FINAL REPORT
NASA CONTRACT NAS8-21484

## Solution of Systems of Nonlinear Equations

Submitted by Victor J. Law, Principal Investigatom for Tulane University, New Orleans, Louisiana
$\underline{A B S T R A C T}$


A generalized method called Diagonal Discrimination for nonlinear algebraic and transcendental equations is described. A computer program which implements the Diagonal. Discrimination technique has been written in IORTRAN IV for the Univac 1108 computer. A users manual is included which gives detailed instructions of how to implement the computer prognam.

I. Diagonal Discrimination for Nonlinear Equation Solving

The method of Diagonal Discrimination (DD) was first described by Fariss and Law (1). A'brief description of the algorithm will now be given.

DD belongs to a class of methods such that the computations always begin from a point $x^{\circ}$ in $n$-dimensional space and a move or increment, $\Delta x$, is computed such that $\underline{x}^{1}=x^{0}+\alpha \Delta x$ forms a. search vector along which a "better" point is sought. This logic is repeated until no further improvement is possible. The choice of the scalar $\alpha$ is made by a one-dimensional search procedure.

The success of the method depends on a property of the $\Delta x$ vector which shall be called truncation convergence. An algorithm for minimization has this property if, for sufficiently small $\alpha$ the objective function $q\left(x^{\circ}+\alpha \Delta x\right)$ takes on a smaller value than $q\left(x^{\circ}\right)$. What this means is that $\Delta x$ must point in a direction such that $q$ decreases at least locally. Hence, a better point can always be found by truncating $\alpha$ to some small positive value.

DD uses a unique combination of the method of weighted steepest descent and the Gauss-Newton method to minimize a sum of squares function. A brieff review of these methods will now be given.
I. 1. Formulation as a Minimization Problem. The specific problem to which attention is now given is that of finding the value of an n-vector $x$ such that the equations

$$
\begin{gather*}
\cdot-2- \\
f_{j}(x)=0 ; \quad, j=1,2, \ldots, n \tag{1-1}
\end{gather*}
$$

are satisfied. The $\underset{\text { f }}{ }$ functions, in general, are nonlinear. This problem may easily be formulated as a minimization problem by forming the sum of squares of residual as an objective function.

$$
\begin{equation*}
q=1 / 2 \sum_{j=1}^{n} f_{j}^{2} \tag{1-2}
\end{equation*}
$$

Clearly, if $q$ is minimized to its absolute minimum of zero, then a solution to the original problem has been obtained.
I. 2. Ordinary Steepest Descent. Perhaps the oldest and still very popular method for unconstrained minimization is the method of steepest descent (SD). Strictly speaking, this method is a continuous one rather than a discrete one in that the path of steepest descent is a continuous curve. For practical use, however, the direcrion of steepest descent is found at the base point, $x^{\circ}$, and this direction is used to form the search vector.

The direction of steepest- descent is given by

$$
\begin{equation*}
\Delta x=\frac{-\gamma g\left(x^{0}\right)}{\left|\underline{g}\left(x^{0}\right)\right|} \tag{1-3}
\end{equation*}
$$

```
where \(g=\frac{\partial q}{\partial x}\), i.e., the gradient of \(q\)
    \(\gamma=\) steepest descent distance factor
```

The search vector then becomes

$$
\begin{equation*}
\underline{x}=\underline{x}^{\circ}+\alpha \underline{\Delta x} \tag{1-4}
\end{equation*}
$$

The value of $\alpha$ is usually selected by performing a one-dimensional search for a minimum in $q$ with respect to $\alpha$. Perhaps the most popular one-dimensional search is the Golden Section Search (2). A more sophisticated method is described by Fariss and Law (1) and is the one implemented in the computer programs listed in this report. It is also possible to simply find a value of $\alpha$ for which $q$ is smaller than at the base point rather than finding the minimum. There is no generally "best" procedure to use. Judicious selection of $\gamma$ can greatly enhance the performance of (SD). It is efficient to select this value of $\gamma$ based on the optimal $\alpha$ from the previous iteration. One means of accomplishing this is to use the following relationship:

$$
\frac{\gamma^{(i+1)}}{\gamma^{(i)}}=\ldots \ldots \cdot \begin{cases}a & ; \alpha_{0}^{(i)}=0  \tag{1-5}\\ b & ; \alpha_{0}^{(i)}>b \\ \alpha_{0}^{(i)} ; & ; 0<\alpha_{0}^{(i)} \leq b\end{cases}
$$

there $\alpha_{0}^{(i)}=$ optimal value of $\alpha$ from the $i$ th iteration
$\gamma^{(i)}=\gamma$ from iteration $i$
$\gamma^{(i+1)}=\gamma$ for iteration $i+1$
a $\quad=$ lower limit on correction factor
b $\quad=$ upper limit on correction factor

This fommula is purely arbitrary and merely attempts to update $\gamma$ based on past experience while not allowing large changes in it from one iteration to the next. The previous $\gamma$ is multiplied by a factor between . $a$ and $b$, with a corresponding to $\alpha_{0}{ }^{(i)} \dddot{=} \ddot{0}$ and $b$ to $\alpha_{0}^{(i)}>b$. If $\alpha_{0}{ }^{(i)}=1, \gamma$ is not changed. Recommended values for $a$ and $b$ are $1 / 4$ and 4 , nespectively.

The $\Delta x$ vector of Equation (1-3) is normal to the objective function contour at the base point and is guaranteed to have the truncation convergence property. This is obvious in that

$$
\begin{equation*}
\left.\frac{d q}{d \alpha}\right|_{\alpha \rightarrow 0}=\left.\underline{g}^{T} \frac{d x}{d \alpha}\right|_{\alpha \rightarrow 0}=-\underline{g}^{T} \underline{g}=-|g|^{2} \tag{1-6}
\end{equation*}
$$

The most serious drawback of the method of steepest descent is the zig-zag tendency especially when near the solution. This property is best explained in the two-dimensional case. Referring to Figure 1-1, it is easily seen that successive directions of steepest descent will be orthogonal (perpendictular in 2-dimensions).

, Fig. 1-1. Steepest descent for a quadratic function.

In order to overcome this difficulty, several techniques have been suggested [see, for example, Booth (3)]. One particular modificaton is the Method of Parallel Tangents developed by Shah, Buehler and Kempthorne (4). Weighted steepest descent is another modification and will be described in the sequel.
I. 3. The Gauss-Newton Method (GN). This method uses Newton's method as a basis. That is, the equation for determing $\Delta x$ is given by

$$
\begin{equation*}
\frac{\partial^{2} q}{\partial \underline{x}^{2}} \Delta x=\frac{\partial q}{\partial \underline{x}} \tag{1-7}
\end{equation*}
$$

A useful and effective approximation is made for the Hessian, $\frac{\partial^{2} q}{\partial \underline{x}^{2}}$, however. In terms of the sum of squares $q$ function, Equation (I-2),

$$
\begin{equation*}
\frac{\partial q}{\partial x_{i}}=\sum_{j=1}^{n} f_{j} \frac{\partial f_{j}}{\partial x_{i}} \tag{1-8}
\end{equation*}
$$

and

$$
\begin{equation*}
\frac{\partial^{2} q}{\partial x_{i} \partial x_{k}}=\sum_{j=1}^{n} \frac{\partial f_{j}}{\partial x_{i}} \frac{\partial f_{j}}{\partial x_{k}}+f_{j} \frac{\partial^{2} f_{j}}{\partial x_{i} \partial x_{k}} \tag{1-9}
\end{equation*}
$$

Now, for equation solving the $f_{j} \rightarrow 0$ as the solution is approached. Therefore, the so-called Gauss-Newton approximation is to omit the second term in the expression for $\frac{\partial^{2} q}{\partial x_{i} \partial x_{k}}$. That is,

$$
\frac{\partial^{2}}{\partial x_{i} \partial x_{k}} \approx \sum_{j=1}^{n} \frac{\partial f_{j}}{\partial x_{i}} \frac{\partial f_{j}}{\partial x_{k}}=G_{i k}
$$

Thus, the relation for step size determination becomes

$$
\begin{equation*}
G \underline{\Delta x}=-\underline{g} \tag{1-10}
\end{equation*}
$$

In the absence of singularity of the $G$ matrix, the (GN) procedure is very efficient at converging to the solution from a near point (i.e., where. the sum of squares function becomes nearly quadratic and the $f_{j}$ are small). This behavior could be anticipated from the approximate quadratic representation of $q$ by using the $G$ matrix as an approximation to $\partial^{2} q / \partial x^{2}$. Thus, quadratic convergence is obtained in the neighborhood of the solution.

As a protection against very long search trials, it is customary to limit the length of the search vector to some arbitrary value. While this will usually force convergence, it has one distinct disadvantage. In order to clearly understand why this is so, consider a problem where only one variable causes the singularity. In such a case the moves predicted for all other variables would be quite good. Thus, the truncation of the entire search vector is clearly inefficient in that only the move in the
maverick variable need be truncated. This justifies the need for the following two operations:
(I) Sorting those variables which cause the singularity in $G$.
(2) Selectively truncating only the moves for these variables. More will be said about this later.

The (GN) method is equivalent to the well known Newton-Raphson (5) (NR) method for equation solving. The (NR) method uses as its basis the linearization of each equation in the neighborhood of the base point. Thus,

$$
\begin{equation*}
f_{j}(\underline{x}) \simeq \dot{f}_{j}\left(\underline{x}^{\circ}\right)+\sum_{i=1}^{n} \frac{\partial f_{j}\left(x^{\circ}\right)}{\partial x_{i}} \Delta x_{i} ; j=1,2, \ldots, n \tag{1-11}
\end{equation*}
$$

The move is then determined so that each $f_{j}(\underline{x})$ would become zero if all functions are linear. In vector-matrix form, these relations become

$$
\begin{equation*}
\mathrm{J} \underline{\Delta \mathrm{x}}=-\underline{f} \tag{1-12}
\end{equation*}
$$

where

$$
J=\text { the Jacobian matrix with } \bar{J}_{i j}=\frac{\partial f_{i}}{\partial x_{j}}
$$

In equation (1-12) is premultiplied by $J^{T}$ (supenscript $T$ indicates the matrix transpose), there results

$$
\begin{equation*}
\left(J^{T} J\right) \underline{x}=-J_{\underline{f}} \tag{1-13}
\end{equation*}
$$

Which is identical to Equation (1-11). That is,

$$
\begin{align*}
& G=J^{T} J  \tag{1-14}\\
& \frac{\partial q}{\partial \underline{q}}=J^{T} \underline{f} \tag{1-15}
\end{align*}
$$

Equation (i-14) constitutes a proof that $G$ is always at least positive semi-definite in that this property aiways holds for a product of a matrix and its transpose.
I. 4. Weighted Steepest Descent (WSD). Weighted steepest descent is a modification of the method of steepest descent. It arises from modifying the elements of the search vector by non-equal positive multiplicative factors chosen so as to produce a more effective vector. Specifically, these factors may be selected so that, under favorable circumstances (i.e., when $G$ is strongly positive definite), the search vector will coincide with the one produced by Newton's method, provided the method is applied in a coordinate system where, in relation to the quantity being minimized, there is no local interaction of variables. The necessity of using such a coordinate system will be made clear in the sequel.

The coordinate system required may be created by transforming the $G$ matrix into diagonal form. Let $T$ by a non-singular transformation matrix such that

$$
\begin{equation*}
T^{T} G T=D \tag{1-16}
\end{equation*}
$$

where $D$ is diagonal. Methods for computing such transformation matrices are discussed by Wilde and Beightler (6). If this transformation is applied to Equation (1-10), there results

$$
\begin{equation*}
T^{T} G T T^{-1} \Delta x=-T^{T} g \tag{1-17}
\end{equation*}
$$

By defining the new variables (coordinates)

$$
\begin{equation*}
\underline{y}=T^{-1} \Delta x \tag{1-18}
\end{equation*}
$$

Equation (1-17) for the search vector $\underline{y}$ in terms of transformed coordinates may be written as

$$
\begin{equation*}
D \underline{y}=-\mathrm{T}^{\mathrm{T}} \underline{\mathrm{~g}} \tag{1-19}
\end{equation*}
$$

On the same basis, the steepest descent equation becomes

$$
\begin{equation*}
\underline{y}=-T^{T} \underline{g} \tag{I-20}
\end{equation*}
$$

In order for coincidence to exist between Newtonian and steepest descent vectors, Equation (1-20) must be modified to

$$
\begin{equation*}
\underline{y}=-\mathrm{kD}^{-1} \mathrm{~T}^{\mathrm{T}} \underline{\mathrm{~g}} \tag{1-21}
\end{equation*}
$$

in which $\mathrm{kD}^{-1}$ supplies the necessary weighting. The scalar factor k is indeterminate (but positive), since steepest descent is intended to define only the direction of the search vector.

The general form of the weighted steepest descent equations, on a transformed coordinate basis, may then be written as

$$
\begin{equation*}
\underline{y}=-\mathrm{WT}^{\mathrm{T}} \underline{g} \tag{1-22}
\end{equation*}
$$

in which $\underline{W}$ is a diagonal matrix of positive elements. Consequently, coincidence between Newtonian and steepest descent vectors can be achieved only if all diagonal elements $d_{i i}$, are non-zero positive, that is, if the $G$ is positive definite. Furthermore, experience has shown that, if reasonable extemal scaling has been applied, weighting factors which are excessively large to yield an effective search vector will result from Equation (1-2l) when one on more are orders of magnitude sinaller than others. It may be concluded from this that the favorable circumstances under which it is feasible and reasonable to weight steepest descent so as to force coincidence with the Newtonian search vector are confined to cases where all $\mathrm{d}_{\mathrm{ij}}$ are positive and of the same order of magnitude.

An important feature of the weighted steepest descent method is that it can be adapted to deviate from Newton's method when circumstances do not warrant their coincidence. This can be achieved by adapting the calculation of the weight factors. The suggested method for doing this will now be described.

First, assume that the diagonal elements, $d_{i i}$ and the corresponding columns of the transformation matrix, $T$, are rearranged to achieve descending algebraic order. That is, $d_{11}>d_{22}>\ldots>d_{n n}$. If $k$ in equation (3-: is chosen as $d_{11}$, then

$$
w_{1 i}=1
$$

and ideal weighting (i.e., that which causes coincidence with Newton's method) is given by

$$
\begin{equation*}
w_{i i}=\frac{d_{11}}{d_{i i}} ; i=2,3, \ldots, n \tag{1-23}
\end{equation*}
$$

An adaptation of the ideal case which has been used successfully is, for $i=1,2, \ldots, n$

$$
w_{i i}=\left\{\begin{array}{l}
\frac{d_{11}}{d_{i i}} \quad \text { if } \frac{d_{11}}{d_{i i}} \leq \beta \text { and } d_{i i}>0  \tag{1-24}\\
. \beta \quad-i f \frac{d_{11}}{d_{i i}}>\beta \text { or } d_{i i} \leq 0
\end{array}\right.
$$

In Equation (1-24), $\beta$ is the maximum weight factor allowed. Thus, if $d_{i j} \ll d_{11}$ or if $d_{j i}<0, w_{i i}$ deviates from the ideal one and is assigned the value $\beta$. It should be pointed out that Equation (1-24) is not the only choice which can be made for non-ideal weighting but is one which has been used successfully.

An outline of the weighted steepest descent method follows:

1) Select a distance factor $\gamma$ and a base point $\underline{x}^{0}$.
2) Compute $g$ and $G$ at the base point.
3) Find the transformation matrix $T$ and the corresponding diagonal matrix D. Reorder $D$ and columns of $T$ so that the resulting $d_{i i}$ are in descending algebraic order.
4) Compute a set of weighting factors from Equation (1-24).
5) Compute the $\underline{y}$ vector as follows:
a) First compute an interim $\underline{y}$ indicated here by $\underline{y}^{i}$ as follows:
$y_{i}^{i}=-w_{i i} p_{i} ; i=1,2, \ldots n$ where $\underline{p}=T^{T} \underline{g}$
b) Set $y_{\max }^{i}=\max \cdot\left|y_{i}^{i}\right|$ $\mu=\gamma / y_{\max }^{i}$
c) Then, $y_{i}=-\mu w_{i i} p_{i}$

Note that the $y_{i}$ of largest magnitude has a magnicude equal to $\gamma$.
6) Perform a one-dimensional search along the vector

$$
\begin{equation*}
\underline{x}=\underline{x}^{0}+\alpha \underline{\Delta x} \tag{1-28}
\end{equation*}
$$

$$
\begin{equation*}
\text { where } \quad \underline{\Delta x}=T \underline{y} \tag{I-29}
\end{equation*}
$$

7) Update $\gamma$ based on the experience of the one-dimensional search [ see Equations (1-5)].
8) Go to step (2) and repeat until convergence is achieved. Equations like (1-22) and (1-23) could also be applied to define a weighted steepest descent algorithm for the original, untransposed coordinate system. The best analogy to Equation (1-24) would involve using the diagonal elements of the $G$ in place of the $d_{i i}$. This approach requires substantially less computation to determine a search vector, since only the diagonal of the $G$ is required, and the diagonalization calculation is not used. However, if interaction is at all significant, there will be no coincidence at all wïth the Newtonian vector, and this approach will be subject to most of the inefficient convergence problems usually experienced with ordinary steepest descent. One significant reason for this is that it is impossible, in general, to weight the $\Delta \mathrm{x}_{\mathrm{i}}$ to make the $S D$ search vector coincide with the NM search vector. To see this more clearly, consider a quadratic function with contours as shown in Figure 1-2.


Fig. 1-2. SD and Exact Search Vectors.

The SD direction produces positive values for both $\Delta x_{1}$ and $\Delta x_{2}$ whereas the exact search vector requires $\Delta x_{I}$ to be negative and $\Delta x_{2}$ to be positive. Thus, no positive weighting factors exist. It is not proper to use negative weight factors since the property of truncation convergence would be destroyed. To see this, consider a diagonal matrix of weights, V. Then a weighted SD direction would be

$$
\begin{equation*}
\Delta \mathrm{x}=-\mathrm{Vg} \tag{1-30}
\end{equation*}
$$

Then Equation (1-6) becomes

$$
\begin{equation*}
\left.\frac{\mathrm{dq}}{\mathrm{~d} \alpha}\right|_{\alpha \rightarrow 0}=\left.\underline{g}^{\mathrm{T}} \frac{\mathrm{dx}}{\mathrm{~d} \mathrm{\alpha}}\right|_{\alpha \rightarrow 0}-\underline{g}^{\mathrm{T}} \mathrm{Vg} \tag{1-3i}
\end{equation*}
$$

Thus $\left.\frac{d q}{d \alpha}\right|_{\alpha \rightarrow 0}$ will be always negative only if $V$ is positive definite. Hence, each scale factor must be positive.

Therefore, scaling the $\underline{y}$ vector is another means of accelerating the convergence of steepest descent. As long as the $d_{i i}$ are positive, the ideal weighting factons given by Equation (1-23) can be quite effective toward this end. However, if there is a large difference in the magnitudes of the $d_{i i}$ or if any $d_{i i}$ are negative or zero, the ideal weighting is either dangerous or not possible. In these situations it has been found that the concept of weighting the $\underline{y}$ vector is still useful but must be suitably modified. One practical means of accomplishing this is that given by Equation (1-24).
I. 5. Diagonal Discrimination (DD). It was discovered that (WSD) had one significant shortcoming. That is, no consideration is given to the relation between the $d_{i i}$ and the comesponding elements of the transformed gradient, $p$. Thus, even if $d_{i i}$ is such that ideal weighting is allowed, the $y_{i}$ produced by (WSD) might be excessively large because of a large $p_{i}$. The quadratic approximation of $q$ at $x^{\circ}$ upon which Newton's method is based is likely to be'poor at points far removed from the base' point. Therefore, only in the very fortunate but, unlikely (at least in practical applications) case that $q$ is almost exactly quadratically dependent upon a particular $y_{i}$ would a long move in that variable be warranted. In order to account for this situation the weighted steepest descenr method was modified into another method called diagonal discrimination (DD).
(DD) is a method which involves choosing a superior search vector based upon the following two principles:

1) Transformation of the coordinate axes in order to remove local interaction between variables.
2) Computing the elements of the search vector discriminantly depending upon the relationships between the diagonal elements of the transformed Hessian and the length of a move as predicted by Newton's method in the transformed coordinate system.

The development of the method is conveniently begun by considering Equation (1-19). Assume again that the columns of $T$ have been reamanged so that the $d_{i i}$ appear in descending algebraic onder.

The logic of. DD presupposes that reasonable external scaling of the
variables has been accomplished so that the following assumptions are likely to be valid:

1) The ac̣ceptability of a move (from a base point) calculated by Newton's method can be judged from its length--long moves are suspect.
2) In regard to the diagonalized $G$, positive $d_{i i}$ which are several orders of magnitude smaller than the largest one are indicative of a linear dependence of $q$ on the associated $y$ coordinate. Negative $d_{i i}$ are associated with those $y$ coordinates for which the objective function tends to exhibit a local maximum rather than a minimum.

Diagonal Discrimination then attempts to combine the best features of the ( $G-N$ ) method and (WSD) by computing the $\underline{y}$ vector components discriminantly That is, if the ( $G-N$ ) move for a particular $y_{i}$ is not too long and if the associated $d_{i i}$ is positive and not greatly different from $d_{11}$, then the ( $G-N$ ) calculation is used. Otherwise, $y_{i}$ is found by (WSD) logic. A step-wise presentation of the logic for $D D$ follows:

1) Select a base point; $\underline{x}^{\circ}$, a steepest descent distance factor, $\gamma$, and a steepest descent threshold factor, $\delta$.
2) Compute $g$ and $G$ at the base point.
3) Find the $T$ and $D$. Order $D$ and $T$ so that the $d_{i i}$ are in descending algebraic order.
4) Starting with $y_{1}$, the elements of $y$ are computed by (G-N) logic, that is

$$
\begin{equation*}
y_{i}=-p_{i} / d_{i i} \tag{1-32}
\end{equation*}
$$

until either
a) $i=n$
b) $\quad d_{i i} \leq \varepsilon d_{i l}$
c) $y_{i}>\delta$
5) When the ( $G-N$ ) sequence is terminated for reason (b) or (c) above, at the kth $y$ variable, then a switch is made to weighted steepest descent logic treating $y_{K}$ as the first (WSD) variable. That is, $\mathrm{y}_{\mathrm{K}}$ is assigned a weighting factor of l and $\mathrm{y}_{\mathrm{K}}, \mathrm{y}_{\mathrm{K}+1}, \ldots, \mathrm{y}_{\mathrm{n}}$ are computed by the (WSD) method as described above.
6) If any $d_{i i} \leq \varepsilon d_{11}$, set $y_{i}=0$ (these are called null-effect parameters and are discussed in the next subsection).
7) The resulting $\underline{y}$ vector is converted back into a $\Delta x$ vector by

$$
\begin{equation*}
\Delta \underline{x}=T \underline{y} \tag{1-33}
\end{equation*}
$$

and a one-dimensional search is performed along the $\underline{x}$ vector as given by Equation (1-28).
8) The distance factor, $\gamma$, is updated using Equation (1-5).
9). Go to step (2) and repeat until convergence is achieved.

It can be shown in that the DD search vector has the property of truncation convergence. A discussion of parameter selection ( $\beta, \gamma, \delta, \varepsilon$ ) is given below.

Selection of $\beta$ :
The choice of $\beta$, the maximum allowable weight factor in Equation (1-24) is not critical. That is, the sensitivity of WSD to $\beta$ is relatively slight. A value of about $10^{4}$ has been used satisfactorily in almost all applications of DD.

Selection of $\gamma$ :
If the variables have been scaled (see next section) in any reasonable way their expected value should be about unity, at least within a few orders of magnitude. A change of $20-50 \%$ would, therefore, be considered relatively lange. Therefore, if $\gamma$ is chosen to be 0.2 , the length of the $S D$ search vector will limit each individual variable increment to be less than $20 \%$ of unity. The actual initial choice of $\gamma$ is not too critical since it is updated at each iteration.

Selection of $\delta$ :
The steepest descent threshold factor, $\delta$, should be chosen with the philosophy that if some $y_{i}$ as calculated by Newton's method is greater than $\delta$, then a switch should be made to WSD logic. Again, if reasonable external scaling has been applied, a value of about 0.2 for $\delta$ has been found to be satisfactory.

Selection of $\varepsilon$ :
The parameter $\varepsilon$ is used to distinguish between small $d_{i i}$ and those which should be treated as zero. The choice of $\varepsilon$ depends upon the number of significant digits carried in the arithmetic. For digital computer which carry d digits, it is recommended that $\varepsilon$ be chosen as follows:

$$
\begin{equation*}
=10^{-\mathrm{d}} \tag{1-34}
\end{equation*}
$$

I. 6. Singularity of $G$ and Null Effect Variables. The singularity of the $G$ matrix is usually caused by a phenomenon which shall be called null effect. Those variables which cause this condition will be called null effect variables. Null effect occurs when perturbation of a parameter or of some linear combination of parameters has no significant effect on any of the residuals in the sum of squares function. In a well posed problem, null effect should not, of course, be present at the solution point. For systems of nonlinear equations, however, null effect is common at points removed from the solution. This is caused by local redundancies on inconsistencies in the linearized versions of the equations. In either case, two or more linear equations become parallel to each other and hence no . solution exists. As an example, consider the problem illustrated in Fig. 1-3.


Figure 1-3. Illustration of Null Effect
I. 7. External Scaling. External scaling of variables is desirable (and highly recommended), primarily because of the problem of heterogeneous physical dimensions. Variables which have different physical dimensions will naturally have different ranges of variation and different impact on the objec function. In any numerical procedure it is desirable that the variables have similar ranges.of variation because of round-off considerations. Vastly different effects on the objective function lead to very elongated contours of the objective function. The latter problem results in large differences in the diagonal elements of the transformed Gauss-Newton matrix, G.

The most natural way to scale variables is by way of dimensional analysis. That is, some variables may be made dimensionless by choosing a natural scale factor. An example of this for temperature as the variables would be to set

$$
\begin{equation*}
t^{*}=\frac{t-t_{L}}{t_{U}-t_{L}} \tag{1-35}
\end{equation*}
$$

where $t^{*}=$ dimensionless temperature
$t_{U}$, $t_{L}=$ upper and iower bounds on the range of temperature to be considered

In addition to having no physical units, the dimensionless temperature would also, conveniently, vary between zero and one.

However, it is not always possible or practical to scale the problem as suggested on the basis of dimensional analysis. Another way of arriving at scaled variables which has proven to be effective and satisfactory is to
choose an increment for numerical differentiation which will cause a change in the objective function in the last three on four digits (as available on the computer to be used). Then choose an external scale factor which will cause the scaled increment for differentiation to be the same for all variables. That is,

$$
\begin{equation*}
e_{i}=\frac{\delta y_{i}}{c_{d}} \tag{1-36}
\end{equation*}
$$

where $\quad c_{d}=$ increment for differentiation of scaled variables $e_{i}=$ external scale factor for $x_{i}$
$\delta y_{i}=$ chosen increment for numerical differentiation of unscaled variables

This technique can usually be carried out easily if the problem solver has some basic knowledge of the problem and if sound engineering judgement is applied. Variable scaling is automatically provided in the computer programs described herein.
II. User's Manual for Implementing Computer Programs.
II. 1. Basic Characteristics. The basic characteristics of the system are as follows:

1) The user sets up arrays containing starting values or simply a range for the parameters being sought, and then starts DD action by the appropriate subroutine calils.
2) Thereafter, until the iterative procedure is terminated by the subroutines, the values contained in the parameter amay are under control of the subroutines. The user must provide programming to calculate residuals for the equations for any set of values in the parameter array, on demand by the $D D$ algorithm.
3) When control is released, the parameter array will contain "best" values, i.e., those which cause the sum of squares of residuals to be a minimum.
4) All required partial derivatives are obtained numerically.
5) There are no internal limits on the size of problems due to dimensions, etc. It is possible to solve problems as large as about 200 parameters and 200 equations.
6) No special NASA system features are required.
II. 2. Skeletal Illustration. A skeletal illustration of a main program which meets the basic requirements for using the subroutines which implement diagonal discrimination is shown in the following:
```
DIMENSION XX(`), SCALE (`),S(\cdots),SX(\cdots),G(\cdots), XX(\cdots), MODE (`),
                                    RHS(`.), DXBAR(`), DXX(..) DQ(..), DQX(..), YBASE(..),
```



```
.COMMON / DDCOM / COM(9), KCOM(13)
EQUIVALENCE (KCOM(1), NCASE), (KCOM(2), NVAR), (KCOM(3), IPRNT)
NCASE = (specify a number of equations)
NVAR = (specify number of unknown parameters)
IPRNT = (specify output option)
XMIN( )=
X( ) = (make initial guesses) or }\quad\textrm{XMAX( )= (specify range on each variable)
KODE = 0 KODE = 1
```

DY, X, SCALE, D2Q, SCALR)
IF (KCOM(5) .LE. O) GO TO 100
CALL NEWSCL (X, SCALĖ).
$Y()=$ (calculate residuals for all equations)
GO TO 20
100 IF (KCOM( 8) .EQ. 0) GO TO 200
GO TO 10
200 ... (problem completed)

The dimension statements establish memoxy requirements for all arrays. Minimum space requirements for these are outlined below:

| Array Name | Minimum Dimension |
| :---: | :---: |
| X,SCALE, $\mathrm{XX}, \mathrm{M} \mathrm{\emptyset DE}, \mathrm{RHS}, \mathrm{DXBAR}, \mathrm{DXX}, \mathrm{DQ}, \mathrm{DQX}, \mathrm{D} 2 \mathrm{Q}, \mathrm{SCALR}$, | NVAR |
| XMIN, XMAX |  |
| $Y$ Y YBASE | NCASE |
| S,SX,G | $\left(\right.$ NAR) ${ }^{2}$ |
| $D Y$ | VAR)X(NCASE) |

where NVAR $=$ Number of parameters to be determined
NCASE $=$ Number of equations

The CøMMØN statement creates Iinkage of the listed variables between the main program and the D.D. subroutines.

The regression calculations are initiated by calling subroutine DDRG, statement 20 above. However, prior to this, four things must have been done:

1) Specify NVAR and NCASE.
2) Set the print control, IPRNT.
3) Put starting values (initial guesses) for the parameter in the $X$ amray or specify lower and upper bounds on each parameter (XMIN and XMAX).
4) Set regression controls by calling subroutines CONST and DDSET.

The action required after the return from subroutine DDRG depends on the value of $\operatorname{KCOM}(5)$ as set by that subroutine:

When $\operatorname{KCOM}(5)=0$, the subroutines have probably completed 'their operation. If a normal finish was obtained, $\mathrm{KCOM}(8)$ will be returned as a zero. At this point the X array will contain the parameter values corresonding to the problem solution, and the final sum of squares, SIGMA, will be $\leq 10^{-9}$. If $K C O M(8) \neq 0$, a false solution was found and a restart (from a jifferent initial guess) should be initiated by returning to the call of CONST. When $K C O M(5)>0$, the main program must calculate the values of the "residuals" and place these values in the Y array. If the equations are written so that all terms appear on only one side of the equal sign (e.g., $x_{I}{ }^{2}+2 x_{2}{ }^{2}=0$ ), then the $Y$ array contains the current value of that side of the equation [e.g., $Y(1)=X(1) * 2+2.0 * X(2) * * 2$ ].

If it is desired to start the solution of another set of equations, NVAR, NCASE, IPRNT and MASCT should be reset and subroutines CONST and DDSET should be called again before the next initial call of DDRG.
II. 3. Internal Print Control and Output Description. The D.D. subroutines brought into action by the call of DDRG furnish considerable information during the progress of the calculation if the contained print statements are not suppressed. Suppression can be accomplished by setting IPRNT=0. Information for only the last iteration is obtained if IPRRNT=1. However,
when new problems are run, it is strongly recommended that the complete intemal print facility be used by setting IPRNT=2.

Output items appear in the following sequence:
I) Heading line
2) Table of starting values and scale factors for parameters
3) Table of regression controls
4) Iteration Iine
5) Panameter, Intemal variable table
6) Search table
7) - Termination explanation

Items 4 through 6 are repeated for each iteration when IPRNT=2.
Items described in the Iteration Line are as follows:
ITER NO Iteration number, beginning with one
INACTIVE Number of "null effect" parameters for this iteration (see Section I.6.)

DIST This is the distance of the desired first try along the search vector, in the scaled parameter space.

FSD

SIGMA Current value of the sum of squares
Under PARAMETERS ane three columns giving respectively the parameter number, the current value and the parameter increment corresponding to the desired first try along the search vector.

Under INTERNAL VARIABLES are four columns, all referring to the transformed coordinates. They are respectively the diagonalized Gauss-Newton
matrix, elements, $d_{i i}$, first partial derivatives of the sum of squares, coordinate increment for first try along the search vector and the MODE of incrementing the transformed coordinates. These MODES are as follows:

0 - Gauss-Newton
1 - Steepest Descent
-2 - Null Effect
The one-dimensional search table under SRCHS contains two columns which contain

FX - The factor multiplied by the search vector increment to obtain a trial point. The last one printed is the optimum one.

ACTION - Hollerith information which describes the progress of the search

SIGMA - The value of the sum of squares for the corresponding value of FX

The termination explanation is self-explanatory.
II. 4. Storage Requirements. The large bulk of the required storage for the D.D. program is taken up by the arrays contained in the DIMENSION statement. If there are $N$ equations and $N$ unknowns, this requirement calls for $15 \mathrm{~N}+3 \mathrm{~N}^{2}$ words of storage. For a 200 variable problem, this would require about 43,000 words. Thus, this size problem should be well within the storage limits of the Univac 1108 computer.
II. 5. Special Features. There are two special features of the D.D. prognam which deserve mention. These are an automatic restart procedure and the ability to treat sets of underdetermined on overdetermined nonlinear equations.

The automatic restart procedure operates as follows:
When $\operatorname{KCOM}(5)=0$, the problem has been solved satisfactorily only if $\operatorname{KCOM}(8)=0$. If this is not the case, then a retum to the call of CONST and DDSET causes a new (random) starting point to be generated within the feasible space (i.e., $X M A X(I) \geq X(I) \geq X M N(I)$, $I=1,2, \ldots$, NVAR). Note, therefore, that XMIN and XMAX must be specified even if the $K \emptyset D E=0$ option were used on the initial start of the problem. This restart procedure will be generated a maximum of five times. The restart procedure can be suppressed by deleting the test of $\operatorname{KCOM}(8)$. That is, at statement 100 , the problem is complete.

In order to solve sets of underdetermined equations (i.e., more variables than equations) or overdetemined equations (i.e., fewer variables than equations), it is necessary only to provide the proper values for NVAR and NCASE. In the case of overdetemined sets, a compromise solution (i.e.s, a minimum sum of squares solution) is all that can be expected. Therefore, the automatic restart procedure should be suppressed as described above.

## SAMPLE PROBLEM

A sample problem is included to illustrate further the programming requirements and usage. In particular, the system of linear algebraic equations

$$
\begin{aligned}
2 x_{1}-x_{2} & =1 \\
x_{1}+x_{2} & =I
\end{aligned}
$$

are solved from the initial guess $x_{1}=1 / 2, x_{2}=1 / 2$. The answers are $x_{I}=2 / 3, x_{2}=1 / 3$. The solution was not reached in one iteration because the required step size exceeded the limits normally provided and WSD was used.

A listing of the main program and printed output are given below.



LISTING OF FORTRAN IV COMPUTER PROGRAMS



1. SUBROUTINE TDVEC (N, NRES, SIGHA, DQ, ND2Q, 'D2Q, FSD, FSCAL,




1 SUBROUTINE CONRE ( $N$; FOSCR , FIT, SIGMA. NCASE. INACT; NOEX; NSD; 11 LTER, MODE, DX, NDER2, DER2, IEXII)












CALL MATMPY(1,1,NVAR, $1,1,1, D X B A R, D X B A R, D I S T)$

: 325 PRINT 1325, (ACTION(I), I $=2,7)_{2} Q$

| $\begin{aligned} & 1325 \text { FORMAT }(5 \times 16 A 4,1 P E 19.5) \\ & 326 \text { : } F \text { FICON) } 122,122,331 \end{aligned}$ | $\ldots$ | 1. |  |
| :---: | :---: | :---: | :---: |
| O 331 D0 $332 \mathrm{I}=1, \mathrm{NVAR}$ | - | ! | ! |
| 332 X $\begin{aligned} \text { X }\end{aligned}$ | $\cdots$ | $\cdots$ | - ! |
| C GO T0 998 | + | 1 | $\cdots$ |
| $\begin{array}{r} \text { CALL MATMPY }(1,1, \text { NCASE:-NCASE, } \\ Q Q=Q-Q Q \end{array}$ | $\text { NCASE }, 1, Y, Y, Q)$ | $i$ | ; : |
| $\therefore \quad \therefore \quad \mathrm{GO} \quad 10317$ | + | ; | ! |
| 122 IF (NSD) 124, I24,123 <br> 123 CALL NEWFSD (FX,Q,FSD) | $\cdots$ | ! | $\stackrel{\square}{3}$ |
| $\begin{aligned} & 124 \\ & \therefore 125 \\ & \end{aligned}$ | : | ; | - |
| CALL MOVE (NSQU,G;SX) |  | ; |  |

CALL RESCAL (NVAR,O,D2Q,DSCAL,SCALR)
XVAR $=$ NVAR
$X S D=N S D+1$.
IF (FXX * DÏST - FDERV * SQRT(XVAR) / XSD): $126,126,127$
126 KICK $=$ KICK + 1
LKICK ${ }^{--}=1$
127 _ Continue
ITER = ITER $+\overline{1}$
GO TO 41



|  | SUBROUTINE NENSCG(X.SCALE) <br> DIMENSION X(1),SCALE(1) |
| :---: | :---: |
|  | COMMON TTUCOM/ COM( 9 ), KCOM |
|  | INEW $=\operatorname{KCOM}(6)$ |
|  | NVAR $=\operatorname{KCOM}(2)$ <br> IF.(INEW) $8,3,8$ |
| 3 | $\begin{aligned} & \text { DO } 7 \mathrm{I}=\mathrm{NVAR} \\ & \text { IF }(\mathrm{ABS}(X(I))-1 . E-08) 60,6,5 \end{aligned}$ |
| 5 | $\begin{aligned} & 5 \quad \begin{array}{l} \text { SCALE(I) } \\ \text { GOTO } 7 \end{array} .=\operatorname{ABS}(X(\bar{I})) \\ & \end{aligned}$ |
|  | 6. $\mathrm{SCALE}(\mathrm{T}) \equiv 1 . \mathrm{E}-08$ |
| 7 | 7. CONTINUE |
| 8 | 8 RETURN |
|  | $\therefore$ END |















$3^{\circ}$ CONTINUE
$0043=1, N$
$J L=\operatorname{LOC}(J, L)$
$4 \mathrm{~A}(\mathrm{JL})=0.0$
IF(L.GE. NN). GO TO 8
$L=L+1$
GO 101

$M M=\operatorname{LOC}(M, M)$

$12 \mathrm{NSW}=\mathrm{NSW}+1$
ZERO DIAGONAL FOUND UPSTREAM--NON-ZERO DIAGONAL FOUND DOWNSTREAM SWAP OF ROW-COLUMN TÓ GET NON-ZERO DEAGONAL UPSTREAM
$13005 \mathrm{~K}=\mathrm{L}, \mathrm{N}$
$K L=-\operatorname{LOC}(K, L)$
$K M=\operatorname{LOC}(K, M)$
$\mathrm{SCR}(K)=A(K L)$
$A(K L)=A(K M)$
$.5-A^{-}(K M)=\operatorname{SCR}(K)$
DO $49 \mathrm{~K}=1 \mathrm{~N}$
$K L=\operatorname{LOC}(K, L)$
$K M=\operatorname{LOC}(K, M)$
$\operatorname{SCR}(K)=T(K L)$
$T(K L)=T(K M)$
$49^{\circ} \mathrm{T}(K M)=\operatorname{SCR}(K)$
DO $6 \mathrm{~K}=\mathrm{L}, \mathrm{N}$
$L K=\operatorname{LOC}(L, K)$
$M K=L O C(M, K)$
SCR $(K)=A(L K)$
$A(L K)=A(M K)$
6 A(MK ) = SCR(K)
GO TO 11
20 II = $1-1$.
$25^{\prime \prime}$ II $=11+1$
IF ( II .LE. N) GO TO 26
$C$ ALI RENAINING: DIAGONALS ARE-ZERO
C ALSO ALL OFF DIAGGNAL ELEMENTS OF ROW ARE ZERO
C ON DAGONAL AND GO TO NEXT RON TO SEARCH FOR NON-ZERO OFF-DIAGONAL ELEMENT
$L=L+1$
$I=L+1$
IF I GT. N GO TO-8
GO TO 20
$26 \mathrm{LI}=\operatorname{LOC}(\mathrm{L} I \mathrm{I})$
IF( ABS (A(LII )). LT. TF) GO TO 25
ALL REMAINING DIAGONALS ARE ZERO BUT OFF DIAGONAL FOUND TO BE NON-ZERO-=- ROTATION PERFORMED

[^0]



[^0]:    CO $42 \mathrm{~K}=\mathrm{L}, \mathrm{N}$
    KII $=\operatorname{LOC}(K, I I)$
    $K L=\operatorname{LOC}\left(K_{r} L\right)$

