FLORIDA STATE UNIVERSITY

ITERATIVE METHODS FOR THE SOLUTION
OF LINEAR EQUATIONS

By

STANLEY R. BENDER

A Paper
Submitted to the Graduate Council of Florida State University in partial fulfillment of the requirements for the degree of Master of Science.

Approved: Paul J. McCarthy
Professor Directing Paper

none
Minor Professor

Nicholas Reissner
Representative of Graduate Council

W. Carothers
Dean of the Graduate School

May, 1958
ACKNOWLEDGEMENTS

I wish to thank Dr. Paul J. McCarthy for his patient assistance in the preparation of this paper.
# TABLE OF CONTENTS

## ACKNOWLEDGEMENTS

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ii</td>
</tr>
</tbody>
</table>

## Chapter

- **I. INTRODUCTION**
  
<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

- **II. ITERATIVE METHODS**
  
  1. Analytic basis 
  2. Method of steepest descent 
  3. Gauss-Seidel method 
  4. Method of relaxation 
  5. Summary 
  6. Examples 

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>13</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>16</td>
</tr>
</tbody>
</table>

## BIBLIOGRAPHY

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

The numerical solutions of many types of problems are generally obtained by solving approximating linear algebraic systems. Moreover, in solving a nonlinear problem, one may replace it by a sequence of linear systems providing progressively improved approximations.1 For the study of these linear systems of equations a geometric termin-
ology with the compact symbolism of vectors and matrices is useful. A resume of the basic principles of higher algebra necessary for the development of the material to follow is therefore included.

A vector $\mathbf{x}$ of order $n$ is defined to be a set of $n$ real numbers $x_1, x_2, \ldots, x_n$ arranged in a definite sequence and is written

$$\mathbf{x} = (x_1, x_2, \ldots, x_n).$$

The vector $\mathbf{y}$ is written

$$\mathbf{y} = (y_1, y_2, \ldots, y_n).$$

Equality of vectors $\mathbf{x} = \mathbf{y}$ is defined to mean simultaneous equality of similarly numbered real number elements, $x_i = y_i$ for $i = 1, 2, \ldots, n$.

The multiplication of $\mathbf{x}$ by any real number $c$ is defined as the vector

$$(1) \quad c\mathbf{x} = (cx_1, cx_2, \ldots, cx_n) = (x_1c, x_2c, \ldots, x_nc) = \mathbf{xc}.$$

Addition of two vectors \( \mathbf{x} + \mathbf{y} \) is defined by

\[
(2) \quad \mathbf{x} + \mathbf{y} = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n).
\]

The totality of all vectors for a fixed integer \( n \) with the operations (1) and (2) is called an \( n \)-dimensional vector space over the field of real numbers \( \mathbb{R} \) and is written \( \mathbb{V}_n(\mathbb{R}) \). A subset of the vectors in the vector space \( \mathbb{V}_n(\mathbb{R}) \) which is closed with respect to the two basic operations of vector algebra (1) and (2) is called a vector subspace of \( \mathbb{V}_n(\mathbb{R}) \).

From (2) it follows that there is a unique null vector

\[
\mathbf{0} = (0, 0, \ldots, 0)
\]

with the property that for any vector \( \mathbf{x} \),

\[
\mathbf{x} + \mathbf{0} = \mathbf{x} \quad \text{and} \quad \mathbf{c} \cdot \mathbf{0} = \mathbf{0}
\]

for all \( \mathbf{x} \) in \( \mathbb{V}_n(\mathbb{R}) \) and all \( \mathbf{c} \) in \( \mathbb{R} \). For any vector \( \mathbf{x} \) in \( \mathbb{V}_n(\mathbb{R}) \) there is a unique vector \( \mathbf{x}' \) such that \( \mathbf{x} + \mathbf{x}' = \mathbf{0} \), namely the vector

\[
\mathbf{x}' = (-x_1, -x_2, \ldots, -x_n).
\]

This vector \( \mathbf{x}' \) is called the negative of \( \mathbf{x} \) and is usually denoted by \( -\mathbf{x} \). Since \( \mathbf{x}' = (-1)\mathbf{x} \), the definition of \( -\mathbf{x} \) may be written \( -\mathbf{x} = (-1)\mathbf{x} \).

Subtraction of vectors is defined by the equation \( \mathbf{x} - \mathbf{y} = \mathbf{x} + (-1)\mathbf{y} \).

Vectors also have the following properties, valid for all vectors \( \mathbf{x}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \) in \( \mathbb{V}(\mathbb{R}) \) and all scalars \( c, c_1, c_2, \ldots, c_n \) in \( \mathbb{R} \):

\[
(x_1 + x_2) + x_3 = x_1 + (x_2 + x_3),
\]

\[
x_1 + x_2 = x_2 + x_1,
\]

\[
c(x_1 + x_2 + \cdots + x_n) = c\mathbf{x}_1 + c\mathbf{x}_2 + \cdots + c\mathbf{x}_n,
\]

\[
(c_1 + c_2 + \cdots + c_n)\mathbf{x} = c_1\mathbf{x} + c_2\mathbf{x} + \cdots + c_n\mathbf{x} \quad \text{and}
\]

\[
(c_1 c_2)\mathbf{x} = c_1 (c_2 \mathbf{x}).
\]
If any \( n \) vectors \( \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \) lie in a vector space \( V_n(\mathbb{R}) \), they are called linearly dependent if there are scalars \( c_1, c_2, \ldots, c_n \), not all zero, such that
\[
c_1 \mathbf{x}_1 + c_2 \mathbf{x}_2 + \cdots + c_n \mathbf{x}_n = \mathbf{0}.
\]
If no such scalars exist, the set \( \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \) is called linearly independent. A vector space is \( m \)-dimensional if \( m \) is the maximum number of linearly independent vectors in the space. Any \( m \) linearly independent vectors in the space then constitute a coordinate system for the space and any vector in the space can be written as a linear combination of these linearly independent vectors.

We shall refer to the following set of \( n \)-dimensional vectors of \( V_n(\mathbb{R}) \) as the \( \mathbf{e} \) coordinate system for \( V_n(\mathbb{R}) \):
\[
\begin{align*}
\mathbf{e}_1 &= (1, 0, 0, \ldots, 0) \\
\mathbf{e}_2 &= (0, 1, 0, \ldots, 0) \\
& \vdots \\
\mathbf{e}_n &= (0, 0, \ldots, 0, 1).
\end{align*}
\]
With regard to the vector \( \mathbf{x} \) the scalars \( x_i \) for \( i = 1, 2, \ldots, n \) are called the coordinates of \( \mathbf{x} \) in the \( \mathbf{e} \) system and the vectors \( x_1 \mathbf{e}_1 \) are its components.

When the coordinates of \( \mathbf{x} \) in the \( \mathbf{e} \) system are arranged in column form:
\[
\begin{pmatrix}
x_{11} \\
x_{21} \\
\vdots \\
x_{n1}
\end{pmatrix}
\]
the arrangement is known as a column matrix which we shall designate by \( \mathbf{x}_e \). The arrangement \( (x_{11}, x_{12}, \ldots, x_{1n}) \) is called a row matrix. Any rectangular array of numbers is called a rectan-
throughout the discussion upper case Latin letters will be used to designate matrices. The double subscripts indicate the two-way ordering of matrices. The first subscript is the row index and the second subscript is the column index.

Given any two matrices \( A \) and \( B = \{b_{ij}\}_{pq} \) where \( i = 1, 2, \ldots, p \) and \( j = 1, 2, \ldots, q \); the product \( AB \) is defined as the matrix

\[
C = \{c_{ik}\}_{pq} \text{ where } c_{ik} = \sum_{j=1}^{n} a_{ij} b_{jk} \text{ if } n = p.
\]

The product \( BA = D = \{d_{ik}\}_{pq} \) where \( d_{ik} = \sum_{j=1}^{n} b_{ij} a_{jk} \) if \( m = q \). In general \( AB \neq BA \). The identity matrix under multiplication, \( I = \{k_{ij}\}_{pq} \), is such that its elements \( k_{ij} \) is equal to one if \( i = j \) and is equal to zero if \( i \neq j \) for all values of \( i \) and \( j \). The matrix \( I \) is therefore a diagonal matrix; that is, it is square (has the same number of columns and rows) and all elements off the main diagonal are zero.

The transpose of the matrix \( A \), designated by attaching a superscript \( t \) to the matrix notation \( A^t = \{a_{ij}\}_{mn} \) is such that for all \( a_{ij} \) of \( A \) and all \( a_{ji}^t \) of \( A^t \), \( a_{ij} = a_{ji}^t \) for all values of \( i \) and \( j \). The transpose of the column matrix \( x_e \) is the row matrix \( x_e^t = (x_{11}, x_{12}, \ldots, x_{1n}) \). The transpose of a row matrix is a column matrix, \((x_e^t)^t = x_e\).
\[ AB = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} b_{jk} = \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ik} = C \] where

\[ (AB)^t = C^t = \sum_{k=1}^{n} \sum_{l=1}^{n} c_{kl} \]

Let \( \vec{f}_1, \vec{f}_2, \ldots, \vec{f}_n \) be another set of linearly independent vectors in this \( n \)-dimensional vector space; then each \( f_j \) is expressible as a linear combination of the \( e_i \)'s and vice versa. If we consider \( e \) and \( f \) to be row matrices,

\[ e = (e_1, e_2, \ldots, e_n) \]
\[ f = (f_1, f_2, \ldots, f_n) \]

in which the coordinates of the \( e_i \)'s and \( f_j \)'s, with regard to the \( \vec{e} \) and \( \vec{f} \) coordinate systems respectively, are arranged column-wise.

Then we can write in abbreviated form,

\[ e = fe^F, \quad f = eF, \quad \text{and} \quad et_e = I. \]

Then \( \vec{f}^t \vec{f} = (eF)^t eF = FteteF = FtIF = FtF = H = H^t = ff^t \). Since \( H = H^t \), \( H \) is called a symmetric matrix. Relating the \( e \) and \( f \) coordinate systems we have

\[ I = e^te = (eF)^t eF = EtF^t eF = EtFE. \]

If both \( e \) and \( f \) are unit (the only nonzero element of the vector is equal to unity) orthogonal coordinate systems then \( H = I \) and

\[ I = EtFE = EtE. \]

The geometric vector \( \vec{x} \) can be expressed as the numerical vector \( x_f \), that is, as an \( n \)-tuple of coordinates \( x_j' \) in the \( f_j \) coordinate system, and we can write \( x_f = Fx_e \).

We define the scalar or dot product of two vectors \( \vec{x} \) and \( \vec{y} \) in the \( f \) coordinate system by

\[ \vec{x} \cdot \vec{y} = x_f^t H_y f, \]
and it is a pure number. If \( \mathbf{x} \) and \( \mathbf{y} \) are the same vector then the scalar product is the square of the length,

\[
\mathbf{x} \cdot \mathbf{x} = |\mathbf{x}|^2 = \mathbf{x}^\top \mathbf{x}.
\]

The matrix \( \mathbf{H} \) is said to be positive definite since \( \mathbf{x}^\top \mathbf{H} \mathbf{x} \geq 0 \) and equality can hold only when \( \mathbf{x} = 0 \). Two vectors \( \mathbf{x} \) and \( \mathbf{y} \) are said to be orthogonal to each other if their scalar product is zero.

The scalar product may be thought of geometrically as the product of the length of one vector by the projection of the other upon it.

In the \( \mathbf{e} \) coordinate system,

\[
\mathbf{x} \cdot \mathbf{y} = x_1 e_1 y_1 e_1 = x_1 y_1 e_1 e_1 = x_1 y_1 = y_e x_e = \mathbf{y} \cdot \mathbf{x}.
\]

Similarly we can prove that the scalar product is commutative in the \( \mathbf{f} \) coordinate system.

The determinant of a square matrix \( \mathbf{A} \), \( m = n \), is defined as

\[
\det(\mathbf{A}) = \sum_{\text{all permutations}} (-1)^t a_{i_1 i_2} a_{i_2 i_3} \cdots a_{i_n i_1},
\]

where \( t \) is the number of transpositions required to restore \( i_1, i_2, \ldots, i_n \) to their natural ordering \( 1, 2, \ldots, n \). The matrix \( \mathbf{A} \) is said to be nonsingular if the value of its determinant is nonzero.

If the row and column of a particular element \( a_{ij} \) of a square matrix \( \mathbf{A} \) are deleted, what remains is called the \( (n-1) \times (n-1) \) submatrix \( \mathbf{M}_{ij} \). The scalar \( d_{ij} = (-1)^{i+j} |\mathbf{M}_{ij}| \) is called the cofactor of \( a_{ij} \) in \( \mathbf{A} \). The \( n \times n \) matrix \( \mathbf{D} = \{d_{ij}\} \) is called the adjoint of \( \mathbf{A} \) and is denoted by \( \mathbf{A}^\top \). The inverse of the matrix \( \mathbf{A} \) (for \( m = n \)) may now be defined as

\[
\mathbf{A}^{-1} = \frac{\mathbf{A}^\top}{|\mathbf{A}|}.
\]
CHAPTER II

ITERATIVE METHODS

1. Analytic basis. In solving any equation or system of equations there are two possible approaches. We may be able to find a solution by means of a direct method which prescribes only a finite sequence of operations whose completion yields an exact solution. When the number of unknowns is great, the algebraic methods become exceedingly laborious. The solution of a system of this type is then most easily found by the use of an iterative method. Iterative methods are especially valuable when machines for handling the successive iterations effectively are available, when a good approximation to the solution is known or when we wish to obtain solutions of adjusted equations.

In solving a set of $m$ equations in $m$ unknowns $y - Ax = 0$ an iterative method is defined as a rule for operating on an approximate solution $x_p$ in order to obtain an improved solution $x_{p+1}$; such that the sequence $\{x_p\}$ so defined has the solution $x$ as its limit.

Since the approximation $x_0$ with which one may begin an iterative method does not need to be "close" to the solution $x$. We may start with an arbitrary $x_0$, however a good approximation to the solution will appreciably diminish the number of iterations necessary to achieve a desired accuracy. If no approximation with which to start the iterative method is available take $x_0 = 0$ unless the main diagonal of the coefficient
matrix features relatively large terms; then take \( x_0 = \begin{bmatrix} y_1/a_{11} \\ y_2/a_{22} \\ \vdots \\ y_m/a_{mm} \end{bmatrix} \).

A large class of iterative methods is based upon the following geometric idea: take any vector \( b_0 \) and a sequence of vectors \( u_p \) and define the sequence \( \{b_p\} \) by \( b_0, b_p = b_{p-1} - \phi_p u_p \) where the scalar \( \phi_p \) is chosen so that \( b_p \) is orthogonal to \( u_p \). If the vectors \( u_p \) "fill out" some \( m \)-dimensional subspace of the vector space \( V_n(\mathbb{R}) \) then the vectors \( b_p \) approach as a limit a vector \( b \) which is orthogonal to this \( m \)-space. A possible choice for the vectors \( u_p \) is any set of reference vectors \( f_i \) of the \( m \)-space taken in order and then repeated \( u_{vm+i} = f_i \), where \( v = 1, 2, \ldots \). Let \( f = eF \) and \( H = FF^t \); then we have

\[
 b_p = b_{p-1} - \phi_p u_p ,
\]

and since \( b_p \) is orthogonal to \( u_p \)

\[
 0 = u_p \cdot b_p = u_p \cdot (b_{p-1} - \phi_p u_p) = u_p \cdot (u_p \cdot H b_p - \phi_p u_p) = u_p \cdot (u_p \cdot H b_p - 1) = \phi' u_p \cdot H u_p,
\]

where \( \phi' = u_p \cdot H b_p - 1/u_p \cdot H u_p \).

If we let \( y = Ax \) represent the set of \( m \) equations in \( m \) unknowns to be solved and \( x_p \) represent any approximation to the solution \( x \), then we have

\[
 r_p = y - Ax_p = Ax - Ax_p = A(x - x_p) = A x_p - A x_p = As_p .
\]

The coordinates of the vector \( r_p \) are called the residuals of the equations. The symbol \( r_p \) and equivalently \( As_p \) are representations
of the deviation of the approximation from the true solution. Hence either $r_p$ or $s_p$ can be taken as $b_p$.

Take the matrix $A$ to be positive definite. Any system of equations $w = Bx$, $B$ nonsingular, can be transformed by a premultiplication of $B^t$ on both sides into a system in which the coefficient matrix is positive definite. The solution of

$$y = B^t w = B^t B x = A x$$

as also the solution of $w = B x$.

Let $r_p = b_p$, then

$$y - A x_p = b_p - 1 - \phi_p u_p = y - A x_p - 1 - \phi_p u_p,$$

and

$$A x_p = A x_p - 1 + \phi_p u_p,$$

and

$$x_p = x_p - 1 + \phi_p u_p.$$

Since $u_p$ is an arbitrary set of vectors we may substitute $u_p = A v_p$ to obtain

$$x_p = x_p - 1 + \phi_p v_p.$$

Now $\phi_p = u_p^t H b_p - 1/ u_p^t H u_p$

$$= v_p^t A^t H b_p - 1/ v_p^t A^t H A v_p.$$

Take $H = A^{-1}$, this gives us

$$\phi_p = v_p^t r_p - 1/ v_p^t A v_p.$$

If on the other hand we take $s_p = b_p$ then

$$y - A x_p = A b_p = A (b_p - 1 - \phi_p u_p) = y - A x_p - 1 - \phi_p A u_p,$$

and

$$A x_p = A x_p - 1 + \phi_p A u_p,$$

and

$$x_p = x_p - 1 + \phi_p u_p.$$

Now $\phi_p = u_p^t H b_p - 1/ u_p^t A v_p.$
Take $H = A$ then $f_p = u_p \cdot r_p - 1/u_p^T A u_p$.

When the vector $f_p u_p$ is subtracted from $x_p$ the new residual $r_p$ may be thought of as a leg of a right triangle of which $r_p - 1$ is the hypotenuse (provided $u_p$ and $r_p - 1$ were not orthogonal).

Any rule for selecting the vectors $u_p$ (or equivalently the $v_p$) at each step defines a particular iterative process; of these the following three are in common use.

2. Method of steepest descent. This method prescribes that we take $u_p = r_p - 1$. Since $A$ in the equation $y - Ax = 0$ will be taken to be positive definite, this implies that the matrix $A$ can be expressed as the product of a matrix $C$ multiplied by its transpose $C^T$, $A = C^T C$, where the matrix $C$ is nonsingular.

If we let $C x = z$; then $y = Ax = C^T C x = C^T z$. Defining the function $g(x)$ we have,

$$g(x) = (C x - z)^T (C x - z)$$
$$= x^T C^T C x - x^T C z - z^T C x + z^T z$$
$$= x^T C^T C x - x^T C z - x^T C z + x^T C^T C x$$
$$= 2(x^T A x - x^T y).$$

This function being a sum of squares is never negative and has a minimum value of zero for $x = A^{-1} y = C^{-1} z$. The function $g(x)$ is a function of $m$ variables $f_1, f_2, \ldots, f_m$. By taking the partial derivatives of $g(x)$ with respect to these variables we find the coordinates of the gradient vector. Then

$$g(x) = \sum_{i=1}^{m} \frac{\partial}{\partial f_i} f_i \frac{\partial}{\partial f_i} g_i \frac{\partial}{\partial f_i} g_i \frac{\partial}{\partial f_i} - 2 \sum_{i=1}^{m} \frac{\partial}{\partial f_i} g_i f_i + z^T z,$$

and since $y$ and $z$ are known constants
\[ g_x(x) = \sum_{i,j=1}^{m,n} a_{ij} \frac{\partial f_i}{\partial j} + \sum_{i,j=1}^{m,n} b_{ij} \frac{\partial g_i}{\partial j} - 2y. \]

Now \[ \phi_{11} = \phi_{jj}; \] therefore
\[ g_x(x) = 2 \sum_{i,j=1}^{m,n} a_{ij} \frac{\partial f_i}{\partial j} - y = 2(Ax - y). \]

The function \( g(x) \) evaluated at the point \( x = x_p - 1 \) is undergoing its most rapid variation in the direction \( r_p - 1 \) since,
\[ g_x(x_p - 1) = 2(Ax_p - 1 - y) = 2(-r_p - 1) = -2r_p - 1. \]

If we think of our problem as one of minimizing \( g(x) \) by successive steps, it is natural to take each step in the direction of most rapid decrease. Hence we have justified the selection of \( r_p - 1 \) as the vector \( u_p \) for most rapid minimization of the residuals. \(^2\)

The method of steepest descent then proceeds as follows:
\[ y - Ax_0 = r_0 \]
\[ x_1 = x_0 + \phi_1 r_0 \text{ where } \phi_1 = r_0^t r_0 / r_0^t A r_0 \]
\[ y - Ax_1 = r_1 \]
\[ \vdots \]
\[ x_p = x_p - 1 + \phi_p r_p - 1 \text{ where } \phi_p = r_p^t r_p - 1 / r_p^t A r_p - 1 \]
\[ y - Ax_p = r_p \]
\[ \vdots \]

Since this is a self correcting method it is desirable to sacrifice temporary accuracy for ultimate speed; an error of calculation made in any step continually corrects itself in subsequent iterations.

\(^2\)Ibid., p. 47.
3. **Gauss-Seidel method.** This method is the classical iterative method for solving linear equations that feature symmetry and relatively large diagonal terms, however it may also be used for certain other systems of equations.

Using this method we let \( u_p = u_{vm+1} + e_1 \) where \( v = 1, 2, \ldots \).

Then \( x_p \) differs from \( x_{p-1} \) only in the \( i \)th element. In selecting \( u_p \) we do so in such a manner that the \( i \)th element of \( r_p \) (the \( i \)th residual) is zero. This means that the \( i \)th equation is satisfied exactly. This process will require more steps than either of the other two methods; however the simplicity of choice and the ease of computation give it a great advantage in automatic machinery work.

In using this method for hand computation it is usually best to set up a \(-A^t\) table and a residual table. The residuals are then successively reduced to zero in the following manner: divide the residual \( r_1 \) to be eliminated by the corresponding main diagonal element of \( A \), \( a_{11} \), and use the quotient \( \phi_p e_1 (\phi_p = e_1^t r_p - 1/e_1^t A e_1) \) as the change in \( x_1 \). Change each of the residuals \( r_1, r_2, \ldots, r_m \) by \( \phi_p e_1 a_{1j} \) where \( j = 1, 2, \ldots, m \); that is, add \( \phi_p e_1 (-a_{1j}) \) to each of the residuals. Each step of the calculation thereafter involves a single application of an operation of the "basic kind". The residual table merely records in tabular form the operation used, the extent to which it is used and the consequent effects on the residuals. One operation of this type corresponds to each unknown in the problem and is called

---

The residuals, one of which stands identically for the non-zero side of each equation in the system are such that for any values assigned to the original unknowns $x_1, x_2, \ldots, x_m$ corresponding values can be calculated for $r_1, r_2, \ldots, r_m$ directly from the defining relations. The method would then be self-checking. To save time, however, use is made of the $-A^T$ table and each of the residuals $r_i$ is changed by $p_{pe1}(-a_{ji})$ as previously indicated. A check on the residuals at periodic intervals and at the end of the computation is therefore necessary.

Under certain circumstances changing the coefficient of $x_i$ in the $i$th equation to one is a useful preliminary step. The techniques for handling the computation in the Gauss-Seidel method also apply to the handling of the computation in the following method.

4. Method of relaxation. This method is an adaptation of the Gauss-Seidel method, and therefore always takes $u_p$ to be some $e_i$, but the selection is made as the computation progresses. Since the choice of $u_p = e_i$ eliminates the $i$th element of $r_p$ one can choose to eliminate the largest residual (or make this residual assume any desirable value in view of future iterations). Selecting the largest residual is not necessarily the best choice. The magnitude of the correction is measured by the correction vector $p_{pu_p}$ and this magnitude is

$$u_p^T r_p - 1(u_p^T A u_p)^{-1/2}$$

Now when $u = e$ this becomes $e_1^T r_p - 1(a_{ii})^{-1/2}$. Hence one should examine the quotients of the residual components divided by
the corresponding $\sqrt{a_{11}}$ and eliminate the largest quotient (in absolute value).

Various refinements have been made on the method of relaxation to make it more practical for hand computation. These refinements include: overrelaxation, reduction of the residuals beyond the zero value; underrelaxation; block relaxation, changing several variables by the same amount; and group relaxation, changing several variables by varying amounts. Group relaxation is sometimes used to produce a modification of the $A^t$ table that has a main diagonal of relatively large elements. This table is constructed with the idea that each change in the unknowns listed will produce a relatively large change in only one of the residuals and a different residual for each entry. Relaxation with this table should show much quicker convergence to a solution than with the $A^t$ table.

5. Summary. Using the method of relaxation causes the $x_i$'s to converge to a solution more rapidly than the method of Seidel. This feature together with the advantages incurred with the use of under and overrelaxation make this method the best of the three discussed for systems with a "reasonable" number of unknowns. The ease of computation distinguishes the Gauss-Seidel method. The method of steepest descent involves the most difficult computations of the three, but relatively rapid convergence to a solution for systems in which the number of unknowns is "large" is its reward. The method to use in solving any particular system is then chosen with the special characteristics of the system, the requirements of the solution and the capabilities of the computer in mind.
In using an iterative method a computer ceases to look upon the problem as that of finding those values of $x_1$, $x_2$, ..., $x_m$ which satisfy the equations, but regards it instead as the determination of a set of values which make the residuals zero.

Since each iteration is a check on the progress of the method, it is usually possible to see when things are getting "out of hand". After one or two iterations the changes in the computed values should be gradual. Erratic changes will mean that the method is either not applicable to this type of problem or that an error of some kind has been made in obtaining the erratic value and therefore the value should be checked before proceeding further.

To illustrate the essential characteristics of each of the three methods discussed the same problem will be worked by each method. The last two methods described were originally self checking, but the adaptations made here were not, therefore a check was taken at periodic intervals, that is, substitution of $x_p$ was made in each of the original equations and the values of the residuals so computed were taken as the starting point of the next series of iterations.

Each method was begun with $x_0 = 0$ for purposes of comparison. The solution of this system (by direct methods) is -1, 1, 2. The approximate solutions achieved are such that they are stable to three decimal places (in the methods of Gauss-Seidel and relaxation). This can be verified by examination of the previous values for the unknowns and the observation that the total of the residuals whether added or subtracted from the values of the unknowns would not influence the third decimal place.
of the unknowns.

Only seventeen steps were needed for this result with the method of relaxation while twenty steps were needed for the Gauss-Seidel method. By under and overrelaxation the solution may be arrived at much faster; see the last example. The method of steepest descent which converged so slowly in this example is really the fastest converging when \( n \) is "large".

In certain iterations the residual had to be changed by amounts too large or too small. An automatic machine would probably be designed to change the residuals by too large an amount. A better procedure would be to consistently round the change in the variable to the nearest even or odd figure. This occurred in the fourteenth iteration of the Gauss-Seidel method. At times the values of the unknowns or the residuals may even worsen rather than improve; see the sixteenth iteration of the Gauss-Seidel method.

### 6. Examples

Let the equations be:

\[
\begin{align*}
  f_1(x_1, x_2, x_3) &= 4 + 5x_1 + x_2 = 0 \\
  f_2(x_1, x_2, x_3) &= -1 - x_1 - 8x_2 + 4x_3 = 0 \\
  f_3(x_1, x_2, x_3) &= 13 = 2x_1 + 3x_2 - 7x_3 = 0
\end{align*}
\]

\[
y - Ax = 0 \quad y = \begin{bmatrix} 4 \\ -1 \\ 13 \end{bmatrix} \quad A = \begin{bmatrix} -5 & -1 & 0 \\ 1 & 8 & -4 \\ -2 & -3 & 7 \end{bmatrix} \quad x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix}
\]

Since a positive definite matrix is needed for most applications of the method of steepest descent we will premultiply both sides of the preceding equation by \( A^t \).
Letting \( z = A^t y \) and \( B = A^t A \) we have an equation in which the coefficient matrix \( B \) is positive definite:

\[ z - Bx = 0. \]

The solution of this equation is also the solution of the equation

\[ y - Ax = 0. \]

The transformed equations being:

\[
\begin{align*}
    g_1(x_1, x_2, x_3) &= -47 - 30x_1 - 19x_2 + 18x_3 = 0 \\
    g_2(x_1, x_2, x_3) &= -51 - 19x_1 - 74x_2 + 53x_3 = 0 \\
    g_3(x_1, x_2, x_3) &= 95 + 18x_1 + 53x_2 - 65x_3 = 0.
\end{align*}
\]
The method of steepest descent

<table>
<thead>
<tr>
<th>( x_p )</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( r_p )</th>
<th>( r_1 )</th>
<th>( r_2 )</th>
<th>( r_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_0 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>( r_0 )</td>
<td>-47</td>
<td>-51</td>
<td>95</td>
</tr>
<tr>
<td>( x_1 )</td>
<td>0.40369</td>
<td>0.43804</td>
<td>0.81596</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( x_2 )</td>
<td>0.38079</td>
<td>1.03636</td>
<td>0.36799</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( x_3 )</td>
<td>-0.30079</td>
<td>1.03636</td>
<td>0.36799</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( x_4 )</td>
<td>0.90138</td>
<td>0.44898</td>
<td>0.59832</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( x_5 )</td>
<td>-0.78914</td>
<td>0.59832</td>
<td>0.59832</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( x_6 )</td>
<td>0.95273</td>
<td>0.81596</td>
<td>0.81596</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( x_7 )</td>
<td>0.95273</td>
<td>0.81596</td>
<td>0.81596</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( x_8 )</td>
<td>0.95273</td>
<td>0.81596</td>
<td>0.81596</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( x_9 )</td>
<td>0.95273</td>
<td>0.81596</td>
<td>0.81596</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( x_{10} )</td>
<td>0.95273</td>
<td>0.81596</td>
<td>0.81596</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( x_{11} )</td>
<td>0.95273</td>
<td>0.81596</td>
<td>0.81596</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The Gauss-Seidel method

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$\phi_{p}e_1$</th>
<th>$r$</th>
<th>$r_1$</th>
<th>$r_2$</th>
<th>$r_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$x_1 = x_2 = x_3 = 0$</td>
<td>4</td>
<td>-1</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.025</td>
<td>.025</td>
<td>$\phi_{g}e_1 = -0.8$</td>
<td>0</td>
<td>-0.2</td>
<td>11.4</td>
<td></td>
</tr>
<tr>
<td>6.1786</td>
<td>1.61786</td>
<td>1.61786</td>
<td>$\phi_{g}e_2 = 0.25$</td>
<td>0.025</td>
<td>0</td>
<td>6.4714</td>
<td>-0.00002</td>
</tr>
<tr>
<td>795</td>
<td>7833</td>
<td>1.96570</td>
<td>$\phi_{g}e_1 = 0.005$</td>
<td>0</td>
<td>6.4664</td>
<td>.00998</td>
<td>2.43488</td>
</tr>
<tr>
<td>95666</td>
<td>97713</td>
<td>2.00271</td>
<td>$\phi_{g}e_2 = 0.16166$</td>
<td>0</td>
<td>1.55306</td>
<td>-0.3232</td>
<td></td>
</tr>
<tr>
<td>.9549</td>
<td>1.00079</td>
<td>2.00163</td>
<td>$\phi_{g}e_1 = 0.00004$</td>
<td>19413</td>
<td>.00002</td>
<td>.25907</td>
<td></td>
</tr>
<tr>
<td>9551</td>
<td>1.00016</td>
<td>2.00031</td>
<td>$\phi_{g}e_2 = 0.00003$</td>
<td>19413</td>
<td>.00002</td>
<td>.25907</td>
<td></td>
</tr>
<tr>
<td>.99549</td>
<td>1.00018</td>
<td>2.00003</td>
<td>$\phi_{g}e_3 = 0.00004$</td>
<td>19413</td>
<td>.00002</td>
<td>.25907</td>
<td></td>
</tr>
<tr>
<td>99549</td>
<td>1.00004</td>
<td>2.00002</td>
<td>$\phi_{g}e_1 = 0.00004$</td>
<td>19413</td>
<td>.00002</td>
<td>.25907</td>
<td></td>
</tr>
</tbody>
</table>
The **method of relaxation**

<table>
<thead>
<tr>
<th>(x_1)</th>
<th>(x_2)</th>
<th>(x_3)</th>
<th>(\phi_{p e_1})</th>
<th>(r_1)</th>
<th>(r_2)</th>
<th>(r_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(x_1 = x_2 = x_3 = 0)</td>
<td>4</td>
<td>-1</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(\phi_{e_3}^1 = \phi_{e_3}^2 = 1.85714)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.96071</td>
<td></td>
<td></td>
<td>(\phi_{e_3}^2 = 0.80357)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.92366</td>
<td></td>
<td></td>
<td>(\phi_{e_3}^1 = -0.96071)</td>
<td>0.0002</td>
<td>96071</td>
<td>48951</td>
</tr>
<tr>
<td>1.97851</td>
<td>2.00451</td>
<td>1.00186</td>
<td>(\phi_{e_2}^3 = 0.02600)</td>
<td>18079</td>
<td>0.0003</td>
<td>18203</td>
</tr>
<tr>
<td>0.99687</td>
<td></td>
<td></td>
<td>(\phi_{e_1}^3 = -0.03616)</td>
<td>0.0001</td>
<td>14019</td>
<td>0.07229</td>
</tr>
<tr>
<td>-1.00037</td>
<td></td>
<td></td>
<td>(\phi_{e_2}^1 = -0.0282)</td>
<td>0.01751</td>
<td>-0.01125</td>
<td>0.00001</td>
</tr>
<tr>
<td>1.00089</td>
<td>2.00028</td>
<td>1.99998</td>
<td>(\phi_{e_2}^1 = -0.0282)</td>
<td>0.0033</td>
<td>-0.00070</td>
<td>0.00014</td>
</tr>
<tr>
<td>-1.00004</td>
<td></td>
<td></td>
<td>(\phi_{e_2}^1 = -0.0282)</td>
<td>0.0003</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The **method of relaxation** with under and over-relaxation

<table>
<thead>
<tr>
<th>(x_1)</th>
<th>(x_2)</th>
<th>(x_3)</th>
<th>(\phi_{p e_1})</th>
<th>(r_1)</th>
<th>(r_2)</th>
<th>(r_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(x_1 = x_2 = x_3 = 0)</td>
<td>4</td>
<td>-1</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(\phi_{e_3}^1 = 2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td></td>
<td>(\phi_{e_2}^1 = 5)</td>
<td>1</td>
<td>-1</td>
<td>2</td>
</tr>
<tr>
<td>-1</td>
<td></td>
<td></td>
<td>(\phi_{e_1}^2 = -1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
BIBLIOGRAPHY


Dwyer, Paul S. Linear Computations. New York: John Wiley and Sons, Inc., 1951.


