# A PARALLEL VARIABLE METRIC
# OPTIMIZATION ALGORITHM

*by Terry A. Straeter*

*Langley Research Center*
*Hampton, Va. 23665*

| 1. Report No.<br>NASA TN D-7329 | 2. Government Accession No. | 3. Recipient's Catalog No. |
|---|---|---|
| 4. Title and Subtitle<br><br>A PARALLEL VARIABLE METRIC OPTIMIZATION<br>ALGORITHM | | 5. Report Date<br>December 1973 |
| | | 6. Performing Organization Code |
| 7. Author(s)<br><br>Terry A. Straeter | | 8. Performing Organization Report No.<br>L-8986 |
| | | 10. Work Unit No.<br>501-06-01-09 |
| 9. Performing Organization Name and Address<br><br>NASA Langley Research Center<br>Hampton, Va. 23665 | | 11. Contract or Grant No. |
| | | 13. Type of Report and Period Covered<br>Technical Note |
| 12. Sponsoring Agency Name and Address<br><br>National Aeronautics and Space Administration<br>Washington, D.C. 20546 | | 14. Sponsoring Agency Code |

15. Supplementary Notes

16. Abstract

    An algorithm is introduced which is designed to exploit the parallel computing or vector streaming (pipeline) capabilities of computers with such advanced features. If $p$ is the degree of parallelism, then one cycle of the parallel variable metric algorithm is defined as follows: first, the function and its gradient are computed in parallel at $p$ different values of the independent variable; then the metric is modified by $p$ rank-one corrections; and finally, a single univariant minimization is carried out in the Newton-like direction. Several properties of this algorithm are established in the paper. In addition, the convergence of the iterates to the solution is proved for a quadratic functional on a real separable Hilbert space; in fact, for a finite-dimensional space the convergence is in one cycle when $p$ equals the dimension of the space. Results of numerical experiments indicate that the new algorithm will exploit parallel or pipeline computing capabilities to effect faster convergence than serial techniques currently in use. In fact, the experiments indicate that even when the computations are done serially, the new algorithm is very competitive with the widely used Davidon-Fletcher-Powell technique.

| 17. Key Words (Suggested by Author(s))<br>Numerical analysis<br>Optimization<br>Parallel computers | | 18. Distribution Statement<br>Unclassified – Unlimited | | |
|---|---|---|---|---|
| 19. Security Classif. (of this report)<br>Unclassified | 20. Security Classif. (of this page)<br>Unclassified | | 21. No. of Pages<br>21 | 22. Price*<br>Domestic, $2.75<br>Foreign, $5.25 |

*For sale by the National Technical Information Service, Springfield, Virginia 22151

# A PARALLEL VARIABLE METRIC OPTIMIZATION ALGORITHM

By Terry A. Straeter
Langley Research Center

## SUMMARY

An algorithm is introduced which is designed to exploit the parallel computing or vector streaming (pipeline) capabilities of computers with such advanced features. If $p$ is the degree of parallelism, then one cycle of the parallel variable metric algorithm is defined as follows: first, the function and its gradient are computed in parallel at $p$ different values of the independent variable; then the metric is modified by $p$ rank-one corrections; and finally, a single univariant minimization is carried out in the Newton-like direction. Several properties of this algorithm are established in the paper. In addition, the convergence of the iterates to the solution is proved for a quadratic functional on a real separable Hilbert space; in fact, for a finite-dimensional space the convergence is in one cycle when $p$ equals the dimension of the space. Results of numerical experiments indicate that the new algorithm will exploit parallel or pipeline computing capabilities to effect faster convergence than serial techniques currently in use. In fact, the experiments indicate that even when the computations are done serially, the new algorithm is very competitive with the widely used Davidon-Fletcher-Powell technique.

## INTRODUCTION

In order to solve optimization problems efficiently by using computers with parallel computing or vector streaming (pipeline) capabilities, it is necessary either to develop new algorithms or to induce a high degree of parallelism into existing techniques. It may be that for these new computers many currently fashionable and serially efficient algorithms will be replaced by new parallel techniques. For simplicity, hereafter the term parallel is used to mean both parallel and vector streaming operations.

For a number of years it has been very popular and quite efficient to solve optimization problems on serial computers by the conjugate gradient (ref. 1) or variable metric (refs. 2 and 3) methods. But, as Miranker points out in his survey article (ref. 4) discussing parallelism in numerical analysis, these methods are inherently serial as each new search direction requires the result of the previous search. Hence, most researchers have concentrated their efforts on developing parallel univariant minimization techniques

(see Avriel and Wilde (ref. 5) and Karp and Miranker (ref. 6)) and on modifying Powell's method (ref. 7) with Zangwill's modification (ref. 8). (See Chazan and Miranker (ref. 9).)

The purpose of this paper is to introduce the parallel variable metric (PVM) algorithm, a new technique with a high degree of parallelism for use on the new computers. The basis of the algorithm is an observation by Powell (ref. 10, p. 93) on rank-one variable metric algorithms (refs. 11, 12, and 13) concerning the form of the rank-one update and some work by the author on an early version of the algorithm (ref. 14). If $p$ is the degree of parallelism, then one cycle of the parallel variable metric algorithm is defined as follows: first, the function and its gradient are computed in parallel at $p$ distinct values of the independent variable; then the metric $\left( V^{(n)} \text{ as defined later} \right)$ is modified by $p$ rank-one corrections; and finally, a single univariant minimization is carried out in the Newton-like direction. Herein the following results for the PVM algorithm are established: (1) If the function to be minimized is quadratic, defined on a finite-dimensional space, then the iterates converge to the location of the minimum in one cycle. (2) For strictly convex functions on a finite-dimensional space, convergence of the iterates to the location of the minimum follows if the metrics are uniformly bounded. (3) Convergence of the iterates to the minimum is established for the problem of minimizing a strongly positive quadratic functional on a real separable Hilbert space. (4) Finally, the results of numerical experiments involving the application of the parallel variable metric algorithm to sample problems are included. These results are compared with other investigators' results obtained by using a sequential or serial technique (Davidon-Fletcher-Powell) on the same problems. These results indicate that the new algorithm will exploit parallel computing capabilities to effect faster convergence (in terms of total time required to solve the problem) than serial techniques currently in use. In fact, those results indicate that for some problems − even when the PVM computations are done in a serial fashion − the PVM method is competitive, in terms of the number of function and gradient evaluations required to locate the minimum, with the Davidon-Fletcher-Powell (DFP) method, a widely used serial method.

## SYMBOLS

| | |
|---|---|
| $A$ | self-adjoint, positive, linear operator from $H$ into $H$ |
| $A^{-1}$ | inverse of $A$ |
| $B^{(n)}, V^{(n)}, V_i^{(n)}$ | sequence of linear operators |
| $b$ | fixed element of $H$ |

| | |
|---|---|
| C,V | linear operator |
| g | gradient of J |
| $g_i$ | gradient J at $x_i$ |
| H | real Hilbert space |
| I | identity operator |
| i,j,k,q | integers |
| J | functional defined on H |
| $J^O, M, m, M_O, m_O$ | real numbers |
| n | iteration number |
| p | degree of parallelism |
| R | set of all real numbers |
| $r_n$ | nth residual vector, element of H |
| $s_n$ | direction of nth step, element of H |
| u,z | elements of H |
| $\tilde{x}$ | element of H at which J is minimized |
| $x^n$ | nth iterate, element of H |
| $x, \bar{x}, x_i$ | element of H |
| $y_i$ | element of H defined by $g_i - g(x^n)$ |
| $\alpha, \beta, \mu$ | scalars |

| | |
|---|---|
| $\alpha_n$ | step size, a real number |
| $\sigma_j$ | basis element of $H$ |
| $\tau_j$ | scalars defined by equations (2) |
| $(\cdot,\cdot)$ | inner product on $H$ |
| $\|\cdot\|$ | norm on $H$ defined as $(\cdot,\cdot)^{1/2}$ |

Abbreviations:

| | |
|---|---|
| DFP | Davidon-Fletcher-Powell |
| dim H | dimension of space $H$ |
| PVM | parallel variable metric algorithm |
| SPQF | strongly positive quadratic functional |

## PARALLEL VARIABLE METRIC (PVM) ALGORITHM

Consider the problem of finding the element $\tilde{x} \in H$, a real separable Hilbert space with inner product $(\cdot,\cdot)$ which minimizes a differentiable function $J:H \rightarrow R$ with gradient $g$. Let $x^1 \in H$ be the initial estimate of the location of the minimum of $J$ and let $V^{(0)}$ be a self-adjoint linear operator from $H$ to $H$. Moreover, let $M_0 \geq m_0 > 0$ be such that for all $x \in H$, $m_0(x,x) \leq \left(x,V^{(0)}x\right) \leq M_0(x,x)$; that is, $V^{(0)}$ is strongly positive. If $J$ happens to be a quadratic functional $\left(\text{i.e.,} \quad J(x) = J^0 + (b,x) + \frac{1}{2}(x,Ax)\right.$ with $b \in H, J^0 \in R$ and $A$ strongly positive$\Big)$ then $V^{(0)}$ can be interpreted as an initial estimate of $A^{-1}$. Further, let $p$ be a positive integer which will be called the degree of parallelism. If $H$ is finite dimensional, it is advantageous to let $p = \dim H$. Let $\Sigma = \{\sigma_j\} \subseteq H$ represent a Schauder basis for $H$ (ref. 15). If $\dim H = p < \infty$, then let $\sigma_{p+1} = \sigma_1$ and $\sigma_{p+2} = \sigma_2$ and so on. That is, for a finite-dimensional space the $\sigma_j$ vectors represent a recycling through the $p$ basis vectors. The quantities $J(x^1)$ and $g(x^1)$ are computed and the first ($n = 1$) and successive iterations are obtained as follows:

Step 1.  The function and its gradient are computed at  $p$  distinct values of the independent variable. For  $i = 1, 2, \ldots, p$  let  $j = (n - 1)p + i$  and

$$x_j = x^n + \sigma_j$$

and compute  $g_j = g(x_j)$  and  $J(x_j)$  in a parallel fashion at the  $p$  values for  $x_j$.  If  $(g_j, g_j) = 0$  then  $x_j$  satisfies the necessary condition for a minimum and computation is stopped provided  $J(x_j) \leq \{$All other computed values of  $J\}$.

Step 2.  The metric is modified by  $p$  rank-one corrections.  Compute the vectors

$$y_j = g_j - g(x^n)$$

for  $j = (n - 1)p + i$  with  $i = 1, 2, \ldots, p$.  Next the residual vectors must be computed. (The reason for this terminology is explained later.)

Define

$$r_{(n-1)p+1} = V^{(n-1)} y_{(n-1)p+1} - \sigma_{(n-1)p+1}$$

and

$$r_j = V^{(n-1)} y_j - \sigma_j - \sum_{k=1}^{i-1} \frac{\left(r_{(n-1)p+k}, y_j\right)}{\left(r_{(n-1)p+k}, y_{(n-1)p+k}\right)} r_{(n-1)p+k} \tag{1}$$

where  $j = (n - 1)p + i$  and  $i = 2, 3, \ldots, p$.  If the denominator in equation (1) is zero for any term, that term is deleted from the summation.

Compute the  $p$  scalars

$$\tau_j = \begin{cases} -\left(y_j, r_j\right)^{-1} & \left(\left(y_j, r_j\right) \neq 0\right) \\ 0 & \left(\left(y_j, r_j\right) = 0\right) \end{cases} \tag{2}$$

$$V^{(n)} = V^{(n-1)} + \sum_{j=(n-1)p+1}^{np} \tau_j B^{(j)} \tag{3}$$

where  $B^{(j)}: H \rightarrow H$  for  $j = (n - 1)p + i$  and  $i = 1, 2, \ldots, p$  is defined such that for all  $x \in H$

$$B^{(j)} x = \left(x, r_j\right) r_j \tag{4}$$

5

Step 3. A single univariant minimization is carried out in the Newton-like direction of search. Let

$$s_n = -V^{(n)}g(x^n) \tag{5}$$

where $s_n$ is called the direction of search. The step size $\alpha_n$ (a scalar) is computed and the next iterate is defined by $x^{n+1} = x^n + \alpha_n s_n$ and then $J(x^{n+1})$ and $g(x^{n+1})$ are computed. The step size herein is chosen by means of a one-dimensional minimization. This could be done by a number of techniques (refs. 2, 5, and 16). If $\|g(x^{n+1})\|$ is sufficiently small, then stop; otherwise, let $n = n + 1$ and go to step 1.

Each pass through the algorithm (steps 1 to 3) is called a cycle. Notice that steps 1 and 2 each entail a high degree of parallelism. Step 3 does not directly involve any parallel computation of $J$ or its gradient. However, in the one-dimensional minimization of step 3, any parallel or pipeline structure of the computer can be exploited (ref. 5). For most optimization problems the time required for the evaluation of the function and gradient is much greater than that required for all the other calculations of the algorithm. It is for this reason that optimization algorithms are judged by the total number of function and gradient evaluations required to solve the problem. The parallel computations of $J$ and $g$ called for in step 1 are the main time-saving element of the algorithm, not the other parallel computations of $y_j, x_j$ and, to a lesser degree, $r_j$.

Figure 1 gives a two-dimensional illustration of the progress of the algorithm with $p = 2$. The figure depicts the level curves of the function $J$, $x^1$ the initial estimate of the location of the minimum, and $g(x^1)$ the gradient of $J$ at $x^1$. Also shown are $x_1$ and $x_2$, as defined in step 1 by $x_1 = x^1 + \sigma_1$ and $x_2 = x^1 + \sigma_2$. At $x_1$ and $x_2$ the vectors of $g(x_1)$ and $g(x_2)$ are shown. With this information, $r_1$, $r_2$, and $V^{(1)}$ can be defined. This then defines the search direction $s_1 = -V^{(1)}g(x^1)$. Finally, the point $x^2$ $\left(\text{found by minimizing } J(x_1 + \alpha s_1) \text{ with respect to } \alpha\right)$ is shown.

## PROPERTIES OF THE $V^{(n)}$ OPERATORS

For any real separable Hilbert space $H$ and differentiable functional $J$, the following theorems can be established:

Theorem 1: $B^{(j)}$, as defined by equation (4), is a self-adjoint positive operator $\left(\text{i.e., for all } x \in H \left(x, B^{(j)}x\right) \geq 0\right)$.

Proof: Clearly $B^{(j)}$ is linear by the linearity of the inner product. If $x \in H$, then $\left(x, B^{(j)}x\right) = \left(x, \left(x, r_j\right)r_j\right) = \left(x, r_j\right)^2 \geq 0$.
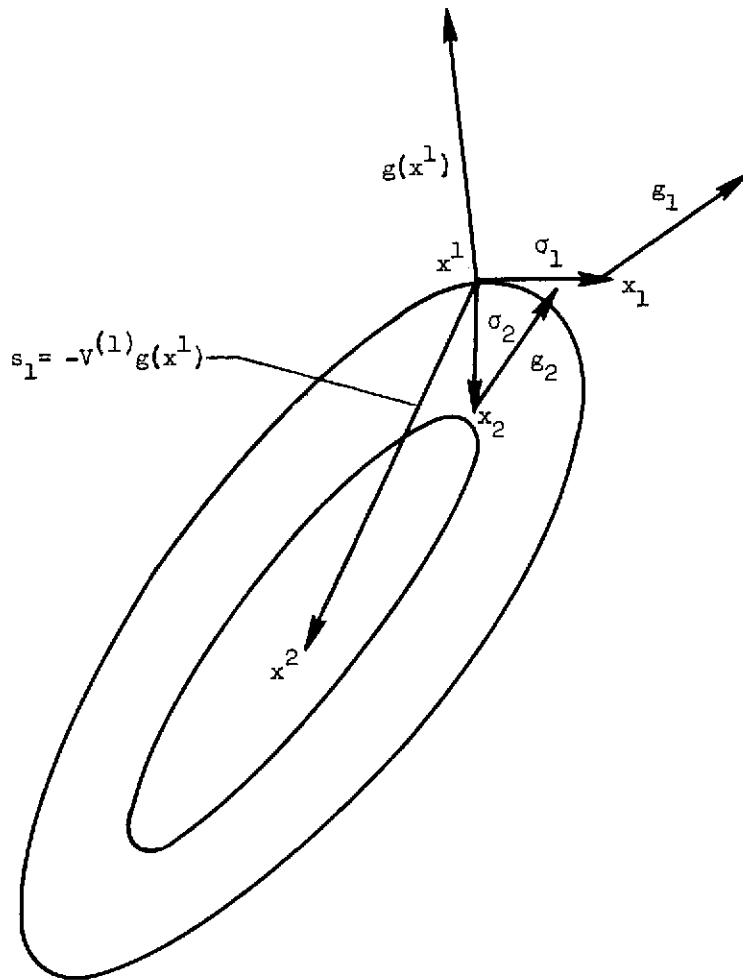
Figure 1.– PVM method progress.

If $x, y \in H$, then

$$\left(x, B^{(j)} y\right) = \left(x, (y, r_j) r_j\right) = (y, r_j)(x, r_j) = \left(y, (x, r_j) r_j\right) = \left(y, B^{(j)} x\right)$$

Theorem 2: $V^{(n)}$ is self adjoint for all $n$.

Proof: By definition, $V^{(0)}$ is self adjoint; by theorem 1, the $B^{(i)}$ operators are self adjoint; and by equation (3), $V^{(n)}$ is a finite linear combination of $V^{(0)}$ and the $B^{(i)}$ operators.

To facilitate the proof of later results, define an additional $p - 1$ linear operators from $H$ to $H$ for each cycle $n$ as follows:

$$V_0^{(n-1)} = V^{(n-1)}$$

and

$$V_i^{(n-1)} = V_0^{(n-1)} + \sum_{j=(n-1)p+1}^{(n-1)p+i} \tau_j B^{(j)} \qquad (i = 1, 2, \ldots, p) \qquad (6)$$

Clearly then

$$V^{(n)} = V_p^{(n-1)}$$

where $\tau_j$ and $B^{(j)}$ are as in equations (2) and (4). Also, it is clear that $V_i^{(n-1)}$ is given by

$$V_i^{(n-1)} = V_{i-1}^{(n-1)} + \tau_{(n-1)p+i} B^{((n-1)p+i)} \qquad (7)$$

and equation (1) can be rewritten as

$$r_j = V_{i-1}^{(n-1)} y_j - \sigma_j \qquad (8)$$

where $j = (n - 1)p + i$ and $i = 1, 2, \ldots, p$. Because of the definitions of $V^{(n)}$, $y_j$, and $\sigma_j$, theorem 3 follows.

Theorem 3: If $\tau_j \neq 0$ or $r_j = 0$, then $V_i^{(n-1)} y_j = \sigma_j$ for $j = (n - 1)p + i$ for each $n$, with $i = 1, 2, \ldots, p$.

Proof: Let $i$ be a positive integer between 1 and $p$ and $n$ be some positive integer. Then, if $r_j = 0$, $\tau_j = 0$ by definition of $\tau_j$ so that $V_i^{(n-1)} = V_{i-1}^{(n-1)}$ and the theorem is true by equation (8). Otherwise, if $\tau_j \neq 0$

$$V_i^{(n-1)} y_j = V_{i-1}^{(n-1)} y_j + \tau_j B^{(j)} y_j \qquad \text{(by eq. (7))}$$

$$= V_{i-1}^{(n-1)} y_j - \left(r_j, y_j\right)^{-1} \left(r_j, y_j\right) r_j \qquad \text{(by eq. (4))}$$

$$= V_{i-1}^{(n-1)} y_j - \left(V_{i-1}^{(n-1)} y_j - \sigma_j\right) \qquad \text{(by eq. (8))}$$

$$= \sigma_j$$

8

At this point it is advantageous to define a strongly positive quadratic functional (SPQF). A functional of the form

$$J(x) = J^O + (b,x) + \frac{1}{2}(x,Ax) \tag{9}$$

is an SPQF when $J^O \in R, b \in H$, and $A$ is a strongly positive, self-adjoint, linear operator from $H$ into $H$ $\left(\text{i.e., there exist } m, M > 0 \text{ such that } m(x,x) \leqq (x,Ax) \leqq M(x,x)\right)$. It is well known that the location of the minimum of $J$, denoted by $\tilde{x}$, exists and is the equal to $-A^{-1}b$. Also, the gradient of an SPQF is $g(x) = Ax + b$ (ref. 17). Another useful result for an SPQF is that if $\bar{x} = x + \sigma$ and $y = g(\bar{x}) - g(x)$, then

$$A^{-1}y = \sigma \tag{10}$$

Using equation (10) in equation (8) gives $r_j = V_{i-1}^{(n-1)}y_j - A^{-1}y_j$ where $j = (n-1)p + i$. This form of $r_j$ clearly shows the reason for calling $r_j$ the residual vector; that is, $r_j$ is the error at $y_j$ in the approximation of $A^{-1}$ by $V_{i-1}^{(n-1)}$. With these fundamental relations for an SPQF theorem 4 can now be established.

Theorem 4: If $J$ is an SPQF and $u \in H$ such that $A^{-1}u = V_{i-1}^{(n-1)}u$, and $C:H \rightarrow H$ is a linear operator such that $C = V_{i-1}^{(n-1)} + \mu B^{(j)}$ for some real $\mu$, for $j = (n-1)p + i$, then $A^{-1}u = Cu$.

Proof: Since $A^{-1}y_j = \sigma_j$ and $x_j = x^n + \sigma_j$ then

$$r_j = V_{i-1}^{(n-1)}y_j - \sigma_j = V_{i-1}^{(n-1)}y_j - A^{-1}y_j$$

$$r_j = \left(V_{i-1}^{(n-1)} - A^{-1}\right)y_j$$

Since

$$\left(V_{i-1}^{(n-1)} - A^{-1}\right)u = 0$$

then

$$\left(r_j,u\right) = \left(\left(V_{i-1}^{(n-1)} - A^{-1}\right)y_j,u\right)$$

$$= \left(y_j,\left(V_{i-1}^{(n-1)} - A^{-1}\right)u\right)$$

$$= \left(y_j,0\right) = 0$$

By hypothesis

$$\left(C - A^{-1}\right)u = \mu B^{(j)}u$$

$$= \mu\left(r_j, u\right)r_j$$

$$= \mu \cdot 0 \cdot r_j$$

$$= 0$$

Since $V_i^{(n-1)} = V_{i-1}^{(n-1)} + \tau_j B^{(j)}$ the following corollaries can be obtained.

Corollary 1: If $V_{i-1}^{(n-1)}u = A^{-1}u$ for some $u \in H$ and $J$ is an SPQF, then $V_i^{(n-1)}u = A^{-1}u$.

Corollary 2 (fundamental property of $V^{(n-1)}$): If $J$ is an SPQF and $\tau_q \neq 0$ for all $q \leq (n-1)p + i$, then $V_i^{(n-1)}y_k = A^{-1}y_k = \sigma_k$ for all $k \leq (n-1)p + i$.

Proof: For any fixed but arbitrary positive integer $n$, recall from theorem 3 that

$$V_1^{(n-1)}y_{(n-1)p+1} = \sigma_{(n-1)p+1}$$

and

$$V_{i-1}^{(n-1)}y_k = \sigma_k \qquad\qquad \left(k = (n-1)p + i - 1\right)$$

Assume $V_{i-1}^{(n-1)}y_k = \sigma_k$ for all $k \leq (n-1)p + i - 1$. Since $\tau_k \neq 0$, then for $k = (n-1)p + i$, $V_i^{(n-1)}y_k = \sigma_k$ by theorem 3. By corollary 1 of theorem 4 and the inductive hypothesis, $V_i^{(n-1)}y_k = \sigma_k$ for all $k \leq (n-1)p + i$. Since $n$ was fixed but arbitrary, the corollary is established. Corollary 2 is most useful in later convergence arguments and, hence, it is called the fundamental property of $V_i^{(n-1)}$.

Although at this point several convergence results could be established, first other properties of the sequence of operators $V_i^{(n)}$ will be derived. For the remainder of this section it is assumed that $J$ is an SPQF. In that case Goldfarb (ref. 18) has observed that $\tau_j$ for $j = (n-1)p + i$ can be expressed as

$$\tau_j = -\left(\left(V_{i-1}^{(n-1)} - A^{-1}\right)y_j, y_j\right)^{-1} \qquad\qquad \left(\tau_j \neq 0\right)$$

10

Hence equation (7) can be written as

$$V_i^{(n-1)} = V_{i-1}^{(n-1)} - \frac{B^{(j)}}{\left(\left(V_{i-1}^{(n-1)} - A^{-1}\right)y_j, y_j\right)} \qquad (11)$$

or if $\tau_j = 0$, $V_i^{(n-1)} = V_{i-1}^{(n-1)}$.

The preceding observation yields the following theorem which is proved by induction using the Cauchy-Schwarz inequality and equation (11). A similar theorem and proof for the serial rank-one algorithm is given in reference 19.

Theorem 5: If $V^{(0)} \geq A^{-1}$ $\left(V^{(0)} \leq A^{-1}\right)$ then $V_i^{(n)} \geq A^{-1}$ $\left(V_i^{(n)} \leq A^{-1}\right)$ for all $n$ and $i = 1, 2, \ldots, p.$ $\left(V^{(0)} \geq A^{-1}$ means $\left(x, V^{(0)}x\right) \geq \left(x, A^{-1}x\right)$ for all $x \in H.\right)$

Again, as in reference 19, by using equation (11), the Cauchy-Schwarz inequality, and theorem 1 a condition can be established under which the $V_i^{(n)}$ and hence the $V^{(n)}$ operators form a monotone sequence of self-adjoint bounded operators.

Theorem 6: If $V^{(0)} \geq A^{-1}$ $\left(V^{(0)} \leq A^{-1}\right)$ then $A^{-1} \leq V_i^{(n)} \leq V_{i-1}^{(n)} \leq \ldots \leq V_0^{(n)} = V^{(n)} = V_p^{(n-1)} \leq \ldots \leq V^{(0)}$ for all $n$ and $i$ $\left(V^{(0)} = V_0^{(0)} \leq V_1^{(0)} \leq \ldots \leq V_i^{(n)} \leq A^{-1}\right).$

Corollary: If $V^{(0)} \leq A^{-1}$ or $V^{(0)} \geq A^{-1}$ then the $V^{(n)}$ operators form a monotone sequence of strongly positive, self-adjoint linear operators bounded by $V^{(0)}$ and $A^{-1}$. Moreover, there exists a strongly positive self-adjoint operator $V$ such that $\lim_{n \to \infty} \left\| V^{(n)}x - Vx \right\| = 0$ for all $x \in H$.

Proof: By the hypothesis on $V^{(0)}$, theorem 6, and theorem 2, it is well known (ref. 20, p. 189) that the conclusion to the theorem follows.

If $\dim H$ is finite and $p = \dim H$, theorem 7 can be proved.

Theorem 7: If $\tau_i \neq 0$ for $i = 1, 2, \ldots, p$ and $\dim H = p$, then $V^{(1)} = A^{-1}$.

Proof: Since $A$ is strongly positive, for each $x \in H$ there exists $z \in H$ such that $x = Az$; hence, $z = A^{-1}x$, and since the $\sigma$ vectors form a basis, there exist scalars $\beta_i$ such that

$$z = \sum_{i=1}^{p} \beta_i \sigma_i$$

Thus

$$x = A \sum_{i=1}^{p} \beta_i \sigma_i = \sum_{i=1}^{p} \beta_i A \sigma_i = \sum_{i=1}^{p} \beta_i y_i$$

Now by the fundamental property of $V_p^{(0)}$

$$V_p^{(0)} y_i = \sigma_i \qquad\qquad (i = 1, 2, \ldots, p)$$

and

$$V_p^{(0)} = V^{(1)}$$

hence

$$V^{(1)} x = \sum_{i=1}^{p} \beta_i V^{(1)} y_i = \sum_{i=1}^{p} \beta_i \sigma_i = z$$

and also

$$A^{-1} x = z$$

Hence

$$A^{-1} x = V^{(1)} x$$

for each $x \in H$.

If dim $H$ is not finite then it can be shown that if $\tau_j \neq 0$ for all $j$, $V^{(n)} \to A^{-1}$ pointwise as in the following theorem.

Theorem 8: If $\tau_j \neq 0$ for $j = 1, 2, \ldots$ and the $V^{(n)}$ operators are uniformly bounded, then $\lim_{n \to \infty} \left\| V^{(n)} x - A^{-1} x \right\| = 0$.

Proof: Let $x \in H$; then there exists $z \in H$ such that $x = Az$ $\left(\text{i.e., } z = A^{-1} x\right)$ and $z = \sum_{i=1}^{\infty} \beta_i \sigma_i$ since the $\sigma$ vectors form a basis for $H$.

Now, since $A$ is bounded,

$$x = A \sum_{i=1}^{\infty} \beta_i \sigma_i = \sum_{i=1}^{\infty} \beta_i y_i$$

Hence

$$v^{(n)}x = v^{(n)} \sum_{j=i}^{\infty} \beta_j y_j = \sum_{j=1}^{np} \beta_j v^{(n)} y_j + v^{(n)} \sum_{j=np+1}^{\infty} \beta_j y_j$$

$$= \sum_{j=1}^{np} \beta_j \sigma_j + v^{(n)} \sum_{j=np+1}^{\infty} \beta_j y_j$$

Therefore

$$\left\| v^{(n)}x - A^{-1}x \right\| = \left\| (v^{(n)} - A^{-1}) \sum_{j=np+1}^{\infty} \beta_j y_j \right\| \leq \left\| v^{(n)} - A^{-1} \right\| \left\| \sum_{j=np+1}^{\infty} \beta_j y_j \right\|$$

Because the $v^{(n)}$ operators are uniformly bounded, $\left\| v^{(n)} - A^{-1} \right\|$ is bounded, and
since $\sum_{i=1}^{\infty} \beta_i y_i$ converges, then $\sum_{i=np+1}^{\infty} \beta_i y_i \to 0$ as $n \to \infty$; therefore, the right-hand
side $\to 0$ as $n \to \infty$.

Corollary: If $v^{(0)} \leq A^{-1}$ or $v^{(0)} \geq A^{-1}$ and $\tau_j \neq 0$ for all $j$, then
$$\lim_{n \to \infty} \left\| v^{(n)}x - A^{-1}x \right\| = 0.$$

## CONVERGENCE RESULTS

By utilizing the previous results, the following convergence theorems can now be
established.

Theorem 9: If $J$ is an SPQF, $\tau_i \neq 0$ for $i = 1, 2, \ldots, p$ where $\dim H = p < \infty$,
then the algorithm converges to the location of the minimum of $J$ in one cycle.

Proof: At step 3 $s_1 = -v^{(1)}g(x^1)$ and by theorem 7 $v^{(1)} = A^{-1}$ so

$$s_1 = -A^{-1}g(x^1) = -A^{-1}(Ax^1 + b) = -x^1 - A^{-1}b$$

Hence, $x^2 = x^1 + \alpha_1(-x^1 - A^{-1}b)$, and $\alpha_1 = 1$ is clearly the proper choice of $\alpha$; hence,
$x^2 = -A^{-1}b = \tilde{x}$.

Theorem 10: If $J$ is an SPQF, the $V^{(n)}$ operators are uniformly strongly positive — that is, there exist $\alpha, \beta > 0$ such that $\alpha I \leq V^{(n)} \leq \beta I$ — and $H$ is an infinite-dimensional Hilbert space, then the algorithm converges to the location of the minimum $J$.

Proof: It is sufficient to show that $g(x^n) \to 0$ as $n \to \infty$. Since the step sizes are chosen as a result of a one-dimensional minimization, it is well known (e.g., ref. 19) that necessarily

$$\lim_{n \to \infty} \frac{\left(s_n, g(x^n)\right)^2}{\left(s_n, As_n\right)} = 0$$

Hence, using the fact that $A$ is stongly positive and the $V^{(n)}$ operators are uniformly strongly positive gives

$$\frac{\left(s_n, g(x^n)\right)^2}{\left(s_n, As_n\right)} \geq \frac{\left(s_n, g(x^n)\right)^2}{M\|s_n\|^2} \geq \frac{\left(V^{(n)}g(x^n), g(x^n)\right)^2}{M\|V^{(n)}g(x^n)\|^2} \geq \frac{\alpha^2\|g(x^n)\|^4}{M\|V^{(n)}g(x^n)\|^2} \geq \frac{\alpha^2}{M\beta^2}\frac{\|g(x^n)\|^4}{\|g(x^n)\|^2} = \frac{\alpha^2}{M\beta^2}\|g(x^n)\|^2 \geq 0$$

Therefore $\|g(x^n)\|^2 \to 0$, so $g(x^n) \to 0$ and $x^n \to -A^{-1}b$.

Corollary: If $J$ is an SPQF and $V^{(0)} \leq A^{-1}$ or $V^{(0)} \geq A^{-1}$, then the algorithm converges to the location of the minimum.

Proof: By theorem 5 and the corollary to theorem 6 the $V^{(n)}$ operators are uniformly strongly positive.

Now consider strictly convex twice-differentiable functions on a finite-dimensional Hilbert space.

Theorem 11: Suppose $J$ is strictly convex with bounded second partial derivatives. That is, if

(a) $m(x,x) \leq (x, J''(x)x) \leq M(x,x)$ for all $x \in H$ where $\infty > M \geq m > 0$ and $J''$ denotes the second derivative of $J$ (i.e., Hessian of $J$)

(b) $J''(x)$ is nondecreasing along any path of nonincreasing function values

(c) For all $x_1, x_2 \in H$, and $\alpha \in (0,1)$, $J\left(\alpha x_1 + (1 - \alpha)x_2\right) < \alpha J(x_1) + (1 - \alpha)J(x_2)$.
Therefore, the iterates generated by the PVM algorithm converge to the global minimum of $J$ if the $V^{(n)}$ operators are uniformly positive definite $\left(\text{i.e.,}\right.$ there exist positive constants $\alpha, \beta$ such that $\alpha I \leq V^{(n)} \leq \beta I$ for $\left.n\right)$.

Proof: This theorem follows from theorem 1 in reference 18.

For a serial version of the rank-one algorithm with step size chosen by a one-dimensional minimization, Goldfarb (ref. 18) has shown that for a strictly convex function

14

on $R^n$ the $V$ operators are uniformly positive definite whenever $V^{(0)} \leqq J''(x^1)$ or $J''(x^1) \geqq V^{(0)}$. However, this result has not been extended to the parallel algorithm described herein.

## EXAMPLE PROBLEMS AND RESULTS

To illustrate the performance of the parallel variable metric minimization algorithm, numerical experiments were conducted on several standard example problems. Although the experiments were conducted on a serial computer, the results of the computations were used as if they had been done in a parallel fashion. Five sample functions are used for the numerical experiments.

A simple quadratic function of three variables

$$f_1(x,y,z) = x^2 - 2xy + 2y^2 + 5z^2$$

was the first function with $x^1 = (1,1,1)$. For this example $p$ was chosen to be 3 and, as theorem 9 indicated, convergence was achieved in one cycle.

The second function was the well-traveled Rosenbrock's parabolic valley function

$$f_2(x,y) = \left(x^2 - y\right)^2 + 0.01(x - 1)^2$$

This is a particularly difficult function to minimize because of the long parabolic valley $y = x^2$ along which the minimization must travel. The traditional starting point is $(-1.2, 1.0)$ and the minimum is located at $(1,1)$.

The third function, known as Powell's function, is

$$f_3(x_1,x_2,x_3,x_4) = \left(x_1 - 10x_2\right)^2 + 5\left(x_3 - x_4\right)^2 + \left(x_2 - 2x_3\right)^4 + 10\left(x_1 - x_4\right)^4$$

with starting values at $(3,-1,0,1)$. This function is particularly difficult for a variable metric algorithm to minimize because at the minimum the Hessian is singular.

The fourth test function is called the 4-D banana or Wood's function and is defined by

$$f_4(x_1,x_2,x_3,x_4) = 100\left(x_1^2 - x_2\right)^2 + \left(1 - x_1\right)^2 + 90\left(x_3^2 - x_4\right)^2 + \left(1 - x_3\right)^2$$
$$+ 10.1\left[\left(x_2 - 1\right)^2 + \left(x_4 - 1\right)^2\right] + 19.8\left(x_2 - 1\right)\left(x_4 - 1\right)$$

The traditional starting estimate is at (-3,-1,-3,-1). This function is difficult to minimize because the quadratics $x_1^2 - x_2$ and $x_3^2 - x_4$ make the level surfaces banana shaped.

The fifth function is the helical valley function defined by

$$f_5(x_1,x_2,x_3) = 100\left[(x_3 - 10\theta)^2 + \left(\sqrt{x_1^2 + x_2^2} - 1\right)^2\right] + x_3^2$$

where

$$2\pi\theta = \begin{cases} \tan^{-1}\dfrac{x_2}{x_1} & (x_1 > 0) \\[2em] \pi + \tan^{-1}\dfrac{x_2}{x_1} & (x_1 < 0) \end{cases}$$

The usual starting estimate is at (-1,0,0).

On a present-day serial computer, numerical experiments were conducted in which the PVM algorithm was applied to the five standard test functions. For these experiments the convergence criterion was the absolute value of the largest component of the gradient less than $\epsilon = 10^{-7}$. The basis vectors were defined as $\sigma_i = 10^3 \epsilon e_i$, where $e_i$ is the ith elementary vector. The one-dimensional minimization required in step 3 of the PVM algorithm was carried out by Davidon's cubic interpolation method with initial estimate of the step size $\lambda = \min\left(2., -2\left(J(x^i) - J_{min}\right)\big/\left(g(x^i),s_i\right)\right)$, where $J_{min}$ is the estimated minimum value of J. The degree of parallelism p was chosen to be the number of variables in the test function and $V^{(0)} = I$.

Table 1 gives the results of these numerical experiments. For each test function the number of cycles required to achieve convergence is listed. Also listed is the total number of function and gradient evaluations required for convergence on the serial computer. The third column of results in table 1 presents the situation as if the computations had been done on an advanced computer with stream or parallel computing capabilities. Then, the computation of the p gradients in step 1 would have been done by utilizing these capabilities. Therefore, ignoring overhead costs, these p gradients would take essentially the same time as the computation of one gradient. It is for this reason that in table 1 the number of evaluations for the parallel case is the same as for the serial case minus (p - 1) times the number of cycles. Columns four and five give measures of the accuracy of the minimization.

TABLE 1.- RESULTS OF PVM METHOD APPLIED TO

FIVE SAMPLE PROBLEMS

| Function | Number of cycles | Total number of function and gradient evaluations (serial) | Total number of function and gradient evaluations (parallel) | Function value at convergence | Largest componentwise error in solution at convergence |
|---|---|---|---|---|---|
| Quadratic | 1 | 7 | 5 | $1.5 \times 10^{-21}$ | $5 \times 10^{-11}$ |
| Rosenbrock | 13 | 58 | 45 | $7.7 \times 10^{-16}$ | $5 \times 10^{-7}$ |
| Powell | 19 | 118 | 61 | $4.6 \times 10^{-15}$ | $2 \times 10^{-4}$ |
| 4-D banana | 30 | 185 | 95 | $9.7 \times 10^{-18}$ | $2 \times 10^{-9}$ |
| Helical valley | 14 | 75 | 47 | $1.2 \times 10^{-20}$ | $6 \times 10^{-12}$ |

Table 2 shows the progress to the minimum of the PVM for the first three test functions. Also recorded in table 2 is the number of function and gradient evaluations taken with each cycle. Notice that the convergence to the minimum occurs in one cycle for the quadratic as predicted by theorem 9. Also notice the apparently superlinear rate of convergence for Rosenbrock's function. It is conjectured that as $\sigma_i$ is made small $V^{(n)} \to (J'')^{-1}$. This conjecture is based on the fundamental property of $V^{(n)}$. The slowing convergence on Powell's function is due to the singularity of $J''$ at the minimum.

TABLE 2.- PROGRESS PER CYCLE OF PVM ON THREE EXAMPLE PROBLEMS

| Cycle | Quadratic | | Rosenbrock's function | | Powell's function | |
|---|---|---|---|---|---|---|
| | Function value | Number of function evaluations | Function value | Number of function evaluations | Function value | Number of function evaluations |
| 0 | 6.0 | 1 | $4.8 \times 10^{-1}$ | 1 | $8.7 \times 10^{2}$ | 1 |
| 1 | $1.5 \times 10^{-21}$ | 6 | $4.7 \times 10^{-2}$ | 4 | $2.1 \times 10^{1}$ | 7 |
| 2 | | | $4.2 \times 10^{-2}$ | 5 | 2.1 | 6 |
| 3 | | | $1.9 \times 10^{-2}$ | 5 | $2.1 \times 10^{-1}$ | 7 |
| 4 | | | $1.6 \times 10^{-2}$ | 4 | $2.1 \times 10^{-2}$ | 6 |
| 5 | | | $6.0 \times 10^{-3}$ | 5 | $2.1 \times 10^{-3}$ | 6 |
| 6 | | | $3.9 \times 10^{-3}$ | 4 | $2.1 \times 10^{-4}$ | 6 |
| 7 | | | $1.8 \times 10^{-3}$ | 5 | $2.1 \times 10^{-5}$ | 6 |
| 8 | | | $6.4 \times 10^{-4}$ | 4 | $2.1 \times 10^{-6}$ | 6 |
| 9 | | | $2.6 \times 10^{-4}$ | 4 | $2.1 \times 10^{-7}$ | 6 |
| 10 | | | $2.3 \times 10^{-5}$ | 5 | $2.1 \times 10^{-8}$ | 6 |
| 11 | | | $4.7 \times 10^{-8}$ | 4 | $2.1 \times 10^{-9}$ | 6 |
| 12 | | | $2.8 \times 10^{-11}$ | 4 | $2.2 \times 10^{-10}$ | 6 |
| 13 | | | $7.7 \times 10^{-16}$ | 4 | $2.4 \times 10^{-11}$ | 6 |
| 14 | | | | | $3.1 \times 10^{-12}$ | 6 |
| 15 | | | | | $4.6 \times 10^{-13}$ | 6 |
| 16 | | | | | $8.6 \times 10^{-14}$ | 6 |
| 17 | | | | | $1.9 \times 10^{-14}$ | 6 |
| 18 | | | | | $1.8 \times 10^{-14}$ | 6 |
| 19 | | | | | $4.6 \times 10^{-15}$ | 7 |
| Total | | 7 | | 58 | | 118 |

Table 3 illustrates a comparison of the performance of the DFP algorithm with the PVM algorithm on the two example functions, $f_2$ and $f_3$. The DFP method was chosen as the standard of comparison because of its wide use and generally it has compared favorably with other techniques. The key factor in the performance of any minimization technique is the number of function and gradient evaluations required to converge, as the computation required for evaluating the function is usually much greater than that involved in the algorithm. Table 3 lists the number of function evaluations required by the DFP method to locate the minimum of $f_2$ and $f_3$ starting from the same initial conditions. These results have been obtained by other investigators (refs. 21 and 22). Also listed in table 3 are results on these problems reported by Jacobson and Oksman (ref. 23) for a DFP subroutine. Finally, in table 3 the performance of the parallel variable metric method is given for two cases. In the first case, the method is used on a serial computer, hence the operations in step 1 (i.e., gradient evaluations) are not done in parallel. For the second case, it is assumed that the operations in step 1 are carried out in parallel. Hence the p gradient evaluations of step 1 will require only the time to carry out one evaluation. Thus the entry in table 3 for the parallel case is merely the same as that for the serial case minus (p - 1) times the number of cycles.

TABLE 3.- COMPARISON OF NUMBER OF FUNCTION EVALUATIONS REQUIRED TO ACHIEVE CONVERGENCE FOR PVM WITH RESULTS OF OTHER RESEARCHERS

| Method | Number of function evaluations required to achieve convergence for — | |
|---|---|---|
| | Rosenbrock's function | Powell's function |
| Davidon-Fletcher-Powell (Greenstadt) | 192 | 134 to $f = 10^{-11}$ |
| DFP (Straeter) | 101 | Not reported |
| DFP (Jacobson) | 165 | 80 to $f = 10^{-8}$ |
| PVM in serial | 58 | 63 to $f = 10^{-8}$ <br> 81 to $f = 10^{-11}$ <br> 118 to $f = 10^{-15}$ |
| PVM in parallel | 44 | 61 |

## CONCLUDING REMARKS

An algorithm designed to exploit the parallel computing or vector streaming (pipeline) capabilities of computers with these special features has been presented. Several properties of this algorithm were established herein. In addition, the convergence of the iterates to the solution has been proved for a quadratic functional on a real separable Hilbert space; in fact, for a finite-dimensional space the convergence is achieved in one cycle, when the degree of parallelism equals the number of independent variables. Results of numerical experiments indicate that the new algorithm will exploit the parallel or pipeline computing capabilities of the new computers to effect faster convergence than serial techniques currently in use. In fact, the experiments indicated that even when the computations are done serially, the new algorithm is very competitive with the widely used Davidon-Fletcher-Powell technique.

Langley Research Center,
National Aeronautics and Space Administration,
Hampton, Va., September 6, 1973.

# REFERENCES

1. Fletcher, R.; and Reeves, C. M.: Function Minimization by Conjugate Gradients. Computer J., vol. 7, no. 2, July 1964, pp. 149-154.

2. Davidon, William C.: Variable Metric Method for Minimization. ANL-5990 Rev. (Contract W-31-109-eng-38), Argonne Nat. Lab., Nov. 1959.

3. Fletcher, R.; and Powell, M. J. D.: A Rapidly Convergent Descent Method for Minimization. Computer J., vol. 6, no. 2, July 1963, pp. 163-168.

4. Miranker, W. L.: A Survey of Parallelism in Numerical Analysis. SIAM Rev., vol. 13, no. 4, Oct. 1971, pp. 524-547.

5. Avriel, M.; and Wilde, D. J.: Optimal Search for Maximum With Sequences of Simultaneous Function Evaluations. Manage. Sci., vol. 12, no. 9, May 1966, pp. 722-731.

6. Karp, Richard M.; and Miranker, Willard L.: Parallel Minimax Search for a Maximum. J. Combinatorial Theory, vol. 4, 1968, pp. 19-35.

7. Powell, M. J. D.: An Efficient Method for Finding the Minimum of a Function of Several Variables Without Calculating Derivatives. Comput. J., vol. 7, no. 2, July 1964, pp. 155-162.

8. Zangwill, W. I.: Minimizing a Function Without Calculating Derivatives. Comput. J., vol. 10, Nov. 1967, pp. 293-296.

9. Chazan, D.; and Miranker, W. L.: A Nongradient and Parallel Algorithm for Unconstrained Minimization. SIAM J. Contr., vol. 8, no. 2, May 1970, pp. 207-217.

10. Powell, M. J. D.: A Survey of Numerical Methods for Unconstrained Optimization. SIAM Rev., vol. 12, no. 1, Jan. 1970, pp. 79-97.

11. Broyden, C. G.: A Class of Methods for Solving Nonlinear Simultaneous Equations. Math. Comput., vol. 19, no. 92, Oct. 1965, pp. 577-593.

12. Davidon, William C.: Variance Algorithms for Minimization. Computer J., vol. 10, no. 4, Feb. 1968, pp. 406-410.

13. Davidon, William C.: Variance Algorithms for Minimization. Optimization, R. Fletcher, ed., Academic Press, 1969, pp. 13-20.

14. Straeter, Terry Anthony: On the Extension of the Davidon-Broyden Class of Rank One, Quasi-Newton Minimization Methods to an Infinite Dimensional Hilbert Space With Applications to Optimal Control Problems. Ph. D. Thesis, North Carolina State Univ. at Raleigh, 1971.

15. Goffman, Casper; and Pedrick, George: First Course in Functional Analysis. Prentice-Hall, Inc., c.1965.

16. Kowalik, J.; and Osborne, M. R.: Methods for Unconstrained Optimization Problems. Amer. Elsevier Pub. Co., Inc., 1968.

17. Daniel, James W.: The Approximate Minimization of Functionals. Prentice-Hall, Inc., c.1971.

18. Goldfarb, D.: Sufficient Conditions for the Convergence of a Variable Metric Algorithm. Optimization, R. Fletcher, ed., Academic Press, 1969, pp. 273-283.

19. Straeter, Terry A.: Davidon-Broyden Rank-One Minimization Methods in Hilbert Space With Application to Optimal Control Problems. NASA TR R-394, 1972.

20. Kantorovich, L. V.; and Akilov, G. P. (D. E. Brown, trans.): Functional Analysis in Normed Spaces. A. P. Robertson, ed., Macmillan Co., 1964.

21. Greenstadt, J. (With appendix by Y. Bard): Variations on Variable-Metric Methods. Math. Comput., vol. 24, no. 109, Jan. 1970, pp. 1-22.

22. Straeter, Terry A.; and Hogge, John E.: A Comparison of Gradient Dependent Techniques for Minimization of an Unconstrained Function of Several Variables. AIAA J., vol. 8, no. 12, Dec. 1970, pp. 2226-2229.

23. Jacobson, D. H.; and Oksman, W.: An Algorithm That Minimizes Homogeneous Functions of N Variables in N+2 Iterations and Rapidly Minimizes General Functions. Tech. Rep. No. 618 (Contract N00014-67-A-0298-0006), Div. Eng. & Appl. Phys., Harvard Univ., Oct. 1970.