

NATIONAL BUREAU OF STANDARDS
MICROCOPY RESOLUTION TEST CHART



Systems
Optimization
Laboratory

AD-A157 587

SPARSE QUASI-NEWTON METHODS
AND THE CONTINUATION PROBLEM

by

Floyd F. Chadee

TECHNICAL REPORT SOL 85-8

June 1985

Acco
NTIS
DTIC
Un
JAN

DTIC FILE COPY

This document has been approved
for public release and its
distribution is unlimited.

DTIC
ELECTE
AUG 6 1985
S
f
D

Department of Operations Research
Stanford University
Stanford, CA 94305

85 7 30 00 8

SYSTEMS OPTIMIZATION LABORATORY
DEPARTMENT OF OPERATIONS RESEARCH
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

SPARSE QUASI-NEWTON METHODS
AND THE CONTINUATION PROBLEM

by

Floyd F. Chadee

TECHNICAL REPORT SOL 85-8

June 1985

Accession For	
NTIS GRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
Distribution/	
Availability Codes	
Avail and/or	
Dist	Special
A-1	

Research and reproduction of this report were partially supported by the Department of Energy Contract DE-AM03-76SF00326, PA# DE-AT03-76ER72018; National Science Foundation Grants DMS-8420623, ECS-8312142 and DMS-8404121; Office of Naval Research Contract N00014-85-K-0343.



Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do NOT necessarily reflect the views of the above sponsors.

Reproduction in whole or in part is permitted for any purposes of the United States Government. This document has been approved for public release and sale; its distribution is unlimited.

ACKNOWLEDGMENTS

I would like to thank my advisor, Professor B.C. Eaves, for having introduced me to this area of research and for his support of my efforts during the period of work on my dissertation. I am also grateful to the other members of my dissertation committee, Professor Richard W. Cottle and Dr. Michael A. Saunders, for their painstaking review of the first complete draft, and to Dr. Philip E. Gill for being on my Oral Examination Committee. In addition, I would like to thank Michael Saunders for his invaluable help with various computational problems that arose during the course of the dissertation research.

I am also indebted to my friends and fellow students in the department for their support, both academic and non-academic, and for providing the distractions that are so important during a period of solitary research. In particular, I would like to mention Mark and Beverly Beltramo, Fred Krueger, John Stone and Alex Svoronos. To the "gang" of graduate students at Stern Hall, whose dinner-time conversation proved to be such an entertaining and educational experience, I express my sincerest gratitude; the lively discussions and heated debates, that we shared over the years, have changed my perspective on so many things in a more deeply meaningful way than would have been possible in any formal learning process.

I would like to express my gratitude to Audrey Stevenin and Sumi Kawasaki for their assistance during my stay at Stanford, and also to Gail Stein for having done an incredibly efficient job of transforming a mass of handwritten symbols into this readable dissertation.

Finally, I would like to thank my mother and brothers for their continuous support and encouragement over the years.

My expression of gratitude to the above-mentioned persons should not be construed as attributing to them the responsibility for any shortcomings in the present work; any such shortcomings are entirely due to the author.

Floyd Chadee

TABLE OF CONTENTS

CHAPTER		PAGE
	ACKNOWLEDGMENTS	iv
	ABSTRACT	v
	LIST OF FIGURES	vii
1	THE PATH FOLLOWING PROBLEM	1
	1.1 Introduction	1
	1.2 Applications of Path-following	2
	1.3 Historical Background	5
	1.4 Path Existence	8
	1.5 Simplicial Path-following Methods	9
	1.6 The Davidenko Differential Equation	15
	1.7 The Predictor-Corrector Method	17
2	THE PREDICTOR-CORRECTOR METHOD	18
	2.1 Introduction	18
	2.2 Predictor Considerations	22
	2.3 Corrector Considerations	24
	2.4 Orientation	29
	2.5 Steplength Strategy	31
	2.6 Estimation of Tangent Directions	35
	2.7 Terminating the Predictor-Corrector Algorithm	38
	2.8 Quasi-Newton Correctors	43
3	DIRECT SECANT UPDATES OF SPARSE MATRIX FACTORS	44
	3.1 Introduction	44
	3.2 Updating Techniques	48
	3.2.1 Notation	48
	3.2.2 Update I	49
	3.2.3 Algorithm I	52
	3.2.4 Update II	54
	3.2.5 Algorithm II	56
	3.2.6 Pivoting Considerations	56
	3.3 Some Sparsity Relationships	59
	3.3.1 Non-cancellation Assumption	59
	3.3.2 Sparsity Results	63
	3.3.3 Proof of Previous Lemmas	73
	3.4 Convergence Analysis of Algorithm I	75
	3.4.1 Properties of Function $F(\cdot)$	75
	3.4.2 Convergence Results	76
	3.5 Convergence Analysis of Algorithm II	88
	3.6 Concluding Remarks	96
4	COMPUTATIONAL EXPERIENCE	99
	4.1 Introduction	99
	4.2 Local Comparison	99
	4.3 The Continuation Problem	101
	REFERENCES	117

LIST OF FIGURES

FIGURE	PAGE
Figure 1.2.1	3
Figure 1.2.2	3
Figure 2.1.1	20
Figure 2.3.1	27
Figure 2.3.2	27
Figure 2.3.3	28
Figure 2.3.4	28
Figure 2.6.1	37
Figure 2.7.1	42

CHAPTER 1

THE PATH FOLLOWING PROBLEM

1. Introduction

The problem addressed in this dissertation can be described as follows. Assume that we are given,

$$H : \Omega \subset \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$$

$H \in C^2$, i.e., H is twice continuously differentiable

$$H(y^0) = 0$$

$$\det[H'(y^0)] \neq 0.$$

Under these conditions, we know, by the implicit function theorem, that for some neighborhood \mathbb{N} of y^0 , $\mathbb{C} \cap \mathbb{N}$ is a smooth one-dimensional manifold, where \mathbb{C} is the maximal connected subset of $H^{-1}(y^0) = \{y \in \Omega : H(y) = 0\}$ containing y^0 . The problem of interest may then be stated as the numerical task of tracing \mathbb{C} in some specified direction, starting at y^0 and continuing until some point of interest is encountered or until \mathbb{C} ceases to be a smooth one-dimensional manifold. We shall be mainly concerned with the case in which n is large and $H'(y)$ is sparse.

2. Applications of Path-following

Applications of the path-following problem fall roughly into two categories: parametric problems and homotopy problems. By "parametric problems" we mean those problems in which motion along the curve may represent the variation of some system under study, as some naturally occurring parameter in the system is changed. For these problems, the entire curve is normally of interest and numerical methods for tracing it are usually required to do so very closely. Into this category fall the so-called continuation problems (Wacker (1978)). Homotopy problems are generally directed towards the solution of some nonlinear system where this solution is represented by a specific point of C . Algorithms for tracing C in this case are concerned only with reaching this point of interest; C is useful only as a guide to get this solution and hence it may be quite loosely followed until we get close to the desired point.

Example 1: Parametric Structural Problem (Rheinboldt (1981))

Figure (1.2.1) represents a simple plane structure in which two identical rods with longitudinal elastic modulus γ are pin-jointed to the supports A and B and together at C . The vertical displacement, x , under load p satisfies the equation

$$H(x,p) = \gamma \left[\sqrt{\frac{1+h^2}{1-(h-x)^2}} - 1 \right] (h-x) - p = 0 ,$$

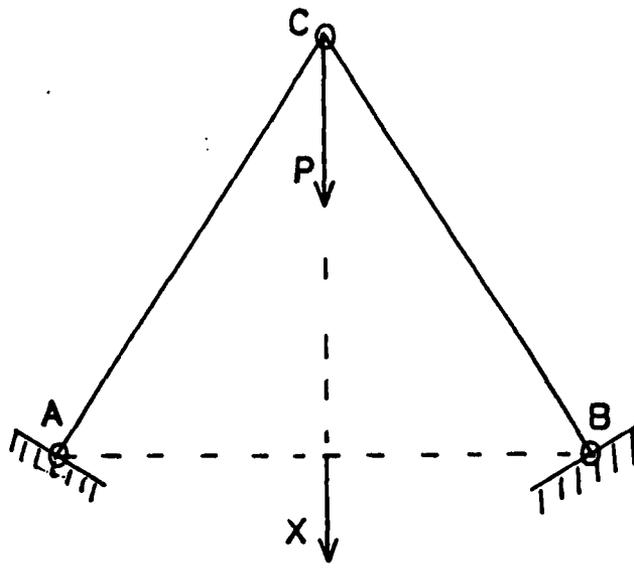


Figure 1.2.1

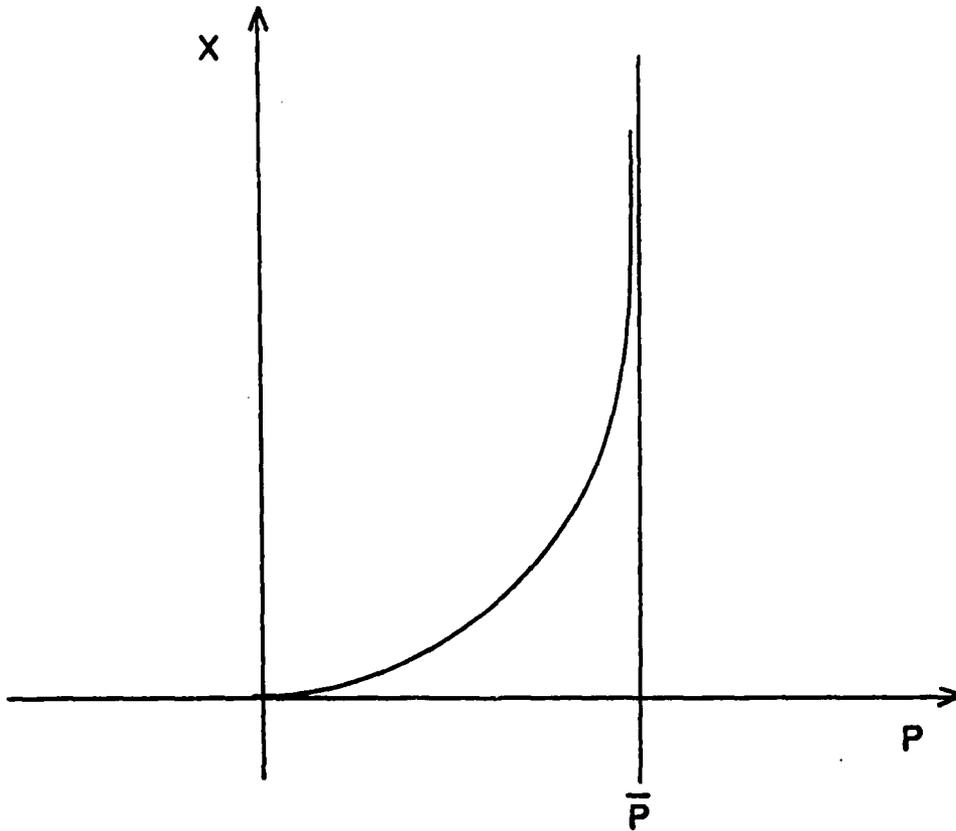


Figure 1.2.2

where h is the vertical distance of C from AB under zero load. The structural engineer, who wishes to study the behavior of the system under varying load p , is interested in all values (x,p) satisfying $H(x,p) = 0$ and $p \geq 0$. Hence he is faced with the problem of tracing $C = \{(x,p) : H(x,p) = 0\}$ from the point $(0,0)$ in the direction of increasing p . The curve C is shown in Figure (1.2.2), where the load $p = \bar{p}$ is an asymptote of C and represents a buckling point of the system. For larger and more complex structures, the system of equations is larger and more complex but the underlying path tracing problem is the same. □

Example 2: Homotopy problem

By Brouwer's Fixed Point Theorem, we know that $f(\cdot)$ has a fixed point in $[1/e, e]$ where $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$

$$f_i(x) = \exp\left[\cos\left(\sum_{j=1}^n x_j\right)\right] \quad \text{for } i = 1, 2, \dots, n.$$

We can locate this fixed point by tracing the path $C = \{(x,t) : H(x,t) = 0\}$ where

$$H(x,t) = x - tf(x)$$

from the point $(0,0)$, starting initially in the direction of increasing t and then continuing along C until some point $(\bar{x},1)$ is encountered. Then \bar{x} is a fixed point of $f(\cdot)$. □

Note that homotopy problems sometimes do require close path following. In Example (2) above, the path C possesses very high curvature through almost its entire length and any path-following technique that is used must follow it very closely or risk the danger of losing it altogether. Also in the homotopy problem for solving for all solutions of polynomial systems (see Garcia and Zangwill (1981), Rosenberg (1983)) several separate paths are followed to different solutions; each path must be closely followed in order to minimize the danger of slipping from one path to another.

3. Historical Background

The path-following problem has developed historically from two completely separate directions. In one direction lies the classical continuation problem as discussed in Ortega and Rheinboldt (1970), Smale (1976) and Wacker (1978); in the other lie the piecewise-linear simplicial techniques as discussed in Eaves (1976) and Todd (1976a).

The classical continuation technique was developed as a globalization of Newton's method and was used for problems in which good starting points were not available. Given a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ a homotopy is set up to produce a smooth path from the available starting point x^0 . This path is followed in the hope that it will lead to a solution of $f(x) = 0$. Examples of possible homotopies are

$$H(x,t) = (1-t)f(x) + tA(x-x^0) \quad A \in \mathbb{R}^{n \times n} \quad (\text{regularizing homotopy})$$

$$H(x,t) = f(x) - (1-t)f(x^0) \quad (\text{defect reducing homotopy})$$

The classical continuation method used "t" as the independent variable so that the algorithm proceeded along the following lines:

- (i) Choose some partition $0 = t^0 < t^1 < \dots < t^{k-1} < t^k = 1$.
- (ii) Solve $H(x, t^i) = 0$ by Newton's method using as starting point x^{i-1} where x^{i-1} is the solution of $H(x, t^{i-1}) = 0$.

Note that x^k solves $f(x) = 0$. As we shall see later, this algorithm is a special -- and inefficient -- implementation of the predictor-corrector method. It breaks down if t does not increase monotonically along the curve C. This is one important difference between the classical continuation method and more recent path-following techniques. The classical method relinquished the path-following task if the curve "turned backwards"; the later methods continue along the curve around such bends by using "t" as a dependent variable, choosing instead the arclength, s, as the independent variable as introduced by Haselgrove (1961). One important problem, to which many papers have been addressed (e.g., Wacker et al. (1978), Deufhard (1979), Den Heijer and Rheinboldt (1981)) is the question of efficient adaptive choice of the partition $\{t^i\}$ as the algorithm progresses. Efforts to attack this problem were directed to the analysis of radii of convergence of Newton's method. Leder (1970) presented another adaptive technique by reformulating the path-following problem as an optimization problem:

$$\min \|G(x,t)\|$$

where

$$G(x,t) = \begin{bmatrix} H(x,t) \\ m(1-t) \end{bmatrix}, \quad m \neq 0, \quad m \text{ constant}.$$

Adaptive incrementing of "t" was achieved by a steplength control based on a monotonicity test in the sense of Goldstein-Armijo (see Armijo (1966) and Goldstein (1967)).

The simplicial techniques originated with Scarf (1967a) and were based on the ideas of complementary pivoting as presented in Lemke and Howson (1964). The relationship between these techniques and homotopy methods was studied by Eaves (1972) and Merrill (1972). For an extensive bibliography on simplicial techniques see Eaves (1976), Todd (1976a) and Allgower and Georg (1980). Scarf's paper was motivated by the search for a fixed point of a mapping. Kellog, Li and Yorke (1976), using a non-retraction principle, provided a constructive proof of Brouwer's fixed-point theorem for smooth mappings, and thus established a link between differentiable homotopy techniques and simplicial methods. This led to a revitalization of interest in the continuation technique. Later followed the studies on differentiable homotopies by Chow et al. (1978), Garcia and Gould (1978) and a host of other publications including the extensive numerical results of Watson (1981).

for more predictor-corrector cycles to trace C . Increasing the predictor steplength leads to more work being needed in the corrector phase, and we may also obtain corrector sequences which fail completely. An efficient implementation of a predictor-corrector algorithm requires a dynamic resetting of the current steplength and an effective strategy for handling those cases in which the corrector sequence diverges. Any such steplength strategy depends in turn on how well we expect the predictor to behave and how powerful the corrector technique is. Ideally we would like to allow for adaptive choice of predictor and corrector techniques and corresponding dynamic reevaluation of steplength strategies, provided this can be obtained economically.

The predictor-corrector method is in sharp contrast to simplicial path-following techniques. While the latter is not conceptually as simple as the predictor-corrector method, once a triangulation has been chosen implementation is a relatively easy task. The strength of the predictor-corrector method versus simplicial techniques is its adaptive ability to move rapidly through well-behaved portions of C by using large steplengths and to slow down at more difficult portions of C . It is the attempt to exploit this capability that leads to difficulties in implementation. The statement of L.F. Shampine (see Watson (1979c)) in reference to numerical techniques for the solution of differential equations ("... how a method is implemented may be more important than the method itself") is clearly applicable to the predictor-corrector method. It should be noted that predictor-corrector methods and

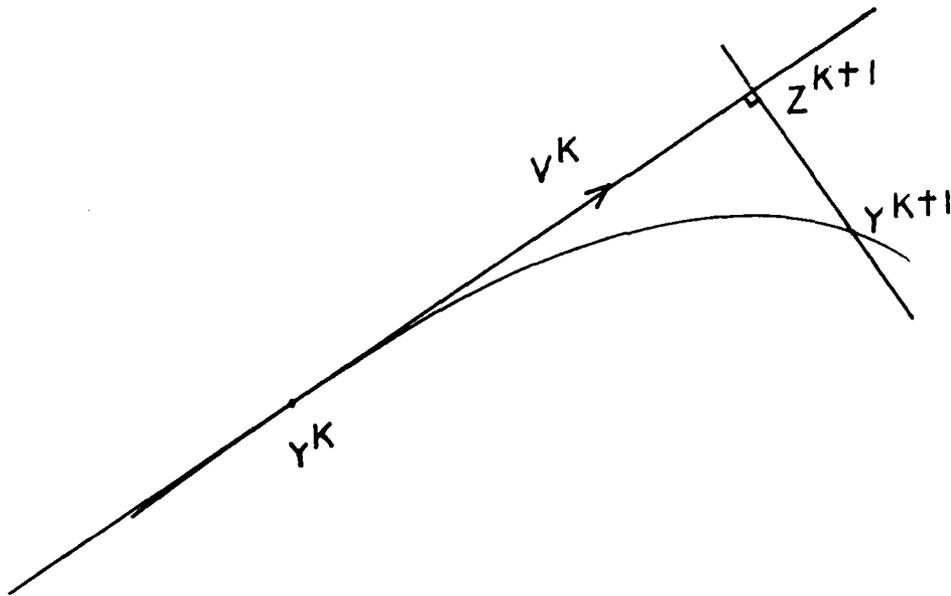


Figure 2.1.1

(i) $z^{k+1} = y^k + \lambda^k v^k$ where $H'(y^k) v^k = 0$, $\|v^k\| = 1$ and λ^k is some predetermined step length.

(ii) Set $w^0 = z^{k+1}$ and

$$w^{j+1} = w^j - \begin{bmatrix} H'(w^j) \\ (v^k)^T \end{bmatrix}^{-1} \begin{bmatrix} H(w^j) \\ 0 \end{bmatrix}$$

(iii) y^{k+1} is taken to be the last w^l in the sequence in (ii)

where the iteration is terminated when some stopping criteria are

satisfied, e.g., y^{k+1} may be taken as z_1^{k+1} where $\|w^l - w^{l+1}\| \leq \epsilon$

and

$$\|w^j - w^{j-1}\| > \epsilon \quad \text{for } j < l, \text{ for some specified } \epsilon > 0.$$

The iteration in (ii) represents the refinement of the initial estimate z^{k+1} in the hyperplane through z^{k+1} perpendicular to the tangent direction at y^{k+1} , see Figure (2.1.1).

The conceptual simplicity of the predictor-corrector method is misleading; in practice an efficient implementation is a quite difficult process. There are many problems which may arise. At the heart of these problems lies the decision, at each predictor step, of what step-length should be used in estimating the next point on the curve. If a small steplength is used, we can expect the next iterative sequence of corrector steps to be quickly convergent. However, there is a tradeoff involved in the use of small predictor steps since this incurs the need

CHAPTER 2

THE PREDICTOR-CORRECTOR METHOD

1. Introduction

Recall the path-following problem stated in Chapter 1. Given $H \in C^2$, $H : \Omega \subset \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$, $H(y^0) = 0$ and $\det[H'(y^0)] \neq 0$ we wish to trace C , the one-dimensional manifold containing y^0 , which is contained in $H^{-1}(0) = \{y \in \mathbb{R}^{n+1} : H(y) = 0\}$. In this chapter we shall discuss the computational considerations involved in using a predictor-corrector method to trace C . Although we are mainly concerned with the specific case when n is large and $H'(\cdot)$ is sparse, most of the following discussion is more generally applicable.

The predictor-corrector algorithm traces C by moving from one point in C to another by the following conceptually simple process. Assume successive points y^0, y^1, \dots, y^k in C have already been located. Then y^{k+1} is obtained by

(i) Predict

Obtain z^{k+1} as an approximation to y^{k+1} by using some extrapolation technique based on the points y^0, y^1, \dots, y^k and possible other information at these points, e.g., $H'(y^0), H'(y^1), \dots, H'(y^k)$ etc.

(ii) Correct

Refine the estimate z^{k+1} by using some local iterative scheme.

For example, if an Euler predictor and a Newton corrector are used then y^{k+1} is obtained as follows:

7. The Predictor-Corrector Method

Equation (6.2) has essentially two separate components. The first term on the right-hand side propels the algorithm along the tangent of the curve, while the second term pushes toward the curve so as to decrease the error incurred in the forward motion. The predictor-corrector method involves an explicit implementation of this idea. Motion along the curve is achieved by predictor-corrector cycles which first predict a point further along on the curve and then use a local iterative technique to correct for the error in prediction. Further details are given in Chapter 2.

The most expensive part of the predictor-corrector algorithm is the local iterative sequence used in the corrector phase. Traditionally, Newton's method has been the choice for the iterative technique, but this may turn out to be an expensive overkill. We shall turn instead to the use of quasi-Newton methods. These are discussed in Chapter 3 for the specific case of large sparse systems.

has suggested approximating $H'(y)$ by the use of quasi-Newton updates. Georg (1981) also uses quasi-Newton updates, but first replaces (6.1) with another system with better stability properties. For $A \in \mathbb{R}^{n \times (n+1)}$, let $t(A) \in \mathbb{R}^n$ be the unique vector satisfying

$$At(A) = 0, \quad \|t(A)\| = 1, \quad \det \begin{bmatrix} A \\ t(A)^T \end{bmatrix} > 0.$$

Then (6.1(a)) is replaced by

$$\frac{dy}{ds} = \delta \cdot t(H'(y)) - [H'(y)]^+ H(y), \quad \delta > 0,$$

where $A^+ = A^T(AA^T)^{-1}$ is the Moore-Penrose inverse of A . (6.2)

The second term on the right-hand side of (6.2) introduces a damping element into the vector field, which makes for a more stable integration. Choice of δ may be made adaptively over different parts of C with δ large where the curvature is small and vice versa.

Tracing C by the use of formulation (6.1) or (6.2) is, in general, not a very efficient technique, even though it does have the convenient property of being able to make use a readily available software. The predictor-corrector technique which is the main focus of this research is more efficient than either the differential equation approach or the simplicial continuation technique.

6. The Davidenko Differential Equation

Tracing C can be accomplished by first converting the problem into the following system of differential equations (Davidenko (1953)):

$$(a) \quad H'(y) \frac{dy}{ds} = 0$$

$$(b) \quad \left| \frac{dy}{ds} \right| = 1$$

$$(c) \quad \det \begin{bmatrix} H'(y) \\ \left(\frac{dy}{ds} \right)^T \end{bmatrix} > 0$$

$$(d) \quad y(0) = y^0 \tag{6.1}$$

The first equation is derived from differentiation of $H(y) = 0$, the second from the use of arclength, s , as the independent variable and the third equation chooses the sign of dy/ds to obtain a consistent orientation. The system (6.1) can now be integrated using any of the efficient and available computer packages for solving the Initial Value Problem. Watson (1981a) has demonstrated the effectiveness of this approach.

Integration of (6.1) is, in general, a quite expensive process. Obtaining dy/ds involves solving the linear system (6.1(a)) and, hence, requires the calculation of $H'(y)$ at each step. Schmidt (1979)

Beginning from a simplex along the chain of simplices which define the piecewise linear path, $G^{-1}(0)$, an attempt is made to predict a simplex further along on this chain by the use of polynomial extrapolation. Usually the predicted simplex does not lie in the chain but just off it. A correction technique, based on topological perturbations, is then used to eliminate this prediction error by locating a simplex on the chain close to the predicted simplex. This technique has been incorporated into the Scout Continuation Package which has been used in the study of various continuation problems (Peitgen and Prufer (1979)).

The main disadvantage of the simplicial technique is its inability to handle problems of large dimensions. The number of simplices that need to be traversed grows rapidly with increasing dimension. In Todd (1980a) and Saigal (1981), techniques are presented for alleviating this problem of dimensionality by exploiting structure and sparsity, which may be encountered in large systems. When separability or sparsity is encountered in the system $H(\cdot)$ and certain specific triangulations are used, the piecewise linear function $G(\cdot)$ is linear over regions which may span groups of adjacent simplices. With the use of appropriate data structures, the simplicial pivot required in moving from one simplex to another within these pieces of linearity can be reduced to a trivial amount of work; the corresponding function evaluation is also virtually eliminated. While these techniques have done a lot towards expanding the range of applicability of simplicial techniques, much work still needs to be done before simplicial techniques can be used as a general path-following technique for large systems.

Two published techniques designed to attack this problem of inefficiency in simplicial continuation methods do so by relaxing the inflexibility of the algorithm. The basic standard technique is retained at points of high curvature so as to take advantage of its robust properties. In other regions of the path, where high curvature is not encountered and robustness is not absolutely necessary, more flexible techniques are used to move rapidly and inexpensively along the path.

The first technique is the "flex simplicial algorithm" of Garcia and Zangwill (1980). Instead of working with a fixed triangulation which is imposed at the start of the procedure, the algorithm allows for the formation of simplices of varying sizes as it moves along the path. Large simplices are used in regions where the path is well-behaved; when high curvature is encountered, the algorithm switches to a fixed triangulation. In this way efficiency and robustness are adaptively traded. The present computational status of the algorithm is unclear. To date, no significant computational experience has been reported. Effective exploitation of the basic idea of the method may still require much research into setting up efficient decision rules for the formation of simplices along the path as the algorithm progresses.

The second technique is the "simplicial predictor-corrector method" of Saupe (1982). Here a fixed triangulation is used and at points of high curvature the standard simplicial pivoting algorithm is in force. In more well-behaved regions, however, the algorithm attempts to skip simplices along the path by using predictor-corrector techniques.

Theorem (5.2)

Let $H : \Omega \subset \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$ where $\Omega = D \times T$, D is the closure of an open bounded set in \mathbb{R}^n and $H^{-1}(0)$ is a one-dimensional manifold. If $H(\cdot)$ satisfies a Lipschitz condition with constant K then for y in the interior of any simplex of the triangulation with simplicial grid size ϵ

$$\|G'(y) - H'(y)\| \leq K\epsilon.$$

Proof: See Saigal (1977). □

Simplicial techniques which employ fine triangulations are very expensive. Each simplex encountered along the piecewise linear path incurs a cost of one function evaluation and one linear programming pivot. Decreasing the grid size results in more simplices being encountered and this leads to an expensive algorithm if the path is very long. On the other hand simplicial path-following techniques are very robust and can be used on highly nonlinear problems for which other path-following methods would fail. They remain unaffected by high curvature of the path, which may cause other methods to lose the correct sense of direction or to cycle. The price of this robustness, though, is very high if the underlying path is very long and it has to be closely followed.

developed to solve for fixed points of nonlinear systems. Hence, the main concern of these algorithms was to get to an endpoint of the path rather than approximating the path closely throughout. The refining triangulations of Eaves (1972) are designed specifically for this goal; $H^{-1}(0)$ is very loosely approximated when we are far away from the point of interest and very closely approximated as we move towards the level $t = 1$. The continuation problem has somewhat different requirements; here, it may be necessary to follow $H^{-1}(0)$ very closely throughout, as in the study of nonlinear eigenvalue problems (Peitgen and Prufer (1979)) or in the problem of finding all solutions of polynomial systems of equations. A fine triangulation must be used throughout so that $G^{-1}(0)$ is close to $H^{-1}(0)$. The following results relate $G^{-1}(0)$ and $H^{-1}(0)$. The simplicial grid is defined as the length of the largest edge over all simplices in the triangulation.

Theorem (5.1)

Let $H : \Omega \subset \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$ where $\Omega = D \times T$ and D is the closure of an open bounded set in \mathbb{R}^n . For any $\delta > 0$, if the simplicial grid of the triangulation is small enough and $y \in G^{-1}(0)$, then

$$\text{dist}(y, H^{-1}(0)) \equiv \min\{\|y-z\| : z \in H^{-1}(0)\} < \delta .$$

Proof: See Garcia and Zangwill (1981), Chapter 12. □

Example 3 (Freudenthal Triangulation) K_1

Denote

$$N \equiv \{1, 2, \dots, n\}$$

$$\pi \equiv \text{some permutation of } N$$

$$K_1^0 \equiv \{y \in \mathbb{R}^n : \frac{y_i}{\delta} \text{ is an integer for } i \in N\},$$

for some $\delta > 0$

$$K_1(y^0, \pi) \equiv \text{the simplex spanned by } \{y^0, y^1, \dots, y^n\}$$

where

$$y^0 \in K_1^0, \quad y^i = y^{i-1} + \delta e^{\pi(i)}, \quad \text{for } 1 < i \leq n$$

and

$$e_k^j = \begin{cases} 1 & \text{for } k = j \\ 0 & \text{otherwise} \end{cases}$$

K_1 is the set of all simplices $K_1(y^0, \pi)$ when $\delta = 1$. We can choose other values for δ to scale the triangulation and it can be made as fine as we want by taking δ sufficiently small. \square

The simplicial technique follows $G^{-1}(0)$ by using the pivoting techniques of linear programming. The principle of complementary pivoting and the use of perturbative techniques ensure that this path is well-defined. Traditionally, simplicial path-following techniques were

with many curves C_e (for e in some region $D_0 \subset \mathbb{R}^n$) most of which are, by Sard's theorem, smooth one-dimensional manifolds. On the other hand, the set of numbers which can be stored on a computer is countable and hence has measure zero; in such a situation, it is not clear what the value of a probabilistic statement is.

For the homotopy problem, it is useful to know, a priori, whether the curve C reaches the level $t = 1$. For a discussion of boundary-free conditions which are sufficient to ensure that this will occur, and which can be shown to be true in some cases, see Eaves (1976) and Garcia and Zangwill (1981). In most cases, no such result can be verified and the implementation of a homotopy method needs to follow the advice of Alexander (1978b): "Have faith."

5. Simplicial Path-Following Methods

Simplicial pivoting techniques provide an effective and robust method for following a piecewise linear approximation to C . Moreover, this method can work with weaker differentiability conditions on the function $H(\cdot)$; in fact, all that is needed is upper semi-continuity of $H(\cdot)$. The first step of the simplicial technique is to impose a triangulation on \mathbb{R}^n ; for details see Eaves (1976) or Todd (1976a). This divides \mathbb{R}^n into a countable number of simplices. The algorithm then works with the piecewise linear function $G(\cdot)$, instead of with $H(\cdot)$, where G agrees with H at the nodes of the simplices and is linear within simplices.

4. Path Existence

It is of interest to know whether we can make any statements, prior to attempting to trace the curve numerically, concerning the existence of such a curve and its smoothness and boundedness properties. As noted in Section (1), all that we can guarantee is that since $\det [H'(y^0)] \neq 0$, there is some neighborhood N containing y^0 such that $N \cap C$ is a smooth one-dimensional manifold. However if we proceed along the curve outside N we may discover that it ceases to be a one-dimensional manifold. Sard's theorem provides a global probabilistic statement concerning the nature of C .

Sard's Theorem (Sard (1942))

Let $G : D \subset \mathbb{R}^q \rightarrow \mathbb{R}^n$, where D is the closure of an open set, and let G be k times continuously differentiable where $k \geq 1 + \max\{0, q-n\}$. Then for almost all e

$$\text{rank } G'(y) = n \text{ for all } y \in G^{-1}(e) \equiv \{y \in D : G(y) = e\}. \quad \square$$

This result tells us that even though C may not be a one-dimensional manifold, we can use an arbitrarily small perturbation, e , so that, by the implicit function theorem, $C_e = \{y : H(y) = e\}$ is a one-dimensional manifold. In practice, the roundoff error encountered by a numerical technique for tracing C automatically provides perturbations in the problem, so we can expect, in some sense, to be dealing

simplicial techniques are not mutually exclusive, as demonstrated in Saupe (1982).

2. Predictor Considerations

The techniques used to obtain the initial estimate, z^{k+1} , of the next point, y^{k+1} , in C can vary from being very simple to quite sophisticated. The choice of predictor depends mainly on how powerful we can expect the corrector technique to be. For example, if Newton's method is used for correcting then simple and crude predictors can work quite well. Simple predicting methods involve fewer arithmetic calculations and also require that less information be retained from previous points, thus incurring a lower storage cost.

(i) The Elevator Predictor (Garcia and Zangwill (1981))

This is a simple predictor which obtains z^{k+1} from y^k by moving parallel to one coordinate axis. If \bar{v}^k satisfies $H'(y^k)\bar{v}^k = 0$, $\|\bar{v}^k\| = 1$ and v^k is some approximation to \bar{v}^k then

$$z_j^{k+1} = \begin{cases} y_j^k, & j \neq i \\ y_i^k + \lambda^k, & j = i \end{cases} \quad (2.1)$$

where

$$i = \operatorname{argmax}_j \{\|v_j^k\|\}, \quad (2.2)$$

and λ^k is some predetermined steplength. If it is known, a priori, that the path is monotonic in one coordinate, say y_p , then we may choose $i = p$ throughout instead of using (2.2), and thus avoid the need to approximate \vec{v}^k . For example, if $y = (x, t)$ and $H'_x(x, t)$ is nonsingular at all points of C , then we can deduce that either $dt/ds > 0$ or $dt/ds < 0$ for all points in C (where $s = \text{arclength}$). In such a case we may simply set $i = n+1$ (since $t = y_{n+1}$). Monotonic curves arise quite often in practice; an important case involves certain homotopies used to solve polynomial systems of equations (Rosenberg (1983)).

(ii) Polynomial and Hermite Predictors

More sophisticated predictors obtain z^{k+1} by fitting polynomials to past points y^k, y^{k-1}, \dots, y^0 of C and derivatives to the curve at these points. In particular, Adams-Bashforth predictors (Shampine and Gordon (1975)) are important for their excellent stability properties and their capacity to make use of approximations of the tangent directions \vec{v}^k at points y^k of C . Using arclength, s , as the independent variable for parameterization of C , with $y(0) = y^0$, we have, at the start of the $(k+1)^{\text{st}}$ predictor step the following approximations:

$$y(s_j) \approx y^j, \quad y'(s_j) \approx v^j \quad \text{for } j = 0, 1, \dots, k \quad (2.3)$$

The Adams-Bashforth predictor based on p points obtains the polynomial $P_{p,k}(t)$ which satisfies

$$P_{p,k}(s_j) = v^j \quad \text{for } j = k-p+1, k-p+2, \dots, k \quad (2.4)$$

and then for some predetermined steplength $\Delta_k = s_{k+1} - s_k$ obtains z^{k+1} as

$$z^{k+1} = y^k + \int_{s_k}^{s_{k+1}} P_{p,k}(t) dt. \quad (2.5)$$

It is easily shown that if a constant steplength Δ is used then the asymptotic error $\tau^k = \|z^{k+1} - y(s_{k+1})\|$ satisfies

$$\tau^k = O(\Delta^{p+2}). \quad (2.6)$$

The use of the Adams-Bashforth predictor is useful in predictor-corrector methods because of its insensitivity to small errors in the tangent directions at y^j , $0 \leq j \leq k$. These errors arise because we may try to estimate the tangent directions instead of directly calculating them and also because the points y^0, y^1, \dots, y^k are not exactly in C . In fact, the previously estimated points y^0, y^1, \dots, y^k may be quite loosely arranged around C since a strategy of loose and inexpensive path following may be employed until we get to the region of C that is of primary interest.

3. Corrector Considerations

The usual and convenient corrector technique involves the use of a locally convergent iterative scheme in a hyperplane through z^{k+1} , that

is sufficiently transversal to C . As in the example of section (1), we may choose the hyperplane $P = \{y : v^k(y - z^{k+1}) = 0\}$. In this case we use an iterative technique to obtain y^{k+1} as the solution of

$$G(y) \equiv \begin{bmatrix} H(y) \\ v^k(y - z^{k+1}) \end{bmatrix} = 0. \quad (3.1)$$

Note that since $\text{rank}[H'(y^k)] = n$ and $H'(y^k)v^k = 0$ we have that

$$G'(y^k) = \begin{bmatrix} H'(y^k) \\ (v^k)^T \end{bmatrix}$$

is nonsingular. Hence using a continuity argument we can deduce that for $\Delta_k = s^{k+1} - s^k$ small enough, there exists a neighborhood $\bar{\Omega}$ such that y^k, z^{k+1} and $y(s_{k+1}) \in \bar{\Omega}$ and $G'(y)$ is nonsingular for $y \in \bar{\Omega}$. This establishes the feasibility of using Newton-type iterations in the solution of (3.1).

We shall be concerned mainly with the use of Newton and quasi-Newton methods in solving (3.1) for y_{k+1} . Each corrector step will be a full Newton step, i.e., no steplength control is imposed on the corrector sequence. The basic philosophy of the predictor corrector method, as implemented here, is that if full Newton steps do not lead to convergence of the corrector sequence then it probably means that z^{k+1} is too far away from C and hence we should restart the predictor-corrector cycle with a shorter steplength, Δ_k . The alternative strategy of using a steplength control to make it more likely that the corrector sequence will converge introduces certain problems. First, if we use

large steps and then encourage convergence by the use of a corrector steplength control strategy then it become more likely that we may converge to a solution of (3.1) other than the one we are seeking (see Figure (2.3.1)). Also in those cases in which a sharp turn causes z^{k+1} to be very far away from C (Figure (2.3.2)) or even leads to the infeasibility of (3.1) (Figure (2.3.3)), then more work is expended before the sequence is eventually relinquished as divergent. For the purpose of safe and effective curve following it seems better to retain a full-step strategy on corrector sequences; steplength controls may increase the likelihood of convergence but not necessarily to the desired point and hence make it more likely that the algorithm may run into cycles (Figure (2.3.4)) or switch to points in $\{y : H(y) = 0\} \setminus C$ (Figure (2.3.1)).

In the classical continuation method (Wacker (1978)), Newton's method is used on the corrector steps. This algorithm was known as the Global Newton Method since the continuation technique was introduced specifically to enlarge the domain of convergence of Newton's method by the construction of a homotopy. The use of quasi-Newton methods on the corrector steps, while mentioned by several authors (Schmidt (1979), Deuffhard (1979), Rheinboldt (1974)), was first discussed in Georg (1981b). Here the focus was on small problems, for which Broyden's Update proved to be a quite effective tool for maintaining a useful approximation of $H'(y)$ throughout the entire length of C , with only an occasional need to re-evaluate $H'(y)$ from scratch.

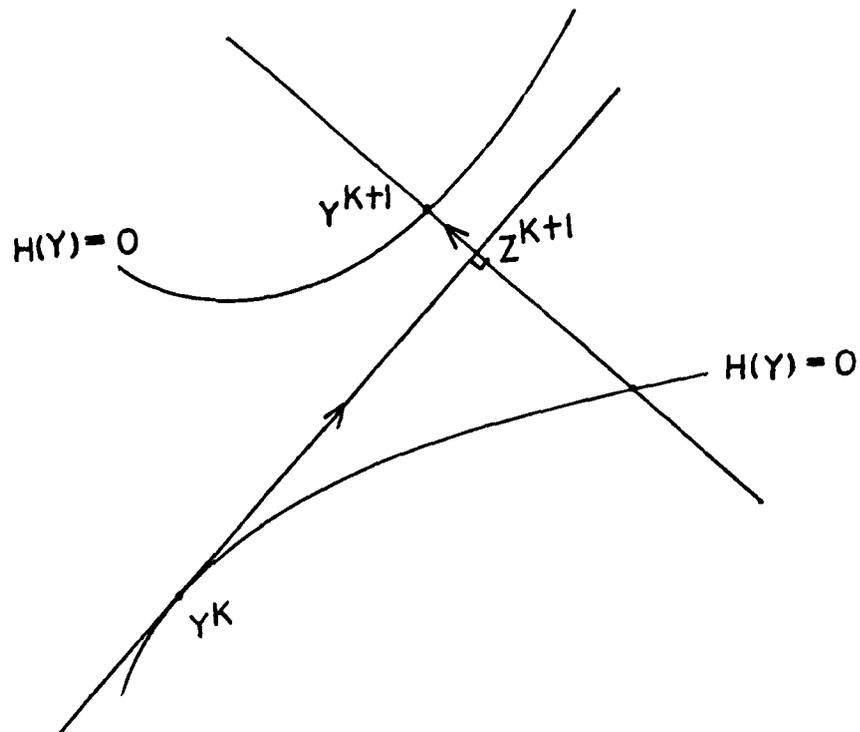


Figure 2.3.1

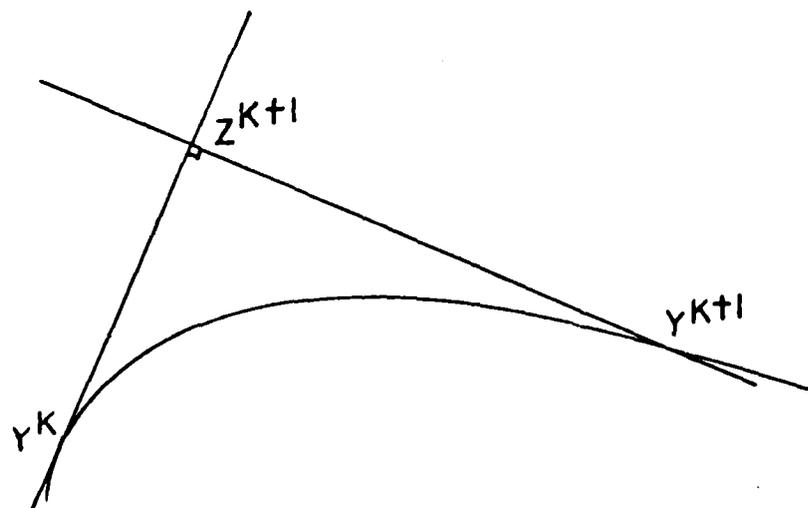


Figure 2.3.2

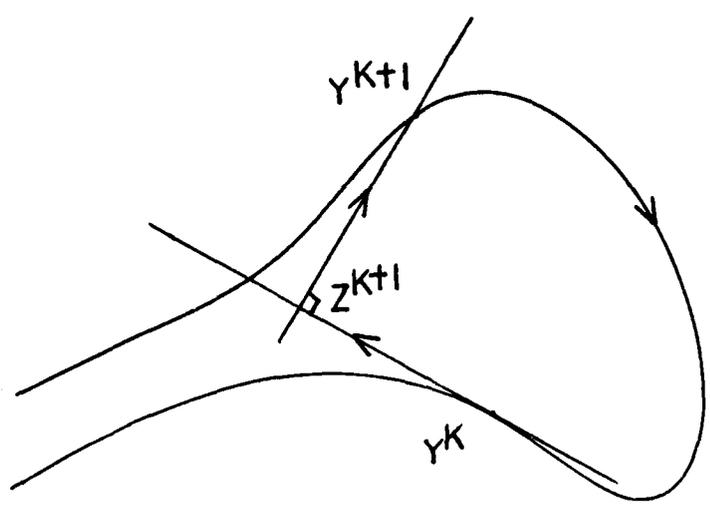


Figure 2.3.3

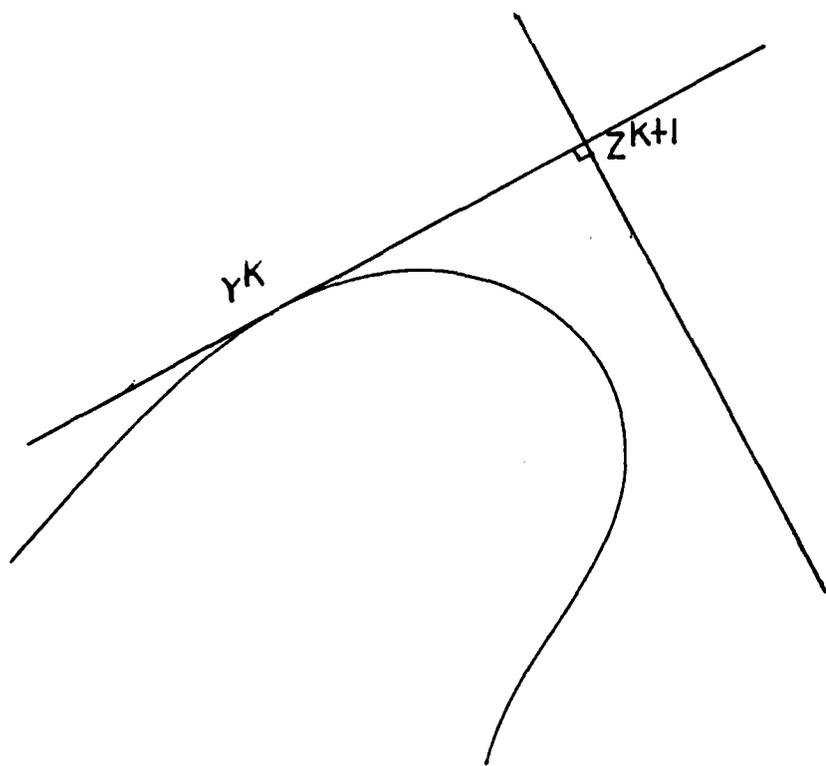


Figure 2.3.4

For large sparse systems the problem is somewhat more difficult. The sparse Broyden method may be used as in the case for small systems. However, as discussed in Chapter 3, this incurs a considerable storage cost since both the current matrix and its LU factors must be maintained explicitly. An alternative strategy is to use the direct secant updates of LU factors which will be discussed in Chapter 3. Note, however, that since these updates can be expected to behave properly only on local convergence problems--because of their inability to update pivoting strategies--their use introduces the need to recalculate the Jacobian, $H'(y)$, from scratch at start of each corrector sequence.

4. Orientation

At the start of each predictor step we need to establish the correct orientation of the curve C , i.e., given $H'(y^k) \bar{w}^k = 0$, $|\bar{w}^k| = 1$ and w^k is some approximation of \bar{w}^k , then the decision has to be made of whether $v^k = -w^k$ or $v^k = +w^k$, where v^k is the direction in which we shall proceed.

We assume that the path is parameterized by arclength, that is

$$C = \{y(s) : 0 \leq s \leq \bar{s}\}$$

where \bar{s} is the length of the curve up to the point of interest. Under the assumption that $\text{rank } [H'(y(s))] = n$ for $0 \leq s \leq \bar{s}$, we have the Basic Differential Equation (Garcia and Zangwill (1981)) which describes C :

$$\frac{dy_i}{ds} = \frac{u_i}{|u|} \quad \text{for } i = 1, 2, \dots, n+1 \quad (4.1)$$

where

$$u_i = q(-1)^i \det[H'_{-i}(y)] \quad \text{for } i = 1, 2, \dots, n+1 \quad (4.2)$$

and $q = \pm 1$ depending on original orientation. $H'_{-i}(y) \in \mathbb{R}^{n \times n}$ is obtained by eliminating the i th column of $H'(y)$. We choose q so that (4.1) is satisfied at $y(0) = y^0$, and then it remains constant thereafter.

We know by Sard's Theorem that $\text{rank } [H'(y(s))] = n$ for $0 \leq s \leq \bar{s}$ is an event of probability one. Hence in order to establish the correct orientation we can choose $v = u/|u|$ where u is given by (4.2).

However, in practice, interesting problems do arise in which the curve C does not have the full rank property throughout. In such cases the path C may intersect other curves $C_i \subset \{y : H(y) = 0\}$. At such points of intersection -- known as bifurcation points -- we have $\det[A(s)] = 0$ where

$$A(s) = \begin{bmatrix} H'(y(s)) \\ [dy/ds]^T \end{bmatrix} \quad (4.3)$$

A bifurcation point, $y(\tilde{s})$, is characterized as odd or even depending on the number of eigenvalues of $A(\tilde{s})$ that go to zero (Crandall and

Rabinowitz (1971)). If the predictor-corrector algorithm jumps over an odd bifurcation point then we need to change the sign of q if we wish to continue along C . Keeping q constant will cause the algorithm to double back along that part of C from which it came.

In order to ensure that the predictor-corrector method follows the curve safely, we monitor the orientation as given by (4.1), and we employ the following orientation strategy. The point y^{k+1} is taken to be an acceptable next point along C if

$$\arccos(v^k \cdot w^{k+1}) \leq \theta \quad \text{or} \quad \arccos(-v^k \cdot w^{k+1}) \leq \theta \quad (4.4)$$

where θ is some given acute angle and w^{k+1} approximates \bar{w}^{k+1} where \bar{w}^{k+1} is defined by $H'(y^{k+1}) \bar{w}^{k+1} = 0$, $\|\bar{w}^{k+1}\| = 1$. If (4.4) is satisfied, then let

$$v^{k+1} = \text{sign}(v^k \cdot w^{k+1}) w^{k+1} . \quad (4.5)$$

This strategy, along with (4.2), results in a safe curve following algorithm and can be used to detect when an odd bifurcation has been encountered.

5. Steplength Strategy

The predictor steplength strategy is the most crucial control problem which arises in an implementation of a predictor-corrector algorithm. An over-conservative steplength strategy, which employs small steps, leads to successful corrector sequences but to many

predictor-corrector cycles. A more ambitious steplength strategy results in longer corrector sequences or to failed corrector sequences for which the predictor-corrector cycle must be restarted with smaller steplength. Denote by λ^k the current value of the steplength for the k th cycle. One basic steplength strategy is the following:

Steplength Strategy (A):

- i) $\lambda^{k+1} = \min\{\lambda_{\max}, \alpha\lambda^k\}$ for some $\alpha > 1$.
- ii) If the corrector sequence fails then set $\lambda^{k+1} = \lambda^{k+1}/\beta$ for $\beta > 1$, and restart the predictor-corrector cycle.
- iii) If necessary repeat (ii) until the corrector sequence converges or until $\lambda^{k+1} < \lambda_{\min}$.

Maximum and minimum steplengths are denoted by λ_{\max} , λ_{\min} . The maximum allowable steplength prevents dangerously large steplengths being taken which may lead to loss of the curve; if the steplength required for convergence falls below λ_{\min} , then the algorithm has failed. Strategy (A) attempts to increase the steplength at the start of each cycle. Several consecutive successful corrector sequences will cause the steplength to grow at an exponential rate. A more conservative strategy is:

Steplength Strategy (B):

$$i) \lambda^{k+1} = \begin{cases} \min\{\lambda_{\max}, \alpha\lambda^k\} & \text{if } \lambda^k \geq \lambda^{k+1} \text{ or } k = 2 \text{ where } \alpha > 1 \\ \lambda^k & \text{otherwise} \end{cases}$$

ii) Same as in Strategy (A).

iii) Same as in Strategy (A).

Strategies (A) and (B) are very simple procedures and are based on the idea of becoming more ambitious when things appear to be going well and more conservative when difficulties are encountered. More sophisticated strategies take into consideration the specific predictor and corrector techniques being employed. Most such strategies attempt to minimize the number of predictor-corrector cycles by estimating, at the start of each cycle, the maximum possible steplength that could be taken and still allow convergence of the next corrector sequence. Den Heijer and Rheinboldt (1981) showed that while a finite upper bound on the radius of convergence of the next corrector sequence cannot be derived solely from information based on previous corrector iterates but requires more global information, we can still use the accumulated information from past corrector sequences to derive a useful estimate of this radius. Deuflhard (1979) and Hackl et al. (1980) also give steplength strategies for the case in which Newton's method is used as the corrector technique. These methods are all based upon the use of information derived from the local behavior of the algorithm to estimate global constants associated with the convergence of Newton's Method.

When quasi-Newton -- rather than Newton -- correctors are used, it is not clear how applicable the above techniques are. Moreover, these techniques are based on the questionable assumption that it is desirable for each predictor step to be as large as possible while still maintaining corrector convergence. They do not take into consideration the possibility that such a strategy might actually lead to an increase in the total work involved in traversing C by increasing the number of corrector steps needed for convergence. In our implementation we shall turn to the more heuristic approach of Georg (1981).

Assume that the predictor is an Adams-Bashforth extrapolation based on the last m points. Then

$$\|z^{k+1} - y^{k+1}\| = O(\Delta_k^p) \quad \text{where } p = m+2, \Delta_k = s_{k+1} - s_k. \quad (5.1)$$

Now if the Jacobian is accurate -- that is, calculated either analytically or by finite differences -- at the start of the corrector step, then we can show the following (Georg (1981)):

$$\kappa_{k+1} \equiv \frac{\|H(w^0)\|}{\|H(w^1)\|} = O(\Delta_k^p), \quad \text{where } w^j \text{ is as defined in Section (1)(5.2)}$$

$$\bar{\kappa}_{k+1} \equiv \frac{\|a_1\|}{\|a_0\|} = O(\Delta_k^p), \quad \text{where } a_i = \begin{bmatrix} H'(w^i) \\ v^i \end{bmatrix}^{-1} \begin{bmatrix} H(w^i) \\ 0 \end{bmatrix} \quad (5.3)$$

for $i = 0, 1$

$$d_{k+1} \equiv \|H(w^0)\| = O(\Delta_k^p) \quad (5.4)$$

$$\bar{d}_{k+1} \equiv \|a_0\| = O(\Delta_k^p) \quad (5.5)$$

$$\alpha_{k+1} \equiv \arccos(v^k, v^{k+1}) = O(\Delta_k) . \quad (5.6)$$

By monitoring the variables defined in (5.2-5.6) a simple and effective heuristic is developed as follows. The user of the algorithm decides, a priori, what would be ideal values for each of these variables and then chooses the steplength, Δ , to make it likely that these values are attained. For example, we have from (5.2)

$$\frac{\kappa_k}{\kappa_{k+1}} \approx \left(\frac{\Delta_k}{\Delta_{k+1}} \right)^p . \quad (5.7)$$

Therefore, in order to obtain an ideal value κ_{ideal} on the (k+1)st step we should set

$$\Delta_{k+1} \approx \left(\frac{\kappa_{ideal}}{\kappa_k} \right)^{1/p} \Delta_k . \quad (5.8)$$

Similarly other values for the steplength can be estimated using (5.3-5.6). We then take Δ_{k+1} to be the minimum of all these estimates.

6. Estimation of Tangent Directions

The use of Adams-Bashforth predictors requires that at the end of each predictor-corrector cycle we estimate the tangent direction, v^k , at the recently located point, y^k , of C . The safest, but somewhat expensive, technique is to calculate it precisely by solving for the

Also let

$S^\Delta \equiv$ the orthogonal projection operator into the subspace Δ

$\Psi_k \equiv \text{span}\{e_1, e_2, \dots, e_k\}$

$\Phi_k \equiv \text{span}\{e_k, e_{k+1}, \dots, e_n\}$

$\chi(\Delta) \equiv \{1 \leq i \leq n : \text{there exists a } v \in \Delta \text{ such that } v_i \neq 0\}$

for any subspace $\Delta \subset \mathbb{R}^n$ (2.1.2)

2.2. Update I

Consider the updating problem encountered at each step. Ignoring superfluous subscripts in this section, we may state it as follows. There is some current approximation, A , to the Jacobian matrix. It is available in the form of the LU factors where $A = LU$. The last Newton step has provided new information which we will use to update the current approximate Jacobian. We wish to update (L,U) directly to (\bar{L}, \bar{U}) where, for $s, y \in \mathbb{R}^n$, we require

$$\bar{L}\bar{U}s = y \quad \text{where } y = F(x+s) - F(x)$$

\bar{L} is a unit lower triangular matrix (2.2.1)

\bar{U} is an upper triangular matrix .

The system (2.2.1) may be rewritten as follows:

step. In Section 2 a new update is described, which may be viewed as a generalization of this update and which allows both L and U to vary. Johnson and Austria (1983) presented an algorithm for full matrices which maintains the factors L^{-1} and U explicitly and updates these factors directly at each iteration. In Section 2 a sparse variant of this update is presented. In Section 3 some sparsity results relating the sparsity pattern of A^k to the sparsity patterns of L , L^{-1} and U are examined. In Sections 4 and 5 local convergence analyses of the two updates are presented. Section 6 concludes with a comparison of the two updates and suggestions are made for overcoming their present disadvantages.

• Updating Techniques

•1. Notation

The following sparsity notation, which generalizes the notation used in Section 1, is convenient and will be used throughout the rest of the chapter. For any matrix-valued function $B : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$ define

$$Z_1^B \equiv \{v \in \mathbb{R}^n : e_j^T v = 0 \text{ for all } j \text{ such that } e_1^T B(x) e_j = 0 \\ \text{for all } x \in \Omega\}$$

$$\bar{Z}_1^B \equiv Z_1^B \cap \{v \in \mathbb{R}^n : v_1 = 0\} \quad (2.1.1)$$

$$Z^B \equiv \{M \in \mathbb{R}^{n \times n} : M^T e_i \in Z_1^B \text{ for } i = 1, 2, \dots, n\}$$

$$\bar{Z}^B \equiv \{M \in \mathbb{R}^{n \times n} : M^T e_i \in \bar{Z}_1^B \text{ for } i = 1, 2, \dots, n\}.$$

suffers disadvantages not associated with its full lower dimensional counterpart. For the full Broyden update the $O(n^3)$ operations involved in solving for the Newton step, s^k , at each stage of the iteration (1.1.1) can be reduced to $O(n^2)$ operations by using the techniques of Gill et al. (1974). These techniques make use of the fact that for the full Broyden update $(A^{k+1} - A^k)$ is a rank one matrix. The QR factors of A^k are maintained explicitly and are updated to give the QR factors of A^{k+1} . The Jacobian approximation A^k is never actually explicitly represented in storage. For the sparse Broyden update, however, $(A^{k+1} - A^k)$ is generally of rank n and the techniques of Gill et al. (1974) are not applicable. A^k must be explicitly maintained so that it can be updated to A^{k+1} according to (1.1.6). Solution of (1.1.1) for the Newton step s^k , therefore, incurs both the need to refactorize the new matrix A^k at each step and the need for extra storage to hold the factors of A^k . If these storage costs and the cost of factorization, relative to the cost of function evaluations, are high enough then other updating techniques for sparse problems may become competitive even when they may have slower convergence rates than the sparse Broyden method.

It is assumed for the rest of this chapter that the storage requirements for the problem allow for the solution of the Newton step, s^k , in (1.1.1) by factorization techniques. We focus on methods which maintain sparse LU factors of A^k which are directly updated by incorporating the quasi-Newton information at each step. Dennis and Moré (1982) introduced an update which holds L fixed and updates U only at each

$$Z_i = \{v \in \mathbb{R}^n : e_j^T v = 0 \text{ for all } j \text{ such } e_i^T F'(x) e_j = 0$$

for all $x \in \Omega\}$

and

$$Z = \{M \in \mathbb{R}^{n \times n} : M^T e_i \in Z_i \text{ for } i = 1, 2, \dots, n\} \quad (1.1.4)$$

Z_i represents the sparsity pattern of the i^{th} row of $F'(x)$, while Z represents the sparsity pattern of $F'(x)$. A^{k+1} is chosen to be the solution of the optimization problem

$$\min\{\|A - A^k\| : A \in Q(y^k, s^k) \cap Z\}. \quad (1.1.5)$$

Let S^Δ be the orthogonal projection operator into the subspace $\Delta \subset \mathbb{R}^n$. Then the solution, A^{k+1} , of (1.1.5) may be written explicitly as

$$A^{k+1} = A^k + \sum_{i=1}^n [(S^{Z_i s^k})^T (S^{Z_i s^k})]^\dagger e_i^T (y^k - A^k s^k) e_i (S^{Z_i s^k})^T$$

where

$$(a)^\dagger = \begin{cases} 0 & \text{for } a = 0 \\ a^{-1} & a \neq 0 \end{cases}. \quad (1.1.6)$$

In Broyden, Dennis and Moré (1973), local superlinear convergence of Broyden's updating technique was demonstrated; later, Marwil (1978) showed that the same is true for the sparse Broyden technique. In an actual computer implementation, however, the sparse Broyden method

Schnabel (1979)) -- adaptively define A^k from A^{k-1} as the iteration proceeds. For example, Broyden's method (Broyden (1965)) derives A^k for $k > 0$ through the following updating technique. After completion of the steps in (1.1.1), A^{k+1} is obtained as

$$A^{k+1} = A^k + (y^k - A^k s^k)(s^k)^T / (s^k \cdot s^k)$$

where $y^k = F(x^{k+1}) - F(x^k)$. (1.1.2)

A^{k+1} is chosen to solve the optimization problem

$$\min\{\|A - A^k\| : A \in Q(y^k, s^k)\}$$

where $Q(y^k, s^k) = \{M \in \mathbb{R}^{n \times n} : Ms^k = y^k\}$

and $\|\cdot\|$ represents the Frobenius norm . (1.1.3)

A^{k+1} is thus derived from A^k by incorporating the new slope information obtained in moving from x^k to x^{k+1} . " $A^{k+1} \in Q(y^k, s^k)$ " is known as the quasi-Newton condition.

For large sparse problems the update (1.1.2) is inappropriate since, in general, it results in an updated Jacobian approximation which is dense. The sparse Broyden update (Broyden (1971), Schubert (1970)) avoids this drawback by imposing a further condition on the optimization problem (1.1.3) so that sparsity is maintained. Using the notation of Dennis and Moré (1982), one obtains the sparse Broyden update as follows:

Let

CHAPTER 3

DIRECT SECANT UPDATES OF SPARSE MATRIX FACTORS

1. Introduction

Consider the general Newton-type iterative technique used to solve for a zero of a non-linear system of equations. Given a function $F : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$, we wish to locate $x^* \in \Omega$ such that $F(x^*) = 0$. In this chapter we shall be concerned mainly with the local convergence problem, i.e., given x^0 and A^0 as initial estimates of x^* and $F'(x^0)$, respectively, we seek to iteratively refine this estimate, x^0 , until it is within some prescribed distance $\epsilon > 0$ from x^* . The $(k+1)^{\text{st}}$ step of the iteration takes the following general form:

$$\begin{aligned} \text{solve} \quad & A^k s^k = -F(x^k), \quad A^k \in \mathbb{R}^{n \times n} \\ \text{and set} \quad & x^{k+1} = x^k + \lambda^k s^k. \end{aligned} \quad (1.1.1)$$

A^k is chosen as some approximation to the Jacobian $F'(x^k)$. Dennis and Moré (1974) demonstrated that, in algorithms of this type, local super-linear convergence of x^k to x^* requires $\lim_{k \rightarrow \infty} \lambda^k = 1$. Hence for x^k close to x^* we may wish to take $\lambda^k = 1$. For the remainder of this chapter, we shall set $\lambda^k = 1$ for all $k \geq 0$.

There are many variations on the specification of A^k . Setting $A^k = A^0$ for $k > 0$ results in the pseudo-Newton method, while $A^k = F'(x^k)$ gives Newton's method. Quasi-Newton methods -- also known as least-change secant methods (Dennis and Moré (1977)), Dennis and

8. Quasi-Newton Correctors

Most of the work involved in a predictor-corrector algorithm is incurred in the corrector phase of the algorithm. As a corrector, Newton's method is very reliable but also very expensive. In our algorithm we attempt to reduce this expense by using quasi-Newton correctors instead. Quasi-Newton techniques for large sparse systems do not, in general, have a very strong reputation (Thapa (1981)). However, there are certain differences between the corrector convergence problem and the general root-finding convergence problem which suggest that there are advantages to be gained from the use of quasi-Newton techniques, rather than the more robust and expensive Newton's method, in predictor-corrector algorithms.

First, the level of difficulty of each corrector problem is an open choice; it can be varied by choosing different predictor steplengths. Hence less robust convergence techniques are feasible. Secondly, it may actually be better for the algorithm to take two short predictor steps and use a less expensive corrector technique than to take one large step and then use a powerful and expensive corrector technique to solve the resulting difficult corrector problem. Not only is it possible that less total work may be required, but also from the point of view of safe path following the second strategy has the disadvantage that it is more likely to lead to corrector divergence or, even worse, to convergence to a point on another curve.

In the next chapter we examine the theoretical aspects of quasi-Newton methods for large sparse systems.

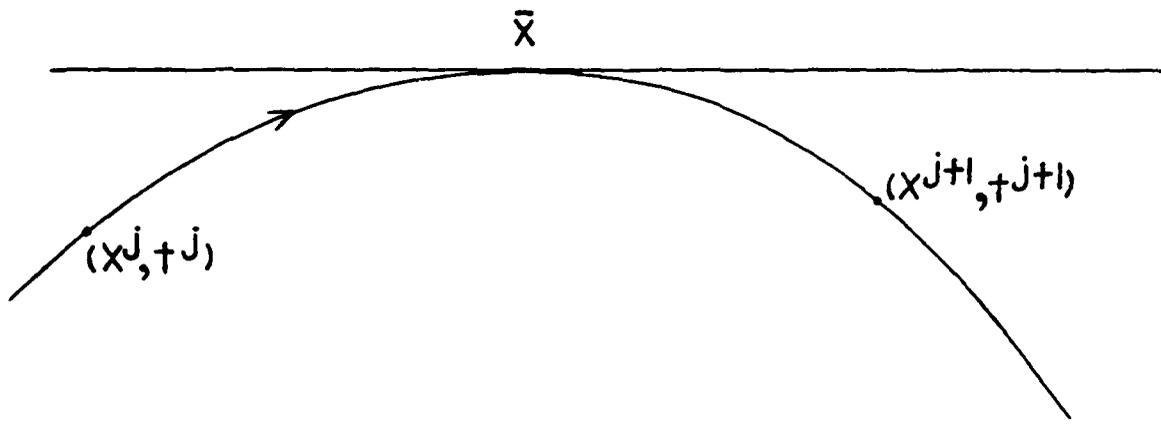


Figure 2.7.1

Hence if M is large, dy_{n+1}/ds is very small and a near-tangential approach is observed. To distinguish between this case and the case involving singularity of $f'(\bar{x})$, we calculate the condition number of $f'(\bar{x}^0)$. A Newton-type iteration at level $t = 1$ is implemented only if this condition number $\sigma[f'(\bar{x}^0)]$ is less than $\bar{\sigma}$ for some prescribed $\bar{\sigma}$.

Whatever the reason for tangential or near-tangential approach, it is clear that if C runs close to the hyperplane $P = \{(x,t) : t = 1\}$ over a significant distance, then it is necessary to trace C very carefully in this region if we want to obtain a good estimate of the point of intersection $\bar{x} = C \cap P$. For those cases in which $\sigma[f'(\bar{x}^0)] > \bar{\sigma}$, we refine the estimate \bar{x}^0 by restarting with homotopy (7.4) and then carefully tracing C , rather than using an iterative technique at level $t = 1$.

Another problem caused by tangential approach is the situation illustrated in Figure (2.7.1). Here C touches P tangentially at \bar{x} , so that for the points (x^j, t^j) , (x^{j+1}, t^{j+1}) we have $t^j < 1$, $t^{j+1} < 1$. To recognize such an occurrence in practice we check for changes in the sign of dt/ds for t close to 1. When this situation arises, \bar{x} is located by carefully retracing C between (x^j, t^j) and (x^{j+1}, t^{j+1}) using smaller predictor steps.

$$H(x,t) = f(x) - (1-t) f(\bar{x}^0) . \quad (7.4)$$

If $f'(\bar{x})$ is singular then this strategy is not appropriate since it is quite likely that the final iteration at level $t = 1$ will diverge again. Fortunately, this situation can be recognized by the algorithm in practice. If $f'(\bar{x})$ is singular then by (4.2) the path C , obtained from either homotopy (7.1) or (7.2), approaches the hyperplane $\{(x,t) : t = 1\}$ tangentially. Tangential--or near tangential--approach may arise, though, for other reasons. For example, if homotopy (7.2) is used and x^0 is very far away from \bar{x} so that $|f_i(x^0)| > M$ for $1 \leq i \leq n$ and $M > 0$ very large, then

$$\frac{dy}{ds} = \frac{u(s)}{\|u(s)\|} \quad \text{where } u(s) = [-[f'(x(s))]^{-1} f(x^0) : 1] .$$

But

$$\|[f'(x(s))]^{-1} f(x^0)\|$$

$$\geq \|[f'(\bar{x})]^{-1} f(x^0)\| - \| \{ [f'(\bar{x})]^{-1} - [f'(x(s))]^{-1} \} f(x^0) \|$$

$$\geq k_1 M - \epsilon \quad \text{for constants } k_1 > 0, \epsilon > 0$$

where ϵ is very small for $x(s)$ close to \bar{x} ,

$$\geq kM \quad \text{for some constant } k.$$

Hence

$$\frac{dy_{n+1}}{ds} = \frac{1}{\|u(s)\|} \leq \frac{1}{\|[f'(x(s))]^{-1} f(x^0)\|} \leq \frac{1}{kM} .$$

- ii) As the hyperplane $\{(x,t) : t = 1\}$ is approached, tighten up the error tolerances of the corrector sequences and the predictor steplength control heuristics so that the curve is traced more closely.
- iii) Continue the algorithm until some point $y^l = (x^l, t^l)$ is located, where $t^l \geq 1$. Now if $y(s) \equiv (x(s), t(s)) = Q(s)$ is the most recent polynomial predictor which was used to locate z^l (where $s = \text{arclength measured along } C$), then we solve for $x(\hat{s})$ where

$$(x(\hat{s}), 1) = Q(\hat{s}) \quad (7.3)$$

If Q is a low-order polynomial we can solve (7.3) explicitly for $x(\hat{s})$; otherwise, we obtain an estimate of $x(\hat{s})$ by interpolation between y^l and y^{l-1} and then use the one-dimensional Newton's method to solve the polynomial system (7.3) for $x(\hat{s})$.

- iv) Now taking $\bar{x}^{-0} = x(\hat{s})$ as an initial estimate of \bar{x} where $f(\bar{x}) = 0$, we use Newton's method or a quasi-Newton method to solve $f(x) = 0$.

Divergence of the final Newton-type iteration at level $t = 1$, in step (iv) above, results in failure of this basic technique. Such divergence means that the path following algorithm did not serve its purpose, which was to obtain an estimate of the solution, \bar{x} , of $f(x) = 0$ which is within the domain of convergence of the final iterative technique. One possible strategy, in this situation, is to restart the entire path following algorithm at $(\bar{x}^{-0}, 0)$ using a new homotopy, e.g.,

For updating techniques which do not allow good approximations to the tangent directions by use of this method, we switch to either direct calculation of \bar{v}^{-k} or the use of a simpler predictor which does not require knowledge of the tangent directions.

7. Terminating the Predictor-Corrector Algorithm

The task of tracing C may be carried out with one of two purposes in mind: either all of C is of interest or only some specific point or region of C is. If the first reason is true, then the algorithm will be relatively conservative with short predictor steps and tight error tolerances on the corrector steps so as to allow for close curve following. If, on the other hand, the algorithm is being used only to get to some endpoint then it attempts to follow C as loosely as possible without losing the curve until the region of interest is encountered.

The homotopy problem is an example of the case in which we are interested only in a specific point of C . For example, if

$$H(x,t) = tf(x) + (1-t) A(x-x^0), \quad A \in \mathbb{R}^{n \times n} \quad (7.1)$$

$$H(x,t) = f(x) - (1-t) f(x^0) \quad (7.2)$$

then we wish to locate $\bar{y} = (\bar{x}, \bar{t}) \in C$ where $\bar{t} = 1$. We then have \bar{x} as a solution of $f(x) = 0$.

The basic technique for locating \bar{y} is as follows:

- 1) Begin the predictor-corrector algorithm at $(x^0, 0)$ moving along C in the direction of increasing t .

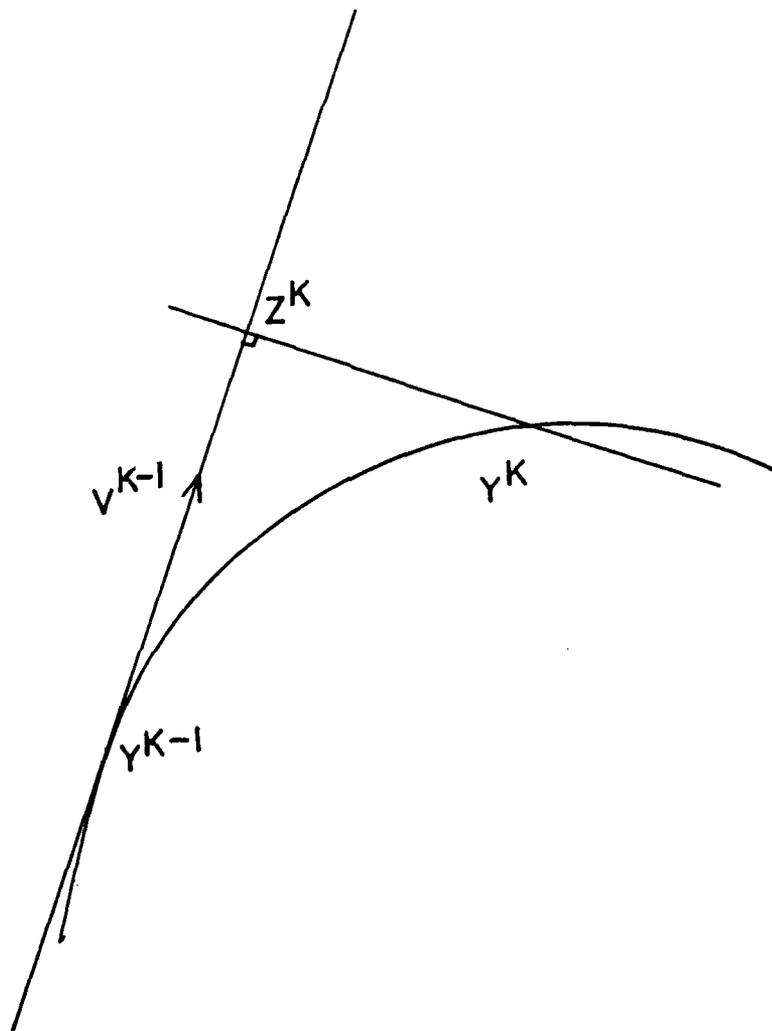


Figure 2.6.1

kernel of $H'(y^k)$, where $H'(y^k)$ is calculated analytically or by finite differences. In the case of quasi-Newton correctors, a less expensive technique which works well in some cases (as discussed in Chapter 4) is to update the currently available approximation to the Jacobian so that it is correct in the two-dimensional subspace spanned by v^{k-1} and $(y^k - z^k)$ (see Figure 2.6.1). Note that this technique requires two extra function evaluations at the end of the corrector phase.

Example:

If Broyden's update is being used, and the current approximation to the Jacobian is B , then let

$$\bar{B} = B + (y_1 - Bs_1) \frac{s_1^T}{s_1^T s_1} + (y_2 - Bs_2) \frac{s_2^T}{s_2^T s_2}$$

where

$$s_1 = \delta v^{k-1}, \quad \text{for some } \delta > 0$$

$$s_2 = \delta(y^k - z^k), \quad \text{for some } \delta > 0$$

$$y_1 = H(y^k + s_1) - H(y^k),$$

$$y_2 = H(y^k + s_2) - H(y^k).$$

Note that $s_1 \cdot s_2 = 0$ since all corrector steps are perpendicular to v^k . It is easily seen that if δ is small enough, the matrix \bar{B} is effectively correct in the subspace spanned by $\{s_1, s_2\}$.

$$\bar{\phi}_i \omega^{(i)} = y_i, \quad \text{for } i = 1, 2, \dots, n \quad (2.2.2)$$

where we denote

$$\omega^{(i)} = \begin{bmatrix} \bar{U}_{11} & \bar{U}_{12} & \dots & \bar{U}_{1n} \\ & \cdot & & \\ & & \cdot & \\ & & & \bar{U}_{i-1,i-1} & \dots & \bar{U}_{i-1,n} \\ & & & & & 1 \\ & \circ & & & & & \circ \\ & & & & & & & \cdot \\ & & & & & & & & \cdot \\ & & & & & & & & & 1 \end{bmatrix} \quad s \quad (2.2.3)$$

for $i = 2, 3, \dots, n$

and

$$\begin{aligned} \phi_1 &= (U_{11}, U_{12}, \dots, U_{1n}) \\ \bar{\phi}_1 &= (\bar{U}_{11}, \bar{U}_{12}, \dots, \bar{U}_{1n}) \\ \phi_i &= (L_{i1}, L_{i2}, \dots, L_{i,i-1}, U_{i1}, \dots, U_{in}) \quad \text{for } 2 \leq i \leq n \\ \bar{\phi}_i &= (\bar{L}_{i1}, \bar{L}_{i2}, \dots, \bar{L}_{i,i-1}, \bar{U}_{i1}, \dots, \bar{U}_{in}) \quad \text{for } 2 \leq i \leq n. \end{aligned} \quad (2.2.4)$$

Using (2.2.2-2.2.4), a least change secant update is used to determine the rows of \bar{L} , \bar{U} sequentially. Under the assumption that rows $k = 1, 2, \dots, i-1$ of \bar{L} , \bar{U} have already been determined, then $\bar{\phi}_i$, whose components are the same as those of the i^{th} rows of \bar{L} , \bar{U} , is chosen as the solution of

$$\min_{\hat{\phi}_i} \|\hat{\phi}_i - \phi_i\| \quad (2.2.5(a))$$

subject to

$$\hat{\phi}_i \omega^{(i)} = y_i \quad (2.2.5(b))$$

$$\hat{\phi}_i \in \text{Span}\{\bar{z}_i^L, z_i^U\}. \quad (2.2.5(c))$$

Condition (2.2.5(c)) maintains the sparsity patterns of L, U while (2.2.5(b)) is the quasi-Newton condition for the i^{th} row. Note that the update of Dennis and Moré (1982) is obtained if we impose on the optimization problem (2.2.5) the further condition

$$S^{-1}(\bar{z}_i^L (\hat{\phi}_i - \phi_i)) = 0 \quad (2.2.6)$$

which holds L fixed. The next theorem gives the solution of (2.2.5); the proof is deferred until Section 3.

Theorem 2.1:

Let $Y_i \equiv \text{Span}\{\bar{z}_i^L, z_i^U\}$. The solution of (2.2.5) may be written explicitly as

$$\bar{\phi}_i = \phi_i + [(S^{Y_i(\omega^{(i)})})^T (S^{Y_i(\omega^{(i)})})]^{-1} \times (y_i - \phi_i - \omega^{(i)}) (S^{Y_i(\omega^{(i)})})^T \quad (2.2.7)$$

where $\omega^{(i)}$ is determined iteratively, according to (2.2.3), by

$$\omega_j^{(i)} = \begin{cases} \sum_{\alpha=i-1}^n \bar{u}_{i-1,\alpha} s_\alpha & \text{if } j = i-1 \\ \omega_j^{(i-1)} & \text{otherwise} \end{cases} \quad \text{for } i = 2, 3, \dots, n \quad (2.2.8)$$

□

The convergence analysis of this update -- presented in Section 4 -- seems to suggest that it should be implemented in a cautious type of algorithm which allows for periodic restarts, similar to that used by Dennis and Moré (1982).

2.3. Algorithm I:

(1) Choose

$x^0 \in \mathbb{R}^n$ as an approximation to x^*

m a fixed positive integer

$k \leftarrow 0$.

(2) Evaluate $F(x^0)$ and A^0 , a finite difference (or analytic) approximation to $F'(x^0)$.

(3) Factorize $P^0 A^0 Q^0 = L^0 U^0$ using a threshold pivoting strategy, where P^0, Q^0 are permutation matrices.

(4) Solve

$$L^k w^k = -P^0 F(x^k),$$

$$U^k t^k = w^k,$$

$$s^k = Q^0 t^k.$$

(5) $x^{k+1} = x^k + s^k$

Evaluate $F(x^{k+1})$

If stopping criteria are met, then STOP.

(6) If $k = m-1$, set $x^0 = x^{k+1}$, $k = 0$ and go to (2).

(7) Set $y^k = F(x^{k+1}) - F(x^k)$ and update the rows of L, U according to (2.2.7).

(8) $k \leftarrow k + 1$; go to (4).

For simplicity, step (4) is stated under the assumption that U^k is non-singular. If U^k is singular after updating, the entire algorithm is restarted with x^0 as the current value, x^k . The threshold pivoting strategy of step (3) is used to enhance the sparsity

of L^0 and U^0 by relaxing the usual stability test; for further details see Duff (1977). As with the update presented in Dennis and Moré (1982), periodic recalculation of the Jacobian matrix seems to be necessary to ensure convergence.

2.4. Update II

The somewhat unsatisfactory theoretical requirement of Update I, that periodic restarts are necessary after some fixed number of steps, may be avoided if we work explicitly with L^{-1} instead of L . We employ a sparse version of the update in Johnson and Austria (1983). Denoting $N = L^{-1}$, i.e., $NA = U$, where N is a unit lower triangular matrix and U is upper triangular, we consider the problem of directly updating (N,U) to (\bar{N},\bar{U}) such that

$$\bar{N}y = \bar{U}s \quad . \quad (2.4.1)$$

Condition (2.4.1) is the quasi-Newton condition, equivalent to $\bar{A}s = y$. This problem is treated in Johnson and Austria (1983) where it is assumed that \bar{N},\bar{U} are dense triangular matrices. For the sparse case we choose (\bar{N},\bar{U}) to be the solution of the following optimization problem:

$$\min \|\bar{N} - N + \bar{U} - U\|$$

subject to

$$\bar{N}y = \bar{U}s$$

$$\bar{N} \in Z^N, \bar{U} \in Z^U$$

$$\bar{N}_{ii} = 1 \quad \text{for } i = 1, 2, \dots, n \quad (2.4.2)$$

The next theorem gives the solution of (2.4.2); the proof is deferred until Section 3.

Theorem 2.4

The solution of (2.4.2) is given by

$$\bar{\tau}_i = \tau_i - \alpha_i [(S^{X_i} v^{(i)})^T (S^{X_i} v^{(i)})]^+ (S^{X_i} v^{(i)})^T$$

$$\text{for } i = 1, 2, \dots, n \quad (2.4.3)$$

where

$$v^{(i)} = \begin{cases} -s & \text{if } i = 1 \\ (y_1, y_2, \dots, y_{i-1}, \dots, -s_1, \dots, -s_n) & \text{if } i > 1 \end{cases} \quad (2.4.4)$$

$$\tau_1 = (U_{11}, U_{12}, \dots, U_{1n})$$

$$\bar{\tau}_1 = (\bar{U}_{11}, \bar{U}_{12}, \dots, \bar{U}_{1n})$$

$$\tau_i = (N_{i1}, \dots, N_{i,i-1}, U_{ii}, \dots, U_{in}) \quad \text{for } i = 2, 3, \dots, n$$

$$\bar{\tau}_i = (\bar{N}_{i1}, \dots, \bar{N}_{i,i-1}, \bar{U}_{ii}, \dots, \bar{U}_{in}) \quad \text{for } i = 2, 3, \dots, n$$

$$\alpha = Ny - Us$$

$$X_i = \text{Span}\{\bar{z}_i^N, z_i^U\}$$

□

2.5. Algorithm II

This is similar to Algorithm I, except we may take $m = \infty$.

(1) Choose $x^0 \in \mathbb{R}^n$ as an approximation to x^* .

$k \leftarrow 0$.

(2) Evaluate $F(x^0)$ and A^0 , a finite difference (or analytic) approximation to $F'(x^0)$.

(3) Factorize A^0 as $N^0 P^0 A^0 Q^0 = U^0$ using a threshold pivoting strategy, where P^0 and Q^0 are permutation matrices.

(4) Solve $U^k t^k = -N^k P^0 F(x^k)$,

$$s^k = Q^0 t^k.$$

(5) $x^{k+1} = x^k + s^k$; evaluate $F(x^{k+1})$.

If stopping criteria are met, then STOP.

(6) Let $y^k = P^0(F(x^{k+1}) - F(x^k))$ and update the rows of N and U according to Theorem 2.4.

(7) $k \leftarrow k+1$; go to (4).

2.6. Pivoting Considerations

We first note that Updates I and II are stated under the assumption that the pivoting strategy is the identity, i.e., $P^0 A^0 Q^0 = L^0 U^0$ where $P^0 = Q^0 = I$. It should be clear that if P^0 and Q^0 are nontrivial, then we can permute the independent and dependent variables

so that we obtain a system with the trivial pivoting strategy, I, and to which we can then apply the Updates I and II as stated above. This will not be done explicitly here since it merely amounts to an exercise in notation.

Updates I and II attempt to go directly from an approximation (L^k, U^k) -- or (N^k, U^k) -- of the factors of $F'(x^k)$ to an approximation (L^{k+1}, U^{k+1}) -- or (N^{k+1}, U^{k+1}) -- of the factors of $F'(x^{k+1})$. In essence this requires assumptions on the continuity of the factors of $F'(x)$ and on the persistence of the pivoting strategy as we move from one matrix approximation to another. The following two theorems establish the validity of this process.

Theorem 2.6.1

Let $A : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$, and suppose that for some $x^0 \in \mathbb{R}^n$, $A(x^0)$ is nonsingular, and that there exist $\epsilon_0 > 0$ and $\gamma_{ij} \geq 0$ such that

$$|e_1^T [A(x) - A(y)] e_j| \leq \gamma_{ij} \|x - y\|,$$

for $i, j = 1, 2, \dots, n$ and for all $x, y \in \mathbb{N}(x^0, \epsilon_0) = \{z \in \mathbb{R}^n : \|z - x^0\| \leq \epsilon_0\}$. If the LU decomposition without pivoting exists at x^0 , $A(x^0) = L(x^0)U(x^0)$, then there exists $\epsilon > 0$ such that decomposition without pivoting exists at all $x \in \mathbb{N}(x^0, \epsilon)$. Furthermore, there exist constants $c_0, d_0 > 0$ such that

$$\|L(x) - L(x^0)\| \leq c_0 \|x - x^0\| \quad \text{and} \quad \|U(x) - U(x^0)\| \leq d_0 \|x - x^0\|$$

for all $x \in \mathbb{N}(x^0, \epsilon)$. □

Proof: See Dennis and Marwil (1982).

Theorem 2.6.2

Let $A : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$ be continuous and nonsingular at x^* .

For any threshold pivoting strategy, there exists $\eta_T > 0$ such that if $x^0 \in \mathbb{N}(x^*, \eta_T)$ and if (P^0, Q^0) is a pivot sequence for which $P^0 A(x^0) Q^0$ has an LU decomposition without further pivoting, $P^0 A(x^0) Q^0 = L(x^0) U(x^0)$, then $P^0 A(x^*) Q^0$ can be factored without further pivoting. \square

Proof: Similar to Dennis and Marwil (1982), Corollary 3.5.

Note that by the Banach Perturbation Lemma (stated below), which establishes the continuity of L^{-1} as a function of L , we obtain as a corollary of Theorem 2.6.1 the following

Corollary 2.6.3

With the same hypotheses as Theorem 2.6.1, there exists $\bar{c}_0 > 0$ such that $\|N(x) - N(x^0)\| = \|L^{-1}(x) - L^{-1}(x^0)\| \leq \bar{c}_0 \|x - x^0\|$ for all $x \in \mathbb{N}(x^0, \epsilon)$. \square

Theorem 2.6.4 (Banach Perturbation Lemma [15])

Let $A, C \in \mathbb{R}^{n \times n}$ and assume A is invertible with $\|A^{-1}\| \leq \alpha$. If $\|A - C\| \leq \beta$ and $\alpha\beta < 1$, then C is also invertible and

$$\|C^{-1}\| \leq \alpha/(1-\alpha\beta) .$$

□

3. Some Sparsity Relationships

3.1. Non-cancellation Assumption

The updates in Section 2 are defined in terms of the sparsity patterns of L, N and U . The sparsity patterns of these factors are generally not given as input to the algorithm but instead are defined during the factorization process. The sparsity pattern of $F'(x)$ and A^0 is, however input which is given to the algorithms of Section 2; it may be specified by the user or predetermined by some separate subroutine. In this section, some results concerning the sparsity patterns of the factors are established. But first, we need an assumption on the way these patterns are defined during the factorization process.

Consider the reduction of a matrix A to upper triangular form by Gaussian elimination. Without loss of generality, assume here that $P = I$ where P is the permutation matrix representing the pivoting strategy, i.e., $A = LU$, where L is unit lower triangular and U is upper triangular with non-zero diagonal elements. The process takes place in n steps as follows:

$$\begin{aligned} A^{(1)} &= \Gamma_0 A & \text{where } \Gamma_0 &= I \\ A^{(i+1)} &= \Gamma_i A^{(i)} & \text{for } i &= 1, 2, \dots, n-1 \end{aligned} \quad (3.1.1)$$

where

$$\Gamma_i = I + \xi^i e_i^T$$

$$\xi^i \in \mathbb{R}^n, \quad \xi_j^i = 0 \quad \text{for } j \leq i$$

and

$$A^{(n)} = U \tag{3.1.2}$$

Thus

$$N = \Gamma_{n-1} \Gamma_{n-2} \cdots \Gamma_0 = \prod_{i=1}^{n-1} (I + \xi^{n-i} e_{n-i}^T)$$

$$L = \Gamma_1^{-1} \Gamma_2^{-1} \cdots \Gamma_{n-1}^{-1} = \prod_{i=1}^{n-1} (I - \xi^i e_i^T) = I - \sum_{i=1}^{n-1} \xi^i e_i^T. \tag{3.1.3}$$

At the $(i+1)^{\text{st}}$ stage of the reduction, the elementary matrix Γ_i premultiplies $A^{(i)}$ to reduce to zero all elements of the i^{th} column of $A^{(i)}$, which are in rows r for $r > i$. However, unexpected zeroes may arise through cancellation in other columns in positions where, formerly, there were non-zero elements. The non-cancellation assumption says that, in defining the sparsity pattern of $A^{(i+1)}$ at the $(i+1)^{\text{st}}$ stage, accidental zeroes which arise are treated as small non-zero elements. No advantage is taken of unexpected zeros to reduce the density of the matrix $A^{(i+1)}$ and of subsequent matrices $A^{(\ell)}$ for $\ell > i+1$. This is equivalent to defining the sparsity patterns of L and U as the sparsity pattern of LU factors of maximum density that can be generated from all matrices that have the same sparsity pattern as A .

Example 3.1.1

$$A^{(1)} = \begin{bmatrix} 1 & 2 & 1 & 2 \\ 1 & 4 & 2 & 2 \\ 0 & 2 & 9 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\Gamma_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A^{(2)} = \begin{bmatrix} 1 & 2 & 1 & 2 \\ 0 & 2 & 1 & \theta \\ 0 & 2 & 9 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\Gamma_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A^{(3)} = \begin{bmatrix} 1 & 2 & 1 & 2 \\ 0 & 2 & 1 & \theta \\ 0 & 0 & 8 & \theta \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\Gamma_3 = I$$

and

$$A^{(4)} = A^{(3)}$$

$$\text{hence } U = A^{(3)} .$$

The symbol θ is used in the above example to denote zero elements which are treated as non-zeroes for the sake of defining the sparsity patterns of the factors. The acceptable sparsity pattern for U in the above example is given by

$$\begin{bmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \\ 0 & 0 & 0 & * \end{bmatrix} \quad \text{and not by} \quad \begin{bmatrix} * & * & * & * \\ 0 & * & * & 0 \\ 0 & 0 & * & 0 \\ 0 & 0 & 0 & * \end{bmatrix}$$

here * denotes a non-zero element.

The non-cancellation assumption is also extended to the generation of the sparsity pattern of N . We assume that no cancellations occur in the generation of N according to (3.1.3), i.e., the sparsity pattern of N has the maximum density that can be obtained from all possible factors having the same sparsity patterns as Γ_i for $i = 1, 2, \dots, m-1$.

Example 3.1.2

$$A^{(1)} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 2 & 0 & 0 & 3 \end{bmatrix} \quad \Gamma_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{bmatrix}$$

$$A^{(2)} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & -2 & 3 \end{bmatrix} = A^{(3)} \quad \Gamma_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -2 & 1 \end{bmatrix}$$

$$A^{(4)} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad N = \Gamma_3 \Gamma_1 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

The sparsity pattern of N is taken to be

$$\begin{bmatrix} * & 0 & 0 & 0 \\ * & * & 0 & 0 \\ * & 0 & * & 0 \\ * & 0 & * & * \end{bmatrix} \quad \text{and not} \quad \begin{bmatrix} * & 0 & 0 & 0 \\ * & * & 0 & 0 \\ * & 0 & * & 0 \\ 0 & 0 & * & * \end{bmatrix} .$$

Here N_{41} was set to zero by an unexpected cancellation while multiplying the factors Γ_i , so we shall consider $N_{41} \neq 0$ in defining the sparsity pattern of N .

3.2. Sparsity Results

Let

$$Y_1 \equiv \text{Span}\{\bar{z}_1^L, z_1^U\} \quad \text{and} \quad X_1 \equiv \text{Span}\{\bar{z}_1^N, z_1^U\} .$$

The following theorem establishes the basic relationships in the sparsity patterns which we shall need.

Lemma 4.2.1

If $\|\bar{U}^{(1)} - U^{*(1)}\| \leq 2\delta$, then for δ small enough there exists a $\rho > 0$ such that $\rho \|S^{Y_1(\omega^{(1)})}\| \geq \|S^{Y_1(s)}\|$.

Proof:

$$\begin{aligned} T^{Y_1(\omega^{(1)})} &= [\hat{T}^{Y_1(\bar{U}^{(1)})}] [T^{Y_1(s)}], && \text{by Lemma 3.2.2} \\ T^{Y_1(s)} &= [\hat{T}^{Y_1(\bar{U}^{(1)})}]^{-1} [T^{Y_1(\omega^{(1)})}]. \end{aligned} \quad (4.2.1)$$

Now $\|\bar{U}^{(1)} - U^{*(1)}\| \leq 2\delta$ implies $\|\hat{T}^{Y_1(\bar{U}^{(1)})} - \hat{T}^{Y_1(U^{*(1)})}\| \leq 2\delta$ and we know that $\hat{T}^{Y_1(U^{*(1)})}$ is nonsingular. Therefore by the Banach Perturbation Lemma, for δ small enough and $\|\bar{U}^{(1)} - U^{*(1)}\| \leq 2\delta$, there exists $\rho_1 > 0$ such that $[\hat{T}^{Y_1(\bar{U}^{(1)})}]^{-1}$ exists and $0 < \|[\hat{T}^{Y_1(\bar{U}^{(1)})}]^{-1}\| \leq \rho_1$. Substituting into (4.2.1), we get

$$\begin{aligned} \|T^{Y_1(s)}\| &\leq \|[\hat{T}^{Y_1(\bar{U}^{(1)})}]^{-1}\| \cdot \|T^{Y_1(\omega^{(1)})}\| \\ &\leq \rho \|T^{Y_1(\omega^{(1)})}\| \quad \text{where } \rho = \max_i \{\rho_i\}. \end{aligned}$$

The required result now follows by noting that for any $v \in \mathbb{R}^n$, $\|T^\Delta(v)\| = \|S^\Delta(v)\|$ for any subspace $\Delta \subset \mathbb{R}^n$. \square

Denote $\phi_1^* \equiv (L_{11}^*, \dots, L_{1,i-1}^*, U_{1,i}^*, \dots, U_{1n}^*)$ and let $\bar{x} = x + s$.

$$\frac{1}{\mu} \|v-u\| \leq \|F(v) - F(u)\| \leq \mu \|v-u\| \quad \text{for some } \mu > 0 .$$

The following lemma gives a result analogous to (4.1.1), but obtains a tighter bound by exploiting the sparsity pattern.

Lemma 4.1.1

There exist $\epsilon > 0$ such that if $\sigma(u,v) < \epsilon$, then

$$|F_1(v) - F_1(u) - F_1'(x^*)(v-u)| \leq \kappa \sigma(u,v) \|S_1^{Z_1^A}(v-u)\| \quad \text{for some } \kappa > 0 .$$

Proof: Let $y = F(v) - F(u)$ and $s = v - u$.

$$|y_1 - F_1'(x^*)s| = |[F_1'(u+ts) - F_1'(x^*)]s|$$

for some $0 \leq t \leq 1$ by the mean value theorem

$$= |[F_1'(u+ts) - F_1'(x^*)] (S_1^{Z_1^A}(s))| \quad \text{since } Z_1^A = Z_1^{F'(x)}$$

$$\leq \kappa_1 \|u+ts-x^*\| \|S_1^{Z_1^A}(s)\| \quad \text{by assumption (c)}$$

$$\leq \kappa \sigma(u,v) \|S_1^{Z_1^A}(s)\| \quad \text{where } \kappa = \max_1 \{\kappa_1\} .$$

4.2 Convergence Results

Denote $A^* = F'(x^*) = L^* U^*$. We assume, without loss of generality, that the pivoting strategy at x^* is the identity. The following lemma uses notation introduced in Lemma 3.2.2.

$$Y_1 \cap \{v \in \mathbb{R}^n : v^T \omega^{(1)} = y_1\}.$$

Proof of Theorem 2.4

We verify easily that $\bar{v}_1 \in X_1$ and $\bar{v}_1^T v^{(1)} = 0$. But these imply the feasibility of \bar{L}, \bar{U} . Optimality and uniqueness now follow as in the proof of Theorem 2.2.

4. Convergence Analysis of Algorithm I

4.1 Properties of Function $F(\cdot)$

Let $\|\cdot\|$ denote the λ_2 vector norm or the Frobenius matrix norm. Assume F satisfies the following conditions:

- (a) $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is continuously differentiable on an open convex set $\Omega \subset \mathbb{R}^n$.
- (b) There exists $x^* \in \Omega$ such that $F(x^*) = 0$ and $F'(x^*)$ is nonsingular.
- (c) There exists $\bar{\kappa} > 0$ such that $\|F'(x) - F'(x^*)\| \leq \bar{\kappa} \|x - x^*\|$ for $x \in \Omega$. Or, equivalently, we have that there exist $\kappa_i > 0$ for $i = 1, 2, \dots, n$ such that $\|F'_i(x) - F'_i(\bar{x})\| \leq \kappa_i \|x - \bar{x}\|$ for all $x, \bar{x} \in \Omega$.

Denote $\sigma(u, v) \equiv \max\{\|u - x^*\|, \|v - x^*\|\}$. By Lemma 3.1 of Broyden et al. (1973) we then have that there exists $\epsilon > 0$ such that if $\sigma(u, v) < \epsilon$, then

$$\|F(v) - F(u) - F'(x^*)(v-u)\| \leq \bar{\kappa} \sigma(u, v) \|v-u\| \quad (4.1.1)$$

and

where $\hat{T}^{Y_1}(\bar{U}^{(1)})$ is nonsingular since it is an upper triangular matrix with non-zero diagonal elements. Therefore $T^{Y_1}(s) = 0$ which implies $S^{Y_1}(s) = 0$. Since $Z_1^A \subset Y_1$ by Lemma 3.2.1(a), we get $S^{Z_1^A}(s) = 0$. Now

$$\begin{aligned}
 y_1 &= F_1(x + s) - F_1(x) \\
 &= e_1^T F'(x + ts)s && \text{for some } 0 \leq t \leq 1 \text{ by mean} \\
 & && \text{value theorem} \\
 &= e_1^T F'(x + ts) [S^{Z_1^A}(s)] && \text{since } A \text{ and } F'(x) \text{ have same} \\
 & && \text{sparsity pattern} \\
 &= 0 && \text{since } S^{Z_1^A}(s) = 0.
 \end{aligned}$$

Hence $\bar{\phi}^{(1)} \omega^{(1)} = 0 = y_1$ and $\bar{\phi}^{(1)}$ is feasible.

We verify the optimality of $\bar{\phi}_1$ by considering any other feasible vector $\hat{\phi}_1$

$$\begin{aligned}
 \|\bar{\phi}_1 - \hat{\phi}_1\| &= \|[(S^{Y_1}(\omega^{(1)}))^T (S^{Y_1}(\omega^{(1)}))]^+ (y_1 - \hat{\phi}_1 \omega^{(1)}) (S^{Y_1}(\omega^{(1)}))^T] \| \\
 &= \|[(S^{Y_1}(\omega^{(1)}))^T (S^{Y_1}(\omega^{(1)}))]^+ (\hat{\phi}_1 - \bar{\phi}_1) (S^{Y_1}(\omega^{(1)})) (S^{Y_1}(\omega^{(1)}))^T] \| \\
 & && \text{since } \hat{\phi}_1 \omega^{(1)} = y_1 \\
 &\leq \|\hat{\phi}_1 - \bar{\phi}_1\|.
 \end{aligned}$$

Uniqueness of the solution follows from the convexity of the feasible region

since $\tilde{w} \neq 0$. But $A^{(1)}$ is non-singular since A is LU factorizable without pivoting. This is a contradiction. Therefore $\tilde{T}^{Y_1(A^{(1)})}$ must be non-singular.

3.3 Proof of Previous Lemmas

Now that the sparsity patterns of L , N and U have been clearly specified, we proceed to verify the assertions in Theorems 2.1 and 2.4.

Proof of Theorem 2.1

We first check feasibility of the stated solution. Since $\phi_1 \in Y_1$ and $S^{Y_1(\omega^{(1)})} \in Y_1$, then $\bar{\phi}_1 \in Y_1$ by (2.2.7). If $S^{Y_1(\omega^{(1)})} \neq 0$, then from (2.2.7)

$$\begin{aligned} \bar{\phi}_1 \omega^{(1)} &= \phi_1 \omega^{(1)} + [(S^{Y_1(\omega^{(1)})})^T (S^{Y_1(\omega^{(1)})})] (y_1 - \phi_1 \omega^{(1)}) \\ &\quad (S^{Y_1(\omega^{(1)})})^T (S^{Y_1(\omega^{(1)})}) \\ &= \phi_1 \omega^{(1)} + (y_1 - \phi_1 \omega^{(1)}) \\ &= y_1 . \end{aligned}$$

If $S^{Y_1(\omega^{(1)})} = 0$, then $\bar{\phi}_1 = \phi_1$ and $\bar{\phi}_1 \omega^{(1)} = 0$. This means that $\bar{\phi}^{(1)}$ is feasible only if $y_1 = 0$. But $S^{Y_1(\omega^{(1)})} = 0$ implies $T^{Y_1(\omega^{(1)})} = 0$. Thus

$$0 = T^{Y_1(\omega^{(1)})} = T^{Y_1(\bar{U}^{(1)}s)} = [\tilde{T}^{Y_1(\bar{U}^{(1)})}] [T^{Y_1(s)}]$$

Proof:

$$\begin{aligned} S^{X_1(A^{(1)})} s &= [\bar{S}^{X_1(A^{(1)})}] s \\ &= [\tilde{S}^{X_1(A^{(1)})}] s \end{aligned}$$

since $\tilde{S}^{X_1(A^{(1)})} = \bar{S}^{X_1(A^{(1)})}$ by Lemma 3.2.1(b)

$$= [\tilde{S}^{X_1(A^{(1)})}] [S^{X_1(s)}].$$

It now follows trivially that $T^{X_1(A^{(1)})} s = [\tilde{T}^{X_1(A^{(1)})}] [T^{X_1(s)}]$. Now let $\chi(X_1) = \{i_1, i_1, \dots, i_m\}$ and assume $\tilde{T}^{X_1(A^{(1)})}$ is singular. Then there exists a $\tilde{w} \in \mathbb{R}^m$ such that $\tilde{w} \neq 0$ and

$$\tilde{v}^T = \tilde{w}^T [\tilde{T}^{X_1(A^{(1)})}] = 0$$

Define $w \in \mathbb{R}^n$ by

$$w_k = \begin{cases} \tilde{w}_r & \text{if } k \in \chi(X_1) \text{ and } k = i_r \\ 0 & \text{otherwise.} \end{cases}$$

Let $v = w^T A^{(1)}$. If $s \in \chi(X_1)$, $s = i_q$ then

$$v_s = \sum_{r \in \chi(X_1)} w_r A_{rs}^{(1)} = \tilde{v}_q = 0,$$

and if $s \notin \chi(X_1)$, then by Lemma 3.2.1(b), $A_{rs}^{(1)} = 0$ for all $r \in \chi(X_1)$ and hence $v_s = 0$. Therefore $v = w^T A^{(1)} = 0$ where $w \neq 0$

Now eliminating the zero rows and columns j for $j \notin Y_1$ we get

$$T^{Y_1}(U^{(i)}_s) = [\tilde{T}^{Y_1}(U^{(i)})] [T^{Y_1}(s)] .$$

Lemma 3.2.3

If

$$A^{(i)} = \begin{bmatrix} A_{11} & & & & & & & & A_{1,n} \\ & \cdot & & & & & & & \cdot \\ & \cdot & & & & & & & \cdot \\ & \cdot & & & & & & & \cdot \\ A_{i-1,1} & \cdot & \cdot & \cdot & A_{i-1,i-1} & \cdot & \cdot & \cdot & A_{i-1,n} \\ & & & & & & & & -1 \\ & & & & & & & & -1 \\ & & & & & & & & \bigcirc \\ & & & & & & & & \cdot \\ & & & & & & & & \cdot \\ & & & & & & & & \cdot \\ & & & & & & & & -1 \\ & & & & & & & & \bigcirc \\ & & & & & & & & -1 \end{bmatrix} ,$$

then

$$S^{X_1}(A^{(i)}_s) = [\tilde{S}^{X_1}(A^{(i)})] [S^{X_1}(s)] ,$$

and

$$T^{X_1}(A^{(i)}_s) = [\tilde{T}^{X_1}(A^{(i)})] [T^{X_1}(s)] .$$

Moreover $\tilde{T}^{X_1}(A^{(i)})$ is non-singular.

Lemma 3.2.2

If

$$U^{(1)} = \begin{bmatrix} U_{11} & U_{12} & \dots & & U_{1,n} \\ & \cdot & & & \\ & & \cdot & & \\ & & & U_{i-1,i-1} & \dots & U_{i-1,n} \\ & \bigcirc & & & & \\ & & & & 1 & \bigcirc \\ & & & & & \cdot \\ & & & & & \cdot \\ & & & & & 1 \end{bmatrix},$$

then

$$S^{Y_1(U^{(1)})}(s) = [\tilde{S}^{Y_1(U^{(1)})}] [S^{Y_1}(s)] ,$$

and

$$T^{Y_1(U^{(1)})}(s) = [\tilde{T}^{Y_1(U^{(1)})}] [T^{Y_1}(s)] .$$

Proof:

$$S^{Y_1(U^{(1)})}(s) = [\bar{S}^{Y_1(U^{(1)})}] s \quad \text{by definition}$$

$$= [\tilde{S}^{Y_1(U^{(1)})}] s$$

$$\text{since } \tilde{S}^{Y_1(U^{(1)})} = \bar{S}^{Y_1} U^{(1)} \text{ by Lemma 3.2.1(a)}$$

$$= [\tilde{S}^{Y_1(U^{(1)})}] [S^{Y_1}(s)] .$$

the projection operator which sets to zero those elements of the matrix with row index i for $i \notin \chi(\Delta)$, i.e., for $M \in \mathbb{R}^{n \times n}$,

$$[\bar{S}^\Delta(M)]_{\alpha\beta} = \begin{cases} 0, & \text{if } \alpha \notin \chi(\Delta) \\ M_{\alpha\beta}, & \text{otherwise} \end{cases}.$$

Similarly define $\tilde{S}^\Delta : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n}$

$$[\tilde{S}^\Delta(M)]_{\alpha\beta} = \begin{cases} 0, & \text{if } \alpha \notin \chi(\Delta) \text{ or } \beta \notin \chi(\Delta) \\ M_{\alpha\beta}, & \text{otherwise} \end{cases}.$$

We also define the associated collapsing operators $T^\Delta, \tilde{T}^\Delta$. If

$$\chi(\Delta) = \{i_1, i_2, \dots, i_m\}$$

where $m = \text{rank } \Delta$ and $i_1 < i_2 < \dots < i_m$

then define $T^\Delta : \mathbb{R}^n \rightarrow \mathbb{R}^m$ by

$$T^\Delta(v) = (v_{i_1}, v_{i_2}, \dots, v_{i_m}) \text{ where } v = (v_1, v_2, \dots, v_n)$$

and $\tilde{T}^\Delta : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{m \times m}$ by

$$\tilde{T}^\Delta(M) = \begin{bmatrix} M_{i_1, i_1} & \dots & M_{i_1, i_m} \\ \vdots & & \vdots \\ M_{i_m, i_1} & \dots & M_{i_m, i_m} \end{bmatrix}$$

$$= \text{Span}\{Z_j^A^{(p+1)} : j \in \chi(\bar{Z}_1^{L^p}) \cup \{i, p\}\}$$

$$= \text{Span}\{Z_j^A^{(p+1)} : j \in \chi(Z_1^{L^p})\}.$$

Therefore (c) is true for $k = p$.

Similarly

$$X_1^p = \text{Span}\{\bar{Z}_1^{N^p}, Z_1^A^{(p+1)}\}$$

$$= \text{Span}\{X_1^{(p-1)}, X_p^{(p-1)}\}$$

$$= \text{Span}\{Z_j^A : j \in \chi(Z_1^{N^{p-1}}) \cup \chi(Z_1^{N^p})\} \quad \text{by induction hypothesis}$$

$$= \text{Span}\{Z_j^A : j \in \chi(Z_1^{N^p})\}$$

which verifies (d) for $k = p$.

Now deduce by induction that (c) and (d) are true for $k = 0, 1, \dots, n-1$ and hence (a) and (b) are true. \square

We now prove some sparsity results which will be used later in the convergence analyses of Sections 4 and 5.

Notation

We already have $S^\Delta : \mathbb{R}^n \rightarrow \mathbb{R}^n$, for some subspace Δ , defined as the projection operator into the subspace Δ . This notation is extended to matrices as follows: Let $\tilde{S}^\Delta : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n}$ be

Case (iii): $i > p$ and $e_p \in Z_i^{A(p)}$

Since $e_p \in Z_i^{A(p)}$ we have

$$\begin{aligned}
 Y_i^p &= \text{Span}\{\bar{Z}_i^{L^p}, Z_i^{A(p+1)}\} \\
 &= \text{Span}\{\text{Span}\{\bar{Z}_i^{L^{p-1}}, e_p\}, \text{Span}\{Z_i^{A(p)}, Z_p^{A(p)}\} \cap \{x : x_p = 0\}\} \\
 &= \text{Span}\{\text{Span}\{\bar{Z}_i^{L^{p-1}}, e_p\}, \text{Span}\{Z_i^{A(p)}, Z_p^{A(p)}\}\} \\
 &= \text{Span}\{\text{Span}\{\bar{Z}_i^{L^{p-1}}, Z_i^{A(p)}\}, \text{Span}\{e_p, Z_p^{A(p)}\}\} \\
 &= \text{Span}\{\text{Span}\{Z_j^{A(p)} : j \in \chi(Z_i^{L^{p-1}})\}, Z_p^{A(p)}\}
 \end{aligned}$$

by induction hypothesis and $e_p \in Z_p^{A(p)}$

$$\begin{aligned}
 &= \text{Span}\{\text{Span}\{Z_j^{A(p)} : j \in \chi(\bar{Z}_i^{L^{p-1}})\}, Z_i^{A(p)}, Z_p^{A(p)}\} \\
 &= \text{Span}\{\text{Span}\{Z_j^{A(p+1)} : j \in \chi(\bar{Z}_i^{L^p})\}, \text{Span}\{Z_i^{A(p)}, Z_p^{A(p)}\}\}
 \end{aligned}$$

since $j \in \chi(\bar{Z}_i^{L^p})$ implies $j \leq p$

$$= \text{Span}\{\text{Span}\{Z_j^{A(p+1)} : j \in \chi(\bar{Z}_i^{L^p})\}, \text{Span}\{Z_i^{A(p+1)}, Z_p^{A(p+1)}\}\}$$

Case (ii): $i > p$ and $e^p \notin Z_1^{A(p)}$

Since $e^p \notin Z_1^{A(p)}$ we have

$$e_i^{T_A(p+1)} = e_i^{T_A(p)}, e_i^{T_N^p} = e_i^{T_N^{p-1}}, e_i^{T_L^p} = e_i^{T_L^{p-1}}$$

and

$$Z_1^{A(p+1)} = Z_1^{A(p)}, Z_1^{N^p} = Z_1^{N^{p-1}}, Z_1^{L^p} = Z_1^{L^{p-1}}.$$

Then

$$\begin{aligned} Y_1^p &= \text{Span}\{\bar{Z}_1^{L^p}, Z_1^{A(p)}\} \\ &= \text{Span}\{\bar{Z}_1^{L^{p-1}}, Z_1^{A(p)}\} \\ &= \text{Span}\{Z_j^{A(p)} : j \in \chi(Z_1^{L^{p-1}})\} \text{ by induction hypothesis} \\ &= \text{Span}\{\{Z_j^{A(p)} : j \in \chi(\bar{Z}_1^{L^{p-1}})\}, Z_1^{A(p)}\} \\ &= \text{Span}\{\{Z_j^{A(p+1)} : j \in \chi(\bar{Z}_1^{L^p})\}, Z_1^{A(p+1)}\}, \\ &\quad \text{since } j \in \chi(\bar{Z}_1^{L^{p-1}}) \text{ implies } j \leq \alpha \text{ and } Z_j^{A(p)} = Z_j^{A(p-1)} \\ &= \text{Span}\{Z_j^{A(p+1)} : j \in \chi(Z_1^{L^p})\}. \end{aligned}$$

Therefore (c) is true for $k = p$. Similarly (d) is true for $k = p$.

$$Y_1^0 = Z_1^A = \text{Span}\{Z_j^A : j \in \chi(Z_1^{L^0})\},$$

$$X_1^0 = Z_1^A = \text{Span}\{Z_j^A : j \in \chi(Z_1^{N^0})\}.$$

Thus (c) and (d) are true for $k = 0$.

Assume (c) and (d) are true for $k = 0, 1, \dots, p-1 < n$.

We shall consider three cases. Note that $\xi_\alpha^p = 0$ for $\alpha \leq p$ where ξ^j is defined in (3.1.2). Therefore

$$e_\alpha^{T_A(p+1)} = e_\alpha^{T_A(p)}, \quad e_\alpha^{T_N^p} = e_\alpha^{T_N^{p-1}}, \quad e_\alpha^{T_L^p} = e_\alpha^{T_L^{p-1}}, \quad \text{for } \alpha \leq p$$

and

$$Z_\alpha^{A(p+1)} = Z_\alpha^{A(p)}, \quad Z_\alpha^{N^p} = Z_\alpha^{N^{p-1}}, \quad Z_\alpha^{L^p} = Z_\alpha^{L^{p-1}}, \quad \text{for } \alpha \leq p$$

Case (i): $i \leq p$

$$\begin{aligned} Y_i^p &= \text{Span}\{\bar{Z}_i^{L^p}, Z_i^{A(p+1)}\} \\ &= \text{Span}\{\bar{Z}_i^{L^{p-1}}, Z_i^{A(p)}\} \\ &= \text{Span}\{Z_j^{A(p)} : j \in \chi(Z_i^{L^{p-1}})\} \quad \text{by induction hypothesis} \\ &= \text{Span}\{Z_j^{A(p+1)} : j \in \chi(Z_i^{L^p})\} \quad \text{since } j \in \chi(Z_i^{L^{p-1}}) \text{ implies } j \leq p. \end{aligned}$$

Therefore (c) is true for $k = p$. Similarly (d) is true for $k = p$.

Theorem 3.2.1

$$(a) \quad Y_1 = \text{Span}\{Z_j^U : j \in \chi(Z_1^L)\} ,$$

$$(b) \quad X_1 = \text{Span}\{Z_j^A : j \in \chi(Z_1^N)\} .$$

Proof: Let

$$N^j = \Gamma_j \Gamma_{j-1} \cdots \Gamma_0 , \quad 0 \leq j \leq n-1$$

$$L^j = \Gamma_0^{-1} \Gamma_1^{-1} \cdots \Gamma_j^{-1} , \quad 0 \leq j \leq n-1$$

$$Y_1^j = \text{Span}\{Z_1^{L^j}, Z_1^{A^{(j+1)}}\}$$

$$X_1^j = \text{Span}\{Z_1^{N^j}, Z_1^{A^{(j+1)}}\} .$$

Then $Y_1 = Y_1^{n-1}$ and $X_1 = X_1^{n-1}$.

We shall show by induction that

$$(c) \quad Y_1^k = \text{Span}\{Z_j^{A^{(k+1)}} : j \in \chi(Z_1^{L^k})\} , \quad k = 0, 1, \dots, n-1$$

and

$$(d) \quad X_1^k = \text{Span}\{Z_j^A : j \in \chi(Z_1^{N^k})\} , \quad k = 0, 1, \dots, n-1 .$$

The statement of the theorem is (c) and (d) for $k = n-1$.

Consider the case $k = 0$

Lemma 4.2.2

There exist $\varepsilon > 0$, $\delta > 0$ such that if $\sigma(x, \bar{x}) < \varepsilon$ and $\|\bar{U}^{(i)} - U^{*(i)}\| < 2\delta$, then there exists a $\rho > 0$ (depending on δ) such that

$$\|\bar{\phi}_i - \phi_i^*\|^2 \leq \|\phi_i - \phi_i^*\|^2 + 2\rho^2 [(\kappa\sigma)^2 + \|L_i^*\|^2 \|\bar{U}^{(i)} - U^{*(i)}\|^2]$$

for $i = 1, 2, \dots, n$. Here L_i^* denotes the i^{th} row of L^* and $\sigma = \sigma(x, \bar{x})$.

Proof: Let $\hat{w}^{(i)} = S^{Y_i(\omega^{(i)})}$. Then from (2.2.6) we obtain

$$\bar{\phi}_i - \phi_i^* = (\phi_i - \phi_i^*) \left[I - \frac{\hat{w}^{(i)} \hat{w}^{(i)T}}{\hat{w}^{(i)T} \hat{w}^{(i)}} \right] + (y_i - \phi_i^* \hat{w}^{(i)}) \frac{\hat{w}^{(i)T}}{\hat{w}^{(i)T} \hat{w}^{(i)}}. \quad (4.2.2)$$

Now

$$\begin{aligned} \|(\phi_i - \phi_i^*) \left[I - \frac{\hat{w}^{(i)} \hat{w}^{(i)T}}{\hat{w}^{(i)T} \hat{w}^{(i)}} \right]\| &= \|\phi_i - \phi_i^*\|^2 - \frac{\|(\phi_i - \phi_i^*) \hat{w}^{(i)}\|^2}{\|\hat{w}^{(i)}\|^2} \\ &\leq \|\phi_i - \phi_i^*\|^2 \end{aligned} \quad (4.2.3)$$

and

$$\begin{aligned}
& \|(y_1 - \phi_1^* \hat{w}^{(1)}) \frac{\hat{w}^{(1)T}}{\hat{w}^{(1)T} \hat{w}^{(1)}}\| \\
&= \|[(y_1 - A_1^* s) + (\phi_1^* U^{*(1)} s - \phi_1^* \hat{w}^{(1)})]\| \frac{\hat{w}^{(1)T}}{\hat{w}^{(1)T} \hat{w}^{(1)}} \\
&\quad \text{where } A_1^* \text{ is the } i^{\text{th}} \text{ row of } A^* \\
&\leq \frac{\|y_1 - A_1^* s\|}{\|\hat{w}^{(1)}\|} + \frac{\|L_1^*(\bar{U}^{(1)} - U^{*(1)})_s\|}{\|\hat{w}^{(1)}\|} . \tag{4.2.4}
\end{aligned}$$

Now using Lemmas 4.1 and 4.2.1 we get

$$\frac{\|y_1 - A_1^* s\|}{\|\hat{w}^{(1)}\|} \leq \frac{\kappa \sigma(x, \bar{x}) \|S^{Z_1^A}(s)\|}{\rho^{-1} \|S^{Y_1}(s)\|} \leq \kappa \rho \sigma \quad \text{since } Z_1^A \subset Y_1 \tag{4.2.5}$$

and using Lemma 4.2.1, we get

$$\begin{aligned}
\frac{\|L_1^*(\bar{U}^{(1)} - U^{*(1)})_s\|}{\|\hat{w}^{(1)}\|} &\leq \frac{\|L_1^*\| \|S^{Y_1}(\bar{U}^{(1)} - U^{*(1)})\| \|S^{Y_1}(s)\|}{\rho^{-1} \|S^{Y_1}(s)\|} \\
&\leq \rho \|L_1^*\| \|\bar{U}^{(1)} - U^{*(1)}\| . \tag{4.2.6}
\end{aligned}$$

Substituting (4.2.5) and (4.2.6) into (4.2.4) and then substituting this result with (4.2.3) into (4.2.2) we get

$$\|\bar{\phi}_1 - \phi_1^*\|^2 = \|(\phi_1 - \phi_1^*) \left(I - \frac{\hat{w}^{(1)} \hat{w}^{(1)T}}{\hat{w}^{(1)} \hat{w}^{(1)T}} \right)\|^2 + \|(y_1 - \phi_1^* w^{(1)}) \frac{\hat{w}^{(1)T}}{\hat{w}^{(1)} \hat{w}^{(1)}}\|^2$$

by orthogonality of vectors on right hand side of (4.2.2)

$$\leq \|\phi_1 - \phi_1^*\|^2 + [\kappa\rho\sigma + \rho\|L_1^*\| \|\bar{U}^{(1)} - U^{*(1)}\|]^2$$

$$\leq \|\phi_1 - \phi_1^*\|^2 + 2\rho^2 [(\kappa\sigma)^2 + \|L_1^*\|^2 \|\bar{U}^{(1)} - U^{*(1)}\|^2]$$

$$\text{since } \|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2. \quad \square$$

The next lemma converts this result into a convenient form.

Lemma 4.2.3.

There exist $\delta > 0$ and $\epsilon > 0$ such that if $\sigma(x, \bar{x}) < \epsilon$ and $\|\bar{U}^{(1)} - U^{*(1)}\| < 2\delta$ then there exist constants $k_1, k_2 > 0$ (depending on δ) such that

$$\|\bar{D}_1 - D_1^*\| \leq k_1 \|D_1 - D_1^*\| + k_2 \sigma, \quad \text{for } 1 \leq i \leq n$$

where

$$D_1 = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_1 \end{bmatrix} \quad \bar{D}_1 = \begin{bmatrix} \bar{\phi}_1 \\ \bar{\phi}_2 \\ \vdots \\ \bar{\phi}_1 \end{bmatrix} \quad \text{and} \quad D^* = \begin{bmatrix} \phi_1^* \\ \phi_2^* \\ \vdots \\ \phi_1^* \end{bmatrix}$$

Proof: $\|\bar{U}^{(1)} - \bar{U}^{*(1)}\| \leq 2\delta$ implies $\|\bar{U}^{(j)} - \bar{U}^{*(j)}\| \leq 2\delta$

for $j = 1, 2, \dots, i$. Hence by Lemma 4.2.3,

$$\|\bar{\phi}_j - \phi_j^*\| \leq \|\phi_j - \phi_j^*\|^2 + 2\rho^2 \kappa^2 \sigma^2 + 2\|L_j^*\|^2 \rho^2 \|\bar{U}^{(j)} - U^{*(j)}\|^2,$$

$$\text{for } j = 1, 2, \dots, i. \quad (4.2.7)$$

Let

$$R_i = \sum_{j=1}^i \|\bar{\phi}_j - \phi_j^*\|^2 = \|\bar{D}_i - D_i^*\|^2$$

Noting that $\|\bar{U}^{(j)} - U^{*(j)}\|^2 \leq R_{j-1}$, we get from (4.2.7),

$$R_j - R_{j-1} \leq r_j + \bar{k} R_{j-1} \quad (4.2.8)$$

where $r_0 = 0$ and $r_j = \|\phi_j - \phi_j^*\|^2 + 2\rho^2 \kappa^2 \sigma^2$ for $j > 0$ and $\bar{k} > 2\|L_j^*\|^2 \rho^2$

for $j \geq 1$.

Now iterating (4.2.8), we obtain

$$\begin{aligned} R_j &\leq r_j + (1+\bar{k}) R_{j-1} \\ &\leq \sum_{j=0}^i (1+\bar{k})^j r_{i-j} \\ &\leq K \sum_{j=1}^i r_{i-j}, \quad \text{where } K > \max_j \{(1+\bar{k})^j\}. \end{aligned}$$

That is

$$\begin{aligned} \|\bar{D}_1 - D_1^*\|^2 &\leq K \left\{ \sum_{j=0}^1 \|\phi_j - \phi_j^*\|^2 + 2n \rho^2 \kappa^2 \sigma^2 \right\} \\ &\leq k_1^2 \|D_1 - D_1^*\|^2 + k_2^2 \sigma^2 \end{aligned}$$

$$\text{where } k_1^2 = K \text{ and } k_2^2 = 2n \rho^2 \kappa^2 K .$$

Now using the fact that $a, b, c \geq 0$ and $a^2 \leq b^2 + c^2$ imply $a < b + c$, we get the desired result.

Theorem 4.2.4 (Linear Convergence)

Given m a fixed positive integer and $r \in (0, 1)$, there exists $\varepsilon > 0$ such that if $\|x^0 - x^*\| \leq \varepsilon$ then Algorithm I produces a sequence x^k for $k = 0, 1, 2, \dots$, which satisfies

$$\|x^{k+1} - x^*\| \leq r \|x^k - x^*\|, \quad \text{for } k = 0, 1, 2, \dots$$

Proof: By Theorems 2.6.1 and 2.6.2 we can choose $\varepsilon_1 > 0$ such that if $\|x - x^*\| \leq \varepsilon$ then there exists a constant $c_0 > 0$ such that

$$\|L(x) - L^* + U(x) - U^*\| \leq c_0 \|x - x^*\| \quad (4.2.8)$$

and if $P_0 A(x) Q_0$ is LU factorizable for some pivoting strategy (P_0, Q_0) , then $P_0 A^* Q_0$ is LU factorizable. Without loss of generality, we shall assume that the pivoting strategy, (P_0, Q_0) , is the identity throughout. Now choose $0 < \varepsilon < \varepsilon_1$ and $\delta > 0$ to satisfy

$$\gamma(1+r) [\bar{k}\epsilon + 2\delta(\delta+\gamma)] \leq r(1-r) \quad \text{where } \gamma = \max\{\|L^*\|, \|U^*\|\} \quad (4.2.9)$$

$$\{[\max(1, k_1)]^m c_0 + k_2 \left(\frac{1 - k_1^m}{1 - k_2}\right)\} \epsilon \leq 2\delta \quad (4.2.10)$$

Note since k_1, k_2 depend on δ we choose δ first and then take ϵ small enough so (4.2.9) and (4.2.10) are satisfied.

Since

$$\begin{aligned} x^{k+1} - x^* &= x^k - x^* - (A^k)^{-1} F(x^k) \\ &= (A^k)^{-1} \{ [F(x^k) - F(x^*) - F'(x^*) (x^k - x^*)] \\ &\quad + [A^k - F'(x^*)] (x^k - x^*) \} \end{aligned}$$

we have

$$\begin{aligned} \|x^{k+1} - x^*\| &\leq \|(A^k)^{-1}\| \{ \|F(x^k) - F(x^*) - F'(x^*) (x^k - x^*)\| \\ &\quad + \|A^k - F'(x^*)\| \|x^k - x^*\| \} \quad (4.2.11) \end{aligned}$$

Using the notation $F'(x) = A(x) = L(x) U(x)$, we note that if $\|x - x^*\| < \epsilon < \epsilon_1$, then $\|L(x) - L^* + U(x) - U^*\| \leq C_0 \epsilon \leq 2\delta$ by (4.2.8) and (4.2.10). Hence $\|L(x) - L^*\| \leq 2\delta$ and $\|U(x) - U^*\| \leq 2\delta$ and $\|L(x)\| \leq \|L(x) - L^*\| + \|L^*\| \leq 2\delta + \gamma$. Hence

$$\begin{aligned}
\|A(x) - A^*\| &= \|L(x) U(x) - L^* U^*\| \\
&\leq \|L(x)\| \|U(x) - U^*\| + \|L(x) - L^*\| \|U^*\| \\
&\leq (2\delta + \gamma)\delta + \delta\gamma = 2\delta(\delta + \gamma) . \tag{4.2.12}
\end{aligned}$$

Now using the Banach Perturbation Lemma (Theorem 2.6.4) and $2\delta(\delta + \gamma) \gamma(1+r) \leq 1$ from (4.2.9), we obtain

$$\| [A(x)]^{-1} \| \leq \gamma \left(\frac{1+r}{1-r} \right) . \tag{4.2.13}$$

We now use a triple induction to show that $H(k)$ is true for $k = 0, 1, 2, \dots$, where

$$\begin{aligned}
H(k) &\equiv \{ \|x^{k+1} - x^*\| \leq r \|x^k - x^*\| \quad \text{and} \quad \|D_n^k - D_n^*\| \leq 2\delta \} \\
&\quad \text{with } D_1^k \text{ as defined in Lemma 4.2.3 .} \tag{4.2.14}
\end{aligned}$$

First we show $H(0)$ is true.

By (4.2.8) and (4.2.10), $\|D_n^0 - D_n^*\| \leq 2\delta$ and from (4.2.11) and (4.2.13)

$$\begin{aligned}
\|x^1 - x^*\| &\leq \gamma \left(\frac{1+r}{1-r} \right) \{ \|F(x^0) - F(x^*) - F'(x^*) (x^0 - x^*)\| + 2\delta(\delta + \gamma) \|x^0 - x^*\| \} \\
&\leq \gamma \left(\frac{1+r}{1-r} \right) \{ \bar{\kappa}\epsilon + 2\delta(\delta + \gamma) \} \|x^0 - x^*\| \quad \text{using (4.1.1)} \\
&\leq r \|x^0 - x^*\| \quad \text{using (4.2.9) .}
\end{aligned}$$

Hence $H(0)$ is true.

Now assume $H(k)$ is true for $k = 0, 1, \dots, pm-1$. We shall first verify that $H(pm)$ is true and then, by induction, that $H(k)$ is true for $k = pm+1, pm+2, \dots, (p+1)m-1$.

Since $\|x^{k+1} - x^*\| \leq r\|x^k - x^*\|$ for $k = 0, 1, \dots, pm-1$, we have that $\|x^{pm} - x^*\| \leq \varepsilon$. Now since $A^{pm} = A(x^{pm})$ by definition of m , we can show $H(pm)$ is true in exactly the same way as was done for $H(0)$ above.

Now assume $H(k)$ is true for $k = pm, pm+1, \dots, pm+j-1$ with $j < m$. Denoting $q = pm+j-1$, we first show $\|D_n^q - D_n^*\| \leq \delta$ by induction on row i .

Note that by induction hypothesis on k we have

$$\|D_n^k - D_n^*\| \leq 2\delta, \quad \text{for } 0 \leq k \leq q.$$

Hence

$$\|D_i^k - D_i^*\| \leq 2\delta, \quad \text{for } 0 \leq k \leq q \text{ and } 1 \leq i \leq n.$$

Thus

$$\|(U^k)^{(i)} - (U^*)^{(i)}\| \leq 2\delta \quad \text{for } 0 \leq k \leq q \text{ and } 1 \leq i \leq n. \quad (4.2.15)$$

Therefore by Lemma 4.2.13

$$\|D_i^{k+1} - D_i^*\| \leq k_1 \|D_i^k - D_i^*\| + k_2 \sigma \quad \text{for } 0 \leq k < q-1, 1 \leq i \leq n \quad (4.2.16)$$

For $i = 1$,

$$\begin{aligned}
\|D_1^q - D_1^*\| &\leq k_1 \|D_1^{q-1} - D_1^*\| + k_2 \delta \\
&\leq k_1^{q-pm} \|D_1^{pm} - D_1^*\| + k_2 (\sigma_q + k_1 \sigma_{q-1} + \dots + k_1^{q-pm} \sigma_{pm}) \\
&\leq [\max(1, k_1)]^m c_0 + k_2 \left(\frac{1 - k_1^m}{1 - k_1} \right) \epsilon \\
&\leq 2\delta \quad \text{by (4.2.10)} .
\end{aligned}$$

Now assume $\|D_\lambda^q - D_\lambda^*\| \leq 2\delta$ for $1 \leq \lambda < i$. Then $\|(U^q)^{(1)} - (U^*)^{(1)}\| \leq 2\delta$ and by Lemma 4.2.3

$$\begin{aligned}
\|D_i^q - D_i^*\| &\leq k_1 \|D_i^{q-1} - D_i^*\| + k_2 \sigma \\
&\leq [\max(1, k_1)]^m c_0 + k_2 \left(\frac{1 - k_1^m}{1 - k_1} \right) \epsilon \quad \text{by iteration} \\
&\leq 2\delta \quad \text{by (4.2.10)} .
\end{aligned}$$

Hence by induction on row i we now deduce $\|D_n^q - D_n^*\| \leq 2\delta$. Now as in (4.2.12) and (4.2.13) above, we deduce that

$$\|A^q - A^*\| \leq 2\delta(\delta + \gamma) \quad \text{and} \quad \|(A^q)^{-1}\| \leq \gamma \left(\frac{1+r}{1-r} \right)$$

since $\|L^q - L^* + U^q - U^*\| = \|D_n^q - D_n^*\|$. Hence from (4.2.11)

$$\begin{aligned}
\|x^{q+1} - x^*\| &\leq \gamma \left(\frac{1+r}{1-r} \right) \{ \|F(x^q) - F(x^*) - F'(x^*)(x^q - x^*)\| \\
&\quad + 2\delta(\delta+\gamma) \|x^q - x^*\| \} \\
&\leq \gamma \left(\frac{1+r}{1-r} \right) (\tilde{k}\epsilon + 2\delta(\delta+\gamma)) \|x^q - x^*\| \\
&\leq r \|x^q - x^*\| \quad \text{from (4.2.9) .}
\end{aligned}$$

Therefore $H(q)$ is true and hence by induction $H(k)$ is true for $k = pm, pm+1, \dots, pm+m-1$. Completing the induction we deduce that $H(k)$ is true for $k = 0, 1, 2, \dots$. \square

Theorem 4.2.5

With the same hypothesis as in Theorem 4.2.4, the sequence x^k for $k = 0, 1, 2, \dots$ is m -step Q -superlinearly convergent to x^* .

Proof: Setting $p = \alpha m$ for $\alpha = 0, 1, 2, \dots$ we have

$$\frac{\|(A^p - A^*)(x^{p+1} - x^p)\|}{\|x^{p+1} - x^p\|} \leq \|A^p - A^*\| \rightarrow 0 \quad \text{as } p \rightarrow \infty .$$

By Theorem 3.1 of Dennis and Moré (1974), this implies

$$\frac{\|x^{p+1} - x^*\|}{\|x^p - x^*\|} \rightarrow 0 \quad \text{as } p \rightarrow \infty . \quad \square$$

Convergence Analysis for Algorithm II

Update II is a sparse version of an update presented in Johnson and Stria (1983). There it was demonstrated that if a certain bounded-deterioration property was satisfied then the algorithm is linearly convergent. The following theorem is paraphrased from Johnson and Stria (1983) to conform with our notation.

Theorem 5.1.1

If $F(\cdot)$ possesses the properties of Section 4.1 and the Update II satisfies

$$\begin{aligned} \|U^{k+1} - U^* + N^{k+1} - N^*\| \leq [1 + \alpha_1 \sigma(x^k, x^{k+1})] \|U^k - U^* + N^k - N^*\| \\ + \alpha_2 \sigma(x^k, x^{k+1}) \end{aligned} \quad (5.1.1)$$

then there exist $\varepsilon = \varepsilon(r)$ and $\delta = \delta(r)$ such that if

$$\begin{aligned} \|x^0 - x^*\| &\leq \varepsilon \\ \|U^0 - U^* + N^0 - N^*\| &\leq \delta \\ (N^k, U^k) &\text{ is defined by Update II for } k > 0, \end{aligned} \quad (5.1.2)$$

then the sequence $\{x^k\}$ defined by Algorithm II satisfies

$$\|x^{k+1} - x^*\| \leq r \|x^k - x^*\|, \quad \text{for } k = 0, 1, 2, \dots$$

Moreover, $\|N^k\|$, $\|U^k\|$, $\|(N^k)^{-1}\|$ and $\|(U^k)^{-1}\|$ are uniformly bounded. \square

Hence, to establish linear convergence, it would be sufficient to verify that the bounded-deterioration property (5.1.1) is true for the sparse Update II.

Lemma 5.1.2

There exists a constant $k_0 > 0$ such that if $\sigma(x, \bar{x}) \leq \epsilon$ for ϵ small enough, $\bar{x} = x + s$ and $y = F(\bar{x}) - F(x)$, then

$$\|N_i^* y - U_i^* s\| \leq k_0 \sigma(x, \bar{x}) \|S_i^{X_1}(s)\|$$

where N_i^* , U_i^* are the i^{th} rows of N^* , U^* respectively.

Proof:

$$\|N_i^* y - U_i^* s\| = \|N_i^* (y - A^* s)\|$$

$$= \left\| \sum_{j \in \eta} N_{ij}^* (y_j - A_j^* s) \right\|$$

where $\eta = \chi(Z_i^N)$ and A_j^* is the j^{th} row of A^*

$$\leq k_1 \left\| \sum_{j \in \eta} (y_j - A_j^* s) \right\| \quad \text{where } k_1 = \max_{j \in \eta} \{|N_{ij}^*|\}$$

$$\leq k_1 \sum_{j \in \eta} \kappa \sigma(x, \bar{x}) \|S_i^{Z_i^A}(s)\| \quad \text{by Lemma 4.1.1}$$

$$\leq k_1 \kappa \sigma(x, \bar{x}) \sum_{j \in \eta} \|S_i^{Z_i^A}(s)\|$$

$$\leq k_1 \kappa \sigma(x, \bar{x}) \cdot \text{rank}(Z^L) \cdot \|S^{\Delta_1}(s)\|$$

$$\text{where } \Delta_1 = \text{Span}\{Z_j^A : j \in \chi(Z_1^N)\}$$

$$\leq k_0 \sigma(x, \bar{x}) \|S^{X_1}(s)\|$$

where $k_0 = k_1 \kappa n$ and $X_1 = \Delta_1$ by Theorem 3.2.1(b).

5.2.2

There exist constants $\rho_1 > 0$, $\rho_2 > 0$ such that if $\sigma(x, \bar{x}) \leq \varepsilon$ ε small enough, then

$$\frac{1}{\rho_1} \leq \frac{\|S^{X_1}(s)\|}{\|S^{X_1}(v^1)\|} \leq \rho_2, \quad \text{for } 1 \leq i \leq n.$$

: Let $A^{(1)}$ be as defined in Lemma 3.2.3. Then using Lemma , we have

$$v^1 = A^{*(1)} s + G^1(x, \bar{x}, s) \tag{5.1.3}$$

$$G_j^1(x, \bar{x}, s) = \begin{cases} g_j(x, \bar{x}, s), & \text{for } j \leq i-1 \\ 0, & \text{otherwise} \end{cases} \tag{5.1.4}$$

$$|g_j(x, \bar{x}, s)| \leq \kappa_j \sigma(x, \bar{x}) \|S^{Z_j^A}(s)\|. \tag{5.1.5}$$

From (5.1.3)

$$T^{X_1}(v^1) = T^{X_1}[A^*(1)_s] + T^{X_1}[G^1(x, \bar{x}, s)] \quad (5.1.6)$$

where

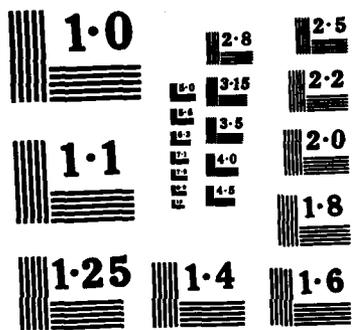
$$T^{X_1}[A^*(1)_s] = [\tilde{T}^{X_1}(A^*(1))] [T^{X_1}(s)] \quad \text{from Lemma 3.2.3 (5.1.7)}$$

and

$$\begin{aligned} & \|T^{X_1}[G^1(x, \bar{x}, s)]\| \\ & \leq \max_j \{\kappa_j\} \sigma(x, \bar{x}) \left\{ \sum_{j \in \eta} \|Z_1^A(s)\|^2 \right\}^{1/2} \quad \text{where } \eta = \chi(Z_1^N) \\ & \leq \bar{k} \sigma(x, \bar{x}) \|T^{\Delta_1}(s)\| \quad \text{where } \bar{k} = n \max_j \{\kappa_j\} \\ & \quad \text{and } \Delta_1 = \text{Span}\{Z_1^A : j \in \chi(Z_1^N)\} \\ & \leq \bar{k} \sigma(x, \bar{x}) \|T^{X_1}(s)\| \quad \text{since } \Delta_1 = X_1 \text{ by Theorem 3.2.1(b)} \end{aligned} \quad (5.1.8)$$

substituting from (5.1.7) and (5.1.8) into (5.1.6) we obtain

$$\begin{aligned} \|T^{X_1}(v^1)\| & \leq [\|\tilde{T}^{X_1}(A^*(1))\| + \bar{k} \sigma(x, \bar{x})] \|T^{X_1}(s)\| \\ & \leq \rho_1^1 \|T^{X_1}(s)\| \quad \text{where } \rho_1^1 \text{ is a constant chosen greater} \\ & \quad \text{than } \|\tilde{T}^{X_1}(A^*(1))\| + \bar{k} \sigma \\ & \leq \rho_1^1 \|T^{X_1}(s)\| \quad \text{where } \rho_1^1 = \max_i \{\rho_i^1\}. \end{aligned} \quad (5.1.9)$$



NATIONAL BUREAU OF STANDARDS
MICROCOPY RESOLUTION TEST CHART

Now let $M_1 = \tilde{T}^{X_1}(A^*(1))$. By Lemma 3.2.3, M_1 is nonsingular. Thus from (5.1.7)

$$T^{X_1}(s) = M_1^{-1} [T^{X_1}(v^1)] - M_1^{-1} T^{X_1}[G^1(x, \bar{x}, s)]$$

$$\|T^{X_1}(s)\| \leq \|M_1^{-1}\| \|T^{X_1}(v^1)\| + \|M_1^{-1}\| \bar{k}\sigma(x, \bar{x}) \|T^{X_1}(s)\| \quad \text{using (5.1.8)}$$

Hence

$$[1 - \|M_1^{-1}\| \bar{k}\sigma(x, \bar{x})] \|T^{X_1}(s)\| \leq \|M_1^{-1}\| \|T^{X_1}(v^1)\| .$$

If σ is small enough so that $1 - \|M_1^{-1}\| \bar{k}\sigma(x, \bar{x}) > 0$, then

$$\|T^{X_1}(s)\| \leq \rho_2 \|T^{X_1}(v^1)\| ,$$

where

$$\rho_2 = \max_i \{ \|M_i^{-1}\| / (1 - \|M_i^{-1}\| \bar{k}\sigma(x, \bar{x})) \} . \quad (5.1.10)$$

The result now follows from (5.1.9) and (5.1.10) by noting that for any $v \in \mathbb{R}^n$, $\|S^\Delta(v)\| = \|T^\Delta(v)\|$ for any subspace Δ . \square

Theorem 5.2.3

Update II satisfies the bounded-deterioration condition (5.1.1) of Theorem 5.1.1.

Proof: We show the result for successive pairs $(N,U), (\bar{N}, \bar{U})$ where $x, \bar{x} = x+s$ are the corresponding successive points and $y = F(\bar{x}) - F(x)$.

Denote

$$\begin{aligned} D &= N - N^* , & \bar{D} &= \bar{N} - N^* \\ E &= U - U^* , & \bar{E} &= \bar{U} - U^* \\ \alpha^* &= N^*y - U^*s , & \alpha &= Ny - Us . \end{aligned}$$

Then from (2.4.3)

$$(\bar{D}_1 + \bar{E}_1) = \begin{cases} (D_1 + E_1) & \text{if } \|S^{X_1}(\nu^1)\| = 0 \\ (D_1 + E_1) - \alpha_1 (S^{X_1}(\nu^1))^T / \|S^{X_1}(\nu^1)\|^2 & \text{otherwise .} \end{cases} \quad (5.1.11)$$

Since

$$\begin{aligned} \alpha_1 &= N_1 y - U_1 s \\ &= (D_1 y - E_1 s) + \alpha_1^* \\ &= (D_1 + E_1) \nu^1 + \alpha_1^* \\ &= (D_1 + E_1) S^{X_1}(\nu^1) + \alpha_1^* \end{aligned} \quad (5.1.12)$$

we have, for $\|S^{X_1}(\nu^1)\| \neq 0$,

$$\begin{aligned}
(\bar{D}_1 + \bar{E}_1) &= (D_1 + E_1) - [(D_1 + E_1) S^{X_1}(v^1) + \alpha_1^*] [S^{X_1}(v^1)]^T / \|S^{X_1}(v^1)\|^2 \\
&= (D_1 + E_1) \left(I - \frac{v_1 v_1^T}{v_1^T v_1} \right) - \alpha_1^* \frac{v_1^T}{v_1^T v_1} \quad \text{where } v_1 = S^{X_1}(v^1) .
\end{aligned}
\tag{5.1.13}$$

Thus for $\|S^{X_1}(v^1)\| \neq 0$,

$$\begin{aligned}
\|\bar{D}_1 + \bar{E}_1 + \alpha_1^* \left(\frac{v_1^T}{v_1^T v_1} \right)^2\| &= \|(D_1 + E_1) \left(I - \frac{v_1 v_1^T}{v_1^T v_1} \right)\|^2 \\
&= \|D_1 + E_1\|^2 - \frac{\|(D_1 + E_1)v_1\|^2}{\|v_1\|^2}
\end{aligned}
\tag{5.1.14}$$

and for $\|S^{X_1}(v^1)\| = 0$,

$$\|\bar{D}_1 + \bar{E}_1\| = \|D_1 + E_1\| .
\tag{5.1.15}$$

Let $W \in \mathbb{R}^{n \times n}$ satisfy

$$W_1 = \begin{cases} \bar{D}_1 + \bar{E}_1 & \text{if } \|v^1\| = 0 \\ \bar{D}_1 + \bar{E}_1 + \alpha_1^* \frac{v_1^T}{v_1^T v_1} & \text{otherwise} \end{cases}$$

and let

$$\pi = \{i : \|v^1\| \neq 0\} .$$

Then

$$\|\bar{D}_1 + \bar{E}_1\| \leq \|W\| + \left[\sum_{1 \in \pi} \frac{|\alpha_1^*|^2}{\|v_1\|^2} \right]^{1/2} . \quad (5.1.16)$$

But from (5.1.14) and (5.1.15)

$$\begin{aligned} \|W\|^2 &= \|D + E\|^2 - \sum_{1 \in \pi} \frac{\|(D_1 + E_1) v_1\|^2}{\|v_1\|^2} \\ &= (1-\theta) \|D + E\|^2 , \end{aligned}$$

where

$$\theta = \left[\sum_{1 \in \pi} \frac{\|(D_1 + E_1) v_1\|^2}{\|v_1\|^2} \right] / \|D + E\|^2 , \quad 0 \leq \theta \leq 1 \quad (5.1.17)$$

and from Lemmas 5.2.1 and 5.2.2, for $1 \in \pi$,

$$\frac{|\alpha_1^*|}{\|v_1\|} \leq \frac{k_0 \sigma(x, \bar{x}) \|S^{X_1}(s)\|}{\rho_2^{-1} \|S^{X_1}(s)\|} \leq k_0 \rho_2 \sigma(x, \bar{x}) . \quad (5.1.18)$$

Substituting from (5.1.17) and (5.1.18) into (5.1.16), we get

$$\|\bar{D} + \bar{E}\| \leq \sqrt{(1-\theta)} \|D + E\| + \beta \sigma(x, \bar{x}) \quad \text{where} \quad \beta = \sqrt{n} k_0 \rho_2 \quad (5.1.19)$$

Thus the bounded-deterioration property (5.1.1) is satisfied with

$$\alpha_1 = 0 \quad \text{and} \quad \alpha_2 = \beta .$$

Theorem 5.2.4

Algorithm II is Q -superlinearly convergent.

Proof: The strong form of the bounded-deterioration condition (5.1.19), with $\alpha_1 = 0$, is sufficient to ensure Q -superlinear convergence. The proof proceeds among the same lines as Theorem 3.5 of Johnson and Austria (1983) and will be omitted here. \square

6. Concluding Remarks

The non-cancellation assumption is essentially a non-degeneracy assumption, similar to the case for linear programming. Without it, Algorithm I is probably still convergent since periodic restarts are necessary anyway. However, it seems unlikely that the Q -superlinear convergence of Algorithm II will be preserved if we underestimate the density of N and U . It is interesting that the use of the non-cancellation assumption eliminates a problem encountered in the update of Dennis and Moré (1982). There it was necessary to forgo updating any row for which the norm of the projection of the previous Newton step, s , onto the sparsity pattern of that row was too small. For Updates I and II we were able to demonstrate that these projections never get too small relative to s .

The theoretical justification for Update II appears to be much stronger than for Update I, since Q -superlinear convergence is ensured without the need for periodic recalculation of the Jacobian matrix from

- (i) Many Newton-type problems must be solved so it is desirable to use an inexpensive quasi-Newton method.
- (ii) The level of difficulty of successive Newton problems can be adaptively chosen and, hence, use of Newton's method may be an expensive overkill.
- (iii) The global problem is naturally broken down into a series of local problems, each of which can be solved separately using either Algorithm I or II, thus overcoming the problem of having a fixed pivoting strategy in the updates.

CHAPTER 4

Computational Experience

1. Introduction

The ideas of the preceding three chapters have been implemented in a Fortran program. Data structures appropriate for sparse systems are used throughout. For example, only the non-zero elements of the Jacobian matrix M are stored, and these are maintained in three arrays A (double precision), $INUM$ (integer) and $JNUM$ (integer), where $M(I,J) \neq 0$ if and only if there is some K such that $A(K) = M(I,J)$, $INUM(K) = I$ and $JNUM(K) = J$. The LU factors of M are obtained by Gaussian elimination using a threshold pivoting strategy with both row and column permutations allowed. The subroutine, LU1FAC, which is part of the LUSOL package (see Gill, et al. (1984)) was used. Obtaining the NU factors explicitly (where $N = L^{-1}$) consists essentially of inverting the matrix L which results from the Gaussian elimination. This is a quite expensive process requiring $O(n^3)$ operations; it is hoped that sparsity will reduce this to $O(n^2)$ operations. The Jacobian matrix, M , is obtained by a finite difference approximation; the graph coloring heuristics of Coleman and Moré (1981) are used to reduce the number of function evaluations necessary for each matrix approximation.

2. Local Comparison

We first compare the local behaviors of the updates discussed in Chapter 3. The following two problem types are taken from Broyden

(1971), where they were used to compare the behavior of the sparse Broyden method with Newton's method.

$$\text{Type 1: } f_i(x) = (3 - k_1 x_1) x_i + 1 - x_{i-1} - 2x_{i+1}, \quad 1 \leq i \leq n$$

$$\text{Type 2: } f_i(x) = (k_1 + k_2 x_1^2) x_i + 1 - k_3 \sum_{\substack{j=i-r_1 \\ j \neq 1}}^{i+r_2} (x_j + x_j^2), \quad 1 \leq i \leq n$$

For both types we have $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ and $x_j = 0$ for $j < 1$ or $j > n$. The initial estimate of the solution in each case was taken to be x^0 , where $x_1^0 = -1$ for $1 \leq i \leq n$. The iterations were stopped at x^ℓ where $\|x^\ell - x^{\ell-1}\| \leq \epsilon$ and $\|x^j - x^{j-1}\| > \epsilon$ for $j < \ell$ and ϵ given. Tables 1, 2, and 3 give the results for varying dimension and values of the parameters k_1 , k_2 and k_3 . The following code is used for type of iterative technique:

- 1 : Newton's Method
- 2 : Dennis/Marwil Update
- 3 : Update I of Chapter 3 for LU factors
- 4 : Update II of Chapter 3 for NU factors

Only full Newton steps were taken. All three updates compared favorably, in terms of the number of function evaluations, with respect to Newton's Method. A comparison with the results of Broyden (1971) -- where a slightly different stopping criterion is used -- shows that these updates require about the same number of iterations as the sparse Broyden Update. Here, however, we have eliminated the need for a matrix factorization at each iterate.

3. The Continuation Problem

The following seven test problems were used.

Problems 1 and 2

The homotopy

$$h(x,t) = f(x) - (1-t) f(x^0) ,$$

is traced from $(x^0,0)$ to $(\bar{x},1)$. Problems 1 and 2 correspond respectively to choosing $f(\cdot)$ from Types 1 and 2 of the previous section.

Problem 3 (Watson (1979d))

We solve the linear complementary problem by the use of Mangasarian's transformation (Mangasarian (1976)).

We have

$$f : \mathbb{R}^n \rightarrow \mathbb{R}^n$$

$$f(x) = Ax + q , A \in \mathbb{R}^{n \times m}$$

$$A_{ii} = 6, A_{ij} = -4 \text{ for } |i-j| = 1$$

$$A_{ij} = 1 \text{ for } |i-j| = 2 \text{ and } A_{ij} = 0 \text{ for } |i-j| > 2$$

$$q = \lambda(-1, 0, \dots, 0) , \lambda > 0$$

We wish to solve the following problem:

$$x \geq 0, f(x) \geq 0, x^T f(x) = 0. \quad (3.1)$$

Define $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ by

$$g_i(x) = |f_i(x) - x_i|^3 - (f_i(x))^3 - x_i^3, \quad 1 \leq i \leq n$$

and $h : \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$ by

$$h(x, \lambda) = -\lambda g(x) + (1-\lambda)(x-x^0).$$

We now trace the path $\{(x, \lambda) : h(x, \lambda) = 0\}$ from $(x^0, 0)$ to $(\bar{x}, 1)$ where \bar{x} solves the linear complementary problem (3.1) (see Watson (1979d)).

Problem 4 (Kellog, Li and Yorke (1976))

$$f : \mathbb{R}^n \rightarrow \mathbb{R}^n$$

$$F_i(x) = a_i + b_i x_{\alpha_i} x_{\beta_i} x_{\gamma_i}$$

where $0 \leq a_i, b_i \leq 1$ and

$$\alpha_i, \beta_i, \gamma_i \in \{1, 2, \dots, n\}.$$

The data $a_i, b_i, \alpha_i, \beta_i, \gamma_i$ are obtained by random number generation. A fixed point of $f(\cdot)$ is located by tracing the zero curve of the homotopy

$$h(x, t) = x - t f(x)$$

from the point $(x^0, 0) = (0, 0)$ to $(\bar{x}, 1)$.

Problem 5 (Saigal (1981))

The following boundary-value problem is solved by discretization

$$u'' = (2u - 0.5t + 1)^3$$

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

On a mesh of n points we have $x_j = u(jh)$, $1 \leq j \leq n$ where $h = 1/(n+1)$ and

$$f_j(x) \equiv (x_j - 2x_j + x_{j-1}) - 2h^2(x_j - \frac{jh}{2} + 1)^3 = 0, \quad 1 \leq j \leq n.$$

We solve $f(x) = 0$ by tracing the zero curve of the homotopy

$$h(x,t) = tf(x) + (1-t)x$$

from $(x^0, 0) = (0, 0)$ to $(\bar{x}, 1)$.

Problem 6

As in problem 5, we solve the following boundary value problem by discretization on n points

$$u'' + \lambda e^u = 0,$$

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

This is an example of a parametric problem in which we may be interested in all solutions for λ in some range.

References

- Abraham, R. and Robbin, J. (1967), *Transversal mapping and flows*, W.A. Benjamin.
- Alexander, J.C. (1978), "The topological theory of an embedding method," In: *Continuation Methods*, H.Wacker (ed.), Academic Press, 37-68.
- Alexander, J.C. (1979), "Numerical continuation methods and bifurcation," In: *Functional Differential Equations and Approximation of Fixed Points*, H.O. Peitgen, H.O. Walther (eds.), *Lecture Notes in Math.* 730, Springer-Verlag, 1-15.
- Allgower, E.L. and Georg, K. (1980), "Simplicial and continuation methods for approximating fixed points and solutions to systems of equations," *SIAM Rev.* 22, 28-85.
- Allgower, E.L. and Georg, K. (1981), "Predictor corrector and simplicial methods for approximating fixed points and zero points of non-linear mappings," *Proc. of the XI International Symposium on Mathematical Programming*, Springer Verlag.
- Armijo, L. (1966), "Minimization of functions having Lipchitz-continuous first partial derivatives," *Pacific J. Math.* 16, 1-3.
- Avila, J.H. (1974), "The feasibility of continuation methods for non-linear equations," *SIAM J. Numer. Anal.* 11, 102-122.
- Broyden, C.G. (1965), "A class of methods for solving non-linear simultaneous equations," *Math. Comp.* 19, 577-593.
- Broyden, C.G. (1969), "A new method for solving non-linear simultaneous equations," *Computer Journal* 12, 94-99.

Table 10: Problem 7

Type of Update	Dimension	No. of Predictor-Corrector Cycles	No. of Final Iterates at Level $t = 1$	No. of Function Evaluations
11	9	13	3	438
22	9	18	4	204
33	9	12	7	165
44	9	12	7	164
41	9	13	4	264
11	39	11	3	451
22	29	18	5	291
33	39	*	*	*
31	39	13	5	289
44	39	*	*	*
41	39	13	5	289

* denotes failure of the algorithm to converge.

We used different starting points for the thirty-nine dimensional problem and the nine-dimensional problem (see Watson (1980b)).

Table 9: Problem 6

Type of Update	Dimension	No. of Predictor-Corrector Cycles	No. of Final Iterates at Level $t = 1$	No. of Function Evaluations
11	100	4	1	40
22	100	4	1	40
33	100	4	1	40
44	100	4	1	40

Table 8: Problem 5

Type of Update	Dimension	No. of Predictor-Corrector Cycles	No. of Final Iterates at Level $t = 1$	No. of Function Evaluations
11	100	4	3	90
22	100	4	3	50
33	100	4	3	50
44	100	4	3	50

Table 7: Problem 4

Type of Update	Dimension	No. of Predictor-Corrector Cycles	No. of Final Iterates at Level $t = 1$	No. of Function Evaluations
11	50	5	3	196
22	50	5	5	90
21	50	5	5	135
33	50	5	4	89
31	50	5	4	134
44	50	5	4	89
41	50	5	4	134
11	100	6	3	247
22	100	6	7	107
33	100	6	4	104
44	100	6	4	104

Table 6: Problem 3

Type of Update	Dimension	No. of Predictor-Corrector Cycles	No. of Final Iterates at level $t = 1$	No. of Function Evaluations
11	10	16	4	681
22	10	33	7	510
21	10	20	6	441
33	10	*	*	*
31	10	18	9	407
44	10	26	7	440
41	10	18	11	408

* denotes failure of the algorithm to converge.

Table 5: Problem 2

$$k_1 = k_2 = k_3 = 1.0; r_1 = r_2 = 1, n = 50$$

$$x^0 = (-1, -1, \dots, -1)$$

Type of Update	No. of Predictor-Corrector Cycles	No. of Final Iterates at $t = 1$	No. of Function Evaluations
11	5	4	168
22	8	6	75
44	8	7	76
40	9	5	88
41	5	6	100

Table 4: Problem 1

$$k_1 = 1.0, x^0 = (-1, -1, \dots, -1)$$

Type of Update	Dimension	No. of Predictor-Corrector Cycles	No. of Final Iterates at Level $t = 1$	No. of Function Evaluations
11	5	3	3	145
11	20	9	4	195
11	100	30	4	270
22	100	30	4	120
33	100	30	4	120
44	100	30	4	120
41	100	30	4	175

Table 3, (continued)

Type of Update	k_1	k_2	k_3	No. of Iterates $\epsilon = 10^{-6}$	No. of Function Evaluations $\epsilon = 10^{-6}$	No. of Steps $\epsilon = 10^{-10}$	No. of Function Evaluations $\epsilon = 10^{-10}$
1	2	2	2	5	66	6	79
2	2	2	2	12	26	20	34
3	2	2	2	10	24	20	34
4	2	2	2	10	24	16	30
1	2	3	2	5	66	6	79
2	2	3	2	11	25	20	34
3	2	3	2	11	25	21	35
4	2	3	2	10	24	21	35
1	2	4	1	5	66	6	79
2	2	4	1	12	26	22	36
3	2	4	1	16	30	31	45
4	2	4	1	16	30	42	56
1	2	5	1	6	79	6	79
2	2	5	1	13	27	24	38
3	2	5	1	22	36	36	50
4	2	5	1	22	36	48	62
1	3	4	1	6	79	6	79
2	3	4	1	14	28	27	41
3	3	4	1	18	32	38	52
4	3	4	1	22	36	45	59
1	3	5	1	6	79	7	92
2	3	5	1	16	30	31	45
3	3	5	1	20	34	42	56
4	3	5	1	24	38	57	71

Table 3: Type 2, dimension = 50

$$r_1 = r_2 = 5$$

Type of Update	k_1	k_2	k_3	No. of			
				Iterates $\epsilon = 10^{-6}$	Function Evaluations $\epsilon = 10^{-6}$	Iterates $\epsilon = 10^{-10}$	Function Evaluations $\epsilon = 10^{-10}$
1	1	1	1	4	53	5	66
2	1	1	1	8	22	17	31
3	1	1	1	8	22	13	27
4	1	1	1	7	21	13	27
1	2	1	1	5	66	6	79
2	2	1	1	10	24	18	32
3	2	1	1	9	23	16	30
4	2	1	1	9	23	16	30
1	1	2	1	5	66	6	79
2	1	2	1	8	22	29	33
3	1	2	1	10	24	19	23
4	1	2	1	9	23	17	21
1	3	2	1	5	66	6	79
2	3	2	1	23	27	39	43
3	3	2	1	13	17	26	40
4	3	2	1	14	18	26	40
1	2	3	1	5	66	6	79
2	2	3	1	11	25	20	34
3	2	3	1	13	27	28	42
4	2	3	1	16	30	26	40
1	3	3	1	5	66	6	79
2	3	3	1	12	26	55	69
3	3	3	1	16	30	39	43
4	3	3	1	15	29	32	36
1	2	2	1	5	66	6	79
2	2	2	1	17	31	26	40
3	2	2	1	11	25	22	36
4	2	2	1	11	25	25	39
1	1	2	2	5	66	6	79
2	1	2	2	13	27	26	40
3	1	2	2	12	26	20	34
4	1	2	2	10	24	25	39

Table 2: Type 2, $\epsilon = 10^{-10}$

$$k_1 = k_2 = k_3 = 1.0$$

Type of Update	Dimension	r_1	r_2	No. of Iterates	No. of Function Evaluations
1	50	1	1	6	31
2	50	1	1	27	32
3	50	1	1	19	24
4	50	1	1	18	23
1	50	3	3	6	57
2	50	3	3	14	23
3	50	3	3	19	28
4	50	3	3	17	26
1	50	5	1	6	55
2	50	5	1	15	24
3	50	5	1	18	27
4	50	5	1	18	27

Table 1: Type 1, $\epsilon = 10^{-10}$

Type of Update	Dimension	k_1	No. of Iterates	No. of Function Evaluations
1	10	0.5	5	26
2	10	0.5	13	18
3	10	0.5	14	19
4	10	0.5	14	19
1	10	2.0	6	31
2	10	2.0	16	21
3	10	2.0	16	21
4	10	2.0	17	22

Efficient implementation of such a strategy requires further research into the monitoring and control of the algorithm and was not attempted here. We have demonstrated, as intended, that the use of quasi-Newton methods for the sparse continuation problem can lead to an increase in efficiency over more traditional methods.

Problem 7 (Watson (1980b))

The following boundary value problem arises in the study of the motion of a fluid squeezed between two parallel plates.

$$S(\eta f'''' + 3f'' + mf' f'' - ff''') = f^{(4)},$$

$$f(0) = f''(0) = 0, \text{ for } (1) = 1, f'(1) = 0,$$

$$m = 0 \text{ (axisymmetric case) .}$$

We discretize on p points to obtain an n dimensional problem, where $n = 2p + 1$. See Watson (1980b) for further details.

The first column of Tables 4 - 10 contains a two-digit number. The first digit is the code for the update used on the corrector iterative sequence; the second refers to the updating technique used at the end of the corrector sequence to approximate the tangent direction. The tangent approximation technique provided some favorable results (see Table 7), but in general proved to be unreliable (as seen in Table 10).

The results of Tables 4 - 10 illustrate that substantial savings can be achieved by using the less expensive updating techniques rather than Newton's method. As with other quasi-Newton methods for non-linear systems the extent of this usefulness depends on the degree of non-linearity of the equations, with quasi-Newton methods becoming less useful the more highly non-linear the system is. In practice, a truly adaptive continuation strategy would allow for switching between quasi-Newton and Newton methods or even sparse simplicial techniques.

- Broyden, C.G. (1971), "The convergence of an algorithm for solving sparse non-linear systems," Math. Comp. 25, 285-294.
- Broyden, C.G., Dennis, J.E. and Moré, J.J. (1973), "On the local and super linear convergence of quasi-Newton methods," J. Inst. Math. Appl. 12, 223-245.
- Chow, S.N., Mallet-Paret, J. and Yorke, J.A. (1978), "Finding zeroes of maps: Homotopy methods that are constructive with probability one," Math. Comp. 32, 887-899.
- Coleman, T. and Moré, J. (1981), "Estimation of sparse Jacobian matrices and graph coloring problems," Tech. Report ANL-81-39, Argonne National Laboratory.
- Crandall, M.G. and Rabinowitz, P.H. (1971), "Bifurcation from simple eigenvalues," J. Func. Anal. 8, 321-340.
- Curtis, A.R., Powell, M.J.D. and Reid, J.K. (1974), "On the estimation of sparse Jacobian matrices," J. Inst. Math. Appl. 13, 117-119.
- Curtis, A.R. and Reid, J.K. (1974), "The choice of steplength when using differences to approximate Jacobian matrices," J. Inst. Math. Appl. 13, 121-126.
- Davidenko, D. (1953), "On a new method of numerical solution of systems of nonlinear equations," Doklady Akad. Nauk USSR 88, 601-602 (in Russian).
- Den Heijer, C. and Rheinboldt, W.C. (1981), "On steplength algorithms for a class of continuation methods," SIAM J. Numer. Anal. 18, 925-948.

- Dennis, J.E. and Moré, J.J. (1974), "A characterization of superlinear convergence and its applications to quasi-Newton methods," *Math. Comp.* 28, 549-560.
- Dennis, J.E. and Moré, J.J. (1977), "Quasi-Newton methods, motivation and theory," *SIAM Rev.* 19, 46-89.
- Dennis, J.E. and Moré, J.J. (1982), "Direct secant updates of matrix factorizations," *Math. Comp.* 38, 459-474.
- Dennis, J.E. and Schnabel, R.B. (1979), "Least change secant updates for quasi-Newton methods," *SIAM Rev.* 21, 443-459.
- Deuflhard, P. (1979), "A stepsize control for continuation methods and its special applications to multiple shooting techniques," *Numer. Math.* 33, 115-146.
- Deuflhard, P. and Heindl, G. (1979), "Affine invariant convergence theorems for Newton's method and extensions to related methods," *SIAM J. Numer. Anal.* 16, No. 1.
- Deuflhard, P., Pesch, H.J. and Rentrop, P. (1976), "A modified continuation method for numerical solution of non-linear two-point boundary value problems by shooting techniques," *Numer. Math.* 26, 327-343.
- Duff, I.S. (1977), "A survey of sparse matrix research," *Proc. IEEE* 65, 500-535.
- Eaves, B.C. (1972), "Homotopies for computation of fixed points," *Math. Prog.* 3, 1-22.
- Eaves, B.C. (1976), "A short course in solving equations with PL homotopies," *SIAM-AMS Proc.* 9, 73-143.

- Erisman, A.M. and Reid, J.K. (1974), "Monitoring the stability of the triangular factorization of a sparse matrix," Numer. Math. 22, 183-186.
- Garcia, C.B. and Gould, F.J. (1978), "A theorem on homotopy paths," Math of Op. Res. 3, 282-289.
- Garcia, C.B. and Zangwill, W.I. (1979), "Finding all solutions to polynomial systems and other systems of equations," Math. Prog. 16, 159-176.
- Garcia, C.B. and Zangwill, W.I. (1980), "The flex simplicial algorithm," In: Numerical Solution of Highly Non-linear Problems, W. Forster (ed.), North-Holland, 71-92.
- Garcia, C.B. and Zangwill, W.I. (1981), Pathways to Solutions, Fixed Points and Equilibria, Prentice Hall.
- Georg, K. (1981a), "Numerical integration of a Davidenko Equation," In: Numerical Solution of Non-linear Equations, E. Allgower, K. Glashoff, H.O. Peitgen (eds.), Springer Lecture Notes in Math. 878, 128-161.
- Georg, K. (1981b), "On tracing an implicitly defined curve by quasi-Newton steps and calculating bifurcation by local perturbation," SIAM J. SSC 2, 35-50.
- Georg, K. (1981c), "A note on stepsize control for numerical curve following," In: Homotopy Methods and Global Convergence, B.C. Eaves, F.J. Gould, H.O. Peitgen (eds.), NATO Conference Series
- Gill, P.E., Golub, G.H., Murray, W. and Saunders, M.A. (1974), "Methods for modifying matrix factorizations," Math. Comp. 28, 505-535.

- Gill, P.E., Murray, W., Saunders, M.A., Wright, M.H. (1984), "LUSOL: A package for updating LU factors of a sparse matrix," presented at the Gatlinburg 9 Conference on Numerical Linear Algebra, University of Waterloo.
- Goldstein, A. (1967), "Constructive Real Analysis," Harper and Row.
- Hackl, J., Wacker, H.J. and Zulehner, W. (1980), "An efficient stepsize control for continuation methods," BIT 20, 475-485.
- Haselgrove, C.B. (1961), "The solution of non-linear equations and differential equations with two-point boundary conditions," Comput. J. 4, 255-259.
- Hirsch, M.W. and Smale, S. (1979), "On algorithms for solving $f(x) = 0$," Comm. Pure and Appl. Math. 32, 381-312.
- Johnson, G.W. and Austria, N.H. (1983), "A quasi-Newton method employing direct secant updates of matrix factorizations," SIAM J. Numer. Anal. 20, 315-325.
- Jorgens, H., Peitgen, H. and Saupe, D. (1980), "Topological perturbations in the numerical study of non-linear eigenvalue and bifurcation problems," In: Proc. Symposium on Analysis and Computation of Fixed Points, S.M. Robinson (ed.), Academic Press, 139-181.
- Kearfott, R.B. (1981), "A derivative-free arc continuation method and a bifurcation technique," In: Numerical Solution of Non-linear Equations, E. Allgower, K. Glashoff, H.O. Peitgen (eds.), Springer Lecture Notes in Math. 878, 182-198.
- Kearfott, R.B. (1983), "Some general bifurcation techniques," SIAM J. Sci. Stat. Comp. 4, No. 1, 52-68.

- Keller, H.B. (1977), "Numerical solution of bifurcation and nonlinear eigenvalue problems," In: Applications of bifurcation theory, P.H. Rabinowitz (ed.), Academic Press.
- Kellog, R.G., Li, T.Y. and Yorke, J.A. (1976), "A constructive proof of the Brouwer fixed point theorem and computational results," SIAM J. Numer. Anal. 4, 473-483.
- Kubicek, M. (1976), "Algorithm 502, Dependence of solutions of nonlinear systems on a parameter," ACM-TOMS 2, 98-107.
- Laasonen, P. (1970), "An imbedding method of iteration with global convergence," Computing 5, 253-358.
- Leder, D. (1970), "Automatische Schrittweitensteuerung bei global konvergenten Einbettungsmethoden," ZAMM 54, 319-342.
- Lemke, C.E. and Howson, J.T. (1964), "Equilibrium points of bimatrix games," SIAM J. Appl. Math 12, 413-423.
- Li, T.Y. and Yorke, J.A. (1979), "Path following approach for solving nonlinear equations: homotopy, continuous Newton and projection," In: Functional differential equations and approximation of fixed points, H. Peitgen, H. Walther (eds.), Springer Lecture Notes in Math. 730, 257-264.
- Li, T.Y. and Yorke, J.A. (1980), "A simple reliable numerical algorithm for following homotopy paths," In: Analysis and Computation of Fixed Points, S.M. Robinson (ed.), Academic Press, New York, 73-91.
- Mangasarian, O.L. (1976), "Equivalence of the complementarity problem to a system of nonlinear equation," SIAM J. Appl. Math 31, No. 1.
- Marwil, E.S. (1978), "Exploiting sparsity in Newton-like methods," Ph.D. Thesis, Cornell University, Ithaca, NY.

- Marwil, E.S. (1984), "Some numerical results using direct secant updates of matrix factorizations," preprint.
- Merrill, O.H. (1972), "Applications and extensions of an algorithm that computes fixed points of a certain upper semi-continuous point to set mapping," Ph.D. Thesis, University of Michigan.
- Moore, R.E. (1978), "A computational test for convergence of iterative methods for nonlinear systems," SIAM J. Numer. Anal. 15, No. 6.
- Ortega, J.M. and Rheinboldt, W.C. (1970), "Iterative solution of nonlinear equations in several variables," Academic Press.
- Peitgen, H.O. and Prufer, M. (1979), "The Leray-Schauder continuation method is a constructive element in the numerical study of nonlinear eigenvalue and bifurcation problems," In: Functional Differential Equations and Approximation of Fixed Points, H.O. Peitgen, H.O. Walther (eds.) Springer Lecture Notes in Math. 730, 326-409.
- Peitgen, H.O. and Schmitt, K. (1981), "Positive and spurious solutions of nonlinear eigenvalue problems," In: Numerical Solution of Nonlinear Equations, E.L. Allgower, K. Glashoff, H.O. Peitgen (eds.) Springer Lecture Notes in Math. 878, 257-324.
- Powell, M.J.D. (1970), "A hybrid method for nonlinear equations," In: Numerical Methods for Nonlinear Algebraic Equations, P. Rabinowitz (ed.), Gordon and Breach.
- Reid, J.K. (1971), "A note on the stability of Gaussian Elimination," J. Inst. Math. Appl. 8, 374-375.

- Rheinboldt, W.C. (1974), "On the solution of large, sparse sets of nonlinear equations," Technical Report TR-324, University of Maryland, Computer Science Center.
- Rheinboldt, W.C. (1977), "Numerical continuation on methods for finite element applications," In: Formulation and Algorithms in Finite Element Analysis, K.J. Bahte, J.T. Oden, W. Wunderlich (eds.), MIT Press, 599-631
- Rheinboldt, W.C. (1977), "An adaptive continuation process for solving systems on nonlinear equations," Polish Academy of Sciences, Banach Ctr. Publ. 3, 129-142.
- Rheinboldt, W.C. (1978), "Numerical methods for a class of finite dimensional bifurcation problems," SIAM J. Numer. Anal. 15, 1-11.
- Rheinboldt, W.C. (1980), "Solution fields of nonlinear equations and continuation methods," SIAM J. Numer. Anal. 17, 221-237
- Rheinboldt, W.C. (1981), "Numerical analysis of continuation methods for nonlinear structural problems," Computers and Structures, 103-113.
- Rosenberg, A. (1983), "Numerical solutions of systems of simultaneous polynomial equations using continuous homotopy methods," Ph.D. Thesis, Stanford University, California.
- Saigal, R. (1976), "On paths generated by fixed point algorithms," Math of Op. Res. 1, 359-380.
- Saigal, R. (1977), "On the convergence rate of algorithms for solving equations that are based on methods of complementary pivoting," Math. of Op. Res. 2, 108-124.
- Saigal, R. (1981), "A homotopy for solving large, sparse and structured fixed point problems," Preprint, Northwestern University.

- Saigal, R. and Todd, M.J. (1978), "Efficient acceleration techniques for fixed point algorithms," *SIAM J. Numer. Anal.* 15, 997-1007.
- Sard, A. (1942), "The measure of the critical values of differential maps," *Bull. AMS* 48, 882-890.
- Saupe, D. (1982), "On accelerating PL continuation algorithms by predictor corrector methods," *Math. Prog.* 23, 87-110.
- Scarf, H.E. (1967), "The approximation of fixed points of a continuous mapping," *SIAM J. Appl. Math.* 15, 1328-1343.
- Scarf, H.E. and Hansen, T. (1973), "Computation of economic equilibria," Yale University Press, New Haven.
- Schmidt, C.P. (1979), "Approximating differential equations that describe homotopy paths," Tech. Report 7931, University of Santa Clara, California.
- Schubert, L.K. (1970), "Modification of a quasi-Newton method for nonlinear equations with a sparse Jacobian," *Math. Comp.* 24, 27-30.
- Shampine, L.F. and Gordon, M.K. (1975), *Computer Solution of Ordinary Differential Equations: The Initial Value Problem*, W.H. Freeman and Company.
- Smale, S. (1976), "A convergent process of price adjustment and global Newton Methods," *Journal of Mathematical Economics* 3 (1976), 107-120.
- Thapa, M. (1981), "Optimization of unconstrained functions with sparse Hessian matrices," Ph.D. Thesis, Stanford University, California.
- Todd, M.J. (1976), "The computation of fixed points and applications," *Spainger Verlag, Lecture Notes in Econ. and Math. Systems* 124.

- Todd, M.J. (1980a), "Exploiting structure in piecewise-linear homotopy algorithms for solving equations." Math. Prog. 18, 233-247.
- Todd, M.J. (1980b), "Traversing large pieces of linearity in algorithms that solve equations by following piecewise-linear paths," Math. of Op. Res. 5, 242-257.
- Toint, Ph.L. (1977), "On sparse and symmetric matrix updating subject to a linear equation," Math. Comp. 31, No. 140, 954-961.
- Toint, Ph.L. (1978), "Some numerical results using a sparse matrix updating formula in unconstrained optimization," Math. Comp. 32, No.143, 839-851.
- Wacker, H.J. (1978) (ed.), Continuation Methods, Academic Press.
- Wacker, H.J., Zarzer, E. and Zulehner, W. (1978), "Optimal stepsize control for the globalized Newton method," In: Continuation Methods, H.J. Wacker (ed.), Academic Press, 249-276.
- Watson, L.T. (1979a), "An algorithm that is globally convergent with probability one for a class of nonlinear two-part boundary value problems," SIAM J. Numer. Anal. 16, 394-401.
- Watson, L.T. (1979b), "Fixed points of C^2 maps," J. Comput. Appl. Math. 5, 131-140.
- Watson, L.T. (1979c), "A globally convergent algorithm for computing fixed points of C^2 maps," Appl. Math. Comput. 5, 297-311.
- Watson, L.T. (1979d), "Solving the nonlinear complementary problem by a homotopy method," SIAM J. Control and Optimization 17, 36-46.
- Watson, L.T. (1980a), "Computational experience with a Chow-Yorke algorithm," Math. Prog. 19, 92-101.

Watson, L.T. (1980b), "Solving finite difference approximations to non-linear two point boundary value problems by a homotopy method," SIAM J. Sci. Stat. Comput. 1, 467-480.

Watson, L.T. (1981), "Engineering applications of the Chow-Yorke algorithm," Appl. Math. and Comp. 9, 111-133.

Wilkinson, J.H. (1965), The Algebraic Eigenvalue Problem, Clarendon Press, Oxford.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER SOL 85-8	2. GOVT ACCESSION NO. AD-A157 587	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Sparse Quasi-Newton Methods and the Continuation Problem	5. TYPE OF REPORT & PERIOD COVERED Technical Report	
	6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s) Floyd F. Chadee	8. CONTRACT OR GRANT NUMBER(s) N00014-85-K-0343	
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Operations Research - SOL Stanford University Stanford, CA 94305	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-047-064	
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research - Dept. of the Navy 800 N. Quincy Street Arlington, VA 22217	12. REPORT DATE June 1985	
	13. NUMBER OF PAGES 127 pp.	
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report) UNCLASSIFIED	
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE	
16. DISTRIBUTION STATEMENT (of this Report) This document has been approved for public release and sale; its distribution is unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) homotopy predictor-corrector		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) PLEASE SEE ATTACHED.		

DD FORM 1 JAN 73 1473

EDITION OF 1 NOV 65 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

DL 85-8: Sparse Quasi-Newton Methods and the Continuation Problem,
by Floyd F. Chadee

The problem of tracing a smooth path arises in many engineering problems, the solution of parametric differential equations and eigenvalue problems; it also finds application in the solution of nonlinear systems of equations by homotopy techniques. In many instances, the path is defined implicitly as the solution of a system of equations whose Jacobian matrix is large and sparse. Robust simplicial path-following techniques cannot be applied to large problems since the work involved rises rapidly with increasing dimension. This dissertation addresses the numerical problems involved in tracing the path for large sparse systems by the use of a predictor-corrector algorithm.

The corrector phase of a predictor-corrector algorithm is very expensive if Newton's method is used as the corrector. We investigate the use of sparse quasi-Newton techniques to reduce this expense. In order to avoid the drawbacks of the sparse Broyden method -- the need for a matrix factorization on each iterate and the need to store both the Jacobian matrix and its factors -- we examine techniques for directly updating the L factors of the approximation to the Jacobian matrix. Under reasonable assumptions on the systems of equations to be solved, a proof of local superlinear convergence is presented for two sparse updating techniques.

A predictor-corrector algorithm employing these sparse updating techniques is implemented in a Fortran code and numerical results are obtained demonstrating the advantages to be gained from the use of quasi-Newton methods for the large sparse continuation problem.

the expense of the corrector phase of the predictor-corrector algorithm inherent in Newton's methods

END

FILMED

9-85

DTIC