AN ANALYSIS OF "BOUNDARY VALUE TECHNIQUES" FOR PARABOLIC PROBLEMS*

by

Alfred Carasso & Seymour V. Parter

Technical Report #39

August 1968

*Prepared under Contract Number N00014-67-A-1028-0004 at the University of Wisconsin. "This research was supported by the Office of Naval Research. Reproduction in whole or in part is permitted for any purpose of the United States Government."

AN ANALYSIS OF "BOUNDARY VALUE TECHNIQUES" FOR PARABOLIC PROBLEMS

1.1 INTRODUCTION

Consider the one-dimensional "heat equation" in a strip:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \qquad 0 < x < 1, \quad t > 0 \tag{1.1}$$

subject to the Dirichlet conditions

$$\begin{cases} u(x, 0) = f(x), & 0 \le x \le 1, \\ u(0, t) = u(1, t) = 0 & t \ge 0. \end{cases} \tag{1.2}$$

It is well known that, provided $f(x)$ is "smooth", there is a unique solution $u(x, t)$ and

$$|u(x, t)| \le K e^{-\pi^2 t} \quad \text{where } K \text{ is a constant.} \tag{1.3}$$

Let $\Delta x = \frac{1}{M+1}$, M a positive integer, let $\Delta t > 0$, let $v_k^n \equiv v(k\Delta x, n\Delta t)$ and consider the following finite difference approximation of (1.1), (1.2):

$$\begin{cases} \dfrac{v_k^{n+1} - v_k^{n-1}}{2\Delta t} = \dfrac{v_{k+1}^n - 2v_k^n + v_{k-1}^n}{\Delta x^2} & k = 1, \ldots M, \quad n = 1, 2, \ldots \\ \text{with} \quad v_0^n = v_{M+1}^n = 0, & n = 0, 1, 2, \ldots \\ v_k^0 = f(k\Delta x), & k = 0, 1, \ldots M+1 \end{cases} \tag{1.4}$$

Rather than use these equations as a marching procedure, D. Greenspan recently (see [10], [11]) proposed an alternative approach: Choose N

large and solve the system

$$(1.5)\qquad \begin{cases} \dfrac{v_k^{n+1} - v_k^{n-1}}{2\Delta t} = \dfrac{v_{k+1}^n - 2v_k^n + v_{k-1}^n}{\Delta x^2} & k = 1, \ldots, M \\ & n = 1, \ldots, N \\ v_0^n = v_{M+1}^n = 0 & n = 0, 1, 2, \ldots N+1 \\ v_k^0 = f(k\Delta x), \qquad v_k^{N+1} \equiv 0, & k = 0, 1, \ldots M+1 \end{cases}$$

of MN linear equations in MN unknowns. Indeed, Greenspan suggested this method for a general class of parabolic problems, linear and non-linear, and carried out several interesting computational experiments.

The scheme selected by Greenspan is the leap-frog scheme discussed in Richtmyer [15]. When used as a marching procedure with parabolic problems, this scheme leads to an improperly posed numerical problem as data on the line $t = \Delta t$ must be supplied, in addition to the usual data, in order to start the calculation. [This is why it is possible to use it as a boundary value procedure.] However, even if this extra data were exactly known, the scheme would in general be useless as a marching procedure: it is unconditionally unstable and therefore always diverges whenever the solution to the analytic problem contains arbitrarily high frequencies. We will show, however, that as a boundary value procedure, for linear problems with time-independent coefficients, the scheme is unconditionally uniformly convergent, and the rate of convergence is $O(\mu^2)$ as the "mesh size" $\mu \to 0$,

$T \to \infty$, under minimal smoothness of the solution. Indeed, for linear problems with time dependent coefficients and for mildly nonlinear problems, one has uniform convergence at the rate of $O(\Delta t^{3/2})$ as $\Delta t \to 0$, $T \to \infty$, $\Delta x = O(\Delta t)$, and at the rate of $O(\Delta t^2)$ for sufficiently smooth exponentially decaying solutions. These results will appear in a later report (see [4] also).

We also analyze an example with which Greenspan had computational difficulty and which points out a peculiar feature of the boundary value method. We then discuss the convergence of the usual iterative methods for solving the systems of linear equations which occur in this method, observing that, unlike the case of systems of elliptic difference equations, line iterative methods may diverge even if the related point iterative methods converge.

At least one reason why such a method may prove useful in practice, especially if one is computing for large times, is provided by its behavior towards round-off error. For, as observed by Southwell [16], marching procedures tend to accumulate round-off error, whereas "jury" methods do not.

1.2 NOTATION AND DEFINITIONS

Let Δx, Δt be small increments in the variables x, t, and let $T = (N+1)\Delta t$ where N is a positive integer. Let M be a positive integer so that $1 = (M+1)\Delta x$. Introduce a mesh over $R_T \equiv \{(x, t);\ 0 < x < 1,\ 0 < t < T\}$ by means of the lines $x = k\Delta x$, $k = 1, \ldots, M$, $t = n\Delta t$ $n = 1, \ldots, N$.

We will be dealing with functions $v(x, t)$ defined at the mesh points of R_T and we adopt the notation

$$(1.6) \qquad v_k^n \equiv v(k\Delta x,\ n\Delta t)$$

Denote by V^n the M component vector, or M-vector

$$(1.7) \qquad V^n = \begin{bmatrix} v_1^n \\ v_2^n \\ \vdots \\ v_M^n \end{bmatrix}$$

and let V be the "block" vector of MN components

$$(1.8) \qquad V = \begin{bmatrix} V^1 \\ V^2 \\ \vdots \\ V^N \end{bmatrix}$$

Let ξ denote an N-component vector

$$(1.9) \qquad \xi = \begin{bmatrix} \xi^1 \\ \xi^2 \\ \vdots \\ \xi^N \end{bmatrix}$$

We define the following norms and scalar products for complex-valued mesh functions:

For any 2M-vectors, X^n, Y^n let their scalar product be defined by

$$(1.10) \qquad \langle X^n, Y^n \rangle = \Delta x \sum_{k=1}^{M} x_k^n - y_k^n$$

and let the corresponding norm be

$$(1.11) \qquad \langle X^n, X^n \rangle^{\sim} = \Delta x \sum_{k=1}^{M} |x_k^n|^2 = \|X^n\|_2^2$$

For N-vectors ξ, ψ define

$$(1.12) \qquad [\xi, \psi] = \Delta t \sum_{n=1}^{N} \xi^{n-n} \bar{\psi}$$

and

$$(1.13) \qquad \|\xi\|_{2,N}^2 = \Delta t \sum_{n=1}^{N} |\xi^n|^2$$

We will also use the norms:

$$(1.14) \qquad \|X^n\|_\infty = \operatorname*{Max}_{j=1\ldots M} \{|x_j^n|\}$$

$$(1.15) \qquad \|V\|_\infty = \operatorname*{Max}_{n=1\ldots N} \{\|V^n\|_\infty\}$$

$$(1.16) \qquad \|V\|_{2,\infty} = \operatorname*{Max}_{n=1\ldots N} \{\|V^n\|_2\}$$

Given any vector X and square matrix A of appropriate size we define

$$\|A\| = \operatorname*{Sup}_{\|X\|=1} \|AX\|$$

the supremum being taken over all complex vectors.

Given a function $u(x, t)$ we sometimes write $u(t_0)$ to denote the function of x obtained from u when t is fixed at the value t_0. Also $u^n(x)$ stands for $u(x, n\Delta t)$.

2. ABSTRACT INITIAL BOUNDARY PROBLEMS OF PARABOLIC TYPE

Let H be a separable Hilbert space of complex valued functions defined on the open interval $0 < x < 1$ with scalar product (u, v) and corresponding norm $\|u\|_H$. Let $\|u\|_\infty$ be the essential supremum norm for such functions and assume that there exists a constant K such that

$$\|u\|_H \leq K\|u\|_\infty \quad \text{for every } u \in H . \tag{2.1}$$

Let A be a linear operator with domain and range contained in H and let b_0, b_1 be linear boundary operators acting at $x = 0$, $x = 1$ respectively. Consider the eigenvalue problem

$$\begin{cases} Av = \lambda v & 0 < x < 1 \\ b_0 v = b_1 v = 0 \end{cases} \tag{2.2}$$

We assume that the problem (2.2) has a complete set of orthonormal eigenfunctions $\{\phi_k\}$ corresponding to strictly positive eigenvalues $\{\lambda_k\}$ with the property that

$$\sum_k \frac{\|\phi_k\|_\infty}{\lambda_k} < \infty \tag{2.3}$$

Let R be the strip $\{(x, t) \mid 0 < x < 1,\ t > 0\}$ in the (x, t) plane and let f be a real valued function on R such that $f(t) \in H$, as a function of x, for each fixed t.

Let $\chi(x)$ be a real valued function on $[0,1]$ belonging to H and let $\psi_0(t)$, $\psi_1(t)$ be defined and real for $t \geq 0$. Consider the following abstract initial boundary value problem on R, associated with the linear operator A:

Find a real valued function $u(x,t)$ defined on R such that for each fixed t, $u(t) \in$ the domain of A as a function of x, and u is differentiable as a function of t, for each fixed x, and

$$
(2.4) \qquad \begin{cases} \dfrac{\partial u}{\partial t} = -Au + f & 0 < x < 1, \quad t > 0 \\ u(x,0) = \chi(x) & 0 \leq x \leq 1 \\ b_0 u = \psi_0(t) \qquad b_1 u = \psi_1(t) & t > 0 \end{cases} .
$$

We assume that the above problem has a unique solution $u(x,t)$ which reaches a known steady state value $u^*(x)$ as $t \to \infty$, in such a way that $\| u(t) - u^* \|_H \to 0$ as $t \to \infty$, and so we speak of problems of parabolic type. Our main concern in this section is to describe a uniformly convergent semi-discrete finite difference approximation to this abstract problem.

2.1 SEMI-DISCRETE APPROXIMATION TO (2.4)

Let $\Delta t > 0$ be a fixed "small" time increment. Let K_1 be a suitable positive constant. Choose T so that for some positive integer N we have

$$
(2.6) \qquad T = (N+1)\Delta t \quad \text{and} \quad \| u(T) - u^* \|_H \leq K_1 \Delta t^3 .
$$

Consider the following semi-discrete[1] approximation to the analytic problem (2.4):

$$(2.7)\qquad \begin{cases} \dfrac{v^{n+1}(x) - v^{n-1}(x)}{2\Delta t} = -A v^{n}(x) + f^{n}(x), & n = 1 \cdots N \\ v^{0}(x) = \chi(x) \qquad\qquad v^{N+1}(x) = u^{*}(x) & \\ b_0 v^{n} = \psi_0^{n} \qquad\qquad b_1 v^{n} = \psi_1^{n}, & n = 1 \cdots N \end{cases}$$

The system of linear equations (2.7) is an approximation to the analytic problem in the following sense:

If $u(x,t)$ is the solution to (2.4), then u satisfies the equations

$$(2.8)\qquad \begin{cases} \dfrac{\hat{u}^{n+1}(x) - \hat{u}^{n-1}(x)}{2\Delta t} = -A\hat{u}^{n} + f^{n} + \tau^{n} & n = 1 \cdots N \\ u^{n} = \hat{u}^{n} & n = 1 \cdots N \\ \hat{u}_0(x) = \chi(x) \qquad\qquad \hat{u}^{N+1}(x) = u^{*}(x) & \\ b_0\hat{u}^{n} = \psi_0^{n} \qquad\qquad b_1\hat{u}^{n} = \psi_1^{n} & n = 1 \cdots N \end{cases}$$

where $\tau^{n}(x)$ is an error term. For $n = 1 \cdots N-1$, $\tau^{n}(x)$ is the "truncation error" $- \left(\dfrac{\partial u}{\partial t}\right)^{n} + \left[\dfrac{u^{n+1}(x) - u^{n-1}(x)}{2\Delta t}\right]$

---

[1] Semi-discrete approximations, where only the time is discretized have been considered from time to time in the literature: In Varga [17], p. 279, the author notes that such a procedure was used by Hartree and Womersley in 1937 to obtain a numerical solution to the heat equation; in Garabedian [9], p. 493, they are used to prove the existence of a solution to the heat equation and the author remarks on the connection with methods in the abstract theory of semi-groups.

For $n = N$,

$$\tau^N(x) = \frac{u^*(x) - u^{N-1}(x)}{2\Delta t} - \left(\frac{\partial u}{\partial t}\right)^N$$

$$= \frac{u(T) - u^{N-1}(x)}{2\Delta t} - \left(\frac{\partial u}{\partial t}\right)^N + \frac{u^* - u(T)}{2\Delta t}$$

We will assume that u is such that

$$\|\tau^n\|_H \le K_4 \Delta t^2 \qquad n = 1 \cdots N \tag{2.9}$$

$$K_4 = \text{constant}.$$

This condition will be satisfied in particular if $u(x, t)$ has bounded continuous third order time derivatives on R, for in that case Taylor's theorem with remainder shows that

$$\left\| -\left(\frac{\partial u}{\partial t}\right)^n + \left[\frac{u^{n+1}(x) - u^{n-1}(x)}{2\Delta t}\right] \right\|_\infty \le K_2 \Delta t^2 \quad n = 1 \cdots N$$

where K_2 is a bound on u_{ttt} in $\bar{R}$. Also T was chosen so that $\|u(T) - u^*\|_H \le K_1 \Delta t^3$. Hence, since the supremum norm is stronger than the H norm we have (2.9).

Because $\|\tau^n\|_H \to 0$ as $\Delta t \to 0$, we say that (2.7) is <u>consistent</u> with the analytic problem. We rewrite (2.7) as

$$\begin{bmatrix} A & \sigma & & \bigcirc \\ -\sigma & \ddots & \ddots & \\ & \ddots & \ddots & \sigma \\ \bigcirc & & -\sigma & A \end{bmatrix} \begin{bmatrix} v^1(x) \\ \vdots \\ v^N(x) \end{bmatrix} = \begin{bmatrix} \frac{v^0(x)}{2\Delta t} + f^1(x) \\ f^2(x) \\ \vdots \\ f^N(x) - \frac{u^*(x)}{2\Delta t} \end{bmatrix} \tag{2.10}$$

$$0 < x < 1$$

with $b_0 v^n = \psi_0^n \qquad b_1 v^n = \psi_1^n \qquad n = 1 \cdots N$ and where $\sigma = \frac{1}{2\Delta t}$.

We assume that $u^*(x)$ is known a-priori, so that the right hand side of (2.10) is known. Having effectively replaced the problem (2.4) by a coupled system of linear equations for N functions of x we must consider two questions:

a) Does the system (2.10) have a solution? Is it unique?

b) Does the solution of (2.10) converge to that of (2.4) as $\Delta t \to 0$?

If so, in which norm, and at what rate does this convergence take place? We will show the following:

Theorem. The system (2.10) has a unique solution $V(\Delta t) = \begin{bmatrix} v^1(x) \\ \vdots \\ v^N(x) \end{bmatrix}$.

Moreover, if U is the exact solution to (2.4) on the lines $t = n\Delta t$, i.e.,

$$U = \begin{bmatrix} u^1(x) \\ \vdots \\ u^N(x) \end{bmatrix} \quad \text{then,}$$

$$\| V(\Delta t) - U \|_{\infty,\infty} \equiv \sup_{n=1 \cdots N} \| v^n - u^n \|_\infty \leq K_0 \Delta t^2$$

so that $V(\Delta t)$ converges uniformly to U at the rate of $O(\Delta t^2)$ as $\Delta t \to 0$, $T \to \infty$.

We begin our analysis with the following key result.

II.2 The matrix $T_N(\sigma_j)$

Let $T = (N+1)\Delta t$, $\lambda_j > 0$ $\sigma_j = \frac{1}{2\lambda_j \Delta t}$ and consider the following $N \times N$ matrix $T_N(\sigma_j)$:

$$T_N(\sigma_j) = \begin{bmatrix} 1 & \sigma_j & & \bigcirc \\ -\sigma_j & \ddots & \ddots & \\ & \ddots & \ddots & \ddots \\ & & \ddots & \ddots & \sigma_j \\ \bigcirc & & & -\sigma_j & 1 \end{bmatrix}$$

If X is an N-dimensional complex vector we define

$$\| T_N \|_\infty = \sup_{\|X\|_\infty = 1} \| T_N X \|_\infty$$

where $\|X\|_\infty = \max_{n=1 \ldots N} |x^n|$. It is well known that with this definition,

$$\| T_N \|_\infty = \max_{i=1 \ldots N} \sum_{j=1}^{N} |a_{ij}|$$

where a_{ij} is the element in the i^{th} row and j^{th} column of the matrix T_N. We will now prove the following lemma:

Lemma 2.1

$T_N(\sigma_j)$ is always invertible and $\| T_N^{-1} \|_\infty$ remains bounded independently of N, σ_j, as $N \to \infty$, $\Delta t \to 0$, $\lambda_j \to \infty$. In fact, if t_{sr} denotes an element in the s^{th} row, r^{th} column of $T_N^{-1}(\sigma_j)$ we have

(a) $$|t_{sr}| \le \frac{4}{(1+4\sigma_j^2)^{1/2}} \; \frac{e^{-\frac{\lambda_j |(s-r)| \Delta t}{1+2\lambda_j \Delta t}}}{1 - e^{\frac{-2\lambda_j T}{1+2\lambda_j \Delta t}}}$$

(b) $$\sum_{r=1}^{N} |t_{sr}| \le \frac{4(1+2\lambda_j \Delta t)}{(1+\lambda_j \Delta t^2)^{1/2}} \; \frac{\left[2 + \Delta t - e^{-\frac{\lambda_j s \Delta t}{1+2\lambda_j \Delta t}} - e^{\frac{\lambda_j (T - s\Delta t)}{1+2\lambda_j \Delta t}}\right]}{\left[1 - e^{-\frac{2\lambda_j T}{1+2\lambda_j \Delta t}}\right]}$$

Proof: We prove this lemma by explicitly computing $T_N^{-1}(\sigma_j)$.

The determinant Δ_N of T_N satisfies the recurrence relation

$$\Delta_{n+1} = \Delta_n + \sigma^2 \Delta_{n-1} \qquad n = 1, \ldots N-1. \qquad (\sigma = \sigma_j)$$

with

$$\Delta_1 = 1 \quad \text{and} \quad \Delta_0 = 1 .$$

Hence if $\alpha = \frac{1}{2} + \frac{1}{2}\sqrt{1+4\sigma^2}$, $\beta = \frac{1}{2} - \frac{1}{2}\sqrt{1+4\sigma^2}$, are the two roots of $x^2 - x - \sigma^2 = 0$, we see that

$$\Delta_N = \frac{\alpha^{N+1} - \beta^{N+1}}{\alpha - \beta} \quad \text{on using } \Delta_1 = \Delta_0 = 1 .$$

Now the co-factor of the element a_{ij} of T_N is

$$A_{ij} = (-)^{i+j} \Delta_{i-1} \Delta_{N-j} (-\sigma)^{j-i} \qquad \text{if } j > i$$

$$= (-)^{i+j} \Delta_{j-1} \Delta_{N-i} (\sigma)^{i-j} \qquad \text{if } i > j$$

and both formulae hold if $i = j$.

If t_{sr} is the element in the s row r^{th} column of T_N^{-1} we then have

$$|t_{sr}| = \frac{\Delta_{r-1}}{\Delta_N} \sigma^{s-r} \Delta_{N-s} \qquad \text{if} \quad s \geq r$$

$$= \frac{\Delta_{s-1}}{\Delta_N} \Delta_{N-r} \sigma^{r-s} \qquad \text{if} \quad s \leq r$$

Since $\|T_N^{-1}\|_\infty = \underset{s=1\ldots N}{\text{Max}} \sum_{r=1}^{N} |t_{sr}|$, we will first estimate $|t_{sr}|$ by means of these formulae and then proceed to estimate $\Sigma|t_{sr}|$.

Since α, β are roots of $x^2 - x - \sigma^2 = 0$ and $\alpha > 0$ whereas $\beta \leq 0$, we have

$$\beta^2 = \sigma^2 + \beta \leq \sigma^2$$

$$\alpha^2 = \sigma^2 + \alpha > \sigma^2$$

Hence $$\alpha - 1 = |\beta| \leq \sigma < \alpha .$$

Consider first $|t_{sr}|$ for $s > r$. Using the formula for the determinants Δ_k we obtain

$$|t_{sr}| = \frac{1}{\sqrt{1 + 4\sigma^2}} \; (\alpha^r - \beta^r) \frac{\sigma^s}{\sigma^r} \; \frac{\alpha^{N-s+1} - \beta^{N-s+1}}{\alpha^{N+1} - \beta^{N+1}}$$

$$\leq \frac{2\alpha^r}{\sqrt{1 + 4\sigma^2}} \frac{\sigma^s}{\sigma^r} \frac{2\alpha^{N-s+1}}{\alpha^{N+1}(1 - \frac{|\beta|^{N+1}}{\alpha^{N+1}})}$$

$$= \frac{4}{\sqrt{1 + 4\sigma^2}} \; (\frac{\sigma}{\alpha})^{s-r} \; \frac{1}{1 - (\frac{|\beta|}{\alpha})^{N+1}}$$

Since $|\beta| = \alpha - 1$,

$$\left(\frac{|\beta|}{\alpha}\right)^{N+1} = \left(1 - \frac{1}{\alpha}\right)^{N+1} = \left(1 - \frac{2}{1 + \sqrt{1 + 4\sigma^2}}\right)^{N+1}$$

and $\sqrt{1 + 4\sigma^2} \leq 1 + 2\sigma$ since $\sigma \geq 0$.

Hence

$$\left(1 - \frac{2}{1 + \sqrt{1 + 4\sigma^2}}\right)^{N+1} \leq \left(1 - \frac{1}{1 + \sigma}\right)^{N+1} = \left(1 - \frac{\frac{N+1}{1+\sigma}}{N+1}\right)^{N+1}$$

which shows that $\left(\frac{|\beta|}{\alpha}\right)^{N+1} \leq \exp - \frac{(N+1)}{1+\sigma}$ using the well-known fact that for $0 \leq x \leq n$, $\left(1 - \frac{x}{n}\right)^{n} \leq e^{-x}$.

Substituting $\sigma = \sigma_j = \frac{1}{2\lambda_j \Delta t}$, $T = (N + 1)\Delta t$ we obtain

$$\left(\frac{|\beta|}{\alpha}\right)^{N+1} \leq e^{\frac{2\lambda_j T}{1 + 2\lambda_j \Delta t}}$$

Hence,

$$\frac{1}{1 - \left(\frac{|\beta|}{\alpha}\right)^{N+1}} \leq \frac{1}{1 - e^{-\frac{2\lambda_j T}{1 + 2\lambda_j \Delta t}}}$$

Let us now examine $\left(\frac{\sigma}{\alpha}\right)^{s-r}$. We have

$$\left(\frac{\sigma}{\alpha}\right)^{s-r} = \left(1 - \frac{(\alpha - \sigma)}{\alpha}\right)^{s-r} \leq \left(1 - \frac{1}{2\alpha}\right)^{s-r} \quad \text{since } \alpha - \sigma \geq \frac{1}{2}$$

$$\leq \left(1 - \frac{1}{2 + 2\sigma}\right)^{s-r} \quad \text{using } \sqrt{1 + 4\sigma^2} \leq 1 + 2\sigma .$$

Hence by a similar device,

$$\left(\frac{\sigma}{\alpha}\right)^{s-r} \leq \exp \frac{-\lambda_j (s-r)\Delta t}{1 + 2\lambda_j \Delta t}$$

Now if $r \geq s$ all formulae still hold with r and s interchanged. We conclude that

$$|t_{sr}| \leq \frac{4}{\sqrt{1 + 4\sigma_j^2}} \frac{e^{-\frac{\lambda_j |s-r| \Delta t}{1 + 2\lambda_j \Delta t}}}{1 - \exp\left(\frac{-2\lambda_j T}{1 + 2\lambda_j \Delta t}\right)} \qquad (s, r = 1 \ldots N)$$

Let us now extimate $\sum_{r=1}^{N} |t_{sr}|$.

We have

$$\sum_{r=1}^{N} |t_{sr}| \frac{4}{1 - e^{\frac{-2\lambda_j T}{1+2\lambda_j \Delta t}}} \frac{\Delta t}{\sqrt{\Delta t^3 + 4\sigma^2 \Delta t^2}} \sum_{r=1}^{N} e^{-\frac{\lambda_j |s-r| \Delta t}{1 + 2\lambda_j \Delta t}}$$

Let $\rho = \frac{1}{1 + 2\lambda_j \Delta t}$ and consider

$$\Delta t \sum_{r=1}^{N} e^{-\rho \lambda_j |s-r| \Delta t} = \Delta t \sum_{p=0}^{s-1} e^{-\rho \lambda_j p \Delta t} + \Delta t \sum_{p=1}^{N-s} e^{-\rho \lambda_j p \Delta t}$$

We may use a geometric argument (the integral test) to show

$$\Delta t \sum_{p=0}^{s-1} e^{-\rho \lambda_j p \Delta t} = \Delta t \quad \Delta t \sum_{p=1}^{s-1} e^{-\rho \lambda_j p \Delta t} < \Delta t + \int_0^t e^{-\rho \lambda_j u} du \quad t = s\Delta t$$

and similarly

$$\Delta t \sum_{p=1}^{N-s} e^{-\rho\lambda_j p\Delta t} < \int_0^{T-t} e^{-\rho\lambda_j u}\, du \qquad T = (N+1)\Delta t$$

Hence

$$\Delta t \sum_{r=1}^{N} e^{-\rho\lambda_j |s-r|\Delta t} \le \frac{\Delta t + 2 - e^{-\rho\lambda_j t} - e^{-\rho\lambda_j (T-t)}}{\rho\lambda_j}$$

$$\text{i.e.,} \quad \sum_{r=1}^{N} |t_{sr}| \le \frac{(1+2\lambda_j\Delta t}{(1+\lambda_j^2\Delta t^2)^{1/2}} \; \frac{4[\Delta t + 2 - e^{\frac{-\lambda_j s\Delta t}{1+2\lambda_j\Delta t}} - e^{\frac{-\lambda_j (T-s\Delta t)}{1+2\lambda_j\Delta t}}]}{[1 - e^{\frac{-2\lambda_j T}{1+2\lambda_j\Delta t}}]}$$

Notice that as $\lambda_j \to \infty$ $1 - e^{-\frac{2\lambda_j T}{1+2\lambda_j\Delta t}} \to 1 - e^{-\frac{T}{\Delta t}} = 1 - e^{-(N+1)}$.

Clearly, the above sum is bounded as $\lambda_j \to \infty$, $\Delta t \to 0$, or both, and the bound is independent of s. This proves the lemma.

Remark. In a subsequent discussion we will also need the following result: Let $\lambda_1 \sim \beta\Delta t^2$ with β fixed $\neq 0$ as $\Delta t \to 0$. Then $\| T_N^{-1}(\sigma_1) \|_\infty$ remains bounded as $\Delta t \to 0$. We may see this as follows:

$$\text{Since} \quad |t_{sr}| \le \frac{4}{\sqrt{1+\sigma_1^2}} \; \frac{e^{-\frac{\lambda_1 |s-r|\Delta t}{1+2\lambda_1\Delta t}}}{1-e^{-\frac{2\lambda_1 T}{1+2\lambda_1\Delta t}}} \le \frac{4}{\sqrt{1+4\sigma_1^2}\;(1 - e^{\frac{-2\lambda_1 T}{1+2\lambda_1\Delta t}})}$$

we have on substituting $\lambda_1 = \beta\Delta t^2$ in the last expression

$$|t_{sr}| \leq \frac{4\beta\Delta t^3}{[1+\beta^2\Delta t^6]^{1/2}\left[1 - e^{\frac{-2\beta\Delta t^2 T}{1+2\beta\Delta t^3}}\right]} \leq \frac{4\beta\Delta t^3}{1 - e^{\frac{-2\beta\Delta t^2 T}{1+2\beta\Delta t^3}}}$$

and both the numerator and denominator of the last expression approach zero as $\Delta t \to 0$. Differentiating with respect to Δt, using L'Hospital's rule we obtain

$$\lim_{\Delta t \to 0} \frac{4\beta\Delta t^3}{1 - e^{\frac{-2\beta\Delta t^2 T}{1+2\beta\Delta t^3}}} = \lim_{\Delta t \to 0} \frac{(12\beta\Delta t^3)(1+2\beta\Delta t^3)^2}{(4\beta T\Delta t - 4\beta^2 T\Delta t^4)\, e^{\frac{-2\beta\Delta t^2 T}{1+2\beta\Delta t^3}}}$$

and, since the last expression tends to $\frac{3\Delta t}{T} = \frac{3}{N+1}$ as $\Delta t \to 0$, we have

$$\sum_{r=1}^{N} |t_{sr}| \leq (N+1) \max_{r,s} |t_{sr}| \to 3 \text{ as } \Delta t \to 0 .$$

Lemma 2.2

The system of equations (2.10) has a unique solution $V(\Delta t)$.

Proof: Let M be the matrix of linear operators occurring in (2.10). In an obvious notation we may write (2.10) as

$$MV = F, \quad b_0 V = \psi_0 \quad b_1 V = \psi_1 \tag{2.11}$$

Observe that F is such that each of its components belongs to H. Observe also that it is sufficient to prove that given any G whose components $g^n(x)$ belong to H, $n = 1 \ldots N$, the system

$$MV = G \quad b_0V = b_1V = 0$$

always has a unique solution. Indeed, if this is the case then (2.11) has at most one solution, and, furthermore, we may always construct a solution to (2.11) as follows:

Fix a vector X whose components belong to the domain of A and such that $b_0X = \psi_0 \quad b_1X = \psi_1$. Such an X must exist, otherwise there can be no solution to the analytic problem. Since MX belongs to H, we may solve the problem

$$MY = -MX \quad b_0Y = b_1Y = 0$$

Next solve the problem $MZ = F \quad b_0Z = b_1Z = 0$. Letting $V = X + Y + Z$ we can always uniquely solve

$$MV = G \quad b_0V = b_1V = 0.$$

To do this expand in the eigenfunctions of the problem (2.2) above:

Set $v^n(x) = \sum_{j=1}^{\infty} c_j^n \phi_j \quad g^n(x) = \sum_{j=1}^{\infty} d_j^n \phi_j$. Then if $\sigma_j = \frac{1}{2\lambda_j \Delta t}$, we obtain the following equations expressing the c_j^n in terms of the known d_j^n

$$\sigma_j(c_j^{n+1} - c_j^{n+1}) + c_j^n = \frac{d_j^n}{\lambda_j}, \quad n = 1 \ldots N, \quad j = 1, 2, \ldots$$

with

$$c_j^0 = c_j^{N+1} = 0 \quad \forall j.$$

Hence if $T_N(\sigma_j)$ is the matrix of lemma (2.1), we may write

$$(2.12) \qquad [T_N(\sigma_j)] \begin{bmatrix} c_j^1 \\ \vdots \\ c_j^N \end{bmatrix} = \frac{1}{\lambda_j} \begin{bmatrix} d_j^1 \\ \vdots \\ d_j^N \end{bmatrix} \qquad j = 1, 2, \ldots$$

Since $T_N(\sigma_j)$ is invertible for every j, (2.12) uniquely defines the c_j^n so that the reduced problem above always has a unique solution. Q.E.D.

We are now ready to prove the convergence theorem of section II.1.

Let $w^n = v^n - \hat{u}^n$, then $w^n(x)$ satisfies

$$w^0 = w^{N+1} = 0$$

$$\frac{w^{n+1} - w^{n-1}}{2\Delta t} = Aw^n + \tau^n \qquad b_0 w^n = b_1 w^n = 0, \quad n = 1 \ldots N$$

where

$$\tau^n(x) \in H \quad \text{and} \quad \|\tau^n\|_H \le K_4 \Delta t^2$$

Setting $$w^n = \sum_{j=1}^{\infty} c_j^n \phi_j \qquad n = 1 \ldots N$$

and $$\tau^n = \sum_{j=1}^{\infty} d_j^n \phi_j \qquad n = 1 \ldots N$$

we have $|d_j^n| = |(\tau^n, \phi_j)| \le \|\tau^n\|_H \|\phi_j\|_H \le K_4 \Delta t^2$

and

$$[T_n(\sigma_j)] \begin{bmatrix} c_j^1 \\ \vdots \\ c_j^N \end{bmatrix} = \frac{1}{\lambda_j} \begin{bmatrix} d_j^1 \\ \vdots \\ d_j^N \end{bmatrix} \qquad j = 1, 2, \ldots$$

By Lemma (2.1), $\| T_N^{-1} \|_\infty$ is bounded as $\Delta t \to 0$, $N \to \infty$, $\lambda_j \to \infty$.

Hence

$$\max_{n=1 \ldots N} |c_j^n| \leq K_5 \frac{\Delta t^2}{\lambda_j}$$

Therefore,

$$\| w^n \|_\infty \leq \sum_{j=1}^{\infty} |c_j^n| \, \| \phi_j \|_\infty \leq K_5 \, \Delta t^2 \sum_{j=1}^{\infty} \frac{\| \phi_j \|_\infty}{\lambda_j} .$$

Since by assumption $\sum_j \frac{\| \phi_j \|_\infty}{\lambda_j} < \infty$ we have

$$\max_{n=1 \ldots N} \| w^n \|_\infty \leq K_6 \Delta t^2 \qquad K_6 = \text{constant},$$

and this proves the theorem.

Examples of such operators A are provided by regular Sturm-Liouville differential operators, operating in $H = L^2[0,1]$. Thus for the problem*

$$(2.13) \qquad \begin{cases} [a(x)u']' + b(x)u' - c(x)u + \lambda u = 0 & 0 < x < 1 \\ u(0) = u(1) = 0 \end{cases}$$

where $a(x) \geq a_0 > 0$ and $c(x) \geq 0$, it is known that the eigenvalues are real and form a countably infinite set, $\lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \ldots$. Moreover $\lambda_1 > \inf_{0<x<1} c(x)$ [Protter and Weinberger [14].

*A standard transformation, puts (2.13) in self-adjoint form.

It is a standard result that the eigenvalues of (2.13) can be characterized as the zeroes of an entire function (Coddington and Levinson [6]) and as observed by Atkinson in [1] this function is of order at most $1/2$ so that,

$$\sum_k \frac{1}{(\lambda_k)^{1/2+\varepsilon}} < \infty$$

for every $\varepsilon > 0$. Also, the normalized eigenfunctions may be shown to be uniformly bounded in the supremum norm, i.e.,

$$\|\phi_k\|_\infty \leq \text{constant}$$

see (Courant-Hilbert [7]).

We remark, however, that A may be a singular differential operator and still satisfy property (2.3). Thus consider in $L_2[0,1]$, the problem

$$(2.14) \qquad \begin{cases} Au \equiv -[(1-x^2)u']' = \lambda u, & 0 < x < 1 \\ u(0) = 0, & (1-x^2)u'(x) \to 0 \quad \text{as} \quad x \uparrow 1 . \end{cases}$$

If $P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2-1)^n$ are the Legendre polynomials for $n = 0, 1, 2, \ldots$, then the eigenfunctions for this problem are

$$P_{2n+1}, \qquad n = 0, 1, 2, \ldots \text{ corresponding to the eigenvalues}$$

$$\lambda_n = (2n+1)(2n+2) \qquad n = 0, 1, 2, \ldots$$

This follows because the P_n satisfy

$$[-(1-x^2)P_n']' = n(n+1)P_n \qquad n = 0, 1, 2, \ldots$$

and $P_{2n+1}(0) = 0$ for $n = 0, 1, 2, \ldots$. (See Courant-Hilbert [7].)

Also, the set $\{P_{2n+1}\}$ spans $L_2[0, 1]$ since the complete set of Legendre polynomials spans $L_2[-1, 1]$ and $P_n(x)$ is an even function if n is even. As defined above, the P_n are not normalized but satisfy $\|P_n\|_\infty \leq 1$ for $|x| \leq 1$. However, if

$$\tau_n(x) = \sqrt{\frac{2n+1}{4}}\, P_n(x) \qquad n = 0, 1, 2, \ldots$$

then the τ_n are orthonormal on $(0, 1)$ and

$$\|\tau_n\|_\infty \leq \sqrt{\frac{2n+1}{4}}$$

thus

$$\sum_{n \text{ odd}} \frac{\|\tau_n\|_\infty}{\lambda_n} < \infty .$$

Finally, we remark that although we have emphasized one dimensional problems, similar problems may be formulated in R^n with H, for example, being a Sobolev space of functions on some bounded domain Ω and A a uniformly elliptic operator of sufficiently high order with say Dirichlet boundary conditions. Indeed A need not even be a differential operator.

3.1 UNIFORMLY PARABOLIC LINEAR INITIAL BOUNDARY PROBLEMS: FULLY DISCRETE METHODS

We are concerned here with the numerical computation of problems of the following kind:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left[a(x)\frac{\partial u}{\partial x}\right] + b(x)\frac{\partial u}{\partial x} - c(x)u + h(x,t) \qquad 0 < x < 1,\ t > 0$$

$$u(x,0) = \chi(x) \qquad 0 \le x \le 1$$

(3.1)

$$u(0,t) = \phi_1(t), \qquad u(1,t) = \phi_2(t)$$

$$\chi(0) = \phi_1(0) \qquad \chi(1) = \phi_2(0)$$

We assume that $a(x) \ge a_0 > 0$ and $c(x) \ge 0$; a having three or more continuous derivatives, and b one or more continuous dirivatives in R. We assume further that a, b, c, h, χ, ϕ_1, ϕ_2 are bounded and sufficiently smooth that a solution $u(x,t)$ exists having three continuous time derivatives and four continuous space derivatives. All of the above mentioned derivatives as well as u itself will be assumed bounded on R. [For existence, uniqueness and regularity theorems for parabolic equations, consult Friedman [8]). As in the previous section, the amount of smoothness that we assume will suffice for the truncation error τ^n to be of the order of $\Delta t^2 + \Delta x^2$ in the discrete L_2 norm and this is all we need. Thus our assumptions may be weakened somewhat. We assume that h, ϕ_1, ϕ_2 reach a steady state as $t \to \infty$. Since $c(x) \ge 0$, $u(x,t)$ converges to a steady state value $u^*(x)$ (see Friedman [8]), and we assume this convergence to be uniform in x. We may suppose that $u^*(x)$ is known without loss of generality. Indeed, we only require its values at mesh points, and these can be obtained with sufficient accuracy by existing numerical techniques, since $u^*(x)$ satisfies

an inhomogeneous boundary value problem for an ordinary differential equation. Finally, we assume that given any $\varepsilon > 0$ it is possible to estimate how large T must be chosen so that

$$\| u(T) - u^* \|_\infty < \varepsilon$$

e.g. by means of asymptotic formulae.

3.2 DISCRETE APPROXIMATION TO THE ANALYTIC PROBLEM

Choose T so that for some positive integer N we have $T = (N+1)\Delta t$ and

$$\| u(T) - u^* \|_2 \equiv \{\Delta x \sum_{k=1}^{n} | u(k\Delta x, T) - u(k\Delta x) |^2\}^{1/2} \leq K\Delta t^3$$

where K is a fixed positive constant independent of Δt. Introduct a mesh over R_T as in section 1.2.

Our finite-difference approximation to (3.1) will be

$$\frac{v_k^{n+1} - v_k^{n-1}}{2\Delta t} = \frac{a_{k+\frac{1}{2}}(v_{k+1}^n - v_k^n) - a_{k-\frac{1}{2}}(v_k^n - v_{k-1}^n)}{\Delta x^2} + \frac{b_k(v_{k+1}^n - v_{k-1}^n)}{2\Delta x} - c_k v_k^n + h_k^n \qquad n = 1 \ldots N, \quad k = 1 \ldots M$$

(3.2)

$$\text{with} \quad v_k^0 = \chi(K\Delta x) \qquad v_k^{N+1} = u^*(K\Delta x) \quad k = 0, 1, \ldots M+1$$

$$v_0^n = \phi_1(n\Delta t) \qquad v_{n+1}^n = \phi_2(n\Delta t) \quad n = 1, \ldots N.$$

As before, this approximation is consistent with the problem (3.1), i.e., the exact solution u satisfies (3.2) if we add an error term τ_k^n on the right hand side. For $n = 1 \ldots N-1$, τ_k^n is the truncation error due to replacing derivatives by finite difference quotients. Taylor's theorem shows that τ_k^n is a linear combination of $\Delta t^2 u_{ttt}$, $\Delta x^2 a_{xxx}$ and $\Delta x^2 u_{xxxx}$ these derivatives being evaluated at points of the (x, t) plane intermediate to successive meshpoints. For $n = N$, there is an additional error due to prescribing $u^*(x)$ on the line $t = T$ instead of the exact solution $u(T)$. Since we assume the above derivatives to be bounded on $\bar{R}$ and we chose T so that $\| u(T) - u^* \|_2 \leq K\Delta t^3$, there results the estimate

$$\| \tau^n \|_2 \equiv \{\Delta x \sum_{k=1}^{M} |\tau_k^n|^2\}^{1/2} \leq K_0(\Delta x^2 + \Delta t^2)$$

where K_0 is a fixed constant independent of Δx and Δt, and n.

Let $\alpha_k = [a_{k+\frac{1}{2}} + a_{k-\frac{1}{2}}] + c_k \Delta x^2$, $\beta_k = -[a_{k+\frac{1}{2}} + \frac{b_k \Delta x}{2}]$ and $\gamma_k = [\frac{b_k \Delta x}{2} - a_{k-\frac{1}{2}}]$ $k = 1 \ldots M$ and define the tridiagonal matrix L of order M by

$$L = \frac{1}{\Delta x^2} \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \gamma_2 & \cdot & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \beta_{M-1} \\ & & & \gamma_M & \alpha_M \end{bmatrix}$$

We may write the approximation (3.2) in the form

$$(3.3) \qquad \frac{V^{n+1} - V^{n-1}}{2\Delta t} + LV^n = F^n \qquad n = 1 \ldots N$$

where V^0, V^{N+1} are given and F^n is an M-vector containing the known lateral boundary data and the inhomogeneous term h_k^n. We may also write (3.3) in "block" form. Let M be the $MN \times MN$ block tridiagonal matrix

$$M = \begin{bmatrix} L & \sigma I & & \\ -\sigma I & \ddots & \ddots & \\ & \ddots & \ddots & \sigma I \\ & & -\sigma I & L \end{bmatrix}$$

where I is the $M \times M$ unit matrix and $\sigma = \frac{1}{2\Delta t}$.

Then we have

$$(3.4) \qquad MV = F$$

where F consists now of data at the four boundaries as well as the inhomogeneous term $h(x,t)$.

Lemma 3.1

There exists a nonsingular diagonal matrix D such that $D^{-1}LD = \hat{L}$ is a symmetric matrix and $\|D\|_\infty$, $\|D^{-1}\|_\infty \le K_0 < \infty$ as $M \to \infty$, $\Delta x \to 0$, $(M+1)\Delta x = 1$.

Proof: See ([4], [5]) .

Remark. The change of variables $X = DZ$ is the discrete analog of the transformation

$$u(x) = e^{-\frac{1}{2}\int_0^x \frac{b(t)}{a(t)}\,dt} v(x)$$

which changes the linear differential operator

$$\mathcal{L}[u] \equiv -(au')' - bu' + cu \qquad 0 < x < 1$$

$$u(0) = u(1) = 0$$

into the formally self-adjoint operator

$$\hat{\mathcal{L}}[v] \equiv -(av')' + [c + \frac{1}{2}b' + \frac{1}{4}\frac{b^2}{a}]v \quad 0 < x < 1$$

$$v(0) = v(1) = 0$$

Lemma 3.2

The eigenvalues of L are strictly positive and remain bounded away from zero as $M \to \infty$, $\Delta x \to 0$, $(M + 1)\Delta x = 1$. Let $\{\lambda_j\}_{j=1}^{M}$ be the eigenvalues of L arranged in increasing order; then there exists a positive integer j_0, independent of M as $M \to \infty$, such that for all $j_0 < j \leq M$, we have

$$c_1 j^2 \leq \lambda_j \leq c_2 j^2 \quad \text{where } c_1, c_2$$

are positive constants.

Finally, let V^j be the eigenvector of $\tilde{L}$ corresponding to the eigenvalue λ_j and normalized so that

$$\Delta x \sum_{K=1}^{M} |v_k^j|^2 = 1 .$$

Then, there exists a positive constant K_0 and a positive integer j_1, independent of M, such that for all $j_1 < j \leq M$,

$$\|V^j\|_\infty = \sup_{K=1\ldots M} |v_k^j| \leq K_0 (j)^{1/2}$$

Proof:

In the self-adjoint case (i.e. $b(x) \equiv 0$) these results are to be found in Bückner [3]. In this more general case, the lemma follows from the discrete maximum principle, and from lemma 3.1 together with Bückner's argument. A different proof may be found in [4] and [5].

Theorem

Let $\mu^2 = (\Delta t^2 + \Delta x^2)$ and let $\{V^n\}_{n=1}^N$ be the solutions of equations (3.3), or equivalently of (3.4). Let $\{U^n\}_{n=1}^N$ be the vector obtained from evaluations of $u(x,t)$ at the mesh points. Finally, assume

$$\|\tau^n\|_2 \leq K_1 \mu^2 . \tag{3.5}$$

Then, there is a constant K_2 such that

$$\|V^n - U^n\|_\infty \leq K_2 \mu^2$$

Proof:

The argument is very similar to the proof of the main theorem of the preceding section. Let $W = V - U$. With D the diagonal matrix which

symmetrizes L, let $X^n = D^{-1} W^n$ and substitute into (3.3), to obtain

$$(3.6) \qquad \frac{X^{n+1} - X^{n-1}}{2\Delta t} + \tilde{L} X^n = D^{-1} \tau^n \quad n = 1, \ldots N, \; X^0 = X^{N+1} = 0 .$$

Since $\tilde{L}$ is real symmetric, it has a complete set of orthonormal eigenvectors Z^j $j = 1, \ldots M$. We may solve by expanding interms of the Z^j. Thus if $X^n = \sum_{j=1}^{M} c_j^n Z^j$, $D^{-1} \tau^n = \sum_{j=1}^{M} d_j^n Z^j$, we obtain on substituting into (3.6), MN equations for the coefficients c_j^n:

$$(3.7) \qquad \frac{c_j^{n+1} - c_j^{n-1}}{2\Delta t} + \lambda_j c_j^n = d_j^n \quad n = 1, \ldots N, \quad j = 1, \ldots M$$

where $c_j^0 = c_j^{N+1} = 0$, $j = 1, \ldots M$.

Let $\sigma_j = \frac{1}{2\lambda_j \Delta t}$; then

$$(3.8) \qquad [T_N(\sigma)] \begin{bmatrix} c_j^1 \\ \vdots \\ c_j^N \end{bmatrix} = \frac{1}{\lambda_j} \begin{bmatrix} d_j^1 \\ \vdots \\ d_j^N \end{bmatrix} \qquad j = 1, \ldots M$$

where $T_N(\sigma_j)$ is the $N \times N$ matrix of lemma (2.1). The proof now follows from lemma 3.2 and the fact that

$$\sum_{j=1}^{\infty} (j)^{-3/2} < \infty$$

Remarks on the round-off error

The main reason for considering finite difference approximations is that they lead to systems of equations which can be solved using a high speed computer. However, rounding-off errors must inevitably be introduced in any automatic computation. In our proof of convergence we have assumed that all computations are carried out with infinite precision in solving the difference equations. Let us consider the effect of such errors: Suppose that ρ^n is the M-vector consisting of the round-off error $\{\rho_k^n\}$, $k = 1 \ldots M$ at time $t = n\Delta t$. We must then replace τ^n by $\tau^n + \rho^n$ in equation (3.6) and d_j^n by $d_j^n + r_j^n$ in (3.7), where $D^{-1}\rho^n = \sum_{j=1}^{M} r_j^n Z^{(j)}$.

This leads to the estimate

$$\max_{n=1\ldots N} |c_j^n| \le \|T_N^{-1}(\sigma_j)\|_\infty \max_{n=1\ldots N} \frac{|d_j^n| + |r_j^n|}{\lambda_j}$$

and hence

$$\|X^n\|_\infty \le K(\mu^2 + \|\rho^n\|_2) \sum_{j=1}^{M} \frac{\|Z^{(j)}\|_\infty}{\lambda_j}$$

If W^n is the error between the solution to the analytic problem, and the solution of the difference equations obtained from the computer, we have

$$\|W\|_\infty \le K_0\mu^2 + K_1 \sup_{n=1\ldots N} \|\rho^n\|_2 . \tag{3.9}$$

It is instructive to compare the above estimates with the error estimate for a general class of marching procedures in parabolic problems

which we may represent as

$$AV^{n+1} = BV^n + F^n \qquad 0 \le n\Delta t \le T$$

where A, B are $M \times M$ matrices, V^n is an M vector, the solutions of the difference equations at time $t = n\Delta t$, F^n contains the inhomogeneous term and lateral boundary data and

$$\|A^{-1}\|_2 \le K_0 \qquad \|(A^{-1}B)^m\|_2 \le K_1$$

for all $0 < m\Delta t \le T$, so that the method is stable. (For example, the Crank-Nicolson scheme leads to a procedure of this form.) The error equation for such a method has the form

$$AW^{n+1} = BW^n + \tau^n + \rho^n \tag{3.10}$$

where τ^n is the truncation error, which we assume is of order $\Delta t^2 + \Delta x^2$, and ρ^n is the round-off error. Hence

$$W^{n+1} = (A^{-1}B)^{n+1} W^0 + \sum_{k=0}^{n} (A^{-1}B)^k A^{-1} \tau^{n-k} + \sum_{k=0}^{n} (A^{-1}B)^k A^{-1} \rho^{n-k} . \tag{3.11}$$

Since $W^0 = 0$ we have the estimate

$$\|W^{n+1}\|_2 \le K_0 K_1 \left\{ \sum_{k=0}^{n} \|\tau^{n-k}\|_2 + \sum_{k=0}^{n} \|\rho^{n-k}\|_2 \right\} \tag{3.12}$$

and if $\Delta x = O(\Delta t)$ so that $\|\tau^j\|_2 \le K_2 \Delta t^2$ we would then have

$$\|W^{n+1}\|_2 \le (K_0 K_1 K_2 T)\Delta t + (n+1) K_0 K_1 \sup_{k=0,\,1\ldots n} \|\rho^k\|_2 \tag{3.13}$$

The factor (n+1) in the second term on the right hand side of (3.13) illustrates a common property of marching procedures: They tend to accumulate round-off error. No such factor is present in the estimate (3.9) for the boundary value method and this is an advantage if one is computing for large times.

4.1 AN EXAMPLE

Consider now the problem

$$
\begin{aligned}
&u_t = u_{xx} + \pi^2 u + \sin \pi x \cos t \qquad 0 < x < 1, \quad t > 0 \\
&u(x,0) = 0 \quad 0 \le x \le 1, \quad u(0,t) = u(1,t) = 0 \quad t \ge 0
\end{aligned} \tag{4.1}
$$

This problem has the unique solution $u = \sin \pi x \sin t$. It differs from the class of problems considered in the previous sections in that $c(x)$ is negative and the related Sturm-Liouville problem has the eigenvalue $\lambda = 0$. Nevertheless, since $u \equiv 0$ at $t = \pi$ and at $t = 2\pi$ we may select <u>either</u> of these lines as the line $t = T$ and prescribe the <u>exact</u> solution $u \equiv 0$ on $t = T$ in our difference approximation to (4.2). Thus, if H is the tridiagonal matrix of order M given by

$$
H = \frac{1}{\Delta x^2} \begin{bmatrix} 2 & -1 & & & & \\ -1 & \ddots & \ddots & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & -1 \\ & & & & -1 & 2 \end{bmatrix} \tag{4.2}
$$

with eigenvalues $0 < \mu_1 < \mu_2 < \cdots < \mu_M$ and if W^1 is the M-vector $w_k^1 = \sin k\pi\Delta x \quad k = 1 \ldots M$, our approximation may be written as

$$\begin{cases} \dfrac{V^{n+1} - V^{n-1}}{2\Delta t} + (H - \pi^2 I)V^n = W^1 \cos n\Delta t \qquad n = 1 \ldots N \\ V^0 = V^{N+1} = 0 \end{cases} \tag{4.3}$$

On expanding in eigenvectors of H, we easily see that (4.3) has the unique solution

$$V^n = c^n W^1$$

where the c^n's satisfy

$$\frac{c^{n+1} - c^{n-1}}{2\Delta t} + (\mu_1 - \pi^2)\, c^n = \cos n\Delta t \qquad n = 1 \ldots N$$

$$c^0 = c^{N+1} = 0$$

The computation of this example was attempted by Greenspan in [10] with $T = 2\pi$. However, he was not able to solve the system of difference equations by point successive over-relaxation for any value of ω. Apart from that, the above example has another interesting property: It makes a difference whether one selects $T = \pi$ or $T = 2\pi$. With $T = \pi$, the unique aolution V of the system (4.3) (even though it remains uniformly bounded as $\Delta x, \ \Delta t \to 0$, with $\Delta x = O(\Delta t)$,) does not converge to the analytic solution U, unless $N \to \infty$ through even integers.

Lemma 4.1

Let S be the skew-symmetric $N \times N$ matrix

$$S = \frac{1}{2\Delta t} \begin{bmatrix} 0 & 1 & & & \\ -1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & -1 & 0 \end{bmatrix}$$

then S has the distinct eigenvalues $\lambda_s = \frac{1}{\Delta t} \cos \frac{s\pi}{N+1}$, $s = 1 \ldots N$ with corresponding eigenvectors

$$\xi_s = \{\xi_s^n\} \qquad s = 1 \ldots N \text{ where}$$

$$\xi_s^n = i^n \sin \frac{ns\pi}{N+1} \qquad n = 1 \ldots N .$$

If N is odd, $\lambda_{(\frac{N+1}{2})}$ is the only zero eigenvalue and the corresponding eigenvector may be taken to be

$$\psi_{\frac{N+1}{2}} = \delta \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \text{ where } \delta \text{ is chosen so that } \left\| \psi_{\frac{N+1}{2}} \right\|_{2,N} = 1 .$$

Proof. Direct verification.

Lemma 4.2

Let $V = \{V^n\}_{n=1}^N$ be the solution of (4.3), and let $\Delta t = \gamma \Delta x$, $\gamma =$ constant as $\Delta x, \Delta t \to 0$, then $\| V(\Delta t) \|_\infty$ remains bounded as $\Delta t \to 0$.

Proof. Let U^n be the analytic solution at the mesh points of the line $t = n\Delta t$, and let $E^n = U^n - V^n$. Then E^n satisfies

$$(4.4)\qquad \begin{cases} \dfrac{E^{n+1} - E^{n-1}}{2\Delta t} + (H - \pi^2 I)E^n = \tau^n, & n = 1 \ldots N \\ E^0 = E^{N+1} = 0 \end{cases}$$

where τ^n is the truncation error and $\|\tau^n\|_2 = O(\Delta t^2)$.

Expanding in the orthonormal eigenvectors of H, we are led to a system of MN linear equations for the Fourier coefficients c_j^n of E^n, in terms of the coefficients d_j^n of τ^n viz

$$(4.5)\qquad [T_N(\sigma_j)] \begin{bmatrix} c_j^1 \\ \vdots \\ c_j^N \end{bmatrix} = \frac{1}{\mu_j - \pi^2} \begin{bmatrix} d_j^1 \\ \vdots \\ d_j^N \end{bmatrix} \qquad j = 1 \ldots M$$

where T_N is the matrix of lemma 2.1 with

$$\sigma_j = \frac{1}{2\Delta t(\mu_j - \pi^2)}$$

From the fact that the eigenvalues μ_j of H are distinct and $\mu_j \to j^2\pi^2$ (see [2]) as $\Delta x \to 0$, j fixed, we have

$$\mu_j - \pi^2 \geq \mu > 0 \quad \text{for all } j \geq 2$$

if Δx is sufficiently small.

And $\mu_j - \pi^2 \geq \frac{1}{2}\left[1 - \frac{\sqrt{3\pi}}{6}\right] j^2 \pi^2$ for all sufficiently large j. Furthermore, if $\tilde{W}^j$, $j = 1 \ldots M$ are the orthonormal eigenvectors of H, then $\|\tilde{W}^j\|_\infty \leq K$ (see [13]).

Using the estimate

$$\max_{n=1\ldots N} |c_j^n| \leq \|T_N^{-1}(\sigma_j)\|_\infty \frac{O(\Delta t^2)}{\mu_j - \pi^2} \qquad j = 1 \ldots M$$

obtained by inverting (4.5) and using lemma 2.1, we have

$$\|E^n\|_\infty \leq \sum_{j=1}^{M} |c_j^n| \, \|\tilde{W}^j\|_\infty \leq O(\Delta t^2) \left(\sum_{j=1}^{M} \frac{\|\tilde{W}^j\|_\infty}{\mu_j - \pi^2} \right)$$

or

$$\|E^n\|_\infty \leq \frac{O(\Delta t^2)\|\tilde{W}^1\|_\infty}{\mu_1 - \pi^2} + O(\Delta t^2)$$

since $\sum_{j=2}^{M} \frac{\|\tilde{W}^j\|_\infty}{\mu_1 - \pi^2}$ is bounded independently of M. Thus

$$\sup_{n=1\ldots N} \|E^n\|_\infty \leq \text{constant}, \tag{4.6}$$

because $\mu_1 - \pi^2 = O(\Delta x^2)$ and we assume $\Delta x = O(\Delta t)$. Since the exact solution U is bounded, it follows from (4.6) that

$$\|V(\Delta t)\|_\infty \leq \text{constant as } \Delta t \to 0.$$

This proves the lemma.

Let us now examine the convergence of $V(\Delta t)$ to U. Let I_N be the $N \times N$ unit matrix.

Let $\underline{c} = \{c^n\}_{n=1}^N$ and $\underline{p} = \{p^n\}_{n=1}^N$ where $p^n = \cos n\Delta t$. Then, equation (4.3) takes the form

$$[S + (\mu_1 - \pi^2) I_N]\, \underline{c} = \underline{p}. \tag{4.7}$$

Let N be odd. Whether $T = \pi$ or $T = 2\pi$ we have

$$[\underline{p}, \psi_{\frac{N+1}{2}}] = \Delta t \sum_{n=1}^{N} \cos n\Delta t\, \psi^n_{\frac{N+1}{2}} = 0$$

Hence, if we solve (4.7) by expanding in the orthonormal eigenvectors ψ_s of S, we see immediately that the solution $\underline{c}$ satisfies

$$[\underline{c}, \psi_{\frac{N+1}{2}}] = 0 \tag{4.8}$$

Suppose now that

$$\| V(\Delta t) - U \|_2 \to 0 \text{ as } N \to \infty. \tag{4.9}$$

Since $U^n = \sin n\Delta t\, W^1$, this means that

$$\Delta t \sum_{n=1}^{N} |c^n - \sin n\Delta t|^2 \to 0 \text{ as } N \to \infty. \tag{4.10}$$

However, if $T = \pi$, $\sin t$ is positive on $(0, \pi)$, and

$$\Delta t \sum_{n=1}^{N} \psi^n_{(\frac{N+1}{2})} \sin n\Delta t \geq \beta > 0$$

and therefore, using (4.8),

$$\Delta t \sum_{n=1}^{N} (\sin n\Delta t - c^n)\, \psi^n_{(\frac{N+1}{2})} \geq \beta > 0$$

By Schwarz's inequality

$$0 < \beta \leq \| \psi_{(\frac{N+1}{2})} \|_{2,N} \{\Delta t \sum_{n=1}^{N} |c^n - \sin n\Delta t|^2\}^{1/2}$$

so that (4.10) is impossible, if $T = \pi$ and N is odd. In fact $V(\Delta t)$ cannot converge to U in any of the previously defined norms since this would imply (4.9).

On the other hand, if $T = 2\pi$, N odd, then we have

$$\Delta t \sum_{n=1}^{N} \psi^n_{\frac{N+1}{2}} \sin n\Delta t = 0 \tag{4.11}$$

Let $\underline{b}$ be the N vector $\{b^n\}$ with $b^n = \frac{\Delta t}{\sin \Delta t} \sin n\Delta t$, $n = 1 \ldots N$ then it is easily verified that $\underline{b}$ satisfies $S\underline{b} = \underline{p}$.

Using (4.7) we then have

$$S(\underline{b} - \underline{c}) = (\mu_1 - \pi^2)\underline{c}$$

Expanding $\underline{b} - \underline{c}$ in the orthonormal eigenvectors ψ_s of S, we have

$$\underline{b} - \underline{c} = \sum_{s=1}^{N} a_s \psi_s .$$

Observe that by (4.8) and (4.11), we have $a_{\frac{N+1}{2}} = 0$. Since $S(\underline{b} - \underline{c}) = \sum_s a_s \lambda_s \psi_s$ we have

$$\| S(\underline{b} - \underline{c}) \|_{2,N}^2 = \sum_s |a_s|^2 |\lambda_s|^2$$

$$= (\mu_1 - \pi^2)^2 \| \underline{c} \|_{2,N}^2 \leq K \Delta t^4$$

where K is a constant, because $\| \underline{c} \|_{2,N}$ is bounded and $(\mu_1 - \pi^2) = O(\Delta t^2)$.

Also, for $s = 1, \ldots N$, $s \neq \frac{N+1}{2}$, the eigenvalues λ_s of S which are smallest in absolute value are given by

$$\lambda = \pm \frac{i}{\Delta t} \sin \frac{\Delta t}{2} \text{ since } \Delta t = \frac{2\pi}{N+1} .$$

Therefore,

$$\frac{\sin^2 \frac{\Delta t}{2}}{\Delta t^2} \max_s |a_s|^2 \leq \sum_{s=1}^{N} |a_s|^2 |\lambda_s|^2 \leq K \Delta t^4$$

i.e.,

$$\max_s |a_s|^2 \leq K_1 \Delta t^4$$

Consequently,

$$\| \underline{b} - \underline{c} \|_{2,N}^2 = \sum_{s=1}^{N} |a_s|^2 \leq K_1 \Delta t^3 \sum_{s=1}^{N} \Delta t \leq 2\pi K_1 \Delta t^3$$

and hence

$$\Delta t \max_n |b^n - c^n|^2 \leq \Delta t \sum_{n=1}^{N} |b^n - c^n|^2 \leq 2\pi K_1 \Delta t^3$$

Thus

$$\max_n |b^n - c^n| \leq K_2 \Delta t$$

Now,

$$\|V^n - U^n\|_\infty \leq \|W^1\|_\infty \{|b^n - c^n| + |b^n - \sin n\Delta t|\}$$

from which we obtain

$$\|V - U\|_\infty \to 0 \quad \text{as} \quad \Delta t \to 0$$

Thus $V(\Delta t)$ converges uniformly to U if N is odd provided $T = 2\pi$.

On the other hand, <u>if N is even</u>, S has no zero eigenvalues and

$$\operatorname{Min}_s |\lambda_s| = O(1) \quad \text{as} \quad \Delta t \to 0$$

Hence as before

$$\operatorname{Max}_{s=1 \ldots N} |a_s|^2 \leq K_1 \Delta t^4$$

We will see later, however, that whether $T = \pi$ or 2π and whether N is even or odd, it is not possible to solve the system of difference equations (4.3) by either the point Jacobi or the point successive over relaxation method. We conclude this section with an observation on the Moore-Penrose pseudo-inverse (or general reciprocal), of a matrix [see Householder [12]], in relation to the semi-discrete approximation for the analytic problem (4.2).

If we discretize only the time variable in (4.2), as was done in Section 2, we obtain the system

$$\begin{cases} \dfrac{v^{n+1}(x) - v^{n-1}(x)}{2\Delta t} = \dfrac{\partial^2 v^n}{\partial x^2} + \pi^2 v^n + \sin \pi x \cos n\Delta t, \quad n=1, \ldots N \\ \text{with} \quad v^n(0) = v^n(1) = 0 \\ \text{and} \quad v^0(x) = v^{N+1}(x) = 0 \end{cases}$$

Clearly any solution of the above system must have the form

$$v^n(x) = c^n \sin \pi x \qquad n = 1 \ldots N$$

where the c^n's satisfy the equation

$$S\underline{c} = \underline{p}$$

in the previously defined notation.

Now let N be odd, so that S is singular. Since $\underline{p}$ is orthogonal to the null space of S, there always exists a solution to the last equation and, in fact, all solutions of $S\underline{c} = \underline{p}$ have the form

$$\underline{c} = \underline{b} + \beta \psi_{(\frac{N+1}{2})}$$

where β is an arbitrary constant and where $\underline{b}$ is the vector $\underline{b} = \{b^n\}_{n=1}^N$ with

$$b^n = \frac{\Delta t}{\sin \Delta t} \sin n\Delta t \qquad n = 1 \ldots N \ .$$

The "pseudo-inverse" of S defines a unique solution of $S\underline{c} = \underline{p}$ by the requirement that $\underline{c}$ be orthogonal to the null space of S.

Suppose now that $T = \pi$. Then as previously noted $[b, \psi_{\frac{N+1}{2}}]$ is positive so that the solution obtained via the pseudo-inverse must be such that

$$\underline{c} = \underline{b} + \beta \psi_{\frac{N+1}{2}} \quad \text{with} \quad |\beta| \geq \beta_0 > 0 \ ,$$

and with this $\underline{c}$, $v^n(x) = c^n \sin \pi x$ does not coverge to $\sin \pi x \sin n\Delta t$.

On the other hand, if $T = 2\pi$, then $[b, \psi_{\frac{N+1}{2}}] = 0$ and the pseudo-inverse gives the "correct" solution

$$\underline{c} = \underline{b} .$$

4.2 SOLUTION OF THE DIFFERENCE EQUATIONS BY ITERATIVE METHODS

The system of difference equations occurring in Sections 3 and 4.1 may be written in the block form

$$AV = F \tag{4.13}$$

where A is a block tridiagonal matrix of the form

$$A = \begin{bmatrix} \Lambda & \sigma I & & & \\ -\sigma I & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \sigma I \\ & & & -\sigma I & \Lambda \end{bmatrix} \quad \text{with} \quad \sigma = \frac{1}{2\Delta t} \quad \text{and}$$

where Λ is a nonsingular $M \times M$ matrix with distinct real eigenvalues λ_j, $j = 1 \ldots M$.

In the iterative solution of linear equations, one distinguishes between point iterative and block iterative methods. The systems of linear equations which arise in the numerical solution of elliptic boundary value problems are usually such that block iterative methods are more efficient than point iterative methods, i.e., they have a larger asymptotic rate of convergence [see Varge [17]]. Such is not the case for the system (4.13) above.

Consider the characteristic pairs of A. Let $|\lambda_1| \le |\lambda_2| \le \cdots \le |\lambda_M|$ be the eigenvalues of Λ and let Y^j $j = 1 \ldots M$ be the corresponding eigenvectors. For fixed s, j define the M vector $X^n_{s,j}$ by

$$X^n_{s,j} = i^n [\sin s (\frac{n\pi}{N+1})] Y^{(j)} \qquad n = 1 \ldots N$$

Let $X_{s,j}$ be the block vector

$$X_{s,j} = \begin{bmatrix} X'_{s,j} \\ \vdots \\ X^N_{s,j} \end{bmatrix} \qquad s = 1 \ldots N, \quad j = 1 \ldots M$$

Let

$$\mu_s = \frac{i}{\Delta t} \cos s (\frac{\pi}{N+1}) \qquad s = 1 \ldots N, \quad i = \sqrt{-1} \ .$$

Then as we readily verify

$$AX_{s,j} = (\lambda_j + \mu_s) X_{s,j} \qquad s = 1 \ldots N, \quad j = 1 \ldots M$$

so that the $X_{s,j}$ are the eigenvectors of A corresponding to the eigenvalues $(\lambda_j + \mu_s) = 0_{s,j}$ respectively.

Let us specialize the matrix A for the moment to be that which arises in connection with the system (4.3) above. Here $\Lambda = H - \pi^2 I$ has $\lambda_1 = O(\Delta t^2)$ as its smallest eigenvalue. Since $\mu_{\frac{N+1}{2}} = 0$ whenever N is odd, it follows that λ_1 is the smallest eigenvalue of A for N odd and hence $\|A^{-1}\|_2 = \frac{1}{|\lambda_2|} = O(\frac{1}{\Delta x})^2$. On the other hand, if N is even, the smallest

eigenvalue of A is $\lambda_1 \pm \frac{1}{\Delta t} \sin \frac{\pi}{2(N+1)}$ and we have $\Delta t = \frac{\pi}{N+1}$ or $\frac{2\pi}{N+1}$ depending on whether $T = \pi$ or $T = 2\pi$. Thus, if N is even, $\|A^{-1}\|_2$ remains bounded as $\Delta t \to 0$.

For the problem discussed in Section 3, i.e., where $\Lambda = L$ has positive eigenvalues which remain bounded away from zero as $\Delta x \to 0$ we also have $\|A^{-1}\|_2 \leq K$ as $\Delta x, \Delta t \to 0$.

The successive line over relaxation method (S.L.O.R)

In the SLOR method, P and N are defined as follows:

$$P = \frac{1}{\omega}[D + \omega E] \qquad N = \frac{1}{\omega}[(1-\omega)D - \omega F]$$

where ω is a nonzero real parameter and D, E, F are the following block matrices:

$$D = \begin{bmatrix} \Lambda & & \bigcirc \\ & \ddots & \\ \bigcirc & & \Lambda \end{bmatrix} \quad F = \begin{bmatrix} 0 & \sigma I & & \bigcirc \\ & \ddots & \ddots & \\ & & \ddots & \sigma I \\ \bigcirc & & & 0 \end{bmatrix} \quad E = \begin{bmatrix} 0 & & & \bigcirc \\ -\sigma I & \ddots & & \\ & \ddots & \ddots & \\ \bigcirc & & -\sigma I & 0 \end{bmatrix}$$

so that $A = D + E + F$.

The choice $\omega = 1$ in the SLOR method is known as the line Gauss-Seidel method. The line Jacobi method corresponds to the splitting $A = P' - N'$ where $P' = D$ and $N' = -(E + F)$ and $(P')^{-1}N'$ is called the line Jacobi matrix.

The following results are known for matrices such as A which are so-called consistently ordered 2-cyclic matrices (see Varga [17] and D. Young [18]).

a) If the SLOR method coverges, then $0 < \omega < 2$.

b) Let ρ be an eigenvalue of $P^{-1}N$, the SLOR matrix, and if χ satisfies

$$(\rho + \omega - 1)^2 = \chi^2 \omega^2 \rho \qquad \omega \neq 0, \tag{4.14}$$

then χ is an eigenvalue of the line Jacobi matrix. Conversely if χ is an eigenvalue of the line Jacobi matrix and if ρ satisfies (4.14), then ρ is an eigenvalue of the SLOR matrix. Hence, if the line Jacobi method converges, so does the line Gauss-Seidel and vice versa.

c) Starting from (4.14) and using conformal mapping arguments, D. Young [18] has proved the following:

Theorem. There exists an ω such that the SLOR method converges if and only if all the eigenvalues χ of the line Jacobi matrix satisfy $|\mathrm{Re}(\chi)| < 1$.

If $\beta > 0$ and if no eigenvalue of the line Jacobi matrix is contained in the closed exterior of the ellipse

$$[\mathrm{Re}(\chi)]^2 + \frac{[\mathrm{Im}(\chi)]^2}{\beta^2} = 1$$

and if $0 < \omega \leq \frac{2}{1+\beta}$, the SLOR method converges.

Let us apply these results to our situation:

Since $A = D + E + F$ has the eigenvalues $\mu_s + \lambda_j$, it follows that

$$\chi_{s,j} = \frac{\mu_s}{\lambda_j} \qquad s = 1 \ldots N, \qquad j = 1 \ldots M$$

are the eigenvalues of the line Jacobi matrix $D^{-1}(E+F)$. Hence if χ is the spectral radius of $D^{-1}(E+F)$, we have

$$|\chi| = \frac{\cos\left(\frac{\pi}{N+1}\right)}{|\lambda_1|\,\Delta t} \gtrsim O\left(\frac{1}{\Delta t}\right) \quad \text{as} \quad \Delta t \to 0$$

so that for all Δt sufficiently small the line Jacobi and Gauss-Seidel methods diverge for the matrix A. On the other hand, since $D^{-1}(E + F)$ has only pure imaginary eigenvalues, Young's theorem shows that if

$\beta = \dfrac{(1+\varepsilon)\cos\left(\frac{\pi}{N+1}\right)}{|\lambda_1|\Delta t}$ for any $\varepsilon > 0$ then the SLOR method converges for all $0 < \omega \le \dfrac{2}{1+\beta}$, i.e., for $0 < \omega \le \dfrac{2|\lambda_1(\Delta t)|\Delta t}{|\lambda_1(\Delta t)|\Delta t + (1+\varepsilon)\cos\frac{\pi}{N+1}}$.

Point Iterative methods for the "model problem" $\Lambda = H$

We consider now point iterative methods for the case $\Lambda = H$ corresponding to the heat equation. We will assume that Δt, Δx approach zero in such a way that $\Delta t = \gamma \Delta x$ where γ is a positive constant.

We will show that there always exists an interval $0 < \omega < \omega_3$ such that the point successive over relaxation method converges, but that the point Jacobi (and hence the point Gauss-Seidel) method converges if and

only if $\gamma \leq \gamma_c$ where γ_c is a constant which depends on the range of the space variable x in the analytic problem.

In the point Jacobi method, A is again split so that $A = P' - N'$ where now P' is the matrix obtained from A by deleting all but the main diagonal elements of A. If L and U are respectively the lower and upper triangular parts of N', the point successive over-relaxation method corresponds to the splitting $A = P - N$ with

$$P = \frac{1}{\omega}[P' + \omega L] \qquad N = \frac{1}{\omega}[(1-\omega)P' - \omega U]$$

where ω is a nonzero real parameter.

Moreover, the convergence results a), b), c) stated for line iterative methods remain valid if we replace line by point.

Consider first the eigenvalues of $(P')^{-1}N'$, given by

$$x_{s,j} = \frac{\lambda_j + \mu_s - d}{d} \qquad s = 1 \ldots N \qquad j = 1 \ldots M$$

where

$$\lambda_j = \frac{4}{\Delta x^2} \sin^2 \frac{j\pi\Delta x}{2} \qquad j = 1 \ldots M$$

are the eigenvalues of H, and

$$\mu_s = \frac{1}{\Delta t} \cos s \left(\frac{\pi}{N+1}\right) \qquad s = 1 \ldots N$$

and where $d = \frac{2}{\Delta x^2}$ are the constant diagonal elements of H.

If χ is the spectral radius of $(P')^{-1}N\lambda$ then

$$\chi^2 = \max_{s,j} \frac{(2 - \lambda_j \Delta x^2)^2 + \Delta x^4 |\mu_s|^2}{4} \tag{4.15}$$

and the maximum is attained for $s = j = 1$. Hence if $\Delta t = \gamma \Delta x$,

$$\chi^2 = (2 - 4 \sin^2 \frac{\pi \Delta x}{2})^2 + \frac{\Delta t^2}{\gamma^4} \cos^2 (\frac{\pi}{N+1}) \tag{4.16}$$

By Taylor's theorem we have

$$(2 - 4 \sin^2 \frac{\pi \Delta x}{2}) = 2 \cos \pi \Delta x = 2 - \pi^2 + \frac{\pi^4 \Delta x^4}{12} + O(\Delta x^6)$$

Hence

$$(2 - 4 \sin^2 \frac{\pi \Delta x}{2})^2 = 4 - 4\pi^2 \Delta x^2 + \frac{4}{3} \pi^4 \Delta x^4 + O(\Delta x^6)$$

$$= 4 - \frac{4\pi^2 \gamma^2 \Delta t^2}{\gamma^4} + \frac{4}{3} \frac{\pi^4 \Delta t^4}{\gamma^4} + O(\Delta t^6)$$

on using $\Delta t = \gamma \Delta x$. Therefore

$$\chi^2 = 1 - \Delta t^2 \frac{[4\pi^2 \gamma^2 - \cos^2 (\frac{\pi}{N+1})]}{4\gamma^4} + \frac{4}{3} \frac{\pi^4 \Delta t^4}{\gamma^4} + O(\Delta t^6) . \tag{4.17}$$

This shows that the point Jacobi method converges for all sufficiently small Δt if and only if

$$\gamma = \frac{\Delta t}{\Delta x} \geq \frac{1}{2\pi} \tag{4.18}$$

and the same is true of the point Gauss-Seidel method.

The eigenvalues $\chi_{s,j}$ of $(P')^{-1}N'$ satisfy

$$[\mathrm{Im}(\chi_{s,j})]^2 \leq \frac{\Delta t^2}{4\gamma^4} \tag{4.19}$$

$$[\mathrm{Re}(\chi_{s,j})]^2 \leq \frac{\pi^2\Delta t^2}{\gamma^2} + O(\Delta t^4) \tag{4.20}$$

Hence

$$\frac{1}{1 - [\mathrm{Re}(\chi_{s,j})]^2} \leq \frac{\gamma^2}{\pi^2\Delta t^2}\left[\frac{1}{1 + O(\Delta t^2)}\right]$$

and therefore

$$\frac{[\mathrm{Im}(\chi_{s,j})]^2}{1 - [\mathrm{Re}(\chi_{s,j})]^2} \leq \frac{1}{4\pi^2\gamma^2}\left[1 + O(\Delta t^2)\right]$$

Consequently, given any $\varepsilon > 0$, $\delta(\varepsilon)$ such that if $0 \leq \Delta t < \delta$

$$\frac{[\mathrm{Im}(\chi_{s,j})]^2}{1 - [\mathrm{Re}(\chi_{s,j})]^2} < \frac{1 + \varepsilon}{4\pi^2\gamma^2} \tag{4.21}$$

Hence if $\beta^2 = \frac{1 + \varepsilon}{4\pi^2\gamma^2}$ we have

$$[\mathrm{Re}(\chi_{s,j})]^2 + \frac{[\mathrm{Im}(\chi_{s,j})]^2}{\beta^2} < 1 \tag{4.22}$$

We see then that even if (4.18) is not satisfied, Young's theorem shows that the point successive over relaxation method converges for all ω such that

$$0 < \omega \leq \frac{2}{1 + \sqrt{\frac{1+\varepsilon}{4\pi^2\gamma^2}}}$$

Point iterative methods for the system (4.3)

Suppose now that $\Lambda = H - \pi^2 I$. In this case the eigenvalues of $(P')^{-1}N'$ are given by

$$\chi_{s,j} = \frac{\lambda_j - \pi^2 + \mu_s - d}{d} \qquad s = 1 \ldots N \quad j = 1 \ldots M$$

where now $d = \frac{2}{\Delta x^2} - \pi^2$. Hence

$$\max_{s,j} \left| \mathrm{Re}(\chi_{s,j}) \right| = \frac{(2 - 4\sin^2 \frac{\pi\Delta x}{2})}{2 - \pi^2 \Delta x^2} = \frac{2\cos \pi\Delta x}{2 - \pi^2\Delta x^2}$$

Again by Taylor's theorem

$$\cos \pi\Delta x = 1 - \frac{\pi^2\Delta x^2}{2} + \frac{\pi^4\Delta x^4}{24} + O(\Delta x^6)$$

and therefore

$$\frac{\cos \pi\Delta x}{1 - \frac{\pi^2\Delta x^2}{2}} = [1 - \frac{\pi^2\Delta x^2}{2} + \frac{\pi^4\Delta x^4}{24} + O(\Delta x^6)]\,[1 + \frac{\pi^2\Delta x^2}{4} + \frac{\pi^4\Delta x^4}{4} + O(\Delta x^6)]$$

$$= 1 + \frac{\pi^4\Delta x^4}{24} + O(\Delta x^6) > 1$$

if Δx is sufficiently small.

Consequently the point successive over-relaxation method diverges for every ω by Young's theorem. In particular, the Gauss-Seidel method (and therefore the point Jacobi method) diverges.

1. Atkinson, F. V. Discrete and Continuous Boundary Problems. Academic Press, New York (1964).

2. Bellman, R. E. Introduction to Matrix Analysis. McGraw-Hill, New York (1960).

3. Büchner, H. "Über Konvergensätze, die sich bei der Anwendung eines Differenzen-verfahrens auf ein Sturm-Liouvillesches Eigenwertproblem ergeben" Math. Z. 51, 423-465 (1948).

4. Carasso, A. "An analysis of numerical methods for parabolic problems over long times," Ph.D. Thesis, University of W sconsin, Madison (1968).

5. Carasso, A. "Finite difference methods and the eigenvalue problem for ordinary differential operators of the second order in non-self adjoint form." (to appear).

6. Coddington, E. A. and Levinson, N. Theory of Ordinary Differential Equations. McGraw-Hill, New York (1955).

7. Courant, R. and Hilbert, D. Methods of Mathematical Physics. Vol. I, Interscience, New York (1953).

8. Friedman, A. Partial Differential Equations of Parabolic Type. Prentice-Hall, Englewood Cliffs, N. J. (1964).

9. Garabedian, P. R. Partial Differential Equations. Wiley, New York (1964).

10. Greenspan, D. "Approximate Solution of Initial-Boundary Parabolic Problems by Boundary Value Techniques," MRC Technical Summary Report #782, August 1967. U. S. Army Mathematics Research Center, University of Wisconsin.

11. Greenspan, D. Lectures on the numerical solution of linear, singular, and non-linear differential equations. Prentice-Hall (In press).

12. Householder, A. S. The theory of matrices in numerical analysis. Blaisdell, New York (1964).

13. Milne, W. E. Numerical Calculus. Princeton University Press, Princeton (1949).

14. Protter, M. H., and Weinberger, H. F. Maximum Principles in Differential Equations. Prentice-Hall, Englewood Cliffs, N. J. (1967).

15. Richtmyer, R. D. Difference Methods for Initial Value Problems. Interscience, New York (1957). 2nd edition with K. W. Morton (1967).

16. Southwell, R. V. Relaxation Methods in Theoretical Physics, Vol. II Clarendon Press, Oxford (1956).

17. Varga, R. S. Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, N. J. (1962).

18. Young, D. "Iterative methods for solving partial difference equations of elliptic type," Trans. AMS. 76, (1954) pp. 92-111.