference approximation

$$\delta_x y(x) = \frac{y(x+h/2) - y(x-h/2)}{h} = \nabla y(x) + O(h^2),$$

giving

$$Lu_i \approx \frac{-1}{h^2}\left[\alpha(x_i+h/2)(u_{i+1}-u_i) - \alpha(x_i-h/2)(u_i-u_{i-1})\right] + \beta(x_i)u_i.$$

In this case **B** is simply the identity matrix. When $\alpha(x)=1$ and $\beta(x)=0$, the matrix **A** takes the well known form:

$$\frac{1}{h^2}\begin{bmatrix} 2 & -1 & & & & 0 \\ -1 & 2 & -1 & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & -1 & 2 & -1 \\ 0 & & & & -1 & 2 \end{bmatrix}$$

Central difference is one of the most popular approximations, and we will adopt it for our “model“ operator. A second major class of methods are the finite element approximations, particularly Galerkin and collocation (see Mitchell and Wait [25]). The important point is that these approximations can be expected to have the following properties in general:

1. The matrix **A** is positive semi-definite for Neumann boundary conditions, and positive definite for Dirichlet boundary conditions. The matrix **B** is positive definite.
2. **A** and **B** will often be symmetric (or can be made symmetric) and thus have real non-negative eigenvalues.
3. The matrices will be sparse and highly structured. In the one-dimensional case, for instance, **A** is typically tri- or pentadiagonal. In the two-dimensional case, **A** is banded and block tri- or pentadiagonal.

It will often happen that **B** is the identity matrix. Since the essential characteristics of the problem are preserved, we will assume that this is case from now on. In section 1.4 it will be shown how a more complicated operator can be introduced during implementation.

We have reduced (1.1.1) to a system of ordinary differential equations (ODE's) which, due to the properties of **A** and **B**, is well posed. It is now entirely feasible to submit (1.1.2) to one of the many robust and well-developed numerical ODE solvers. This is inadvisable for at least two reasons:

1. The semi-discrete system often displays a property called stiffness, for which many standard packages will be inefficient.
2. A competitive scheme should take full advantage of the special properties of the problem. In particular, the sparseness of the matrices must be exploited.

A class of numerical methods which are designed specifically for the efficient solution of the semi-discretized heat equation is the subject of this thesis.

1.2 The Method of Rational Approximations

Numerical methods for the solution of (1.1.2) reflect those for general ODE's (for a general survey, see Seward, Fairweather, and Johnston [33]). Our approach, which is in fact equivalent to the class of semi-implicit Runga-Kutta schemes, will be as follows:

Consider the homogeneous problem

$$\frac{d\mathbf{u}(t)}{dt} = -\mathbf{A}\mathbf{u}(t); \quad \mathbf{u}(0) \text{ given.} \tag{1.2.1}$$

The exact solution is

$$\mathbf{u}(t) = \exp(-t\mathbf{A})\ \mathbf{u}(0),$$

where $\exp(\mathbf{C})$ represents the convergent series $\mathbf{I} + \mathbf{C} + \mathbf{C}^2/2! + \cdots$. Since time marching techniques are used, this will be written in the incremental form

$$\mathbf{u}(t+\tau) = \exp(-\tau\mathbf{A})\ \mathbf{u}(t).$$

The symbol τ will be used throughout to represent the time step.

An approximation to the exponential is now required. For reasons that we will discuss later on, rational polynomials will be used, so that

$$\exp(-\tau\mathbf{A}) \approx R(\tau\mathbf{A}) = q(\tau\mathbf{A})^{-1}p(\tau\mathbf{A}),$$

where p and q are polynomials. In practice, of course, matrix inverses are never actually calculated. The numerical method can now be written

$$q(\tau\mathbf{A})\mathbf{v}_{k+1} = p(\tau\mathbf{A})\mathbf{v}_k, \tag{1.2.2}$$

where $\mathbf{v}_k$ is the calculated solution after the k'th time step.

In view of the properties of $\mathbf{A}$, we may consider instead the approximation to $\exp(-z)$

$$R(z) = q(z)^{-1}p(z)$$

where z is a complex scalar with non-negative real part. Some useful properties for $R(z)$ will now be given.

Definition 1.2.1

A rational approximation $R(z)$ to $\exp(-z)$ is said to be

1. Of order r if $\exp(-z) - R(z) = 0(z^{r+1})$ as $z \to 0$.
2. A-acceptable if $|R(z)| < 1$ whenever $Re[z] > 0$.
3. A_α-acceptable if $|R(z)| < 1$ whenever $|arg(z)| < \alpha$.
4. A_0-acceptable if $|R(z)| < 1$ whenever z is real and positive.
5. L-acceptable if $R(z)$ is A-acceptable and $R(z) \to 0$ as $|z| \to \infty$.

The problem of stiffness often arises in connection with the solution of (1.2.1). A detailed description of this phenomenon will not be attempted here; however, a good introduction is given by Shampine and Gear [34], and Finlayson [8] discusses stiff systems arising from partial differential equations in particular.

Stiffness is caused by eigenvalues in the Jacobian (in this case, the matrix $-\mathbf{A}$) which have real parts that are large and negative relative to an appropriate time scale for the problem. In these cases the time increment for many numerical schemes must often be kept unduly small because of stability, rather than accuracy, considerations. As an example, the largest negative eigenvalue of the matrix $-\mathbf{A}$ resulting from the central difference approximation, assuming a uniform mesh, is proportional to $1/h^2$. The time scale is determined here by the smallest eigenvalue, which is roughly constant for any mesh selection, and so the system becomes stiffer as the spatial mesh is refined. For instance, to ensure the stability of the Forward Euler scheme it is required that

$$\tau \leq \frac{h^2}{2}.$$

The method can be very inefficient when fine spatial meshes are taken.

This leads naturally to the consideration of methods which are stable regardless of step size; that is, the A-stable methods. Dahlquist defines A-stability in terms of the scalar equation

$$y' = -\lambda y, \quad y(0) = y_0, \quad Re(\lambda) > 0,$$

resulting in the numerical scheme

$$v_{k+1} = R(\tau\lambda)v_k, \quad v_0 = y(0). \tag{1.2.3}$$

Definition 1.2.2

Scheme (1.2.3) is

1. A-stable (A_α-stable) (A_0-stable) if $v_k \to 0$ as $k \to \infty$ for any step size τ whenever $Re(\lambda) > 0$ ($|arg(\lambda)| < \alpha$) (λ is real and positive).
2. L-stable if it is A-stable and $R(\tau\lambda) \to 0$ as $|\tau\lambda| \to \infty$, with $Re(\lambda) > 0$.

The definitions can be extended to the system (1.2.2) if **A** is non-defective (has a full set of eigenvectors), which we will assume. In this case λ is considered to represent the eigenvalues of the matrix **A**. Since for the current problem these will normally be real, we will primarily be concerned with A_0-stability.

The relationship between stability and acceptability is as follows:

Scheme (1.2.3) is A-/A_α-/A_0-/L-stable if R is A-/A_α-/A_0-/L-acceptable.

The advantage of rational over polynomial approximations is now apparent; A-stability is possible only when the degree of the denominator is at least that of the numerator. The price, of course, is that we must now solve systems of equations at each time step. This can be partially ameliorated by the careful selection of the rational approximation.

Lack of L-stability (assuming the method is A-stable) is normally not a problem unless $|R(\tau\lambda)| \to 1$ as $|\tau\lambda| \to \infty$, in which case the larger eigenvalues can become (very nearly) parasitic. A classic example is the Crank-Nicolson scheme, although it occurs for all (n,n) Padé approximants. Various measures have been suggested for overcoming this difficulty. Morris and Lawson [19], for instance, recommend that time steps for Crank-Nicolson be no larger than lh/π, where l is the spatial interval length. A major goal when developing numerical methods is to avoid such stability-related restrictions.

1.3 Approximations to the Exponential

We now address the problem of choosing an approximation to the exponential. One of the best studied to date is the (m,n) Padé approximant, defined as the rational

$$R(z) = (\sum_{i=0}^{n} d_i z^i)^{-1} (\sum_{i=0}^{m} c_i z^i),$$

where the coefficients are determined solely to maximize the order at zero. The resulting method is of order $m+n$. The numerator and denominator are unique up to a common scaling factor, which we resolve by assuming $d_0 = 1$.

The usual manner of displaying these methods is through the Padé table:

m n	0	1	2
0	1	$1-z$	$1-z+z^2/2$
1	$\frac{1}{1+z}$	$\frac{1-z/2}{1+z/2}$	$\frac{1-2z/3+z^2/6}{1+z/3}$
2	$\frac{1}{1+z+z^2/3}$	$\frac{1-z/3}{1+2z/3+z^2/6}$	$\frac{1-z/2+z^2/12}{1+z/2+z^2/12}$

Table 1.3.1: Padé approximants to $\exp(-z)$ for $0 \leq m,n \leq 2$.

The entries for the first row $(n=0)$ are simply the truncated Maclaurin series of $\exp(-z)$. Some well known Padé approximants are:

(1,0) Padé (Forward Euler)

(0,1) Padé (Backward Euler)

(1,1) Padé (Trapezoidal or Crank-Nicolson)

The stability properties are as follows:

Theorem 1.1

The (m,n) Padé approximant to the exponential is

(1) A_0-acceptable iff $m \leq n$.

(2) A-acceptable iff $n-2 \leq m \leq n$.

(3) L-acceptable iff $n-2 \leq m \leq n-1$.

Proof

(1) Varga [38], (2)-(3) Ehle [5] and Wanner, et al [40].

In direct contrast to Padé are the Chebyshev approximations, where the coefficients of $R(z)$ are determined to minimize the uniform error on the positive real axis:

$$\| \exp(-z) - R(z) \|_\infty = \sup_{z>0} |\exp(-z) - R(z)|$$

These were first studied by Cody, Meinardus, and Varga [4], and the principle application is for methods which give reasonable results using a single large time step. Otherwise they would seem to be of little use, since the Chebyshev property is not preserved under compounding (Lawson and Swayne [22]).

A disadvantage of the pure Chebyshev methods is that they have no order at zero (although they can be manipulated to give a 0-order method). As we shall see in the following section, order can be valuable for the integration of forcing terms. In response, Lawson and Lau [18] proposed the order constrained approximations, which determine the rational $R(z)$ with minimum uniform error that satisfies given order criteria at zero. To date, little work has been done on these (but see Lawson and Swayne [21]).

The major drawback with all of the above methods lies in the implementation. From (1.2.2), we are faced with the solution of the linear system

$$q(\tau\mathbf{A})\mathbf{v} = \mathbf{c}$$

The matrix $\mathbf{A}$ is generally very sparse; however when $q(\tau\mathbf{A})$ is formed explicitly severe fill-in can occur, to the point where solving the system becomes infeasible.

In addition, the system can become very poorly conditioned. If, on the other hand, q has only real zeroes, we can perform the factorization

$$q(\tau\mathbf{A}) = (\mathbf{I}+b_1\tau\mathbf{A}) \cdots (\mathbf{I}+b_n\tau\mathbf{A})$$

and only the solution of systems similar in structure to **A** is required.

Unfortunately, the approximations discussed so far lead invariably to denominators with complex roots whenever $n > 1$. Two approaches have been considered here. The first is to factor q fully, and then carry out the solution in complex arithmetic. A second approach is to factor q into linear and quadratic real factors. Normally, squaring **A** does not produce too much fill-in, and complex arithmetic is avoided. The two methods are roughly competitive for simple examples (see Swayne [36]), although complex factorization is perhaps preferred, since it leads to better conditioned systems and avoids the more pathological examples of matrix fill-in.

These problems can be avoided altogether by selecting approximations which have only real poles. In particular, consider rationals of the form

$$R^m_{n,b}(z) = \frac{c_0 + c_1 z + \cdots + c_m z^m}{(1+bz)^n}, \quad m \leq n. \tag{1.3.1}$$

The parameter b is chosen real and positive, so that the approximation is analytic in the right-half complex plane.

An immediate question is then, why is only a single n'th order pole considered? As it happens, allowing more than one distinct real pole seems to offer no advantages, based on either Chebyshev (Lau [16]) or order (Norsett and Wolfbrandt [32]) criteria. As a bonus, the implementation of a single repeated pole approximation can be expected to be more efficient in both time and memory space; for instance, if Gaussian elimination is used to solve the linear systems, only one LU decomposition is required.

The parameters b and $\{c_i\}$ can be selected using the same criteria as the previous approximations. Norsett [30], for example, considered the case where the order of the approximation at zero is maximized, resulting in an order $m+1$ method. This class includes the Crank-Nicolson and Backward Euler methods. A simple Norsett approximation which is not also Padé is the L21 approximation of Swayne [36]. Lau [16] studied approximations of the type $R^m_{n,b}(z)$ where the parameters are chosen to minimize the uniform error on the positive real axis, both with and without order constraints. Since both the Norsett and Lau approximations will be discussed in the next chapter, no more will be said at this point.

This brings us finally to the approximations which will be studied in chapter 2; that is, where b is selected without constraint, and the numerator coefficients are then chosen to give maximum order at zero. The resulting method is at least order m. The main question posed is this: What values of b are of most benefit in the numerical solution of the heat conduction problem? A number of possibilities come to mind, including order, Chebyshev, and stability criteria, and this provides a good basis for comparison.

1.4 The Integration of Forcing Terms

In this section we discuss how the method of rational approximations can be extended to include forcing terms. The full semi-discrete problem is

$$\frac{\partial \mathbf{u}}{\partial t} = -\mathbf{A}\mathbf{u}(t) + \mathbf{s}(t), \quad \mathbf{u}(0) \text{ given.} \tag{1.4.1}$$

The exact solution is given by (in incremental form)

$$\mathbf{u}(t+\tau) = \exp(-\tau\mathbf{A})\mathbf{u}(t) + \int_t^{t+\tau} \exp[(\theta-t-\tau)\mathbf{A}]\mathbf{s}(\theta)d\theta$$

or with the appropriate change of variable,

$$\mathbf{u}(t+\tau) = \exp(-\tau\mathbf{A})\mathbf{u}(t) + \tau\int_0^1 \exp[(\sigma-1)\tau\mathbf{A}]\mathbf{s}(t+\sigma\tau)d\sigma$$

The method shown here, first proposed by Norsett [31a], and later adopted for rational approximations to the exponential by Lawson [17], allows the efficient evaluation of the integral term using the classic criterion that it be exact whenever $\mathbf{s}(t)$ is a polynomial of specified degree r or less. This is essentially an extension of Hamming's uniform method for scalar operators ([13], chapter 10).

Define the i'th moment $m_i(z)$ as

$$m_i(z) = \int_0^1 \exp[(\sigma-1)z]\sigma^i d\sigma$$

Using integration by parts, we can derive the following recursion relationship

$$m_0 = z^{-1}[1-\exp(-z)]$$

$$m_i(z) = z^{-1}[1-im_{i-1}(z)], \quad i=1,\cdots,r.$$

Replacing the exponential with a rational approximation provides the impetus for the following method:

1. Choose an order r rational approximation to $\exp(-z)$

$$R(z) = q(z)^{-1}p(z) \tag{1.4.2}$$

2. Compute the coefficients of the $r+1$ rational moment functions from the recursion

$$M_0(z) = z^{-1}[1-R(z)] \tag{1.4.3a}$$

$$M_i(z) = z^{-1}[1-iM_{i-1}(z)], \quad i=1, \cdots, r. \tag{1.4.3b}$$

3. Compute the coefficients of the $r+1$ rational weighting functions from

$$W_i(z) = \sum_{j=0}^{r} \nu_{ij} M_j(z), \quad i=0, \cdots, r. \tag{1.4.4}$$

   where ν_{ij} are the elements of the inverse Vandermonde matrix associated with the nodes a_i, $i=0, \cdots, r$.

The method can now be written

$$\mathbf{v}_{k+1} = R(\tau\mathbf{A})\mathbf{v}_k + \tau\sum_{i=0}^{r} W_i(\tau\mathbf{A})\mathbf{s}(t+a_i\tau). \tag{1.4.5}$$

For easily calculated source terms, the nodes $\{a_i\}$ are usually selected to be uniformly spaced between 0 and 1, including the endpoints. When the source terms are difficult to calculate, or involve the values of $\mathbf{u}$, the nodes often correspond to the previous $r+1$ time steps.

The over-riding advantage to this method lies in the following fact:

> If $R(z)$ is an order r approximation to $\exp(-z)$, then the rational functions $R(z)$ and $\{W_i(z)\}$ all have the same denominator.

Except for the right hand side, the systems of equations to be solved are the same as for the homogeneous system (1.2.2), and so the remarks of the previous section still hold.

In implementation, Swayne [37] argues that the natural expression for (1.4.5) is a partial fraction ordering. When R is chosen to be a repeated pole approximation (1.3.1), this takes the form

$$R^m_{n,b}(z) = \alpha_0 + \alpha_1(1+bz)^{-1} + \cdots + \alpha_n(1+bz)^{-n}$$

$$W_i(z) = \omega_{i1}(1+bz)^{-1} + \cdots + \omega_{in}(1+bz)^{-n}, \quad i=0,\cdots,r.$$

Combined with a Horner-type scheme, $\mathbf{v}_{k+1}$ can then be evaluated from the previous time step $\mathbf{v}_k$ as follows:

$$\mathbf{v}^* \leftarrow 0; \tag{1.4.6a}$$

$$\text{for } i = 0\ (1)\ n-1 \text{ do} \tag{1.4.6b}$$

$$\{\mathbf{v}^* \leftarrow (\mathbf{I}+b\tau\mathbf{A})^{-1}\,[\mathbf{v}^* + \alpha_{n-i}\mathbf{v}_k + \tau\sum_{j=0}^{r}\omega_{j,n-i}\mathbf{s}(t+a_j\tau)]\}; \tag{1.4.6c}$$

$$\mathbf{v}_{k+1} \leftarrow \mathbf{v}^* + \alpha_0\mathbf{v}_k; \tag{1.4.6d}$$

The efficiency of the method is obvious. When the forcing terms are simple it is not much more expensive than the homogeneous problem. The source terms at each node can be calculated once at the beginning of the loop and stored, or recalculated each iteration; the proper choice is usually apparent from the application. In any case, at least two vectors ($\mathbf{v}_k$ and $\mathbf{v}^*$) require concurrent storage.

Calculation of the moments, weights, partial fraction coefficients, and so on, is tedious, and so FORTRAN routines have been provided in appendix A. The code handles only the single repeated pole approximation, although most routines would be useful for other rationals.

We can return to the more general problem (1.1.2), where the matrix $\mathbf{B}$ is not the identity matrix, by substituting $\mathbf{B}^{-1}\mathbf{A}$ for $\mathbf{A}$ and $\mathbf{B}^{-1}\mathbf{s}(t)$ for $\mathbf{s}(t)$ in (1.4.6c), giving

$$\{(\mathbf{I}+b\tau\mathbf{B}^{-1}\mathbf{A})\mathbf{v}^* \leftarrow \mathbf{v}^* + \alpha_{n-i}\mathbf{v}_k + \tau\sum_{j=0}^{r}\omega_{j,n-i}\mathbf{B}^{-1}\mathbf{s}(t+a_j\tau)\};$$

Multiplying both sides by $\mathbf{B}$ results in the following replacement for (1.4.6c):

$$\{\mathbf{v}^* \leftarrow (\mathbf{B} + b\tau\mathbf{A})^{-1}\,[\mathbf{B}(\mathbf{v}^* + \alpha_{n-i}\mathbf{v}_k) + \tau\sum_{j=0}^{r}\boldsymbol{\omega}_{j,n-i}\mathbf{s}(t+a_j\tau)]\};$$

An interesting property of this method is the following:

Theorem 1.2

Let $R(z)$ and $W_i(z)$, $i=0,\cdots,r$ be computed as in (1.4.2)-(1.4.4). Then (1.4.5) is exact for the polynomial particular solution of (1.4.1) whenever $\mathbf{s}(t)$ is a polynomial of degree r or less, and $\mathbf{A}$ is a real non-singular square matrix.

Proof

See Lawson [17].

The important point is that high order is a very useful property when complicated forcing terms are present. If a method is both exact for the polynomial particular solution and A-stable, we are guaranteed that the asymptotic solution is correct, regardless of step size.

1.5 The Integration of Transients

The previous section discussed the integration of forcing terms, and concluded that the order of the method played a significant role in the evaluation of the particular solution. This is not necessarily the case for the evaluation of transients, as the following example demonstrates.

Gourlay and Morris [11] investigated a number of stable third and fourth order extrapolation schemes using the following problem for comparison:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 \le x \le 2, \quad t > 0, \tag{1.5.1}$$

$$u(0,t) = u(2,t) = 0, \quad u(0,x) = 1.$$

Time steps of $\tau = .025$ and $\tau = .1$ were taken, and the solution examined at $t = 1.2$. The lowest uniform error achieved with any of the methods was 0.13×10^{-4}. Yet using the same operator **A**, the simple 0-order scheme

$$[\mathbf{I} + b\tau\mathbf{A}]\mathbf{v}_{k+1} = \mathbf{v}_k \tag{1.5.2}$$

with $\tau = .1$ and $b = 1.13454$, achieves an error of 0.76×10^{-6}. Of course, the 0-order scheme would not be useful if complicated forcing terms were present, so this is not a criticism of the extrapolation schemes. The conclusion, however, is that the requirements for the efficient evaluation of the transients are quite different than for that of forcing terms. In particular, high order is of little use if large time steps are to be taken.

The above example may be understood by considering the exact solution to (1.5.1)

$$u(x,t+\tau) = \sum_{i=1}^{\infty} a_i \sin(\sqrt{\mu_i}\, x)\exp(-\mu_i \tau)$$

where

$$\mu_i = \frac{\pi^2 i^2}{4}, \qquad a_i = \int_0^2 u(x,t)\sin(\sqrt{\mu_i}\, x)\,dx, \qquad i = 1,2, \cdots .$$

The eigenvalues of the matrix **A** will be approximations to the eigenvalues μ_i, $i = 1, \cdots, N-1$, and the associated eigenvectors discrete approximations to the functions $\sin(\sqrt{\mu_i}x)$, $i=1,\cdots,N-1$. To dampen each eigenvector at a rate commensurate with the decay of the exact solution we require

$$R(\tau\mu_i) = \exp(-\tau\mu_i), \quad i=1,\cdots,N-1 \quad \text{(approximately)}. \qquad (1.5.3)$$

Note that the behaviour of R at zero, or anywhere but at the above points, is irrelevant. In fact, condition (1.5.3) can be relaxed at all but the first few eigenvalues, since only these are contained in any quantity in the exact solution (except very near time zero). It is enough that the larger eigenvalues be attenuated at a rate greater than the first few. Scheme (1.5.2), for example, is exact only for the value μ_1.

Methods which are exact for the exponential at one or more eigenvalues have been termed *exponentially fitted*. The main result for real eigenvalues is the maximal interpolation theorem of Iserles [14]:

Theorem 1.3

If $p(z)$ and $q(z)$ are polynomials of degree m and n, respectively, then $f(z) = \exp(-z) - p(z)/q(z)$ has at most $m+n+1$ zeroes (counting multiples) on the real axis.

We shall refer to rationals which achieve this upper bound on the non-negative real axis as *generalized Padé approximants*. Methods in this class have been proposed by Liniger and Willoughby [24], Ehle [6], and Ehle and Picel [7].

Unfortunately, the generalized Padé suffer from the same implementation difficulties as the (pure) Padé approximants; that is, of complex factors in the denominator $q(z)$ whenever $n > 1$. It would seem worthwhile, if exponential fitting is to be employed for the current problem, to develop high order methods involving only real poles. One of the few examples to date is a fourth order scheme by Cash [3].

That most schemes do not solve the complementary problem efficiently is reflected in the well known heuristic that one must take small time steps when transients are present. Since this is usually a temporary condition, a case must be made that schemes which calculate transients both efficiently and accurately are worth developing. We make the following points:

1 In a survey, Shampine and Gordon [35] found that about half the problems involving stiff differential equations require the accurate solution of rapid transients, in addition to slowly varying portions.

2. Theorem 1.3 states, in effect, that if the forcing terms are smooth, and the method is of high enough order, the time step is restricted by consideration of the complementary problem alone. Once the transients have died out, any time step whatsoever may be taken (provided the method is stable enough). Under such circumstances, most of the work can be taken up solving the transient portion of the problem.

3. Certain problems have transients continually arising due to the forcing terms being only piece-wise smooth in time (for instance, in periodic heating, or petroleum reservoir simulations).

Nevertheless, exponential fitting will generally be useful for only a small part of the time in most problems. When the steady state solution is approached, it is best to consider other methods. In addition, the *a priori* estimation of eigenvalues is required.

Chapter 2
Selecting the Rational Approximation

2.1 Introduction

In this chapter we study a class of approximations to the exponential function as they apply to the complementary, or transient, solution of the heat equation. Our one dimensional model problem is as follows:

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2}, \ 0 \le x \le l, \ t \ge 0, \\ u(x,0) &= f(x), \ 0 < x < l, \qquad (2.1.1) \\ u(0,t) &= u(l,t) = 0, \ t \ge 0. \end{aligned}$$

Following chapter 1, the system is discretized into N equally spaced intervals, and the central difference operator used to approximate the second derivative. The solution of the resulting ODE is (in incremental form)

$$\mathbf{u}(t+\tau) = \exp(-\tau\mathbf{A})\ \mathbf{u}(t) \qquad (2.1.2)$$

where $\mathbf{A}$ is symmetric and positive definite. An approximation to the exponential is now required. For the most part we will restrict ourselves to rationals of the form

$$R_{n,b}^{m}(z) = \frac{c_0 + \cdots + c_m z^m}{(1+bz)^n}, \quad m \le n. \qquad (2.1.3)$$

The numerator coefficients are determined so as to maximize the order at zero; that is

$$\exp(-z) - R_{n,b}^{m}(z) = O(z^{m+1}). \qquad (2.1.4)$$

It is understood that we always choose $n \ge 1$. As well, the parameter b is chosen positive so that the rational has an n'th order pole at $z=-1/b$ but is otherwise analytic.

Some facts will now be presented which will be of use later on. If $(1+bz)^n = d_0 + ... + d_n z^n$ then

$$d_i = \begin{pmatrix} n \\ i \end{pmatrix} b^i, \quad i=0, \cdots, n.$$

To derive an expression for the numerator coefficients we multiply (2.1.4) by $(1+bz)^n$ to get

$$(1+bz)^n \left(\sum_{i=0}^{\infty} \frac{(-z)^i}{i!}\right) - (c_0 + \cdots + c_m z^m) = O(z^{m+1}).$$

Equating the coefficients of z^i to zero for $i \le m$ gives

$$c_i = \sum_{k=0}^{min(i,n)} \begin{pmatrix} n \\ k \end{pmatrix} b^k \frac{(-1)^{i-k}}{(i-k)!}, \quad i=0, \cdots, m.$$

Also note that

$$\exp(-z) - R_{n,b}^{m+1}(z) = \exp(-z) - R_{n,b}^{m}(z) - \frac{c_{m+1} z^{m+1}}{(1+bz)^n} = O(z^{m+2})$$

and so the approximation error has the form

$$\exp(-z) - R_{n,b}^{m}(z) = \frac{c_{m+1} z^{m+1}}{(1+bz)^n} + O(z^{m+2}) = c_{m+1} z^{m+1} + O(z^{m+2}) \quad (2.1.5)$$

Laguerre polynomials will play a major role in deriving many of the results, and a number of their properties are listed in chapter 3. For now, however, we will need only the following:

Two Properties Of The Laguerre Polynomials

1. When λ is a non-negative integer, the generalized Laguerre polynomial of degree n is

$$L_n^\lambda(y) = \sum_{k=0}^{n} \binom{n+\lambda}{k+\lambda} \frac{(-y)^k}{k!}, \quad n \geq 0.$$

2. $L_n^\lambda(y)$ has n distinct simple positive roots.

See, for instance, Lebedev [23].

The following theorem is due to Norsett [29]. We have generalized somewhat and used different notation.

Theorem 2.1

$$c_i = b^i L_i^{n-i}(1/b), \quad i \leq n,$$

$$c_i = (-1)^{i-n} b^n L_n^{i-n}(1/b) \,/\, \prod_{j=n+1}^{i} (j), \quad i > n,$$

$$d_i = b^i L_i^{n-i}(0), \quad i \leq n.$$

The remainder of this chapter will be concerned with finding the approximation $R_{n,b}^m(z)$ which is "best" during the transient phase of the heat conduction problem. The goal is a method which gives accurate results even for moderately large time steps.

The most interesting problem lies in selecting the parameter b. The following three sections each describe a possible criterion for doing so. The first two have previously been studied. The third criterion, which will be shown to be best, is new for this class of approximations. The remaining sections will develop this criterion more fully, as well as introducing an important variation.

2.2 Selecting b to give Maximal Order

We have seen that $R_{n,b}^{m}(z)$ is an approximation to the exponential function whose error is at least $O(z^{m+1})$. For given m and n, however, there exists certain values of b such that

$$exp(-z) - R_{n,b}^{m}(z) = O(z^{m+2}). \tag{2.2.1}$$

These approximations were first studied by Norsett in [30] and [31], hence we will refer to them as being of *Norsett type*.

From (2.1.5), the error for general $R_{n,b}^{m}(z)$ has the form $c_{m+1}z^{m+1} + O(z^{m+2})$. For an approximation to be of Norsett type, b must be chosen so that $c_{m+1}=0$. From theorem 2.1, then, $1/b$ is the zero of:

$$\begin{cases} L_{m+1}^{n-m-1}(y) & \text{if } m<n, \\ L_{n}^{m+1-n}(y) & \textit{otherwise.} \end{cases} \tag{2.2.2}$$

By the second property of the Laguerre polynomials, the number of such points is $\min\{m+1,n\}$. These values of b, which we shall call *Norsett points*, have been tabulated below for $n \leq 6$ and $0 \leq m \leq n$.

It will be of interest later on to note that if $R_{n,b}^{m}(z)$ is of Norsett type, then it is also of the class $R_{n,b}^{m+1}(z)$, and so the error has the form $c_{m+2}z^{m+2} + O(z^{m+3})$.

m n	0	1	2	3	4	5	6
1	1.00000	0.50000					
2	0.50000	0.29289 1.70711	0.21133 0.78867				
3	0.33333	0.21133 0.78867	0.15898 0.43587 2.40515	0.12889 0.30253 1.06858			
4	0.25000	0.16667 0.50000	0.12889 0.30253 1.06858	0.10644 0.22043 0.57282 3.10032	0.09129 0.17448 0.38886 1.34537		
5	0.20000	0.13820 0.36180	0.10904 0.23193 0.65903	0.09129 0.17448 0.38886 1.34537	0.07911 0.14113 0.27805 0.70751 3.79420	0.07013 0.11906 0.21688 0.47327 1.62066	
6	0.16667	0.11835 0.28165	0.09485 0.18813 0.46702	0.08027 0.14487 0.29304 0.81515	0.07013 0.11906 0.21688 0.47327 1.62066	0.06257 0.10165 0.17316 0.33414 0.84109 4.48739	0.05667 0.08901 0.14453 0.25795 0.55670 1.89513

Table 2.2.1: Norsett points for $0 \leq m \leq n$, $1 \leq n \leq 6$.

2.3 Selecting b to Minimize the Uniform Error

Selecting b to minimize the norm

$$\sup_{z>0} \left| \exp(-z) - R^m_{n,b}(z) \right| \tag{2.3.1}$$

was studied in a more general form by Lau [16]. The uniform error as a function of b shows a number of local minima, which we shall refer to as *Lau points*. It was conjectured that the number of local minima was $m+1$ when $m<n$ and n otherwise. We note this is the same as for approximations of Norsett type. The author's own numerical tests indicate that the Lau and Norsett points separate each other, with the smallest being a Norsett point.

The following plot, with $m=2$ and $n=4$, is typical. Line N is the coefficient $|c_3|$ as it varies with b ($c_3=0$ occurs at a Norsett point). Line L is the uniform error (2.3.1) of the approximation as it varies with b (a local minimum occurs at a Lau point).

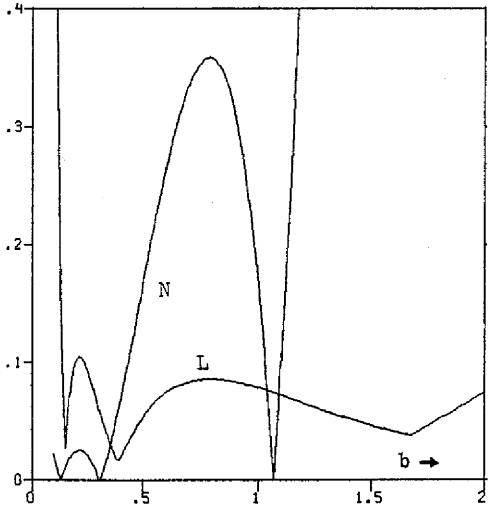

Figure 2.3.1: Norsett and Lau points for $R^2_{4,b}(z)$.

2.4 An Exactness Criterion for b

The third criterion will be developed for general rational approximations (denoted R) to the exponential. The criterion equates the primary eigenvalue of $R(\tau\mathbf{A})$ with the attenuation of the primary Fourier component of the true solution. It has been used by Lawson, Morris, and Wilkes [20] to select an optimum time step τ for the Crank-Nicolson scheme. For the current schemes we will show this can be done through the proper selection of the parameter b.

Consider the model problem (2.1.1). The exact solution is given by (in incremental form)

$$u(x,t+\tau) = \sum_{i=1}^{\infty} a_i \sin(\frac{i\pi x}{l}) \exp(\frac{-i^2\pi^2\tau}{l^2})$$

where

$$a_i = \int_0^l u(x,t) \sin(\frac{i\pi x}{l})\, dx.$$

The damping of the most persistent component $a_1 \sin(\frac{\pi x}{l})$ after a single time step is

$$\exp(\frac{-\pi^2\tau}{l^2})$$

Consider also the exact solution (2.1.2) of the discretized problem. If we replace the exponential with an approximation R, the solution becomes

$$\mathbf{v}_{k+1} = R(\tau\mathbf{A})\,\mathbf{v}_k.$$

For the central difference operator, the eigenvalues of $\tau\mathbf{A}$ are known to be

$$\lambda_i = \frac{4\tau}{h^2}\sin^2(\frac{i\pi h}{2l}), \quad i=1,\ldots,N-1.$$

We shall refer to the smallest eigenvalue λ_1 as λ_D (for discrete). The damping of λ_D for a single time step is

$$R(\lambda_D) = R(\frac{4\tau}{h^2}\sin^2(\frac{\pi h}{2l}))$$

The criterion can now be stated as satisfying

$$\exp(-\lambda_P) = R(\lambda_D) \tag{2.4.1}$$

where (for the model problem)

$$\lambda_P = \frac{\pi^2 \tau}{l^2}, \quad \lambda_D = \frac{4\tau}{h^2}\sin^2(\frac{\pi h}{2l})$$

Essentially, the error due to the spatial discretization cancels the error due to the approximation of the exponential; thus the most persistent components of the continuous and numerical systems are damped at an equal rate. For this reason, criterion 2.4.1 will be referred to as *spatial error compensation,* or *SEC.*

When R is the Crank-Nicolson approximation, Lawson, Morris, and Wilkes show that (2.4.1) is satisfied when

$$\tau \approx \frac{lh}{\pi}.$$

For the current approximations we can list at least two strategies; prescribing the time step τ and then selecting the parameter b to satisfy (2.4.1), or prescribing b and selecting τ appropriately. The former is in keeping with our aim of removing restrictions on the step size. The latter strategy, however, might allow us to satisfy some other criterion which depends on b alone; the Chebyshev criterion of section 2.3, for example. Here we will adopt the first strategy of prescribing the time step.

For a given time step τ, then, the question arises as to whether there exists a value of b for which $R^m_{n,b}(z)$ satisfies (2.4.1). The answer is generally yes, as figure 2.4.1 demonstrates. The function

$$E(b) = \exp(-\lambda_P) - R^2_{4,b}(\lambda_D) \tag{2.4.2}$$

has been plotted as a function of b, and for various values of τ. The interval length l is 1. Condition (2.4.1) is satisfied when $E(b)=0$.

Note for a given time step there may exist a number of values of b satisfying (2.4.1). Also note we have chosen a rather course discretization in space, with $h=.1$. Section 2.6 will show that a slight variation of (2.4.1), particularly

valid for a fine mesh, can always be satisfied.

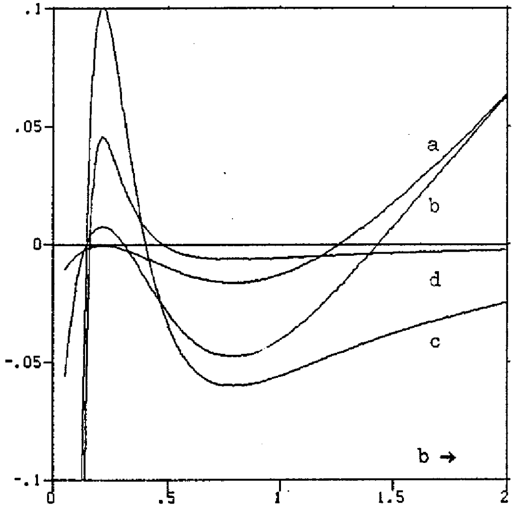

Figure 2.4.1: $E(b)$ **for** $\tau =$ **(a) .05, (b) .1, (c) .5, (d) 1. (**h **= .1,** l **= 1.)**

2.5 Numerical Tests on the Transient Problem

In the previous sections, three possible criteria for selecting the parameter b have been discussed. Here we present some simple numerical tests to determine which might be best for the following problem:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 \leq x \leq 1, \; t \geq 0.$$

$$u(x,0) = f(x), \quad u(0,t) = u(1,t) = 0.$$

The system is discretized into 20 equally spaced intervals ($h = .05$), replacing the second derivative with the central difference operator. The rational $R^2_{4,b}(z)$ is selected as an approximation to the exponential. The system to be solved is then

$$\mathbf{v}_{k+1} = R^2_{4,b}(\tau \mathbf{A})\, \mathbf{v}_k.$$

The scheme is order 2 in space, order 2 in time, and order 3 in time when the approximation is of Norsett type. A numerical examination shows the scheme to be A_0-stable when $b > .1$. For $\tau = .0125, .05, .2, .5$, and 1., the following critical values of b were calculated (all values are approximate):

Norsett Points	Lau Points	SEC τ=.0125	SEC τ=.05	SEC τ=.2	SEC τ=.5	SEC τ=1.
.129	.150	.210[1]	.144	.137	.145	.153
.303	.385	1.16	.294	.345	.400	.447
1.07	1.67		1.23	1.91	5.15	≈34.

Table 2.5.1: Critical values of b for the approximation $R^2_{4,b}(z)$.

[1] This point represents only a local minimum of $|E(b)|$.

Test 1: $f(x) = \sin(\pi x)$

For time steps .025, .05, .2, .5, and 1., the numerical scheme was solved for values of b in the range .1 to 2. The following error function was plotted as a function of b:

$$\log_{10}\left(\frac{||\mathbf{v}(1)-\mathbf{u}(1)||_\infty}{||\mathbf{u}(1)||_\infty}\right) \tag{2.5.1}$$

$\mathbf{u}(1)$ is the exact solution at time 1., calculated spatially at the discrete sample points. $\mathbf{v}(1)$ is the computed numerical solution at time 1.

The results are shown in figures 2.5.1 through 2.5.5. Perhaps not surprisingly, the SEC criterion gave dramatically superior results.

The $\tau = .5$ case has an extra spike (the second one) which has an interesting explanation. At this point we have

$$\exp(-\lambda_P) = -R_{4,b}^2(\lambda_D).$$

After two time steps, however, the damping factors of both the discrete and continuous systems have been squared, and hence are now equivalent. In this instance, then, SEC occurs every second time step. This phenomenon will not be investigated further, but may be useful for approximations where SEC is not possible in the usual sense.

Test 2: $f(x) = \sum_{i=1}^{19}\sin(i\pi x)$

The spectrum of the initial condition is "white"; that is, it contains every Fourier component representable by the discrete system (see Hamming [13], chapter 21). In the exact solution, all but the primary component are quickly damped. A numerical scheme must mimic this behaviour to give accurate results.

Test 1 was repeated with the new initial condition. The error function (2.5.1) has been plotted in figures 2.5.6 through 2.5.8. The two smallest time steps displayed no appreciable differences between tests, and so these results are not shown. The time step $\tau = .2$ was slightly less accurate at the first SEC point. The two largest time steps, however, showed considerable loss of accuracy. It would appear that the larger time steps behave poorly when higher frequencies are present; nevertheless, the exactness criterion of section 2.4 remained optimum. This will be discussed further in section 2.7.

Test 3: $f(x) = \sum_{i=1}^{N-1} \sin(i\pi x), \quad N = 40, 80$

In the third test, the spatial mesh was increased to $N = 40$ and $N = 80$. The initial condition remained white. As discussed in the introduction, the problem becomes progressively stiffer as the mesh is refined.

The results are shown only for $\tau = .0125$ in figure 2.5.9. In this case two additional SEC points appeared, and so the implementation of the criterion benefitted by taking a finer mesh. For other time steps there was no significant change in the error curves from test 2, except for an expected shifting of the critical points. These methods, then, were insensitive to the increased stiffness of the problem.

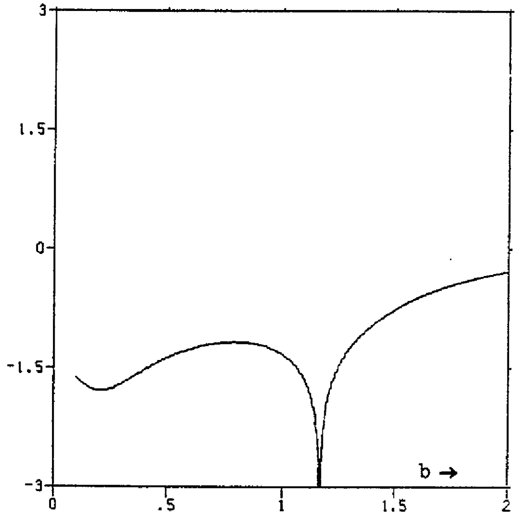

Figure 2.5.1: Test 1, (2.5.1) with $\tau = .0125$.

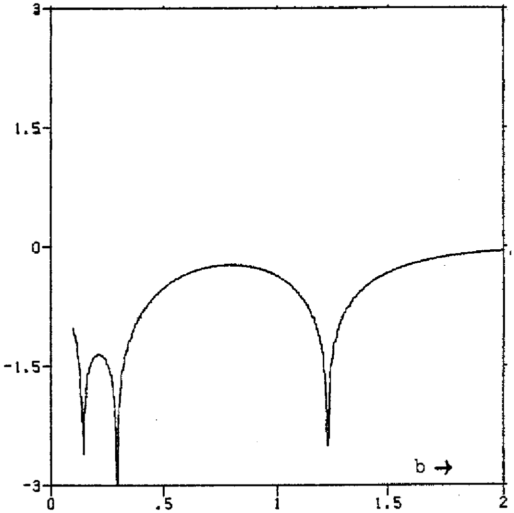

Figure 2.5.2: Test 1, (2.5.1) with τ= .05.

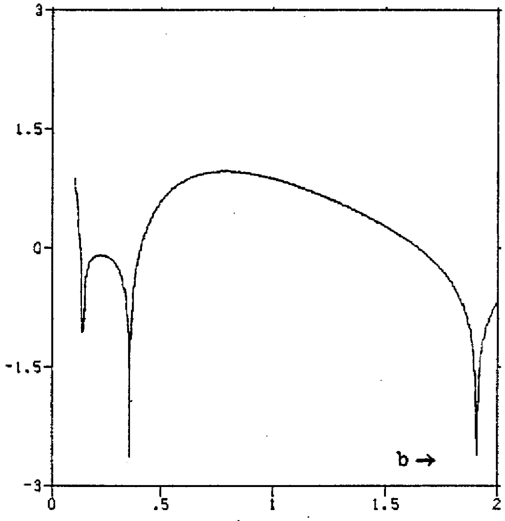

Figure 2.5.3: Test 1, (2.5.1) with τ= .2.

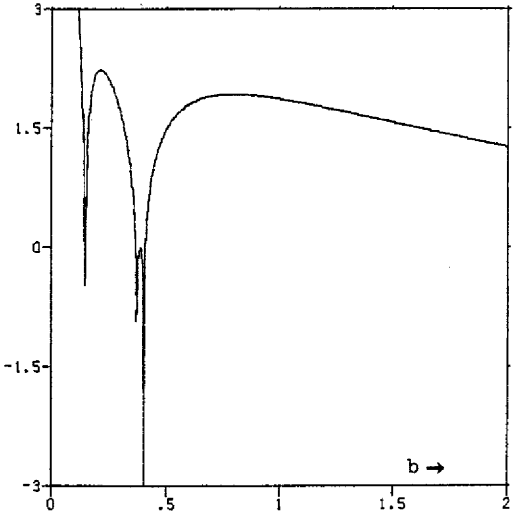

Figure 2.5.4: Test 1, (2.5.1) with $\tau = .5$.

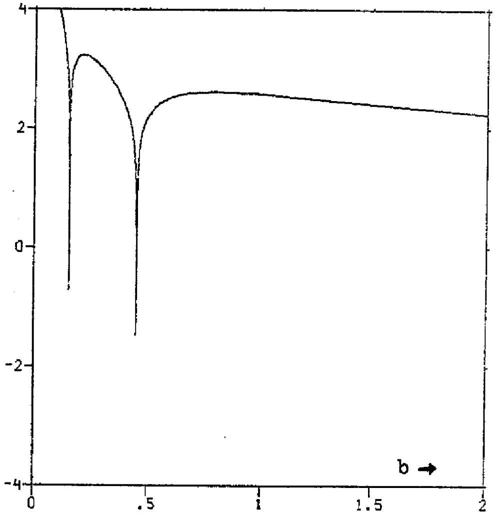

Figure 2.5.5: Test 1, (2.5.1) with $\tau = 1$.

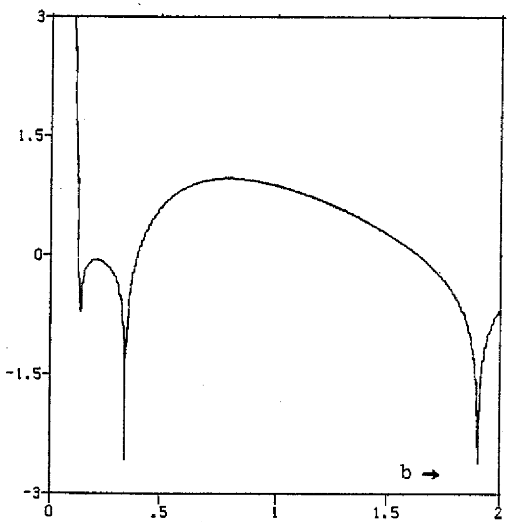

Figure 2.5.6: Test 2, (2.5.1) with $\tau = .2$.

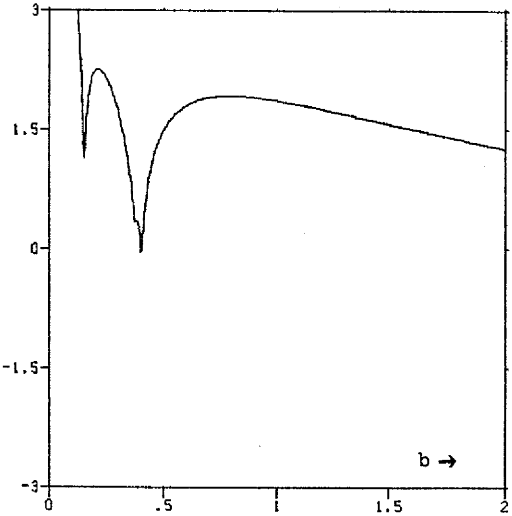

Figure 2.5.7: Test 2, (2.5.1) with $\tau = .5$.

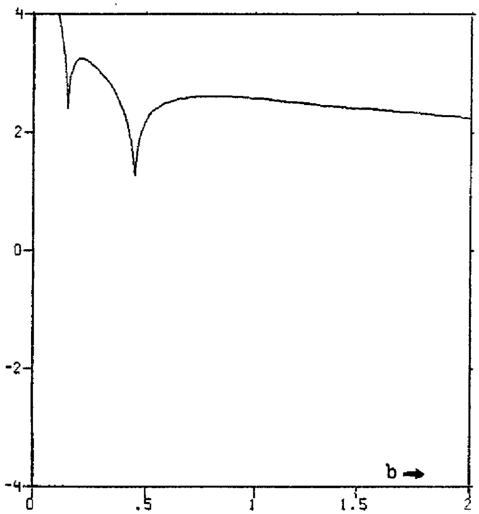

Figure 2.5.8: Test 2, (2.5.1) with $\tau = 1$.

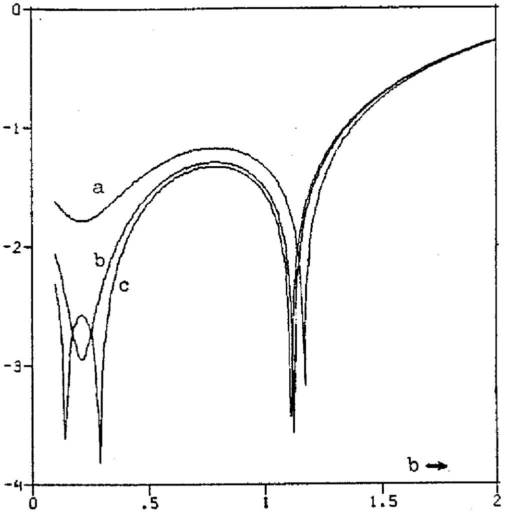

Figure 2.5.9: Test 3, (2.5.1) with $\tau = .0125$, (a) $N = 20$ (b) $N = 40$ (c) $N = 80$.

2.6 Exponential Fitting with b

In section 2.4 we presented the criterion

$$\exp(-\lambda_P) = R(\lambda_D) \tag{2.6.1}$$

where

$$\lambda_P = \frac{\pi^2\tau}{l^2}, \quad \lambda_D = \frac{4\tau}{h^2}\sin^2(\frac{\pi h}{2l})$$

From the Taylor series expansion

$$\left|\frac{\lambda_D - \lambda_P}{\lambda_P}\right| = \frac{(\pi h)^2}{6} + O(h^4),$$

so for small h we have $\lambda_P \approx \lambda_D$. If it is assumed this holds exactly, that is $\lambda_P = \lambda_D$, (2.6.1) can be written

$$\exp(-\lambda_P) = R(\lambda_P), \quad \lambda_P = \frac{\pi^2\tau}{l^2}. \tag{2.6.2}$$

It will be convenient to define the error function

$$E^m_{n,b}(z) = \exp(-z) - R^m_{n,b}(z),$$

so for the current methods, (2.6.2) becomes

$$E^m_{n,b}(\lambda_P) = 0, \quad \lambda_P = \frac{\pi^2\tau}{l^2}. \tag{2.6.3}$$

Hence we have arrived at the criterion of exponential fitting. The suitability of this criterion for the current approximations is shown by the following two theorems. Proofs are given in the next chapter.

Theorem 2.2 ($m < n$ *case*)

Let

$z > 0$, m, and n be given, with $m < n$,

$\delta_1 < \delta_2 < \cdots < \delta_{m+1}$ be the values of b such that $R^{m}_{n,b}(z)$ is of Norsett type,

$\sigma_1 < \sigma_2 < \cdots < \sigma_m$ be the values of b such that $R^{m-1}_{n,b}(z)$ is of Norsett type.

Then

1. $\delta_1 < \sigma_1 < \delta_2 < \cdots < \sigma_m < \delta_{m+1}$,
2. for a given i, $1 \le i \le m$, there exists exactly one value of $b \in (\delta_i, \sigma_i)$ such that $E^{m}_{n,b}(z) = 0$,
3. there exists exactly one $b > \delta_{m+1}$ such that $E^{m}_{n,b}(z) = 0$,
4. for all other positive values of b, $E^{m}_{n,b}(z) \neq 0$.

The following result has been shown in Norsett [29]. We will give an alternate proof in section 3.3.

Theorem 2.3 ($m = n$ *case*)

Let

$z > 0$ and n be given,

$\delta_1 < \delta_2 < \cdots < \delta_m$ be the values of b such that $R^{n}_{n,b}(z)$ is of Norsett type,

$\sigma_1 < \sigma_2 < \cdots < \sigma_n$ be the values of b such that $R^{n-1}_{n,b}(z)$ is of Norsett type.

Then

1. $\delta_1 < \sigma_1 < \delta_2 < \cdots < \delta_n < \sigma_n$,
2. for a given i, $1 \le i \le n$, there exists exactly one value of $b \in (\delta_i, \sigma_i)$ such that $E^{n}_{n,b}(z) = 0$,
3. for all other positive values of b, $E^{n}_{n,b}(z) \neq 0$.

Definition 2.6.1

Given m and n, the ***i'th interpolation interval*** will be defined as the interval $[\delta_i,\sigma_i]$ from theorem 2.2 or 2.3 (as appropriate). When $m < n$, the *m+1'st,* or ***infinite interpolation interval*** is defined as the interval $[\delta_{m+1},\infty)$ from theorem 2.2.

Figure 2.6.1 shows typical behaviour for the approximation error. The function $E^2_{4,b}(z)$ has been plotted for $0 \leq z \leq 15$, and for five values of b on the second interpolation interval $[\delta_2,\sigma_2]$. The curves (a) and (e) correspond to the Norsett points $b = \delta_2$ and $b = \sigma_2$, respectively. Curve (c) corresponds to a Lau point (note the "equi-ripple" property).

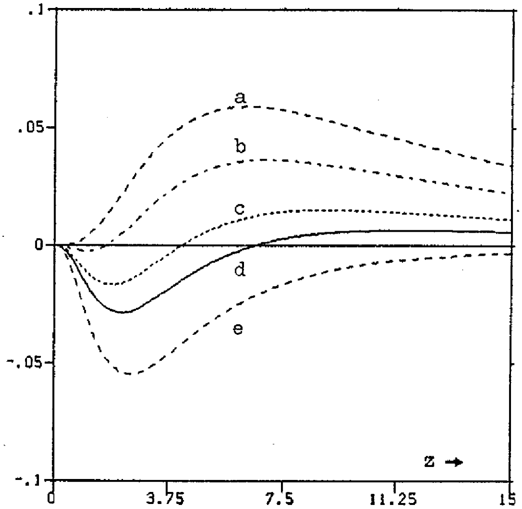

Figure 2.6.1: $E^2_{4,b}(z)$ for b = (a) .30253 (b) .34 (c) .388 (d) .42 (e) .5.

Two Questions

1. When can the exponential fitting criterion (2.6.3) be used in place of criterion (2.6.1)? That is, when can the spatial discretization error be ignored?
2. How does one calculate the value of b in some given interpolation interval

$[\delta_i, \sigma_i]$ such that $R^m_{n,b}(z)$ is exact for given λ_P?

The first question is important since, while the eigenvalue λ_P of the exact solution may be well known, it will often happen that λ_D cannot be determined without considerable calculation. Unfortunately, the answer depends on the approximation to $-\nabla^2$, the time step, the values of m and n, the interpolation interval for b, and so on. However, it is generally to be expected that if the spatial mesh is "reasonably fine", criterion (2.6.3) can be used with confidence.

Table 2.6.1 below shows values for the function

$$\left| \, 1 - \left(\frac{R^2_{4,b}(\lambda_D)}{\exp(-\lambda_P)} \right)^{\frac{1}{\tau}} \right| \tag{2.6.4}$$

for various values of h and τ. The b values were chosen from the second interpolation interval to satisfy (2.6.3) in each case. The function indicates the relative error at $t = 1$ introduced by using exponential fitting over SEC in the solution of the model problem.

$h \diagdown \tau$	.5	.05	.005
.1	.164	$.844 \times 10^{-1}$	$.844 \times 10^{-1}$
.01	$.155 \times 10^{-2}$	$.813 \times 10^{-3}$	$.812 \times 10^{-3}$
.001	$.377 \times 10^{-5}$	$.817 \times 10^{-5}$	$.817 \times 10^{-5}$

Table 2.6.1: (2.6.4) for various values of h and τ $(l = 1)$.

The second question is easier to answer. We shall use a result from section 3.3 and state that $E^m_{n,\delta_i}(z)$ and $E^m_{n,\sigma_i}(z)$ are of opposite sign for any $z > 0$. Since $E^m_{n,b}(\lambda_P)$ depends continuously on b, the point where $E^m_{n,b}(\lambda_P) = 0$ can be found by a simple iterative technique; the bisection method, for example. When the interpolation interval is infinite (as in $[\delta_{m+1}, \infty)$), one might first find an upper bound $b_1 > \delta_{m+1}$ such that $E^m_{n,b_1}(\lambda_P)$ is of opposite sign to $E^m_{n,\delta_{m+1}}(\lambda_P)$, but this presents no real difficulties.

2.7 The Non-Primary Components and Stability

It has been shown that a set of rational approximations to the exponential of the type $R^m_{n,b}(z)$ can be found which are exact (in the sense of section 2.6) for the primary component of the transient solution. In this section we discuss ways to ensure the remaining components, which in the true solution are quickly damped, are properly handled. From section 1.5 we ask at least

$$R^m_{n,b}(\lambda_P) > |R^m_{n,b}(z)| \quad \text{whenever} \quad z > \lambda_P. \tag{2.7.1}$$

One method is to require that $R^m_{n,b}(z)$ strictly decrease to some non-negative limit on the interval $[0,\infty)$. Obviously, such an approximation will also be A_0-acceptable. When $m<n$ we only require that $R^m_{n,b}(z)$ be strictly decreasing, since it is asymptotically zero. Surprisingly, such approximations are easily found.

Theorem 2.4

Given n and $m<n$, let b be contained in the $m+1$'st (infinite) interpolation interval of $R^m_{n,b}(z)$. Then $R^m_{n,b}(z)$ strictly decreases with z on the non-negative real axis.

The proof will be left until chapter 3.

Corollary 2.1

Given n, $m<n$, and $\lambda_P>0$, there exists exactly one value of b in the $m+1$'st interpolation interval such that

1. $R^m_{n,b}(z)$ is exact for $z = \lambda_P$,
2. $R^m_{n,b}(z)$ is A_0-acceptable, and
3. condition (2.7.1) is satisfied.

Theorem 2.4 also holds when $m=n$, and b is restricted to the n'th interpolation interval; however the approximation is asymptotically negative here, and so (2.7.1) is not assured. However, Theorem 2 of Norsett [30] can easily be extended to the following:

Theorem 2.5

The approximation $R^n_{n,b}(z)$ is A_0-acceptable when b is contained in the n'th interpolation interval.

Other interpolation intervals may still provide usable approximations. Figure 2.7.1 shows plots of the logarithms of $\exp(-z)$ and $|R^2_{1,b}(z)|$, where the b's where chosen from the first interpolation interval so that the approximation is exact for λ_P = .296, 1.09, and 2.96. When $l=1.$, this satisfies criterion (2.6.3) for time steps $\tau =$ .03, .1, and .3, respectively.

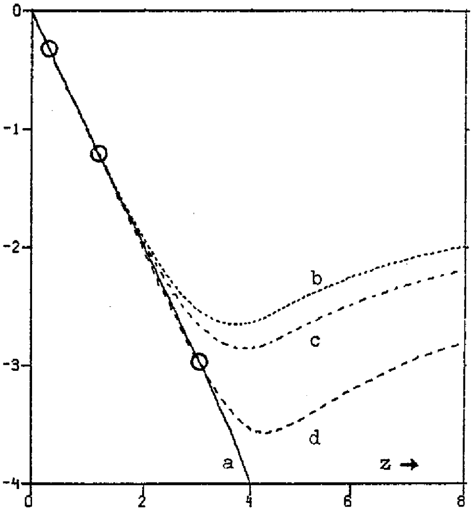

Figure 2.7.1: (a) -z, and $\ln[R^2_{1,b}(z)]$ fitted at
λ_P = (b) .296 (c) 1.09 (d) 2.96 (interpolation points are circled).

It is apparent that (2.7.1) is most likely to be violated when λ_P (proportional to the time step) is large. This suggests calculating, for given m and n, the maximum value λ_S which satisfies the following:

$$\textit{If } 0<\lambda_P<\lambda_S \text{ and } E^m_{n,b}(\lambda_P)=0 \text{ then (2.7.1) holds.} \qquad (2.7.2)$$

Of course, there will be min{$m+1,n$} interpolation intervals from which to choose the values of b, and it is worth considering each interval separately. An interesting feature is that when b is restricted to a single interpolation interval, (2.7.2) appears to be strict; that is, all calculations have shown that when $\lambda_P > \lambda_S$, (2.7.1) *does not* hold.

The values of λ_S which satisfy (2.7.2) are tabulated below for $n \leq 6$ and $0 \leq m \leq n$. Each pair (m,n) has min{$m+1,n$} entries, corresponding to each of the interpolation intervals in order. An entry of 0 means that $\lambda_S < .1$, with the possibility that no value exists which satisfies the condition. An entry of ∞ refers to the results of theorem 2.4.

m n	0	1	2	3	4	5	6
1	∞	.029					
2	∞	2.42 ∞	0 .447				
3	∞	3.60 ∞	1.17 2.92 ∞	0 0 .617			
4	∞	4.50 ∞	2.33 4.24 ∞	0 2.06 3.15 ∞	0 0 .032 .711		
5	∞	5.26 ∞	3.15 5.28 ∞	1.11 3.30 4.54 ∞	0 1.04 2.52 3.28 ∞	0 0 0 .282 .769	
6	∞	5.94 ∞	3.80 6.15 ∞	2.02 4.22 5.66 ∞	0 2.38 3.80 4.73 ∞	0 0 1.71 2.80 3.37 ∞	0 0 0 0 .449 .810

Table 2.7.1: Maximum values of λ_S which satisfy (2.7.2).

Some Remarks

1. For the model problem, condition (2.7.2) implies the constraint

$$\tau \leq \frac{l^2 \lambda_S}{\pi^2}$$

Except for those cases delineated in theorem 2.4, we have a restriction on the time step (typically between $\tau \leq .1$ and $\tau \leq .5$ when $l = 1.$). In particular, the results of test 2 (section 2.5) are explained. Most importantly, *the restriction is independent of the spatial mesh.*

2. For given m and n it is the later interpolation intervals, that is, the larger values of b, which are best able to dampen the higher frequencies.

3. For given n, and ignoring the cases of theorem 2.4, it is the smaller values of m which best dampen the higher frequencies. In particular, *one should not choose $m = n$ if large time steps are to be taken.* This directly opposes the goal of maximizing the order of a method.

To complete the discussion on stability we present the following result. Again the proof is left to the following chapter.

Theorem 2.6

Let n and $m \leq n$ be given, and let $[\delta, \sigma]$ be some finite interpolation interval for $R^m_{n,b}(z)$. Then $R^m_{n,b}(z)$ is A_0-acceptable if $b \in [\delta, \sigma]$ and both $R^m_{n,\delta}(z)$ and $R^m_{n,\sigma}(z)$ are A_0-acceptable.

Table 2.7.2 shows the stability of each interpolation interval for $1 \leq n \leq 6$ and $m \leq n$. The intervals have been organized as in table 2.7.1. The letter S means the interval is A_0-acceptable throughout. Otherwise the interval has the entry U.

m n	0	1	2	3	4	5	6
1	S	S					
2	S	S S	U S				
3	S	S S	S S S	U U S			
4	S	S S	S S S	U S S S	U U S S		
5	S	S S	S S S	S S S S	U S S S S	U U U S S	
6	S	S S	S S S	S S S S	U S S S S	U U S S S S	U U U U S S

Table 2.7.2: Stable (S) and unstable (U) interpolation intervals.

2.8 Further Remarks

We have examined two closely related criteria for deriving approximations which will efficiently solve for transients in the heat equation. Of these, spatial error compensation (SEC) is theoretically the "correct" approach. Exponential fitting (EF) is but a close approximation to the first if spatial discretization is carried out accurately.

However, the EF approach merits attention for the following reasons: First, it is independent of how the spatial discretization is performed. Second, there do exist problems for which there is no discretization error associated with the most persistent component of the complementary problem. An example will be given in chapter 4. Third, EF is useful for general ODE's which are not the result of some spatial discretization. Finally and most important, there exist strong theoretical results concerning both existence and stability. Similar results for SEC would be difficult to derive, and probably the best approach is to assume that the results for exponential fitting extend.

Nevertheless there will be instances when SEC should be used. This is especially true for problems in 2 and 3 spatial dimensions, where accurate spatial discretizations can only be achieved at great expense.

Further remarks should also be made concerning the selection of interpolation intervals for exponential fitting. Apparently, the later interpolation intervals are superior. In fact the so-called "infinite" interpolation intervals result in approximations whose properties would appear to be ideal, and one would assume any size time step can be taken. One problem does arise, however.

Recall that the implementation of our methods (section 1.4) reduces to the repeated solution of systems of the form

$$[\mathbf{I} + \gamma \mathbf{A}]\mathbf{v} = \mathbf{c}, \quad \gamma = b\tau, \tag{2.8.1}$$

where $\mathbf{v}$ is to be determined. For the model problem, the condition number ρ of $\mathbf{I} + \gamma\mathbf{A}$ is roughly

$$\frac{1 + 4\gamma/h^2}{1 + \pi^2\gamma/l^2}.$$

When γ is small and h large, ρ approaches 1. When the opposite is true we can expect a condition number of about $4N^2/\pi^2$. From this point of view, then, small time steps and the earlier interpolation intervals are to be preferred.

A more extreme case occurs when Neumann boundary conditions are introduced in the model problem. Here **A** will be singular, giving

$$\rho = 1 + \frac{4\gamma}{h^2}.$$

Consider the exponential fitting of $R^0_{1,b}(z) = 1/(1+bz)$ at $\lambda_P = \pi^2\tau/l^2$. The appropriate value of γ is

$$\frac{l^2}{\pi^2}(\exp(\frac{\pi^2\tau}{l^2}) - 1).$$

The exponential growth of γ with τ can be expected whenever an infinite interpolation interval is used. We now have

$$\rho = 1 + \frac{4N^2}{\pi^2}(\exp(\frac{\tau\pi^2}{l^2}) - 1).$$

Obviously the time step must be restricted if a reasonably well conditioned system is to result.

Chapter 3
Proof of Theorems

3.1 Introduction

The purpose of this chapter is to prove the results stated in chapter 2. Beforehand, however, an important property of Norsett approximations is required; specifically that the error is uni-modal on the positive real axis. This is presented as a corollary to a general result on exponentially fitting in section 3.2. Section 3.3 then proves theorems 2.2 and 2.3 which comprise the main result. The final section establishes the two stability-related theorems 2.4 and 2.6. This last result is based on a rather interesting property of the current approximations; that is, that they vary monotonically with the parameter b when b is restricted to a single interpolation interval.

In the main we will be concerned with approximations of the form

$$R_{n,b}^{m}(z) = \frac{c_0 + ... + c_m z^m}{(1+bz)^n}$$

where $n \geq 1$, $0 \leq m \leq n$, $b > 0$, and the numerator coefficients chosen so as to maximize the order. z will always be assumed to be real. Define

$$E_{n,b}^{m}(z) = \exp(-z) - R_{n,b}^{m}(z)$$

$E_{n,b}^{m}(z)$ is at least $O(z^{m+1})$ at $z=0$. $R_{n,b}^{m}(z)$ is said to be of *Norsett type* if $E_{n,b}^{m}(z)$ is $O(z^{m+2})$. We say $R_{n,b}^{m}(z)$ is *exact*, or *exponentially fitted*, at z if $E_{n,b}^{m}(z)=0$.

This chapter will use Laguerre polynomials extensively, so the required properities are listed here and then referenced by number when needed.

Some Properties Of The Laguerre Polynomials

1. When λ is a non-negative integer, the generalized Laguerre polynomial of degree n is

$$L_n^\lambda(y) = \sum_{k=0}^{n} \binom{n+\lambda}{k+\lambda} \frac{(-y)^k}{k!}, \quad n \geq 0.$$

2. $L_n^\lambda(y)$ has n distinct simple positive roots.
3. The zeroes of $L_n^\lambda(y)$ and $L_{n-1}^{\lambda+1}(y)$ separate each other.
4. The zeroes of $L_n^\lambda(y)$ and $L_n^{\lambda+1}(y)$ separate each other, with the *i'th* zero of $L_n^\lambda(y)$ preceding the *i'th* zero of $L_n^{\lambda+1}(y)$.
5. $-L_n^{\lambda+1}(y) + L_n^\lambda(y) + L_{n-1}^{\lambda+1}(y) = 0, \quad n \geq 1.$
6. $y\dfrac{d^2L_n^\lambda(y)}{dy^2} + (\lambda+1-y)\dfrac{dL_n^\lambda(y)}{dy} + nL_n^\lambda(y) = 0.$
7. $\dfrac{dL_n^\lambda(y)}{dy} = -L_{n-1}^{\lambda+1}(y), \quad n \geq 1.$
8. $y\dfrac{dL_n^\lambda(y)}{dy} = nL_n^\lambda(y) - (n+\lambda)L_{n-1}^\lambda(y), \quad n \geq 1.$

See, for instance, Lebedev [23].

The main relationship between Laguerre polynomials and the current approximations is given by theorem 2.1.

3.2 Exponential Fitting of Rational Approximations

Lemma 3.1

Suppose $f(z)$ is continuous on the finite interval $[a,b]$ and differentiable on (a,b). Also suppose $f(a) = f(b) = 0$, and $f(z) \neq 0$ if $z \in (a,b)$. Then there exists some point $c \in (a,b)$ such that $f(c)+f'(c) = 0$.

Proof

Assume that $f(z)>0$ on (a,b), since otherwise we would prove the result for $-f(z)$. By Rolle's theorem there exists a point z_1 in (a,b) such that $f'(z_1)=0$, so that

$$f(z_1) + f'(z_1) > 0. \tag{3.2.1}$$

Consider the interval $I = [max\{a,b-.5\},b]$. By compactness, $f(z)$ achieves its maximum on I at some point $z_m < b$. From the mean value theorem there exists a point z_2 in (z_m,b) such that

$$f'(z_2) = \frac{f(b)-f(z_m)}{b-z_m} = \frac{-f(z_m)}{b-z_m} < -f(z_m),$$

and since $f(z_m) \geq f(z_2)$,

$$f(z_2) + f'(z_2) < 0. \tag{3.2.2}$$

By (3.2.1), (3.2.2), and continuity of $f(z)+f'(z)$, there exists a point c between z_1 and z_2 such that $f(c)+f'(c) = 0$. □

Lemma 3.2

Suppose $f(z)$ is analytic on some open, possibly infinite, interval (a,b) on the real axis. If $f(z)$ has M zeroes on (a,b), then $f(z)+f'(z)$ has at least $M-1$ zeroes on (a,b) (including multiples).

Proof

Suppose $f(z)$ has N distinct zeroes at the points $\{\alpha_1, \ldots, \alpha_N\}$, ordered so that $\alpha_i < \alpha_{i+1}$. If $f(z)$ has a zero of order ν at α_i, then $f'(z)$, hence $f(z)+f'(z)$, has a zero of order $\nu-1$ at α_i. This accounts for $M-N$ zeroes. Consider the intervals

$$[\alpha_i, \alpha_{i+1}], \quad i=1,\ldots,N-1.$$

Each interval satisfies the conditions of lemma 3.1; hence $f(z)+f'(z) = 0$ at some point inside the interval, and we have found $N-1$ more zeroes. □

Theorem 3.1

Let

(a,b) be some open, possibly infinite, interval on the real axis,

$p(z)$ be a polynomial of degree m,

$q(z)$ be a polynomial of degree n with at most d distinct zeroes, none of which are on (a,b).

Then $f(z) = \exp(-z) - p(z)/q(z)$ has at most $m+d+1$ zeroes (counting multiples) on (a,b).

Proof

That $f(z)$ has finitely many zeroes is given by theorem 1.3. Now

$$f(z) + f'(z) = \frac{p(z)q'(z) - [p(z) + p'(z)]q(z)}{q^2(z)} = \frac{\bar{p}(z)}{\bar{q}(z)}.$$

In the numerator, $q(z)$ and $q'(z)$ will have $n-d$ common factors which will cancel with factors in the denominator. Hence in reduced form $\bar{p}(z)$ will be of degree $(m+n)-(n-d) = m+d$. It follows that $f(z)+f'(z)$ can have at most $m+d$ zeroes on (a,b). By lemma 3.2, the result now follows. □

Simply put, there exists a direct trade-off between the order of accuracy at the origin, and the ability to satisfy exactness criteria elsewhere on the real axis. By setting the interval of theorem 3.1 to $(-1/b,\infty)$, we have the following results:[1]

Corollary 3.1

The current approximations $R^m_{n,b}(z)$ can be exact for at most one positive value of z. Moreover, the zero of $E^m_{n,b}(z)$ at this point is simple.

The following will be of use in the next section.

Corollary 3.2

Approximations of Norsett type are not exact for any positive value of z.

[1] It has been drawn to the author's attention that these corollaries can also be derived from theorem 3 of Iserles [15], a result based on the order star theory. Nevertheless, theorem 3.1 is still of general interest, since there are implementation advantages associated with repeated factors in the denominator.

3.3 Proof of Theorems 2.2 and 2.3

Lemma 3.3

For fixed $z > 0$, m, and n, there exists at most $\min\{m+1, n\}$ positive values of b such that $E_{n,b}^{m}(z) = 0$.

Proof

Consider

$$\begin{aligned} h(b) &= (1+bz)^n E_{n,b}^{m}(z) \\ &= (1+bz)^n exp(-z) - (c_0 + \ldots + c_m z^m). \end{aligned}$$

Obviously $E_{n,b}^{m}(z) = 0$ only if $h(b) = 0$. From (2.1.4) we see that c_i is an i'th degree polynomial of b when $i \leq n$, and an n'th degree polynomial of b when $i \geq n$. Hence $h(b)$ is an n'th degree polynomial of b, and has at most n roots.

Also, if $m < n$ then

$$\frac{d^{m+1} h(b)}{db^{m+1}} = exp(-z)(1+bz)^{n-m-1} z^{m+1} n \cdots (n-m)$$

for which all roots are negative. It follows from Rolle's theorem that $h(b)$ can have at most $m+1$ positive roots. □

Lemma 3.4

If α_i is the i'th zero of $L_{n-1}^{\lambda+1}(y)$ then

(a) $L_n^{\lambda}(\alpha_i) \begin{cases} < 0 & \text{if } i \ \underline{odd} \\ > 0 & otherwise, \end{cases}$

(b) similarly for $L_n^{\lambda+1}(\alpha_i)$.

If β_i is the i'th zero of $L_n^{\lambda}(y)$ then

(c) $L_{n-1}^{\lambda+1}(\beta_i) \begin{cases} < 0 & \text{if } i \ \underline{even} \\ > 0 & otherwise, \end{cases}$

(d) similarly for $L_n^{\lambda+1}(\beta_i)$.

Proof

Property 3 of the Laguerre polynomials states that the zeroes of $L_n^\lambda(y)$ and $L_{n-1}^{\lambda+1}(y)$ separate each other. Part (a) now follows by noting that $L_n^\lambda(0) > 0$. Part (b) follows from (a), since by the fifth property of Laguerre polynomials

$$L_n^{\lambda+1}(\alpha_i) = L_n^\lambda(\alpha_i).$$

Parts (c) and (d) have a similar proof. □

We will now restate and prove the main results.

Theorem 2.2 (*$m < n$ case*)

Let

$z > 0$, m, and n be given, with $m < n$,

$\delta_1 < \delta_2 < \cdots < \delta_{m+1}$ be the values of b such that $R_{n,b}^m(z)$ is of Norsett type.

$\sigma_1 < \sigma_2 < \cdots < \sigma_m$ be the values of b such that $R_{n,b}^{m-1}(z)$ is of Norsett type.

Then

1. $\delta_1 < \sigma_1 < \delta_2 < \cdots < \sigma_m < \delta_{m+1}$,
2. for a given i, $1 \leq i \leq m$, there exists exactly one value of $b \in (\delta_i, \sigma_i)$ such that $E_{n,b}^m(z) = 0.$,
3. there exists exactly one value of $b > \delta_{m+1}$ such that $E_{n,b}^m(z) = 0.$,
4. for all other positive values of b, $E_{n,b}^m(z) \neq 0$.

Proof

Part 1: From (2.2.2), $\{\delta_i\}$ are the values of b such that

$$L_{m+1}^{n-m-1}(1/b) = 0.$$

Similarly, $\{\sigma_i\}$ are the values of b such that

$$L_m^{n-m}(1/b) = 0.$$

From property 3 of the Laguerre polynomials, the zeroes of $L_{m+1}^{n-m-1}(y)$ and $L_m^{n-m}(y)$ separate each other.

Part 2: Let $b=\delta_i$ where $1 \leq i \leq m$. From section 2.2,

$$E_{n,\delta_i}^m(z) = c_{m+2}z^{m+2} + O(z^{m+3}).$$

To determine the sign of $E_{n,\delta_i}^m(z)$ we may choose any $z>0$ from corollary 3.2. By allowing z to be arbitrarily close to zero we have that $E_{n,\delta_i}^m(z)$ is of the same sign as c_{m+2}. If $m<n-1$, then from theorem 2.1

$$c_{m+2} = (\delta_i)^{m+2}L_{m+2}^{n-m-2}(1/\delta_i),$$

and when $m=n-1$

$$c_{m+2} = -(n+1)(\delta_i)^n L_n^1(1/\delta_i).$$

Now $1/\delta_i$ is the $(m+2-i)$'th zero of $L_{m+1}^{n-m-1}(y)$, so in either case it follows from lemma 3.4 (a), (d)

$$E_{n,\delta_i}^m(z) \quad \begin{cases} <0 & \text{if } m+2-i \ \underline{odd} \\ >0 & \textit{otherwise.} \end{cases}$$

In a similar fashion, and noting that $R_{n,\sigma_i}^m(z) = R_{n,\sigma_i}^{m-1}(z)$ (since $c_m = 0$),

$$E_{n,\sigma_i}^m(z) = E_{n,\sigma_i}^{m-1}(z) \quad \begin{cases} <0 & \text{if } m+1-i \ \underline{odd} \\ >0 & \textit{otherwise.} \end{cases}$$

It follows that $E_{n,\delta_i}^m(z)$ and $E_{n,\sigma_i}^m(z)$ are of opposite sign. Since $E_{n,b}^m(z)$ depends continuously on b, there exists at least one b_1 between δ_i and σ_i such that $E_{n,b_1}^m(z) = 0$.

Part 3: From part 2, $E_{n,\delta_{m+1}}^m(z) < 0$. Now the numerator of $R_{n,b}^m(z)$ can be considered a polynomial of b of degree m, and the denominator a polynomial of degree n. Since $m<n$

$$E_{n,b}^m(z) \to exp(-z) \ \ as \ \ b \to \infty;$$

$$E^m_{n,b}(z) \rightarrow exp(-z) \;\; as \;\; b \rightarrow \infty;$$

that is to say, $E^m_{n,b}(z)$ becomes positive at some point. Again by continuity there exists at least one $b_1 > \delta_{m+1}$ such that $E^m_{n,b_1}(z)=0$.

Part 4: We have found $m+1$ intervals, each of which contains at least one value of b where $E^m_{n,b}(z)=0$. Lemma 3.3 states there can be at most $m+1$ such values; hence each interval contains at most one. Moreover there exist no others. □

Theorem 2.3 ($m=n$ *case*)

Let

$z > 0$ and n be given,

$\delta_1 < \delta_2 < \cdots < \delta_n$ be the values of b such that $R^n_{n,b}(z)$ is of Norsett type,

$\sigma_1 < \sigma_2 < \cdots < \sigma_n$ be the values of b such that $R^{n-1}_{n,b}(z)$ is of Norsett type.

Then

1. $\delta_1 < \sigma_1 < \delta_2 < \cdots < \delta_n < \sigma_n$,
2. for a given i, $1 \le i \le n$, there exists exactly one value of $b \in (\delta_i, \sigma_i)$ such that $E^n_{n,b}(z) = 0$,
3. for all other positive values of b, $E^n_{n,b}(z) \neq 0$.

Proof

Part 1: From (2.2.2), $\{\delta_i\}$ are the values of b such that

$$L^1_n(1/b) = 0.$$

Similarly $\{\sigma_i\}$ are the values of b such that

$$L^0_n(1/b) = 0.$$

Part 1 now follows from property 4 of the Laguerre polynomials.

$$E^n_{n,\delta_i}(z) = c_{n+2}z^{n+2} + O(z^{n+3}).$$

As before, $E^n_{n,\delta_i}(z)$ has the same sign as c_{n+2}. From theorem 2.1

$$c_{n+2} = (n+1)(n+2)(\delta_i)^n L_n^2(1/\delta_i).$$

Now $1/\delta_i$ is the $(n+1-i)$'th zero of $L_n^1(y)$, so from lemma 3.4 (d)

$$E^n_{n,\delta_i}(z) \quad \begin{cases} <0 & \text{if } n+1-i \ \underline{\textit{even}} \\ >0 & \textit{otherwise}, \end{cases}$$

and from the proof of theorem 2.2,

$$E^n_{n,\sigma_i}(z) = E^{n-1}_{n,\sigma_i}(z) \quad \begin{cases} <0 & \text{if } n+1-i \ \underline{\textit{odd}}, \\ >0 & \textit{otherwise}. \end{cases}$$

It follows that $E^n_{n,\delta_i}(z)$ and $E^n_{n,\sigma_i}(z)$ are of the opposite sign, and the result follows as before.

Part 3: Similar to part 4 of the previous theorem. □

3.4 Proof of the Stability Theorems

Theorem 2.4

Given n and $m < n$, let b be contained in the $m+1$'st interpolation interval of $R_{n,b}^m(z)$. Then $R_{n,b}^m(z)$ strictly decreases with z on the non-negative real axis.

Proof

Let

$$R_{n,b}^m(z) = \frac{c_0 + ... + c_m z^m}{(1+bz)^n} = \frac{p(z)}{(1+bz)^n}.$$

Then

$$\frac{dR_{n,b}^m(z)}{dz} = \frac{(1+bz)p'(z) - nbp(z)}{(1+bz)^{n+1}} = -\left[\frac{\bar{c}_0 + ... + \bar{c}_m z^m}{(1+bz)^{n+1}}\right]. \quad (3.3.1)$$

We will first consider the coefficients $\bar{c}_i$ for $i \leq m-1$. Since $R_{n,b}^m(z)$ is an order m approximation to $\exp(-z)$, (3.3.1) will be an order $m-1$ approximation to $-\exp(-z)$. By theorem 2.1, then,

$$\bar{c}_i = b^i L_i^{n+1-i}(1/b), \quad i = 0, \cdots, m-1.$$

From definition 2.6.1, b is in the $m+1$'st interpolation interval if $1/b \leq 1/\delta$, where $1/\delta$ is the smallest zero of $L_{m+1}^{n-m-1}(y)$. By applying property 3 of the Laguerre polynomials $m-i+1$ times, and then property 4 once, it is seen that $1/\delta$ is smaller than the zeroes of $L_i^{n+1-i}(y)$. Hence when $y \leq 1/\delta$, $L_i^{n+1-i}(y)$ will be of one sign, and therefore positive since $L_i^{n+1-i}(0) > 0$. It follows that $\bar{c}_i > 0$ when $b \geq \delta$.

For the remaining coefficient

$$\bar{c}_m = (n-m)bc_m.$$

Now $c_m = b^m L_m^{n-m}(1/b)$, which is also positive when $b \geq \delta$ (using only property 3 this time). Since all the coefficients $\{\bar{c}_i\}$ are positive, it follows that

$$\frac{dR_{n,b}^m(z)}{dz} < 0 \textit{ when } z \geq 0. \quad \square$$

Lemma 3.5

Let n, $m \le n$, and $z > 0$ be given. If b is positive then

$$\frac{\partial R_{n,b}^m(z)}{\partial b} = 0$$

if and only if $1/b$ is a zero of

$$\begin{cases} L_m^{n-m-1}(y) & \text{if } n < m, \\ L_{n-1}^1(y) & \text{if } m = n. \end{cases}$$

Proof

$$\frac{\partial R_{n,b}^m(z)}{\partial b} = \frac{\bar{c}_0 + \cdots + \bar{c}_{m+1} z^{m+1}}{(1+bz)^{n+1}}$$

where

$$\bar{c}_i = \begin{cases} 0 & \text{if } i=0, \\ c'_i + bc'_{i-1} - nc_{i-1} & \text{if } 1 \le i \le m, \\ bc'_m - nc_m & \text{if } i = m+1, \end{cases}$$

$$c_i' = \frac{dc_i}{db}, \quad \text{and} \quad c_i = b^i L_i^{n-i}(1/b).$$

We will first consider the case where $1 \le i \le m$. By property 7 of the Laguerre polynomials,

$$c_{i-1} = -b^{i-1} \frac{dL_i^{n-i}(1/b)}{d(1/b)}$$

hence

$$\bar{c}_i = b^{i-1} \frac{d^2 L_i^{n-i}(1/b)}{d(1/b)^2} + [-b^{i-2} + (n-i+1)b^{i-1}] \frac{dL_i^{n-i}(1/b)}{d(1/b)} + ib^{i-1} L_i^{n-i}(1/b).$$

Dividing both sides by b^{i-1} and setting $y = 1/b$ gives

$$\frac{\bar{c}_i}{b^{i-1}} = y\frac{d^2L_i^{n-i}(y)}{dy^2} + [(n-i)+1-y]\frac{dL_i^{n-i}(y)}{dy} + iL_i^{n-i}(y)$$

$$= 0$$

by property 6, and we have that $\partial R_{n,b}^m(z)/\partial b = 0$ iff $\bar{c}_{m+1}=0$. Proceeding similarly for $\bar{c}_{m+1}$ gives

$$\frac{\bar{c}_{m+1}}{b^m} = (m-n)L_m^{n-m}(1/b) + (1/b)\frac{dL_m^{n-m}(1/b)}{d(1/b)}$$

If $m=n$ then $\bar{c}_{m+1} = b^{n-1}L_{n-1}^1(1/b)$ by property 7. Otherwise by property 8

$$\frac{\bar{c}_{m+1}}{b^m} = -nL_m^{n-m}(1/b) + nL_{m-1}^{n-m}(1/b),$$

and by property 5 we have that $\bar{c}_{m+1} = nb^m L_m^{n-m-1}(1/b)$. □

Lemma 3.6

Let n, $m \leq n$, and $z > 0$ be given, and let I be an interpolation interval of $R_{n,b}^m(z)$. Then $R_{n,b}^m(z)$ is strictly monotonic as a function of b when $b \in I$.

Proof

We will only consider the case $m < n$, since the argument for $m = n$ differs only in detail.

Let δ_i and σ_i, $1 \leq i \leq m$, be the Norsett points as in theorem 2.2. From (2.2.2), $1/\delta_i$ is the $m+2-i$'th zero of $L_{m+1}^{n-m-1}(y)$, and $1/\sigma_i$ is the $m+1-i$'th zero of $L_m^{n-m}(y)$. From properties 3 and 4 of the Laguerre polynomials, the zeroes of $L_m^{n-m-1}(y)$ are contained within the intervals $(1/\delta_{i+1}, 1/\sigma_i)$, $1 \leq i \leq m$. Hence from lemma 3.5, $\partial R_{n,b}^m(z)/\partial b = 0$ only when $b \in (\sigma_i, \delta_{i+1})$. The result now follows by noting these intervals are disjoint from the interpolation intervals. □

The final theorem of chapter 2 now follows.

Theorem 2.6

Let n and $m \leq n$ be given, and let $[\delta,\sigma]$ be some finite interpolation interval for $R^m_{n,b}(z)$. Then $R^m_{n,b}(z)$ is A_0-acceptable if $b \in [\delta,\sigma]$ and both $R^m_{n,\delta}(z)$ and $R^m_{n,\sigma}(z)$ are A_0-acceptable.

Proof

Obvious from lemma 3.6 and the definition of A_0-acceptability. □

Chapter 4
A Study of a Thermal Printhead

4.1 Introduction

In this chapter we present some numerical studies with the following purposes in mind: first, to demonstrate how the methods developed in chapter 2 might be utilized under more interesting circumstances, including two spatial dimensions and Neumann boundary conditions; second, to carry out comparisons between these and more standard methods; and finally, to show that problems exist where exponential fitting (or spatial error compensation) is of obvious utility.

The problem is taken from Morris ([26],[27]). Although we shall describe it completely in section 2.2, only one portion, the two-dimensional thin film, will be examined here. Numerical methods for the thin film problem are discussed in section 2.3. In the final section we present the results of numerical tests.

All computations were carried out in single precision FORTRAN on the Honeywell DPS-8/49 computer at the University of Waterloo Faculty of Mathematics. The only exception is the double precision calculation of certain coefficients, the source code for which is listed in appendix A.

4.2 Description of the Physical Problem

The problem to be studied involves the heating of a print head from a thermal line printer. A print head is composed of a matrix of individual elements, as shown in figure 4.2.1. Printing takes place when thermally sensitive paper is pressed against the print head and some or all of the elements are heated. Each heated element reacts with the paper to create a single dot, the dots from a single print head together forming a single character. When the image is formed, the elements are allowed to cool until the next character. The principle advantage to this system is that it requires few moving parts.

The goal is to develop a printer head which prints as quickly as possible. If not enough time is allowed between characters, the elements overheat and a blurred image results. Conversely when the elements are underheated the paper will remain blank. Designing a print head which avoids these problems is expensive if each prototype must be built in order to determine its properties; hence the need for numerical modelling.

The construction of a single element is shown in figure 4.2.2. The x and y coordinates run in the horizontal directions in the figure. The z coordinate is assumed to run in the vertical direction, increasing downwards. In this example the dimensions of an element are taken to be $0 \le x,y,z \le 1$, the boundary for which is denoted $\partial\Omega$. Morris argued in [26] that each element of the print head may be modelled independently if Neumann boundary conditions are assumed in the x and y directions.

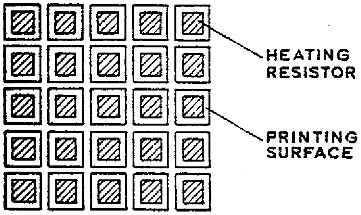

Figure 4.2.1: A 5 x 5 matrix thermal print head - top view (from Morris [26]).

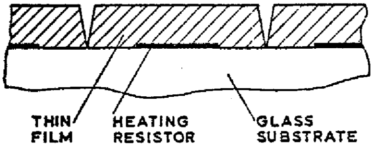

Figure 4.2.2: A single element of the matrix - side view (from Morris [26]).

Heat flow is examined in two separate parts of the element; the thin film (assumed here to be silver), which serves as the point of contact for the paper, and the glass substrate upon which the film lies. The temperature gradient of the film in the z-direction is assumed to be negligible, resulting in a two-dimensional model. The film is heated underneath by a square electrical resistor of dimension $\alpha \leq x,y \leq \beta$ whose current switches on or off every t_0 seconds. Cooling of the plate is assumed to take place through contact with the surrounding air, which is held at a constant ambient temperature u_∞. The modelling heat equation for the thin film is then as follows:

$$\frac{\partial u}{\partial t} = \sigma(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}) - \nu(u - u_\infty) + \epsilon sw(t)H(x,y) \tag{4.2.1}$$

$$\frac{\partial u}{\partial x} = 0 \ on \ \partial\Omega_{x=0,1}, \ \frac{\partial u}{\partial y} = 0 \ on \ \partial\Omega_{y=0,1}.$$

where

$$\sigma = \frac{K}{\rho C}$$

$$\nu = \frac{h_0}{\rho C D}$$

$$\epsilon = \frac{q}{\rho C}$$

$$sw(t) = \begin{cases} 1 & \text{if } t \in [2nt_0,(2n+1)t_0], \ n = 0,1,2,\cdots \\ 0 & otherwise \end{cases}$$

$$H(x,y) = \begin{cases} 1 & \text{if } (x,y) \in [\alpha,\beta]\text{x}[\alpha,\beta] \\ 0 & \textit{otherwise.} \end{cases}$$

The constants have the following meanings and typical values:

K	Thermal conductivity coefficient	1.
ρ	Density	10.49
C	Specific heat	.0556
q	Heat generated	10.
h_0	Convective heat transfer coefficient	.000053
D	Thickness of film	.000015
u_∞	Ambient temperature	0.

The glass substrate effectively acts as a heat sink, with the boundary condition at $z = 0$ defined as the temperature of the thin film. The modelling equation of the glass substrate is fully three-dimensional:

$$\frac{\partial u}{\partial t} = \sigma_1(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}) \tag{4.2.2}$$

$$\frac{\partial u}{\partial x} = 0 \text{ on } \partial\Omega_{x=0,1}, \quad \frac{\partial u}{\partial y} = 0 \text{ on } \partial\Omega_{y=0,1}, \quad \frac{\partial u}{\partial z} = 0 \text{ on } \partial\Omega_{z=1}.$$

where

$$\sigma_1 = \frac{K_1}{\rho_1 C_1}$$

and with constant values:

K_1	Thermal conductivity coefficient	.0028
ρ_1	Density	2.4
C_1	Specific heat	.2

When the electric current is fixed at one position, either on or off, the system will approach a steady state condition. However, if printing is carried out as rapidly as possible then switching occurs long before, and the system remains in a constant state of transience. Figure 4.2.3 shows a typical situation. Note that after a few switchings, a kind of steady state is approached where the response becomes very near periodic. We are in the unusual position that the failure of a numerical method to follow the transient phase accurately can result in the wrong long term solution.

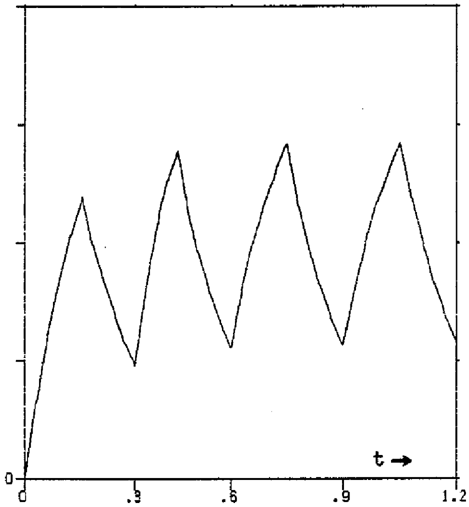

Figure 4.2.3: Temperature of the thin film at a single point ($t_0 = .15$).

As remarked earlier, we will examine only the two-dimensional thin film problem. Furthermore, since Morris [26] found no serious loss of accuracy due to the switching of the electric current, we limit our discussion to the transient portion of problems whose source terms are held constant in time.

4.3 Numerical Methods

We will discuss various numerical approaches for the two-dimensional thin film problem described in the previous section.

All methods are based on the same spatial discretization. A rectilinear grid is superimposed on the film plate, dividing the x and y directions into N equally spaced intervals. For Neumann boundary conditions the nodes at the boundaries are included, resulting in a system of $(N+1)^2$ mesh points. In all numerical tests we will take $N = 10$ ($h = .1$). The central difference operator will be used as an approximation to the second derivative at each point; that is

$$u_{xx}(x,y) \approx \delta_x^2 u(x,y) = \frac{u(x+h,y) - 2u(x,y) + u(x-h,y)}{h^2} \qquad (4.3.1)$$

and similarly for $u_{yy}(x,y)$. When $x = 0$ or 1, (4.3.1) involves mesh points which lie outside the plate. However, from the boundary conditions we set

$$u_x(x,y) \approx \frac{u(x+h,y) - u(x-h,y)}{2h} = 0.$$

At $x = 0$, for instance, this can substituted into (4.3.1), giving

$$u_{xx}(0,y) \approx \frac{2u(h,y) - 2u(0,y)}{h^2},$$

and so the point $u(-h,y)$ does not appear.

Define $\mathbf{W}_{N+1}$ as the $(N+1)$ by $(N+1)$ matrix

$$\frac{1}{h^2}\begin{bmatrix} 2 & -2 & & & & 0 \\ -1 & 2 & -1 & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & -1 & 2 & -1 \\ 0 & & & & -2 & 2 \end{bmatrix},$$

$\mathbf{I}_{N+1}$ as the $(N+1)$ by $(N+1)$ identity matrix, and

$$\mathbf{U} = \mathbf{W}_{N+1} \otimes \mathbf{I}_{N+1}$$

$$\mathbf{V} = \mathbf{I}_{N+1} \otimes \mathbf{W}_{N+1}$$

where $\otimes$ is the usual tensor product. The problem can then be written in the semi-discrete form

$$\frac{d\mathbf{u}(t)}{dt} = -\sigma\,(\mathbf{U} + \mathbf{V})\mathbf{u}(t) - \nu\mathbf{u}(t) + \mathbf{s}(t); \quad \mathbf{u}(0) \text{ given.} \tag{4.3.2}$$

Splitting Methods

Splitting methods are based on the following observation; that $\mathbf{U}$ and (under re-ordering) $\mathbf{V}$ are tridiagonal, and hence can be inverted at comparatively little expense. Consider the problem (4.3.2) when $\nu = \mathbf{s}(t) = 0$. The exact solution is then

$$\mathbf{u}(t+\tau) = \exp(-\tau\sigma(\mathbf{U}+\mathbf{V}))\;\mathbf{u}(t). \tag{4.3.3}$$

If $\mathbf{U}$ and $\mathbf{V}$ commute (that is, if $\mathbf{UV} = \mathbf{VU}$), which is true for this problem, then (4.3.3) can be written

$$\mathbf{u}(t+\tau) = \exp(-\tau\sigma\mathbf{U})\exp(-\tau\sigma\mathbf{V})\mathbf{u}(t),$$

and we may then proceed as in section 1.2, solving at worst tridiagonal systems.

It is not clear how these methods can best be adopted for complicated forcing terms. Morris, however, examined the following second order splitting method, known as Alternating Direction Implicit, or ADI.

$$\mathbf{v}^* \leftarrow (\mathbf{I} + \alpha\mathbf{U})^{-1}\,[(\mathbf{I} - \alpha\mathbf{U})(\mathbf{I} - \alpha\mathbf{V})\mathbf{v}_k - \tau\nu\mathbf{v}_{k+1/2} + \tau\mathbf{s}(t+\tau/2)];$$

$$\mathbf{v}_{k+1} \leftarrow (\mathbf{I} + \alpha\mathbf{V})^{-1}\mathbf{v}^*;$$

where

$$\mathbf{v}_{k+1/2} = (\mathbf{I} - \alpha[\mathbf{U} + \mathbf{V}])\mathbf{v}_k - \frac{\tau}{2}\nu\mathbf{v}_k + \frac{\tau}{2}\mathbf{s}(t),$$

$$\alpha = \frac{\tau\sigma}{2}.$$

In addition Morris considered the Locally One Dimensional (LOD) method.

However, since its performance is nearly identical to that of ADI, this method will not be mentioned again.

It is interesting that the radiation term νu was treated as a source term rather than part of the elliptic operator Lu. With the splitting

$$\exp(-\tau(\sigma\mathbf{U} + \sigma\mathbf{V} + \nu\mathbf{I})) = \exp(-\tau\sigma\mathbf{U})\exp(-\tau\sigma\mathbf{V})\exp(-\tau\nu\mathbf{I})$$

we arrive at a method which does not require an explicit predictor formula:

$$\mathbf{v}^{*} \leftarrow (\mathbf{I}+\alpha\mathbf{U})^{-1}\,[\exp(-\tau\nu)(\mathbf{I}-\alpha\mathbf{U})(\mathbf{I}-\alpha\mathbf{V})\mathbf{v}_k + \tau\exp(-\tau\nu/2)\mathbf{s}(t+\tau/2)];$$

$$\mathbf{v}_{k+1} \leftarrow (\mathbf{I}+\alpha\mathbf{V})^{-1}\mathbf{v}^{*};$$

This scheme (which Morris did not consider) will be referred to as ADI1. Although based on the (1,1) Padé approximation, the method is L-stable whenever $\nu > 0$. By comparison, the ADI scheme can be shown to be unstable for $\tau \geq 2/\nu$.

A General Approach

Alternatively we can take the approach described in sections 1.2 - 1.4. In this case the semi-discrete form is written

$$\frac{d\mathbf{u}(t)}{dt} = -\mathbf{A}\mathbf{u}(t) + \mathbf{s}(t); \quad \mathbf{u}(0)\ \textit{given},$$

where

$$\mathbf{A} = \sigma(\mathbf{U} + \mathbf{V}) + \nu\mathbf{I}$$

is a discrete approximation to the operator L where

$$Lu = -\sigma(u_{xx} + u_{yy}) + \nu u.$$

The procedure is then identical to the 1-dimensional case discussed earlier, and affords far more flexibility in the treatment of source terms than the splitting methods. Moreover, serious complications are avoided when **U** and **V** fail to commute. However, the resulting systems are banded and block tridiagonal, and in general are much more difficult to solve for than a simple tridiagonal matrix.

Two classes of methods have been developed for the solution of sparse systems of this type. The first, the iterative methods (Varga [38], or Hageman and Young [12]), require little memory space, but cannot take advantage of work carried out in previous solutions. For this reason a direct method has been adopted; that is, the quotient minimum degree ordering scheme implemented on the SPARSPAK package.

SPARSPAK (see George, et al [9], [10]) is a user's interface to various sparse matrix solvers. The matrix is assumed to be structurally symmetric and to require no pivoting for numerical stability. Both assumptions are commonly met by systems arising from the heat conduction problem. The method of solution is divided into six separate stages:

1. The non-zero structure of the matrix is specified.
2. Rows and columns are reordered so as to reduce the amount of fill-in occurring during factorization.
3. The values of the matrix elements are specified.
4. The matrix is factorized using LU decomposition.
5. The right hand side vector is specified.
6. The system is solved.

The important feature is that any stage may be repeated without redoing previous stages. In the current schemes, for instance, steps 1 through 4 are performed once at time zero, and steps 3 and 4 are redone whenever the parameter b or time step is changed (hopefully rarely). Most time steps, then, involve only the repeated execution of steps 5 and 6. The savings are critical, since matrix factorization is more than 5 times as expensive as a single solve in our example.

The current approximations, whose denominators are composed of a single repeated linear factor of **A**, reduce the amount of storage and LU decomposition required to a minimum. Schemes requiring the inversion of two or more different linear systems per time step complicate matters considerably. Since SPARSPAK can keep at most one system in core at a time, the factorized matrices must be stored on disk files and read back in before each solve. It is possible that the use of rationals with multiple poles (for instance, high order

Padé approximants) could preclude direct methods entirely.

We will consider only those methods based on the repeated pole approximation (2.1.3). For convenience these will be referred to as the RP schemes. The parameter b will be chosen so as to give an approximation of Norsett type, or to exponentially fit (EF) some prescribed eigenvalue.

Exactness Properties

Consider the problem (4.2.1) with $sw(t) = H(x,y) = 1$. Given $u(x,y,t)$, the exact solution at the following time step is

$$u(x,y,t+\tau) = \frac{\epsilon}{\nu} + \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} a_{ij}\phi_{ij}(x,y)\psi_{ij}(t),$$

where

$$\phi_{ij}(x,y) = cos(i\pi x)cos(j\pi y),$$

$$\psi_{ij}(t) = \exp[-\tau(\sigma\pi^2(i^2 + j^2) + \nu)],$$

$$a_{ij} = \int_0^1\int_0^1 [u(x,y,t) - \frac{\epsilon}{\nu}]\phi_{ij}(x,y)dxdy.$$

The methods ADI, ADI1, and those of type RP, can be considered in the light of various exactness criteria. Specifically, we ask if the methods are exact for the polynomial particular solution

$$u_p(x,y,t) = \frac{\epsilon}{\nu}.$$

In this respect the ADI and RP methods are correct, but not ADI1. Of course if the source term was polynomial and non-constant in time, only the RP methods (if properly chosen - Theorem 1.2) would be exact.

In addition we consider the complementary problem. The RP methods can be exponentially fitted at one positive eigenvalue, and the obvious choice is that associated with the most persistent component $\phi_{0,0}(x,y) = 1$. It can easily be shown that the ADI1 method is also exact in this sense. This is summarized in the following chart:

	ADI	ADI1	RP(Norsett)	RP(EF)
Polynomial particular soln.	E		E	E
First Fourier component		E		E

Table 4.3.1: Exactness properties of the numerical methods (E = Exact).

An interesting property is that there is no spatial discretization error associated with the particular solution u_p or with the first Fourier mode. In this last instance, spatial error compensation is not required; exponential fitting is the correct approach.

4.4 Numerical Tests on the Thin Film Problem

In this section we present the results of numerical tests on the thin film problem. The problem examined has only a constant heat source:

$$\frac{\partial u}{\partial t} = \sigma(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}) - \nu u + \epsilon,$$

and initial condition

$$u(x,y,0) = \cos(\pi x)\cos(\pi y).$$

The exact solution is known to be

$$u = \frac{\epsilon}{\nu}(1 - \exp(-\nu t)) + \cos(\pi x)\cos(\pi y)\exp(-(\nu + 2\sigma\pi^2)t).$$

Errors are displayed using the function

$$\log_{10} ||\mathbf{u}(t) - \mathbf{v}(t)||_\infty$$

where $\mathbf{u}(t)$ is the exact solution sampled at the mesh points, and $\mathbf{v}(t)$ is the computed solution, both at time t.

Figure 4.4.1 shows the consequences of fitting each of the two Fourier components which make up the complementary problem. The uniform error of an RP scheme based on the $R_{1,b}^1(z)$ approximation has been plotted as a function of b for various times in the solution. The best choice for fitting can be seen to vary with time. Earlier in the solution it is best to fit (or actually, spatial error compensate) the faster decaying component. If accuracy at larger times is required, fitting the slower decaying component is preferable.

One solution to this dilemma is to develop approximations which fit both Fourier components at once; however, such methods are beyond the scope of this thesis. A second approach is to fit different components at different times in the solution. Probably the best solution is to fit the most persistent component, but ensure the error near time zero is small through small time increments and high order. This is similar to what is done using conventional methods, although the time steps can now be increased at a much earlier stage.

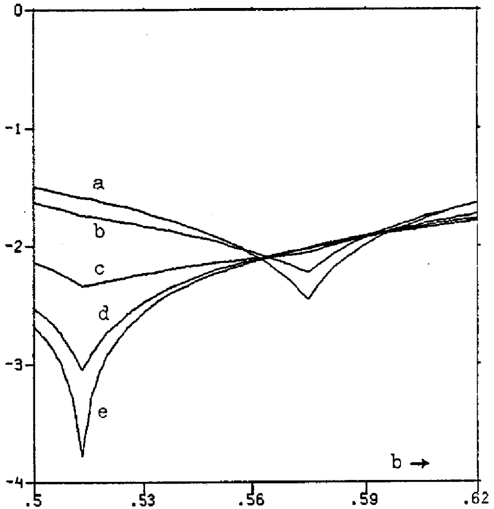

Figure 4.4.1: Error of $R^1_{1,b}(z)$ as a function of b at t = (a) .025 (b) .05 (c) .1 (d) .15 (e) .2, τ = .025.

Numerical tests were carried out using the following methods:

A	ADI	Order 2
B	ADI1	Order 2, EF

and RP methods based on the following approximations to the exponential:

C	$R^0_{1,b}$	Order 0, EF
D	$R^1_{1,.5}$	Order 2, Norsett (Crank-Nicolson)
E	$R^1_{1,b}$	Order 1, EF
F	$R^1_{2,.29289}$	Order 2, 1st Norsett (L21)
G	$R^1_{2,b}$	Order 1, 1st EF
H	$R^1_{2,1.70711}$	Order 2, 2nd Norsett
I	$R^1_{2,b}$	Order 1, 2nd EF
J	$R^2_{2,.78867}$	Order 3, 2nd Norsett
K	$R^2_{2,b}$	Order 2, 2nd EF

All exponential fitting was carried out for the first Fourier mode $\phi_{0,0}(x,y)$. "1st EF" means the parameter b was chosen from the first interpolation interval, although the exact value remains undetermined until the time step is specified.

Figure 4.4.2 shows the uniform error as a function of time for methods A - E using a constant time step of $\tau = .025$. Time steps of .05 and .01 were also studied, but no significant changes in the qualitative performance were found.

Note that the results of comparisons between the methods depend on where the error is analyzed. For instance, ADI1 is superior to ADI for $t \leq .4$, but behaves poorly when the steady state solution is approached. This was expected from the exactness properties. The advantage of the exponentially fitted RP methods becomes clear only when the unfitted component has disappeared (about $t = .2$). Soon after this point, however, the error becomes essentially zero.

It is interesting to compare methods based on approximations with the same form. Methods D and E, for instance, have values for b of .5 and .51262, respectively. This apparently slight change gives rather marked differences in behaviour. Figure 4.4.3 compares errors for the methods F through I. The first interpolation interval of the $R_{2,b}^1$ approximation was found to be considerably better than the second except when very large time steps were taken. Methods J and K are examined in figure 4.4.4. High order was found useful among the Norsett approximations, but not among the exponentially fitted methods (compare K with G or E). The added expense makes method K less desirable for problems with constant source terms.

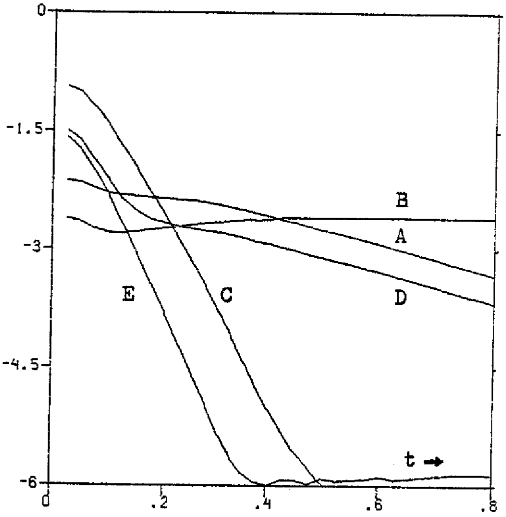

Figure 4.4.2: Error for methods A - E, τ = .025.

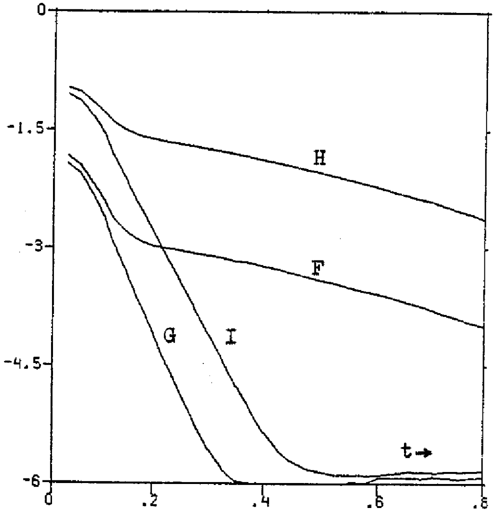

Figure 4.4.3: Error for methods F - I, τ = .025.

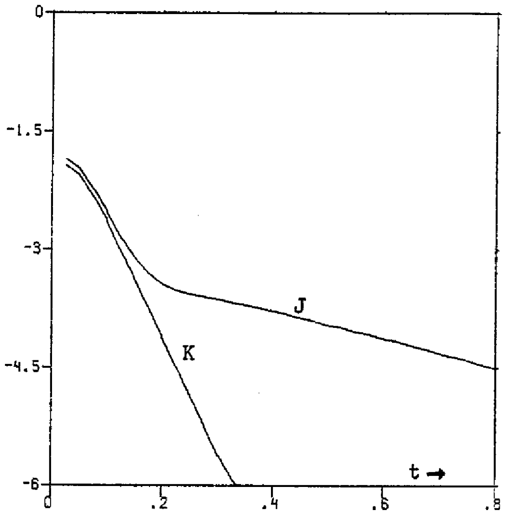

Figure 4.4.4: Error for methods J and K, $\tau = .025$.

Cost Comparisons

We have not yet attempted to compensate for differences in the costs of the various methods. Doing so is not at all straightforward, since the cost per time step is highly dependent on the implementation. For instance, there exists substantial tradeoffs between execution time and memory space when choosing among the options of the SPARSPAK package. In addition it must be assumed that the values of b and τ in the RP schemes are changed only rarely, since otherwise the cost of refactorization must be taken into account.

For simplicity we have chosen implementations which optimize the execution speed, measured in operations per time step. An operation is defined as a single precision floating point multiply or divide. Memory space is measured in the number of floating point variables requiring concurrent storage. Assuming an RP scheme is based on the approximation $R_{n,b}^{m}(z)$, rough estimates of the costs per time step are given below:

	ADI	ADI1	RP
Operations	1200	960	$1540n + 120(m+1)$
Storage	500	300	6000

Table 4.4.1: Approximate costs per time step.

Comparisons between methods were made by examining the amount of work required to achieve a given error tolerance at a given time in the solution. Computation began at $t = 0$ for each entry in the tables below, with the time increment held constant during the course of a single run.

Time Method	$t = .1$	$t = .3$	$t = .6$
A	14.4	28.8	31.2
B	4.8	18.2	38.4
C	166	9.9	5.0
D	17.8	28.5	32.0
E	14.2	7.1	10.7
F	19.9	39.8	43.2
G	19.9	13.3	13.3
H	206	206	216
I	166	16.6	6.6
J	20.6	27.5	34.4
K	20.6	13.8	17.2

Table 4.4.2: Work (operations/1000) required for error $\leq .001$.

Again the results of comparisons vary with time. The splitting methods appear to be most efficient very early on. At $t = .3$ the exponentially fitted methods are superior. At the latest time level, those exponentially fitted methods which possess the monotonicity property of theorem 2.4 (methods C and I) are clearly the best. This last result supports what we have said in section 2.7; that is, an approximation must have special properties if it is to properly attenuate the unfitted components in the presence of large time increments.

The following tables compare a Norsett method (F) with an exponentially fitted method (G) under a range of error tolerances. It appears that exponentially fitting becomes progressively more valuable as error tolerances are reduced.

Error Method	.01	.001	.0001
F	13.3	39.8	113.
G	10.	13.3	16.6

Table 4.4.3: Work required for various error tolerances at $t = .3$.

Error Method	.01	.001	.0001
F	10.	43.2	126.
G	10.	13.3	16.6

Table 4.4.4: Work required for various error tolerances at $t = .6$.

Appendix A

Section 1.4 describes an efficient means of handling source terms in the method of rational approximations. It has been found, however, that the calculation of parameters can be tedious even for very simple cases. This appendix provides FORTRAN routines to carry out the task for approximations involving a single n'th order real pole.

The two user interfaces PALPHA and POMEGA return the values of parameters $\{\alpha_i\}$ and $\{\omega_{ij}\}$, respectively. Since these routines are fully commented, they will not be described here. All other routines are subordinate and need not be accessed directly by the user. The rational approximation is assumed to be of type $R^m_{n,b}(z)$ (section 2.1). Approximations of the same form, but having numerator coefficients not determined solely by order constraints, may be accommodated by appropriate changes to routine RCOEFF. All routines were tested on the Honeywell DPS-8/49 computer at the University of Waterloo Faculty of Mathematics.

```
C***********************************************************************
C   PALPHA    Calculate the coefficients ALPHA(i), i=1,...,N+1, where
C
C            M                    ALPHA(2)            ALPHA(N+1)
C          R  (z) = ALPHA(1) + --------- + ... + ------------
C           N,B                  (1 + B z)           (1 + B z)**N
C
C    Input Parameters:
C         M, N, B        Parameters describing a repeated pole
C                        approximation of the type (2.1.3), where the
C                        numerator coefficients are determined so as to
C                        maximize the order at zero.
C
C         WORK           Work array of M+1 double precision variables.
C
C    Output Parameters:
C         ALPHA          In the context of section 1.4,
C                        ALPHA(i)  <=>  <alpha>
C                                              i-1
C                        Note that if M < N then ALPHA(i)=0, i=1,...,N-M.
C
C    Restrictions:  M >= 0, N >= max{1,M}, B > 0.
C
C    All FP variables (B,WORK,ALPHA) are double precision.
C***********************************************************************
          SUBROUTINE PALPHA (M,N,B,ALPHA,WORK)
          DOUBLE PRECISION B,ALPHA(1),WORK(1)
          CALL RCOEFF (M,N,B,WORK)
          CALL PARTF (M,N,B,WORK,ALPHA)
          RETURN
          END


C***********************************************************************
C    POMEGA   Calculate OMEGA(i+1,j), i=0,...,IORDER, j=1,...,N, where
C
C                  OMEGA(i+1,1)          OMEGA(i+1,N)
C          W (z) = ------------ + ... + ------------,  i = 0,...,IORDER
C           i        (1 + B z)            (1 + B z)**N
C
C              are the weight functions described in (1.4.2).
C
C     Input Parameters:
C          M, N, B        Parameters describing a repeated pole
C                         approximation of the type (2.1.1), where the
C                         numerator coefficients are determined so as to
C                         maximize the order at zero.
C
C          IORDER         No. of weighting functions to be calculated - 1.
C                         (<= the order of the approximation).
C
C          NODES(i+1),    Interpolation nodes (in the context of section 1.4,
C          i=0,...,       NODES(i)  <=>   a  ).  Nodes must be distinct!
C          IORDER                          i-1
```

```
C
C         NDIM          Row dimension of OMEGA. (>= IORDER+1)
C
C         WORK          Work array of 3*P*P double precision variables
C                       where P = max{N+1,IORDER+1}.
C
C     Output Parameters:
C         OMEGA         In the context of section 1.4,
C                       OMEGA(i,j)  <=>  w
C                                         i-1,j
C
C     Restrictions:   M <= 0, N >= max{1,M}, B > 0, 0 <= IORDER <= M+1.
C
C     All FP variables (B,NODES,WORK,OMEGA) are double precision.
C***********************************************************************
        SUBROUTINE POMEGA (M,N,B,IORDER,NODES,OMEGA,NDIM,WORK)
        DOUBLE PRECISION B,NODES(1),OMEGA(NDIM,1),WORK(1)
        INTEGER WDIM,WDIM2
C.......All we do here is allocate working space for POMEG1.
        WDIM = MAX (N+1,IORDER+1)
        WDIM2 = WDIM**2
        CALL POMEG1 (M,N,B,IORDER,NODES,OMEGA,NDIM,
     *              WORK,WORK(WDIM2+1),WORK(2*WDIM2+1),WDIM)
        RETURN
        END


C***********************************************************************
C     POMEG1
C     Same as POMEGA, except three work arrays are given,
C     each of size WDIM * WDIM, where WDIM >= max{N,IORDER} + 1.
C***********************************************************************
        SUBROUTINE POMEG1 (M,N,B,IORDER,NODES,OMEGA,NDIM,
     *                     WORK1,WORK2,WORK3,WDIM)
        INTEGER WDIM
        DOUBLE PRECISION B,NODES(1),OMEGA(NDIM,1),
     *                   WORK1(WDIM,WDIM),WORK2(WDIM,WDIM),
     *                   WORK3(WDIM,WDIM),SUM,BINOML

C.......Calculate numerator coefficients
        CALL RCOEFF (M,N,B,WORK1)
C.......Calculate denominator coefficients
        DO 100 I = 1,N+1
            WORK1(I,2) = BINOML(N,I-1) * B**(I-1)
 100    CONTINUE
C.......Get numerator coefficients of moments
        CALL MCOEFF (IORDER,WORK1,M,WORK1(1,2),N,WORK2,WDIM)

C.......Multiply by the inverse Vandermonde
C.......to get the numerator coefficients of the weight functions.
        CALL VDMINV (IORDER+1,NODES,WDIM,WORK1)
        DO 400 I = 1,IORDER+1
            DO 300 L = 1,N
                SUM = 0.D0
                DO 200 J = 1,IORDER+1
```

```
                         SUM = SUM + WORK1(I,J) * WORK2(L,J)
 200                CONTINUE
                    WORK3(L,I) = SUM
 300           CONTINUE
 400      CONTINUE

C.......  Calculate partial fractions of weights
          DO 700 I = 1,IORDER+1
               CALL PARTF (N-1,N,B,WORK3(1,I),WORK1(1,I))
 700      CONTINUE

C.......  Place solution in proper format in OMEGA
          DO 800 I = 1,IORDER+1
          DO 850 J = 1,N
               OMEGA(I,J) = WORK1(J+1,I)
 850      CONTINUE
 800      CONTINUE

          RETURN
          END


C***********************************************************************
C    RCOEFF
C                                                   M
C    Calculate the numerator coefficients of       R  (Z)    given by
C                                                   N,B
C                    max{N,i}
C                      ---                 j     (i-j)
C        C(i+1)  =     >   BINOML(N,j) B   (-1)      / (i-j)!,   i = 0,...,M
C                      ---
C                      j=0
C
C    Restrictions:  M >= 0,  N >= 1,  B > 0.
C***********************************************************************
          SUBROUTINE RCOEFF (M,N,B,C)
          DOUBLE PRECISION B,C(1),FACTRL,BINOML
          C(1) = 1.D0
          IF (M.EQ.0) RETURN
          DO 200 I = 1,M
               C(I+1) = (-1.D0)**I / FACTRL(I)
               NI = MIN (N,I)
               DO 100 J = 1,NI
                    C(I+1) = C(I+1) + BINOML(N,J) * B**J * (-1)**(I-J)
     *                                 / FACTRL(I-J)
 100           CONTINUE
 200      CONTINUE
          RETURN
          END


C***********************************************************************
C    Calculate FACTRL = N!
C***********************************************************************
          DOUBLE PRECISION FUNCTION FACTRL (N)
```

```
      FACTRL = 1.D0
      IF (N.LT.1) RETURN
      DO 100 I = 1,N
          FACTRL = I * FACTRL
 100  CONTINUE
      RETURN
      END


C***********************************************************************
C    Calculate BINOML = N! / (M! * (N-M)!)
C***********************************************************************
      DOUBLE PRECISION FUNCTION BINOML (N,M)
      DOUBLE PRECISION FACTRL
      BINOML = FACTRL(N) / (FACTRL(M) * FACTRL(N-M))
      RETURN
      END


C***********************************************************************
C    MCOEFF
C
C    We are given the 'IORDER' order rational approximation to exp(-z)
C
C                NUMC(1) + ... + NUMC(M+1) z**M
C        R(z) = ------------------------------,    M <= N.
C                DENC(1) + ... + DENC(N+1) z**N
C
C    Calculate the values C(j,i), j=1,...,N, i=1,...,IORDER+1, such that
C
C                  C(1,i) + ... + C(N,i) z**(N-1)
C        M  (z) = ------------------------------
C         i-1      DENC(1) + ... + DENC(N+1) z**N
C
C    is the (i-1)'th moment function defined by (1.4.3a,b).
C
C    The row dimension of C is assumed to ROWDIM.
C***********************************************************************
      SUBROUTINE MCOEFF (IORDER,NUMC,M,DENC,N,C,ROWDIM)
      INTEGER IORDER,M,N,ROWDIM,I,J
      DOUBLE PRECISION NUMC(1),DENC(1),C(ROWDIM,1)
C.......  Case i = 0
      DO 100 J = 1,N
          C(J,1) = DENC(J+1)
          IF (M.GE.J) C(J,1) = C(J,1) - NUMC(J+1)
 100  CONTINUE
      IF (IORDER.LT.1) RETURN
C.......  Case i = 1,...,IORDER
      DO 300 I = 1,IORDER
          DO 200 J = 1,(N-1)
              C(J,I+1) = DENC(J+1) - I * C(J+1,I)
 200      CONTINUE
          C(N,I+1) = DENC(N+1)
 300  CONTINUE
      RETURN
      END

C***********************************************************************
C     PARTF
C
C     We are given the rational
C
C                        C(1) + ... + C(M+1) z**M
C                        ------------------------,   M <= N.
C                             (1 + B z) ** N
C
C     Calculate the coefficents PFC(i), i=1,...,N+1 such that the above
C     rational has the partial fraction decomposition
C
C                    PFC(1) +  PFC(2)    + ... +    PFC(N+1)
C                              ---------            -------------
C                              (1 + B z)            (1 + B z) ** N
C
C     Note that if M < N, then PFC(i), i=1,...,N-M, will be zero.
C***********************************************************************
      SUBROUTINE PARTF (M,N,B,C,PFC)
      DOUBLE PRECISION B,C(1),PFC(1),BINOML
      DO 100 I = 1,N+1
           PFC(I) = 0.D0
 100  CONTINUE
      DO 300 I = 1,N+1
           NI1 = N - I + 1
           IF (NI1.LE.M) PFC(I) = C(NI1+1) / B**NI1
           IF (I.EQ.1) GO TO 300
           DO 200 J = 1,I-1
                PFC(I) = PFC(I) - PFC(J) * BINOML (N+1-J,NI1)
 200       CONTINUE
 300  CONTINUE
      RETURN
      END


C***********************************************************************
C     VDMINV
C
C     Calculate the NxN inverse Vandermonde matrix A given the N nodes
C     NODES(1),....,NODES(N).
C
C     A(i,j)  =  C(i,j)  /  FP (NODES(i))
C                             i
C
C     where C(i,j) is the (j-i)'th coefficient of FP (x), and FP (x) is the
C     fundemental polynomial given by:              i            i
C
C     (x-NODES(1)) ... (x-NODES(i-1)) (x-NODES(i+1)) ... (x-NODES(N))
C***********************************************************************
      SUBROUTINE VDMINV (N,NODES,ROWDIM,A)
      INTEGER N,ROWDIM,ROW,COEFF
      DOUBLE PRECISION NODES(1),A(ROWDIM,1),FPDIV
```

```
C.......  Initialize first coefficient of each row
          DO 100 ROW = 1,N
              A(ROW,1) = 1.D0
 100      CONTINUE
C.......  If N=1 that's it!
          IF (N.EQ.1) RETURN

C.......  Calculate C(i,k), i=1,...,N, k=1,...,N.
          DO 500 ROW = 1,N
              DO 400 J = 1,N-1
                    IF (J.LT.ROW) INODE = J
                    IF (J.GE.ROW) INODE = J + 1
C.................. Multiply coefficent array by the factor (x - NODES(INODE))
C.................. Coefficients are processed in reverse order.
                    A(ROW,J+1) = 1.D0
                    DO 300 K = 1,J
                         COEFF = J - K + 1
                         A(ROW,COEFF) = -NODES(INODE) * A(ROW,COEFF)
                         IF (COEFF.NE.1) A(ROW,COEFF) = A(ROW,COEFF) +
     *                                                  A(ROW,COEFF-1)
 300                CONTINUE
 400           CONTINUE
 500      CONTINUE

C.......  Divide by FPi (NODES(i))
          DO 900 ROW = 1,N
              FPDIV = 1.D0
              DO 700 J = 1,N-1
                    IF (J.LT.ROW) INODE = J
                    IF (J.GE.ROW) INODE = J + 1
                    FPDIV = FPDIV * (NODES(ROW)-NODES(INODE))
 700           CONTINUE
               DO 800 COEFF = 1,N
                    A(ROW,COEFF) = A(ROW,COEFF) / FPDIV
 800           CONTINUE
 900      CONTINUE

          RETURN
          END
```

Bibliography

[1] Bird, R. B., Stewart, W. E., and Lightfoot, E. N., *Transport Phenomena,* John Wiley, New York, 1960.

[2] Carslaw, H. S., and Jaeger, J. C., *Conduction of Heat in Solids,* 2nd Edition, Oxford University Press, 1959.

[3] Cash, J. R., *On the design of high order exponentially fitted formulae for the numerical integration of stiff systems,* Numer. Math., 36 (1981), pp.253-266.

[4] Cody, W., Meinardus, G., and Varga, R. S., *Chebyshev rational approximations to e^{-x} in $[0,\infty)$ and applications to heat-conduction problems,* J. Approx. Theory, 2 (1969), pp.50-65.

[5] Ehle, B. L., *A-stable methods and Padé approximations to the exponential,* SIAM J. Math. Anal., 4 (1973), pp.671-680.

[6] ___________, *Some results on exponential approximation and stiff equations,* University of Victoria Res. Rep. No. 77, Victoria, B.C., 1974.

[7] Ehle, B. L., and Picel, Z., *Two parameter, arbitrary order, exponential approximations for stiff equations,* Math. Comp., 29 (1975), pp.501-511.

[8] Finlayson, B. A., *Solution of stiff equations resulting from partial differential equations,* Proc. Int'l Conf. on Stiff Computation, Park City, Utah, 1982.

[9] George, J. A., and Liu, J. W. H., *Computer Solution of Large Sparse Positive Definite Systems,* Prentice-Hall, Englewood Cliffs, New Jersey, 1981.

[10] George, J. A., Liu, J. W. H., and Ng, E., *User's guide for SPARSPAK: Waterloo sparse linear equations package,* Res. Rep. CS-78-30, University of Waterloo, 1980.

[11] Gourlay, A. R., and Morris, J. Ll., *The extrapolation of first order methods for parabolic partial differential equations II,* SIAM J. Numer. Anal., 17 (1980), pp.641-655.

[12] Hageman, L. A., and Young, D. M., *Applied Iterative Methods,* Academic Press, New York, 1981.

[13] Hamming, R. W., *Numerical Methods for Scientists and Engineers,* McGraw-Hill, New York, 1962.

[14] Iserles, A., *On the generalized Padé approximations to the exponential function,* SIAM J. Numer. Anal., 16 (1979), pp.631-636.

[15] __________, *Generalized order star theory,* Padé Approximation and its Applications - Amsterdam 1980 (ed. M. G. de Bruin and H. van Rossum), LNiM 888, Springer-Verlag, Berlin, 1981, pp.228-238.

[16] Lau, T. C., *A Class of Approximations to the Exponential Function for the Numerical Solution of Stiff Differential Equations,* Ph.D. Thesis, University of Waterloo, 1974.

[17] Lawson, J. D., *Some numerical methods for stiff ordinary and partial differential equations,* Proc. Second Manitoba Conf. on Numerical Math., pp.741-746., 1972.

[18] Lawson, J. D., and Lau, T. C., *Order constrained Chebyshev rational approximation,* Res. Rep. CS-80-10, University of Waterloo, 1980.

[19] Lawson, J. D., and Morris, J. Ll., *The extrapolation of first order methods for parabolic differential equations I,* SIAM J. Numer. Anal., 15 (1978), pp.1212-1224.

[20] Lawson, J. D., Morris, J. Ll., and Wilkes, M. V., *On the exact solution of linear constant coefficient parabolic partial differential equations using (1,1) Padé approximations,* Utilitas Mathematica, 16 (1979), pp.189-196.

[21] Lawson, J. D., and Swayne, D. A., *High-order near best uniform approximations to the solution of the heat conduction problem,* Proc. IFIP Congress 80, 1980, pp.741-746.

[22] __________, *Compounding of rational approximations with applications to heat conduction problems,* Proc. Thirteenth Manitoba Conf. on Numerical Math. (submitted for publication).

[23] Lebedev, N. N., *Special Functions and their Applications,* Prentice-Hall, Englewood Cliffs, New Jersey, 1965.

[24] Liniger, W., and Willoughby, R. A., *Efficient integration methods for stiff systems of ordinary differential equations,* SIAM J. Numer. Anal., 7 (1970), pp.47-66.

[25] Mitchell, A. R., and Wait, R., *The Finite Element Method in Partial Differential Equations,* John Wiley, New York, 1977.

[26] Morris, J. Ll., *On the numerical solution of a heat equation associated with a thermal print head,* J. of Comp. Phys., 5 (1970), pp.208-228.

[27] __________, *On the numerical solution of a heat equation associated with a thermal print head II,* J. of Comp. Phys., 7 (1971), pp.102-119.

[28] Muskat, M., *The Flow of Homogeneous Fluid Through Porous Media,* McGraw-Hill, New York, 1937.

[29] Norsett, S. P., *Multiple Padé-approximations to the exponential function,* Report No. 4/74, NTH, Trondheim, Norway, 1974.

[30] __________, *One-step methods of Hermite type for numerical integration of stiff systems,* BIT, 14 (1974), pp.63-77.

[31] __________, *Restricted Padé approximations to the exponential function,* SIAM J. Numer. Anal., 15 (1978), pp.1008-1029.

[31a] __________, *An A-stable modification of the Adams-Bashforth methods,* Conf. on the Numerical Solution of Stiff Differential Equations (ed. A. Dold and B. Eckmann), LNiM 109, Springer-Verlag, Berlin, 1969, pp.214-219.

[32] Norsett, S. P., and Wolfbrandt, A., *Attainable order of rational approximations to the exponential function with only real poles,* BIT, 17 (1977), pp.200-208.

[33] Seward, W. L., Fairweather, G., and Johnston, R. L., *A survey of higher order methods for the numerical integration of semidiscrete parabolic problems* (to appear).

[34] Shampine, L. F., and Gear, G. W., *A user's view of solving stiff differential equations,* SIAM Review, 21 (1979), pp.1-17.

[35] Shampine, L. F., and Gordon, M. K., *Typical problems for stiff differential equations,* SIGNUM Newsletter, 10 (1975), pp.11.

[36] Swayne, D. A., *Computation of Rational Functions with Application to Initial Value Problems,* Ph.D. Thesis, University of Waterloo, 1975.

[37] __________, *Matrix operations with rational approximations,* Proc. Seventh Manitoba Conf. on Numerical Math., 1977, pp.581-589.

[38] Varga, R. S., *On higher order stable implicit methods for solving parabolic partial differential equations,* J. Math. Physics, 40 (1961), pp.220-231.

[39] __________, *Matrix Iterative Analysis,* Prentice-Hall, Englewood Cliffs, New Jersey, 1962.

[40] Wanner, G., Hairer, E., and Norsett, S. P., *Order stars and stability theorems,* BIT, 18 (1978), pp.475-489.