INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or “target” for pages apparently lacking from the document photographed is “Missing Page(s)”’. If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in “sectioning” the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from “photographs” if essential to the understanding of the dissertation. Silver prints of “photographs” may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

Xerox University Microfilms
300 North Zeeb Road
Ann Arbor, Michigan 48106
KENDALL, Richard Parker, 1941-
ASYMPTOTIC ROUNDING ERROR ANALYSIS OF SEVERAL
TECHNIQUES FOR THE SOLUTION OF FINITE DIFFERENCE
ANALOGUES OF SELF-ADJOINT UNIFORMLY ELLIPTIC
DIRICHLET PROBLEMS.

Rice University, Ph.D., 1973
Mathematics

University Microfilms, A XEROX Company, Ann Arbor, Michigan
RICE UNIVERSITY

ASYMPTOTIC ROUNGING ERROR ANALYSIS OF SEVERAL TECHNIQUES FOR THE SOLUTION OF FINITE DIFFERENCE ANALOGUES OF SELF-ADJOINT UNIFORMLY ELLIPTIC DIRICHLET PROBLEMS

by

Richard P. Kendall

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Thesis Director's Signature:

Houston, Texas
November, 1972
# TABLE OF CONTENTS

I. Introduction

II. Chapter 0: Asymptotic Rounding Error Analysis of the Point Jacobi Iteration for Self-Adjoint Uniformly Elliptic Dirichlet Problems

III. Chapter 1: Asymptotic Rounding Error Analysis for Alternating Direction Techniques for Elliptic Self-Adjoint Dirichlet Problems

   A. Section 1: The Optimal Single Parameter Peaceman-Rachford Alternating Direction Iteration

   B. Section 2: The Higher Dimensional Case

   C. Section 3: The Multi-parameter Case in $\mathbb{R}^2$
      
      i. The Peaceman-Rachford 'Good' Parameters

      ii. The Wachspress 'Good' Parameters

IV. Chapter 2: Asymptotic Rounding Error Analysis for the Successive Overrelaxation Technique for Self-Adjoint Uniformly Elliptic Problems

   A. Section 1: Single Parameter Point Iterative Successive Overrelaxation

   B. Section 2: Semi-Iterative Symmetric Successive Overrelaxation
TABLE OF CONTENTS - (Continued)

V. Chapter 3: Asymptotic Rounding Error Analysis
   for the Factorization Procedure of Dupont,
   Kendall and Rachford for Self-Adjoint
   Uniformly Elliptic Dirichlet Problems 124

VI. Bibliography
ASYMPTOTIC ROUNDDING ERROR ANALYSIS OF SEVERAL TECHNIQUES
FOR THE SOLUTION OF FINITE DIFFERENCE ANALOGUES OF
SELF-ADJOINT UNIFORMLY ELLIPTIC DIRICHLET PROBLEMS

Introduction

Rounding error analyses of iterative techniques for elliptic problems has been a neglected topic despite the fact that as early as 1958 Kahan remarked in his Ph.D. dissertation that it is conceivable that the Jacobi iteration for the solution of Laplace's equation might be subject to rounding errors that grow like $\Theta(h^{-2})$, where $h$ is the discretization step, if a computational strategy that avoids this possibility is not followed. The reason for this lack of interest seems to have been the argument that rounding errors are a less significant difficulty for elliptic problems than truncation errors. Typically, this argument has been the following.

Consider the self-adjoint elliptic difference system

(i) \[ Au = b \]

which arises from the usual discretization of Laplace's equation in $\mathbb{R}^2$. Suppose that the consistent, stable, stationary iteration

(ii) \[ u^{(n+1)} = T u^{(n)} + f \]

is employed to solve (i). Furthermore, let us assume that
the error propagator of the technique, $T$, is very well-behaved, say Hermitian (as is the case for the simplest procedure, the point Jacobi method). Moreover, we expect at worst that $\|T\|_2 \leq 1 - c \ h^2$, $c > 0$. Denoting by $r^{(k)}$ the global rounding error after $k$ iterations, and by $e^{(k)}$ the local rounding error introduced in the $k^{th}$ iteration itself, we may write

$$(iii) \quad \hat{u}^{(n+1)} = u^{(n+1)} + r^{(n+1)} = T[u^{(n)} + r^{(n)}] + e^{(n+1)} + f,$$

where a caret over a quantity denotes that it is a computed rather than an exact value. Subtracting (ii) from (iii) yields the recursion

$$(iv) \quad r^{(n+1)} = T \ r^{(n)} + e^{(n+1)}.$$ 

Since we are assuming that $T$ is Hermitian we may claim that in the 2-norm

$$(v) \quad \|r^{(n+1)}\|_2 \leq \|I + T + T^2 + \cdots + T^n\|_2 \max_j \|e^{(j)}\|_2$$

$$\leq \left[ \frac{1 - s(T)^{n+1}}{1 - s(T)} \right] \max_j \|e^{(j)}\|_2$$

where $s(.)$ denotes the spectral radius of its argument. Hence the global rounding error behaves like $h^{-2} \max_j \|e^{(j)}\|_2$ at worst. It is argued that the local error is independent of $h$, which is asserted to be reasonable in view of the fact that the linear system (i) has no more than five entries per
row. In view of this assumption we expect that the local errors are proportional to the precision of the computing machine, say $2^{-t}$ for a binary machine where $t$ is the mantissa length of a single precision word. Consequently, the components of the rounding error vector behave like $\Theta(2^{-t}h^{-2})$ at worst. It is usually asserted at this point that one never undertakes a problem where $2^{-t}h^{-2}$ is not negligible compared to the truncation error, which behaves like $\Theta(h^2)$.

It is clear that the greater the demand for accuracy in the computed solution, the less valid the conclusion above. Indeed, far from being an excuse for ignoring rounding errors, the argument above compels a careful consideration of the asymptotic behavior of the rounding error, for from it we conclude that the global rounding error may grow like $\Theta(h^{-4})$ times the discretization error. From this point of view several of the assumptions above deserve careful scrutiny. Principal among these is the assumption that the local rounding errors are independent of $h$. Whether this is true depends on the properties of the particular error propagator $T$ as well as the properties of the difference system itself. The iterative techniques we shall consider are all based on the idea that we are approximately inverting the coefficient matrix of the elliptic difference system at each step of the iteration. (If we could invert it exactly we would obtain convergence in one step.)
Moreover, the error propagators of these techniques can all be written in the form

\[(vi) \quad T = I - B^{-1}A\]

where \(B^{-1}\) is the \textbf{approximate inverse} of the technique and \(A\) the coefficient matrix of the difference system. For the alternating direction technique for self-adjoint Dirichlet problems, for instance, the approximate inverse is given by \(B^{-1} = 2\rho(\rho + A_2)^{-1}. \ (\rho + A_1)^{-1}\) where \(A_1\) and \(A_2\) are the matrices corresponding to the second spatial differences in the \(x_1\) and \(x_2\) directions, respectively, and \(\rho\) is an iteration parameter chosen to accelerate convergence (see [13]). It is reasonable to expect the condition number of \(B\) to be comparable to that of \(A\), which is \(\Theta(h^{-2})\) for self-adjoint problems. Hence we may anticipate a contribution to the local error from the numerical solution of a system with coefficient matrix \(B\) which not only depends on \(h\), but may be proportional to \(h^{-2}\).

A second consideration ignored in the heuristic argument above is the strategy by which the iteration is advanced in the computing machine. It is clear, for example, that any strategy which requires the use of \(B^{-1}\) is far more likely to propagate rounding errors than one which does not, for \(B^{-1}\) is not a sparse matrix in general. Moreover, we shall discover that there are modifications in computing strategy more subtle than the mere avoidance of the use of \(B^{-1}\) which yield either improved uniform bounds
on the global rounding error or weakened constraints on the machine precision required to insure asymptotic stability.

Before undertaking an asymptotic rounding error analysis of the point Jacobi, alternating direction, successive overrelaxation, and factorization techniques for the solution of the finite difference analogues of self-adjoint uniformly elliptic Dirichlet problems, we should like to make the following remarks.

In the following chapters the underlying self-adjoint uniformly elliptic Dirichlet problem to be considered will be

\[(vii) \quad \mathcal{L} U = - \frac{\partial}{\partial x_1} \left[ a_1(x_1, x_2) \frac{\partial U}{\partial x_1} \right] - \frac{\partial}{\partial x_2} \left[ a_2(x_1, x_2) \frac{\partial U}{\partial x_2} \right] + q(x_1, x_2) U
\]

\[= g(x_1, x_2) \text{ in } \Omega \]

\[U = f(x_1, x_2) \text{ on the boundary of } \Omega \]

where \( \Omega \) is an open bounded subset of \( \mathbb{R}^2 \), \( a_1(x_1, x_2), a_2(x_1, x_2) \geq \eta > 0 \), and \( q(x_1, x_2) \leq 0 \) throughout \( \Omega \). Occasionally \( \Omega \) will be restricted even further and it will always be assumed that the difference system obtained in the discretization process is irreducible.

Employing the usual five point finite difference analogue of (vii) based on interpolation of the coefficients
of the differential operator at the half-step, we obtain the elliptic difference system

\[(viii) \quad A\ u = b\]

where, denoting by \( \Omega_h \) the set of mesh points obtained by superimposing a uniform grid of spacing \( h \) over the region \( \Omega \), that is, \( \Omega_h = \{(jh, kh) : j, k \text{ integers}\} \cap \Omega \), \( A \) operates on \( \mathbb{R}^{\text{card } \Omega_h^0} \), the real vector space whose dimension is the cardinality of \( \Omega_h^0 \), the subset of interior mesh points. Specifically, for the grid function \( u \) at the point \((jh, kh) \in \Omega_h^0\), \( A \) is defined by

\[(ix) \quad A_{j, k}u_{j, k} = b_{j, k}u_{j, k} + c_{j, k}u_{j+1, k} + f_{j, k}u_{j, k+1} + c_{j-1, k}u_{j-1, k} + f_{j, k-1}u_{j, k-1},\]

where

\[(x) \quad b_{j, k} = -(c_{j, k} + c_{j-1, k} + f_{j, k} + f_{j, k-1}) + q(jh, kh)h^2,\]

\[c_{j, k} = -a_1((j + \frac{1}{2})h, kh), \quad f_{j, k} = -a_2(jk, (k + \frac{1}{2})h),\]

\[c_{j-1, k} = -a_1((j - \frac{1}{2})h, kh), \quad f_{j, k-1} = -a_2(jh, (k - \frac{1}{2})h).\]

Note that the system has been scaled by \( h^2 \). For the problem \((vii)\) it is well known that \( A \) is a Stieltjes matrix (see Varga [14]). In Chapter 1 we shall consider the \( p \)-dimensional analogue of \((vii)\), with the obvious generalization of the notation above.
Next we take up the question of how the magnitude of the rounding errors should be measured. It seems clear that either the max-norm or some other norm blind to the size of the difference system should be employed in an asymptotic analysis, for the difference system grows like \( \Theta(h^{-2}) \) as \( h \) tends to zero. In order to avail ourselves of any matrix theoretic properties of the error propagators of the techniques to be considered, we choose a scaled Euclidean norm whose subordinate matrix norm is the 2-norm. Specifically,

\[
(xi) \quad \|u\|_{L^2} = (u,u)^{\frac{1}{2}} = h^p \sum_{p \in \mathcal{O}_h} u(p)^2 \quad p \geq 2
\]

where \( p \) is the dimension of the Euclidean space on which the Dirichlet problem is defined. The designation \( L^2 \) above is motivated by the observation that for smooth functions, this norm tends to the \( L_2 \)-norm as \( h \) tends to zero. In a sense (xi) defines a family of norms indexed by the parameter \( h \). We shall see that the techniques to be considered propagate rounding errors bounded in \( L^2 \)-norm uniformly with respect to the iteration step provided that the mantissa of a machine word is chosen proportional to \( h^2 \). This agrees with the definition for a numerically stable method given in [1].

Turning to the matter of machine precision, we shall consider only the floating point mode of machine arithmetic
and it should be understood that unless precision is specifically mentioned it will be assumed to be single precision. For bounds on the rounding errors in the single precision arithmetic operations we shall accept those obtained by Wilkinson [15]. Denoting by \( f_t(a \circ b) \) the single precision floating point analogue of the arithmetic operation \( a \circ b \), Wilkinson has established that

\[(xii) \quad f_t(a \circ b) = (a \circ b)(1 + \varepsilon) , \quad |\varepsilon| \leq \nu \]

where \( \nu = sN^{-t}, \ s = \frac{1}{2}, 1 \) depending on whether the computing machine rounds before storage or truncates, \( N \) is the base of machine arithmetic, and \( t \) is the mantissa length of a single precision floating point machine word. Not only the bound given in (xii) but also its form depends on the accumulation of floating point sums, products, quotients, and differences in a double precision accumulator, that is, one of length at least \( 2t \). Since we shall not assume that double precision machine operations take place in a quadruple precision accumulator, the expression given in line (xii) is invalid for double precision floating point operations even if \( \nu \) is replaced by \( \nu^2 \). In fact, we must write

\[(xiii) \quad f_{td}(a \circ b) = a(1 + \varepsilon_1) \circ b(1 + \varepsilon_2) , \quad |\varepsilon_1|, |\varepsilon_2| \leq o(\nu^2) \]

where the subscript \( d \) denotes double precision. In the
accumulation of inner products frequent use will be made of what Wilkinson refers to as $f_{l_2}$ arithmetic, the double precision accumulation of an inner product or sum followed by rounding to single precision before storage. On many machines this mode of arithmetic is very little more costly than single precision arithmetic but much more stable with respect to rounding errors. Wilkinson has established that

$$(xiv) \quad f_{l_2} \left( \sum_{i=1}^{p} a_i b_i \right) - \sum_{i=1}^{p} a_i b_i (1 + \epsilon) = \sum_{i=1}^{p} a_i b_i \epsilon_i ,$$

$$|\epsilon| < \nu , \quad |\epsilon_i| \leq \Theta[(p-i+2)\nu^2] .$$

In contrast to this result, the single precision accumulation of an inner product satisfies

$$(xv) \quad f_{l} \left( \sum_{i=1}^{p} a_i b_i \right) = \sum_{i=1}^{p} a_i b_i (1 + \epsilon_i) , \quad |\epsilon_i| < 1.06i \nu .$$

By the symbols $f_{l}[Q \circ S]$ and $f_{l_2}[Qx]$ we shall denote the obvious extensions of the floating point operations above to vectors and matrices.

The stability questions arising from our inability to represent the data of the problem (vii), (viii) exactly in the floating point format of machine arithmetic will be ignored here. It will be assumed that the coefficient matrix A, the iteration parameters and the starting vector are exactly representable in single precision. Since we can regard slight perturbations in the data as the exact
data of a slightly perturbed problem, we may rely on the
inherent stability of (vii) and (viii) to conclude that no
growing error is introduced by assuming that the data is
exactly representable. In the analysis to follow a further
assumption about the starting vector is necessary. It will
be assumed that the starting vector $u^{(0)}$ is sufficiently
smooth that $\frac{Au^{(0)} - b}{h^2}$ remains bounded as $h$ tends to zero in
$L^2$-norm.

We shall borrow freely from the rounding error analyses
given by H. H. Rachford, Jr., for alternating direction
techniques for parabolic problems. Several results, in
particular, will be directly applicable not only to our
analysis of alternating direction techniques for the solution
of (viii) but other techniques as well. Foremost among these
is a lemma designated as lemma 3.1 in [10]. We state this
in slightly more general form as

**Lemma 0:** Let $Q = (q_{i,j})$ be a strictly diagonally dominant
matrix such that

$$(xvi) \quad |q_{i,i}| \geq \sum_{j \neq i} |q_{i,j}| + \delta , \quad \delta > 0 ,$$

then

$$\|Q^{-1}\|_\infty \leq \delta^{-1} .$$
Proof: We note that $Q^{-1}$ exists by Gerschgorin's theorem.

By definition

$$(xvii) \quad \|Q^{-1}\|_{\infty} = \sup_{y \neq 0} \frac{\|Q^{-1}y\|_{\infty}}{\|y\|_{\infty}} = \sup_{x \neq 0} \frac{\|x\|_{\infty}}{\|Qx\|_{\infty}}.$$

Normalizing $x$ so that $\|x\|_{\infty} = 1$ and $|x_k| = 1$, we obtain from (xvi) the bound

$$(xviii) \quad \|Qx\|_{\infty} = \max_i |\sum_j q_{i,j}x_j| \geq \sum_j q_{i,j}x_j \geq |q_{k,k}| - \sum_{j \neq k} |q_{k,j}| \geq 0;$$

however, since

$$(xix) \quad \|Q^{-1}\|_{\infty} = \sup_{\|x\|_{\infty} = 1} \frac{1}{\|Qx\|_{\infty}},$$

the result follows immediately.

In our analysis of the alternating direction iteration a direct appeal will be made to Theorem 3.1 of [10], which we shall designate here as

Theorem 1: [Rachford] Consider the tridiagonal system

$\Gamma w = d$ where $\Gamma$ is an irreducible diagonally dominant matrix

whose elements satisfy

$$(xx) \quad \delta + |a_j| + |c_j| < b_j (1 - 4\gamma + \delta(\gamma^2)),$$

$$j = 1, 2, \ldots, \text{rank } \Gamma$$

$$-1 \leq |a_j|, |c_j|$$
where \( a_j \) and \( c_j \) lie on the sub and superdiagonals, respectively, of the \( j^{th} \) row and \( b_j \) on the diagonal. Then if the solution of this system is carried out in single precision floating point arithmetic by Gaussian elimination without pivoting, the computed solution \( \hat{w} \) satisfies

\[
(\text{xxi}) \quad \frac{\| \hat{w} - w \|_2}{\| w \|_2} < (15 + 2\| \Gamma \|_2) \frac{\nu}{\delta} + \Theta[\frac{\nu}{\delta}],
\]

provided \( \frac{\nu}{\delta} < 1 \).

**Proof:** See [10] or [12].

Finally, we shall assume that all of the iterative techniques to be investigated are executed in residual form, that is, we shall assume that the \((n+1)st\) iterate is computed from the \(n^{th}\) by

\[
(\text{xxii}) \quad u^{(n+1)} = u^{(n)} + d^{(n)},
\]

\[
d^{(n)} = -B(Au^{(n)} - b)
\]

where the matrix \( B \) is the approximate inverse defined previously. The quantity \( Au^{(n)} - b \) is called the residual and should beyond some point in a convergent iteration tend to zero. The quantity \( d^{(n)} \) is referred to as the \(n^{th}\) difference correction and should also tend to zero with increasing \( n \). The choice of this strategy is motivated by the desire to execute what we shall see is the most sensitive step of
the computation, the numerical solution of $By = f(\text{Au}^{(n)} - b)$, with a small right-hand side. This will minimize the effects of an ill-conditioned $B$ (with respect to the solution of systems). In general we avoid computation with large quantities whenever possible to avoid large absolute errors in computed results. In the chapters that follow we shall describe an iteration merely by indicating its approximate inverse or at most its difference correction. Also, we shall employ the '$\circ$' notation liberally to describe the asymptotic behavior of functions. However, this notation will never be used for non-positive functions; hence it should be assumed that if $g(x) = \circ(h(x))$, then both $g(x)$ and $h(x)$ are non-negative.
Chapter 0. Asymptotic Rounding Error Analysis of the Point Jacobi Iteration for Self-Adjoint Uniformly Elliptic Dirichlet Problems

Description of the Technique

In order to motivate what follows, we shall take a brief look at the point Jacobi iteration for the p-dimensional analogue of (vii) of the Introduction. In residual form the point Jacobi iteration can be written as

\[ d^{(n)} = -D^{-1}(A^{(n)} - b) \]

where \( A \) is the coefficient matrix of the difference system described in line (ix) of the Introduction, and \( D \) is a diagonal matrix whose entries are the diagonal elements of \( A \). From (0.1) we can see that the approximate inverse of the point Jacobi technique must be \(-D^{-1}\), a diagonal matrix whose entries are bounded away from zero independently of \( h \) in view of the uniform ellipticity of the differential operator. Recall that the system (viii) has been scaled by \( h^2 \) to avoid computation with large numbers. It is well known that if \( A \) is an irreducible diagonally dominant matrix, as is certainly the case for the system (viii) arising from (vii) of the Introduction, then the point Jacobi iteration converges. In fact, if we decompose \( A \) into

\[ A = D - L - U \]
where $L$ and $U$ are the strictly lower and upper triangular components of $A$, respectively, then the spectral radius of the error propagator, $-D^{-1}(L+U)$, of the point Jacobi iteration satisfies

$$s(-D^{-1}(L+U)) \leq 1 - ch^2, \quad c > 0$$

for the problem (viii) arising from (vii), as we shall prove in Chapter 2.

**Asymptotic Rounding Error Analysis**

Babuska, et al [1] have argued that even if the assumption is made that the local rounding errors are independent of the mesh spacing, the global rounding error may grow proportional to $h^{-2}$ for self-adjoint problems. In fact, they exhibit a problem, $\Delta U = 1$ with zero boundary conditions, for which they assert such growth is obtained, given their assumption about the local rounding errors. We shall show that faster growth is impossible in $L^2$-norm with no assumption about the local errors provided that the iteration is executed in residual form.

Employing the generalized notation of the Introduction and an approach similar to that given in [10], we may write the floating point realization of (0.1) as

$$\hat{u}^{(n)} = (I + \theta_n) \left[ \hat{u}^{(n-1)} - (I + R_n)D^{-1} \left[ (A + \delta A_n)\hat{u}^{(n-1)} - (I + \Lambda_n)b \right] \right]$$

(0.4)
where as before a caret over a quantity denotes the computed value of the corresponding exact quantity. The perturbation matrices \( \theta_n, R_n, A_n, \) and \( \Lambda_n \) above are defined as follows: (i) \( \theta_n \) is a diagonal matrix whose entries are bounded in magnitude by \( \nu \) representing the relative error introduced in the accumulation of the vectors \( \hat{u}^{(n-1)} \) and 
\[-(I + R_n)D^{-1}[(A + \delta A_n)\hat{u}^{(n-1)} - (I + \Lambda_n)b]; \]
(ii) \(-(I + R_n)D^{-1} \)
is the matrix we accept as the computational analogue of the approximate inverse of the point Jacobi technique, \(-D^{-1} \), that is, \( R_n \) is the matrix such that if the computed solution of \( D\hat{x} = \hat{u} \) is denoted by \( \hat{\xi} \), then \( \hat{\xi} = (I + R_n)D^{-1}\hat{u} \);
(iii) \( \delta A_n \) and \( \Lambda_n \) are the perturbations introduced in the accumulation of the residual \( Au^{(n)} - b \) as an inner product, which from (xvi) satisfy the bounds \( \|\delta A_n\|_\infty < (2p - 1)\|A\|_\infty \nu \) and \( \|\Lambda_n\|_\infty \leq \nu \) provided we add \(-b\) to \( Au^{(n-1)} \) as the first step of the computation. Note that we have subscripted \( R_n \) because it is necessary to solve the system \( D\hat{x} = f(Au^{(n-1)} - b) \) at every step of the iteration. We do not actually invert \( D \) in the machine.

Line (0.4) may be rearranged with the aid of the substitutions \( \hat{u}^{(n)} = u^{(n)} + r^{(n)} \) and \( H = -D^{-1}(L + U) \) as

\[
\hat{u}^{(n)} = (I + \theta_n)\left\{ [H - R_nD^{-1}A - (I + R_n)D^{-1}\delta A_n]r^{(n-1)}
\right.
\]
\[
+ [u^{(n)} - R_nD^{-1}(Au^{(n-1)} - b)]
\]
\[
- (I + R_n)D^{-1}(\delta A_nu^{(n-1)} - \Lambda_n b)] \right\}.
\]
Subtracting \( u^{(n-1)} \) from both sides of (0.5) we obtain

\[
(0.6) \quad r^{(n)} = (I + \alpha_n)[H - [R_nD^{-1}A + (I + R_n)D^{-1}\delta A_n]]r^{(n-1)} + \theta_nu^{(n)} - (I + \theta_n)[R_nD^{-1}(Au^{(n-1)} - b)] \\
+ (I + \alpha_n)(I + R_n)D^{-1}(\delta A_nu^{(n-1)} - \Lambda_n b),
\]

which may be written in the form

\[
(0.7) \quad r^{(n)} = [H + \Delta_n]r^{(n-1)} + e^{(n)}
\]

where

\[
(0.8) \quad \Delta_n = \theta_nH - (I + \theta_n)[R_nD^{-1}A + (I + R_n)D^{-1}\delta A_n],
\]

and

\[
(0.9) \quad e^{(n)} = \theta_nu^{(n)} - (I + \theta_n)[R_nD^{-1}(Au^{(n-1)} - b)] \\
- (I + \theta_n)(I + R_n)D^{-1}(\delta A_nu^{(n-1)} - \Lambda_n b).
\]

From this last line it is clear that when we take into account the actual steps of machine computation necessary to advance the iteration, we do not obtain a recursion as simple as the one of line (iv) of the Introduction. Whether the local error can be bounded in \( L^2 \)-norm independent of \( h \) depends on whether the computational analogue of the approximate inverse, \((I + R_n)D^{-1}\), is bounded in 2-norm independent of \( h \). The solution of \( Dz = ft(Au^{(n-1)} - b) \) is very easily accomplished and requires only one division per mesh point.
Consequently, $R_n$ can be taken to be a diagonal matrix all of whose elements are bounded in magnitude by $\nu$, and since the elements of the diagonal matrix $D^{-1}$ are all bounded away from zero independent of $h$, we may indeed claim that the local errors are bounded in $L^2$-norm independent of $h$. This follows from the remark above and a term by term examination of the expression for the local error given in line (0.9). Recall that since the iteration is convergent and consistent, $u^{(n)}$ tends to a $O(h^2)$ approximation to the solution of (vii) and the scaled residual $h^{-2}(Au^{(n-1)} - b)$ tends to zero; hence both quantities are bounded in $L^2$-norm uniformly in $n$. (See page (vi) of the Introduction.)

From the bound

$$(0.10) \quad \|r^{(n)}\|_2^2 \leq \frac{1 - [\|H\|_2^2 + \delta]^n}{1 - [\|H\|_2^2 + \delta]} \max_j \|e(j)\|_2^2,$$

where $\delta$ is a uniform bound for the 2-norm of the perturbations $\Delta_n$, we discover that a sufficient condition for the existence of a uniform $L^2$-norm bound on the global rounding error is the requirement that $s(H) + \delta < 1$. Since we may take $\delta = \nu\|H\|_2 + (1 + \nu)\nu[\|D^{-1}A\|_2 + O(\nu^2)]$, it is clear that a choice of $\nu$ proportional to $h^2$ will suffice to guarantee that a uniform bound for the global rounding error exists. By requiring that this ratio be sufficiently small we may be assured that $\|H\|_2 + \delta < 1 - ch^2$; hence

$$(0.11) \quad \|r^{(n)}\|_2^2 \leq C \left[ \frac{\nu}{h^2} \right].$$
Several remarks about the analysis above are in order now. The approximate inverse here, $D^{-1}$, is unusual in that it is not ill-conditioned in any sense. (It is also true that $D^{-1}$ is not a very good approximation for $A^{-1}$.) For methods whose error propagators yield more efficient iterators this is not the case. Consequently a great deal more effort will be required to obtain 'nice' bounds on the error perturbations which are the analogues of $\Delta_n$ and $e^{(n)}$ above. It will nevertheless be our goal to establish the same sort of result for the alternating direction, successive overrelaxation, and factorization techniques as was obtained here, namely, the existence of a uniform bound for the $L^2$-norm of the global rounding error with no more stringent requirement on the precision of the machine than that $\nu$ be taken proportional to $h^2$ asymptotically.
Chapter 1. Asymptotic Rounding Error Analysis for
Alternating Direction Techniques for Elliptic
Self-Adjoint Dirichlet Problems

Section 1: The Optimal Single Parameter Peaceman-Rachford
Alternating Direction Iteration

Given the open bounded region \( \Omega \) in \( \mathbb{R}^2 \), we shall again
confine ourselves to the self-adjoint uniformly elliptic
Dirichlet problem

\[
\begin{align*}
\mathcal{A} \ U &= f \quad \text{in} \quad \Omega \\
\quad U &= g \quad \text{on} \quad \partial \Omega
\end{align*}
\]

where \( \mathcal{A} \) is the operator described in line (vii) of the
Introduction. It is clear from the discretization of (1.1)
given in line (ix) of the Introduction that we may split \( \mathcal{A} \),
the finite difference approximation for \( \mathcal{A} \), into the components

\[
\begin{align*}
\mathcal{A} &= \mathcal{A}_1 + \mathcal{A}_2 + \Sigma
\end{align*}
\]

where \( h^{-2} \mathcal{A}_i \) is a tridiagonal matrix, for some ordering of
the mesh points, representing the centered second differences
in the \( i^{\text{th}} \) direction and \( \Sigma \) corresponds to the discretization
of \( h^2 q(x_1,x_2) \). In terms of this spatial decomposition of \( \mathcal{A} \)
the approximate inverse of the single parameter Peaceman-
Rachford alternating direction iteration is given by

\[
\begin{align*}
\mathcal{B}^{-1}_\rho &= 2\rho (\rho + \mathcal{A}_2)^{-1} (\rho + \mathcal{A}_1)^{-1}
\end{align*}
\]
where $\rho$ denotes the iteration parameter and $A_i = G_i + \frac{1}{2} \Sigma$, $i = 1, 2$. $B_{\rho}^{-1}$ exploits the simplicity of the spatial components of $A$ without sacrificing, for appropriate choice of $\rho$, a favorable 2-norm comparability to $A^{-1}$. Moreover, we shall show that in addition to being sparse ($\rho + A_i$) is diagonally dominant; hence $B_{\rho}^{-1}$ may be computed by successive steps of Gaussian elimination without pivoting. For sparse diagonally dominant matrices Gaussian elimination is not only accomplished with modest effort but is also stable with respect to the propagation of rounding errors (see [3] and [10]). Since the $G_i$ may be reduced to tridiagonal form by a mere reordering of $\Omega_h^0$ we shall henceforth assume them to be tridiagonal without explicit reference to the reordering. Furthermore, we shall assume that the finite difference system has been scaled if necessary to facilitate the application of Theorem 1.

Although it will not be necessary for the stationary iteration (1.3), we shall in later sections restrict our analysis to the commutative case, that is, the case of commuting $G_1$ and $G_2$. Dirichlet problems giving rise to commuting matrices $G_1$ and $G_2$ have been completely characterized in [2] as those for which (i) $\Omega$ is parallel to the co-ordinate axes, and (ii) the coefficients $a_1(x_1, x_2)$, $a_2(x_1, x_2)$, and $q(x_1, x_2)$ of the differential operator $\mathfrak{A}$ are separable with respect to $x_1$ and $x_2$. For such problems it was shown that the value of $\rho$ which minimizes the spectral radius of the error propagator of the stationary
iteration (1.3) is given by

\[ \rho_{\text{opt}} = (cd)^{\frac{1}{2}} \]

where \([c,d]\) bounds the common spectral interval of \(A_1\) and \(A_2\). The problem of determining an optimal iteration parameter for the more general case of unequal spectral intervals was also considered in [2]. Moreover, for intervals \([a,b]\) and \([\alpha,\beta]\) containing the spectra of \(A_1\) and \(A_2\), respectively, where necessarily \(a > 0\), \(\alpha > 0\), it was established that

\[ \rho_{\text{opt}} = \begin{cases} 
(ab)^{\frac{1}{2}} & \text{if } a \geq \alpha, \text{ or } a \leq \alpha \text{ and } a \beta \geq b \alpha, \\
\frac{ab}{\alpha \beta} & \text{if } \alpha \beta \leq \alpha \beta \\
(\alpha \beta)^{\frac{1}{2}} & \text{if } b \geq \beta, \text{ or } b \leq \beta \text{ and } a \beta \leq b \alpha.
\end{cases} \]

For this choice of parameter the error propagator of the alternating direction iteration (1.3), which we shall denote by \(T_\rho\), satisfies

\[ \|T_\rho^\text{opt}\|_2 = \begin{cases} 
\left( \frac{b^{\frac{1}{2}} - a^{\frac{1}{2}}}{b^{\frac{1}{2}} + a^{\frac{1}{2}}} \right) \left( \frac{\beta - (ab)^{\frac{1}{2}}}{\beta + (ab)^{\frac{1}{2}}} \right) & \text{if } \rho_{\text{opt}} = (ab)^{\frac{1}{2}}, \\
\left( \frac{(\alpha \beta)^{\frac{1}{2}} - a}{(\alpha \beta)^{\frac{1}{2}} + a} \right) \left( \frac{\beta^{\frac{1}{2}} - \alpha^{\frac{1}{2}}}{\beta^{\frac{1}{2}} + \alpha^{\frac{1}{2}}} \right) & \text{if } \rho_{\text{opt}} = (\alpha \beta)^{\frac{1}{2}}.
\end{cases} \]

That the spectral intervals \([a,b]\) and \([\alpha,\beta]\) are essentially those we would obtain for the model problem \(a_1 = a_2 = 1\), \(q = 0\) follows from the simple monotonicity argument given below.
By taking $\bar{A}_1$ to be the matrix obtained by replacing $a_1$ everywhere by its maximum over $\bar{\Omega}$, call it $\overline{a}_1$; $\overline{A}_1$ to be the matrix obtained by replacing $a_1$ by its minimum over $\bar{\Omega}$, $\underline{a}_1$, it is clear that

$$\bar{b} = \max \left\{ \frac{(\bar{A}_1 u, u)}{(u, u)} \right\} = \max \left\{ \frac{(A_1 u, u)}{(u, u)} + \frac{((\bar{A}_1 - A_1) u, u)}{(u, u)} \right\}. \tag{1.7}$$

However, $\bar{A}_1 - A_1$ is non-negative definite; hence

$$\underline{b} \succeq b, \tag{1.8}$$

justifying our notation. By an analogous argument

$$\underline{a} \preceq a. \tag{1.9}$$

We have now reduced the problem of obtaining bounds for the spectral interval of $A_1$ to that of obtaining an estimate for the smallest eigenvalue of $A_1$ and the largest eigenvalue of $\bar{A}_1$. These two matrices differ from the component $A_1$ of the model problem only by a multiplicative constant; hence it is not surprising that

$$\underline{a} = 4\overline{a}_1 \sin^2 \frac{\pi}{2M} + \frac{\sigma}{2}, \quad \bar{b} = 4\underline{a}_1 \cos^2 \frac{\pi}{2M} + \frac{\bar{\sigma}}{2}. \tag{1.10}$$

where $\bar{\Omega}$ is contained in a rectangle of length $[M + 1]h$, $\bar{\sigma} = \sup \limits_{\bar{\Omega}} (-h^2 q)$, and $\underline{a} = \inf \limits_{\bar{\Omega}} (-h^2 q)$. A similar result holds for $A_2$.

**Asymptotic Rounding Error Analysis**

We turn our attention now to the growth of rounding error in the single precision analogue of (1.3).
Specifically, we want to show that a fixed bound for the L²-norm of the global rounding error may be maintained by a choice of $\nu$ proportional to $h^2$ as $h$ tends to zero. This is the result obtained in Chapter 0 for the point Jacobi iteration for self-adjoint uniformly elliptic problems.

Instead of the sequence

(1.11) \[ u^{(n)} = u^{(n-1)} - 2\rho (\rho + A_2)^{-1}(\rho + A_1)^{-1}(Au^{(n-1)} - b), \]

\[ n \geq 1, \quad \rho = \rho_{opt}, \]

the computing machine forms (in single precision)

(1.12) \[ \hat{u}^{(n)} = (I + \xi_n)[\hat{u}^{(n-1)} - 2\rho (I + \xi_n)(I + R_{n,2})(\rho + A_2)^{-1} (I + R_{n,1})(\rho + A_1)^{-1} \hat{f}_{L^2}(A\hat{u}^{(n-1)} - b)], \]

\[ n \geq 1, \quad \rho = \rho_{opt}. \]

The residual $Au^{(n-1)} - b$ is computed in $f_{L^2}$-arithmetic in order to mitigate the effect of an ill-conditioned approximate inverse, and as a precaution against catastrophic cancellation. The matrices $\xi_n$, $R_{n,1}$, and $R_{n,2}$ have the same general meanings as in Chapter 0. In particular, the perturbations $R_{n,1}$ and $R_{n,2}$ represent the relative errors introduced in the single precision solution of linear systems with $(\rho + A_1)$, respectively. Although the approximate
inverse of the point Jacobi iteration can be employed in the
iteration in inverse form, this is not feasible here, for the
inverses of the \((\rho + A_1)\) are positive matrices and thus as
expensive to compute as \(\Theta(h^{-2})\) steps of the entire alter-
nating direction iteration. Instead we compute the approx-
imate inverse by solving successively the systems

\[
(\rho + A_1) \ x = f_{L_2}(A\hat{u}^{(n-1)} - b)
\]

\[
(\rho + A_2) \ y = x ,
\]

then multiplying by \(2\rho\). The perturbation \(\hat{\gamma}_n\) is a diagonal
matrix whose elements are bounded in magnitude by \(\nu\) repre-
senting the relative error introduced in this multiplication.

Of course, rounding errors will force us to accept as
an approximation to the exact solution of the systems of
(1.13) the exact solution of the perturbed systems

\[
[(\rho + A_1) + K_1] \ \hat{x} = f_{L_2}(A\hat{u}^{(n-1)} - b)
\]

\[
[(\rho + A_2) + K_2] \ \hat{y} = \hat{x} .
\]

Consequently, we shall have formed not \((\rho + A_1)^{-1}f_{L_2}(A\hat{u}^{(n-1)} - b)\),
but rather \((I + R_{n,1})(\rho + A_1)^{-1}f_{L_2}(Au^{(n-1)} - b)\) where

\[
R_{n,1} = [(\rho + A_1) + K_1]^{-1} (\rho + A_1) - I .
\]

An expression analogous to the one given in (1.15) holds
for \(R_{n,2}\). To obtain bounds for the 2-norm of the perturba-
tions \(K_1\) and \(K_2\) we must examine carefully the numerical
solutions of (1.13).

Generalizing line (xiv) of the Introduction we obtain

\begin{equation}
\ell_2(A\hat{u}^{(n-1)} - b) = (I + \Psi_n)(A + \delta A_n)\hat{u}^{(n-1)} - (1 + \Xi_n)b
\end{equation}

where

\begin{equation}
\|\Psi_n\|_2 \leq \nu, \hspace{1em} \|\delta A_n\|_2 \leq \Theta(\|A\|_2 \nu^2), \hspace{1em} \|\Xi_n\|_2 \leq \Theta(\nu^2)
\end{equation}

provided the vector \(b\) is accumulated last in the inner-product \(Au^{(n-1)} - b\).

Subtracting line (1.11) from line (1.12) yields

\begin{equation}
\begin{aligned}
r^{(n)} &= \hat{u}^{(n)} - u^{(n)} \\
&= (I + \Theta_n)[I - 2 \rho(\rho + A_2)^{-1}(\rho + A_1)^{-1}A + \Lambda_n]r^{(n-1)} + e^{(n)}
\end{aligned}
\end{equation}

where

\begin{equation}
\Lambda_n = -2\rho[\Gamma_n A + (\rho + A_2)^{-1}(\rho + A_1)^{-1} + \Gamma_n \delta A_n],
\end{equation}

\begin{equation}
\Gamma_n = (\rho + A_2)^{-1}(\rho + A_1)^{-1} - (I + \Psi_n)(I + R_{n,2})(\rho + A_2)^{-1}
\end{equation}

\[ (I + R_{n,1})(\rho + A_1)^{-1}(1 + \Psi_n), \]

and

\begin{equation}
\begin{aligned}
e^{(n)} &= \rho_n u^{(n)} - 2\rho(1 + \Theta_n)[\Gamma_n (Au^{(n-1)} - b) \\
&\hspace{1em} + [(\rho + A_2)^{-1}(\rho + A_1)^{-1} + \Gamma_n] \cdot [\delta A_n u^{(n-1)} - \Xi_n b]}
\end{aligned}
\end{equation}

Recognizing the term underlined in (1.18) as the error
propagator of the alternating direction iteration, \( T_\rho \), we have established that

\[
(1.21) \quad r^{(n)} = (1 + \theta_n)(T_\rho + \Lambda_n)r^{(n-1)} + e^{(n)}
\]

for the residual strategy of computation. From (1.21) follows the \( L^2 \)-norm bound

\[
(1.22) \quad \|r^{(n)}\|_2 \leq \frac{1 - [\|T_\rho\|_2 + \delta]^{(n)}}{1 - [\|T_\rho\|_2 + \delta]} \max_j \|e^{(j)}\|_2
\]

where \( \delta = \sup_n \|\theta_n T_\rho + (I + \theta_n)\Lambda_n\|_2 \), provided \( \delta \) and the rightmost factor in (1.22) exist. The bound \( \delta \) will be shown to be a function of \( \nu \) and \( h \), which if it can be constrained to satisfy \( \|T_\rho\|_2 + \delta < 1 \) will certainly yield a uniform bound on the \( L^2 \)-norm of the global rounding error provided that the local errors \( e^{(j)} \) are sufficiently well behaved.

For the Dirichlet problem (1.1) we obtain from lines (1.10) and (1.7) the bound

\[
(1.23) \quad \|T_\rho\|_2 \leq 1 - ch, \quad c > 0 ;
\]

hence we must be prepared to take \( \delta \) to zero like \( \theta(h) \) to maintain a fixed bound on the \( L^2 \) norm of the global rounding error via (1.22) regardless of the behavior of the local errors.

Consider, then, the problem of obtaining the uniform bound \( \delta \). A uniform bound of the form \( \theta(\nu) \) for the 2-norm of the term \( \theta_n T_\rho \) is immediate. More difficult is a uniform
bound on the 2-norm of $\Lambda_n$. Since the perturbation $\Gamma_n$, which
is recognized as the absolute error in the computed approximate inverse from the definition in line (1.19), plays the
principal role in the perturbation $\Lambda_n$, we shall consider it
first. This error depends not only on the relative errors
associated with the numerical solutions of the systems of
(1.13) but on the inverses as well. Unfortunately, these
inverses have large 2-norms. Moreover, $(\rho + A_i)$ is diag-
onally dominant only by amount $\rho$ since the row sums of $A_i$
corresponding to interior mesh points are zero; hence, from
lemma 0 and lines (1.5) and (1.10) we obtain

\begin{equation}
(1.24) \quad \|(\rho + A_i)^{-1}\|_\infty \leq \frac{1}{\rho} = o(h^{-1}) \quad , i = 1, 2 \quad .
\end{equation}

A consideration of the model problem $a_1 = a_2 = 1$, $q = 0$
gives assurance that this bound is not unduly pessimistic.
The symmetry of $(\rho + A_1)$ permits the replacement of the max-
norm in (1.24) with the 2-norm. Consequently, we conclude
from Theorem 1 that the relative error introduced in the
solution of tridiagonal systems with $(\rho + A_i)$ by Gaussian
elimination without pivoting satisfies

\begin{equation}
(1.25) \quad \|R_{ni,i}\|_2 \leq (15 + 2\|\rho(A_i)\|_2) \frac{\nu}{\rho} + o[\frac{\nu}{\rho}]^2 \quad ,
\end{equation}

$$
\bar{\rho} = \rho - 4[\|\text{diag}(A_i)\|_\infty \nu + o(\nu^2)] \quad , \quad i = 1, 2
$$

provided $\frac{\nu}{\rho} < 1$. A term by term examination of line (1.19)
reveals that

\[(1.26a) \quad \|\Gamma_n\|_2 \leq \delta \left[ \frac{\nu}{\rho \bar{\rho}^2} \right], \]

and thus

\[(1.26b) \quad \|\Lambda_n\|_2 \leq \delta \left[ \frac{\nu}{\rho^2} \right]. \]

Consequently, we may take \( \delta = \Theta \left[ \frac{\nu}{\rho^2} \right] \). If we accept this uniform bound we shall be forced to take \( \nu \) proportional to \( h^3 \) in order to insure that \( [\|T_{\rho}\|_2 + \delta]^n \) decays with increasing \( n \). Such a constraint on the precision of the machine is more severe by a factor of \( h \) than the one obtained in Chapter 0 for the point Jacobi iteration. Of course, the approximate inverse for the alternating direction technique consists of factors which have large condition numbers as is not the case for the point Jacobi technique. The obvious remedy for the influence of this ill-conditioning is to purify the computed solutions of the systems (1.13). From a practical point of view, purification will not be valuable if improved bounds are obtained for the relative errors in the computed solutions of (1.13) only asymptotically. In fact, we would be unwilling to employ purification if more than one iteration is required. Fortunately, this is not the case as we shall show. First we describe the purification technique.
Wilkinson [15] suggests that the accuracy of the computed solution of any linear system

\[(1.27) \quad B \hat{x} = c,\]

denote it by \(\hat{x}\), can be refined by the following procedure. From the residual \(c - B \hat{x}\), computed in \(f_{\ell_2}\)-arithmetic, form a difference correction \(d^{(n)} = f_{\ell_2}(B^{-1}(c - B \hat{x}))\). Compute a new, hopefully more accurate, approximate solution in single precision by adding to the difference correction the first approximate solution \(\hat{x}\). We have not described this recursively because we intend to show that just one such iteration is sufficient to improve the bound on the relative error in the computed solution of a linear system with \((\rho + A_l)\) given in line (1.25) from \(\sigma \left[ \frac{\gamma}{\rho} \right]\) to \(\sigma(\gamma)\) for \(\gamma/h^2\) sufficiently small. Moreover, since we shall find that a single step of the purification procedure is also beneficial in other contexts, we state and prove

**Theorem 2:** Consider the linear system \(Bx = c\). Suppose that one step of the purification algorithm described above is performed on the approximate solution \(\hat{x}^{(1)}\) of this system. Also, assume that whenever a linear system in \(B\) such as \(Bx = c\) is solved numerically, not \(x\) but \(\hat{x}\) is obtained where \(\hat{x}\) is the exact solution of

\[(1.28) \quad (B + K)\hat{x} = c\]

for some perturbation \(K\) satisfying \(\|B^{-1}K\| \leq \frac{1}{2}\). Then denoting
by $\hat{x}^{(2)}$ the result of the purification, we obtain

\begin{equation}
\frac{\Vert \hat{x}^{(2)} - x \Vert}{\Vert x \Vert} \leq 2 \left\{ \Vert B^{-1} K \Vert + \nu \Vert B^{-1} \Vert \left[ N_o \Vert B \Vert \nu (1 + \nu)^2 \right.ight.
\end{equation}

\begin{equation}
+ (1 + \nu) \Vert B \Vert \right\} + \frac{3}{2} \nu \left\{ \frac{\Vert \hat{x}^{(1)} - x \Vert}{\Vert x \Vert}
\end{equation}

\begin{equation}
+ \left\{ \nu + 2 \nu \Vert B^{-1} \Vert \left[ (1 + \nu)^2 (\Vert B \Vert + N_o \Vert B \Vert) \right] \right\}
\end{equation}

where $N_o$ is the length of the longest inner product computed in the evaluation of the residual, and the norm may be taken to be either the 2 or max-norm.

**Proof:** The argument is a straightforward analysis of the purification step described above. Of course, we must take into account rounding errors that are inevitably a part of this procedure.

Consider first the accumulation of the residual $c - B \hat{x}^{(1)}$. In $f_{\ell_2}$-arithmetic we obtain

\begin{equation}
f_{\ell_2}(c - B \hat{x}^{(1)}) = (I + \hat{\epsilon}) [(I + \Lambda)c - (B + \delta B) \hat{x}^{(1)}],
\end{equation}

where $\delta B$, and $\Lambda$ are $\mathcal{O}(\nu^2)$ perturbations arising in the double precision accumulation of the residual and $\hat{\epsilon}$ is the perturbation representing the error introduced in subsequent rounding to single precision. If we add and subtract $\delta B x$ inside the brackets in (1.30) we may write

\begin{equation}
f_{\ell_2}(c - B \hat{x}^{(1)}) = (I + \hat{\epsilon}) [(B + \delta B)(x - \hat{x}^{(1)}) + (\Lambda B - \delta B) x].
\end{equation}
Subtracting the exact residual \(c - B\hat{x}^{(1)}\) from (1.31) we obtain

\[
(1.32) \quad f_\mathcal{L}_2(c - B\hat{x}^{(1)}) - (c - B\hat{x}^{(1)}) = P_1(x - \hat{x}^{(1)}) + P_2x,
\]

where \(P_1 = \delta B + (I + \delta)\delta B\), and \(P_2 = (I + \delta)(\Lambda B - \delta B)\).

Again taking into account rounding errors, we may write the purified solution as

\[
(1.33) \quad \hat{x}^{(2)} = (I + \Gamma)[\hat{x}^{(1)} + (B + K)^{-1}[B(x - \hat{x}^{(1)}) + (f_\mathcal{L}_2(c - B\hat{x}^{(1)}) - (c - B\hat{x}^{(1)}))]].
\]

If we subtract \(x\) from both sides of (1.33) and take the norm, we obtain

\[
(1.34) \quad \|\hat{x}^{(2)} - x\| \leq \|[I - (I + \Gamma)(B + K)^{-1}B]\| \|\hat{x}^{(1)} - x\|
\]
\[
+ \|\Gamma\| \|\hat{x}^{(1)}\| + \|(I + \Gamma)(B + K)^{-1}\|
\]
\[
\cdot \|[f_\mathcal{L}_2(c - B\hat{x}^{(1)}) - (c - B\hat{x}^{(1)})]\|.
\]

The assumption that \(\|B^{-1}K\| < \frac{1}{2}\) permits simplification of the coefficient of \(\|\hat{x}^{(1)} - x\|\) in the line above, to wit:

\[
(1.35) \quad I - (I + \Gamma)(B + K)^{-1}B = (B^{-1}K + \Gamma)(I + B^{-1}K)^{-1}.
\]

Making this substitution, we may rewrite line (1.34) as

\[
(1.36) \quad \|\hat{x}^{(2)} - x\| = \|[B^{-1}K + \Gamma]\|\|(I + B^{-1}K)^{-1}\| + \|\Gamma\|
\]
\[
+ \|(I + \Gamma)\|(B + K)^{-1}\|P_1\|\|\hat{x}^{(1)} - x\|
\]
\[
+ \|\Gamma\| + \|(I + \Gamma)\|(B + K)^{-1}\|P_2\|\|x\|.
\]
The perturbation $\Gamma$ is a diagonal matrix whose elements are bounded by $\nu$ in magnitude, arising in vector addition in single precision. Also, since $(B + K)^{-1} = (I + B^{-1}K)^{-1}B^{-1}$, we may accept the bound

$$\| (B + K)^{-1} \| < 2\|B^{-1}\| .$$

$P_1$ and $P_2$ satisfy

$$\|P_1\| \leq (1 + \nu)N_0\|B\| + \nu^2 + \nu\|B\| ,$$

and

$$\|P_2\| \leq (1 + \nu)\nu^2[\|B\| + N_0\|B\|] ,$$

respectively. These two bounds follow from line (xiv) of the Introduction.

Making substitutions from (1.38) and (1.39) into line (1.36) and dividing by $\|x\|$, the result of the theorem, line (1.29), is obtained.

We may now apply this theorem to the case at hand, namely, the problem of purifying the computed solutions of linear systems with $(\rho + A_i)$, $i = 1, 2$. In [10] it was established that whenever Gaussian elimination is employed to solve numerically tridiagonal systems with diagonal dominance the perturbation $K$ of the theorem above satisfies

$$\|K\|_\infty \leq 1.06 \nu [6 + 7a + 2.12 \nu (1 + 2a)]$$
in single precision where a is the largest element of the diagonal of the system. Denoting the right-hand side of (1.40) by \( k_\nu \), Theorem 2 yields the bound

\[
(1.41) \quad \frac{\| \hat{x}(2) - x \|}{\| x \|} \leq 2 \left\{ \frac{k_\nu}{\rho} + \frac{\nu}{\rho} \left[ 3\| (\rho + A_1) \| (1 + \nu) \right. \right.
\]
\[+ \left. (1 + \nu)\| (\rho + A_1) \| + \frac{3}{2} \nu \right\} \frac{\| \hat{x}(1) - x \|}{\| x \|}
\]
\[+ \left\{ \nu + 2(1 + \nu)\frac{\nu^2}{\rho} \left[ \| (\rho + A_1) \| \right. \right.
\]
\[+ 3\| \rho + A_1 \| \| \right\} .
\]

Replacing the relative error in the initial solution \( \| \hat{x}(1) - x \| / \| x \| \) with the bound of line (1.25) we obtain

\[
(1.42) \quad \frac{\| \hat{x}(2) - x \|}{\| x \|} \leq \delta(\nu)
\]

provided \( \nu \) is chosen proportional to \( h^2 \). By redefining \( R_{n,1} \) and \( R_{n,2} \) to represent the relative error introduced in both the solution step and the purification step, that is, redefining \( R_{n,1} \) and \( R_{n,2} \) so that \( (I + R_{n,1})(\rho + A_1)^{-1} \) maps \( \hat{x}(2) \), the purified solution, onto the right-hand side of the system \( (\rho + A_1)x = z \), we obtain

**Theorem 3:** If the single optimal parameter alternating direction technique for the Dirichlet problem (1.1) is executed in residual form in single precision floating-point
arithmetic where

(i) \( \gamma \) is taken proportional to \( h^2 \), the uniform mesh size ,
(ii) the residual is computed in \( fL_2 \)-arithmetic ,
(iii) a single step of the purification of Theorem 2 is performed after the numerical solution of each of the tridiagonal systems of (1.13),

then a fixed bound may be maintained for the \( L^2 \)-norm of the global rounding error uniformly in \( n \), the iteration step.

**Proof:** The proof follows immediately from the observation that with the improved bound of line (1.42), we may now take \( \delta \), the uniform bound on the 2-norm of the perturbation in the propagator of the rounding error, to be \( \Theta[\frac{\gamma}{h}] \) instead of \( \Theta[\frac{\gamma}{h^2}] \) . Of course, there is one more thing to establish. We must show that the local errors are uniformly bounded in \( L^2 \)-norm.

An examination of line (1.21) reveals that the \((n+1)\)st local error depends on the perturbations \( \theta_n \), \( \Gamma_n \), \( \delta A_n \) and \( \Xi_n \) as well as the exact \((n+1)\)st iterate and the exact \(n\)th residual \( A^{(n-1)}u - b \). The consistency and stability of the exact iteration for (1.1) permits the deduction that the exact iterates converge in \( L^2 \)-norm to an \( \Theta(h^2) \) approximation of the solution of (1.1) for smooth initial guess. Similarly, the residuals \( h^{-2}(A^{(n-1)}u - b) \) tend to zero in \( L^2 \)-norm. (Recall that we have scaled the finite difference
system by $h^2$.) Hence, the first term on the right-hand side of (1.21) behaves in $L^2$-norm like $\Theta(\nu)$, the second term no worse than $\Theta(\nu h)$ for $\nu$ proportional to $h^2$. The uniform bound

$$\|e^{(n)}\|_{L^2} \leq \Theta(\nu)$$

follows.

We note that if the residual had not been computed in $f_l^2$-arithmetic, the terms $\delta A_n$ and $\Gamma_n$ appearing in the last factor on the right-hand side of the expression for the local error in line (1.20) would have 2-norm bounds of the form $\Theta(\nu)$, forcing us to accept a bound on the $L^2$-norm of the local error of the form $\Theta(\frac{\nu}{h})$.

Finally, to obtain a bound on the $L^2$-norm of the global rounding error independent of $n$, the step number, we return to line (1.22). By choosing $\nu$ proportional to $h^2$ we have assured that the perturbation in the error propagator for the rounding error has a 2-norm bounded uniformly by a quantity behaving no worse than $\Theta(h)$; in fact, this perturbation is dominated in 2-norm by the $\Theta(h)$ term in the bound on the 2-norm of $T_\rho$. Hence the denominator in line (1.22) behaves like $\Theta(h)$ and the numerator like $\Theta(\nu)$ uniformly in $n$. Consequently, we have the uniform bound

$$\|r^{(n)}\|_{L^2} \leq \Theta(\frac{\nu}{h}) .$$
The bound given in line (1.44) is necessarily a small one for \( \nu \) proportional to \( h^2 \) and small \( h \). If \( f_{\nu_2} \)-arithmetic had not been employed to accumulate the residuals, we would have obtained a bound above of the form \( \Theta \left[ \frac{\nu}{h^2} \right] \) which is small only for a small ratio of \( \nu \) to \( h^2 \) regardless of the size of \( h \).
Section 2: The Higher Dimensional Case

For the Dirichlet problem (1.1) in $\mathbb{R}^p, p > 2$, the analogue of the approximate inverse of the Peaceman-Rachford 2-dimensional alternating direction iteration is given by

\begin{equation}
B_{p, p}^{-1} = 2\rho (p-1) \prod_{i=0}^{p-1} (\rho + A_{p-i})^{-1}, \quad p \geq 2
\end{equation}

for the splitting

\begin{equation}
A = \sum_{i=1}^{p} A_i
\end{equation}

of the finite difference operator $A$ into spatial components. The stationary iteration with approximate inverse $B_{p, p}^{-1}$ is known as the Douglas alternating direction method. Although it was not emphasized in Section 1, the Peaceman-Rachford technique is not the only possible 'alternating direction' iteration for 2-dimensional self-adjoint Dirichlet problems. Moreover, Vanderschel [13] and Guittet [8] have shown that for real $k$

\begin{equation}
T_{\rho, k} = I - k\rho B_{p-1}^{-1} \prod_{i=0}^{p-1} (\rho + A_{p-i})^{-1} A, \quad p \geq 2
\end{equation}

defines an infinite family of stationary alternating direction error propagators. Some of the members of this family are well known: for $p = 2$, for instance, the value $k = 2$ corresponds to the Peaceman-Rachford method; $k = 2 - \tau$ to Varga's 2-dimensional method [14]. For $p > 2$, the value $k = 2$ corresponds to the Douglas method; $k = 1$ to the
Douglas-Rachford method. Indeed, (1.47) is an economical way to write these techniques. Independently, Vanderschel and Guittet established that the value of \( k \) which yields the error propagator of smallest spectral radius approaches 2 asymptotically. Thus, the Peaceman-Rachford iteration for \( p = 2 \) and the Douglas iteration for \( p > 2 \) are nearly optimal methods among those of the family (1.47). We were not remiss, then, in having analyzed only the Peaceman-Rachford alternating direction technique in Section 1.

In this section we shall extend the analysis of Section 1 to the \( p \)-dimensional Douglas alternating direction iteration for the commutative case. Although both Vanderschel and Guittet have given very elegant analyses of the choice of the optimal iteration parameter for all of the error propagators of the family (1.47), the author offers his own analysis of this choice for the near optimal case \( k = 2 \).

**Theorem 4:** The optimal iteration parameter, \( \rho_{\text{opt}} \), for the stationary \( p \)-dimensional Douglas alternating direction iteration for the commutative case for equal spectral intervals is given by

\[
(1.48) \quad \rho_{\text{opt}} = b \left[ \frac{a}{b} \right]^p, \quad p \geq 2
\]

where \( a, b \) contains these intervals. For this value of \( \rho \),

\[
(1.49) \quad \| T_{p,2} \|_2 \leq 1 - 2p \left[ \frac{a}{b} \right]^p,
\]
provided that $\frac{a}{b} < \left[ \frac{1}{p} \right]^p$. (We shall henceforth denote the spectral set of $A_i$ by $\sigma(A_i)$.)

Proof: Although the spectral intervals of the $A_i$ are not in general identical, they may be made so by a simple change of variables (see [16]); hence there is no loss of generality in assuming that $\sigma(A_i) \subseteq [a,b]$ for all $i$. It is clear from line (1.47) that the eigenvalues of the Douglas alternating direction error propagator are given by

$$
\lambda_{i_1, i_2, \ldots, i_p} = 1 - 2 \frac{\gamma \left[ \lambda_{i_1} + \ldots + \lambda_{i_p} \right]}{\prod (1 + \gamma \lambda_{i_j})} ,
$$

where $\gamma = 1/p$. In order to determine an optimal value for $\gamma$, we employ the standard device of embedding the problem of minimizing the maximum of the eigenvalues described in (1.50) into the continuous minimax problem

$$
\min_{\gamma > 0} \max_{[a,b]} \left[ 1 - \frac{2 \gamma \sum x_{i_j}}{\prod (1 + \gamma x_{i_j})} \right] , \quad \sigma(A_i) \subseteq [a,b] , \quad i = 1, \ldots, p .
$$

The minimax problem posed in (1.51) can be easily solved once the extremal behavior of the quotient in the brackets above is determined as a function of $\gamma$. If we can show that

$$
\frac{\partial}{\partial x_{i_j}} \frac{\gamma \sum x_{i_j}}{\prod (1 + \gamma x_{i_j})}
$$

has one sign in $[a,b]$, then we shall
be able to deduce that as a function of $x_{i_j}$, \( \frac{\gamma \sum x_{i_j}}{\prod (1 + \gamma x_{i_j})} \) must take on its extrema at the end-points a or b of the interval. That this is indeed the case follows from an examination of \( \frac{\partial}{\partial x_{i_j}} \ln \frac{\gamma \sum x_{i_j}}{\prod (1 + \gamma x_{i_j})} \). This derivative is given by

\[
(1.52) \quad \frac{\partial}{\partial x_{i_j}} \ln \frac{\gamma \sum x_{i_j}}{\prod (1 + \gamma x_{i_j})} = \frac{\prod (1 + \gamma x_{i_k})}{\gamma \sum x_{i_k}} \frac{\partial}{\partial x_{i_j}} \frac{\gamma \sum x_{i_k}}{\prod (1 + \gamma x_{i_k})} \\
= \frac{\prod (1 + \gamma x_{i_j})}{\gamma \sum x_{i_j}} \left[ \frac{\frac{\partial}{\partial x_{i_j}} (\prod (1 + \gamma x_{i_k})) \gamma \sum x_{i_k}}{\prod (1 + \gamma x_{i_k})^2} - \frac{\frac{\partial}{\partial x_{i_j}} (\sum x_{i_k}) \prod (1 + \gamma x_{i_k})}{\prod (1 + \gamma x_{i_k})^2} \right] \\
= \frac{1}{(1 + \gamma x_{i_j})} \sum_{k \neq j} x_{i_k} - 1 \right].
\]

Since the numerator of the expression above does not depend on $x_{i_j}$, the derivative of \( \frac{\gamma \sum x_{i_k}}{\prod (1 + \gamma x_{i_k})} \) is independent of $x_{i_j}$, and is a one-signed function of $x_{i_j}$.

Consequently, the minimax problem (1.51) reduces to a consideration of the extrema of the family of functions:

\[
(1.53) \quad \left\{ \frac{ia_\gamma + (p - i)b_\gamma}{(1 + \gamma a)^i(1 + \gamma b)^{p-1}} \right\}, \quad i = 0, 1, 2, \ldots, p,
\]

in the variable $\gamma$. Optimal $\gamma$ is that value which maximizes the minimum of the lower envelope of the family, for this
value will serve to minimize the maximum of \(1 - 2\gamma \frac{\sum x_{ij}}{\Pi(1 + \gamma x_{ij})}\).

It is immediate that (i) all of the functions in the family are zero when \(\gamma = 0\), (ii) all are bounded above, and (iii) all tend to zero asymptotically with increasing \(\gamma\). The "geometry" of this family is depicted in a simple case in Figure 1 and serves to motivate the lemma which follows.

**Lemma 1:** Consider the one parameter family of curves

\[
(1.54) \quad \left\{ \frac{ia\gamma + (p-i)b\gamma}{(1 + a\gamma)^{i}(1 + b\gamma)^{p-i}} \right\}, \quad \gamma > 0.
\]

For \(a/b < \left[\frac{1}{p}\right]^p < 1 < b\), the value of \(\gamma\) which minimizes the lower envelope corresponds to the intersection of the two curves with extreme indices, namely, the \(i = 0\) and \(i = p\) curves. Its value is given by

\[
(1.55) \quad \gamma_{opt} = \frac{1}{b^p - a^p} = \frac{1}{b^{[a/b]^p}}, \quad a/b << \left[\frac{1}{p}\right]^p.
\]

**Proof:** The argument is straightforward. For \(\gamma < \gamma_{opt}\) we shall show that the curve \(\frac{pa\gamma}{(1 + a\gamma)^p}\) lies below every other member of the family (1.54), but that for \(\gamma > \gamma_{opt}\) the curve \(\frac{pb\gamma}{(1 + b\gamma)^p}\) lies below every other member of (1.54) assuming that \(a/b < \left[\frac{1}{p}\right]^p\). Finally, we shall show that the lower
$n = 5$

$\frac{a}{b} = \frac{1}{40}$

**Figure 1**
envelope of the family which must consist of the \( i = p \) curve for \( \gamma < \gamma_{\text{opt}} \) and the \( i = 0 \) curve for \( \gamma > \gamma_{\text{opt}} \) reaches a global maximum at \( \gamma = \gamma_{\text{opt}} \).

Consider the first assertion:

\[
(1.56) \quad \frac{\gamma(ia + (p-i)b)}{(1+\gamma a)^i(1+\gamma b)^{p-i}} \geq \frac{p\gamma}{(1+\gamma a)^p} , \quad i = 0, 1, \ldots, p-1
\]

for \( \gamma < \gamma_{\text{opt}} \). To verify (1.56) we need only show that (i) the \( i = p \) curve does not lie above any other member of (1.54) in some neighborhood of the origin and (ii) the intersection of the \( i = p \) curve with any other member of the family never occurs to the left of \( \gamma_{\text{opt}} \).

We shall establish (i) by showing that the curves in the family have slopes that decrease with increasing \( i \) at the origin. Since all curves begin at the origin this is enough to justify our claim. The slope of the \( i^{\text{th}} \) curve in the family is given by

\[
(1.57) \quad \frac{d}{d\gamma} \left( \frac{\gamma(ia + (p-i)b)}{(1+\gamma a)^i(1+\gamma b)^{p-i}} \right) = \frac{(1+\gamma a)^i(1+\gamma b)^{p-i}(ia+(p-i)b)}{(1+\gamma a)^{2i}(1+\gamma b)^{2(p-i)}}
\]

\[-(p-i)b(1+\gamma a)^i(1+\gamma b)^{p-i-1}(\gamma(ia+(p-i)b)}{(1+\gamma a)^{2i}(1+\gamma b)^{2(p-i)}
\]

\[\quad ia(1+\gamma a)^{i-1}(1+\gamma b)^{p-i}(\gamma(ia+(p-i)b)}{(1+\gamma a)^{2i}(1+\gamma b)^{2(p-i)}}
\]
At $\gamma = 0$ the last two terms on the right vanish and the first term simplifies to $ia + (p-i)b$. The values of this derivative are, therefore, ordered inversely with increasing $i$ provided that $\frac{a}{b} < 1$, which is a fortiori satisfied here.

For (ii) we offer the following argument. The intersections of the $i = p$ curve with the other members of the family occur whenever

$$
(1.58) \quad \frac{\gamma(ia + (p-i)b)}{(1+\gamma a)^{\frac{1}{p}}(1+\gamma b)^{p-1}} = \frac{\gamma p a}{(1+\gamma a)^p}, \quad i = 0,1,\ldots,p-1;
$$

hence the value of $\gamma$ corresponding to the intersection of these two curves is given by

$$
(1.59) \quad \gamma = \frac{\frac{ia + (p-i)b}{pa^{p-1}} - 1}{b - a\left[\frac{ia + (p-i)b}{pa^{p-1}}\right]} , \quad i = 0,1,\ldots,p-1.
$$

We must show that for each $i \in \{0,1,\ldots,p-1\}$ the quantity $\gamma$ above is at least as large as $\gamma_{\text{opt}} = \frac{b^{p-1}}{a^{p-1}b - b^{p-1}a}$. 

We begin by noting that

$$
(1.60) \quad \left[\frac{b}{a}\right]^{\frac{p-i}{p}} - 1 = \left[\frac{b}{a}\right]^{\frac{i}{p}} = \left[\frac{a}{b}\right]^{\frac{i}{p}} < \left[\frac{a}{b}\right]^{\frac{1}{p}} < \frac{1}{p} < \left[\frac{p-i}{p}\right], \quad i \neq 0,p,
$$
from which follows the inequality

\[(1.61) \quad \left[ \frac{b}{a} \right]^p \left[ \frac{1}{p-i} \right] \frac{1}{p-1} < \left[ \frac{1}{p} + \frac{p-i}{p} \frac{b}{a} \right] \frac{1}{p-1}, \quad i \neq 0, p.\]

Subtracting the extreme members of this inequality from \(b\) yields

\[(1.62) \quad b - a \left[ \frac{b}{a} \right]^p > b - a \left[ \frac{1}{p} + \frac{p-i}{p} \frac{b}{a} \right] \frac{1}{p-1}, \quad i \neq 0, p.\]

Multiplying both sides of (1.62) by \(\left[ \frac{1}{b^p} - \frac{1}{a^p} \right]\) yields

\[(1.63) \quad \left[ \frac{1}{b^p} - \frac{1}{a^p} \right] \left[ b - a \left[ \frac{b}{a} \right]^p \right]
\[> \left[ \frac{1}{b^p} - \frac{1}{a^p} \right] \left[ b - a \left[ \frac{1}{p} + \frac{p-i}{p} \frac{b}{a} \right] \frac{1}{p-1} \right], \quad i \neq 0, p.\]

The left-hand side of (1.63) may be rewritten as

\[(1.64) \quad \left[ \frac{1}{b^p} - \frac{1}{a^p} \right] \left[ b - a \left[ \frac{b}{a} \right]^p \right] = \left[ \left[ \frac{b}{a} \right]^p - 1 \right] \left[ \frac{1}{a^p} b - ab^p \right],
\]

hence

\[(1.65) \quad \left[ \frac{b}{a} \right]^p - 1 \left[ \frac{1}{a^p} b - ab^p \right]
\[> \left[ \frac{1}{b^p} - \frac{1}{a^p} \right] \left[ b - a \left[ \frac{1}{p} + \frac{p-i}{p} \frac{b}{a} \right] \frac{1}{p-1} \right], \quad i \neq 0, p.\]

From (1.61) we perceive that by replacing \(\left[ \frac{b}{a} \right]^p\) with

\[\left[ \frac{1}{p} + \frac{(p-i)}{p} \frac{b}{a} \right] \frac{1}{p-1}, \quad i \neq 0, p\] in (1.65) we obtain
\[(1.66) \quad \left[ \frac{1}{p} + \frac{p-i}{p} \left( \frac{1}{p} \right)^{p-1} \right] \left( \frac{1}{a^p b} - \frac{1}{a b^p} \right) \]

\[> \left( \frac{1}{b^p - a^p} \right) \left[ b - a \left( \frac{1}{p} + \frac{p-i}{p} \left( \frac{1}{p} \right)^{p-1} \right) \right], \quad i \neq 0, p. \]

This last result is precisely the stipulation that \( \gamma \) of line (1.59) is at least as large as \( \gamma_{\text{opt}} \), provided that \( \frac{a}{b} < \left( \frac{1}{p} \right)^p \).

We shall now show that the curve \( \frac{\gamma_{\text{pb}}}{(1 + \gamma b)^p} \) lies below every other member of the family for \( \frac{a}{b} < \left( \frac{1}{p} \right)^p \). We have already noted that this curve does not lie below any other curve in the family in some right-handed neighborhood of the origin. To complete the argument, then, we need only prove that the intersections of this curve with other members of the family never occur to the right of \( \gamma_{\text{opt}} \). The intersections of the \( j = 0 \) curve with the \( j = 1 \) curve corresponds to \( \gamma \) given by

\[(1.67) \quad \gamma = \frac{1 - \left( \frac{ia + (p-i)b}{pb} \right)^{1/1}}{b \left( \frac{ia + (p-i)b}{pb} \right)^{1/1} - a}, \quad i \neq 0 . \]

We must show that \( \gamma \) is no larger than \( \gamma_{\text{opt}} \) for \( \frac{a}{b} < \left( \frac{1}{p} \right)^p \). At the outset we note that if \( \frac{a}{b} < \left( \frac{1}{p} \right)^p \), then
(1.68) \[ \left[ \frac{1}{a} \right]^p < \frac{1}{p} < \left[ \frac{1}{1} \right]^1 < \left[ \frac{p-i}{p} \right]^1 < \left[ \frac{p-i + \frac{i}{p} \frac{a}{b}}{p} \right]^1, \quad i \neq 0. \]

Subtracting \( a \) from the extreme terms of (1.68) yields

(1.69) \[ b \left[ \frac{i}{p} \frac{a}{b} + \left( \frac{p-i}{p} \right)^1 \right] - a \geq \left[ \frac{a}{b} \right]^p b - a, \quad i \neq 0. \]

Multiplying both sides of (1.69) by \( b^p - a^p \) produces

(1.70) \[
\left[ \frac{1}{b^p - a^p} \right] \left[ b \left[ \frac{i}{p} \frac{a}{b} + \left( \frac{p-i}{p} \right)^1 \right] - a \right] \\
\geq \left[ \frac{1}{b^p - a^p} \right] \left[ \frac{1}{a^p} \right] b - a, \quad i \neq 0.
\]

The right-hand side of (1.70) may be rearranged as follows

(1.71) \[
\left[ \frac{1}{b^p - a^p} \right] \left[ \frac{1}{a^p} \left[ a \left[ a \frac{1}{b} \right]^p b - a \right] \right] = \left[ 1 - \left[ \frac{a}{b} \right]^p \right] \left[ \frac{1}{a^p} b - a \right] \frac{1}{b^p} - \frac{1}{a^p};
\]

hence

(1.72) \[
\left[ 1 - \left[ \frac{a}{b} \right]^p \right] \left[ \frac{1}{a^p} \left[ a \left[ a \frac{1}{b} \right]^p b - a \right] \right] \geq \left[ 1 - \left[ \frac{i}{p} \frac{a}{b} + \left( \frac{p-i}{p} \right)^1 \right] \right] \left[ \frac{1}{a^p} b - a \right] \frac{1}{b^p} - \frac{1}{a^p}, \quad i \neq 0.
\]

Combining (1.70) and (1.72), we obtain the result

(1.73) \[
\left[ \frac{1}{b^p - a^p} \right] \left[ b \left[ \frac{i}{p} \frac{a}{b} + \left( \frac{p-i}{p} \right)^1 \right] - a \right] \\
\geq \left[ 1 - \left[ \frac{i}{p} \frac{a}{b} + \left( \frac{p-i}{p} \right)^1 \right] \right] \left[ \frac{1}{a^p} b - a \right] \frac{1}{b^p} - \frac{1}{a^p}, \quad i \neq 0.
Dividing both sides of (1.73) by
\[
\left[ \frac{1}{a^p b - ab^p} \right] \left[ b \left[ \frac{i}{p} \frac{a}{b} + \frac{(p-i)}{p} \right] \right] \frac{1}{I} - a \]
yields

\[
\frac{1}{b^p - a^p} \frac{1}{a^p b - ab^p} \leq \frac{1 - \left[ \frac{i}{p} \frac{a}{b} + \frac{(p-i)}{p} \right]}{b \left[ \frac{i}{p} \frac{a}{b} + \frac{(p-i)}{p} \right] - a} I, \quad i \neq 0,
\]

which is precisely the requirement that \( \gamma_{\text{opt}} > \gamma \).

We now turn our attention to the problem of establishing that \( \gamma_{\text{opt}} \) is the largest value that the composite minimal curve consisting of the \( i = p \) curve for values of \( \gamma < \gamma_{\text{opt}} \) and the \( i = 0 \) curve for all \( \gamma > \gamma_{\text{opt}} \) attains. We first observe that the derivative of the \( i = p \) curve is given by

\[
\frac{d}{d\gamma} \frac{\gamma p a}{(1 + \gamma a)^p} = \frac{p a (1 + \gamma a)^p - p^2 a^2 (1 + \gamma a)^{p-1}}{(1 + \gamma a)^2 p}
\]
is positive for \( 0 < \gamma < \frac{1}{a(p-1)} \). Since \( \frac{a}{b} < \left[ \frac{1}{p} \right]^p \), we deduce that

\[
\frac{1}{p-1} > \left[ \frac{1}{p-1} \right]^{p-1} > \left[ \frac{1}{p} \right]^{p-1} > \left[ \frac{a}{b} \right] \frac{p-1}{p}
\]

hence

\[
\frac{1}{a(p-1)} > \frac{1}{b \left[ \frac{a}{b} \right]} = \gamma_{\text{opt}}.
\]
The derivative of the \( i = 0 \) curve is identical to that of the \( i = p \) curve if \( a \) is replaced by \( b \). It is clear that this curve is initially increasing but reaches a global maximum at \( \gamma = \frac{1}{b(p-1)} \), then decreases to 0 asymptotically. Since \( \left| \frac{a}{b} \right|^p < \frac{1}{p} < p-1 \), it follows that

\[
\frac{1}{b(p-1)} < \frac{1}{b \left[ \frac{a}{b} \right]^p} = \gamma_{\text{opt}},
\]

that is, the global maximum of the \( i = 0 \) curve lies to the left of \( \gamma_{\text{opt}} \). This completes the argument.

Having completed the proof of the lemma we can return to the proof of the theorem. With \( \gamma_{\text{opt}} \) we shall now obtain a bound on the 2-norm of the Douglas alternating direction iterator. Since \( \gamma_{\text{opt}} \) lies on the curve \( \frac{\gamma^p a}{(1+\gamma a)^p} \), which as we have already remarked, serves as a lower envelope for the family in the interval \([0, \gamma_{\text{opt}}]\), we can compute

\[
\max_{\gamma} \min_{i} \left\{ \frac{\gamma (ia + (p-i)b)}{(1+\gamma a)^i(1+\gamma b)^{p-i}} \right\} = \frac{p a \gamma_{\text{opt}}}{(1 + a \gamma_{\text{opt}})^p}
\]

\[
= p \left[ \frac{a}{b} \right]^{1 - \frac{1}{p}},
\]

from which we conclude that

\[
(1.80) \quad \|T_{\gamma_{\text{opt}}}\|_2 \leq 1 - 2p \left[ \frac{a}{b} \right]^{1 - \frac{1}{p}}.
\]

This completes the proof of the theorem.
Asymptotic Rounding Error Analysis of the Higher Dimensional Technique

In floating point single precision arithmetic the residual Douglas alternating direction iteration is realized as

\[(1.81) \quad \hat{u}^{(n)} = (I + \theta_n) \{ \hat{u}^{(n-1)} - 2^{p-1}(I + \hat{\delta}_n) \hat{B}_\rho^{-1} (I + \hat{\psi}_n)[(A + \delta A_n) \hat{u}^{(n-1)} - (I + \hat{\xi}_n)b] \}\]

The matrix \( \hat{B}_\rho^{-1} \) is a backward error analysis analogue of \( B_\rho^{-1} \) (depending on the strategy of computation) that is, the operator which maps \( f_{\ell_2}(A\hat{u}^{(n-1)} - b) \) exactly onto \( (I + \hat{\psi}_n) \hat{d}^{(n-1)} \hat{d}^{(n-1)} \) where \( \hat{d}^{(n-1)} \) is the computed difference correction. The perturbations \( \theta_n, \hat{\delta}_n, \hat{\psi}_n, \) and \( \hat{\xi}_n \) are the analogues of those appearing in line (1.12) of Section 1. As in the single parameter case, the residual is assumed to be computed in \( f_{\ell_2} \)-arithmetic; hence both of the perturbations \( \delta A_n \) and \( \hat{\xi}_n \) have max-norm bounds of the form \( o(2) \). For the commutative case in p-dimensions the difference operator \( A \) will have at most \( 2p + 1 \) non-zero entries per row; consequently the computation of the residual will require the accumulation of inner products of length at most \( 2p + 2 \).

The difference correction may be accumulated a dimension at a time, so to speak, by the recursion
\[(\rho + A_1) d_1^{(n-1)} = Au^{(n-1)} - b,\]
\[(\rho + A_2) d_2^{(n-1)} = d_1^{(n-1)}, \]
\[\vdots \]
\[(\rho + A_p) d_p^{(n-1)} = d_p^{(n-1)}, \]

where \(d_p^{(n-1)} = \frac{1}{2\rho p^{-1}} d^{(n-1)}, \) the \(n^{th}\) difference correction. These steps are realized computationally as

\[\hat{d}_1^{(n-1)} = (I + R_{n,1})(\rho + A_1)^{-1} ft_2(Au^{(n-1)} - b),\]
\[\hat{d}_2^{(n-1)} = (I + R_{n,2})(\rho + A_2)^{-1} \hat{d}_1^{(n-1)}, \]
\[\vdots \]
\[\hat{d}_p^{(n-1)} = (I + R_{n,p})(\rho + A_p)^{-1} \hat{d}_{p-1}^{(n-1)}, \]

where \(R_{n,i}\) is the relative error in the solution of any linear system with coefficient matrix \((\rho + A_1)\) in single precision. Since \((\rho + A_1)\) may be tridiagonalized merely by reordering the components of \(\text{card}(\rho^{(0)}_n)\) we sacrifice no generality in treating a linear system with coefficient matrix \((\rho + A_1)\) as though it were tridiagonal insofar as rounding errors are concerned. Consequently, we shall employ Theorem 1 without concern for the location of the diagonals of \((\rho + A_1).\)
Subtracting the exact $n^{th}$ iterate from its single precision analogue yields

$$r^{(n)} = [T_\rho + \Delta_n]r^{(n-1)} + e^{(n)},$$

where

$$\Delta_n = \theta_n T_\rho - 2\rho^{p-1}(1 + \theta_n)\Lambda_n(A + \delta A_n)$$

$$- 2\rho^{p-1}(I + \theta_n)\prod_{i=0}^{p-1}(\rho + A_{p-i})^{-1}\delta A_n,$$

$$\Lambda_n = \phi_n \prod_{i=0}^{p-1}(\rho + A_{p-i})^{-1}(I + \varphi_n) + \prod_{i=0}^{p-1}(\rho + A_{p-i})^{-1}\varphi_n$$

$$- (I + \phi_n) \left\{ \prod_{i=0}^{p-1}(\rho + A_{p-i})^{-1} \right. $$

$$- \left. \prod_{i=0}^{p-1}(I + R_n, p-1)(\rho + A_{p-i})^{-1} \right\} (I + \varphi_n),$$

and finally

$$e^{(n)} = \theta_n u^{(n)} - 2\rho^{p-1}(1 + \theta_n)\Lambda_n(Au^{(n-1)} - b)$$

$$- 2\rho^{p-1}(I + \theta_n) \left[ \prod_{i=0}^{p-1}(\rho + A_{p-i})^{-1} + \Lambda_n \right] \left[ \delta A_n u^{(n-1)} - \varphi_n b \right].$$

To obtain a useful bound from the recursion of line (1.84), it is necessary to establish uniform bounds for the 2 and $L^2$-norms of $\Delta_n$ and $e^{(n)}$, respectively. Since $A_i$ is zero row sum for rows corresponding to interior mesh points, $(\rho + A_i)$ is diagonally dominant by no more than $\rho$; hence
Lemma 0 yields \( \|(\rho + A_1)^{-1}\| \leq \frac{1}{\rho} \). Also, assuming that 
A has been scaled so that no off-diagonal element exceeds unity in magnitude, we obtain the familiar bound from Theorem 1 for the relative error in the single precision solution of any linear system with \((\rho + A_1)\):

\[
(1.88) \quad \|R_{n,i}\|_2 \leq (15 + 2 \max_i \|(\rho + A_1)\|_2) \frac{v}{\rho} + o\left[\frac{v}{\rho}\right]^2,
\]

\[\bar{\rho} = \rho - 4[\|\text{diag}(A_1)\| + o(\rho^2)], \quad i = 1, \ldots, p, \quad \frac{v}{\bar{\rho}} < 1.\]

To bound the 2-norm of \( \Lambda_n \), we examine the difference

\[
(1.89) \quad \prod_{i=0}^{p-1} (I + R_{n,p-i})(\rho + A_{p-i})^{-1} - \prod_{i=0}^{p-1} (\rho + A_{p-i})^{-1}.
\]

With the convention \( \prod_{s=0}^{-1} = I \), this difference may be expressed as the sum

\[
(1.90) \quad \prod_{i=0}^{p-1} (1 + R_{n,p-i})(\rho + A_{p-i})^{-1} - \prod_{i=0}^{p-1} (\rho + A_{p-i})^{-1}
\]

\[= \sum_{i=1}^{p} \prod_{s=0}^{i-1} (1 + R_{n,p-s})(\rho + A_{p-s})^{-1} R_i \prod_{s=0}^{i-1} (\rho + \Lambda_{i-s})^{-1}.\]

From (1.90) we obtain the bound

\[
(1.91) \quad \|\prod_{i=0}^{p-1} (I + R_{n,p-i})(\rho + A_{p-i})^{-1} - \prod_{i=0}^{p-1} (\rho + A_{p-i})^{-1}\|_2
\]

\[\leq \frac{M}{\rho^p} \sum_{i=1}^{p} (1 + M)i \leq \frac{M}{\rho^p} \exp(1 + M)
\]

where \( M \) denotes the largest of the bounds of (1.88).
Thus
\[(1.92) \quad \|A_n\|_2 \leq \frac{\nu}{\rho^p} (1 + \nu) + \frac{\nu}{\rho^p} + (1 + \nu)^2 \frac{M}{\rho^p} \exp (1 + M),\]
and we may take \( \delta \), the uniform bound on \( \|A_n\|_2 \), to be
\[(1.93) \quad \delta = \nu (1 - 2p \left[ \frac{a}{b} \right]^p) + 4 \frac{\nu}{\rho} (1 + \nu)^3 \frac{M}{\rho^p} \exp (1 + M).\]

Returning to (1.84) we perceive that sufficient conditions for the existence of a uniform bound on the \( L^2 \)-norm of the global rounding error are that (i) \( \delta < 2p \left[ \frac{a}{b} \right]^{p-1} \) and (ii) \( \|e^{(n)}\|_2 \leq \Theta(\nu) \). For \( \frac{\nu}{\rho} < 1/(1 + 4 \left[ \max_i \|\text{diag}(A_i)\| + \Theta(\nu) \right]), \)
\( M = \Theta \left[ \frac{\nu}{\rho} \right] \); hence \( \nu \) must be chosen proportional to \( \left[ \frac{a}{b} \right]^{p+1} \) in order to satisfy constraint (i). We shall show that this choice of \( \nu \) also guarantees that the second constraint is satisfied. From the monotonicity argument of Section 1 it is clear that the condition number \( \frac{a}{b} \) satisfies
\[(1.94) \quad \frac{a}{b} \leq \left[ \frac{\min_i \bar{a}_i}{\max_i \bar{a}_i} \right] \tan^2 \left[ \frac{\pi h}{2} \right], \quad k: \text{ independent of } h\]
where the meanings of \( a_i \) and \( \bar{a}_i \) may be inferred from their analogues in Section 1. Thus, in terms of the mesh spacing, \( \nu \) must be chosen proportional to \( [h^2]^{p+1} \) to satisfy (i). If we can show that the local errors \( e^{(j)} \) are uniformly bounded in \( L^2 \)-norm, this choice of \( \nu \) will insure that the global rounding error is uniformly bounded in \( L^2 \)-norm.
Note that for the 2-dimensional case, a bound of the form $\Theta \left[ \frac{\nu}{h^3} \right]$ follows from the analysis above. This is the same result obtained without purification in Section 1 for the Peaceman-Rachford alternating direction technique. Perhaps the purification technique of Theorem 2 will yield an improved bound here.

Suppose we require that $\frac{\nu}{\rho^2}$ be fixed, then the relative error in each purified step of the accumulation of the difference correction satisfies

$$\frac{\|d_i^{(n)} - d_i^{(n)}\|_2}{\|d_i^{(n)}\|_2} \leq \Theta \left[ \frac{\nu}{\rho} \right]^2 = \Theta(\nu),$$

where $d_i^{(n)}$ has the meaning given in line (1.82) and $d_i^{(n)}$ is the result of the purification of $\hat{d}_i^{(n)}$ given in line (1.83). By redefining $R_{n,i}$ to include the purification step we may conclude that $M$, the uniform bound for the 2-norm of the redefined $R_{n,i}$ may be taken proportional to $\nu$ instead of $\frac{\nu}{\rho}$. Consequently, $\delta$, the uniform bound on the 2-norm of the perturbation in the error propagator for the rounding error, may be chosen proportional to $\nu$ instead of the earlier choice given in line (1.93). To satisfy condition (i) we need only take $\nu$ proportional to $p h^2$. Furthermore, even without purification the result we have just obtained is approached asymptotically as $p$ increases. This is in sharp contrast to the choice of $\nu$ required to assure the stability of the
direct solution with respect to the growth of rounding errors (see [12]). Also note that \( \frac{\nu}{\rho^2} \) is indeed bounded for \( \nu \) proportional to \( \rho h^2 \) provided that \( p \geq 2 \). For \( p > 2 \), this ratio tends to zero with \( h \).

Finally, a term by term examination of the expression in line (1.87) for the local error, keeping in mind the bound on \( \Lambda_n \) given in (1.92) and the facts that both \( u^{(n)} \) and \( h^{-2}(Au^{(n-1)}-b) \) are bounded in \( L^2 \)-norm independent of \( n \), yields the result of condition (ii) above, namely, that 
\[
\|e^{(n)}\|_{L^2} \leq \Theta(\nu).
\]
Note that even though \( \|\Lambda_n\|_{L^2} \leq \Theta\left[\frac{\nu}{\rho(\rho-1)}\right] \), the bound we obtain for the second term on the right in line (1.87) is no worse than \( \Theta\left[\frac{2-2}{\rho h^p}\right] \). Also recall that both \( \delta A_n \) and \( z_n \) are \( \Theta(\nu^2) \) perturbations. We summarize this discussion in

**Theorem 5:** If

(i) \( \nu \) is taken proportional to \( h^2 \) in sufficiently small ratio,

(ii) the residual at each step is computed in \( ft_2 \)-arithmetic,

(iii) the purification of Theorem 2 is employed after each step of the accumulation of the difference correction (see line (1.84)),
then the rounding error committed in the single precision execution of the Douglas alternating direction method in residual form for the commutative case in $\mathbb{R}^p$ satisfies an $L^2$-norm bound of the form

\[
(1.96) \quad \|r^{(n)}\|_{L^2} \leq c \left\{ \frac{\nu}{1 - 1/p} \right\}, \quad p \geq 2.
\]
Section 3: The Multiparameter Case in $\mathbb{R}^2$

In this section we shall consider the multiparameter, or m-cyclic, Peaceman-Rachford alternating direction method for the commutative case in $\mathbb{R}^2$ (see Section 1 for a characterization of this case). Two parameter sets will be considered: the Peaceman-Rachford 'good' parameters and the Wachspress 'good' parameters. Although optimal parameters exist, these sets are the ones most commonly employed because of the relative ease with which they may be computed. Moreover, for the model problem these parameters yield an almost optimal convergence rate.

In residual form the m-cyclic Peaceman-Rachford method may be written as

\[
(1.97) \quad d^{(n-1)} = -2\rho_i(\rho_i + A_2)^{-1}(\rho_i + A_1)^{-1}(Au^{(n-1)} - b) ; \\
i = 1, 2, \ldots, m.
\]

As in the previous sections we shall assume that $A$ has been scaled to insure that no off-diagonal element exceeds unity in magnitude and we shall treat both $A_1$ and $A_2$ as tridiagonal matrices insofar as solutions of systems implied by $(\rho_i + A_j)^{-1}$ are concerned, recognizing that both may be reduced to tridiagonal form by similarity transformations with permutation matrices. The Peaceman-Rachford 'good' parameters are given by

\[
(1.98) \quad \rho_i = b \left[ \frac{a}{b} \right]^{2i-1} \quad \frac{2m}{2} \quad i = 1, 2, \ldots, m
\]
where the interval \([a, b]\) contains the union of the spectral intervals of \(A_1\) and \(A_2\). If we define

\[
y(n) = u((n-1)m+1),
\]

\[
r(n) = Ay(n) - b,
\]

and

\[
B_mA = I - \prod_{i=0}^{m-1} T_{\rho_{m-i}}
\]

where

\[
T_{\rho_j} = I - 2\rho_j(\rho_j + A_2)^{-1}(\rho_j + A_1)^{-1}A,
\]

then we may rewrite (1.97) as

\[
y(n) = y(n-1) - B_mr(n-1).
\]

Consequently, we are justified in viewing the multiparameter technique as a stationary residual iteration with approximate inverse \(B_m\) given by

\[
B_m = 2\rho_1(\rho_1 + A_2)^{-1}(\rho_1 + A_1)^{-1}
\]

\[
+ 2\rho_2(\rho_2 + A_2)^{-1}(\rho_2 + A_1)^{-1}T_{\rho_1} + \ldots
\]

\[
+ 2\rho_m(\rho_m + A_2)^{-1}(\rho_m + A_1)^{-1}\prod_{i=1}^{m-1} T_{\rho_{m-i}}.
\]

We shall investigate two strategies for the computation of an \(m\)-cycle. In the first strategy the role of the residual will be emphasized: we shall arrange the computation in such a way that a difference correction is computed at each step of the cycle. Specifically, we shall compute the \(n^{th}\)
cycle by the recursion
\[
\beta(0) = u^{(n-1)}
\]
\[
\beta(1) = \beta(0) + \delta_1 (A\beta(0) - b)
\]
\[
\vdots
\]
\[
u(n) = \beta(m) = \beta(m-1) + \delta_m (A\beta(m-1) - b)
\]

where by our notation we relegate the intermediate steps in the cycle to the role of 'fractional' steps, that is, we imagine that the computation proceeds from \(u^{(n-1)}\) to \(u^{(n)}\) via \(m\) fractional steps instead of taking the view that the \(m\) steps of the cycle advance the iteration from \(u^{(n-1)}\) to \(u^{(n+m-1)}\). The symbol \(\delta_i\) represents what we shall call the **local approximate inverse**

\[
(1.106) \quad \delta_i = -2\rho_i (\rho_i + A_2)^{-1} (\rho_i + A_1)^{-1} .
\]

In the second strategy a new residual will not be computed with each step of the cycle; instead the approximate inverse will be accumulated recursively:

\[
w(i) = w(i-1) + A\beta(i-1)
\]
\[
\beta(i) = -2\rho_i (\rho_i + A_2)^{-1} (\rho_i + A_1)^{-1} w(i) , i = 1, 2, \ldots, m.
\]
\[
\delta_{n-1}^i = \beta(i) + \delta_{n-1}^{i-1}
\]

This is initialized by taking \(w(0) = (Au^{(n-1)} - b)\), \(\beta(0) = 0\), and \(\delta_{n-1}^0 = 0\).
It is important to recognize that the cycle length \( m \) cannot be chosen independently of \( h \) for either of the 'good' parameter sets if the technique is to be effective. It is well known that the error propagator of the \( m \)-cyclic alternating direction technique satisfies

(1.108) \[
    s \left( \prod_{i=0}^{m-1} \tau_i \right) \leq \max_{u \in [a,b]} \prod_{i=1}^{m} \left| \frac{\mu - \rho_i}{\mu + \rho_i} \right|^2
\]

where as before \( s(\cdot) \) denotes the spectral radius of its argument. Moreover, it may be shown that

(1.109) \[
    \max_{\mu \in [a,b]} \prod_{i=1}^{m} \left| \frac{\mu - \rho_i}{\mu + \rho_i} \right| \leq \frac{1 - c^{1/2m}}{1 + c^{1/2m}}, \quad c = \frac{a}{b}
\]

for the Peaceman-Rachford 'good' parameters (see Birkhoff, et al [2]). The \( \frac{2}{m} \) th root of the bound in (1.109), whose limit with increasing \( m \) bounds the asymptotic reduction in 2-norm of the \( m \)-cyclic Peaceman-Rachford technique for these parameters, does not decrease strictly with increasing \( m \) as is the case for optimal parameters, but decreases to a certain point and then increases.

We can discover where, approximately, this change occurs by treating \( m \) as a continuous variable. Such an argument was given by Douglas and is described in Birkhoff, et al [2]. We shall give a variation of it here. Taking \( 1/2m \) as the continuous variable instead of \( m \), we shall find it convenient to consider not the majorizing function \( \frac{1 - c^{1/2m}}{1 + c^{1/2m}} \) but
the logarithm of this function. Since the logarithm is convex, both functions will attain minima for the same values of \( x = \frac{1}{2m} \). Differentiating the logarithm of \( \left[ \frac{1 - c^x}{1 + c^x} \right]^{2x} \) and equating this derivative to zero, we obtain

\[
(1.110) \quad \frac{4x c^x \log(c)}{(1 + c^x)^2} \left[ \frac{1 + c^x}{1 - c^x} \right] - 2 \log \left[ \frac{1 - c^x}{1 + c^x} \right] = 0
\]

which after rearrangement becomes

\[
(1.111) \quad \left[ \frac{1 - c^{2x}}{2} \right] \log \left[ \frac{1 - c^x}{1 + c^x} \right] = c^x \log c^x.
\]

Solving for \( c^x \) we obtain

\[
(1.112) \quad c^x = 2^{\frac{1}{2}} - 1,
\]

from which it follows that

\[
(1.113) \quad m = \frac{1}{2} \log (c)/\log (\sqrt{2} - 1).
\]

For the model problem

\[
(1.114) \quad s \left( \prod_{i=0}^{m-1} T_{\rho_{m-i}} \right) \sim 1 - 4 \left[ \frac{\tau h}{2} \right] \log \tan \left( \frac{\pi h}{2} \right).
\]

**Asymptotic Rounding Error Analysis**

We shall give asymptotic rounding error analyses of both of the strategies (1.103), (1,105) for the computation of the multiparameter iteration scheme. The first question to be considered is: What effect do the small parameters have on numerical stability with respect to rounding errors?
From (1.98) it is clear that the smallest parameter is near \( c = \Theta(h^2) \). However, our study of the single parameter case shows that a small parameter \( \rho \) tends to worsen the condition of the problem implied by the factor \((\rho_i + A_j)^{-1}\) of the local approximate inverse. Specifically, recall that in Theorem 1 the constraint \( \nu/\rho = \Theta(1) \) was needed to obtain an \( \Theta(\frac{\nu}{\rho}) \) bound on the relative error in the single precision solution (by Gaussian elimination) of \((\rho + A_j)x = y, j = 1,2\). Moreover, for the single parameter case \( \rho_{opt} = \Theta(h) \) revealing the need for purification of such a solution. In the multiparameter case the situation is worse since some of the parameters are almost as small as \( c \). Thus, the computed solution of \((\rho_i + A_j)x = y, j = 1,2\) could have a single precision relative error bound almost a power of \( h \) larger than in the single parameter case. Nor is this an artificial result, for the smallest eigenvalue of \( A_j \) is proportional to \( h^2 \). We must consider, then, how to compute the approximate inverse so as to guarantee stability with respect to rounding error growth for \( \nu \) proportional to \( c \). We expect to require at least this precision to recognize the presence of \( \rho_i \) when added to the \( \Theta(1) \) elements of \( A_j \), \( j = 1,2 \).

Two solutions to the problem posed above come to mind. On the one hand, the numerical solutions of the systems with \((\rho_i + A_j)\) could be performed in \( f_{l_2} \)-arithmetic. Since these systems are not triangular their solution by Gaussian
elimination requires factorization into LU form; hence it would be necessary to store an intermediate vector of length \( o(h^{-2}) \) in double precision. We would obtain, however, a bound for the relative error in the solution of \((\rho_i + A_j)x = y\) no worse than \( o[\nu^2/c^{1-\frac{1}{2m}}] \) for any Peaceman-Rachford parameter \( \rho_i \).

On the other hand, the purification technique of Theorem 2 avoids the additional storage problem and the cost of extensive double precision arithmetic if it is applicable. It is clear from a consideration of the smallest Peaceman-Rachford parameter that a single purification will not yield an adequately improved bound on the relative error. The question becomes, then: How many purifications are required to insure that the accumulation of the \( i \)th local approximate inverse, \( \hat{\beta}_i = -2\rho_i(\rho_i + A_2)^{-1}(\rho_i + A_1)^{-1} \) is stable with respect to rounding error growth, provided \( \nu \) is proportional to \( c \)?

To answer this question we must examine the details of the computation of this local approximate inverse. As in the single parameter case, we actually compute not
\[-2\rho_i(\rho_i + A_2)^{-1}(\rho_i + A_1)^{-1} \]
but
\[(1.115) \quad \hat{\beta}_i = -2\rho_i(I + \theta_{i,n})(I + R_{n,i,2})(\rho_i + A_2)^{-1} \]
\[\cdot (I + R_{n,i,1})(\rho_i + A_1)^{-1} ,\]
where the perturbation \( R_{n,m,j} \) is a measure of the relative accuracy of the solution implied by \( (\rho + A)^{-1} \), that is,

\[
(I + R_{n,i,j})(\rho + A)^{-1} = [(\rho + A) + K_{n,i,j}]^{-1}.
\]

Rearranging line (1.115) yields

(1.116) \[ \hat{\varphi}_i = \varphi_i + \Lambda_{n,i} \]

for

(1.117) \[
\Lambda_{n,i} = -2\rho_i \left\{ \varphi_{n,i}(\rho + A_2)^{-1}(\rho + A_1)^{-1} + (I + \rho_{n,i})[R_{n,i,2}(\rho + A_1)^{-1} + (\rho + A_2)^{-1}R_{n,i,1}(\rho + A_1)^{-1} + R_{n,i,2}(\rho + A_2)^{-1}R_{n,i,1}(\rho + A_1)^{-1}] \right\}.
\]

The perturbation \( \Lambda_{n,i} \) is the absolute error in the single precision analogue of the \( i \)th local approximate inverse.

An examination of the terms in line (1.117) reveals that without purification we may be forced to accept a 2-norm bound for \( \Lambda_{n,i} \) as large as \( \Theta[\nu/\rho_i^2] \). For the smallest Peaceman Rachford parameter (see (1.98)) this bound behaves like \( \Theta[h^{-2+2/m}] \) for \( \nu = \Theta(c) \). The analysis following Theorem 2 in Section 1 applies if we replace \( \rho \) there with \( \rho_i \). In fact there is nothing about our argument to limit its applicability to the performance to just one purification. For a sufficiently small ratio of \( \nu/\rho_i \), the right-most term in braces in line (1.31) behaves like \( \nu \). Hence for each purification we are willing to perform on the
initial computed solution of \((\tilde{\rho}_i + \Lambda_i) x = y\), we obtain an improvement in the bound for relative error of the computed solution by a factor of \(\Theta\left(\frac{\nu}{\tilde{\rho}_i}\right)^k\), that is, a sequence of relative error matrices \(R_{n,i,j}^k\) satisfying

\[
(1.118) \quad \|R_{n,i,j}^k\|_2 \leq c_1 \left(\frac{\nu}{\tilde{\rho}_i}\right)^k \|R_{n,i,j}\|_2 + \Theta(\nu),
\]

may be defined where \(c_1\) is a positive constant and \(R_{n,i,j}\) is the relative error without purification. Repeated application of Theorem 1 yields

\[
(1.119) \quad \|R_{n,i,j}^k\| \leq \left[ c_1 \frac{\nu}{\tilde{\rho}_i} \right]^{k+1} + \Theta(\nu),
\]

where we agree to make \(c_1\) larger if necessary to accommodate the coefficient of \(\frac{\nu}{\tilde{\rho}_i}\) in the bound given in Theorem 1.

Since no Peaceman-Rachford parameter is as small as \(c\), in fact, none smaller than \(c \frac{1}{2^m}\), we gain at least a factor of \(\frac{1}{2^m}\) in the bound obtained for the 2-norm of the perturbation \(\Lambda_n,i\) with each purification we perform; hence no more than \(2m-1\) purifications are required to obtain an \(\Theta[1]\) bound on the 2-norm of \(\Lambda_n,i\), which may be made as small as desired by a judicious choice of the ratio of \(\nu\) to \(c\), instead of the \(\Theta\left[\frac{\nu}{c^{2-1/m}}\right]\) bound obtained without any purifications. In general, far fewer than \(2m-1\) purifications are required to obtain such a bound for the 2-norm of the absolute error in the \(i^{th}\) local approximate inverse. None are required for
any value of \( i \) between 1 and \( \frac{m+2}{2} \). Thus, \( m+2 \) purifications have been eliminated in the sense that we would have found it necessary to perform at least \( m+2 \) purifications to obtain an analogous bound on the absolute error in the approximate inverse of the single parameter technique of Section 1 for \( m+2 \) successive steps which need not be performed here. Let \( k(i) \) denote the number of purifications required to obtain a bound (uniform in \( n \)) of the form \( o[1] \) for the 2-norm of \( \Lambda_{n,i} \). It follows from line (1.119) that the choice

\[
(1.120) \quad k(i) = \left\lceil \frac{2i-1}{2(m-i)+1} \right\rceil,
\]

where \( \left\lceil \ \right\rceil \) denotes the greatest integer function, suffices. Summing the consecutive values of \( k(i) \) from \( \frac{m+2}{2} \) to \( m \), we obtain

\[
(1.121) \quad \sum_{i=\frac{1}{2}(m+2)}^{m} k(i) = \left[2m-1\right] + \left\lceil \frac{2m-3}{3} \right\rceil + \left\lceil \frac{2m-5}{5} \right\rceil \\
+ \cdots + \left\lceil \frac{m+1}{2(m-1)+1} \right\rceil.
\]

For \( m \) odd, the lower index of the summation above is not an integer; however, in that case we need not perform any purifications until \( i = \frac{m+3}{2} \). Hence the summation in (1.121) should begin at \( i = \frac{m+3}{2} \) when \( m \) is odd. For \( m \) even we obtain the bounds
(1.122) \( (m+1) \left\{ 1 + \frac{1}{3} + \cdots + \frac{1}{\left\lceil \frac{m-2}{2} \right\rceil + 1} \right\} - \frac{(m-2)}{2} \)

\( \leq \sum_{i=\frac{1}{2}(m+2)}^{m} k(i) \leq (2m-1) \left\{ 1 + \frac{1}{3} + \frac{1}{5} + \cdots + \frac{1}{\left\lceil \frac{m-2}{2} \right\rceil + 1} \right\} . \)

A term by term application of the elementary inequality

(1.123) \( \frac{1}{2n+1} < \log(2n+1) - \log(2n-1) < \frac{2}{2n-1} \)

yields

(1.124) \( \frac{(m+1)}{2} \log(m+1) - \frac{(m-2)}{2} < \sum_{i=\frac{1}{2}(m+2)}^{m} k(i) \)

\( \leq (2m-1)[\log(m-1) + 1] . \)

This bound also suffices for m: odd although for such m the summation in (1.122) should begin with \( i = \frac{1}{2}(m+3) \). Thus, we may guarantee that the absolute errors in the approximate inverses may be made uniformly small (for a sufficiently small ratio of \( \sqrt{\nu} \) to \( c \)) provided no more than

2(2m-1)[\log(m-1) + 1] purification iterations are performed, 2k(i) for each local approximate inverse whose index exceeds \( \frac{1}{2}(m+1) \). This is only a constant multiple of \( \log(m-1) \), the non-bounded elementary function of slowest growth, more purifications than required to obtain an analogous bound for the single parameter alternating direction iteration of Section 1 for \( m+2 \) successive steps.

We now consider the question of whether \( k(i) \) purifications of the solution implied by the factor \((\rho_1 + A_j)^{-1}\) of
the local approximate inverse \( \mathbf{B}_i, i \geq \frac{1}{2}(m+2) \) will insure the
stability of a complete m-cycle for the choice of \( \nu \) made
above, namely, \( \nu = \mathcal{O}(c) \). The answer to this question must
take into consideration how the cycle is actually computed.
For the strategy given in line (1.105) we obtain the single
precision analogue:

\[
(1.125) \quad \hat{\beta}(i+1) = (I + \hat{\psi}_1)[\hat{\beta}(i) + (A_1 + \Lambda_1)(I + \nu_1) + (A + \delta A_1)\hat{\beta}(i) - (I + \xi_1)b], \quad i = 1, 2, \ldots, m
\]

where the meanings of the perturbations are clear from the
context in which they appear. We should point out, however,
that the residuals \( A\beta(i) - b \) are assumed computed in \( \mathfrak{fl}_2 \-
arthmetic. No denotation of cycle number is made to sim-
plify notation. If we define

\[
(1.126) \quad \beta_{\rho} = (I + \hat{\psi}_1)[I + (B_1 + \Lambda_1)(I + \nu_1)(A + \delta A_1)], \quad i = 1, 2, \ldots, m
\]

the recursion of (1.125) may be written in the more compact
form

\[
(1.127) \quad \hat{\beta}(i+1) = \beta_{\rho} \hat{\beta}(i) + (A_1 + \Lambda_1)(I + \nu_1)(I + \xi_1)b, \quad i = 1, 2, \ldots, m-1
\]

This is the single precision analogue of the exact recursion

\[
(1.128) \quad \beta(i+1) = T_{\rho} \beta(i) + \mathbf{B}_1b, \quad i = 1, 2, \ldots, m-1
\]
Consequently,

\( (1.129) \quad \hat{u}(n) = \frac{m-1}{i=0} \rho_{m-i} \hat{u}(n) \)

\[ + \frac{m}{i=1} \frac{m-(i+1)}{k=0} \rho_{m-k} (\theta_i + \lambda_i) (I + \psi_i) (I + \xi_i) b. \]

Subtracting \( u(n) = \frac{m}{i=0} \rho_{m-i} \frac{m-i}{i=0} T_{m-i} \theta_i b \)

from both sides of (1.129), we obtain

\( (1.130) \quad r(n) = \hat{u}(n) - u(n) \)

\[ = \left[ \frac{m}{i=0} \rho_{m-i} + \frac{m}{i=0} \rho_{m-i} - \frac{m}{i=0} T_{m-i} \right] r(n-1) \]

\[ + (\hat{u}(n) - u(n)) , \]

where \( \hat{u}(n) \) is defined by

\( (1.131) \quad \hat{u}(n) = \frac{m-1}{i=0} \rho_{m-i} u(n-1) \)

\[ + \frac{m}{i=1} \frac{m-(i+1)}{k=0} \rho_{m-k} (\theta_i + \lambda_i) (I + \psi_i) (I + \xi_i) b. \]

The quantity \( \hat{u}(n) \) has the same form the machine iteration would assume at the \( n^{th} \) step of the computation if that step were begun with exact data \( u(n-1) \); however, we must view (1.131) as a definition. We shall think of the difference \( \hat{u}(n) - u(n) \) as the local error at the \( n^{th} \) step.
It is clear that the rounding error propagator will not magnify the global rounding error in $L^2$-norm from cycle to cycle, if, independently of the cycle index,

\begin{equation}
|| \prod_{i=0}^{m-1} T_{\rho_{m-i}} + \prod_{i=0}^{m-1} \mathcal{J}_{\rho_{m-i}} - \prod_{i=0}^{m-1} \mathcal{J}_{\rho_{m-i}} ||_2 < 1 .
\end{equation}

This condition will be satisfied a fortiori whenever

\begin{equation}
|| \prod_{i=0}^{m-1} \mathcal{J}_{\rho_{m-i}} - \prod_{i=0}^{m-1} T_{\rho_{m-i}} ||_2 < 1 - \left[ \frac{1 - e^{1/2m}}{1 + e^{1/2m}} \right]^2 .
\end{equation}

We now consider how $\nu$ must be constrained to satisfy (1.133) asymptotically for $m = \frac{\log(c)}{\log(\sqrt{2}-1)}$. To facilitate matters, we write

\begin{equation}
\mathcal{J}_{\rho_i} = T_{\rho_i} + \Delta_i ,
\end{equation}

where

\begin{equation}
\Delta_i = \phi_i T_{\rho_i} + (I + \phi_i) [ \Lambda_i (A + \delta A_i) + \Theta_i \delta A_i ]
\end{equation}

\begin{equation}
+ (\Theta_i + \Lambda_i) \psi_i (A + \delta A_i) .
\end{equation}

By performing $2k(i)$ purifications for $i \geq \frac{m+2}{2}$, none for $i < \frac{m+2}{2}$, we obtain the bound $||\Lambda_i||_2 \leq O \left[ \frac{\sqrt{c}}{c} \right]$; hence $||\Delta_i||_2 = O \left[ \frac{\sqrt{c}}{c} \right]$, $i = 1, 2, \ldots, m$. Denoting the largest of these bounds by $\varnothing$, it is clear that

\begin{equation}
|| \prod_{i=0}^{m-1} \mathcal{J}_{\rho_{m-i}} - \prod_{i=0}^{m-1} T_{\rho_{m-i}} ||_2 < \sum_{i=1}^{m} \binom{m}{i} \varnothing^i < \exp(m\varnothing) - 1 .
\end{equation}
Certainly (1.133) will be satisfied whenever

\[(1.137) \quad m \delta < \log \left\{ 1 + \left( 1 - \left[ \frac{1 - c^{1/2m}}{1 + c^{1/2m}} \right]^2 \right) \right\}. \]

The requirement of line (1.137) can be met by taking $\nu$ proportional to $c/(-\log(c^{1/2}))$ instead of $c$. This is not a great sacrifice, for $-\log(c^{1/2})$ grows more slowly than any non-positive power of $c$. We might be tempted to perform more purifications to improve the bound on $\Lambda_1$. From terms in $\Lambda_1$ like $\mathcal{E}_1 A$, it is clear that the limit of what may be reasonable expected from the purification technique has been achieved, for this term behaves in 2-norm like $O(c^{-1/2m})$ at worst whether we purify or not. Unfortunately,\n
\[
\lim_{c \to 0} (m c^{-1/2m}) = \infty \quad \text{for} \quad m = -\log(c^{1/2}).
\]

Consider the bound

\[(1.138) \quad \left\| r^{(n)} \right\|_{L^2} \]

\[
\leq \frac{1}{1 - \sup_{\text{all cycles}} \left\{ \left\| \prod_{i=0}^{m-1} T \rho_{m-i} + \left( \prod_{i=0}^{m-1} \mathcal{E}_{m-i} \right) - \prod_{i=0}^{m-1} T \rho_{m-i} \right\|_2 \right\}^{n}}
\]

\[
\cdot \max_j \left\| \hat{u}^{(j)} - u^{(j)} \right\|_{L^2}.
\]

On the one hand, the choice $\nu = O[c/(-\log(c^{1/2}))]$ guarantees that the supremum in the numerator of (1.138) decays with
increasing \( n \). On the other hand, since \( s \prod_{i=0}^{m-1} I_{p_{m-i}} \) tends to \( \left[ \frac{1 - (\sqrt{2} - 1)}{1 + (\sqrt{2} - 1)} \right]^2 \), the denominator of (1.138) approaches a constant asymptotically (for a sufficiently small ratio of \( \nu \) to \( c/\log(c^{\frac{1}{2}}) \)). Consequently, we need only examine the \( L^2 \)-norm of the local error vectors, \( \tilde{u}(j) - u(j) \), to establish the following theorem:

**Theorem 6:** If the \( m \)-cyclic Peaceman-Rachford alternating direction technique for the commutative case is executed in single precision machine arithmetic by the strategy of line (1.105) with the Peaceman-Rachford 'good' parameters (see (1.98)) and

(i) \( \nu \) is taken proportional to \( c/(-\log(c^{\frac{1}{2}})) \) in sufficiently small ratio,

(ii) each of the residuals computed in (1.105) is computed in \( \ell_2 \)-arithmetic,

(iii) \( k(i) = \left[ \frac{2i - 1}{2(m-1)-1} \right] \) purifications of Theorem 2 are performed on the solution implied by each factor \((\rho_i + A_j)^{-1}\) of the local approximate inverse \( a_i \) for a total of less than

\[
2(2m-1)[\log(m-1) + 1]
\]

but more than

\[
(m+1)[2 \log (m+1) - 1] + 3 ; \quad i \geq \frac{1}{2}(m+2),
\]

then the global rounding error vector satisfies
(1.139) \[ \| r^{(n)} \|_{L^2} \leq o(-\log(c^{\frac{1}{2}})) = o(c) \]

uniformly in \( n \).

**Proof:** To complete the proof we need only establish that

\[ \text{(1.140)} \sup \| \tilde{u}^{(j)} - u^{(j)} \|_{L^2} \leq o(m) \]

It is clear that (1.127) may be considered to be the propagator of a sequence \( \tilde{\beta}^{(1)}, \ldots, \tilde{\beta}^{(m)} \) whose \( m \)th term is \( \tilde{u}^{(n)} \) if we initialize with

\[ \tilde{\beta}^{(1)} = (I + \delta_1) [u^{(n)} + (\beta_1 + \Lambda_1)(I + \gamma_1) \cdot \left\{ (A + \delta A_1)u^{(n-1)} - (I + \xi_1)b \right\}] . \]

Rearrangement of the term in braces yields

\[ \text{(1.142)} \quad \tilde{\beta}^{(1)} = (I + \delta_1) \left[ \beta^{(1)} + (\beta_1 + \Gamma_1)(\delta A_1 u^{(n-1)} - \xi_1 b) + \Gamma_1 (A u^{(n-1)} - b) \right] . \]

where \( \Gamma_1 \) is defined by

\[ \text{(1.143)} \quad \Gamma_1 = (\beta_1 + \Lambda_1)(I + \gamma_1) - \beta_1 . \]

Since the exact iteration is convergent, \( \beta^{(1)}, u^{(n-1)}, \) and \( h^{-2}(Au^{(n-1)} - b) \) are bounded in \( L^2 \)-norm. Also, since \( \Gamma_1 \) is merely an \( o(\nu) \) perturbation of \( \Lambda_1 \) in 2-norm, \( \| \Gamma_1 \| \leq o(1) \); hence, the third term in (1.143) behaves no worse than \( o(\nu) \) in \( L^2 \)-norm, the second no worse than \( o\left[ \frac{\nu}{c} \right] \).
A relationship similar to the one given in line (1.142) holds for every relevant value of the index $i$. Specifically,

\begin{equation}
\rho(i+1) = \rho(i+1) + \delta(i+1), \quad 1 \leq i \leq m-1 \tag{1.144}
\end{equation}

where

\begin{equation}
\delta(i+1) = \delta(i+1) + \delta(i) + (\beta_{i+1} + \Gamma_{i+1})
\end{equation}

\begin{equation}
\cdot (A + \delta A_{i+1}) \delta(i) + (\beta_{i+1} + \Gamma_{i+1})
\end{equation}

\begin{equation}
\cdot (\delta A_{i+1} \beta(i) - \varepsilon_{i+1} b) + \Gamma_{i+1}(A \delta(i) - b)
\end{equation}

\begin{equation}
= \delta(i+1) + \epsilon(i+1), \quad i = 2, \ldots, m-1.
\end{equation}

The second expression on the right is obtained from the first by noting that the term underlined is merely $\delta_{i+1}$. The perturbation $\epsilon(i+1)$ is defined to be the sum of the remaining terms. By induction

\begin{equation}
\delta(m) = \prod_{j=0}^{m-1} \delta(j) + \sum_{k=2}^{m} \prod_{j=0}^{m-k} \epsilon(k).
\end{equation}

Having already imposed a sufficiently stringent condition on $\nu$ to insure that $\prod_{j=0}^{m-1} \delta_{m-j}$ has a 2-norm bound of the form $\delta(1)$, we need only establish that the $\epsilon(k)$ above have $L^2$-norm bounds proportional to $\nu$ to conclude that $\rho(m) - \beta(m) = \delta(m)$ has an $L^2$-norm bound of the form $\delta(m\nu)$. To accomplish this we first show that $\beta(i)$ is bounded independently of $h$ in $L^2$-norm for all $i$ less than or equal to $m$. Since
\[(1.147) \quad \beta^{(i)} = u^{(n-1)} + \beta_1(Au^{(n)} - b) + \beta_2 T_{\rho_1} (Au^{(n-1)} - b) + \cdots + \beta_i \prod_{k=1}^{i} T_{\rho_i} (Au^{(n-1)} - b)\]

and \(A\beta_i = T_{\rho_i} - I\), it is clear that

\[(1.148) \quad A\beta^{(i)} = Au^{(n-1)} + (T_{\rho_1} - I)(Au^{(n-1)} - b) + (T_{\rho_2} - I)T_{\rho_1} (Au^{(n-1)} - b) + \cdots + (T_{\rho_i} - I)T_{\rho_{i-1}} \cdots T_{\rho_1} (Au^{(n-1)} - b).\]

This telescopes into

\[(1.149) \quad A\beta^{(i)} = Au^{(n-1)} \left[ \prod_{k=0}^{i-1} T_{\rho_{i-k}} - I \right] (Au^{(n-1)} - b),\]

from which we obtain, after premultiplication by \(A^{-1}\),

\[(1.150) \quad \beta^{(i)} = u^{(n-1)} + \left[ \prod_{k=0}^{i-1} T_{\rho_{i-k}} - I \right] (u^{(n-1)} - u),\]

where \(u\) is the exact solution of \(Au = b\). Thus \(\beta^{(i)}\) is bounded in \(L^2\)-norm uniformly in \(i\), for \(\prod_{k=0}^{i-1} T_{\rho_{i-k}}\) is bounded in 2-norm by unity regardless of the value of \(m\). Moreover, for \(\bar{u} = \max \{ ||u||_{L^2}, ||U||_{L^2} \}\), where \(U\) is the analytic solution of the Dirichlet problem (1.1), it is clear that

\[(1.151) \quad \| \beta^{(i)} \|_{L^2} \leq 5 \bar{u}.\]
Consider the residuals \( A_{\beta}^{(i)} - b \). From the recursion

\[
(1.152) \quad A_{\beta}^{(i)} - b = \prod_{k=0}^{i-1} \mathbf{T}_{i-k} (A u^{(n-1)} - b)
\]

we perceive that

\[
(1.153) \quad \left| \frac{A_{\beta}^{(i)} - b}{h^2} \right|_{L^2} \leq \left| \frac{A u^{(n-1)} - b}{h^2} \right|_{L^2}.
\]

However, for a smooth initial guess the residual \( \frac{A u^{(n-1)} - b}{h^2} \) approaches zero with increasing \( n \); hence \( h^{-2}(A_{\beta}^{(i)} - b) \) is bounded in \( L^2 \)-norm by a constant independent of \( h \).

It is clear from (1.145), (1.151), and (1.153) that

\[
\left\| e^{(k)} \right\|_{L^2} \leq \varepsilon(\nu) \quad \text{from which we deduce, finally, that}
\]

\[
(1.154) \quad \left\| \delta^{(m)} \right\|_{L^2} \leq \varepsilon(m\nu).
\]

This completes the argument. We have assumed in this analysis that terms of the form \( m\nu^2 \) are negligible. Also note that although a parameter ordering is implicit in (1.105), none of the results above depend on this ordering.

We shall now comment on the execution of the Peaceman-Rachford alternating direction technique by the strategy given by line (1.107). The approximate inverse \( B_\rho \) is realized computationally as

\[
(1.155) \quad B_m = \prod_{j=1}^{m-1} (1 + \delta_{m-j})(\rho_1 + \Gamma_1)
\]

\[
+ \sum_{i=2}^{m} \prod_{j=1}^{m-(i-1)} (1 + \delta_{m-j})(\rho_i + \Gamma_i) \prod_{k=1}^{i-1} \delta_{i-k}
\]
where $\beta_i$ and $\lambda_i$ are defined as before, $\theta_i$ is the perturbation introduced in the accumulation of the sum $\delta^i_n$ (see line (1.107)), and $\mathcal{S}_{\rho, \ell}$ is an approximation to $T_{\rho, \ell}$ given by

\begin{equation}
\mathcal{S}_{\rho, \ell} = (1 + \theta_{\ell})[(1 + \psi_{\ell}) + (A + \delta A_{\ell})(\beta_{\ell} + \lambda_{\ell})].
\end{equation}

The form of $\mathcal{S}_{\rho, \ell}$ reveals that we intend to compute this step in $f_{\ell_2}$-arithmetic. The recursion for the $n^{th}$ single precision iterate is given by

\begin{equation}
\hat{u}^{(n)} = (I + \hat{s}_m)(\hat{u}^{(n-1)} - \overline{\beta}_m \hat{\hat{r}}^{(n-1)}),
\end{equation}

where

\begin{equation}
\hat{r}^{(n-1)} = f_{\ell_2}(Au^{(n-1)} - b).
\end{equation}

Subtracting $u^{(n)}$ from both sides yields

\begin{equation}
r^{(n)} = (I + \hat{s}_m)(I - \overline{\beta}_m(I + \chi_n)(A + \delta A_n))r^{(n-1)}
\end{equation}

\begin{equation}
+ \left[ \hat{s}_m - (I + \hat{s}_m)(\overline{\beta}_m(I + \chi_n)\delta A_n) \right]u^{(n-1)}
\end{equation}

\begin{equation}
- (I + \hat{s}_m)\overline{\beta}_m(I + \chi_n)\hat{\hat{r}}^{(n-1)}b
\end{equation}

\begin{equation}
+ \left[ B_m - (I + \hat{s}_m)\overline{\beta}_m(I + \chi_n) \right] (Au^{(n-1)} - b).
\end{equation}

From this last line we see that the global rounding error is propagated by $I - \overline{\beta}_m(I + \chi_n)(A + \delta A_n)$, a perturbation of $I - B_mA$, the error propagator for the Peaceman-Rachford multiparameter iteration. These two propagators differ to
$\varphi(n)$ terms in $L^2$-norm by

\[(1.160)\quad \bar{\varphi}_m (I + \chi_n) (A + \delta A_n) - B_m A\]

\[\approx \sum_{i=1}^{m} \beta_i \left[ \prod_{k=0}^{i-1} T^\rho_{i-k} - \prod_{k=0}^{i-1} \rho^\rho_{i-k} \right] A\]

\[+ \sum_{i=1}^{m} \left( I + \rho_{i-j} \right) \lambda_i \prod_{k=0}^{i-1} T^\rho_{i-k} (I + \chi_n) (A + \delta A_n)\]

Examining the first term on the right-hand side of (1.160), we notice that if $A$ and $\prod \rho^\rho_{i}$ were to commute for all $i$, the behavior of this term would depend essentially on that of the difference $\prod T^\rho_{i-k} - \prod \rho^\rho_{i-k}$, for $T_i - I = \beta_i A$. On the other hand, it is possible, although unlikely, that $\prod \rho^\rho_{i}$ and $\beta_i$ share the eigenvector associated with each of their spectral radii, that $\prod \rho^\rho_{i}$ is symmetric, and that $\prod \rho^\rho_{i}$ does not commute with $A$. In this case the component of $r(n)$ in the direction of this eigenvector is magnified in $L^2$-norm by $\varphi \left( \frac{\nu}{\beta_i} \right)$ which behaves like $\varphi \left[ \frac{\nu}{c^3(1-\frac{1}{2m})} \right]$ at worst. We cannot guarantee that this magnification is small without a much more stringent requirement on the precision of single precision arithmetic than was necessary for the first strategy of computation. Since commutativity fails to be a property of single precision arithmetic, we cannot exploit the fact that all of the exact local approximate inverses operate on vectors of the form $\prod_{k=0}^{i-1} T^\rho_{i-k} (Au(n-1) - b)$. 
Finally, we consider the asymptotic rounding error behavior of the $m$-cyclic alternating direction technique when it is attempted with fixed, and therefore non-optimal, $m$. A bound on the spectral radius of the exact error propagator like the one given in line (1.114) for the model problem is valid in this case. Consequently, the condition of line (1.138) may be replaced by

\[(1.161) \quad m \delta < \log[1 + \delta(h^{1/m})] \, .\]

Since $m$ is fixed, it suffices to perform just one more purification than before (at most) for $i \geq \frac{m+2}{2}$ to satisfy (1.161) provided that $\nu/c$ is sufficiently small. That is, since the performance of $k(i)$ purifications will insure that $\delta = \delta(1)$, the performance of no more than one more purification for each value of $i \geq \frac{m+2}{2}$ at most will guarantee that $\delta = \delta(h^{1/m})$ for $\nu$ proportional to $c = \delta(h^2)$.

Specifically, no more than

\[(1.162) \quad k(i) = \left[ \frac{2i}{2(p-1) + 1} \right], \quad i \geq \frac{1}{2}(m+2)\]

purifications of the computed solution implied by $(\rho_i + A_j)^{-1}$ are required to insure that the bound for $\| \Delta_i \|_2$ behaves no worse than $\delta(h^{1/m})$ provided that $\nu = \delta(c)$, $j = 1, 2$. Thus for fixed $m$ we have the following theorem:

**Theorem 7:** For fixed $m$, the global rounding error in the single precision realization of the Peaceman-Rachford
alternating direction iteration of Theorem 6 has a uniform $L^2$-bound no larger than

\[(1.163) \quad \|r^{(n)}\|_2 < s[m_h^{-1/m}],\]

if all of the relevant hypotheses of Theorem 6 are satisfied.

**Wachspress Parameters**

We shall now obtain the analogues of Theorems 6 and 7 for the Wachspress parameters:

\[(1.164) \quad \rho_i = b \left[ \frac{i-1}{c^{m-1}} \right], \quad m > 2, \ i = 1, 2, \ldots, m.\]

Estimates of the average rate of convergence (see Birkhoff et al [2]) for 'optimal' $m$ reveal that these parameters are superior to the Peaceman-Rachford parameter set given in line (1.100) by a factor of approximately 2 asymptotically. Moreover, it can be shown that

\[(1.165) \quad \max_{\gamma \in [a, b]} \prod_{i=1}^{m} \left| \frac{\gamma - \rho_i}{\gamma + \rho_i} \right| \leq \left[ \frac{1}{1 - c^{2(m-1)}} \right]^{1/2}, \quad \sigma(A_i) \in [a, b].\]

For the model problem, in particular, we obtain the bound

\[(1.166) \quad s \left( \prod_{i=0}^{m-1} T_{\rho_{m-i}} \right) \leq 1 - 8 \left[ \frac{m_h}{2} \right]^{1/m-1},\]

for the choice of the Wachspress 'good' parameters.
From line (1.164) it is clear that the largest and smallest Wachspress 'good' parameters are \( a \) and \( b \), respectively. Moreover, it is also clear that for \( \nu \) proportional to \( c \) the performance of \( k(i) \) purifications of Theorem 2 (where \( k(i) \) is defined in line (1.120)) will not necessarily insure that the absolute errors in the local approximate inverses are small in 2-norm no matter how small the choice of the ratio of \( \nu \) to \( c \). For the Peaceman-Rachford parameters, the effectiveness of the purification depended on the fact that no parameter, not even the smallest, is proportional to \( c \) (as is the case here). Thus, choosing \( \nu \) proportional to \( c \) will not suffice.

We should like to make as small an additional sacrifice to machine precision as possible, such as choosing \( \nu \) proportional to \( \frac{c}{-\log c} \). This will not do, however, for any positive power of \( c \) no matter how small tends to zero faster than \(-\log c\) tends to infinity with decreasing \( c \). Suppose \( \nu \) is chosen proportional to \( c^{1+\alpha} \) for fixed \( \alpha \in (0,1] \). By an analysis similar to the one given for the Peaceman-Rachford parameters we can show that

\[
(1.167) \quad k(i) = \begin{cases} \left[ \frac{i - \frac{1}{m-1} + \alpha(m-1)}{1 + \alpha(m-1)} \right] & , \quad i > \frac{1}{2}(m+2) \\
\frac{1 - \alpha}{\alpha} & , \quad i = m
\end{cases}
\]

purifications are sufficient to insure that the absolute error in the single precision analogues of the local
approximately inverses are uniformly small for a sufficiently
small choice of the ratio $\sqrt{c^{1+\alpha}}$ when measured in 2-norm.
The initial value of the index $i$ above is due to the obser-
vation that regardless of the size of $\alpha$, no purifications
are necessary for the first $\frac{1}{2}(m+2)$ steps of any cycle.
Moreover, for $\alpha = 1$ no purifications are required for any
step in the cycle. However, we are interested in small $\alpha$,
for the smaller $\alpha$ the smaller the sacrifice to machine
precision. For any value of $i$ but $m$, we need not require
that $k(i)$ be any greater than $\left[\frac{i-1}{m-i}\right]$ as $\alpha$ tends to zero,
but for $i = m$, $k(m)$ must tend to infinity with decreasing $\alpha$.
For the choice $\alpha = \frac{1}{2}$ it is clear that $k(i)$ may be taken to
be unity for each value of $i$. On the other hand, since
$\left[\frac{i-1}{m-i}\right]$ is an increasing function of $i$ which never exceeds
$m-2$ for $i \leq m-1$, we may attempt to choose $\alpha$ to yield a
bound of the desired form for $\|A_m\|_2$ without requiring more
than $m-1$ purifications. This quantity is chosen because it
is only one greater than the number of purifications required
for the $i = m-1$ case for arbitrarily small $\alpha$. Obviously the
choice $\alpha = 1/(m-1)$ suffices. For this value of $\alpha$ no more
than
\begin{equation}
(1.168) \quad 2 \sum_{i > \frac{1}{2}(m+2)}^m k(i) < 2(m-1)[\log[\frac{1}{2}(m-3)] + 1]
\end{equation}

purifications per cycle are necessary, but at least
\begin{equation}
(1.169) \quad (m-3)\log[\frac{1}{2}(m-1)] - (m-3) < 2 \sum_{i > \frac{1}{2}(m+2)}^m k(i)
\end{equation}
are required for \( k(i) \) given by (1.167). Moreover, this result is typical in the sense that if \( \alpha \) is chosen proportional to \( 1/m \), then more than a constant multiple of \( m \) purifications are required to assure that no approximate inverse has an absolute error whose 2-norm is not arbitrarily small for a sufficiently small but fixed ratio of \( \nu \) to \( c^{1+\alpha} \).

For small \( c \), the choice of 'optimal' \( m \) is made the same way for the Wachspress 'good' parameters as for the Peaceman-Rachford parameters; namely,

\[
(1.170) \quad m = \frac{\frac{1}{2} \log c}{\log(\sqrt{2} - 1)}.
\]

For this value of \( m \) and the bound given in line (1.165) it follows by analogy with the argument given for the Peaceman-Rachford parameters that the choice of \( \nu \) proportional to \( c^{1+\alpha}/-\log c^{\frac{1}{2}} \) suffices to negate the growth of \( m \) with decreasing \( c \). Hence we have

**Theorem 8:** If the optimal m-cyclic alternating direction iteration for the Wachspress 'good' parameter set is executed in single precision arithmetic by the strategy of line (1.105) for the commutative case in \( R^2 \) and

(i) \( \nu \) is chosen proportional to \( c^{1+\alpha}/-\log c^{\frac{1}{2}} \) for any \( \alpha \in (0,1] \) in sufficiently small but fixed ratio,

(ii) each of the residuals in (1.105) is computed in \( f_{l_2} \)-arithmetic,
(iii) \( k(i) = \left[ \frac{i-1}{(m-1) + \alpha(m-1)} \right], \ k(m) = \frac{1 - \alpha}{\alpha} \) purifications iterations of Theorem 2 are performed on the single precision realization of the solution implied by \((\rho_i + A_j)^{-1}; j = 1, 2; i > \frac{m}{2}(m+1), m: \text{odd} (i > \frac{m}{2}, m: \text{even})\) for a total of not more than
\[ \frac{2}{\alpha} + 2(m-2)[\log \left[ \frac{1}{2}(m-2) + 1 \right]] \]
iterations per cycle but at least \( \frac{2}{\alpha} + m \log \left[ \frac{m}{2} \right] - (m-2) \)
assuming \( \alpha \) is negligible compared to \( m \), then the global rounding vector \( r^{(n)} \) possesses a uniform \( L^2 \)-norm bound of the form
\[ (1.171) \quad \| r^{(n)} \|_{L^2} \leq \Theta(- \nu \log c^{\frac{1}{2}}) = \Theta(c) . \]

If we make the assumption that \( m \) is fixed as \( c \) tends to zero, we obtain
\[ (1.172) \quad \| r^{(n)} \|_{L^2} \leq \Theta(\nu^{1/m}) , \]
from the bound given in (1.167) for \( \nu \) proportional to \( c^{1+\alpha} - \log(c^{\frac{1}{2}}) \). This, of course, is valid only if \( k(i) \) purifications as given by (iii) above are performed to insure the stability of \( \rho_i \) with respect to the growth of rounding errors. Line (1.173) is the analogue of line (1.163) of Theorem 7.

Finally, we remark that an alternate but more costly way to satisfy (1.171) is to require that \( \nu = \Theta(c^{1+b}) \) for \( b > \alpha \).
Chapter 2. Asymptotic Rounding Error Analysis for the Successive Overrelaxation Technique for Self-Adjoint Uniformly Elliptic Problems

Section 1. Single Parameter Point Iterative Successive Overrelaxation

Description of the Technique

Having analyzed the asymptotic behavior of the rounding error in the single precision realizations of the alternating direction techniques for self-adjoint problems, we turn our attention to the successive overrelaxation iteration. As before, the Dirichlet problem for which an approximate solution is sought will be taken to be

\[(2.1) \quad gU = f \quad \text{in} \quad \Omega \]
\[U = g \quad \text{on} \quad \partial \Omega \]

where \( g \) is described in line (vii) of the Introduction. We assume the region \( \Omega \) is such that the discretization of \((2.1)\) described in line (x) of the Introduction leads to a difference system

\[(2.2) \quad Au = b \]

with an irreducible Stieltjes coefficient matrix. In terms of the splitting

\[(2.3) \quad A = D - L - U \]

of the coefficient matrix of this system (see Chapter 0),
the difference correction for the stationary successive overrelaxation iteration for (2.2) is given by

\[ d^{(n)} = -w_b(D - w_bL)^{-1}(Au^{(n)} - b). \]

If at the outset (2.2) is scaled by \( D^{-1} \), the matrices \( D, L \) and \( A \) above may be replaced by \( I, g = D^{-1}L, \gamma = D^{-1}A \), respectively. The vector \( b \) must be replaced with \( c = D^{-1}b \). This scaling insures that the solution of the linear system implied by the approximate inverse can be obtained by Gaussian elimination without divisions. Young [16] has shown that the optimal overrelaxation parameter \( w_b \) is related to the spectral radius \( \mu \) of the point Jacobi error propagator (described in Chapter 0) by

\[ w_b = \frac{2}{1 + (1 - \mu^2)^{\frac{1}{2}}}. \]

For the model problem \( a_1 = a_2 = 1, q = 0, w_b = 2 - 2h + \mathcal{O}(h^2) \). Moreover, we shall show that \( w_b = 2 - ch + \mathcal{O}(h^2) \) for the more general Dirichlet problem (2.1). For \( w_b \), the spectral radius of the error propagator of the stationary successive overrelaxation iteration

\[ \phi_{w_b} = I - w_b(I - w_b\phi)^{-1} \gamma \]

is known to satisfy

\[ s(\phi_{w_b}) \leq \frac{1 - (1 - \mu)^2}{1 + (1 - \mu^2)^{\frac{1}{2}}}. \]
However, the $n^{th}$ power of the error propagator tends to zero like $ns(z_{wb}^n)$ asymptotically in the usual norms. Indeed, Young [16] has recently established that if $A$ is merely a positive definite consistently ordered matrix (and not necessarily a Stieltjes matrix), then

$$(2.8) \quad \|z_{wb}^n\|_D = s(z_{wb})^n \left\{ n(s(z_{wb})^{\frac{1}{2}} + s(z_{wb})^{-\frac{1}{2}} \right. $$

$$+ \left[ n^2(s(z_{wb})^{\frac{1}{2}} + s(z_{wb})^{-\frac{1}{2}})^2 + 1 \right]^{\frac{1}{2}} \right\} ,$$

where the D-norm is the matrix norm related to the 2-norm by

$$(2.9) \quad \|Q\|_D = \|D^{\frac{1}{2}}QD^{-\frac{1}{2}}\|_2$$

for any matrix $Q$. For the model problem these two norms are identical.

As we shall later prove, the spectral radius of the successive overrelaxation error propagator tends to unity as $h$ tends to zero; hence we perceive that it will be more difficult to obtain a uniform $L^2$-norm bound on the rounding error in the single precision realization of successive overrelaxation technique than was the case for the single parameter alternating direction iteration. We shall, in fact, give our analysis in the D-norm, converting to the $L^2$-norm only for the purpose of comparison. Uniform ellipticity insures the existence of $D^{-1}$. Recall that the difference system has been scaled by $h^2$ as in our earlier analyses to avoid computation with large quantities.
Asymptotic Rounding Error Analysis

Since the iteration parameter does not change from step to step, a multiplication per mesh point may be avoided by storing \(-w_b\xi\) instead of \(\xi\). The error generated in the single precision computation of the elements of \(-w_b\xi\) is quite small; hence an appeal to the inherent stability of the successive overrelaxation procedure for (2.1) assures that no growing error is introduced by considering \(-w_b\xi\) to be exact just as we are assuming \(\xi\) and \(\zeta\) to be exact. By storing \(w_bG\) and \(w_bC\) an additional multiplication per mesh point may be saved. No attempt was made to avoid this multiplication in the stationary alternating direction case since it was felt that because the iteration parameter is proportional to \(h\) multiplication by it has a self-correcting effect on the computation of the solution implied by the approximate inverse. In any case, no significant error is introduced in this multiplication. From the remark following line (2.5) it is clear that multiplication by \(w_b\) cannot possibly have a beneficial effect on any error introduced in the computation of the solution implied by the approximate inverse of the successive overrelaxation iteration. Thus, storage of the scaled components of the residual is desirable. For the accumulation of the scaled residual in \(f_{t_2}\)-arithmetic, line (2.4) is realized in single
precision arithmetic as

\[
(2.10) \quad \hat{u}^{(n)} = (I + \xi_n) \{ \hat{u}^{(n-1)} - \omega_b (I + R_n) (I - \omega_b \xi) \}^{-1} (I + \xi_n) \cdot [G + \delta G_n] \hat{u}^{(n-1)} - (I + \psi_n)c \}, \quad n \geq 1.
\]

In (2.10) we note that (i) $\xi_n$ is a diagonal matrix whose elements are bounded in magnitude by $\nu$ corresponding to the relative error introduced in the single precision computation of the vector sum $\hat{u}^{(n-1)}$ and $\hat{d}^{(n-1)}$; (ii) $R_n$ is a measure of the relative accuracy of the single precision solution implied by $(I - \omega_b \xi)^{-1}$; and (iii) $\xi_n$, $\delta G_n$ and $\psi_n$ are generalizations of the perturbations described in line (xiv) of the Introduction. In particular $\delta G_n$ and $\psi_n$ correspond to the relative errors introduced in the double precision accumulation of each component of the scaled residual as an inner product; $\xi_n$ corresponds to the error introduced in the subsequent rounding of these inner products to single precision before storage.

In terms of the absolute error in the solution implied by $(I - \omega_b \xi)^{-1}$,

\[
(2.11) \quad \Lambda_n = (I - \omega_b \xi)^{-1} - (I + R_n) (I - \omega_b \xi)^{-1},
\]

we obtain, subtracting the exact $n^{th}$ iterate from (2.10), the following expression for the global rounding error at
the $n^{th}$ step of the computation:

$$r^{(n)} = \hat{u}^{(n)} - u^{(n)}$$

$$= [s_{wb} + \{\delta_n s_{wb} - \omega_b (I + \alpha_n) (I - \omega_b g)^{-1} \phi_n (\alpha + \delta G_n)$$

$$+ (I - \omega_b g)^{-1} \delta G_n\}] r^{(n-1)}$$

$$+ \{\delta_n u^{(n)} - \omega_b (I + \alpha_n) (I - \omega_b g)^{-1}$$

$$\cdot [\delta_n (\alpha u^{(n-1)} - c) (I + \phi_n) (\delta G_n u^{(n-1)} - \psi_n c)]\}.$$

Referring to the first quantity in braces above as $\Delta_n$ and the second as $e^{(n)}$, this recursion may be written in the by now familiar form

$$r^{(n)} = [s_{wb} + \Delta_n] r^{(n-1)} + e^{(n)} , \quad n \geq 1 .$$

Consider the computation of the linear system implied by the approximate inverse. For the usual ordering of the mesh points $A$ is block tridiagonal. Moreover, no row of $(I - \omega_b g)$ consists of more than three non-zero entries; hence the solution of the system $(I - \omega_b g)$ consists of more than three non-zero entries; hence the solution of the system $(I - \omega_b g) y = z = f_{L,2} (\alpha u^{(n-1)} - c)$ is easily obtained by Gaussian elimination. Since the diagonal of this system is everywhere unity, no divisions are necessary. Furthermore, no pivoting is required, for the matrix $(I - \omega_b g)$ is, as we
shall later show, diagonally dominant. Typically,

\begin{equation}
(2.15) \quad \hat{y}_{i,j} = z_{i,j} - w_b c_{i-1,j} y_{i-1,j} - w_b f_{i,j-1} y_{i,j-1},
\end{equation}

\[(ih,jh) \in \Omega^0_h.\]

Since we are assuming that the quantities \(w_b c_{i-1,j}, w_b f_{i,j-1}\) are exact, that is, determined prior to iteration as a part of the data of the problem, (2.15) is realized in single precision arithmetic as

\begin{equation}
(2.16) \quad \hat{y}_{i,j} = z_{i,j} (1 + \gamma_{i,j}) - w_b c_{i-1,j} \hat{y}_{i-1,j} (1 + \mu_{i-1,j}) \nonumber \\
- w_b f_{i,j-1} \hat{y}_{i,j-1} (1 + \lambda_{i,j-1}).
\end{equation}

Dividing both sides of (2.16) by \((1 + \gamma_{i,j})\) to disassociate any rounding error from the right-hand side of the system, we obtain

\begin{equation}
(2.17) \quad \hat{y}_{i,j} (1 + \delta_{i,j}) = z_{i,j} - w_b c_{i-1,j} \hat{y}_{i-1,j} (1 + \theta_{i-1,j}) \nonumber \\
- w_b f_{i,j-1} \hat{y}_{i,j-1} (1 + \varphi_{i,j}).
\end{equation}

where \(1 + \delta_{i,j} = 1/(1 + \gamma_{i,j}), 1 + \theta_{i-1,j} = (1 + \mu_{i-1,j})/(1 + \gamma_{i,j})\) and \(1 + \varphi_{i,j} = (1 + \lambda_{i,j-1})/(1 + \gamma_{i,j})\). In matrix form this may be written as

\begin{equation}
(2.18) \quad [(I - w_b \xi) + K] \hat{y} = z,
\end{equation}

where \(K\) is a lower triangular matrix which may be obtained from \(I - w_b \xi\) by replacing the diagonal by \(\delta_{i,j}, w_b c_{i-1,j}\) with
\( \theta_{i-1,j} \) and \( w_{b,j}^f, j-1 \) with \( \psi_{i,j} \), respectively. Treating (2.16) as a single precision inner product, we deduce from line (xv) of the Introduction that none of the error perturbations \( \gamma_{i,j}, \mu_{i,j}, \lambda_{i,j} \) exceed \( 3.18 \nu \). Consequently, all of the perturbations \( \delta_{i,j}, \theta_{i,j} \) and \( \psi_{i,j} \) are bounded above in magnitude by \( 3.18 \nu + \mathcal{O}(\nu^2) \). Thus, we have established that

\[
(2.19) \quad \| K \|_\infty \leq 3.18 \| I - w_{b.g} \|_\infty \nu + \mathcal{O}(\nu^2).
\]

The relative error \( R_n \) in the solution implied by \( (I - w_{b.g})^{-1} \) may be expressed in terms of \( K \) as

\[
(2.20) \quad R_n = ((I - w_{b.g}) + K)^{-1}(I - w_{b.g}) - I. \]

Factoring \( (I - w_{b.g}) \) from the first term on the right above yields

\[
(2.21) \quad R_n = (I + (I - w_{b.g})^{-1}K)^{-1} - I. \]

In any norm, then,

\[
(2.22) \quad \| R_n \| \leq \frac{\| (I - w_{b.g})^{-1} K \|}{1 - \| (I - w_{b.g})^{-1} K \|}
\]

if, in this norm, \( \| (I - w_{b.g})^{-1} K \| < 1 \).

From the remark following line (2.18) it is clear that the bound given in line (2.19) also holds for the 1-norm. Moreover, this bound is valid for the 2-norm as well, for
given any matrix \( Q \), \( \| Q \|_2 \leq \left( \| Q \|_\infty \| Q \|_1 \right)^{\frac{1}{2}} \). All we lack to bound the 2-norm of \( R_n \) is an estimate for the 2-norm of \( (I - w_b \xi)^{-1} \). For the model problem it is clear that \( \|(I - w_b \xi)^{-1}\|_2 \leq (1 - \frac{1}{2} w_b)^{-1} \), for in that case \((I - w_b \xi)\) is diagonally dominant by \( (1 - \frac{1}{2} w_b)^{-1} \). For the more general problem (2.1) scaling the coefficient matrix \( A \) of the difference system, which apart from the contribution from the non-derivative term \( q(x_1, x_2) \) is zero row sum for rows corresponding to interior mesh points, by \( D^{-1} \) is sufficient to insure that \( \xi = D^{-1} L \) has no row sum corresponding to an interior mesh point exceeding one-half. Row sums corresponding to points adjacent to the boundary are even smaller than one-half in magnitude. Once the assertion following line (2.5) has been established, an application of Lemma 0 will yield the bound

\[
(2.23) \quad \|(I - w_b \xi)^{-1}\|_\infty \leq (1 - \frac{w_b}{2})^{-1} = O(h^{-1}) \, .
\]

Since the same argument applies to the column sums, the max-norm in (2.23) may be replaced with the 2-norm. Thus, for \( \sqrt{h^{-1}} \) small enough to insure that the diagonal dominance of the perturbed matrix \((I - w_b \xi) + K\) is not sacrificed, we obtain the uniform bound

\[
(2.24) \quad \| R_n \|_2 < O\left[\frac{\sqrt{h}}{n}\right] \, .
\]

Having obtained uniform 2-norm bounds for \( R_n \) and \((I - w_b \xi)^{-1}\), we may now bound the absolute error \( \lambda_b \) in the
approximate inverse. From line (2.11) it follows that \( \Lambda_n \) behaves like the product \( R_n (I - w_b)\)^{-1} to within \( \Theta(1) \) perturbations in 2-norm. Thus, we must accept a 2-norm bound of the form

\[ \| \Lambda_n \|_2 \leq \Theta \left( \frac{\nu}{h^2} \right) . \]

The purification iteration of Theorem 2 should yield an improved bound here, for the computation of the solution of the system implied by \( (I - w_b)\)^{-1} is equivalent to a half-step of the computation of the approximate inverse of the single parameter alternating direction technique.

Specifically, by redefining \( R_n \) to include the rounding error introduced in the purification step as well as the elimination step, we deduce from Theorem 2 that

\[ \| R_n \|_2 \leq \Theta(\nu) , \]

for \( \nu \) proportional to \( h^2 \). This is an improvement by a factor of \( h \) in the bound of (2.25) for the redefined \( R_n \).

The perturbation \( \Delta_n \) of line (2.12) possesses a 2-norm bound no larger than \( \Theta \left( \frac{\nu}{h^2} \right) \) since \( \delta G_n \) has a 2-norm bound of the form \( \Theta(\| G \|_2 \nu^2) \).

Assuming that the initial guess, \( u^{(0)} \), is exactly representable, we obtain the expression

\[ r^{(n)} = \sum_{i=1}^{n} \prod_{j=0}^{n-(i+1)} (\delta w_b + \Delta_{n-j}) e^{(i)} , \]
for the global rounding error at the \( n \)th step of the computation, where we shall adopt the convention that the product from 0 to 0 is \( \varepsilon_b + \Delta_n \), and from 0 to -1 is the identity matrix of rank \( N_h \). If \( \varepsilon_b \) and \( \Delta_j \) were to commute for all \( j \), the right-hand side of (2.27) could be rearranged as suggested by the binomial theorem:

\[
(2.28) \quad r^{(n)} = \sum_{i=1}^{n} \varepsilon_{wb}^{n-i} e^{(i)} + \sum_{i=1}^{n-1} \sum_{j=0}^{n-(i+1)} \varepsilon_{wb}^{n-(i+1)} \Delta_{n-j} e^{(i)} \\
+ \sum_{i=1}^{n-2} \varepsilon_{wb}^{n-(i+2)} \sum_{k > t} \sum_{k \in [0, n-(i+1)]} \Delta_{n-k} \Delta_{n-t} e^{(i)} \\
+ \cdots + \prod_{i=0}^{n-1} \Delta_{n-i} e^{(1)} .
\]

Of course, commutativity cannot be assumed; hence (2.28) should be written as

\[
(2.29) \quad r^{(n)} = \sum_{i=1}^{n} \varepsilon_{wb}^{n-i} e^{(i)} + \sum_{i=1}^{n-1} \sum_{j=0}^{n-(i+j+1)} \varepsilon_{wb}^{n-(i+j+1)} e^{(i)} \\
+ \cdots + \prod_{i=0}^{n-1} \Delta_{n-i} e^{(1)} .
\]

After premultiplication by \( D^{1/2} \), (2.29) yields the \( L^2 \)-norm bound

\[
(2.30) \quad \|D^{1/2} r^{(n)}\|_{L^2} \leq \sum_{i=1}^{n} \|\varepsilon_{wb}^{n-i}\|_{D} e + \sum_{i=1}^{n-1} \binom{n-i}{1} \|\varepsilon_{wb}^{n-(i+1)}\|_{D} \delta e \\
+ \sum_{i=1}^{n-2} \binom{n-i}{2} \|\varepsilon_{wb}^{n-(i+2)}\|_{D} \delta^2 e \\
+ \cdots + \sum_{i=1}^{n-(n-1)} \binom{n-i}{n-1} \delta^{n-1} e ,
\]
where \( \delta = \sup_{i} \| \Delta_i \| \) and \( e = \sup_{i} \| e^{(i)} \| \). The last term in (2.30) has been written as a sum merely as a notational device. Reversing the order of summation in (2.30), we obtain

\[
\| D_{2r}^{(n)} \|_{L2} \leq \sum_{j=0}^{n-1} \| e \|_{w_{b}} \sum_{j=1}^{n-1} \binom{j}{1} \| e \|_{w_{b}} \delta e \\
+ \sum_{j=2}^{n-1} \binom{j}{2} \| e \|_{w_{b}} \delta^2 e + \ldots \\
+ \sum_{j=k}^{n-1} \binom{j}{k} \| e \|_{w_{b}} \delta^k e + \ldots \\
+ \sum_{j=n-1}^{n-1} \binom{j}{n-1} \delta^{n-1} e ,
\]

or, more simply

\[
\| D_{2r}^{(n)} \|_{L2} \leq \sum_{k=0}^{n-1} \sum_{j=k}^{n-1} \binom{j}{k} \| e \|_{w_{b}} \delta^k e .
\]

Since \( s(\| e \|_{w_{b}}) \) is bounded above by unity and below by \( 1 - \text{ch} \), it is easily shown that

\[
\| e \|_{w_{b}} \leq (2 + \sqrt{5}) \text{ns}(\| e \|_{w_{b}})^n ,
\]

hence

\[
\| D_{2r}^{(n)} \|_{L2} < (2 + \sqrt{5}) \sum_{k=0}^{n-1} \sum_{j=k+1}^{n-1} \binom{j}{k} (j - k) \text{ns}(\| e \|_{w_{b}})^{j-k} \delta e .
\]

This bound may be made uniform by replacing \( n-1 \) with infinity. The inner sum begins at \( k+1 \) due to the presence of the factor \( (j-k) \) in the summand. We note that the inner sum may be
written as

\[ (2.35) \sum_{j=k+1}^{\infty} \binom{j}{k} (j-k)s(\gamma_{\omega_b})^{j-k}k_e = \sum_{j=k+1}^{\infty} \left[ \frac{i(j-1)\cdots(j-k)}{k!} s(\gamma_{\omega_b})^{j-k} \right] \delta^k e. \]

The factor in brackets in the sum on the right above is\( s(\gamma_{\omega_b}) \) times the corresponding term of the \( k+1 \)-st derivative of the geometric series \( \sum_{n=0}^{\infty} s(\gamma_{\omega_b})^n \); hence

\[ (2.36) \| D^{\frac{1}{2}}r^{(n)} \|_{L^2} < (2+\sqrt{5}) \sum_{k=0}^{\infty} \frac{\delta^k e}{k!} s(\gamma_{\omega_b})^k D_{k+1} \sum_{j=0}^{\infty} s(\gamma_{\omega_b})^j. \]

However,

\[ (2.37) \quad D_{k+1} \sum_{j=0}^{\infty} s(\gamma_{\omega_b})^j = \frac{(k+1)!}{(1-s(\gamma_{\omega_b}))^{k+2}}; \]

consequently,

\[ (2.38) \quad \| D^{\frac{1}{2}}r^{(n)} \|_{L^2} < \sum_{k=0}^{\infty} \frac{(2+\sqrt{5})s(\gamma_{\omega_b})^k e(k+1)\delta^k}{[1-s(\gamma_{\omega_b})]^{k+2}}. \]

Having already obtained a uniform bound \( \| \Delta_n \|_2 \), a choice for \( \delta \) is immediate. Thus, we need only demonstrate a suitable uniform bound for the \( L^2 \)-norm of \( D^{\frac{1}{2}}e^{(i)} \) to obtain a useful bound for the global rounding error from line (2.38).

Returning to line (2.12), we note that \( e^{(i)} \) depends on the exact \( i \)-th iterate, the exact \( (i-1) \)-st residual \( (\gamma u^{(i-1)} - c) \), and error perturbations for which 2-norm bounds have been established. For a smooth initial guess both \( u^{(n)} \) and \( h^{-2}(\gamma u^{(n-1)} - c) \) are uniformly bounded in \( L^2 \)-norm since the
iteration is convergent. Hence, the first term in the expression for \( e^{(i)} \) in line (2.12) behaves like \( o(\nu) \) in \( L^2 \)-

norm; the second like \( o(\nu h) \) since \( \Lambda \) has a uniform 2-norm 

bound of the form \( o\left[ \frac{\nu}{h} \right] \); and the third term like \( o\left( \frac{\nu^2}{h} \right) \) since 

both \( \delta G_{i} \) and \( \Xi_i \) behave like \( o(\nu^2) \) in 2-norm. Because the 

elements of the diagonal matrix \( D \) are strictly bounded away 

from zero and positive due to the uniform ellipticity of 

(2.1), \( \| D^{\frac{1}{2}} \| _2 = o(1) \); hence 

(2.39) \[ \| D^{\frac{1}{2}} e^{(i)} \| _{L^2} \leq o(\nu) , \]

uniformly in \( i \) for \( \nu \) proportional to \( h^2 \). We remark that we 

may take \( \delta = o\left[ \frac{\nu}{h} \right] \).

Finally, because a uniform \( L^2 \)-norm bound for the global 

rounding error in terms of \( \nu \) and \( h \) is sought, we must investi-
gate the dependence of \( s(\omega_b) \) on \( h \). From line (2.7) it is 
clear that what is needed is an estimate for the spectral 

radius of the Jacobi matrix \( D^{-1}(L + U) \). We shall obtain 

this bound by a reduction argument similar to the one given 
in Young [16]. Assuming initially that \( q(x_1,x_2) = 0 \), we 

may split the coefficient matrix \( A \) of the linear system 

(2.2) into the summands 

(2.40) \[ A = A_1 + A_2 \] ,

where, as in our discussion of the alternating direction 

methods, \( A_1 \) is the matrix operator corresponding to the
second differences in the $x_1$-direction and $A_2$ the second differences in the $x_2$-direction. Both $A_1$ and $A_2$ are weakly diagonally dominant. Furthermore, if we define

\[(2.41) \quad \overline{a}_1 = \max_{\Omega} a_1, \quad \underline{a}_1 = \min_{\Omega} a_1, \quad \overline{a}_2 = \max_{\Omega} a_2, \quad \underline{a}_2 = \min_{\Omega} a_2,\]

then the replacement of $a_1$ with $\overline{a}_1$ in $A_1$ will not diminish the spectral radius of the resulting matrix, call it $\overline{A}_1$, since $\overline{A}_1 - A_1$ is non-negative definite. A similar statement may be made for $A_2$. Let us now embed the region $\Omega$ in the smallest rectangle containing it, one of dimension $[Nh,Mh]$, say. If the longest row in $\Omega^0_h$ has $K-1$ grid points, then certainly $K$ is bounded above by $N$. By indexing the mesh points in this row from left to right by

\[(2.42) \quad x_1 = (x + ih, y),\]

the point $(x,y)$ being the leftmost grid point in the row, it follows easily that

\[(2.43) \quad 2a_1 v(x,y) - a_1 v(x+h,y) - a_1 v(x-h,y) = 4a_1 \sin^2 \left[ \frac{k\pi}{2K} \right] v(x,y), \quad x \in \{x_1, \ldots, x_{k-1}\}, \quad k \in \{1, \ldots, K-1\}\]

where

\[(2.44) \quad v(x,y) = \begin{cases} 
\sin \left[ \frac{k\pi}{2K} \right]; & x = x_i \\
0 & \text{elsewhere}.
\end{cases}\]
Consequently,

\[(2.45) \quad s(\overline{A}_1) \leq 4a_1 \cos^2\left[\frac{\pi}{2N}\right].\]

If we replace \(a_1\) with \(\overline{a}_1\), the smallest eigenvalue of the resulting \(A_1\) is not larger than the smallest eigenvalue of \(A_1\); hence

\[(2.46) \quad 4a_1 \sin^2\left[\frac{\pi}{2N}\right] \leq \lambda_{\min}(\overline{A}_1).\]

By a similar argument

\[(2.47) \quad 4\overline{a}_2 \sin^2\left[\frac{\pi}{2M}\right] \leq \lambda_{\min}(\overline{A}_2) < s(\overline{A}_2) \leq 4\overline{a}_2 \cos^2\left[\frac{\pi}{2M}\right].\]

Young [16] has shown that

\[(2.48) \quad s(D^{-1}(L + U)) = \frac{\kappa[D^{-1/2}AD^{-1/2}] - 1}{\kappa[D^{-1/2}AD^{-1/2}] + 1},\]

where \(\kappa[\cdot]\) is the spectral condition number of its argument, that is, the ratio of the largest to the smallest eigenvalue in magnitude; hence

\[(2.49) \quad s(D^{-1}(L+U)) \leq 1 - \frac{2a_1 \sin^2\left[\frac{\pi}{2N}\right] + 2\overline{a}_2 \sin^2\left[\frac{\pi}{2M}\right]}{\frac{1}{2}(\overline{a}_1 + a_1) + \frac{1}{2}(\overline{a}_2 - a_2) + \frac{1}{2}(\overline{a}_1 - a_1) \cos\left[\frac{\pi}{N}\right] + \frac{1}{2}(\overline{a}_2 - a_2) \cos\left[\frac{\pi}{M}\right]}.\]

In order to generalize to the case \(q \neq 0\), we shall designate the diagonal of \(A\) by \(D_0\) when \(q = 0\). Clearly,

\[(2.50) \quad D^{-1}(L + U) = (D^{-1}D_0)(D_0^{-1}(L + U)).\]
The factor in brackets above is the Jacobi matrix for which the result of line (2.49) was obtained. But the largest element of $D^{-1}D_0$ is certainly bounded above by
\[
\frac{2(\overline{a_1} + \overline{a_2})}{2(\overline{a_1} + \overline{a_2}) + h^2(-q)},
\]
where the bar under the $q$ has the obvious meaning. For the general case, then, we need only multiply the result of line (2.49) by this bound on the elements of $D^{-1}D_0$ to obtain
\[
(2.51) \quad s(D^{-1}(L+U)) \leq 1 - c(a_1, a_2)h^2 + o(h^4).
\]
Consequently, the bound
\[
(2.52) \quad s(\xi_{w_1}) \leq \frac{1 - (1-\mu^2)^{\frac{1}{2}}}{1 + (1-\mu^2)^{\frac{1}{2}}} \approx 1 - 2(1-\mu^2)^{\frac{1}{2}} \leq 1 - 2\sqrt{2k_1} h + o(h^2)
\]
follows from (2.7).

Returning to line (2.38) with this result and the uniform bounds $\delta = O(\frac{\nu}{h^2})$, $e = O(\nu)$, we obtain
\[
(2.53) \quad \|D_{x_1}^2(\gamma(n))\|_{L^2} \leq (2 + \sqrt{5})s(\xi_{w_1}) \sum_{k=0}^{n-1} (k+1)[\frac{c\nu}{h^2}]^{k+1},
\]
where $c$ is a constant independent of $\nu$ and $h$. The right-hand side of (2.53) is recognized, roughly speaking, as the derivative of the geometric series whose generic term is $c\frac{\nu}{h^2}$ (assuming this quantity is less than 1). Specifically,
\[
(2.54) \quad \|D_{x_1}^2(\gamma(n))\|_{L^2} \leq O\left[\frac{z}{(1-z)^2}\right]
\]
where \( z = \frac{\nu}{h^2} < 1 \). The comparability between the \( D \) and \( L^2 \)-norms does not alter this result insofar as dependence on \( h \) is concerned; hence

\[
\| r^{(n)} \|_{L^2} < \frac{z}{(1-z)^2}.
\]  

(2.55)

Summarizing this we have

**Theorem 10:** If the single optimal parameter successive overrelaxation iteration for the uniformly elliptic self-adjoint Dirichlet problem (2.1) is carried out in single precision arithmetic in scaled residual form (see (2.4)) and

(i) the residual \( \bar{u}^{(n)} - c \) is computed in \( f_{L^2} \)-arithmetic at each step,

(ii) the single precision solution by Gaussian elimination of the linear system implied by the approximate inverse is followed by a purification step (see Theorem 2),

(iii) \( \nu \) is taken proportional to the square of the uniform mesh size, \( h \), in sufficiently small ratio ,

then the global rounding error satisfies the uniform bound

\[
\| r^{(n)} \|_{L^2} < \frac{z}{(1 - z)^2},
\]

(2.56)

where \( z = \frac{\nu}{h^2} \) for a constant \( c \) independent of \( \nu \) and \( h \).

Though we have not emphasized the fact, our analysis has been limited to point iterative successive overrelaxation as opposed to block iterative methods. We might
speculate that some advantage ought to accrue from the fact that for the point iterative procedure (2.4), no component of the new iterate at a given step depends on more than four components of the previous iterate for the Dirichlet problem (2.1); thus, we should be able to employ $f_{l_{2}}$-arithmetic throughout the computation without a significant temporary double precision storage requirement. This is not possible for the alternating direction methods without the temporary storage of vector of length $\text{card}(\Omega_n^0)$ in double precision.

In $f_{l_{2}}$-arithmetic the point iterative successive over-relaxation method is realized as

\begin{equation}
\hat{u}^{(n)} = (I + \theta_n)(r + \psi_n)\hat{u}^{(n-1)} - \omega_b (I + \xi_n)(I + R_n)(I - \omega_b \pi)^{-1}\left((\alpha + \delta_n \Omega_n)\hat{u}^{(n-1)} - (I + \xi_n) \cdot c\right)
\end{equation}

where $R_n$ plays the same role as before except that we shall not assume purification is performed; $\delta_n \Omega_n$ and $\xi_n$ are perturbations introduced in the $f_{l_{d}}$ computation of the residual. Because we are not assuming that the machine can accumulate sums or inner products in quadruple precision, the $f_{l_{d}}$ realization of the vector sum $v_1 + v_2$ is expressed by

\begin{equation}
\text{f}_{l_{d}}(v_1 + v_2) = (1 + \kappa_1)v_1 + (1 + \kappa_2)v_2,
\end{equation}

\[\|\kappa_1\|_{\infty}, \|\kappa_2\|_{\infty} \leq \varepsilon(\nu^2),\]
instead of \( f l_d(v_1 + v_2) = (I + \kappa_1)(v_1 + v_2) \), the sort of result expected when the accumulation of a sum is done in twice the precision of the summands. Thus, the role of the perturbations \( \psi_n \) and \( \xi_n \) becomes clear. Finally, \( \theta_n \) corresponds to the relative error resulting when the entire quantity in braces is rounded to single precision for storage. Defining \( \Lambda_n \) implicitly by \( (I + \psi_n)(I + R_n)(I - \omega_b G)^{-1} = (I - \omega_b G)^{-1} + \Lambda_n \), we may rearrange (2.57), after having subtracted \( u^{(n)} \) from both sides, to obtain

\[
(2.59) \quad r^{(n)} = (I + \theta_n)(\omega_b^+ + \Lambda_n)r^{(n-1)} + e^{(n)}
\]

where

\[
(2.60) \quad \Lambda_n = \psi_n - \omega_b(I - \omega_b G)^{-1}\delta G_n - \omega_b \Lambda_n(G + \delta G_n)
\]

and

\[
(2.61) \quad e^{(n)} = \theta_n u^{(n)} + (I + \theta_n)\left[ \psi_n u^{(n-1)} - \omega_b \Lambda_n(Gu^{(n-1)} - c)
- \omega_b(I - \omega_b G)^{-1} + \Lambda_n \right] \left[ \delta G_n u^{(n-1)} - \Xi c \right].
\]

As usual, \( r^{(n)} = \hat{u}^{(n)} - u^{(n)} \).

Certainly \( \psi_n, \xi_n, \) and \( \Xi_n \) all share 2-norm bounds of the form \( \Theta(\nu^2) \), for all are diagonal matrices whose entries are bounded by \( \Theta(\nu^2) \) in magnitude. The perturbation \( \delta G_n \) is given by

\[
(2.62) \quad \delta G_n = (k_{i,j}^n \nu^2 a_{i,j})
\]

where \( G = (a_{i,j}) \) and \( k_{i,j}^n \) is a constant for each \( i,j,n \) which is independent of \( \nu \); hence \( \|\delta G_n\|_2 \leq \max_{i,j} |k_{i,j}^n| \|G\|_2 \nu^2 \). The error
analysis of the solution of the system implied by the approximate inverse given earlier applies here except that the perturbations \( \gamma_j, k, \mu_j, k \), and \( \lambda_j, k \) defined in line (2.16) are bounded in magnitude by \( \Theta(\gamma^2) \) instead of \( \Theta(\gamma) \). Thus,

\[
\| R_n \|_2 = \| (I + (I - \omega_b \xi)^{-1} K_n)^{-1} - I \|_2 \leq \Theta(\gamma^2). \tag{2.63}
\]

Since

\[
\Lambda_n = [R_n + \delta_n(I + R_n)](I - \omega_b \xi)^{-1}, \tag{2.64}
\]

bounds for \( \| \Delta_n^{(n)} \|_2 \) and \( \| e^{(n)} \|_{L^2} \) follow. In particular, the third term on the right in line (2.60) can be written as

\[
R_n + \delta_n(I + R_n)[(\xi - \omega_b(I - \omega_b \xi)^{-1} \delta G_n)], \tag{2.65}
\]

from which a 2-norm bound of the form \( \Theta(\gamma^2) \) is easily deduced. The second term shares this bound while the first is bounded by \( \Theta(\gamma^2) \). Consequently,

\[
\| \Delta_n^{(n)} \|_2 \leq \Theta(\gamma^2). \tag{2.66}
\]

uniformly. By an argument analogous to the one preceding line (2.39) we conclude that

\[
\| e^{(n)} \|_{L^2} \leq \Theta(\gamma). \tag{2.67}
\]

This is precisely the result obtained in the single precision case. We do not obtain a bound of the form \( \Theta(\gamma^2) \) due to the presence of \( \delta_n \), which behaves in 2-norm like \( \Theta(\gamma) \).
Expanding the global error $r^{(n)}$ in terms of the 'local' errors $e(j)$, we obtain, with redefined $\Delta_n$, the same result here that we did in the single precision argument, namely

\begin{equation}
\|D^{\frac{1}{2}}r^{(n)}\|_{L^2} < \sum_{k=0}^{\infty} \frac{(2+\sqrt{5})s(\varphi_{w_b})(k+1)e_k}{[1 - s(\varphi_{w_b})]^{k+2}},
\end{equation}

where $e$ and $\delta$ are uniform bounds for the $D$ and $L^2$-norms of $\Delta_j$ and $D^{\frac{1}{2}}e(j)$, respectively. Actually, we must redefine $\Delta_n$ to include the perturbation $\varphi_{\Delta n} \varphi_{w_b}$ if (2.68) is to be valid.

From lines (2.66) and (2.67) and the uniform ellipticity of the problem, it follows that

\begin{equation}
\|D^{\frac{1}{2}}r^{(n)}\|_{L^2} \leq (2 + \sqrt{5})k_3s(\varphi_{w_b})\sum_{j=0}^{\infty} \frac{(j+1)[k_1 \varphi_{w_b}]}{[k_2h]^{2j+2}},
\end{equation}

where $k_1$ and $k_2$ are constants independent of $h$ and $\varphi$ and $k_3$ is a bound on the comparability of the $D$-norm to the $2$-norm (which is independent of $h$ here). Factoring $\frac{k_1 \varphi_{w_b}}{k_2h^2}$ from each summand yields

\begin{equation}
\|D^{\frac{1}{2}}r^{(n)}\|_{L^2} \leq (2 + \sqrt{5})s(\varphi_{w_b})k_3\sum_{j=0}^{\infty} \left[ \frac{k_1 \varphi_{w_b}}{k_2h^2} \right]^{j+1} \left[ \frac{k_1 \varphi_{w_b}}{k_2h^2} \right]^j.
\end{equation}

For $w = \left[ \frac{k_1 \varphi_{w_b}}{k_2h^2} \right]^2$, the sum on the right above is recognized as the derivative of the geometric series $\sum w^j$, provided $w < 1$; hence

\begin{equation}
\|r^{(n)}\|_{L^2} < c \frac{\varphi_{w_b}}{h^2} \cdot \frac{1}{[1 - w]^2},
\end{equation}
for a sufficiently small ratio of \( \nu \) to \( h \). This result is rather surprising, for it says that we obtain no improvement in the bound exhibited in Theorem 10 insofar as dependence on the ratio \( \frac{\nu}{h^2} \) is concerned even if \( ft_\nu \)-arithmeti is employed throughout the computation and not just in the accumulation of the residual. In summary, then, we have

**Theorem 11:** If

(i) each step of the point iterative successive relaxation technique is performed in double precision arithmetic then stored in single precision, but no purifications are performed in the computation of the solution implied by the approximate inverse \( (I - \omega_b \xi)^{-1} \),

(ii) \( \nu \) is taken proportional to \( h^2 \) in sufficiently small ratio,

then the global rounding error behaves no worse than

\[
\| r^{(n)} \|_{L^2} \leq \Theta \left( \frac{\nu}{h^2} \right),
\]

uniformly in \( n \).

**Remark:** We note that iterative improvement would not yield an improved 2-norm bound for \( R_n \) here, but none is needed since the bound already is no larger than \( \Theta(\sqrt{\nu} h) \) for \( \nu \) proportional to \( h^2 \).
Section 2: Semi-Iterative Symmetric Successive Overrelaxation

If we reverse the ordering of the difference system with each successive step of the successive overrelaxation iteration of the previous section, that is, we compute the \((n+1)\)st and \((n+2)\)nd iterates by

\[
\begin{align*}
\tag{2.73}
\mathbf{u}^{(n+1)} &= (\mathbf{I} - \omega \mathbf{L})^{-1}(\omega \mathbf{U} + (1-\omega)\mathbf{L})\mathbf{u}^{(n)} + (\mathbf{I} - \omega \mathbf{L})^{-1}\mathbf{c} \\
\text{and}
\tag{2.74}
\mathbf{u}^{(n+2)} &= (\mathbf{I} - \omega \mathbf{U})^{-1}(\omega \mathbf{L} + (1-\omega)\mathbf{U})\mathbf{u}^{(n+1)} + (\mathbf{I} - \omega \mathbf{U})^{-1}\mathbf{c}
\end{align*}
\]

respectively, we obtain an iteration over the two steps, known as the SSOR method, whose error propagator

\[
\begin{align*}
\tag{2.75}
\mathbf{S}_\omega &= (\mathbf{I} - \omega \mathbf{U})^{-1}(\omega \mathbf{L} + (1-\omega)\mathbf{U})(\mathbf{I} - \omega \mathbf{L})^{-1}(\omega \mathbf{U} + (1-\omega)\mathbf{L})
\end{align*}
\]

is symmetric for the model problem. For the more general self-adjoint problem (2.1) \(\mathbf{S}_\omega\) is not necessarily symmetric but does still possess a real spectrum. Indeed, similarity transformation of \(\mathbf{S}_\omega\) with \((\mathbf{I} - \omega \mathbf{U})\) yields

\[
\begin{align*}
\tag{2.76}
(\mathbf{I} - \omega \mathbf{U})\mathbf{S}_\omega(\mathbf{I} - \omega \mathbf{U})^{-1} &= (\omega \mathbf{L} + (1-\omega)\mathbf{U})(\mathbf{I} - \omega \mathbf{L})^{-1} \\
&\quad \cdot (\omega \mathbf{U} + (1-\omega)\mathbf{L})(\mathbf{I} - \omega \mathbf{U})^{-1}
\end{align*}
\]

which is recognized as the product of \((\omega \mathbf{L} + (1-\omega)\mathbf{U})(\mathbf{I} - \omega \mathbf{L})^{-1}\) and its transpose. Such matrix products necessarily have real spectra. Consequently, we may semi-iterate this technique (see Varga [14]) to achieve a virtual rate of convergence that is approximately the square root of the rate of convergence of the symmetric successive overrelaxation
technique itself. For sufficiently small h, the improved rate of convergence more than compensates for the two and a half times more work required of the SSOR method per iteration.

**Description of the SSOR-SI Method for Self-Adjoint Elliptic Problems**

We semi-iterate the stationary, single step, completely consistent iteration

\[(2.77) \quad u^{(n)} = G u^{(n-1)} + k \quad , \quad \sigma(G) \subseteq \mathbb{R}\]

by combining the iterates \(\{u^{(j)\}}_{j}^{n}\) linearly to form a new sequence \(\{v^{(j)\}}_{j}^{n}\) whose \(n\)th term is given by

\[(2.78) \quad v^{(n)} = \sum_{k=0}^{n} a_{n,k} u^{(k)}\]

The coefficients \(a_{n,k}\) are chosen to minimize the virtual average spectral radius of the resulting error propagator, which is a polynomial in \(G, P_n(G)\), of degree \(n\) at the \(n\)th step of the iteration. The problem of minimizing the virtual spectral radius of \(P_n(G)\) can be solved with Chebyshev polynomials. It is well known that if the spectral interval of \(G\) is contained in the real interval \(0 \leq \alpha < \beta < 1\), then the polynomial \(P_n(G)\) of smallest virtual spectral radius is given by

\[(2.79) \quad P_n(G) = \frac{T_{n} \left[ \frac{2G - (\alpha + \beta)}{\beta - \alpha} \right]}{T_{n} \left[ \frac{2 - (\alpha + \beta)}{\beta - \alpha} \right]} ,\]
where \( T_n \) is the Chebyshev polynomial of degree \( n \). It would appear that semi-iteration is quite costly to compute, it being necessary not only to perform the iterations of (2.77) but also to compute the coefficients \( a_{n,k} \) of (2.78). Fortunately this is not the case. Moreover, it is possible to show that the \( v^{(n)} \) given by (2.78) satisfy the second order difference equation

\[
(2.80) \quad v^{(n)} = \gamma_n [G v^{(n-1)} + k] + (1 - \gamma_n) v^{(n-2)}
\]

where

\[
(2.81) \quad \gamma_n = \frac{2}{s(G)} \frac{T_{n-1} \left[ \frac{1}{s(G)} \right]}{T_n \left[ \frac{1}{s(G)} \right]},
\]

and, of course, \( v^{(1)} = u^{(1)} \), \( v^{(0)} = u^{(0)} \). This two-step recursion is inherited from the well-known fact that

\[
(2.82) \quad T_n(z) = 2zT_{n-1}(z) - T_{n-1}(z), \quad n \geq 1.
\]

Before we examine the rounding error behavior of the SSOR-SI method, we remark that Young [16] has shown that for positive definite \( A \) the optimal choice of the iteration parameter \( w \) is approximately

\[
(2.83) \quad w_1 = 2/[1 + (1 - 2\mu + 4\gamma)^{1/2}] \]

where \( \mu \) is the spectral radius of the Jacobi matrix associated with the system and \( \gamma = \max \{s(LU), \frac{1}{2}\} \). For this choice of \( w \) we obtain the bound

\[
(2.84) \quad s(\mathcal{A}_{w_1}) \leq \left[ 1 - \left[ \frac{1-\mu}{2} \right]^{1/2} \right] / \left[ 1 + \left[ \frac{1-\mu}{2} \right]^{1/2} \right].
\]
Some Remarks on the Asymptotic Rounding Error Behavior of the SSOR Semi-Iterative Method

The symmetric successive overrelaxation iteration of lines (2.73), (2.74) for the Dirichlet problem (2.1) is given in residual form by

\[(2.85) \quad u(n) = u(n-1) - B^{-1}(Au(n-1) - b)\]

where \(B^{-1} = w_1(2 - w_1)(I - w_1U)^{-1}(I - w_1L)^{-1}D^{-1}\). For this strategy of computation the semi-iterative successive overrelaxation method is realized in single precision arithmetic as

\[(2.86) \quad \hat{v}(n+1) = \gamma_{n+1}(I + \chi_{n+1})(I + \theta_{n+1})\{(I + \psi_{n+1})(\hat{v}(n)\]

\[- \hat{B}^{-1}(I + \xi_{n+1})[(A + \delta A_{n+1})\hat{v}(n) - (I + \xi_{n+1})b]\}

\[+ (1 - \gamma_{n+1})(I + \chi_{n+1})(I + \Sigma_{n+1})\hat{v}(n-1), \quad n \geq 1\]

where (i) \(\hat{B}^{-1}\) is the single precision analogue of the approximate inverse \(B^{-1}\); (ii) the expression in brackets is a generalization of line (xiv) of the Introduction; (iii) \(\psi_{n+1}\) is the relative error introduced when the computed difference correction is added to the \(n^{th}\) iterate; (iv) \(\theta_{n+1}\) and \(\Sigma_{n+1}\) correspond to the relative errors introduced in multiplication by \(\gamma_{n+1}\) and \(1 - \gamma_{n+1}\), respectively, and (v) \(\chi_{n+1}\) is the relative error introduced when the vectors \(f_\ell[\gamma_{n+1}f_\ell(\gamma_{w_1}\hat{v}(n)) + k]\) and \(f_\ell[(1 - \gamma_{n+1})\hat{v}(n-1)]\) are summed. Subtracting the exact \((n+1)^{st}\)
iterate $v^{(n+1)}$ from (2.86), we obtain

\begin{equation}
(2.87) \quad r^{(n+1)} = \gamma_{n+1} [3_{w_1} + \Delta_{n+1}] r^{(n)} + (1 - \gamma_{n+1}) (I + \Gamma_{n+1}) r^{(n-1)} + e^{(n+1)}
\end{equation}

where $\Gamma_{n+1}$ is defined implicitly by $(I + \chi_{n+1})(I + \Sigma_{n+1}) = (I + \Gamma_{n+1})$,

\begin{equation}
(2.88) \quad \Delta_{n+1} = \chi_{w_1} - (I + \chi_{n+1})(I + \eta_{n+1})(I + \psi_{n+1}) \\
\quad \quad \quad \cdot [I - \hat{B}^{-1}(I + \xi_{n+1})(A + \delta A_{n+1})]
\end{equation}

and

\begin{equation}
(2.89) \quad e^{(n+1)} = \gamma_{n+1} (I + \chi_{n+1})(I + \eta_{n+1})(I + \psi_{n+1}) \\
\quad \quad \quad \cdot [v^{(n)} - \hat{B}^{-1}(I + \xi_{n+1})[Av^{(n)} - b] + (\delta A_{n+1} v^{(n)} \\
\quad \quad \quad - \xi_{n+1} b)] + (1 - \gamma_{n+1}) (I + \Gamma_{n+1}) v^{(n-1)} - v^{(n+1)}.
\end{equation}

For any norm we obtain from (2.89) the second order inhomogeneous difference inequality

\begin{equation}
(2.90) \quad \| r^{(n+1)} \| \leq \gamma_{n+1} \left\{ \| 3_{w_1} \| + \| \Delta_{n+1} \| \right\} r^{(n)} \\
\quad \quad \quad + (\gamma_{n+1} - 1) (1 + \| \Gamma_{n+1} \|) \| r^{(n-1)} \| + \| e^{(n+1)} \|.
\end{equation}

Note that the coefficient $1 - \gamma_{n+1}$ must be factored through the norm as $\gamma_{n+1} - 1$. This sequence of normed rounding error vectors is majorized by the sequence $\{g^{(n)}\}$ of real numbers satisfying
\[ g^{(n+1)} = 2[\| s_{w_1} \| + \delta] g^{(n)} + (\gamma_2 - 1)(1 + \gamma) g^{(n-1)} + \| e^{(n+1)} \|, \]

\[ g^{(0)} = \| r^{(0)} \|, \]

\[ g^{(1)} = \| r^{(1)} \|. \]

The quantities \( \delta \) and \( \gamma \) are uniform bounds for the norms of \( \Lambda_{n+1} \) and \( \Gamma_{n+1} \), respectively. The characteristic equation of this second order inhomogeneous difference equation is given by

\[ \lambda^2 - 2[\| s_{w_1} \| + \delta] \lambda - (1 + \gamma)(\gamma_2 - 1) = 0. \]

It is well-known that the roots of the real quadratic \( x^2 + ax + b \) are bounded by unity if and only if \( |b| < 1 \) and \( |a| < 1 + b \). For (2.92) this last condition is equivalent to

\[ 2[\| s_{w_1} \| + \delta] \leq 1 - (1 + \gamma)(\gamma_2 - 1). \]

This constraint cannot be satisfied for the usual choice of norm, that is, one for which \( \| s_{w_1} \| \) approximates \( s(s_{w_1}) \), for the right hand side of (2.93) is less than one for a sufficiently small uniform bound \( \gamma \), but the left-hand side is near 2 assuming that \( \delta \) is sufficiently small. Consequently, we do not expect to be able to extend the analysis of Section 1 to the SSOR-SI procedure.

\*Varga [14] has shown that the parameters \( \gamma_{n+1} \) are bounded above by 2 and tend monotonically to \( \omega_b \) for \( n \geq 1 \).
Instead, we shall investigate the propagation of an error $e_k$ introduced in the $k$th step of the computation assuming that no further error is introduced in successive steps. This will at the very least provide some insight into appropriate necessary conditions for the stability of the SSOR-SI technique with respect to rounding errors. Let us also assume that $e_k$ has occurred late enough in the computation that $y_k$ is indistinguishable from $w_b$ to the precision of the machine.

It is clear that for exact machine computation, the iterates following the $k$th are given by

\[
\hat{u}(k+1) = w_b \left[ 3_{w_1} \hat{u}(k) + \hat{e}_k \right] + (1-w_b)u^{(k-1)} = u^{(k+1)} + w_b 3_{w_1} e_k
\]

\[
\hat{u}(k+2) = w_b \left[ 3_{w_1} \hat{u}(k+1) + \hat{e}_k \right] + (1-w_b)\hat{u}(k)
\]

\[
\vdots
\]

\[
\hat{u}(m) = u^{(m)} + r^{(m-k)},
\]

where

\[
r(j) = w_b 3_{w_1} r^{(j-1)} + (1-w_b)r^{(j-2)},
\]

\[
r^{(0)} = e_k, \quad r^{(1)} = w_b 3_{w_1} e_k
\]

and, as usual, a caret denotes a perturbed value. Since the eigenvalues of $3_{w_1}$ are real, we may consider the propagation of an $e_k$ proportional to the normalized real eigenvector
\( v_s(3_{w_1}) \) corresponding to the spectral radius of \( 3_{w_1} \). For this eigenvector the recursion (2.95) may be written as

\[
(2.96) \quad r(t) = w_b s(3_{w_1}) r(t-1) + (1 - w_b) r(t-2);
\]

\[
r(0) = v_s(3_{w_1}), \quad r(1) = w_b s(3_{w_1}) v_s(3_{w_1}).
\]

Interpreting (2.96) componentwise, we have obtained a second order homogeneous difference whose characteristic equation is given by

\[
(2.97) \quad \hat{\varphi}^2 - w_b s(3_{w_1}) \hat{\varphi} + (w_b - 1) = 0.
\]

This equation is familiar to us as the one arising in the analysis of the optimal choice of iteration parameter \( w \) for the single parameter successive overrelaxation iteration. In particular, the roots of this quadratic are known to be

\[
(2.98) \quad \hat{\varphi} = (w_b - 1)^{\frac{1}{2}}.
\]

Consequently, the solution of (2.96) is given by

\[
(2.99) \quad r(t) = r(0)(t+1)^{\frac{1}{2}} + (r(1) + w_b s(3_{w_1}) r(0)) t^{\frac{1}{2}} - 1,
\]

for \( r(0) = v_s(3_{w_1}) \) and \( r(1) = w_b s(3_{w_1}) v_s(3_{w_1}) \). Since \( w_b \) is given by

\[
(2.100) \quad w_b = 2/(1 + (1 - s(3_{w_1})^2)^{\frac{1}{2}}),
\]

a value very near 2 for small \( h \) (but less than 2), we deduce that such an error as the one we have introduced above is
magnified as long as \((t + 1)\) exceeds \(\hat{\delta}^t\), but eventually peaks and then decreases. Moreover, if all of these errors were to reinforce each other for the initialization \(r^{(0)} = \nu s(3_{w_1})\) we would have

\[
\sum_{t=0}^{\infty} r(t) = \sum_{t=0}^{\infty} (t+1) \hat{\delta}^t \nu s(3_{w_1}) + 2 \nu t \hat{\delta}^{t-1} w_b s(3_{w_1}) v s(3_{w_1})
\]

\[
= \left[ 1 + \frac{1}{(1-\hat{\delta})^2} \right] \nu s(3_{w_1}) + \left[ 1 + \frac{2 w_b s(3_{w_1})}{(1-\hat{\delta})^2} \right] \nu s(3_{w_1}).
\]

Since \(\hat{\delta} = (w_b - 1)^{1/2} = \frac{s(3_{w_1})}{1 + (1 - s(3_{w_1})^2)^{1/2}}\), we deduce that

\((1 - \hat{\delta})^2 = \hat{\delta}(h^3)\). Thus, even under optimal conditions we must take \(\nu\) proportional to \(h^3\) to assure that the introduction of an error at any step in the computation does not grow catastrophically.

We shall now argue that we cannot expect to be able to guarantee that an error introduced at the \(k\)th step will not grow without bound in max-norm when propagated by the perturbed recursion

\[
(2.101) \quad r(j) = w_b (3_{w_1} + \Delta_j) r(j-1) + (1 - w_b)(1 + \Gamma_j) r(j-2), \quad j + 2 \geq k
\]

if that error lies in the \(v s(3_{w_1})\) direction, that is, is proportional to \(v s(3_{w_1})\), in single precision arithmetic.
unless \( \nu \) is taken proportional to \( h^{5/2} \) for the self-adjoint elliptic problem (2.1). The perturbations \( \Delta_j \) and \( \Gamma_j \) are those defined in (2.88). In particular we shall investigate the behavior of \( r(t) \) for the case \( \Gamma_j = \nu I, \Delta_j = c_3 w_1 \) where \( c \) satisfies \( c \| w_1 \|_{\infty} = \sup_j \| \Delta_j \|_{\infty} \) uniformly in \( j \). Introducing the symbols \( R_j, U, R_j, L, \) and \( R_j, D \) to represent the relative errors in the solutions of the linear systems implied by the factors \( (I - w_1 U)^{-1}, (I - w_1 L)^{-1} \) and \( D^{-1} \), respectively, of the approximate inverse, we may express the absolute error in the approximate inverse by

\[
(2.102) \quad \Delta_j = B_U^{-1} B_L^{-1} R_D D^{-1} + B_U^{-1} R_L B_L^{-1} (I + R_D) D^{-1} + R_U B_U^{-1} (I + R_L) B_L^{-1} (I + R_D) D^{-1}
\]

where \( B_U = (I - w_1 U) \), and \( B_L = (I - w_1 L) \). Since \( B_U \) and \( B_L \) are diagonally dominant by at least \( 1 - \frac{1}{2} w_1 \), the max-norm bounds

\[
(2.103) \quad \| B_U^{-1} \|_{\infty}, \| B_L^{-1} \|_{\infty} \leq \frac{1}{1 - \frac{1}{2} w_1}
\]

are an immediate consequence of Lemma 0. For the self-adjoint problem

\[
(2.104) \quad \frac{\partial}{\partial x_1} \left[ a_1(x_1, x_2) \frac{\partial U}{\partial x_1} \right] + \frac{\partial}{\partial x_2} \left[ a_2(x_1, x_2) \frac{\partial U}{\partial x_2} \right] = g
\]

the \( U \) row sum at the point \( (x, y) \) of the mesh is given by
\[
(2.105) \quad \frac{a_1(x+h/2, y) + a_2(x, y+h/2)}{a_1(x+h/2, y) + a_1(x-h/2, y) + a_2(x, y+h/2) + a_2(x, y-h/2)}
\]

\[
= \frac{1}{2} + \frac{1}{4} h \left[ \frac{a_1(x+h/2, y) + a_2(x, y+h/2)}{a_1(x+h/2, y) + a_2(x, y+h/2)} \right] + o(h^2).
\]

Clearly, the L row sum at \((x, y)\) is also given by an expression of the form \(\frac{1}{2} + \frac{1}{4} ch + o(h^2)\); hence \(\kappa = \frac{1}{4} + o(h)\) (see page 113 for the definition of \(\kappa\)) and \(\omega_1 = 2 - o(h^{1/2}) + o(h)\). Consequently, the bounds given in (2.103) behave like \(h^{-1/2}\) asymptotically. Extending the analysis for the analogue of the relative error perturbation \(R_{j,L}\), given in the previous section, to \(R_{j,U}\), we deduce that both of these perturbation matrices have max-norm bounds of the form \(o\left[\frac{\nu}{h^{3/2}}\right]\). The relative error \(R_{j,D}\) is bounded in max-norm by \(o(\nu)\) since \(D\) is a diagonal matrix whose elements are uniformly bounded away from zero independently of \(h\). Thus, the absolute error in the approximate inverse, \(A_j\), has a max-norm bound of the form \(o\left[\frac{\nu}{h^{3/2}}\right]\); the perturbation \(\Delta_j\) a max-norm bound of the form \(o\left[\frac{\nu}{h}\right]\).

Assuming that all of the perturbations \(\Delta_j\) are proportional to \(\omega_1\) with max-norm no greater than \(o\left[\frac{\nu}{h}\right]\), the uniform bound just obtained, it is clear that we should write \(\Delta_j = c\omega_1\) uniformly in \(j\) where \(c = o\left[\frac{\nu}{h}\right]\). For this choice of \(\Delta_j\), \(\Gamma_j = \nu I\) and initial vectors \(v_s(\omega_1)\).
the recursion (2.101) may be written

\[
(2.106) \quad r(j) = w_b s(\tau_{w_1}) + c r(j-1) + (1 - w_b)(1 + 2\nu)r(j-2),
\]

\[ j - 2 \geq k. \]

This difference equation has characteristic polynomial

\[
(2.107) \quad \xi^2 - w_b[s(\tau_{w_1}) + c]\xi + (w_b - 1)(1 + 2\nu). \]

Both roots of this polynomial are bounded by unity in magnitude if and only if

\[
|w_b[s(\tau_{w_1}) + c]| < 1 + (w_b - 1)(1 + 2\nu)
\]

and

\[
|(w_b - 1)(1 + 2\nu)| < 1.
\]

Since \( \mu = 1 - c h^2 + o(h^4) \) we obtain a bound for the spectral radius of the symmetric successive overrelaxation technique from line (2.84) of the form

\[
(2.110) \quad s(\tau_{w_1}) \leq 1 - c h^{3/2};
\]

hence, line (2.108) cannot be satisfied unless \( \nu \) is chosen proportional to \( h^{5/2} \). This completes the argument.

We deduce from the two arguments presented above that the SSOR-SI technique does not possess the same stability with respect to the growth of rounding errors that the SOR technique itself does. We anticipate having to require at
the very least that $v$ be taken proportional to $h^{5/2}$ in sufficiently small ratio to insure that no growing rounding error is propagated in single precision.
Chapter 3. Asymptotic Rounding Error Analysis of the Factorization Procedure of Dupont, Kendall, and Rachford for Self-Adjoint Uniformly Elliptic Dirichlet Problems

Description of the Factorization Technique in $R^2$

Although the factorization technique of Dupont, Kendall, and Rachford [5] has been described and analyzed for $p$-dimensional self-adjoint uniformly elliptic problems by Dupont [7], we shall restrict ourselves to the $R^2$ case, namely

$$\begin{align*}
\mathcal{A} U &= g(x_1, x_2) \text{ in } \Omega \subset R^2 \\
U &= f(x_1, x_2) \text{ on } \partial \Omega
\end{align*}$$

where $\mathcal{A}$ is the self-adjoint uniformly elliptic operator of line (vii) of the Introduction. We shall make the usual five point approximation for this differential operator based on interpolation of its coefficients $a_1(x_1, x_2)$ and $a_2(x_1, x_2)$ at the mid-points of a uniform mesh imposed over the region (see (ix)) to obtain the symmetric linear difference system

$$Au = b.$$ 

It will be assumed that the matrix $A$ operates only on the real vector space $R$ card($\Omega_0^\circ$) where $\Omega_0^\circ$ is the subset of interior points of the mesh $\Omega_h = [(j, k)h] \cap \Omega^\circ$ , and not $R$ card($\Omega_h$) since the Neumann problem will not be considered.
In [5] it was established that the stationary iteration
\[(3.3) \quad u^{(n+1)} = u^{(n)} - \gamma(A + B)^{-1}(Au^{(n)} - b),\]
where \(B\) is a locally second order correct difference approximation for the directional derivative \(\frac{\partial^2}{\partial \xi^2}\) (see Figure 2 below) has the same rate of convergence as the single parameter techniques of the previous chapters for an appropriate choice of the iteration parameter \(\gamma\). Crucial to the success of (3.3) as a practical iteration is the existence of a sparse \(LL^*\) factorization for the approximate inverse \((A+B)^{-1}\).

\[\begin{align*}
\kappa & \quad \xi \\
\downarrow & \quad \downarrow \\
x_2 & \quad x_1
\end{align*}\]

Figure 2.

Moreover, it is the existence of such a factorization which motivates this choice of the approximate inverse. Note that a sparse \(LL^*\) factorization of \(A\) itself is not possible. For the lower triangular matrix \(L\) defined by
\[(3.4) \quad (Lu)_{j,k} = v_j, k^{u_j,k} + t_j, k^{u_j,k} + g_j, k^{u_j,k}, k^{-1}, u_j, k^{-1},
\]
and the three-diagonal matrix \(B\) defined by
\[(3.5) \quad (Bu)_{j,k} = (Bu)_{j,k} + (Du)_{j,k},
\]
\[(\tilde{Bu})_{j,k} = h_j, k^{u_j,k+1} + h_{j+1}, k^{-1}, u_j, k^{-1},
\]
\[- (-h_j, k - h_{j+1}, k^{-1})u_j, k,
\]
\[(Du)_{j,k} = \delta_j, k^{b_j,k}, u_j, k,
\]
this symmetric factorization is given by

\begin{equation}
\begin{split}
    v_{j,k}^2 &= b_{j,k}(1 + \phi_{j,k}) - h_{j,k} - h_{j+1,k-1} - t_{j-1,k}^2 - g_{j,k-1}^2 \\
    g_{j,k} &= f_{j,k} v_{j,k} \\
    t_{j,k} &= c_{j,k} v_{j,k} \\
    h_{j,k} &= t_{j-1,k} g_{j-1,k}
\end{split}
\end{equation}

The quantity \( \phi_{j,k} \) is a parameter introduced to improve the comparability of the matrix operator \( A + B \) to \( A \), and is chosen proportional to \( h^2 \) to yield the best estimates for (3.1). (It is the matrix operator \( B \) which approximates the partial derivative \( \frac{\partial^2}{\partial \xi^2} \) at interior mesh points.)

No matter how desirable factorization into sparse \( LL^* \) form is analytically, it cannot be desirable from a computational view, for a square root must be computed in the recursion for \( v_{j,k} \) in (3.6) for each point of the mesh. In [5] it is remarked that the equivalent factorization

\begin{equation}
\hat{LU} = [L \text{ diag}(L)][\text{diag}(L)^{-1} L^*],
\end{equation}

avoids the computation of square roots. The simplicity of (3.7) becomes clearer if we write

\begin{equation}
A = \hat{L}_1 + D_1 + \hat{L}_1^*,
\end{equation}

where \( \hat{L}_1 \) is the strictly lower triangular part of \( A \) and \( D_1 = \text{diag } (A) \), for in terms of this decomposition
\[ \tilde{LU} = (L_1 + F)(I + F^{-1}L_1^*) \quad , \quad F = \text{diag} \ (L)^2 . \]

Moreover, we perceive from (3.9) that insofar as the actual computation of the difference correction is concerned, the factorization can be thought of as complete once \( F \) has been determined. The elements of \( F \) may be computed directly from the coefficients of the operator \( A \) by the recursion

\[ v_{j,k}^2 = \left\{ \left\lvert \frac{c}{v_{j-1,k}} \right\rvert^2 + \left\lvert \frac{f}{v_{j,k-1}} \right\rvert^2 + \left\lvert \frac{cf}{v_j^2} \right\rvert + \left\lvert \frac{cf}{v_{j-1,k}} \right\rvert \right\} \]

\[ + (1 + \phi_{j,k}) b_{j,k} . \]

For the ordering implicit in (3.10) the values of the \( v_{j,k}^2 \) depend only on the coefficients of \( A \) and the values of \( F \) in the hyperoctant 'behind' the point \((jh,kh)\). These values are never zero since (see lemma 4, [6])

\[ \frac{v_{j,k}^2}{c_{j,k} - f_{j,k}} \geq \beta = 1 + ch , \]

\[ c: \text{positive constant}. \]

Another consequence of (3.11) is that \( U \) is diagonally dominant by \( o(h) \). We shall show that \( \tilde{L} \) is also diagonally dominant by \( O(h) \). It is clear, however, that \( A + B \) itself is not diagonally dominant. Moreover, for the model problem \( a_1 = a_2 = 1, q = 0 \) the difference operator \( h^{-2}(A + B) \) tends, away from the boundary, to the differential operator

\[ -\Delta + \frac{\partial^2}{\partial \xi^2} = -\frac{1}{2} \left[ \Delta + 2 \frac{\partial^2}{\partial x \partial y} \right] \]

with decreasing \( h \). Since this differential operator has a
zero eigenvalue we anticipate an eigenvalue of \((A + B)\) no larger than \(kh^2\), \(k\): independent of \(h\). This, of course, implies that \(A + B\) has a spectral condition number proportional to \(h^{-2}\). In view of the fact that the iterative refinement technique of Theorem 2 is inapplicable, since the single precision analogue of \((A + B)\) is not precisely \((A + B)\) in general, it appears that \((A + B)\) is not a particularly satisfactory approximate inverse insofar as rounding error stability is concerned. Moreover, merely to compute the elements of the diagonal matrix \(F\) by (3.10) requires some care, for in single precision it is conceivable that by the end of the computation rounding errors proportional to \(\sqrt{h^{-2}}\) are introduced in the floating point realization of the \(v_{j,k}^2\). To avoid this possibility we shall accumulate (3.10) in \(fl_2\)-arithmetic. This strategy will insure that no relative error in any \(v_{j,k}^2\) exceeds \(\delta(\sqrt{h^{-2}})\). It will be necessary to store in double precision at most a row of elements of \(F\) 'behind' the particular diagonal element being computed at any given step. This amount of double precision storage is not serious as only \(\delta(h^{-1})\) elements are involved. The \(\delta(h^2)\) arrays of coefficients and iterates will continue to be stored in single precision.

The computation of the elements of \(F\) to within relative errors of \(\delta(\nu)\), for \(\nu\) proportional to \(h^2\) yields not \(A + B\) but rather \(A + B + K\) where the perturbation \(K\) satisfies
\[(3.13) \quad K = (L_1 + F)F^{-1}K - (L_1F^{-1})K(I + F^{-1}K)^{-1}F^{-1}[L_1^* + F] + \hat{K}, \]
\[\hat{K} = \hat{F} - F.\]

An immediate consequence of Lemma 4 to follow is the bound
\[\|F^{-1}(F + L_1^*)\|_{1,\infty} \leq 2; \text{ hence}\]
\[(3.14) \quad \|K\|_2 < 2\|\hat{K}\|_2 \left[1 + \frac{1 + \frac{1}{2}\|F^{-1}\|_2\|\hat{K}\|_2}{1 - \|F^{-1}\|_2\|\hat{K}\|_2}\right] = \Theta(\nu).\]

Certainly we will not worsen the spectral condition number of the approximate inverse of (3.3) by replacing \((A + B)^{-1}\) with \([A + B + K]^{-1} = (L_1^* + \hat{F})^{-1}F(L_1 + \hat{F})^{-1}\). More importantly, such a replacement restores to us the iterative refinement of Theorem 2, for the factors \((L_1 + \hat{F}), \hat{F},\) and \((L_1^* + \hat{F})\) may now be considered to be exact. It is necessary to see, however, if this choice of perturbed approximate inverse possesses a satisfactory comparability estimate with respect to \(A\). In [6] the comparability estimate
\[(3.15) \quad \frac{(Aw, w)}{([A + B]w, w)} \in [e_1, e_2] , \quad e_1 = \Theta(1) , \quad e_2 = \Theta(h^{-1})\]
was established for the Dirichlet problem (3.1). If an estimate of the same form can be demonstrated for \((A + B) + K\), namely,
\[(3.16) \quad \frac{(Aw, w)}{([A + B + K]w, w)} \in [e_1, e_2] , \quad e_1 = \Theta(1) , \quad e_2 = \Theta(h^{-1})\]
we shall be able to apply Lemma 2, [6] to obtain the A-norm
bound

\[(3.17) \quad \| I - [(A + B) + K]^{-1} A \|_A \leq 1 - c h , \]

\[c: \text{ independent of } h, \quad \bar{\gamma} = o(h) \]

for the error propagator, call it \( G \), of the perturbed iteration. Since \( F \) is diagonal, the perturbed approximate inverse \([A + B + K]^{-1}\) is symmetric, a requirement of Lemma 2, [6]. It remains to prove

**Lemma 3:** For

(i) \( \gamma \) proportional to \( h^2 \) in sufficiently small ratio,

(ii) \( \| K \|_2 \leq o(\gamma) \) for any symmetric perturbation \( K \),

(iii) \( \frac{(Aw,w)}{([A+B]w,w)w} \in [o(1), o(h^{-1})] \),

we obtain the comparability estimate

\[(3.18) \quad \frac{(Aw,w)}{([A + B + K]w,w)} \in [o(1), o(h^{-1})] . \]

**Proof:** Since

\[(3.19) \quad \frac{(Aw,w)}{([A + B]w,w)} = \frac{(Aw,w)}{([A + B + K]w,w)} - (Kw,w) \]

it is clear that

\[(3.20) \quad \odot(1) \{ [(A + B) + K]w,w) - (Kw,w) \} \leq (Aw,w) . \]

Dividing both sides by \([A + B + K]w,w)\) we obtain

\[(3.21) \quad \odot(1) \left[ 1 - \frac{(Kw,w)}{([A + B + K]w,w)} \right] \leq \frac{(Aw,w)}{([A + B + K]w,w)} . \]

Denoting by \( \lambda_1 \) any eigenvalue of \( A + B \) and by \( \lambda \) any eigenvalue of \( (A + B) + K \), it follows from a theorem of Bauer and
Fike (see Wilkinson [15]) that

\[
(3.22) \quad \min_i | \lambda - \lambda_i | < 2 \| \mathbf{x} \|_2 \left[ 1 + \frac{1 + \frac{1}{2} \| F^{-1} \| \| \mathbf{x} \|_2}{1 - \| F^{-1} \| \| \mathbf{x} \|_2} \right].
\]

Thus, for \( \nu \leq k \lambda_{\min}(A+B) \), \( k \) sufficiently small, no eigenvalue of \((A+B)+K\) is smaller than, say \( \frac{1}{2} \lambda_{\min}(A+B) \). Consequently, there is a choice of \( \nu \) proportional to \( h^2 \) for which

\[
(3.23) \quad |(Kw, w)/(([A + B] + K)w, w)| < 1.
\]

The bound on the left is immediate. In the other direction, we may certainly write

\[
(3.24) \quad \frac{(Aw, w)}{[([A + B] + K)w, w] - (Kw, w)} \leq \Theta(h^{-1}),
\]

or

\[
(3.25) \quad (Aw, w) \leq [([A + B] + K)w, w] - (Kw, w)] \Theta(h^{-1}).
\]

Again, dividing both sides of (3.25) by \(([A + B] + K)w, w\), we obtain

\[
(3.26) \quad \frac{(Aw, w)}{([A + B] + K)w, w} < \left[ 1 - \frac{(Kw, w)}{([A + B] + K)w, w} \right] \Theta(h^{-1}).
\]

In view of (3.23), the result follows.

Lemma 3 assures us that \((A + B) + K\) is an acceptable substitute for \((A + B)\) provided that the perturbation \( K \) is well enough behaved (specifically, \( \| K \|_2 \leq \Theta(\nu) \)). Henceforth we shall consider the exact iteration to be

\[
(3.27) \quad u^{(n+1)} = u^{(n)} - \gamma [(L_1 + \hat{F})F^{-1}(\hat{F} + L_1^*)]^{-1}(Au^{(n)} - b).
\]
The optimal choice of iteration parameter $\gamma$ is seen to be proportional to $h$ from Lemma 4, [6]. Before undertaking an asymptotic rounding error analysis, we prove

**Lemma 4:** For $F$ given by (3.10), $(L_1 + F)$ is diagonally dominant by $\mathcal{O}(h)$.

**Proof:** The row sum of $L_1 + F$ corresponding to the mesh point $(jh, kh)$ away from the boundary is given by

\begin{equation}
(3.28) \quad c_{j-1,k} + f_{j,k-1} + v_{j,k}^2
\end{equation}

where the elements $c_{j-1,k}$ and $f_{j,k-1}$ are the appropriate off-diagonal terms of $A$. For such mesh points we shall prove that

\begin{equation}
(3.29) \quad \frac{v_{j,k}^2}{c_{j-1,k} - f_{j,k-1}} \geq 1 + c_1 h \quad , c_1: \text{positive constant},
\end{equation}

from which the assertion of the lemma will follow. For row sums corresponding to mesh points adjacent to the boundary one or both of $c_{j-1,k}$, $f_{j,k-1}$ may be missing. In this case diagonal dominance by at least $\mathcal{O}(h)$ is inherited from $A$.

The quantity $\beta$ referred to in line (3.11) is shown in lemma 2, [6] to satisfy

\begin{equation}
(3.30) \quad \beta \geq 1 + \frac{1}{2}(\rho_{j,k} - 1) + |1 - \rho_{j,k}| + \frac{\phi_{j,k}(\rho_{j,k} + 1)^2}{[2\phi_{j,k}(\rho_{j,k} + 1)^2 + (\rho_{j,k} - 1)^2]^{\frac{1}{2}}}
\end{equation}
where \( \rho_{j,k} = (c_{j,k} + f_{j,k})/(c_{j-1,k} + f_{j,k-1}) \). From (3.10) we obtain, dropping the negligible term \( \phi_{j,k} b_{j,k} \), the inequality

\[
\nu_{j,k}^2 > (c_{j,k} + f_{j,k} + c_{j-1,k} + f_{j,k-1}) - f_{j,k-1} \left[ \frac{c_{j,k-1} + f_{j,k-1}}{\nu_{j,k-1}^2} \right] - c_{j-1,k} \left[ \frac{f_{j-1,k} + c_{j-1,k}}{\nu_{j-1,k}^2} \right],
\]

which, after substitution from line (3.11), becomes

\[
\nu_{j,k}^2 > (c_{j,k} + f_{j,k} + c_{j-1,k} + f_{j,k-1}) + \frac{f_{j,k-1}}{\beta} + \frac{c_{j-1,k}}{\beta}
\]

\[
= c_{j,k} + f_{j,k} - \left[ c_{j-1,k} + f_{j,k-1} \right] \left[ 1 - \frac{1}{\beta} \right].
\]

Dividing both sides of (3.32) by \( -c_{j-1,k} - f_{j,k-1} \) yields

\[
\frac{\nu_{j,k}^2}{c_{j-1,k} - f_{j,k-1}} > \frac{1}{\rho_{j,k}} + \left[ 1 - \frac{1}{\beta} \right].
\]

Since \( \rho_{j,k} \) is near unity for sufficiently small \( h \) we shall argue two cases:

**Case 1:** \( 0 < \rho_{j,k} \leq 1 \).

In this case \( \frac{1}{\rho_{j,k}} \geq 1 \). The result follows immediately from (3.33) since \( \beta = 1 + ch \). Moreover, we may take \( c_1 = c \).
Case 2: \( \rho_{j,k} > 1 \).

From the expression for \( \beta \) given in line (3.30) we obtain

\[
\beta > \rho_{j,k} + \frac{1}{2} \frac{\phi_{j,k} (\rho_{j,k} + 1)^2}{[2\phi_{j,k} (\rho_{j,k} + 1)^2 + (\rho_{j,k} - 1)^2]^{1/2}},
\]

for if \( \rho_{j,k} > 1 \), then \( |1 - \rho_{j,k}| = \rho_{j,k} - 1 \). Dividing both sides of (3.34) by \( \rho_{j,k} \) yields

\[
\frac{\beta}{\rho_{j,k}} > 1 + \frac{\phi_{j,k}}{2\rho_{j,k}} \frac{(\rho_{j,k} + 1)^2}{[2\phi_{j,k} (\rho_{j,k} + 1)^2 + (\rho_{j,k} - 1)^2]^{1/2}}
\]

\[
> 1 + \frac{\phi_{j,k}^{1/2}}{2\rho_{j,k}} \frac{(\rho_{j,k} + 1)^2}{[2(\rho_{j,k} + 1)^2 + (\rho_{j,k} - 1)^2]^{1/2} \phi_{j,k}^{1/2}}
\]

\[
> 1 + c_2 h
\]

for \( \phi_{j,k} = k_1 h^2 \). Thus, we can write

\[
\frac{v_{j,k}^2}{c_{j-1,k} - f_{j,k-1}} > \frac{1}{\rho_{j,k}} + \left[ 1 - \frac{1}{\beta} \right]
\]

\[
= \frac{1}{\beta} \left[ \frac{\beta}{\rho_{j,k}} \right] + \left[ 1 - \frac{1}{\beta} \right] \geq 1 + \left[ c_2 \right] h,
\]

which concludes the argument.

By retracing our steps through the argument above it can be shown that \( \frac{v_{j,k}^2}{c_{j-1,k}} > 1 + \mathcal{O}(h) \) in the case that \( f_{j,k-1} \) is missing, that is, for \((jh,kh)\) adjacent to the lower boundary. A similar argument applies to the left-hand
boundary. Since \( L_1 \) is lower triangular, we need not concern ourselves with the right-hand or upper boundaries.

In view of line (3.11) and the result above an application of Lemma 0 yields

\[
(3.37) \quad \| (L_1 + F)^{-1} \|_2, \| (L_1^* + F)^{-1} \|_2 \leq \Theta(h^{-1}).
\]

Also, provided that the \( ft_2 \)-factorization is carried out as described preceding line (3.13) above, the elements \( \hat{v}_{j,k}^2 \) of \( \hat{F} \) satisfy

\[
(3.38) \quad \hat{v}_{j,k}^2 = v_{j,k}^2 (1 + \Theta(v)).
\]

For \( h \) sufficiently small and \( v \) proportional to \( h^2 \) in sufficiently small but fixed ratio it is clear that \( L_1 + \hat{F} \) and \( L_1^* + \hat{F} \) remain diagonally dominant by \( \Theta(h) \); hence

\[
(3.39) \quad \| (L_1^* + \hat{F})^{-1} \|_2, \| (L_1 + \hat{F})^{-1} \|_2 \leq \Theta(h^{-1}).
\]

**Asymptotic Rounding Error Analysis**

Having decided to accept \( (A+B+K)^{-1} = (L_1^* + \hat{F})^{-1} \hat{F}(L_1 + \hat{F})^{-1} \) as exact approximate inverse, (3.27) is realized in single precision floating point arithmetic as

\[
(3.40) \quad \hat{u}^{(n)} = (I + \delta_n)[\hat{u}^{(n-1)} - \nabla(I + \delta_n)(I + R_n, 1)(L_1^* + \hat{F})^{-1}

\cdot(I + R_n, 2)\hat{F}(I + R_n, 3)(L_1 + \hat{F})^{-1}(I + \Gamma_n)

\cdot[(A + \delta A_n)\hat{u}^{(n-1)} - (I + \varepsilon_n b)]).
\]

The meanings of the relative error perturbations \( \delta_n, \varepsilon_n, \Gamma_n \),
\( \delta A_n \) and \( \varepsilon_n \) are apparent from their context. The matrices \( R_{n,1} \), \( R_{n,2} \), and \( R_{n,3} \) represent the relative errors introduced in the solution of the linear system implied by 
\( (L_1^* + \hat{F})^{-1} \), the multiplication by \( \hat{F} \), and the solution of the linear system implied by \( (L_1 + \hat{F})^{-1} \), respectively. Since the matrices \( L_1 + \hat{F} \) and \( L_1^* + \hat{F} \) are sparse, diagonally dominant and triangular, solutions of linear systems for which they are coefficient matrices may be easily obtained by Gaussian elimination without pivoting. If all inner products that arise in the elimination are computed in \( f_{t_2} \)-arithmetic, we shall obtain, instead of the solutions of the systems implied by \( (L_1 + \hat{F})^{-1} \), \( (L_1^* + \hat{F})^{-1} \), the solutions of the perturbed systems

\[
\begin{align*}
(L_1 + \hat{F}) + K_2 & \quad \hat{x} = y \\
(L_1^* + \hat{F}) + K_1 & \quad \hat{z} = w 
\end{align*}
\]

for given right-hand sides \( y \) and \( w \), where \( K_1 \) and \( K_2 \) are sparse matrices (no more than three entries per row) with 2-norm bounds of the form \( \Theta(\sqrt{n}) \). Consequently, from the identities

\[
\begin{align*}
(I + R_{n,1})(L_1^* + \hat{F})^{-1} &= [(L_1^* + \hat{F}) + K_1]^{-1}(L_1^* + \hat{F}) - I \\
(I + R_{n,3})(L_1 + \hat{F})^{-1} &= [(L_1 + \hat{F}) + K_2]^{-1}(L_1 + \hat{F}) - I 
\end{align*}
\]

we deduce the 2-norm bounds

\[
\|R_{n,1}\|_2, \|R_{n,3}\|_2 \leq \Theta(\sqrt{n}) 
\]
for $\nu$ proportional to $h^2$. Actually, bounds of this form can be obtained for a sufficiently small ratio of $\nu$ to $h$, but we may take no advantage of this because it has already been necessary to take $\nu$ proportional to $h^2$ to assure stability in the factorization. Since $\hat{F}$ is a diagonal matrix, $R_{n,2}$ may be taken to be a diagonal perturbation whose elements are bounded by $\Theta(\nu)$.

Because we are considering $(L_1 + \hat{F})$ and $(L_1^* + \hat{F})$ to be exact, the purification iteration of Theorem 2 is applicable. Hence $R_{n,1}$ and $R_{n,3}$ may be redefined so that they measure the relative error introduced in both the Gaussian elimination step and the purification step and satisfy, for $\nu$ proportional to $h^2$, the uniform bounds

$$
\|R_{n,1}\|_2, \|R_{n,3}\|_2 \leq \Theta(\nu).
$$

Subtracting (3.27) from (3.40) yields

$$
r(n) = \hat{u}(n) - u(n) = [G + \Delta_n]r(n-1) + e(n)
$$

where $G$ is the error propagator, $I - \nabla (A+B+K)^{-1}A$. The perturbations $\Delta_n$ and $e(n)$ are given by

$$
\Delta_n = \theta_n G - \nabla (I + \theta_n)[(L_1^* + \hat{F})^{-1}F(L_1 + \hat{F})\delta A_n + \Lambda_n (A + \delta A_n)]
$$

where

$$
\Lambda_n = (L_1^* + \hat{F})^{-1}F(L_1 + \hat{F})^{-1} - (I + \sigma_n)(I + R_{n,1})(L_1^* + \hat{F})^{-1}
$$

$$
\cdot (I + R_{n,2})\hat{F}(I + R_{n,3})(L_1 + \hat{F})^{-1}(I + \Gamma_n),
$$
and
\[ e^{(n)} = \theta_n u^{(n)} - \Phi(I + \theta_n)\{A_n(Au^{(n)} - b) + [L_1^* + \hat{F}]^{-1}\hat{F}(L_1 + \hat{F})^{-1} + \Lambda_n\} (\delta A_n u^{(n)} - \xi_n b)\} \]

Since the efficiency of the error propagator is measured in the A-norm, we shall consider an A-norm bound for the global rounding error. Premultiplying (3.45) by \( A^{\frac{1}{2}} \), we obtain
\[ A^{\frac{1}{2}} r^{(n)} = [A^{\frac{1}{2}}(G + \Delta_n)A^{-\frac{1}{2}}]A^{\frac{1}{2}} r^{(n-1)} + A^{\frac{1}{2}} e^{(n)} .\]

This yields the difference inequality
\[ \|r^{(n)}\|_A \leq \|G + \Delta_n\|_A \|r^{(n-1)}\|_A + \|e^{(n)}\|_A \]
from which follows the bound
\[ \|r^{(n)}\|_n \leq \frac{1 - [\|G\|_A + \delta]^{(n)}}{1 - [\|G\|_A + \delta]} \max_{1 \leq j \leq n} \|e^{(j)}\|_A \]
where \( \delta = \sup_n \|\Delta_n\|_A \).

Similarity transformation of \( \Delta_n \) by \( A^{\frac{1}{2}} \) yields
\[ A^{\frac{1}{2}} \Delta_n A^{-\frac{1}{2}} = A^{\frac{1}{2}} \rho_n GA^{-\frac{1}{2}} - \Phi A^{\frac{1}{2}}(I + \rho_n)\{(L_1^* + \hat{F})^{-1}\hat{F}(L_1 + \hat{F})^{-1} \cdot \delta A_n A^{-\frac{1}{2}} + \Lambda_n(A + \delta A_n)A^{-\frac{1}{2}}\} \]

Since (i) the perturbations \( \rho_n, \xi_n, \Gamma_n, R_n, i \) share uniform 2-norm bounds of the form \( \delta(\nu) \) (see lines (xii) of the Introduction and (3.44) above), (ii) \( \xi_n \) and \( \delta A_n \) share uniform 2-norm bounds of the form \( \delta(\nu^2) \) for the \( \ell_2 \)-accumulation of the residual (see line (xiv) of the Introduction),
(iii) the estimate given in (3.17) holds for $\gamma = e_2 = \phi(h)$ (see Lemma 4, [5]), and (iv) the smallest eigenvalue of $A^{1/2}$ behaves like $\phi(h)$ for (3.1), the 2-norms of the terms on the right-hand side of (3.52) behave no worse than

\[
(a) \quad \| A^{1/2} \phi_n GA^{-1/2} \|_2 \leq \phi[\frac{\gamma}{h}] ,
\]

(3.53) \( (b) \quad \| \nabla A^{1/2}(I + \phi_n)^{-1}(L_1 + F)^{-1} \delta A_n A^{1/2} \|_2 \leq \phi[\frac{\gamma}{h}]^2 , \)

\( (c) \quad \| \nabla A^{1/2}(I + \phi_n)(A + \delta A_n) A^{1/2} \|_2 \leq \phi[\frac{\gamma}{h}] + \phi[\frac{\gamma}{h}]^2 \).

Hence,

(3.54) \[ \| \Delta_n \|_A \leq \phi[\frac{\gamma}{h}] \]

for $\gamma$ proportional to $h^2$ in sufficiently small fixed ratio. Consequently, the uniform bound $\delta$ may be taken proportional to $\frac{\gamma}{h}$.

Turning to the local errors $e^{(j)}$, line (3.48) yields the uniform $A$-norm bound

(3.55) \[ \| e^{(j)} \|_A = \phi(\gamma) . \]

This bound follows in part from the observation that for any sufficiently smooth function $\gamma$, $h^{-2} A \gamma = - \phi \gamma + \phi(h^2)$. We are, of course, assuming that the iteration is initialized with a vector that is both exactly representable to machine single precision and smooth. Returning to line (3.51), it is clear that

(3.56) \[ \| r^{(n)} \|_A = \phi[\frac{\gamma}{h}] \]
uniformly for $\nu$ proportional to $h^2$ in sufficiently small fixed ratio. In view of the comparability between the $L_2$ and $A$-norms, we may replace the $A$-norm above at a cost of squaring the $h$ in the denominator to obtain

**Theorem 12:** For the perturbed approximate inverse

$$(L_1^* + \hat{F})^{-1}\hat{F}(L_1 + \hat{F})^{-1},$$

the global rounding error inherent in the single precision floating point realization of the factorization procedure of Dupont, Kendall, and Rachford behaves no worse than

$$(3.57) \quad \|r^{(n)}\|_{L^2} = \mathcal{O}\left[\frac{\nu}{h^2}\right]$$

uniformly, provided that

(i) $\nu$ is chosen proportional to $h^2$ in sufficiently small but fixed ratio

(ii) the residuals $A\mathbf{u}^{(n)} - \mathbf{b}$ are computed in $f\ell_2$-arithmetic

(iii) sufficient care is taken in the computation of the elements of $F$ to insure that $\hat{F}$, the machine analogue of $F$, satisfies

$$(3.58) \quad \|F - \hat{F}\|_2 \leq \mathcal{O}(\nu).$$

Finally, we remark that we cannot expect the higher dimensional analogue of this factorization technique to be as well-behaved as the higher dimensional alternating direction iteration insofar as the growth of rounding errors
are concerned. This follows from the generalization of (3.10) to n-dimensions:

\[(3.59) \quad f_\alpha = (1+\phi_\alpha) b_\alpha - \sum_{j=1}^{n} c_{\alpha-e_j,e_j} \left[ f^{-1}_{\alpha-e_j} \sum_{i=1}^{n} c_{\alpha-e_j,e_i} \right] \]

where \(\alpha\) is a multi-index for \(\mathbb{R}^n\) and \(e_i\) the unit vector in the \(i^{th}\) direction. The employment of \(f^2\)-arithmetic in the accumulation of (3.59) will insure only that no relative error in any \(f_\alpha\) exceeds \(\Theta(\sqrt{n^2})\) in magnitude. Consequently, to guarantee that the computed approximate inverse of the factorization iteration possesses the same comparability to \(A\) we shall be forced to take \(\gamma\) proportional to \(h^{n/2}\).


