ANALYSIS AND MODIFICATION
OF
NEWTON'S METHOD AT SINGULARITIES

by

A. O. Griewank

A thesis submitted to the

Australian National University

for the degree of Doctor of Philosophy

June, 1980

PREFACE

Large parts of the research for this thesis were done in close cooperation with Dr M.R. Osborne, who suggested in particular the bordering approach developed in Section 4.2. Of two joint papers, one [46] gives a simplified proof of Lemma 2.2 for first order singularities and the other [24] deals mainly with the irregular case which is not considered in this thesis. The main results of Chapter 2 are contained in the technical report [47].

The proof of Lemma 1.3 (iii) was suggested to me by Mark Lukas and Dr J.E. Hutchinson helped me with the proof of Lemma 1.4. Otherwise the results of the thesis are my own.

A. Griewank

ACKNOWLEDGEMENTS

Throughout my postgraduate studies I received continuous support and encouragement from my supervisor Professor Richard Brent, who originally suggested research on nonlinear equations and provided valuable advice. During our joint research I had many stimulating discussions with Dr Mike Osborne, who contributed several ideas and improved my understanding of singular problems and their numerical solution with his constructive criticism. In writing the thesis I was greatly assisted by Mark Lukas who suggested many clarifications and fought a losing battle against my German "Bandwurmsaetze". I wish to thank Dorothy Nash for typing the thesis with her usual proficiency, despite many delays and last minute changes.

The work would have been impossible without the support and patience of my wife Elizabeth. Accepting a traditional role without traditional security she lived through the ups and downs of my research effort and cared for our two children. I am greatly indebted to my Mother, who always encouraged my studies even though it meant long term separation from her only son.

Finally I acknowledge with gratitude the financial support, received from the Australian Government and the Australian National University, and the friendliness and hospitality of the Australian people.

ABSTRACT

For systems of nonlinear equations $f = 0$ with singular Jacobian $\nabla f(x^*)$ at some solution $x^* \in f^{-1}(0)$ the behaviour of Newton's method is analysed. Under a certain regularity condition Q-linear convergence is shown to be almost sure from all initial points that are sufficiently close to x^*. The possibility of significantly better performance by other nonlinear equation solvers is ruled out. Instead convergence acceleration is achieved by variation of the stepsize or Richardson extrapolation. If the Jacobian ∇f of a possibly underdetermined system is known to have a nullspace of a certain dimension at a solution of interest, an overdetermined system based on the QR or LU decomposition of ∇f is used to obtain superlinear convergence.

TABLE OF CONTENTS

INTRODUCTION (i)

NOTATION AND TERMINOLOGY (vi)

CHAPTER 1 : GENERAL RESULTS ON NEWTON'S METHOD AT SINGULARITIES

1. The Determinant Function in the Neighbourhood of a Singularity. 1

2. The Starlike Domain of Invertibility R' . 8

3. Rational Expansion of the Newtonian Iteration Function. 20

4. General Results on Domains of Convergence and Contraction. 29

5. General Results on Rates of Convergence. 41

CHAPTER 2 : STARLIKE DOMAINS OF CONVERGENCE AT REGULAR SINGULARITIES

1. Balanced and Regular Singularities. 48

2. Domains of Convergence at Regular Singularities. 57

3. First Step Analysis and Main Result. 68

CHAPTER 3 : MODIFICATION OF NEWTON'S METHOD AT SINGULARITIES

1. The Numerical Difficulty of Singular Problems. 80

2. Asymptotic Behaviour of Newton's Method at Regular Singularities. 90

3. Variation of the Stepsize at Regular Singularities. 96

CHAPTER 4 : EXTRAPOLATION AND BORDERING

1. Extrapolation at Regular Singularities. 113

2. Bordering of Underdetermined or Singular Systems. 134

DISCUSSION AND CONCLUSION 149

APPENDIX 152

INDEX 164

BIBLIOGRAPHY 166

INTRODUCTION

Whenever the Jacobian $\nabla f(x)$ of a vector function $f : \mathbb{R}^n \to \mathbb{R}^n$ is explicitly available, the Newton iteration

$$x_{j+1} = x_j - \nabla f^{-1}(x_j) f(x_j) \qquad (1)$$

is the natural approach to the numerical solution of the system of nonlinear equations $f(x) = 0$.

If the Jacobian is continuous and nonsingular at some solution point $x^* \in f^{-1}(0)$, the Newton iteration converges to x^* from any initial point x_0 in a sufficiently small ball centred at x^* . The radius of such a ball can only be given explicitly if the Jacobian satisfies a known Lipschitz condition, in which case the rate of convergence is not only superlinear but quadratic.

Locally, the generally good performance of Newton's method can be impaired by either discontinuity or singularity of the Jacobian at x^* . In this thesis we study the latter contingency under the assumption that f is at least twice Lipschitz continuously differentiable.

If systems of simultaneous equations were generated at random singularity would be an extremely unlikely occurrence. However in practice they are derived from models of some usually more complex problem in for instance, science or economics. In this context singularity of some solution x^* (i.e. its Jacobian $\nabla f(x^*)$) is a distinct possibility and may be quite instructive with respect to the model or the underlying problem. For instance some of the model variables may have been chosen badly or a relevant functional relation could have been overlooked. Otherwise the underlying physical or social system may actually be in some transition state or at a bifurcation point.

Therefore our first aim is to study the behaviour of Newton's method in the neighbourhood of singularities such that their existence and kind can be inferred from the properties of the iteration sequences generated by (1). Secondly we attempt to accelerate the generally slow, linear convergence of Newton's method to singularities by suitable modifications.

Throughout the thesis the emphasis is on the theoretical analysis of singular problems and prospective techniques for their solution rather than the development of an efficient and reliable computer routine.

As often in numerical analysis, we expect from the study of the exactly singular case to gain insight into the properties of systems $f = 0$ which are nearly singular, in that the computed values of f are barely distinguishable from those of some function $\tilde{f}$ for which $\tilde{f} = 0$ has singular solutions. Methods which converge fast to exactly singular solutions could be used during the intermediate stages of iterative schemes for the location of nearly singular solutions. Unfortunately this approach has severe limitations because the currently available classification of algorithms and the mathematical tools for their analysis are asymptotic in nature and cannot be applied to the intermediate stage of some iteration.

Except for the scalar case $n = 1$, which has been examined in considerable detail (for instance in Traub [1]), there are only a few, comparatively recent results on singular simultaneous equations. It is mentioned on page 119 in [2] that a solution x^*, whose existence is guaranteed if the famous Newton-Kantorovich Theorem applies at some point x_0, may be singular. However this is only possible under the most extreme conditions. If the theorem applies at x_0 with respect to the Euclidean norm of vectors and spectral norm of matrices, it can be seen that singularity of $\nabla f(x^*)$ requires with $s \equiv x_0 - x^*$

$$f(x^*+\lambda s) = f(x_0)(\lambda/\|s\|)^2 \ .$$

Furthermore $f(x_0)$ and s must be at the same time left and right singular vectors associated with the smallest singular value $\|\nabla f_0^{-1}\|^{-1}(\lambda/\|s\|)$ of the Jacobians

$$\{\nabla f(x^*+\lambda s)\}_{\lambda\in[0,\|s\|]}$$

and the largest singular values $\|\nabla f_0^{-1}\|^{-1}(1-\lambda/\|s\|)$ of the difference matrices

$$\{\nabla f(x^*+\lambda s) - \nabla f(x_0)\}_{\lambda\in[0,\|s\|]} \ .$$

For all its importance in the nonsingular case we can therefore conclude that the Newton Kantorovich Theorem is practically never applicable in the neighbourhood of singular solutions.

Rall [3] and Cavanagh [4] considered the case in which the Jacobian is singular at the solution x^* itself but nonsingular in some neighbourhood of x^* . In this rather special situation Newton's method converges under suitable assumptions from within some ball centred at x^* , as shown in Theorem 2.4 (iv) of this thesis. In general we must expect, according to Lemma 1.1, that x^* is not isolated in the set of points at which the Jacobian is singular.

The first result under these more realistic conditions is due to Reddien [5], who showed the convergence of (1) from within the intersection of some cone with some ball, provided x^* is a regular first order singularity as defined in Section 2.1 and Theorem 1.6. Actually Reddien's assumptions were considerably stronger in the finite dimensional case, but his result applies also to differentiable operators between Banach spaces. In this thesis we will always assume that the number of equations and variables equals some fixed integer n .

$$f(x^*+\lambda s) = f(x_0)(\lambda/\|s\|)^2 .$$

Furthermore $f(x_0)$ and s must be at the same time left and right singular vectors associated with the smallest singular value $\|\nabla f_0^{-1}\|^{-1}(\lambda/\|s\|)$ of the Jacobians

$$\{\nabla f(x^*+\lambda s)\}_{\lambda\in[0,\|s\|]}$$

and the largest singular values $\|\nabla f_0^{-1}\|^{-1}(1-\lambda/\|s\|)$ of the difference matrices

$$\{\nabla f(x^*+\lambda s) - \nabla f(x_0)\}_{\lambda\in[0,\|s\|]} .$$

For all its importance in the nonsingular case we can therefore conclude that the Newton Kantorovich Theorem is practically never applicable in the neighbourhood of singular solutions.

Rall [3] and Cavanagh [4] considered the case in which the Jacobian is singular at the solution x^* itself but nonsingular in some neighbourhood of x^* . In this rather special situation Newton's method converges under suitable assumptions from within some ball centred at x^* , as shown in Theorem 2.4 (iv) of this thesis. In general we must expect, according to Lemma 1.1, that x^* is not isolated in the set of points at which the Jacobian is singular.

The first result under these more realistic conditions is due to Reddien [5], who showed the convergence of (1) from within the intersection of some cone with some ball, provided x^* is a regular first order singularity as defined in Section 2.1 and Theorem 1.6. Actually Reddien's assumptions were considerably stronger in the finite dimensional case, but his result applies also to differentiable operators between Banach spaces. In this thesis we will always assume that the number of equations and variables equals some fixed integer n .

An important aspect of Reddien's result is the departure from the idea that convergence must be established from within some ball about the solution x^*, which is in general impossible in the singular case. In Section 1.2 we introduce the concept of a *starlike domain* centred at x^*, which includes balls, cones and the interior of polytopes in the sense of Householder [6] as special cases. The *density* of such a starlike domain A at x^* is given by a real number $\tau^*(A) \in [0,1]$, which can be thought of as the probability that a point, which is "very" close to x^*, belongs to A.

In Chapter 1 we examine the general properties of the determinant function $\det(\nabla f(x))$ and the Newtonian iteration function $x - \nabla f^{-1}(x) f(x)$, on the starlike domain R' in which the Jacobian is nonsingular. This analysis leads to the definition of the *order* $k \in \mathbb{N}$ and the *degree* $\hat{i} \in \{1,0,-1,\ldots\}$ of a singularity. Depending on the degree we obtain bounds on the density of *domains of convergence*, *domains of bounded convergence*, and *domains of contraction*, respectively (Section 1.4). The analysis of Chapter 1 indicates that only singularities of first degree can be located by Newton's method in a reasonably stable fashion.

In the first Section of Chapter 2 we introduce the concept of balanced and regular singularities, which are necessarily of first degree but may have any order $k \in \mathbb{N}$. As a generalisation of results by Reddien in [5,7] and Decker and Kelley in [8,9], we establish in Section 2.2 Q-linear convergence to any regular singularity from within some starlike domain $W \subseteq R'$. Then we show in Section 2.3 that the first step from within some starlike domain $R \subseteq R'$ with density 1 leads into W and thus to convergence to x^*. Parallel to these developments we establish, under the assumption of *strong regularity*, Q-linear convergence of *approximate Newton sequences*.

Before discussing modifications of Newton's method, we observe in Section 3.1 that any nonlinear equation solver, that is Lipschitz continuous in the values of f, converges at best R-sublinearly to x^* from all initial points within some starlike domain with density 1 at x^*. In view of this result the performance of Newton's method in the regular case seems quite acceptable and its convergence can be considerably accelerated by *variations of stepsize* (Section 3.3) or *extrapolation* (Section 4.1). In the final Section 4.2 we consider the case where, the Jacobian of some possibly underdetermined nonlinear system

$$f(x) = 0 \quad \text{with} \quad f \in \mathbb{R}^{n+n'} \to \mathbb{R}^n$$

has a nullspace of known dimension at a solution $x^* \in f^{-1}(0)$ of interest. This kind of problem can be solved quite efficiently by *bordering* based on the QR or LU decomposition of the Jacobian.

Test calculations with all discussed methods on a family of singular problems in three variables are reported in the Appendix.

NOTATION AND TERMINOLOGY

With the exception of the integers $\mathbb{N}$ and the reals $\mathbb{R}$ sets are denoted by script capitals. The characters $i,j...q$ represent integers. $\mathbb{R}^{i\times\ell}$ denotes the set of real $i\times\ell$ matrices, which are represented by Roman capitals. Vectors are denoted by small Roman letters and scalars usually by small Greek characters but sometimes also by small Roman letters.

The sequence of iterates generated by some iterative scheme is usually denoted by $\{x_j\}_{j\geq 0} \subset \mathbb{R}^n$. The components of $x \in \mathbb{R}^n$ occur almost exclusively in the examples and are written as $x = (\xi,\zeta,\eta)^T$ if $n = 3$ and $x = (\xi_1,\xi_2,..\xi_n)^T$ otherwise.

For any two open subsets $\mathcal{U} \subseteq \mathbb{R}^i$ and $\mathcal{V} \subseteq \mathbb{R}^\ell$ the linear space of all functions, that have a continuous q-th Frechet-derivative in $x \in \mathcal{U}$ is denoted by

$$C^q(\mathcal{U},\mathcal{V}).$$

We will usually assume that the Frechet derivatives are locally Lipschitz continuous (i.e. Lipschitz continuous in some neighbourhood of each $x \in \mathcal{U}$) and denote the corresponding subspace of $C^q(\mathcal{U},\mathcal{V})$ by

$$C^{q,1}(\mathcal{U},\mathcal{V}) .$$

In order to avoid ambiguity of superscripts we use the symbol $\nabla^q h$ to denote the q-th derivative tensor of $h \in C^q(\mathcal{U},\mathcal{V})$. Repeated multiplication of $\nabla^q h$ from the right by some column vector $v \in \mathbb{R}^i$ is defined such that for fixed $x \in \mathbb{R}^i$ and $q \geq \Delta q \in \mathbb{N}$

$$\frac{d^{\Delta q}}{d\lambda^{\Delta q}} \nabla^{q-\Delta q} h(x+\lambda v)\Big|_{\lambda=0} = \nabla^q h(x) v^{\Delta q} \equiv \nabla^q h(x) \underbrace{v\ v...v}_{\Delta q} .$$

In particular we consider the expressions

$$\nabla^q h(x) v^q \in \mathbb{R}^\ell \quad \text{and} \quad \nabla^q h(x) v^{q-1} \in \mathbb{R}^{\ell\times i}$$

as a column vector and a $\ell \times i$ matrix respectively.

Unless otherwise specified $\|\cdot\|$ will always denote the Euclidean norm for vectors and the spectral norm for matrices. For fixed $x^* \in \mathbb{R}^n$ the elements of $\mathbb{R}^n$ are frequently written as $x = x^* + \rho t$, where $\rho = \|x - x^*\|$ and t is an element of the unit sphere

$$S \equiv \{t \in \mathbb{R}^n \mid \|t\| = 1\} .$$

Given a particular converging sequence $x_j = x^* + \rho_j t_j \to x^*$ and some real valued function h on a domain that includes $\{x_j\}_{j \geq 0}$ we abbreviate

$$h = O(\rho_j^{\ell}) \quad \text{if} \quad \limsup_{j \to \infty} |h(x_j)| \rho_j^{-\ell} < \infty$$

and

$$h = o(\rho_j^{\ell}) \quad \text{if} \quad \limsup_{j \to \infty} |h(x_j)| \rho_j^{-\ell} = 0 .$$

The same notation is used for vectors, matrices or tensors, whose entries are real valued functions which satisfy either condition for the same ℓ.

A real valued function h on some domain of the form

$$\mathcal{D} \equiv \{z \in \mathbb{R}^n \mid z = 0 \quad \text{or} \quad z/\|z\| \in T\} \quad \text{with} \quad T \subseteq S$$

is said to be *homogeneous* with the *degree of homogeneity* $i \in \mathbb{N}$ if

$$h(\lambda z) = \lambda^i h(z) \quad \text{for all} \quad z \in \mathcal{D} \quad \text{and} \quad \lambda > 0 .$$

Each homogeneous function has a unique degree of homogeneity except for the trivial function which is homogeneous of any degree. A real polynomial in n variables is homogeneous if and only if each nontrivial term in its expansion has the same degree i. Vector- , matrix- and tensor valued functions are said to be homogeneous if all their entries are homogeneous with the same degree of homogeneity.

Rates of convergence are described in the framework of Ortega and Reinbold [10] and matrix related terms are used in agreement with Stewart [11].

Within each of the four chapters the equations are numbered consecutively from 1 to ca.60 and referred to accordingly. For cross references between chapters the equation number is prefixed by the chapter number, e.g. (3.4) refers to the fourth equation in chapter 3.

Frequently used symbols and expressions are listed in the Index.

CHAPTER 1

GENERAL RESULTS ON NEWTON'S METHOD AT SINGULARITIES

1. The Determinant Function in the Neighbourhood of a Singularity

As a polynomial in the entries of the Jacobian ∇f the *determinant function* $\det(\nabla f(x))$ is one time less differentiable than f itself. The determinant function has no particular structure and could be any scalar function that is as often differentiable as the Jacobian is known to be. To see this we define for any scalar function

$$\delta(x) = \delta(\xi_1, \xi_2, .. \xi_n) \in C^1(\mathbb{R}^n)$$

and arbitrary x^* , $f(x^*) \in \mathbb{R}^n$ the vector function

$$f(x) \equiv f(x^*) + \begin{pmatrix} \int_{\xi_1^*}^{\xi_1} \delta(\xi, \xi_2, .. \xi_n) d\xi \\ \xi_2 - \xi_2^* \\ \\ \xi_n - \xi_n^* \end{pmatrix}, \qquad (1)$$

so that

$$\nabla f(x) = \begin{pmatrix} \delta(x) \ , & \int_{\xi_1^*}^{\xi_1} \frac{\partial}{\partial \xi_2} \delta(\xi, \xi_2, .., \xi_n) d\xi, \ldots, & \int_{\xi_1^*}^{\xi_1} \frac{\partial}{\partial \xi_n} \delta(\xi, \xi_2, .. \xi_n) d\xi \\ \\ 0 \ , & & I \end{pmatrix}$$

which implies

$$\det(\nabla f(x)) = \delta(x) \quad \text{for all} \quad x \in \mathbb{R}^n \ .$$

Consequently the *singular set*

$$\delta^{-1}(0) \equiv \{x \in \mathbb{R}^n \mid \delta(x) = 0\}$$

in the neighbourhood of a *singular solution* or *singularity*

$$x^* \in f^{-1}(0) \cap \delta^{-1}(0)$$

can have a rather complicated structure, even if f is highly differentiable and x^* is an isolated solution.

Consider for instance with some $\bar{k} \geq 1$ the function $f \in C^{\bar{k}+1}(\mathbb{R}^n, \mathbb{R}^n)$ as defined by (1) with

$$f(x^*) = 0 \quad \text{and} \quad \delta(\xi_1, .., \xi_n) = \xi_1 (\xi_1^{\bar{k}} \sin \frac{1}{\xi_1})^2 \ . \tag{2}$$

It can easily be seen that the singular set consists of $(n-1)$-dimensional hyperplanes, of which countably many intersect with any neighbourhood of the unique solution x^*. This peculiarity cannot occur if the determinant function has an expansion of the form

$$\delta(x) = \pi_0(x-x^*) + O(\|x-x^*\|^{p+1}) \tag{3}$$

where $\pi_0 \in C^{\infty}(\mathbb{R}^n)$ is a nontrivial homogeneous polynomial of degree $p \geq 1$. The unique polynomial π_0 is simply the leading term in the Taylor expansion of δ at x^* if δ is *nontrivially differentiable* in that not all its existing Lipschitz continuous derivatives vanish at x^*, as they do in the example defined by (2).

LEMMA 1.1 *Arbitrariness of Determinant Function*

If $\delta \in C^1(\mathbb{R}^n)$ has an expansion of the form (3) at $x^* \in \mathbb{R}^n$ then there is a vector-function $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ such that x^* is an isolated, singular solution of f and

$$\det(\nabla f(x)) = \delta(x) \quad \text{for all} \quad x \in \mathbb{R}^n .$$

Proof. Since π_0 is by assumption nontrivial there is a vector $v \in \mathbb{R}^n$ such that $\pi_0(v) \neq 0$. Assuming without loss of generality that v is the first Cartesian base vector, we see that the assertion is true for f as defined by (1) with $f(x^*) = 0$. ////

In general we do not impose the condition that δ is nontrivially differentiable, as this need not be the case when $\operatorname{rank}(\nabla f(x^*)) < n-1$ even though the singular problem may be otherwise well defined.

From now on we consider a given $(\bar{k}+1) \geq 2$ times Lipschitz continuously differentiable vectorfunction $f \in C^{\bar{k}+1,1}(\mathbb{R}^n, \mathbb{R}^n)$ whose Jacobian ∇f has a nullspace N of dimension $m > 0$ at a solution point $x^* \in f^{-1}(0)$ and the determinant function $\delta \equiv \det(\nabla f) \in C^{\bar{k},1}(\mathbb{R}^n)$.

Newton's method has the important and well known property that it is essentially invariant not only with respect to nonsingular linear transformations on the range of f [12], but also with respect to nonsingular affine ones on the domain of f. With A, B nonsingular $n \times n$ matrices and $\bar{y}$ a vector in $\mathbb{R}^n$ we find that the transformed system

$$\tilde{f}(y) = Af(\bar{y}+B(y-\bar{y})) = 0$$

generates from some $y_0 \in \mathbb{R}^n$ a *Newton sequence* $\{y_j\}_{j\geq 0}$ with

$$y_{j+1} = y_j - \nabla\tilde{f}^{-1}(y_j)\tilde{f}(y_j)$$

which is parallel to the Newton sequence $\{x_j\}_{j\geq 0}$ with

$$x_{j+1} = x_j - \nabla f^{-1}(x_j) f(x_j)$$

in that for all $j \geq 0$

$$x_j = \bar{y} + B(y_j - \bar{y})$$

provided this is true for $j = 0$.

With $\nabla f(x^*) = U \Sigma V^T$, $U^T U = I = V^T V$, $\Sigma = \mathrm{diag}(0,0,..,\sigma_{m+1},..\sigma_n)$ the singular value decomposition [11] of the Jacobian at the singularity x^* we may choose $A = U^T$ and

$$B \equiv V \begin{pmatrix} \sigma_n^{-1} & & & & & 0 \\ & \ddots & & & & \\ & & \sigma_n^{-1} & & & \\ & & & \sigma_{m+1}^{-1} & & \\ & & & & \ddots & \\ 0 & & & & & \sigma_n^{-1} \end{pmatrix}$$

so that

$$\nabla \tilde{f}(y^*) = A \nabla f(x^*) B = \begin{pmatrix} 0,0 \\ 0, I_{n-m} \end{pmatrix}$$

where $y^* \equiv \bar{y} + B^{-1}(x^* - \bar{y})$.

Provided the ratio σ_n / σ_{m+1} between the largest and smallest nonvanishing singular values is of moderate magnitude as we will assume, the conditioning of the problem is essentially unchanged. Furthermore the Euclidean norm of f is preserved as $A = U^T$ is orthogonal, so that statements about the reduction of residual norms during line searches apply equivalently to $\tilde{f}$ and f .

Throughout the thesis we can therefore assume without loss of generality that f itself is in *normal form* at x^* in that

$$\nabla f(x^*) = I - P \quad \text{and} \quad N = \mathbb{R}^m \times \{0\}^{n-m} , \tag{4}$$

where P is the orthogonal projection from $\mathbb{R}^n$ onto the nullspace of the transposed $\nabla f^T(x^*)$. In some proofs we set $x^* = 0$ to simplify the notation even further. However all major results are stated in general terms such that they apply directly to nonnormalised cases as well.

With $\nabla f|_N$ and $\nabla f|_{N^\perp}$ the restrictions of ∇f to the nullspace N and its orthogonal complement respectively we can partition the Jacobian as follows:

$$\nabla f = \begin{pmatrix} B \ , \ C^T \\ \\ D \ , \ E \end{pmatrix} \equiv \begin{pmatrix} P\nabla f|_N \ , & P\nabla f|_{N^\perp} \\ \\ (I-P)\nabla f|_N \ , & (I-P)\nabla f|_{N^\perp} \end{pmatrix} ,$$

where at the singularity

$$B(x^*) = C(x^*) = D(x^*) = 0 \quad \text{and} \quad E(x^*) = I \ . \tag{5}$$

Loosely speaking, we may consider the first m equations to be the *singular equations* as their Jacobian (B,C^T) vanishes identically at x^* and the first m variables to be the *singular variables* since none of them enters linearly into any of the n equations so that the corresponding components of x^* are in some sense weakly determined. Applying the implicit function theorem to the $n-m$ remaining *nonsingular equations*, we can theoretically eliminate the last $n-m$ *nonsingular variables* to obtain a new system of m equations in the first m variables with the *reduced Jacobian*

$$G(x) \equiv B(x) - C^T(x)E^{-1}(x)D(x) \ . \tag{6}$$

Since $E(x^*) = I$, the reduced Jacobian is well defined and $\bar{k}$ times Lipschitz continuous differentiable in some ball $\mathcal{B}_{r_b}$ of radius $r_b > 0$ about x^* . Because of the elementary identity

$$\delta(x) = \det(G(x))\det(E(x)) \tag{7}$$

G determines when the full Jacobian is singular, which leads to the following result.

LEMMA 1.2 *Polynomial Expansion of Determinant Function*

Let $f \in C^{\bar{k}+1,1}(\mathbb{R}^n,\mathbb{R}^n)$ have a singularity at $x^* \in f^{-1}(0) \cap \delta^{-1}(0)$ and be such that

$$\delta(x) \neq 0(\|x-x^*\|^{\bar{k}+m}) \ . \tag{8}$$

Then

(i) There are maximal indices $p \in [m,\bar{k}+m-1]$ and $\Delta p \in [\bar{k}+m-p,\bar{k}]$ such that

$$\delta(x) = \sum_{i=0}^{\Delta p-1} \pi_i(x-x^*) + 0(\|x-x^*\|^{p+\Delta p}) \tag{9}$$

where the π_i are homogeneous polynomials of degree $p+i$ and π_0 is nontrivial.

(ii) The discrepancy $\Delta m \equiv p-m \geq 0$ is zero iff the linear operator

$$P\nabla^2 f(x^*)y\big|_N : N \to P(\mathbb{R}^n)$$

is nonsingular for some $y \in \mathbb{R}^n$, which implies that for some constant $\alpha \neq 0$ and all $y \in \mathbb{R}^n$

$$\pi_0(y) = \alpha \det(P\nabla^2 f(x^*)y\big|_N) \ ,$$

where the determinant on the RHS is calculated with respect to two orthonormal bases of the domain N and the range $P(\mathbb{R}^n)$.

Proof. Since ∇f has a $\bar{k}$-th Lipschitz continuous derivative and $E(x^*) = I$, there are Taylor expansions

$$E(x) = I + \sum_{i=1}^{\bar{k}} E_i(x-x^*) + O(\|x-x^*\|^{\bar{k}+1})$$

and

$$G(x) = G_0 + \sum_{i=1}^{\bar{k}} G_i(x-x^*) + O(\|x-x^*\|^{\bar{k}+1}) ,$$

where the entries of the matrices E_i and G_i are homogeneous polynomials of degree i in $(x-x^*) \in \mathbb{R}^n$. The constant G_0 is zero by (5) so that

$$\det(E(x)) = \det(I + \sum_{i=1}^{\bar{k}} E_i(x-x^*)) + O(\|x-x^*\|^{\bar{k}+1})$$

and

$$\det(G(x)) = \det(\sum_{i=1}^{\bar{k}} G_i(x-x^*)) + O(\|x-x^*\|^{m+\bar{k}}) \qquad (10)$$

$$= O(\|x-x^*\|^m) .$$

Here we have used the fact that $\det(G(x))$ is a sum over products of m entries which are all at least linear in $(x-x^*)$. By assumption (8) the polynomial

$$\det(I + \sum_{i=1}^{\bar{k}} E_i(x-x^*))\det(\sum_{i=1}^{\bar{k}} G_i(x-x^*)) = \delta(x) + O(\|x-x^*\|^{\bar{k}+m})$$

must involve terms of order less than $m+\bar{k}$ which can be ordered to form the expansion (9).

(ii) Using again the Taylor expansions of E and G, we derive that

$$\delta(x) = [\det(G_1(x-x^*)) + O(\|x-x^*\|^{m+1})][1+O(\|x-x^*\|)]$$

$$= \det(G_1(x-x^*)) + O(\|x-x^*\|^{m+1}) ,$$

which implies $p = m$ iff $G_1(y)$ is nonsingular for some $y \in \mathbb{R}^n$.

Since both $C(x)$ and $D(x)$ vanish at x^* we have $G(x) = B(x) + O(\|x-x^*\|^2)$ so that the linear term in the Taylor expansion of G is given by

$$G_1(x-x^*) = \nabla B(x^*)(x-x^*) = P\nabla^2 f(x^*)(x-x^*)\big|_N$$

which completes the proof with $\alpha \neq 0$ allowing for the initial transformation of the problem if it was not in normal form originally. ////

From now on we will always assume that (8) is satisfied so that (9) has at least one meaningful term. This may be of order $\bar{k}+\bar{m}-1$ and is thus not necessarily given by the Taylor expansion of the only $\bar{k}$ times differentiable determinant function δ .

2. The Starlike Domain of Invertibility $\mathcal{R}'$

In the one dimensional case $n = 1$ we have $\delta(x) = df(x)/dx$ so that whenever f is nontrivially Lipschitz continuously differentiable

$$\begin{aligned} f(x) &= \int_{x^*}^{x} (\pi_0(1)y^p + O(y^{p+1}))dy \\ &= \frac{\pi_0(1)}{p+1}(x-x^*)^{p+1} + O(|x-x^*|^{p+2}) , \end{aligned}$$

which means that x^* is an isolated point in both the solution set $f^{-1}(0)$ and the singular set $\delta^{-1}(0)$. Furthermore by Theorem 7.2 in [1] for $s = 1$ there is an open neighbourhood of x^* from where Newton's method converges linearly with Q-factor $p/(p+1)$ to x^* .

In the more interesting cases with $n > 1$ it follows from the mean value theorem that x^* is isolated in $\delta^{-1}(0)$ iff δ attains an isolated extremum at x^* . This strong assumption was used in earlier work by Rall [3] and Cavanagh [4], but as we have seen in Lemma 1.1 there is in general no reason why it should be satisfied. Whenever the singularity

x^* is not a priori known to be an isolated point in $\delta^{-1}(0)$ we may ask for which points in its neighbourhood we can guarantee that they do not belong to the singular set.

From now on we will frequently write the elements of $\mathbb{R}^n$ in the form

$$x = x^* + \rho t \ ,$$

where $\rho = \|x-x^*\|$ and t belongs to the unit sphere

$$S \equiv \{t \in \mathbb{R}^n \mid \|t\|=1\}$$

of *directions* in $\mathbb{R}^n$. Because of (9) there are constants $r_b < 0$ and $\omega > 0$ such that for all $x \in \mathbb{R}^n$ with $\rho < r_b$

$$|\delta(x) - \rho^p \pi_0(t)| \leq \omega \rho^{p+1} \ , \tag{11}$$

which implies that the Jacobian is nonsingular at all points in the open set

$$R' \equiv \{x^*+\rho t \mid t \in S, 0<\rho<\bar{r}(t)\} \ , \tag{12}$$

where $\bar{r}$ is the nonnegative continuous function

$$\bar{r}(t) \equiv \min\{r_b \ , \frac{1}{2} |\pi_0(t)|/\omega\} \tag{13}$$

from S to $\mathbb{R}$.

Any open set $A \subseteq \mathbb{R}^n$ which like R' has the property

$$x \in A \Rightarrow (1-\lambda)x + \lambda x^* \in A \quad \text{for} \quad \lambda \in (0,1) \tag{14}$$

will be called a *starlike domain centred at* x^*. It should be noted that in contrast to the usual definition of star-shaped domains in complex analysis, the *central point* x^* does not in general belong to A. The

concept of a starlike domain is clearly invariant with respect to affine transformations, but not necessarily with respect to nonlinear ones. This seems appropriate as the same is true for the concept of a singularity of a system of equations.

As an immediate consequence of the defining condition (14), we note that finite intersections and finite or infinite unions of starlike domains with the same central point are starlike too, so that there are maximal starlike domains with respect to certain properties of their points. In order to characterise the maximal starlike domain contained in some open set we use the notion of tangential directions or *tangents*.

An element $s \in S$ is said to be *tangential* to a given set $A \subseteq \mathbb{R}^n$ at some point x^* if there is a sequence $\{y_j\}_{j\geq 0} \subseteq A - \{x^*\}$ such that

$$y_j \to x^* \quad \text{and} \quad \frac{y_j - x^*}{\|y_j - x^*\|} \to s \quad .$$

It can be easily seen that $s \in S$ is not tangential to A iff there are constants $\theta > 0$ and $\bar{\rho} > 0$ such that

$$\{x^* + \rho t \mid t \in S, \cos^{-1}(t^T s) < \theta, 0 < \rho < \bar{\rho}\} \cap A = \emptyset \ .$$

Consequently the set of tangents of A at x^* is closed in S for any $A \subseteq \mathbb{R}^n$. Now we can give the following convenient characterisation of starlike domains and subdomains.

THEOREM 1.3 *Starlike Domains and Subdomains*

Let $\hat{A}$ be an open set and x^* a point in $\mathbb{R}^n$. Then

(i) The nonnegative *boundary function*

$$a(t) \equiv \begin{cases} 0 & \text{if } t \text{ is tangential to } \mathbb{R}^n - \hat{A} \text{ at } x^* \\ \sup \{\bar{\rho} \mid \{x^* + \rho t\}_{0<\rho<\bar{\rho}} \subset \hat{A}\} & \text{otherwise} \end{cases} \tag{15}$$

from the unit sphere S to $\mathbb{R} \cup \{\infty\}$ is lower semicontinuous, and the set

$$A \equiv \{x^*+\rho t \mid t\epsilon S, 0<\rho<a(t)\} \tag{16}$$

is the maximal starlike domain centred at x^* and contained in $\hat{A}$.

(ii) The set of *excluded* directions

$$a^{-1}(0) = \{t\epsilon S \mid A\cap\{x^*+\rho t\}_{\rho>0} = \emptyset\} \tag{17}$$

equals the set of tangents of $\mathbb{R}^n-\hat{A}$ at x^* and is closed in S.

(iii) Any starlike domain contains a starlike subdomain with continuous boundary function and the same set of excluded directions.

Proof. Firstly we show (ii) with a defined by (15) and A by (16).

(ii) The identity (17) follows immediately from (16). Since

$$\sup\{\bar{\rho} \mid \{x^*+\rho t\}_{0<\rho<\bar{\rho}} \subset \hat{A}\} = 0$$

implies the existence of a sequence $\rho_j \to 0$ with $x^*+\rho_j t \notin \hat{A}$, any direction in $a^{-1}(0)$ must be tangential to $\mathbb{R}^n - \hat{A}$. The converse holds by definition of a. The fact that $a^{-1}(0)$ is closed follows either from it being a set of tangents or the lower semicontinuity of a which will be established next.

(i) Suppose a is not lower semicontinuous at some $t \in S$. According to the equivalent definitions of lower semicontinuity given on page 40 in [13] there must be a sequence of directions $t_j \to t$ with $a(t_j) \to \alpha < a(t)$. Since a is by definition nonnegative this can only be the case if $a(t) > 0$, and consequently t itself and all but finitely many of the t_j are not tangential to $\mathbb{R}^n-\hat{A}$ at x^*. Then there must

be a sequence of positive numbers $\varepsilon_j \to 0$ such that at most finitely many points of the converging sequence

$$x_j \equiv x^* + (a(t_j) + \varepsilon_j)t_j \to x^* + \alpha t$$

belong to $\hat{A}$. This contradicts the openess of $\hat{A}$ since α must be positive so that $x^* + \alpha t \in \hat{A}$, as otherwise t would be tangential to $\mathbb{R}^n - \hat{A}$. Thus a is lower semicontinuous.

Any set of the form (16) does obviously satisfy (14), so that only the openess and maximality of A remains to be shown. For any converging sequence

$$x_j = x^* + \rho_j t_j \to x^* + \rho t \in A$$

the semicontinuity of a ensures

$$\rho_j \to \rho < a(t) \leq \liminf_{j\to\infty} a(t_j) ,$$

so that all but finitely many of the x_j must belong to A which is therefore open.

Now consider any other starlike domain $\tilde{A} \subseteq \hat{A}$ with the boundary function

$$\tilde{a}(t) \equiv \sup\{\rho \mid x^*+\rho t \in \tilde{A}\} .$$

For any t that is not tangential to $\mathbb{R}^n - \hat{A}$ we derive from (15) that $\tilde{a}(t) \leq a(t)$. For any other $t \in S$ there exists a sequence

$$\{x^*+\rho_j t_j\}_{j\geq 0} \subseteq \mathbb{R}^n - \hat{A}$$

with $\rho_j \to 0$ and $t_j \to t$, so that by the lower semicontinuity of $\tilde{a}$

$$\tilde{a}(t) \le \liminf_{j\to\infty} \tilde{a}(t_j) \le \liminf_{j\to\infty} \rho_j = 0 \ .$$

Consequently the directions tangential to $\mathbb{R}^n - \hat{A}$ are excluded from all starlike domains contained in $\hat{A}$ and A is maximal.

(iii) Since the intersection of a starlike domain with any open ball about the central point is starlike too, we may consider without loss of generality a starlike domain A of the form (16) with $a \le 1$. With the convention $\min(\emptyset) = 180^\circ$ the angle

$$\phi(t) \equiv \frac{1}{2} \min\{\cos^{-1}(t^T s) \mid s \in S \cap a^{-1}(0)\} \le 90^\circ$$

is a nonnegative continuous function from S to $\mathbb{R}^n$ with $\phi^{-1}(0) = a^{-1}(0)$. The starlike domain defined by the boundary function

$$\tilde{a}(t) \equiv \inf\left\{ \frac{a(s)}{1-\cos^{-1}(s^T t)/\phi(t)} \;\middle|\; s \in S, \cos^{-1}(t^T s) < \phi(t) \right\}^{\phi(t)/90^\circ}$$

is obviously contained in A and it can be shown that $\tilde{a}$ is continuous as required. However the proof is rather tedious and we prefer a less constructive approach based on *partitions of unity* as described in [14]. According to Remark 2.1.4 in that book the C^∞ *submanifold* $S - a^{-1}(0)$ of S is *paracompact* and has therefore, by proposition 1.2.1, a *locally finite covering* $\{V_i\}_{i\in I}$ such that

$$S - a^{-1}(0) = \cup\{V_i \mid i \in I\} \ ,$$

$$\bar{V}_i \equiv \text{closure}(V_i) \subseteq S - a^{-1}(0) \quad \text{for all} \quad i \in I \ , \tag{18}$$

and each t has a neighbourhood in $S - a^{-1}(0)$ that is disjoint from all but finitely many of the V_i. Furthermore there exists by Theorem 2.2.14 a family of functions called a partition of unity

$$\{\eta_i\}_{i\in I} \subseteq C^\infty(S - a^{-1}(0), \mathbb{R})$$

such that

$$\eta_i \geq 0 \ , \ \eta_i(t) = 0 \quad \text{if} \quad t \notin V_i \ ,$$

and

$$\sum_{i \in I} \eta_i(t) = 1 \quad \text{for all} \quad t \in S - a^{-1}(0) \ .$$

Since lower semicontinuous functions attain by proposition 2.10 in [15] minima on compact subsets of their domain we have by (18) for all $i \in J$

$$a_i \equiv \min\{a(t) \mid t \in \bar{V}_i\} > 0 \ .$$

Now it can be easily checked that

$$\hat{a}(t) \equiv \sum_{i \in J} a_i \eta_i(t) \leq a(t) \leq 1$$

is a continuous positive function on $S - a^{-1}(0)$. Since $\hat{a}$ is bounded the boundary function

$$\tilde{a}(t) \equiv \begin{cases} 0 & \text{if} \quad t \in a^{-1}(0) \\ \\ \hat{a}(t)\phi(t)/90^\circ & \text{otherwise} \end{cases}$$

is continuous on S and defines a subdomain of A with the same set of excluded directions $a^{-1}(0)$.

By construction R' is a starlike *domain of invertibility*, i.e. a subset of $\mathbb{R}^n - \delta^{-1}(0)$. We know from Theorem 1.3 (ii) that the full domain of invertibility $\mathbb{R}^n - \delta^{-1}(0)$ contains a maximal starlike subdomain at x^* with corresponding minimal set of excluded directions which equals the set of tangents of $\delta^{-1}(0)$ at x^* . In constructing starlike domains with some particular property our main aim is to keep the set of excluded directions as small as possible. The actual values of the boundary function at *included*, (i.e. not excluded) directions depend on the magnitude of higher derivatives as well as technicalities of the mathematical derivation

and therefore are considered to be of lesser importance.

To justify this approach we define for an arbitrary set $A \subseteq \mathbb{R}^n$ its *upper outer density* at x^* as

$$\tau^*(A) \equiv \limsup_{\rho \to 0} L_n^*(A \cap B_\rho^-)/L_n(B_\rho^-) \in [0,1] . \tag{19}$$

Here L_n^* denotes the outer measure induced by the n-dimensional Lebesgue measure L_n such that for any subset $C \subseteq \mathbb{R}^n$

$$L_n^*(C) = \inf\{L_n(\hat{C}) \mid \hat{C} \text{ is measurable and contains } C\} .$$

The concept of upper outer density was taken from [16] where hypercubes are used instead of the balls B_ρ^-. Since we are only interested in the upper outer density of sets at the singular solution point, the explicit reference to x^* will sometimes be omitted. Without using the corresponding concept of lower outer density, we refer to $\tau^*(A)$ as the *outer density* of A if the limit superior in (19) is in fact a limit. This must always be the case if $\tau^*(A) = 0$. If A is measurable the outer measure of $A \cap B_\rho^-$ reduces to the proper Lebesgue measure and $\tau^*(A)$ will be referred to as the *upper density* or *density* of A respectively. In the latter case $\tau^*(A)$ can be interpreted, loosely speaking, as the probability that a given point, which is very close to x^* belongs to A. Starlike domains are measurable as they are open and have a well defined density at their central point which is completely determined by the set of excluded directions.

LEMMA 1.4 *Density of Starlike Domains at* x^*

(i) The density of a starlike domain A with boundary function $a : S \to \mathbb{R}$ is given by

$$\tau^*(A) = 1 - L_{n-1}(a^{-1}(0))/L_{n-1}(S) .$$

(ii) If A has a density at x^*, then

$$\mathbb{R}^n - A \text{ has the density } 1 - \tau^*(A) ,$$

$$A \cap C = \emptyset \Rightarrow \tau^*(C) \leq 1 - \tau^*(A)$$

where C may be any subset of $\mathbb{R}^n$.

Proof. By Fubini's theorem [16] we can obtain the measure of $A \cap \mathcal{B}_{\bar{\rho}}$ by integrating over the intersections of A with spheres of radius $\leq \bar{\rho}$ so that

$$L_n(A \cap \mathcal{B}_{\bar{\rho}}) = \int_0^{\bar{\rho}} L_{n-1}\{x^* + \rho t \mid t \in S, a(t) > \rho\} d\rho .$$

Changing the integration variable from ρ to $\mu \equiv \rho/\bar{\rho}$ and expanding the spheres by a factor of $1/\bar{\rho}$, we find

$$L_n(A \cap \mathcal{B}_{\bar{\rho}}) = \bar{\rho}^n \int_0^1 L_{n-1}\{x^* + \mu t \mid t \in S, a(t) > \mu\bar{\rho}\} d\mu .$$

With $L_n(\mathcal{B}_{\bar{\rho}}) = \bar{\rho}^n L_n(\mathcal{B}_1)$, we derive from the Lebesgue Dominated Convergence Theorem applied to the characteristic functions of the sets in the integrand

$$\begin{aligned}
\tau^*(A) &= \int_0^1 \lim_{\bar{\rho} \to 0} L_{n-1}\{x^* + \mu t \mid t \in S, a(t) > \mu\bar{\rho}\} d\mu / L_n(\mathcal{B}_1) \\
&= \int_0^1 \mu^{n-1} d\mu \; L_{n-1}\{x^* + t \mid t \in S - a^{-1}(0)\} / L_n(\mathcal{B}_1) \\
&= \frac{1}{n} \left[L_{n-1}(S) - L_{n-1}(a^{-1}(0)) \right] / L_n(\mathcal{B}_1) ,
\end{aligned}$$

which completes the proof of (i) as we must obviously have $\tau^*(A) = 1$ if $a^{-1}(0) = \emptyset$.

(ii) The complement $\mathbb{R}^n - A$ is also measurable so that

$$L_n^*((\mathbb{R}^n - A) \cap \mathcal{B}_{\bar{\rho}}) / L_n(\mathcal{B}_{\bar{\rho}}) = 1 - L_n(A \cap \mathcal{B}_{\bar{\rho}}) / L_n(\mathcal{B}_{\bar{\rho}}) ,$$

which gives for $\bar{\rho} \to 0$ the limit $\tau^*(\mathbb{R}^n - A) = 1 - \tau^*(A)$. The second assertion is an immediate consequence as by definition of the outer measure

$$L_n^*(C \cap B_{\bar{\rho}}) / L_n(B_{\bar{\rho}}) \leq L_n((\mathbb{R}^n - A) \cap B_{\bar{\rho}}) / L_n(B_{\bar{\rho}}) \ . \qquad ////$$

By definition of $\bar{r}$ in (13), the set $\bar{r}^{-1}(0)$ of directions that are excluded from R' is the solution set of the restriction of the nontrivial homogeneous polynomial π_0 to S which is a nontrivial analytic function on the smooth manifold S . As stated on page 240 in [17] the solution sets of nontrivial analytic functions have zero Lebesgue measure so that $L_{n-1}(\bar{r}^{-1}(0)) = 0$. Now we compile the properties of R' in the following lemma.

LEMMA 1.5 *The Starlike Domain of Invertibility* R'

Under the assumptions of Lemma 1.2 let $\bar{r}$ and R' be defined by (12) and (13) respectively. Then

(i) The starlike domain of invertibility R' includes the set of *regular directions* $S' \equiv S - \pi_0^{-1}(0)$ and excludes the set of *irregular* directions

$$\bar{r}^{-1}(0) = \{t \in S \mid R' \cap \{x^* + \rho t\}_{\rho > 0} = \emptyset\} = S \cap \pi_0^{-1}(0)$$

which has Lebesgue measure zero in S .

(ii) R' has density 1 at x^* and any set in its complement $\mathbb{R}^n - R'$, in particular the singular set $\delta^{-1}(0)$, has outer density zero.

(iii) At any irregular direction $t \in \pi_0^{-1}(0)$, that is not tangential to the singular set $\delta^{-1}(0)$, the polynomial π_0 attains a local extremum.

(iv) The smallest singular value $\sigma(x)$ of $\nabla f(x)$ satisfies

$$\sigma(x^*+\rho t) = O(\rho^{p/m}) \quad \text{for all} \quad t \in S$$

$$\sigma(x^*+\rho t) = o(\rho^{p/m}) \quad \text{for all} \quad t \in \pi_0^{-1}(0)$$

$$\sigma(x^*+\rho t) = o(\rho^{\Delta m+1}) \quad \text{implies} \quad t \in \pi_0^{-1}(0) \ .$$

Proof.

(i) has already been established. (ii) follows from (i) by Lemma 1.4. For the proof of (iii) assume $x^* = 0$.

(iii) Suppose $\pi_0(t) = 0$ is not a local extremum. Then there must be sequences $\{t_j^-\}_{j\geq 1}$ and $\{t_j^+\}_{j\geq 1}$ in S such that

$$\lim_{j\to\infty} t_j^- = \lim_{j\to\infty} t_j^+ = t \quad \text{and} \quad \pi_0(t_j^-) < 0 < \pi_0(t_j^+) \quad \text{for all} \quad j \ .$$

Because of (9) there must be a sequence of multipliers $\{\rho_j\}_{j\geq 1} \subset \mathbb{R}$ such that

$$\delta(\rho_j t_j^-) < 0 < \delta(\rho_j t_j^+) \ ,$$

which implies by the meanvalue theorem the existence of vectors

$$y_j = \rho_j(\alpha_j t_j^- + (1-\alpha_j) t_j^+) \quad , \quad \alpha_j \in (0,1)$$

with $\delta(y_j) = 0$ for all $j \geq 1$. Since the y_j/ρ_j are convex combinations of the t_j^- and t_j^+, we must have

$$\lim_{j\to\infty} y_j/\|y_j\| = t$$

so that t is tangential to $\delta^{-1}(0)$. Consequently any $t \in \pi_0^{-1}(0)$ that is not tangential must be a minimiser or maximiser of π_0 .

(iv) Since all but the smallest m singular values of ∇f are non-zero at x^*, we derive from (9)

$$\sigma(x^*+\rho t) = O(\delta(x^*+\rho t)^{1/m}) = \begin{cases} O(\rho^{p/m}) & \text{for} \quad t \in S \\ o(\rho^{p/m}) & \text{for} \quad t \in \pi_0^{-1}(0) \end{cases} .$$

Applying Lemma 1.2 (i) to the minors of the Jacobian ∇f which are the entries of the adjugate $\operatorname{adj}(\nabla f^T)$, we find that

$$\operatorname{adj}(\nabla f(x^*+\rho t)) = O(\rho^{m-1}) ,$$

since the nullspace of any minor at x^* must have a dimension greater than or equal to $m-1$. Now we obtain by Cramer's rule for regular $t \in S'$

$$\sigma^{-1}(x^*+\rho t) = \|\nabla f^{-1}(x^*+\rho t)\|$$

$$= \|\operatorname{adj}\nabla f(x^*+\rho t)\| \; / \; |\delta(x^*+\rho t)| = O(\rho^{-\Delta m-1})$$

so that any $t \in S$ must be irregular if $\sigma(x^*+\rho t) = o(\rho^{\Delta m+1})$. ////

Since nontrivial homogeneous polynomials are unbounded and all their stationary points have zero value, it is quite likely that they have no extrema besides possibly the origin. In this case the set of directions excluded from R' is minimal, so that the boundary function of the maximal starlike domain of invertibility differs from $\bar{r}$ only in size but not in sign. If there are irregular directions t at which π_0 attains an extremum, we could theoretically enlarge R' by including either t or $-t$, provided π_1 exists and $\pi_1(t) \neq 0$. However such extensions seem of little use and would complicate the analysis of Newton's method significantly. When $p = m$ or $m = 1$, Lemma 1.5 (iv) implies

$$\sigma(x^*+\rho t) = o(\rho^{p/m}) \quad \Leftrightarrow \quad t \in \pi_0^{-1}(0) ,$$

so that the irregular directions are exactly those along which the Jacobian is particularly illconditioned unless $m = n$ in which case it vanishes

completely at x^* . Thus we can conclude that we do not lose much by confining our analysis of Newton iterations in the neighbourhood of x^* to the starlike domain R' .

3. Rational Expansion of the Newtonian Iteration Function

The convergence of some iteration $x_{j+1} = g(x_j)$ to a fixed point x^* is frequently demonstrated by showing that the *iteration function* g has in the neighbourhood of x^* a Jacobian with spectral radius less than 1 . In our singular case such a contracting linear approximation to the *Newtonian iteration function*

$$g(x) \equiv x - \nabla f^{-1}(x) f(x) \tag{20}$$

does not in general exist, since g is undefined in the singular set $\delta^{-1}(0)$ and usually unbounded on its domain $\mathbb{R}^n - \delta^{-1}(0)$. Using again the adjugate $\mathrm{adj}(\nabla f(x))$, we can write

$$\begin{aligned} g(x) - x^* &= [\delta(x)(x-x^*) - \mathrm{adj}(\nabla f^T(x)]f(x))/\delta(x) \\ &= \mathrm{adj}(\nabla f^T(x))[\nabla f(x)(x-x^*) - f(x)]/\delta(x) \ . \end{aligned}$$

Under the assumptions of Lemma 1.2, the matrices and vectors in the numerator as well as the scalar $\delta(x)$ in the denominator have Taylor like expansions in terms of $(x-x^*)$. Hence we can approximate $g(x) - x^*$ in R' , where it is well defined, by some form of rational expansion as developed below.

THEOREM 1.6 *Rational Expansion of Newtonian Iteration Function*

Under the assumptions of Lemma 1.2 let g be defined by (20). Then

(i) There are $\Delta p \geq 1$ vector functions

$$u_i \in C^\infty(\mathbb{R}^n, \mathbb{R}^n) \quad \text{for} \quad i = 1-\Delta m, \quad , \Delta p - \Delta m$$

whose components are polynomials such that the rational vector functions

$$g_i \equiv u_i / \pi_0^{i+\Delta m} \in C^\infty(\mathbb{R}^n - \pi_0^{-1}(0), \mathbb{R}^n)$$

are homogeneous of degree i, and for all $x^* + \rho t \in R'$

$$\|g(x^*+\rho t) - x^* - \sum_{i=1-\Delta m}^{\Delta p-\Delta m} \rho^i g_i(t)\| \le \gamma \frac{\rho^{\Delta p-\Delta m+1}}{|\pi_0(t)|^{\Delta p+1}}, \tag{21}$$

where γ is a suitable positive constant.

(ii) There is an index $k \in [1,\bar{k}]$ such that

$$P\nabla^{k+1} f(x^*) \neq 0 \quad \text{and} \quad P\nabla^q f(x^*) = 0 \quad \text{for} \quad q \in [1,k],$$

which will be called the *order of the singularity*.

(iii) The *degree of the singularity*, i.e. the lowest index $\hat{i}$ for which $u_{\hat{i}}$ and consequently $g_{\hat{i}}$ are nontrivial, cannot be greater than 1, and $g_{\hat{i}}(t)$ belongs to N for all $t \in S'$.

(iv) For any regular t the vectors $g_i = g_i(t)$ solve the block triangular Toeplitz system

$$\begin{pmatrix} A_1 & 0 & \cdots & 0 \\ A_2 & A_1 & \ddots & \vdots \\ A_3 & A_2 & A_1 & \ddots \\ \vdots & \ddots & \ddots & \ddots \\ A_{\Delta p} & \cdots & A_3 & A_2 & A_1 \end{pmatrix} \begin{pmatrix} g_{1-\Delta m} \\ g_{2-\Delta m} \\ \vdots \\ g_0 \\ g_1 \\ g_2 \\ \vdots \\ g_\ell \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \frac{1}{2}A_2 t \\ \vdots \\ \frac{\ell-1}{\ell}A_\ell t \end{pmatrix} \tag{22}$$

where $\ell = \Delta p - \Delta m$ and $A_i = A_i(t) = \nabla^i f(x^*) t^{i-1}/(i-1)!$

(v) If the vector $(y_{1-\Delta m}^T, y_{2-\Delta m}^T, \ldots y_{q-\Delta m})^T \in \mathbb{R}^{n \cdot q}$ solves the first $q \leq \Delta p$ "rows" of the linear block system (22) for some regular $t \in S'$ then its first $q - \Delta m - 1$ component vectors must be correct in that

$$y_i = g_i(t) \quad \text{for} \quad i = 1-\Delta m, \ldots, q - 2\Delta m - 1 \ .$$

Proof.

(i) Note that $\Delta p \geq 1$ by Lemma 1.2 (i). Deviding (9) by ρ^p, we obtain for $x = x^* + \rho t \neq x^*$

$$\delta(x^*+\rho t)/\rho^p = \sum_{i=0}^{\Delta p-1} \rho^i \pi_i(t) + O(\rho^{\Delta p}) \ . \tag{23}$$

Because of (11) and the definition of $\bar{r}$ in (13) we have for $x = x^* + \rho t \in R'$ the lower bound

$$|\delta(x^*+\rho t)/\rho^p| \ > \ \frac{1}{2} \ |\pi_0(t)| \ . \tag{24}$$

There is a unique polynomial of the form

$$\nu(x^*+\rho t) \equiv \sum_{j=0}^{\Delta p-1} \rho^j \nu_j(t)$$

such that for any regular t

$$\nu(x^*+\rho t) \sum_{j=0}^{\Delta p-1} \rho^j \pi_j(t) = 1 + O(\rho^{\Delta p}) \ , \tag{25}$$

where the remainder on the RHS is usually not uniform in $t \in S'$. Since the π_j are homogeneous polynomials of degree $(p+j)$, we can show by induction that the ν_j are of the form $\varepsilon_j/\pi_0^{j+1}$, where ε_j is a homogeneous polynomial of degree $j(p+1)$. This is true for $j = 0$ as obviously $\nu_0 = 1/\pi_0$. Now suppose the assertion holds for all $q \in [0,j]$ with $j \geq 0$. Identifying terms in (25), we find the recurrence

$$\nu_{j+1} = -\Big(\sum_{q=0}^{j} \pi_{j+1-q}\nu_q\Big)/\pi_0$$

$$= -\sum_{q=0}^{j} \pi_{j+1-q}\varepsilon_q/\pi_0^{q+2}$$

$$= -\Big(\sum_{q=0}^{j} \pi_{j+1-q}\varepsilon_q\pi_0^{j-q}\Big)/\pi_0^{j+2}$$

Each term in the sum has the same degree

$$(j+1-q+p) + q(p+1) + p(j-q) = (p+1)(j+1) ,$$

so that they add up to a homogeneous polynomial ε_{j+1} of degree $(p+1)(j+1)$ as asserted. Since $\bar{r}$ is continuous on the compact domain S it must be bounded, and there is a constant $\gamma_1 > 0$ such that for all $x^*+\rho t \in R'$

$$|\nu(x^*+\rho t)| \le \gamma_1/|\pi_0(t)|^{\Delta p} .$$

Multiplying (23) by $\nu(x^*+\rho t)$ we obtain for some $\gamma_2 > 0$

$$|\nu(x^*+\rho t)\delta(x^*+\rho t)/\rho^p - 1| \le \frac{1}{2}\gamma_2(\rho/\pi_0(t))^{\Delta p} .$$

After division by $\delta(x^*+\rho t)/\rho^p$ we obtain by (24) for $\rho < \bar{r}(t)$

$$|\rho^p/\delta(x^*+\rho t) - \nu(x^*+\rho t)| \le (\gamma_2/\pi_0(t))(\rho/\pi_0(t))^{\Delta p} . \qquad (26)$$

Applying Lemma 1.2 (i) to the entries of the $\bar{k}$ times Lipschitz continuously differentiable adjugate, we derive the expansion

$$\mathrm{adj}(\nabla f^T(x^*+\rho t))/\rho^{m-1} = \sum_{j=0}^{\bar{k}-1} \rho^j F_j(t) + O(\rho^{\bar{k}}) , \qquad (27)$$

where the entries of the matrices F_j are homogeneous polynomials of degree $j+m-1$ in $t \in S$. As a consequence of (ii) at most the first Δm matrices $\{F_j\}_{j=0,\ldots,\Delta m-1}$ can vanish identically. However this fact is

not needed in this proof. With A_j as defined in (iv), we find the Taylor expansion

$$\rho\nabla f(x^*+\rho t)t - f(x^*+\rho t) = \sum_{j=2}^{\bar{k}+1} \rho^j (1-\tfrac{1}{j}) A_j(t)t + O(\rho^{\bar{k}+2}) \qquad (28)$$

so that

$$\nabla f(x^*+\rho t)t/\rho - f(x^*+\rho t)/\rho^2 = \sum_{j=0}^{\bar{k}-1} \rho^j v_j(t) + O(\rho^{\bar{k}}) ,$$

where the components of the vector function

$$v_j(t) \equiv \frac{j+1}{j+2} A_{j+2}(t)t$$

are homogeneous polynomials of degree $j+2$ in $t \in S$. Multiplication of (28) from the left by (27) gives for $x = x^* + \rho t \neq x^*$

$$[(x-x^*)\delta(x) - \mathrm{adj}(\nabla f^T(x))f(x)]/\rho^{m+1} = \sum_{j=0}^{\bar{k}-1} \rho^j w_j(t) + O(\rho^{\bar{k}}) , \qquad (29)$$

where the vector functions

$$w_j(t) \equiv \sum_{q=0}^{j} F_q(t) v_{j-q}(t)$$

are homogeneous polynomials of degree $j+m+1$ in $t \in S$. Since by definition (20) for $x = x^* + t \in R'$

$$g(x^*+\rho t) - x^* = \rho^{1-\Delta m} \frac{[\rho\delta(x)t - \mathrm{adj}(\nabla f^T(x))f(x)]/\rho^{m+1}}{\delta(x)/\rho^p} ,$$

equation (21) follows from (26) and (29) with

$$u_{q+1-\Delta m} \equiv \left(\sum_{j=0}^{q} \nu_j w_{q-j} \right) \pi_0^{q+1} \qquad \text{for } q = 0,\ldots,\Delta p-1$$

$$= \left(\sum_{j=0}^{q} \varepsilon_j \pi_0^{q-j} w_{q-j} \right) .$$

Since each term in the sum has the degree

$$j(p+1) + (q-j)p + (q-j+m+1) = q(p+1) + m + 1 ,$$

their sum is a homogeneous polynomial of degree $q(p+1)+m+1$, so that with $i = q+1-\Delta m$ the rational vector function $g_i = u_i/\pi_0^{i+\Delta m}$ is homogeneous of degree

$$(i+\Delta m-1)(p+1)+m+1 - p(i+\Delta m) = i \ .$$

(ii) We have to exclude the possibility

$$P\nabla^q f(x^*) = 0 \quad \text{for} \quad q = 1,..,\bar{k}+1 \tag{30}$$

which would imply $G(x) = O(\rho^{\bar{k}+1})$, so that by (7) $\delta(x) = O(\rho^{m(\bar{k}+2)})$ which contradicts assumption (8).

(iii) The hypothesis that the $\{g_i\}_{i\in[1-\Delta m,1]}$ vanish identically in $t \in S'$ will be shown by induction to imply (30), which has been ruled out in (ii). If (30), which is obviously true for $q = 1$, holds for all $q \le i \ge 1$ we obtain, multiplying the $(\Delta m+1+i)$-th "row" of the linear system (22) from the left by P,

$$0 = \sum_{q=1}^{i} P A_q(t) g_{2+i-q} = \frac{i}{i+1} P A_{i+1} t$$

$$= \frac{i}{(i+1)!} P \nabla^{i+1} f(x^*) t^{i+1} \ .$$

Since this identity holds for all $t \in S'$ which is open in S we must have $P\nabla^{i+1} f(x^*) = 0$.

The second assertion in (iii) is a consequence of the linear system (22) whose $(\hat{i}+\Delta m)$-th "row" reads simply $A_1 g_{\hat{i}} = 0$.

(iv) Based on the expansions (21), (28) and

$$\nabla f(x^*+\rho t) = \sum_{j=1}^{\bar{k}+1} \rho^{j-1} A_j(t) + O(\rho^{\bar{k}+1})$$

the linear system can be obtained by identifying terms in the equation

$$\nabla f(x^*+\rho t)\,(g(x^*+\rho t)-x^*) = \rho \nabla f(x^*+\rho t)\,t - f(x^*+\rho t) \ .$$

(v) Obviously we must have

$$\nabla f(x^*+\rho t)\left(g(x^*+\rho t) - x^* - \sum_{i=1-\Delta m}^{q-\Delta m} \rho^i y_i\right) = O(\rho^{q-\Delta m+1})$$

so that by Lemma 1.5 (iv) for regular t with (21)

$$\sum_{i=1-\Delta m}^{q-\Delta m} \rho^i (g_i - y_i) + O(\rho^{q-\Delta m+1}) = O(\rho^{q-2\Delta m}) \ ,$$

which implies $g_i = y_i$ for $i = 1-\Delta m, \ldots, q-2\Delta m-1$. ////

The order k of the singularity, as defined in Theorem 1.6 (ii), gives the order of the first nonvanishing term in the Taylor expansion of the Jacobian (B, C^T) of the singular equations. In the scalar case $k+1$ is commonly called the multiplicity of the root x^* . The degree $\hat{i}$ of the singularity in the sense of Theorem 1.6 (iii) has apparently not been discussed in the literature before. At least in the context of numerical methods the degree seems to give a more fundamental classification than the order. In the nonsingular case π_0 is a constant and (21) reduces to a Taylor expansion with leading term of order $\hat{i} = 2$. Thus we can think of a nonsingular solution as a second degree singularity, which suggests some correspondence between the order of the Newton process at a singularity and its degree. This link is certainly tenuous as we can see from the following family of examples which illustrate the results of the Theorem 1.6.

For some integer $\Delta m \geq 0$ let f be defined as

$$f(\xi,\zeta) = \begin{pmatrix} \frac{1}{2}\zeta^2 \\ \zeta - \frac{1}{1+\Delta m}\,\xi^{1+\Delta m} \end{pmatrix} \qquad (31)$$

so that the Jacobian and its determinant are given by

$$\nabla f(\xi,\zeta) = \begin{pmatrix} 0 & , & \zeta \\ -\xi^{\Delta m} & , & 1 \end{pmatrix}$$

and

$$\delta(\xi,\zeta) = \zeta\xi^{\Delta m} .$$

The unique solution $x^* = (\xi^*,\zeta^*)^T = 0$ belongs to the singular set $\delta^{-1}(0)$, which consists of the ξ-axis and the ζ-axis if $\Delta m > 0$. Since δ is of order $p = \Delta m+1$ in $\rho = \|x\| = \sqrt{\xi^2+\zeta^2}$, and $m = \operatorname{rank}(\nabla f(x^*)) = 1$, the use of $\Delta m = p-m$ is consistent with its definition in Lemma 1.2. An elementary calculation yields the Newtonian iteration function as

$$g(\xi,\zeta) = \left(\frac{\Delta m}{1+\Delta m}\,\xi + \frac{1}{2}\,\zeta/\xi^{\Delta m},\ \frac{1}{2}\,\zeta\right)^T , \tag{32}$$

which can be rewritten with $(\xi,\zeta)^T = \rho t^T = \rho(\mu,\lambda)^T$ as

$$g(\xi,\zeta) = \rho^{1-\Delta m} g_{1-\Delta m}(t) + \rho g_1(t) ,$$

where

$$g_{1-\Delta m}(\mu,\lambda) = \left(\frac{1}{2}\,\lambda/\mu^{\Delta m},0\right)^T$$

and

$$g_1(\mu,\lambda) = \left(\frac{\Delta m}{1+\Delta m}\,\mu\,,\frac{1}{2}\,\lambda\right)^T .$$

Since $g_{1-\Delta m}$ is nontrivial the degree is given by $\hat{1} = 1-\Delta m$. For any Newton sequence $\{x_j = (\xi_j,\zeta_j)^T\}_{j\geq 0}$ we obtain from (32) the recurrences

$$\zeta_{j+1} = \frac{1}{2}\,\zeta_j \quad \text{and} \quad \xi_{j+1} = \frac{\Delta m}{1+\Delta m}\,\xi_j + \frac{1}{2}\,\zeta_j\xi_j^{-\Delta m} . \tag{33}$$

We analyze only the case where $(\xi_0,\zeta_0)^T$, and consequently all subsequent iterates belong to the first quadrant

$$Q \equiv \{(\xi,\zeta)^T \mid \xi > 0 < \zeta\}$$

which is a starlike domain. To this end we consider the ratio

$$\alpha_j = \xi_j \zeta_j^{-1/(1+\Delta m)} \quad \text{for} \quad j \geq 0$$

which satisfies the recurrence

$$\alpha_{j+1} = h(\alpha_j) \equiv \left(\frac{\Delta m}{1+\Delta m}\,\alpha_j + \frac{1}{2}\,\alpha_j^{-\Delta m}\right) 2^{1/(1+\Delta m)} .$$

The iteration function h has the derivative

$$h'(\alpha) = \Delta m\left(\frac{1}{1+\Delta m} - \frac{1}{2}\,\alpha^{-\Delta m-1}\right) 2^{1/(1+\Delta m)} ,$$

which vanishes at the minimizer

$$\hat{\alpha} \equiv [(1+\Delta m)/2]^{1/(1+\Delta m)}$$

and increases monotonically in $\alpha > 0$ towards the limit

$$\lim_{\alpha\to\infty} h'(\alpha) = \left[2\left(1 - \frac{1}{1+\Delta m}\right)^{(1+\Delta m)}\right]^{1/(1+\Delta m)} < \left(\frac{2}{e}\right)^{1/(1+\Delta m)} .$$

Here we have used the Taylor expansion of $e^{-1/(1+\Delta m)}$ to obtain the inequality on the right.

Provided $\alpha_0 > 0$ we have for $j \geq 1$

$$\alpha_j = h(\alpha_{j-1}) \geq h(\hat{\alpha}) = (1+\Delta m)^{1/(1+\Delta m)} > \hat{\alpha} ,$$

so that for $j \geq 2$ with some mean values $\varepsilon_j \in (h(\hat{\alpha}),\infty)$

$$(\alpha_{j+1}-\alpha_j)/(\alpha_j-\alpha_{j-1}) = h'(\varepsilon_j) \in [0,(2/e)^{1/(1+\Delta m)}] ,$$

which ensures that the sequence converges at least Q-linearly to the unique fixed point

$$\alpha^* = \left[2^{\Delta m/(1+\Delta m)} - 2\,\frac{\Delta m}{1+\Delta m}\right]^{-1/(1+\Delta m)} .$$

For the Newton iterates $\{(\xi_j,\zeta_j)^T\}_{j\geq 0}$ themselves this has the following consequences. After the first step we have

$$\zeta_j = (\xi_j/\alpha_j)^{1+\Delta m} \leq \frac{1}{(1+\Delta m)}\,\xi_j^{1+\Delta m} ,$$

so that by (33)

$$\xi_{j+1} \leq \frac{(\frac{1}{2}+\Delta m)}{(1+\Delta m)}\,\xi_j .$$

Since ζ is halved at each step, the Newton iteration is contracting in the domain

$$\hat{\mathcal{Q}} \equiv \{(\xi,\zeta)^T \in \mathcal{Q} \mid \zeta < \frac{1}{(1+\Delta m)}\,\xi^{1+\Delta m}\} ,$$

in that $g(\hat{\mathcal{Q}}) \subset \hat{\mathcal{Q}}$ and

$$\rho_{j+1} \leq \rho_j(\tfrac{1}{2}+\Delta m)/(1+\Delta m) \quad \text{for} \quad j \geq 1 .$$

Consequently Newton's method converges from all points in $\mathcal{Q}$ and the asymptotic rate of convergence is linear with the Q-factor

$$\frac{\xi_{j+1}}{\xi_j} = \left(\frac{\alpha_{j+1}}{\alpha_j}\right)\left(\frac{\zeta_{j+1}}{\zeta_j}\right)^{1/(1+\Delta m)} \to \frac{1}{2^{1/(1+\Delta m)}} .$$

4. General Results on Domains of Convergence and Contraction

Let $X_0 \subseteq \mathbb{R}^n - \{x^*\}$ be the set of all initial points from which Newton's method converges to a given solution x^* in a finite or infinite number of steps. Formally X_0 may be written as

$$X_0 = \bigcup_{j=1}^{\infty} g^{-j}(x^*) \cup \bigcap_{q=1}^{\infty} \bigcup_{\ell=1}^{\infty} \bigcap_{j=\ell}^{\infty} g^{-j}(\mathcal{B}_{1/q}) , \tag{34}$$

where $\mathcal{B}_{1/q}$ denotes again the unit ball with radius $1/q$ about x^*, and the inverse image $g^{-j}(\mathcal{D})$ contains all those points from which j Newton steps are well defined and lead to a point in the set $\mathcal{D} \subseteq \mathbb{R}^n$. This rather

unwieldy expression for X_0 will be used below to show that it is measurable and has therefore an upper density at x^* .

Since the iteration (0.1) is only defined as long as $\delta(x_j) = \det(\nabla f(x_j)) \neq 0$, the set X_0 is disjoint from $\delta^{-1}(0)$, so that all its subsets which will be called *domains of convergence* (to x^*) must primarily be domains of invertibility. Iterative methods of any kind are usually expected to have spherical domains of convergence in the neighbourhood of an isolated solution x^*, which implies that the full set of points from which the method converges to x^* is open, provided the iteration function is continuous. Whereas Newton's method has spherical domains of convergence in the nonsingular case, this requirement is in general not satisfied at a singularity x^*, as we know from Lemma 1.1 that x^* is only in special circumstances an isolated point of the singular set. Thus we obtain in general the following result.

LEMMA 1.7 *The Full Domain of Convergence* X_0

For $f \in C^1(\mathbb{R}^n,\mathbb{R}^n)$ and $x^* \in f^{-1}(0)$ let g and X_0 be defined by (20) and (34) respectively. Then

(i) X_0 is a Borel set but need not be open.

(ii) $g(X_0 - X_0^0) \subseteq X_0 - X_0^0$,

where X_0^0 denotes the interior of X_0 .

Proof.

(i) By continuity of g in its domain $\mathbb{R}^n - \delta^{-1}(0)$, the sets $g^{-j}(B_{1/q})$ are open so that the countable unions and intersections on the far right of (34) must define a Borel set. Thus we are left with the set of points from which Newton's method converges in finitely many steps.

By induction over $j \geq 0$ we can show that

$$g^{-j}(x^*) = C_j \cap \bigcap_{\ell=0}^{j-1} g^{-\ell}(\mathbb{R}^n - \delta^{-1}(0)) \tag{35}$$

where the sets C_j are closed in $\mathbb{R}^n$. For $j=0$ we set $C_0 = \{x^*\}$, as g^0 is the identity mapping on $\mathbb{R}^n$.

Now suppose (35) holds for some $j \geq 0$. Since the inverse image of intersections equals the intersection of the corresponding inverse images, we have

$$g^{-j-1}(x^*) = g^{-1}(C_j) \cap \bigcap_{\ell=1}^{j} g^{-\ell}(\mathbb{R}^n - \delta^{-1}(0)) \tag{36}$$

Let $x_i \to x$ be a sequence of points in $g^{-1}(C_j)$. If $x \notin \delta^{-1}(0)$, the point $g(x)$ is well defined and must be the limit of the $g(x_i) \in C_j$ by continuity of g. As C_j is closed, we have either $x \in g^{-1}(C_j)$ or $x \in \delta^{-1}(0)$, so that the set

$$C_{j+1} \equiv g^{-1}(C_j) \cup \delta^{-1}(0)$$

is closed. Since by definition $g^{-1}(C_j) \subseteq \mathbb{R}^n - \delta^{-1}(0)$, we obtain finally

$$C_{j+1} \cap (\mathbb{R}^n - \delta^{-1}(0)) = g^{-1}(C_j) \ ,$$

which substituted into (36) gives (35) for $j+1$. Clearly each $g^{-j}(x^*)$ is a Borel set so that the same is true for their countable union and consequently X_0 itself.

To show that X_0 need not be open we construct the following example. At some point $x_0 \neq x^*$ we can define $f \in C^2(\mathbb{R}^n, \mathbb{R}^n)$ such that the matrices $\nabla f(x_0)$ and $F_0 \equiv \nabla^2 f(x_0)(x_0 - x^*)$ are nonsingular and furthermore $f(x_0) = \nabla f(x_0)(x_0 - x^*)$ which is always possible since f and its derivative at x_0 are independent. Now we can easily calculate that

$$g(x_0) = x^* \quad \text{and} \quad \det(\nabla g(x_0)) = \det(F_0)/\det(\nabla f(x_0)) \neq 0 \ .$$

Assuming that x^* is not isolated in $\delta^{-1}(0)$, we find that x_0 is a clusterpoint of $g^{-1}(\delta^{-1}(0)-\{x^*\}) \subseteq \mathbb{R}^n - \mathcal{X}_0$ since g is because of its nonsingular Jacobian at x_0 locally 1-1 . Thus $\mathcal{X}_0$ cannot be open as it contains the point x_0 which does not belong to its interior $\mathcal{X}_0^0$. ////

The inclusion (ii) in Lemma 1.7 means that each point in some Newton sequence $\{x_i\}_{j\geq 0}$ with $x_0 \in \mathcal{X}_0 - \mathcal{X}_0^0$ belongs to the closure of the complement $\mathbb{R}^n - \mathcal{X}_0$, which consists of all those initial points from which Newton's method does not converge to x^* . Clearly such a theoretically converging iteration would be numerically highly unstable and could hardly be realised on a digital computer in finite precision arithmetic. Therefore we restrict our attention to the interior $\mathcal{X}_0^0$ which could theoretically be empty even if $\mathcal{X}_0$ has positive measure and upper density at x^* . Moreover we contend that (14) is quite a natural condition on the domain of convergence nominated in a local result.

If the line segment

$$\{\lambda x^* + (1-\lambda)x_0 \mid \lambda \in (0,1)\}$$

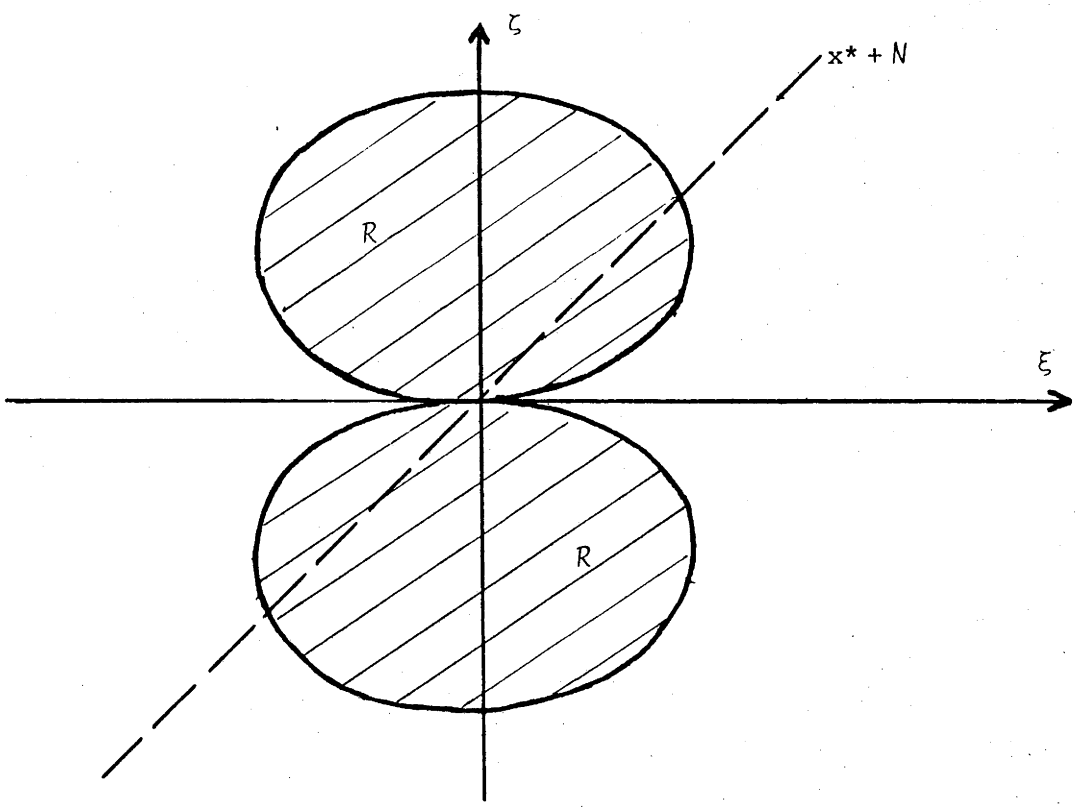
does not fully belong to the interior $\mathcal{X}_0^0$, then the statement that x_0 is an element of $\mathcal{X}_0$ should be considered a global rather than a local convergence result. By Theorem 1.3 (ii) $\mathcal{X}_0$ contains no nonempty starlike subdomains if and only if all directions in S are tangents of its complement $\mathbb{R}^n - \mathcal{X}_0^0$. As we will see later, this may be the case even though the interior $\mathcal{X}_0^0$ is nonempty and has x^* as an accumulation point. Nevertheless our main interest lies with starlike domains of convergence to a central point x^* for which a certain degree of numerical stability

can be expected. Particular examples of such starlike domains of convergence are balls, cones and their intersections which were used by Reddien [5,7] and Decker and Kelley [8] in their work on singular problems.

To give a simple example of a starlike domain of convergence, we consider f as defined by (31) with $\Delta m = 0$. The one dimensional nullspace N of $\nabla f(x^*)$ is spanned by $(1,1)^T$, so that f is not in normal form. The recurrence (33) gives directly for $j \geq 1$

$$\xi_j = \zeta_j = \zeta_0 2^{-j} \quad \text{and} \quad 2^j \rho_{j+1} = \rho_1 \leq \rho_0 ,$$

so that the Newton iteration converges from any initial point in $\{(\xi,\zeta)^T \mid \zeta \neq 0\}$, which is a starlike domain with the excluded directions $\pm(1,0)^T$. In the presence of higher order terms a starlike domain of convergence R has to be bounded and may typically take the form



The actual construction of such a set R will be given in Chapter 2 for the class of regular problems which includes the example considered here. Since obviously $L_1\{\pm(1,0)\} = 0$ the starlike domain R and consequently by Lemma 1.4 (ii), the domain of convergence X_0 have density 1 at x^*.

For $\Delta m > 0$ the family of examples (31) shows that the degree of a singularity x^* can be arbitrarily high negative and we may still have convergence from within a nonempty starlike domain so that X_0 has positive upper density at x^*. However this case is rather special in that it is known that all but the first $1+\Delta m$ derivatives of f vanish identically in $\mathbb{R}^n$, which enables us to calculate the Newton iteration exactly even for points that are arbitrarily far away from the singular solution x^*. In contrast a truly local convergence result has to be based on the values of finitely many derivatives of f at x^* and a nonzero Lipschitz constant of the highest one alone. Under these more realistic conditions nothing can be said about the values of f and its derivatives outside a certain ball so that convergence can only be guaranteed from within domains $\mathcal{D} \subseteq \mathbb{R}^n$ for which the set of intermediate points

$$\bigcup_{i=0}^{\infty} g^i(\mathcal{D})$$

is bounded. Any such $\mathcal{D}$ must be bounded itself and will be called a *domain of bounded convergence*.

In particular the first step from any point $x = x^* + \rho t \in \mathcal{D} \cap R'$ must be bounded by some radius $\beta > 0$. Since $\rho < \bar{r}$ and π_0 are bounded on S, there is a constant $\gamma_{\hat{i}} > 0$ such that by (21)

$$\|g(x^*+\rho t) - x^* - \rho^{\hat{i}} g_{\hat{i}}(t)\| \leq \gamma_{\hat{i}} \rho^{\hat{i}+1} / |\pi_0(t)|^{\hat{i}+\Delta m+1}, \tag{37}$$

which implies by definition of $g_{\hat{i}}$

$$
\begin{aligned}
&(\|u_{\hat{i}}(t)\| - \rho\gamma_{\hat{i}}/|\pi_0(t)|)\rho^{\hat{i}} \\
&\leq \|g(x^*+\rho t)-x^*\| \; |\pi_0(t)|^{\hat{i}+\Delta m} \qquad (38) \\
&\leq (\|u_{\hat{i}}(t)\| + \rho\gamma_{\hat{i}}/|\pi_0(t)|)\rho^{\hat{i}} \quad .
\end{aligned}
$$

Now it follows for $\hat{i} = 1$ that the condition

$$\|g(x^*+\rho t)-x^*\| \leq \beta \qquad (39)$$

is met at all points in the starlike domain

$$\left\{x^*+\rho t \;\middle|\; t\in S, 0<\rho<\min\left\{\bar{r}(t), \frac{\pi_0(t)}{\gamma_1}, \frac{\beta|\pi_0(t)|^{1+\Delta m}}{1+\|u_1(t)\|}\right\}\right\} ,$$

which has the same set of excluded directions as R' itself and thus density 1 at x^* for any $\beta > 0$.

For $\hat{i} = 0$ we find that the condition (39) is violated at all points in the starlike domain

$$T_\beta \equiv \{x^*+\rho t \mid t\in S, 0<\rho<(\|u_0(t)\|-\beta|\pi_0(t)|^{\Delta m})|\pi_0(t)|/\gamma_0\}$$

where $\Delta m \geq \hat{i}+1 = 1$.

Unless $g_0 = u_0/\pi_0^{\Delta m}$ is bounded all T_β with $\beta > 0$ are nonempty so that by Lemma 1.4 (ii) for any domain $\mathcal{D}$ of bounded convergence: $\tau^*(\mathcal{D}) \leq 1 - \tau^*(T_\beta) < 1$. To illustrate this situation we consider the function f defined by (31) with $\Delta m = 1$. Outside the ball B_1 we can modify f such that it equals the linear function $x - (3,3)^T$ for $\|x\| \geq 2$ and is smooth in between. Since the first step from any x_0 with $\|x_0\| \geq 2$ leads to the separate solution $(3,3)^T$, any domain of convergence to $x^* = 0$ must be contained in the set P_2 as defined by

$$P_\beta \equiv \left\{(\xi,\zeta)^T \;\middle|\; \|g(\xi,\zeta)\| = \frac{1}{2}\sqrt{\xi^2+\zeta^2+\zeta^2/\xi^2+2\zeta} < \beta\right\} ,$$

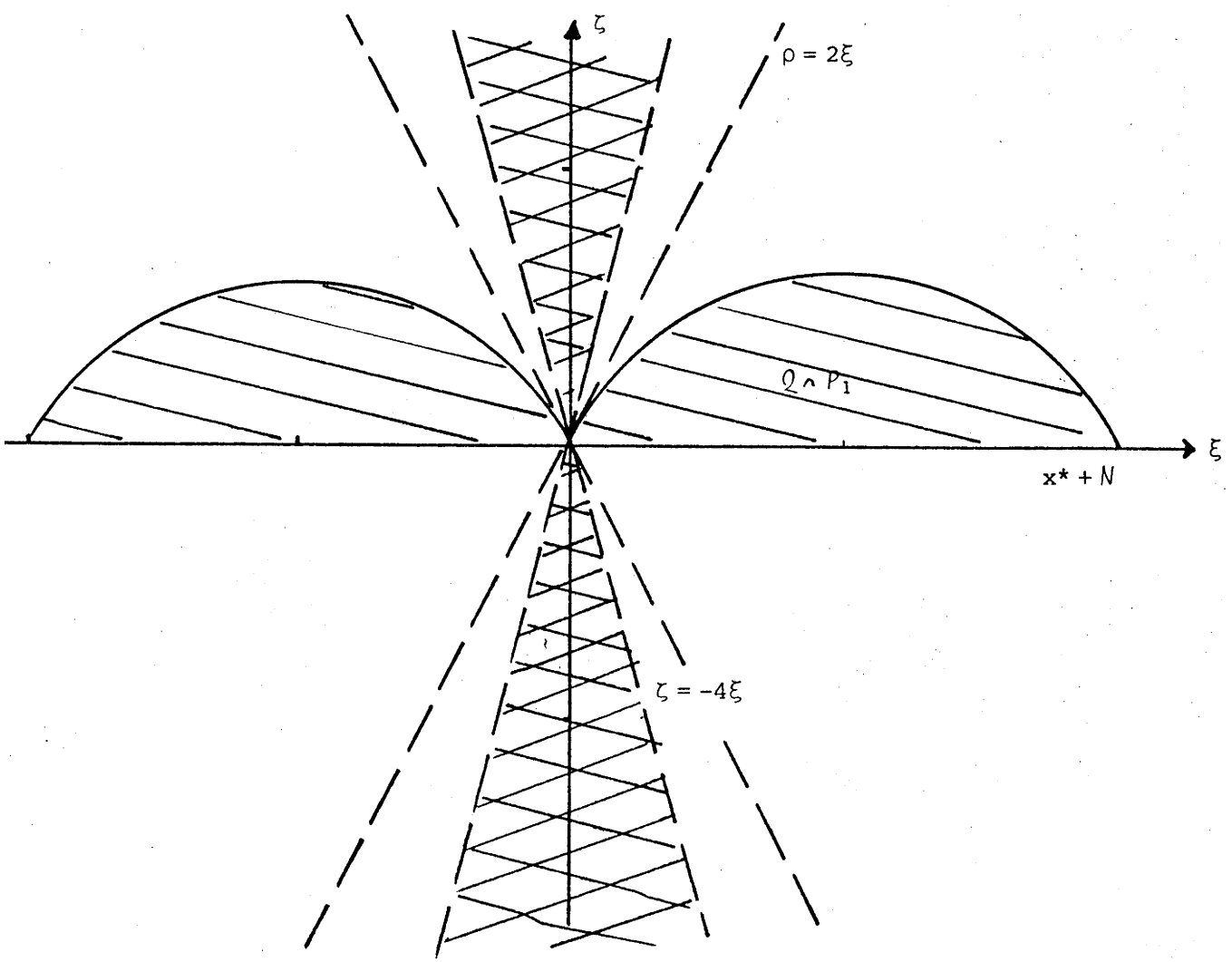
which is a starlike domain with the set of excluded directions

$$\{(\mu,\lambda)^T \in S \mid |\lambda| \geq 2\beta|\mu|\} .$$

We know from the analyses of the unmodified problem function that the Newton iteration is contracting after the first step from any x_0 within the first quadrant Q so that the set $Q \cap P_1$ must still be a starlike domain of convergence in the modified case. $Q \cap P_1$ has the set of included directions

$$\{(\mu,\lambda)^T \in S \mid 0 < \lambda < 2\mu\} ,$$

so that we have the following situation:



In the fourth quadrant the situation is exactly the same as in the first since the recurrence (33) is for $\Delta m = 1$ unaffected by a sign change of ξ. It does not really matter what exactly happens in the second and third quadrant as our main intention is to demonstrate that there are cases in which starlike domains of convergence do exist but some open sets of directions are necessarily excluded. This kind of situation seems to be typical for the case $\hat{i} = 0$ just studied.

If finally $\hat{i} < 0$ the condition (39) implies with some $\omega_\beta > 0$

$$\|u_{\hat{i}}(t)\| \, |\pi_0(t)| \le \beta\rho^{|\hat{i}|}|\pi_0(t)|^{\hat{i}+\Delta m+1} + \rho\gamma_{\hat{i}} \le \rho\omega_\beta$$

which means that any domain of bounded convergence $\mathcal{D}$ is disjoint from some starlike domain

$$T_\beta = \{x^*+\rho t \mid t\in S, 0<\rho<\min\{\|u_{\hat{i}}(t)\| \, |\pi_0(t)|/\omega_\beta, \bar{t}(t)\}\} \, . \tag{40}$$

Since $\|u_{\hat{i}}(t)\|^2 |\pi_0(t)|^2$ is an analytic function from S to $\mathbb{R}$, which does not vanish identically we derive from Lemma 1.4 (ii) and (i) that $\tau^*(\mathcal{D}) \le 1 - \tau^*(T_\beta) = 0$ which means that any domain of bounded convergence has outer density zero.

To illustrate this possibility we consider again the vector function (31), this time with $\Delta m = 2$, and modify it outside the unit ball as described in the case $\hat{i} = 0$. It can be easily checked that the Newton step from any point in the starlike domain

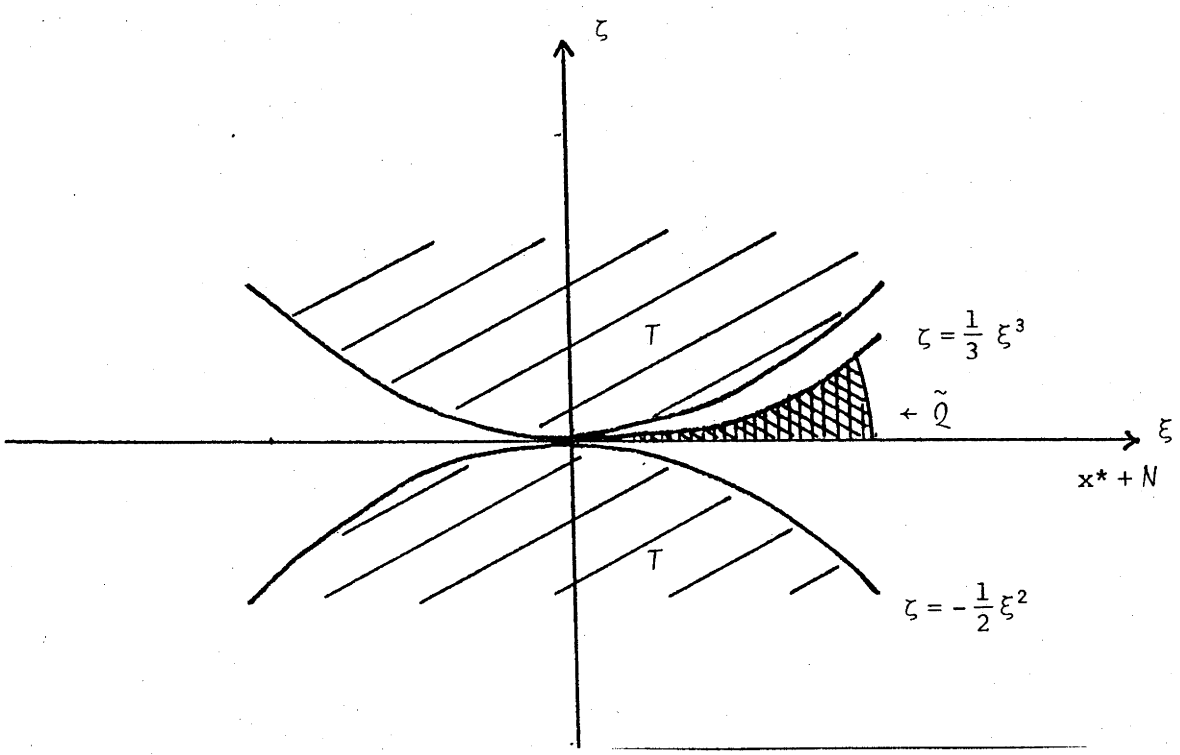
$$T = \{(\xi,\zeta)^T \mid 1 > \xi^2 < |\zeta|\}$$

is either not defined ($\xi=0$) or leads outside the ball with radius 2 about $x^* = 0$. Consequently any domain of convergence to x^* must be disjoint from T which has density 1, since there are only two excluded

directions, namely $\pm(0,1)$. However we know from the analysis of the unmodified problem that the Newton iteration is contracting in the domain $\hat{Q}$, so that

$$\tilde{Q} \equiv \{(\xi,\zeta)^T \mid 0 < \zeta < \tfrac{1}{3}\xi^3 \ , \ \zeta^2 + \xi^2 < 1\}$$

must still be a domain of convergence for the modified function. Thus we have the following picture.



The situation is essentially the same for all $\Delta m \geq 2$, so that whenever the degree is negative there may be a domain of bounded convergence whose closure contains the singularity x^*, but it must always be disjoint from some starlike domain T with density 1 . Since any infinite Newton sequence $x_j \to x^*$ must approach x^* through the narrowing channels excluded from T, it seems intuitively clear that this type of convergence would be numerically rather unstable. This notion can be made more precise in the following way.

THEOREM 1.8 *Instability of Convergence to Singularity with Degree* $\hat{i} < 1$.

Under the assumptions of Lemma 1.2 let $x_j \to x^*$ be any infinite Newton sequence with

$$\limsup_{j\to\infty} \|x_{j+1}-x^*\|/\|x_j-x^*\| = \omega < 1 \ . \tag{41}$$

Then we have if $\hat{i} < 0$

$$\lim_{j\to\infty} \left[\min \left\{ \frac{\|x_{j+1}-y\|}{\|x_{j+1}-x_j\|} \mid \delta(y) \neq 0 \ , \ \|g(y) - x^*\| \geq \beta \right\}\right] = 0 \ , \tag{42}$$

and if $\hat{i} = 0$

$$\lim_{j\to\infty} \left[\min \left\{ \frac{\|x_{j+1}-y\|}{\|x_{j+1}-x_j\|} \mid \delta(y) \neq 0 \ , \ \|g(y) - x^*\| \geq \tilde{\beta}\|y-x^*\| \right\}\right] = 0 \ ,$$

where β and $\tilde{\beta}$ are arbitrarily large positive constants.

Proof. Firstly we note that because of (41) for all but finitely many j

$$\|x_{j+1}-x_j\| \geq \|x_j-x^*\| - \|x_{j+1}-x^*\| \geq \rho_{j+1}(1-\omega)/(1+\omega) \ ,$$

so that the assertions must be true if they hold with the ratio $\|x_{j+1}-y\|/\|x_{j+1}-x_j\|$ replaced by $\|x_{j+1}-y\|/\rho_{j+1}$.

By construction of T_β , as defined in (40) for $\hat{i} < 0$, we have $\|g(y)-x^*\| \geq \beta$ for all $y \in T_\beta$. Since the sphere S is compact, the sequence $\{x_j = x^* + \rho_j t_j\}_{j\geq 0}$ has a nonempty set of tangents which must all be excluded from T_β as none of the iterates can belong to T_β . Now consider any subsequence $\{j_i\}_{i\geq 0}$ with $t_{j_i} \to t \in S$. Since the directions excluded from T_β are nowhere dense there is a sequence of included directions $s_q \to t$. Again we select an index sequence $\{q_i\}_{i\geq 0}$ where some of the q's may be repeated such that $y_i \equiv x^* + \rho_{j_i} s_{q_i} \in T_\beta$ for all but finitely many i . Then we obtain by the triangular inequality.

$$\lim_{i\to\infty} \|x_{j_i}-y_i\|/\rho_i \leq \lim_{j\to\infty} \|t_{j_i}-t\| + \lim_{j\to\infty} \|t-s_{q_i}\| = 0 \ . \tag{44}$$

As any subsequence of $\{x_j\}_{j\geq 0}$ has a subsequence with a unique tangent, the assumption that (42) does not hold must lead to a contradiction to (44). For $\hat{i} = 0$ we derive from (38) that $\|g(y)-x^*\| \geq \tilde{\beta}\|y-x^*\|$ for all y in the starlike domain

$$\{x^*+\rho t \mid t\in S\,,\ 0<\rho<\|u_0(t)\|\ |\pi_0(t)|/(\gamma_0+\tilde{\beta}|\pi_0(t)|^{1+\Delta m})\} \tag{45}$$

whose set of excluded directions has zero measure in S, so that (43) can be shown in the same way as (42). ////

Each Newton correction $g(x_j) - x_j$ is the solution of a linear system in the Jacobian $\nabla f(x_j)$ whose conditioning is deteriorating as we approach the singularity x^*. The best one can possibly hope for is that each step can be calculated with a relative error bounded by some constant $\varepsilon \in (0,1)$, so that the numerically evaluated new point x_{j+1} satisfies

$$\|x_{j+1} - g(x_j)\|/\|x_{j+1} - x_j\| \leq \varepsilon\ . \tag{46}$$

Unless the rate of convergence even of the exact Newton sequence is already less than linear, we know from (42) for $\hat{i} < 0$ and from (43) for $\hat{i} = 0$ that this accuracy is not enough to prevent intermittent steps away from the solution which are arbitrarily large either in absolute terms or at least relative to the previous distance from the solution. Thus even the existence of a domain of bounded convergence with positive upper outer density is not sufficient to ensure that there are Newton sequences which converge in a reasonably robust fashion.

Therefore we introduce the stronger concept of a *domain of contraction* to describe a subset $E \subseteq \mathbb{R}^n - \delta^{-1}(0)$ that satisfies

$$g(E) \subseteq E \quad \text{and} \quad \tilde{\beta}(E) \equiv \sup\{\|g(x)-x^*\|/\|x-x^*\| \mid x\in E\} < 1\ .$$

If $\hat{i} = 0$ any domain of contraction E is disjoint from the corresponding starlike domain (45) with $\tilde{\beta} = \tilde{\beta}(E)$ so that all of them have outer density zero. For $\hat{i} = 1$ we derive from (38) that any E must be disjoint from the starlike domain

$$\{x^*+\rho t \mid t\epsilon S, 0<\rho<(\|u_1(t)\|-\beta|\pi_0(t)|^{1+\Delta m})|\pi_0(t)|/\gamma_1\} ,$$

where again $\tilde{\beta} = \tilde{\beta}(E)$. Unless

$$\sup\|g_1(t)\| = \sup\|u_1(t)\|/|\pi_0(t)|^{1+\Delta m} < 1$$

all these starlike domains are nonempty so that no domain of contraction can have upper outer density 1 . Thus we can compile the following bounds on the upper outer density of a domain of bounded convergence $\mathcal{D}$ and a domain of contraction E at a singularity of degree $\hat{i}$.

	$\hat{i} < 0$	$\hat{i} = 0$	$\hat{i} = 1$	$\hat{i} = 2$
$\tau^*(\mathcal{D})$	0	≤ 1	≤ 1	$= 1$
$\tau^*(E)$	0	0	≤ 1	$= 1$

where $\hat{i} = 2$ represents the nonsingular case. We can usually expect that $\tau^*(\mathcal{D}) < 1$ for $\hat{i} = 0$ and $\tau^*(E) < 1$ for $\hat{i} = 1$, and it will be shown in Chapter 2 that $\tau^*(\mathcal{D}) = 1$ for the wide class of *regular* first degree singularities.

5. General Results on Rates of Convergence

It might be thought that the numerical difficulties of Newton's method if applied to the vector function (31) with $\Delta m = 1-\hat{i} \geq 2$, are somehow related to the fact that x^* is poorly isolated as a solution in that

$$\|f(x^*+\lambda e_1)\| = O(\lambda^{1+\Delta m}) .$$

Clearly $x^* \in f^{-1}(0)$ is a nonsingular solution iff for all $z \in \mathbb{R}^n$

$$\|f(x^*+z)\|^{-1} = O(\|z\|^{-1})$$

so that linear growth of the residual $\|f\|$ along any smooth path emanating from x^* ensures quadratic convergence of Newton's method, provided the initial point is sufficiently close to x^* and the Jacobian is Lipschitz continuous. In the singular case however there is in general no direct correspondence between the degree of isolation of x^*, i.e. the growth order of the residual in its neighbourhood, and the performance of Newton's method.

To see this we append $f = (f_1, f_2)^T$ as defined in (31) by a third component function

$$f_3(\xi,\zeta,\eta) \equiv \frac{1}{2}\xi^2 + \frac{1}{2}\eta^2 ,$$

where η is a new variable in which f_1 and f_2 are considered to be constant. A simple calculation gives

$$\sqrt{3}\,\|(f_1,f_2,f_3)\| \geq |f_1| + |f_2| + |f_3| \geq \frac{1}{2}\|(\xi,\zeta,\eta)\|^2 ,$$

so that $x^* = 0$ is as strongly isolated as possible for a singular solution with Lipschitz continuous Jacobian. However since the Jacobian of the extended system is (permuted) triangular, the recurrences (33) are unaffected and the speed of convergence of Newton's method is therefore at best unchanged.

Nevertheless there is a relation between the behaviour of Newton's method and a certain sufficiency condition for isolation as given in the following theorem.

THEOREM 1.9 *Isolation Condition and Rates of Convergence*

Under the assumption of Lemma 1.2 with $\bar{k} \geq k+\Delta m$ let $x_j \to x^*$ be any converging Newton sequence. Then

(i) If

$$P\nabla^{k+1}f(x^*)t^{k+1} \neq 0 \quad \text{for all} \quad t \in N \cap S \tag{47}$$

then x^* is an isolated solution of f, i.e. an isolated point in $f^{-1}(0)$.

(ii) If $\{x_j\}_{j\geq 0}$ converges Q-quadratically all its regular tangents belong to

$$T \equiv \{t \in S \mid P\nabla^2 f(x^*)tt=0\} .$$

(iii) If $\{x_j\}_{j\geq 0}$ converges Q-subquadratically in that

$$\liminf_{j\to\infty} \rho_j^2/\rho_{j+1} = 0$$

then it has at least one tangent in $N \cap S$.

(iv) If x^* is a first order singularity of degree 1 and the isolation condition (47) is satisfied then $\{x_j\}_{j\geq 0}$ provided it has no tangent in $N \cap S \cap \pi_0^{-1}(0)$ converges either Q-quadratically or

$$\liminf_{j\to\infty} \rho_{j+1}/\rho_j < 0$$

which means at best Q-linear convergence.

Proof.

(i) Suppose there is a sequence of solution points $\{x_j = x^*+\rho_j t_j\} \subseteq f^{-1}(0)$ with

$$\rho_j \to 0 \quad \text{and} \quad t_j \to t \in S .$$

Using the Taylor expansion of f at x^* we derive

$$0 = [f(x_j)-f(x^*)]/\rho_j = \nabla f(x^*)t_j + O(\rho_j)$$

and similarly

$$0 = P[f(x_j)-f(x^*)]/\rho_j^{k+1} = P\nabla^{k+1}f(x^*)t_j^{k+1}/(k+1)! + O(\rho_j) ,$$

so that in the limit as ρ_j tends to zero

$$\nabla f(x^*)t = 0 = P\nabla^{k+1}f(x^*)t^{k+1} .$$

(ii) Since $\{x_j=x^*+\rho_j t_j\}_{j\geq 0}$ converges Q-quadratically we have

$$\limsup_{j\to\infty} \rho_{j+1}/\rho_j^2 = \limsup_{j\to\infty} \|g(x_j)-x^*\|/\rho_j^2 < \infty .$$

For any subsequence $\{t_{j_i}\}_{i\geq 0} \subseteq \{t_j\}_{j\geq 0}$ that converges to a regular direction $t \in S'$ we find

$$\lim_{i\to\infty} \pi_0(t_{j_i}) = \pi_0(t) \neq 0 \quad \text{and} \quad x_{j_i} \in R'$$

for all but finitely many i, so that by (21)

$$\limsup_{i\to\infty} \left\| \sum_{\ell=1-\Delta m}^{1} \rho_{j_i}^{\ell-2} g_\ell(t_{j_i}) \right\| < \infty .$$

Since $\rho_{j_i} \to 0$ this can only be the case if

$$g_\ell(t) = \lim_{i\to\infty} g_\ell(t_{j_i}) = 0 \quad \text{for} \quad \ell = 1-\Delta m,\ldots 1 .$$

Then the first Δm "rows" of the linear system (22) are trivially satisfied, and we obtain multiplying the $(2+\Delta m)$-th "row" from the left by P

$$PA_1 g_2(t) = 0 = \frac{1}{2} PA_2(t)t = \frac{1}{2} P\nabla^2 f(x^*)tt$$

which proves assertion (ii).

(iii) Let $\{x_{j_i}\}_{i\geq 0}$ be a subsequence for which

$$\lim_{i\to\infty} \rho_{j_i}^2/\rho_{j_i+1} = 0 \quad \text{and} \quad \lim_{i\to\infty} t_{j_i} = t \in S .$$

By assumption f is twice differentiable so that

$$\nabla f(x_j)(x_{j+1}-x^*) = -(f(x_j)-\nabla f(x_j)(x_j-x^*)) = O(\rho_j^2) .$$

Dividing by ρ_{j_i+1} we find for the subsequence $\{j_i\}_{i\geq 0}$

$$\lim_{i\to\infty} \nabla f(x_{j_i}) t_{j_i+1} = \lim_{i\to\infty} O(\rho_{j_i}^2/\rho_{j_i+1}) = 0 ,$$

so that all clusterpoints of the subsequence $\{t_{j_i+1}\}_{i\geq 0} \subseteq S$ must belong to N. As S is compact $\{x_j\}$ must have a tangent in N.

(iv) Since $\{x_j\}_{j\geq 0}$ has no tangent in $N \cap S \cap \pi_0^{-1}(0)$ there is a lower bound $\phi_0 > 0$ such that for all but finitely many j

$$\phi_0 < \phi(t_j) \equiv \min\{\cos^{-1}(s^T t_j) \mid s\in S\cap N\cap\pi_0^{-1}(0)\} .$$

The function $\phi(t)$ gives the minimal angle between some direction t and the set of irregular directions in N. Similarly we define

$$\theta(t) \equiv \min\{\cos^{-1}(s^T t) \mid s\in S\cap N\} \leq \phi(t) . \tag{48}$$

With θ_0 such that $\theta_0 \leq 45^\circ$ and

$$\theta_0 \leq \frac{1}{2}\min\{\cos^{-1}(t^T s) \mid s\in N\cap S , t\in T \text{ or } t\in\pi_0^{-1}(0) \text{ and } \phi(t)\geq\frac{1}{2}\phi_0\} ,$$

the set of directions

$$U \equiv \{t\in S \mid \theta(t)<\theta_0 , \phi(t)>\phi_0\}$$

and its closure are disjoint from T and $\pi_0^{-1}(0) \cap S$, so that

$$\varepsilon_1 \equiv \min\{|\pi_0(t)| \mid t\in U\} > 0 ,$$

$$\rho_0 \equiv \inf\{\bar{r}(t) \mid t\in U\} > 0 ,$$

and similarly

$$\varepsilon_2 \equiv \min\{\|g_1(t)\| \mid t\in U\} > 0 .$$

The positiveness of ε_2 follows from the fact that, because $\hat{i} = 1$, the first nontrivial coefficient vector $g_1(t)$ can only vanish at a regular direction if $t \in T$, as shown in the proof of (i).

Whenever $t_j \in U$ and $\rho_j < \bar{\rho}_0$ we have by (37)

$$\|\rho_{j+1} t_{j+1} - \rho_j g_1(t_j)\| \leq \gamma_1 \rho_j^2 / \varepsilon_1^{2+\Delta m} , \tag{49}$$

which implies because $g_1(t_j) \in N$

$$\sin \theta(t_{j+1}) \leq \sin\left[\cos^{-1}(t_{j+1}^T g_1(t_j)/\|g_1(t_j)\|)\right]$$

$$= \min_{\lambda \in \mathbb{R}} \left\| \lambda t_{j+1} - g_1(t_j)/\|g_1(t_j)\| \right\| \leq \gamma_1 \rho_j / (\varepsilon_2 \varepsilon_1^{2+\Delta m}) .$$

Thus we find that for sufficiently large j

$$t_j \in U \quad \text{and} \quad \rho_j < \bar{\rho} \equiv \min\left\{\bar{\rho}_0, \varepsilon_1 \varepsilon_2^{2+\Delta m} \gamma_1^{-1} \sin \theta_0\right\}$$

ensures $t_{j+1} \in U$. Unless $\{x_j\}_{j \geq 0}$ converges quadratically it has by (iii) a tangent in N which must be regular by assumption. Since furthermore $\lim_{j \to \infty} \rho_j = 0 < \bar{\rho}$ all but finitely many iterates must belong to the set

$$\{x^* + \rho t \mid t \in U , \rho < \bar{\rho}\} .$$

Finally we derive from (49) that

$$\liminf_{j \to \infty} \rho_{j+1}/\rho_j \geq \liminf_{j \to \infty} \left[\|g_1(t_j)\| - \gamma_1 \rho_j / \varepsilon_1^{2+\Delta m}\right] \geq \varepsilon_2$$

which completes the proof. ////

Even though it could not be shown conclusively, it seems likely that Newton sequences always have integer order unless f involves fractional powers. At singularities quadratic convergence from certain initial points is theoretically possible but numerically unstable, as

rounding errors will prevent the minimal angles between the t_j and directions in T from becoming arbitrarily small. Without analyzing the nature of such domains of quadratic convergence in any detail, we note that for all Q-superlinearly converging Newton sequences $\{x_j=x^*+\rho_j t_j\}_{j\geq 0}$ with $\lim_{j\to\infty} \rho_{j+1}/\rho_j \to 0$

$$\limsup_{j\to\infty} \frac{\|g(x_j)-x^*\|}{\|g(x_j)-x_j\|} \leq \limsup_{j\to\infty} \frac{\rho_{j+1}/\rho_j}{1-\rho_{j+1}/\rho_j} = 0 \ . \tag{50}$$

Even under the optimistic assumption that each Newton step can be calculated with a uniformly bounded relative error ε, the numerically evaluated new iterate could be any point in the ball

$$\{x\in\mathbb{R}^n \mid \|x-x^*\| < \varepsilon\|g(x_j)-x_j\| - \|g(x_j)-x^*\|\} \ ,$$

which is by (50) nonempty for sufficiently large j. We have noted that in general any neighbourhood of the singularity x^* contains points $x = x^*+\rho t$ where the Newton step is either undefined ($\delta(x)=0$) or leads further away from the solution ($\|g_1(t)\|>1$). Thus superlinear convergence of the unmodified Newton method on a singular problem seems unlikely to occur in practice and is not even desirable.

CHAPTER 2

STARLIKE DOMAINS OF CONVERGENCE AT REGULAR SINGULARITIES

1. Balanced and Regular Singularities

An apparent shortcoming of the theory developed so far is that the crucial degree $\hat{i} \in [1-\Delta m, 1]$ is only defined implicitly through Theorem 1.6 and we have no rule to compute it at any given singularity. Clearly $\hat{i} = 1$ if $\Delta m = 0$ which is by Lemma 1.2 equivalent to the condition that with $k = 1$ the linear operator

$$\bar{B}(y) \equiv \frac{1}{k!} P\nabla^{k+1} f(x^*) y^k \big|_N : N \to P(\mathbb{R}^n) \tag{1}$$

is nonsingular for some $y \in \mathbb{R}^n$. As will be seen later x^* is still a first degree singularity if it is of order $k>1$ as defined in Theorem 1.6 and $\det(\bar{B}(y))$ does not vanish identically. Such singularities will be called *balanced* because the lack of determinacy caused by the singularity of the Jacobian $\nabla f(x^*)$ is essentially compensated at the level of the $(1+k)$-th derivative. Assuming again that f is in normal form at x^* we have the Taylor expansions

$$\begin{aligned} B(x) &= \bar{B}(x-x^*) + O(\rho^{k+1}) = \rho^k \bar{B}(t) + O(\rho^{k+1}) , \\ C(x) &= \bar{C}(x-x^*) + O(\rho^{k+1}) = \rho^k \bar{C}(t) + O(\rho^{k+1}) , \\ D(x) &= O(\rho) \quad \text{and} \quad E(x) = I + O(\rho) . \end{aligned} \tag{2}$$

so that the reduced Jacobian (1.6) satisfies

$$G(x) = \rho^k \bar{B}(t) + O(\rho^{k+1}) . \tag{3}$$

Since the leading terms on the RHS are of order k in ρ we obtain instead of (1.10)

$$\det(G(x)) = \det\left(\sum_{i=k}^{\bar{k}} G_i(x-x^*)\right) + O\left(\rho^{k(m-1)+\bar{k}+1}\right)$$

$$= \rho^{km}\det(\bar{B}(t)) + O(\rho^{mk+1}) \ ,$$

which implies because of (1.7) by comparison with (1.9)

$$p = km \ , \quad \Delta p \in [\bar{k}-k+1, \bar{k}] \tag{4}$$

and

$$\pi_0(y) = \alpha \det(\bar{B}(y)) \ , \tag{5}$$

where $\alpha \neq 0$ allows again for linear transformations into normal form.

At balanced singularities the linear system (1.22) can be solved explicitly provided f is sufficiently often differentiable. Multiplying for $i = 1-\Delta m, \ldots, \ell-k$ the $(i+k)$-th "row" by P and adding the result to the i-th "row" we obtain the block triangular Toeplitz system

$$\begin{pmatrix} \hat{A}_1 & 0 & \cdots & \cdots & 0 \\ \hat{A}_2 & \hat{A}_1 & \ddots & & \vdots \\ \hat{A}_3 & \hat{A}_2 & \hat{A}_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \hat{A}_{\Delta p-k} & \cdots & \hat{A}_3 & \hat{A}_2 & \hat{A}_1 \end{pmatrix} \begin{pmatrix} g_{1-\Delta m} \\ g_{2-\Delta m} \\ \vdots \\ g_0 \\ g_1 \\ g_2 \\ \vdots \\ g_{\ell-k} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \frac{k}{k+1} PA_{k+1}t \\ \frac{1}{2}A_2 t + \frac{k+1}{k+2} PA_{k+1}t \\ \vdots \\ \frac{\ell-k-1}{\ell-k} A_{\ell-k}t + \frac{\ell-1}{\ell} PA_\ell t \end{pmatrix} \tag{6}$$

where $\ell = \Delta p - \Delta m \geq \bar{k} - (m+1)(k-1)$, $A_i = A_i(t) = \nabla^i f(x^*) t^{i-1}/(i-1)!$ as before and

$$\hat{A}_i = \hat{A}_i(t) = (A_i(t) + PA_{i+k}(t)) \quad \text{for} \quad i = 1-\Delta m, \ldots, \ell-k \ .$$

The square matrix $\hat{A}_1$ in the diagonal has by (5) the determinant

$$\det(\hat{A}_1(t)) = \det[\nabla f(x^*) + P\nabla^{k+1} f(x^*) t^k/k!] = \pi_0(t)/\alpha$$

so that for all regular $t \in S'$ the system (6) can be solved by back substitution which yields in particular

$$g_i \equiv 0 \quad \text{for} \quad i<1 \quad \text{and} \quad g_1(t) = \frac{k}{k+1} \hat{A}_1^{-1}(t) P A_{k+1}(t) t \ . \tag{7}$$

Thus all balanced singularities are of first degree. Unfortunately we have to assume within the framework of Theorem 1.6 that f is $\bar{k}+1 \geq 2 + (m+2)(k-1)$ times differentiable, to obtain a first order approximation to g by g_1, whereas the form of the latter suggests that $\bar{k} \geq k$ should be sufficient. This is indeed the case as shown below. Since $\|\cdot\|$ denotes the spectral norm of a matrix, the smallest singular value of $\bar{B}(t)$ is given by the continuous function

$$\nu(t) \equiv \begin{cases} 0 & \text{if } \bar{B}(t) \text{ is singular} \\ \|\bar{B}^{-1}(t)\|^{-1} & \text{otherwise} \end{cases} \tag{8}$$

on the compact domain S. Combining results from Lemma 1.5 and Theorem 1.6 we obtain the following lemma for the balanced case.

LEMMA 2.1 *Properties of* R' *and* g_1 *at Balanced Singularities*

Let $f \in C^{k+1,1}(\mathbb{R}^n, \mathbb{R}^n)$ be in normal form at a k-th order balanced singularity $x^* \in f^{-1}(0)$. Then

(i) At all points $x = x^* + \rho t \in R'$ the inverse Jacobian takes the form

$$\nabla f^{-1} = \begin{pmatrix} G^{-1} & , & -G^{-1}C^T E^{-1} \\ -E^{-1}DG^{-1} & , & E^{-1} + E^{-1}DG^{-1}C^T E^{-1} \end{pmatrix} , \tag{9}$$

with

$$G^{-1}(x) = \rho^{-k}\bar{B}^{-1}(t) + \nu^{-1}(t)O(\rho^{1-k}) = \nu^{-2}(t)O(\rho^{-k}) \quad . \tag{10}$$

(ii) The smallest singular value $\sigma(x)$ of $\nabla f(x)$ satisfies

$$\sigma(x^*+\rho t) = \begin{cases} \rho^k \nu(t)(1+O(\rho)) & \text{if } t \in S' \\ o(\rho^k) & \text{otherwise} \end{cases}$$

(iii) There is a constant d such that for all $x = x^* + \rho t \in R'$

$$\|g(x) - x^* - g_1(x-x^*)\| \leq d(\rho/\nu(t))^2 \quad , \tag{11}$$

where the homogeneous vector function

$$g_1 : (\mathbb{R}^n - \pi_0^{-1}(0)) \to N \subseteq \mathbb{R}^n$$

is given by

$$g_1(x-x^*) = \rho g_1(t) = \frac{k}{k+1} \begin{pmatrix} I & , & \bar{B}^{-1}(t)\bar{C}^T(t) \\ 0 & , & 0 \end{pmatrix} (x-x^*) \quad , \tag{12}$$

with $\bar{C}(t)$ as defined implicitly in (2).

Proof.

(i) By definition of $\bar{r}$ in (1.13) the matrices $\nabla f, E$ and therefore by (1.7) also G are nonsingular at all points in R', so that the inverse Jacobian must take the given form. Equation (10) follows from (3) by the Perturbation Lemma 2.3.2 in [10].

(ii) Using (2) we derive from (i) for $t \in S'$

$$\nabla f^{-1}(x) = \begin{pmatrix} G^{-1}(x) & , & O(\rho^0) \\ O(\rho^{1-k}) & , & O(\rho^0) \end{pmatrix} = \rho^{-k} \left[\begin{pmatrix} \bar{B}^{-1}(t) & , & 0 \\ 0 & , & 0 \end{pmatrix} + O(\rho) \right] \quad .$$

Hence we find for the spectral norm $\|\cdot\|$

$$\sigma^{-1}(x^*+\rho t) = \|\nabla f^{-1}(x^*+\rho t)\| = \rho^{-k}\nu^{-1}(t)(1+O(\rho))$$

which proves assertion (ii) for regular t. For $t \in S \cap \pi_0^{-1}(0)$ the assertion follows from Lemma 1.5 (iv) with $p = km$.

(iii) In order to obtain an approximate expression for $f(x)$ we use the obvious identity

$$f(x) = \left(\int_0^\rho \nabla f(x^*+\mu t)\,d\mu \right) t \ . \tag{13}$$

It follows from (2) and (3) that

$$\int_0^\rho B(x^*+\mu t)\,d\mu = \frac{\rho^{k+1}}{(k+1)} \bar{B}(t) + O(\rho^{k+2}) = \frac{\rho}{(k+1)} G(x) + O(\rho^{k+2})$$

and similarly $\int_0^\rho C(x^*+\mu t)\,d\mu = \frac{\rho}{k+1} C(x) + O(\rho^{k+2})$,

$$\int_0^\rho D(x^*+\mu t)\,d\mu = \frac{\rho}{2} D(x) + O(\rho^3) \quad \text{and} \quad \int_0^\rho E(x^*+\mu t)\,d\mu = \rho E(x) + O(\rho^2) \ .$$

Substituting these expansions into (13) we find

$$f(x) = \begin{pmatrix} \frac{1}{k+1} G(x) + O(\rho^{k+1}) & , & \frac{1}{k+1} C^T(x) + O(\rho^{k+1}) \\ \\ \frac{1}{2} D(x) + O(\rho^2) & , & E(x) + O(\rho) \end{pmatrix} (x-x^*) \ .$$

Multiplying from the left by (9) we obtain

$$\nabla f^{-1}(x) f(x) = F(x)(x-x^*)$$

where $F(x) \equiv$

$$\begin{pmatrix} \frac{1}{k+1} I + \|G^{-1}(x)\| O(\rho^{k+1}) & , & -\frac{1}{k+1} G^{-1}(x) C^T(x) + \|G^{-1}(x)\| O(\rho^{k+1}) \\ \\ \left(\frac{1}{2} - \frac{1}{k+1}\right) E^{-1}(x) D(x) + O(\rho^2) + \|G^{-1}(x)\| O(\rho^{k+2}) & , & I + O(\rho) + \|G^{-1}(x)\| O(\rho^{k+1}) \end{pmatrix} . \tag{14}$$

In general the bottom left submatrix is $O(\rho)$, whereas for $k=1$ it is $O(\rho^2)$, a fact that will be used in Chapter 3 but is as yet unimportant.

Because of (2), (10) and the boundedness of ν we have

$$G^{-1}(x)C^T(x) = \bar{B}^{-1}(t)\bar{C}^T(t) + \nu^{-1}(t)O(\rho)$$

which completes the proof as we may use again (10) to bound $\|G^{-1}(x)\|$. ////

That balanced singularities are not the only ones of first degree can be seen from the following example. The function

$$f(\xi,\zeta) = \left(\frac{1}{3}\xi^3, \frac{1}{2}\zeta^2\right)^T$$

has a second order singularity at $x^* = 0$ with $m=2=n$. The Newtonian iteration function is easily calculated as

$$g(\xi,\zeta) = g_1(\xi,\zeta) = \left(\frac{2}{3}\xi, \frac{1}{2}\zeta\right)^T,$$

so that $\hat{i} = 1$, whereas the matrix

$$\bar{B}(\mu,\lambda) = \begin{pmatrix} 0 & , & 0 \\ & & \\ 0 & , & \lambda \end{pmatrix} \qquad \text{for} \quad (\mu,\lambda)^T \in S,$$

which represents the leading, linear terms of the Jacobian is always singular. Consequently the singularity is not balanced. Nevertheless the Newton iteration converges linearly from all points that do not belong to either axis. The reason for this is that, even though the lack of definition caused by the vanishing Jacobian at x^* is compensated at the level of the second derivative with respect to ζ and the third with respect to ξ, the resulting different speeds of convergence are mutually independent. By adding higher order terms to either component of f, e.g. ζ^3 to the first, this independence can be destroyed which usually makes $\hat{i}$ negative and leads to less regular behaviour of the Newton iteration.

Whereas in the example given above

$$g_1(t) = \alpha t \quad \text{for some} \quad \alpha \in \mathbb{R} \quad \text{iff} \quad \pi_0(t) = 0$$

we have by Lemma 2.1 (iii) at a balanced singularity

$$g_1(t) = t\, k/(k+1) \quad \text{for all} \quad t \in S' \cap N \tag{15}$$

This equation suggests that g may be contracting in some domain including all regular directions in N. An arbitrary set is said to *include* a direction $t \in S$ at a certain point, if it contains a starlike domain that include t. Clearly the statement (15) is void if all directions in N are irregular. Excluding such degeneracy we introduce the following concept of a regular singularity.

A singularity x^* of order k is said to be *regular* if the linear operator $\bar{B}(t)$ as defined by (1) is nonsingular for some $t \in N \cap S$ so that x^* is balanced and

$$\pi_0(N) \neq \{0\} . \tag{16}$$

In those cases where the nullspace of the Jacobian $\nabla f(x^*)$ is spanned by a single vector $t \in S$ we find

$$\bar{B}(t) \text{ is nonsingular} \iff \bar{B}(t)t = P\nabla^{k+1} f(x^*) t^{k+1} \neq 0 .$$

Thus the regularity condition on the LHS is equivalent to the isolation condition introduced in Theorem 1.9 (i). For $m > 1$ neither condition implies the other as we can see from the following examples.

Firstly consider the function

$$f(\xi,\zeta) = (\tfrac{1}{2}(\xi^2+\zeta^2) , \tfrac{1}{4}\zeta^4)^T$$

with $m = 2 = m$, $k = 1$ and $f^{-1}(0) = \{x^* = 0\}$. The linear terms in the Jacobian form the matrix

$$\bar{B}(\mu,\lambda) = \begin{pmatrix} \mu & , & \lambda \\ & & \\ 0 & , & 0 \end{pmatrix} \quad \text{for} \quad (\mu,\lambda)^T \in S$$

which is obviously always singular, whereas

$$\bar{B}(\mu,\lambda)\,(\mu,\lambda)^T = (\mu^2+\lambda^2,0)^T = (1,0)^T$$

can never vanish. Hence the isolation condition but not the regularity condition is satisfied.

Secondly consider the function

$$f(\xi,\zeta) = \left(\frac{1}{2}\xi^2,\xi\zeta\right)^T$$

with $m = 2 = n$, and $k = 1$. Since all points on the ζ-axis are solutions of f the isolation condition cannot be satisfied at $x^* = 0$ but the matrix

$$\bar{B}(\mu,\lambda) = \begin{pmatrix} \mu & , & 0 \\ & & \\ \lambda & , & \mu \end{pmatrix}$$

is nonsingular for all $(\mu,\lambda)^T \in S$ except $\pm(0,1)^T$ so that the problem is regular. By adding higher order terms one can easily make x^* an isolated solution without changing $\bar{B}$.

In what follows we will consider the case of a regular first order singularity as the most important and likely possibility. Even though isolation of x^* as a solution is implied by regularity if $m = 1$ we will otherwise not make the assumption that the singularity is isolated in $f^{-1}(0)$.

As a consequence of Lemma 2.1 (ii) the condition number of $\nabla f(x^*+\rho t)$

is of order ρ^{-k} or larger unless the singularity is *pure* in that the Jacobian vanishes completely at x^* (i.e. m=n). In the well researched scalar case $n = 1$ every singularity must obviously be pure, and correspondingly the conditioning of the 1×1 matrix formed by the derivative of f poses no numerical problem. For $n \geq 1$ a pure singularity of order k is balanced iff it is regular iff

$$\bar{B}(t) = \frac{1}{k!} \nabla^{k+1} f(x^*) t^k$$

is nonsingular for some and consequently almost all $t \in S$. As will be shown in Lemma 2.2 most Newton sequences $\{x_j = x^* + \rho_j t_j\}_{j \geq 0}$ that converge to a regular singularity x^* do so along a regular direction $t = \lim t_j$ so that by (3) and (10) in the pure case

$$\lim_{j \to \infty} \|\nabla f(x_j)\| \, \|\nabla f^{-1}(x_j)\| = \|\bar{B}(t)\| \, \|\bar{B}^{-1}(t)\| \, . \tag{17}$$

Here we have used the fact that $\nabla f = G$ if $m = n$. Hence the conditioning of the Jacobian as such is unlikely to cause numerical difficulties at a pure, regular singularity. However due to cancellation the relative error in the computed values of $f(x_j) = O(\rho_j^{k+1})$ and $\nabla f(x_j) = O(\rho_j^k)$ may grow rapidly as ρ_j tends to zero.

For the general, nonpure case the regularity assumption (16) implies that the pure, reduced system discussed in Section 1.1 is regular too as its Jacobian G is according to (3) dominated by $\bar{B}$. The converse is not true as we can see from the family of examples (1.31) with $\Delta m > 0$. Using the second component of f to eliminate $\zeta = (1+\Delta m)^{-1} \xi^{1+\Delta m}$ we obtain the reduced system of one equation

$$f_1(\xi, (1+\Delta m)^{-1} \xi^{1+\Delta m}) = \frac{1}{2} (1+\Delta m)^{-2} \xi^{2(1+\Delta m)} = 0$$

with the derivative

$$G(\xi,(1+\Delta m)^{-1}\xi^{1+\Delta m}) = (1+\Delta m)^{-1}\xi^{1+2\Delta m} .$$

Since the isolation condition (1.47) is clearly satisfied the reduced system must be regular whereas the full system has degree $1-\Delta m < 1$ and is therefore not even balanced. This situation occurs because the leading terms in both components of f are powers of the nonsingular variable ζ so that $\bar{B}$ does not dominate G.

The regularity assumption amounts to the condition that for the singular equations, the leading terms in the singular variables are at most of the same order as the leading terms in the nonsingular ones and form a system of m homogeneous equations in m variables whose Jacobian is not everywhere singular.

2. Domains of Contraction at Regular Singularities

Firstly we derive from Lemma 2.1 (iii) some useful relations between the iterates of a Newton sequence $\{x_j=x^*+\rho_j t_j\}_{j\geq 0}$ with $x_{j+1} = g(x_j)$ and $\nu_j \equiv \nu(t_j)$.

Provided the ratio (ρ_0/ν_0) is small enough the first step from $x_0 = x^* + \rho_0 t_0 \in R'$ is essentially a projection like mapping to the vector $\rho_0 g_1(t_0)$ in the nullspace N. Whenever $g_1(t_0) \neq 0$ we derive from (11) that the angle $\psi_1(t_0)$ between $g_1(t_0)$ and the exact t_1 is bounded by

$$\sin\psi_1(t_0) = \min_{\lambda\in\mathbb{R}} \left\| \lambda t_1 - \frac{g_1(t_0)}{\|g_1(t_0)\|} \right\| \leq \frac{(k+1)d\rho_0}{k\nu_0^2\|g_1(t_0)\|} . \tag{18}$$

Similarly we obtain for the angles θ_{j+1} between the subsequent iterates t_{j+1} and the subspace N

$$\sin\theta_{j+1} = \min_{y\in N} \|t_{j+1} - y\| \leq d(\rho_j/\nu_j)^2/\rho_{j+1} . \tag{19}$$

Using the uniform bound

$$c \equiv \max\{\|\bar{C}(t)\| + \nu(t) \mid t \in S\}$$

we derive from (11) with (12)

$$\|x_{j+1} - x^* - \frac{k}{k+1}(x_j - x^*)\| \leq \left[\frac{kc}{(k+1)\nu_j} \sin\theta_j + \frac{d}{\nu_j^2}\rho_j\right]\rho_j \tag{20}$$

which implies by the inverse triangular inequality

$$\left|\frac{\rho_{j+1}}{\rho_j} - \frac{k}{k+1}\right| \leq \left[\frac{kc}{(k+1)\nu_j} \sin\theta_j + \frac{d}{\nu_j^2}\rho_j\right] \tag{21}$$

and furthermore

$$\sin\Delta\psi_j = \min_{\lambda \in \mathbb{R}} \|t_j - \lambda t_{j+1}\| \leq \left[\frac{c}{\nu_j} \sin\theta_j + \frac{(k+1)d}{k\nu_j^2}\rho_j\right] \tag{22}$$

where $\Delta\psi_j$ is the angle between two consecutive directions t_j and t_{j+1}.

According to Theorem 1.9 (iii) any subquadratically converging Newton sequence has a tangent in N. Now we can show that the Newton iteration from some initial point $x_1 = x^* + \rho_1 t_1$ must converge to a regular singularity x^* if ρ_1 and the angle between t_1 and some regular $s \in S' \cap N$ are sufficiently small. This result was obtained by Decker and Kelley [9] under the assumption that x^* is strongly regular as will be defined after the next lemma and by Reddien [7] under the assumption that $\det(\nabla^{k+1} f(x^*) s^k) \neq 0$ for some $s \in N$. This is never satisfied if any linear combination of the component functions of f is linear in x. Excluding only those directions in N along which the smallest singular value of ∇f is $o(\rho^k)$ we obtain the following result.

LEMMA 2.2 *Linear Convergence near N at Regular Singularity*

Let $f \in C^{k+1,1}(\mathbb{R}^n, \mathbb{R}^n)$ have a regular singularity of order $k \geq 1$ at x^*. Then there are two nonnegative continuous functions

$$\hat{\phi} : N \cap S \to \mathbb{R} \quad \text{and} \quad \hat{\rho} : N \cap S \to \mathbb{R}$$

such that for any regular direction $s \in S' \cap N$ the starlike domain

$$\hat{W}(s) \equiv \{x^*+\rho t \mid t\in S, \cos^{-1}(t^T s) < \hat{\phi}(s) , 0<\rho<\hat{\rho}(s)\}$$

is nonempty and any Newton sequence $\{x_j=x^*+\rho_j y_j\}_{j\geq 1}$ from some $x_1 \in \hat{W}(s)$ onverges to x^* with

$$|\rho_{j+1}/\rho_j - k/(k+1)| \leq [4(k+1)]^{-1} \quad \text{for all} \quad j \geq 1 \tag{23}$$

and in the limit

$$\rho_{j+1}/\rho_j \to k/(k+1) \quad , \quad t_j \to t \in N \cap S' \ . \tag{24}$$

Proof. With the convention $\min(\emptyset) = 90^\circ$ the angle

$$\phi(s) \equiv \frac{1}{2} \min\{\cos^{-1}(t^T s) \mid t\in S\cap\pi_0^{-1}(0)\} \leq 45^\circ \tag{25}$$

is obviously a nonnegative continuous function in $s \in N \cap S$ with $\phi^{-1}(0) = S \cap N \cap \pi^{-1}(0)$. Consequently the two minima

$$\hat{\nu}(s) \equiv \min\{\nu(t) \mid t\in S , \cos^{-1}(t^T s) \leq \phi(s)\} , \tag{26}$$

$$\hat{r}(s) \equiv \min\{\bar{r}(t) \mid t\in S , \cos^{-1}(t^T s) \leq \phi(s)\} \tag{27}$$

exist and are both nonnegative and continuous on $S \cap N$ with $\hat{\nu}^{-1}(0) = \hat{r}^{-1}(0) = \phi^{-1}(0)$. Abbreviating $\chi(s) \equiv \frac{1}{4} \sin \phi(s) \leq \frac{1}{4}$ we can now define recursively

$$\sin \hat{\phi}(s) \equiv \min\left\{ \frac{\chi(s)}{kc/\hat{\nu}(s)+k-\chi(s)} \ , \ \frac{(k+1)\,d\hat{r}(s)}{(k-\chi(s))\hat{\nu}^2(s)} \right\} , \tag{28}$$

$$\hat{\rho}(s) \equiv \frac{(k-\chi(s))\hat{\nu}^2(s)}{(k+1)d} \sin \hat{\phi}(s) , \tag{29}$$

which ensures $\hat{\rho}(s) \le \hat{r}(s) \le \bar{r}(s)$. Both functions are nonnegative and continuous on $N \cap S$ with $\hat{\phi}^{-1}(0) = \hat{\rho}^{-1}(0) = \phi^{-1}(0)$ so that for $s \in S \cap N$ $\hat{W}(s) = \emptyset$ iff $\pi_0(s) = 0$.

Keeping $s \in N \cap S'$ fixed we show by induction that the sequence of Newton iterates $\{x_j = x^* + \rho_j t_j\}_{j \ge 1}$ from any initial point $x_1 \in \hat{W} \equiv \hat{W}(s)$ maintains the properties

$$\rho_j < \hat{\rho} \equiv \hat{\rho}(s) \ , \ \theta_j < \hat{\phi} \equiv \hat{\phi}(s) \ , \ \psi_j < \phi \equiv \phi(s) \tag{30}$$

where ψ_j is the angle between s and t_j , whose boundedness by ϕ implies $\nu_j \ge \hat{\nu} \equiv \hat{\nu}(s)$ which will be used frequently. For the first iterate x_1 the three conditions must hold by definition of $\hat{W}$ and because of the inequality

$$\sin \theta_1 = \min_{z \in N} \|t_1 - z\| \le \|t_1 - ss^T t_1\| < \sin \hat{\phi} \ . \tag{31}$$

Assuming that (30) holds for all $i \le j$ we obtain with (28)

$$\frac{kc}{(k+1)\nu_i} \sin \theta_i + \frac{d}{\nu_i^2} \rho_i \le \frac{\sin \hat{\phi}}{(k+1)} (kc/\hat{\nu} + k - \chi(s)) \le \frac{\chi}{(k+1)} \equiv \frac{\chi(s)}{(k+1)} \ ,$$

which implies by (21)

$$\frac{k-\chi}{k+1} \le \frac{\rho_{i+1}}{\rho_i} \le \frac{k+\chi}{k+1} \qquad \text{for} \qquad i = 1 \ldots j \ , \tag{32}$$

so that

$$\rho_{i+1} \le \rho_1 \left(\frac{k+\chi}{k+1}\right)^i < \hat{\rho} \quad \text{for} \quad i = 1 \ldots j \ . \tag{33}$$

Using the left inequality in (32) we obtain from (19) and (29)

$$\sin \theta_{j+1} \le \frac{d\rho_j (k+1)}{\hat{\nu}^2 (k-\chi)} < \sin \hat{\phi} \ . \tag{34}$$

In order to obtain an upper bound on ψ_{j+1} we consider the sums

$$\sum_{i=1}^{j} \rho_i \leq \rho_1 \left(\frac{k+1}{1-\chi}\right) \leq \frac{\hat{\nu}^2(k-\chi)}{d(1-\chi)} \sin\hat{\phi} \tag{35}$$

and

$$\sum_{i=1}^{j-1} \sin\theta_{i+1} \leq \frac{d(k+1)}{\hat{\nu}^2(k-\chi)} \sum_{i=1}^{j-1} \rho_i \leq \frac{(k+1)}{(1-\chi)} \sin\hat{\phi} \ . \tag{36}$$

Recalling the definition of $\Delta\psi_i$ in (22) we note

$$\psi_{j+1} \leq \psi_1 + \sum_{i=1}^{j} \Delta\psi_i$$

which implies because of $\psi_1 \equiv \cos^{-1}(s^T t_1) < \hat{\phi}$

$$\sin\psi_{j+1} < \sin\hat{\phi} + \sum_{i=1}^{j} \sin\Delta\psi_i \ .$$

Using (35) and (36) we derive from (22)

$$\sum_{i=1}^{j} \sin\Delta\psi_i \leq \frac{c}{\hat{\nu}} \left[\sin\theta_1 + \left(\frac{k+1}{1-\chi}\right) \sin\hat{\phi}\right] + \frac{(k+1)(k-\chi)}{k(1-\chi)} \sin\hat{\phi} \ .$$

Adding to this sum $\sin\hat{\phi}$ and applying the first inequality implicit in (28) we find

$$\sin\psi_{j+1} \leq \frac{\chi}{(1-\chi)} \left[\frac{(2+k-\chi)c/\hat{\nu} + (k+2k-2k\chi-\chi)/k}{kc/\hat{\nu}+k-\chi}\right] \ .$$

It can be checked easily that the fraction in brackets is always ≤ 3 so that by definition of $\chi \leq \frac{1}{4}$

$$\sin\psi_{j+1} < \sin\phi \ .$$

Thus we have shown that all iterates stay in the set

$$\bar{W}(s) \equiv \{x^*+\rho t \mid t\in S \ , \cos^{-1}(t^T s) < \phi(s) \ , \ 0<\rho<\hat{\rho}(s) \ , \ \theta(t) < \hat{\phi}(s)\}$$

which is by definition of $\hat{\rho}(s)$ a subset of R' . Since (33) and (34) hold for all $j \geq 1$ we see from (21) that

$$\lim_{j\to\infty} \rho_{j+1}/\rho_j = k/(k+1) \ ,$$

so that we must have linear convergence at the asserted rate.

Furthermore the Q-linear decline of ρ_j and θ_j implies by (22) for any $\tilde{j} \geq j$

$$\cos^{-1}(t_j^T t_{\tilde{j}}) \leq \sum_{i=j}^{\tilde{j}-1} \Delta\psi_i = O(\rho_j) \ . \tag{37}$$

Hence the $\{t_j\}$ form a Cauchy sequence in S whose limiting directions t satisfies

$$\cos^{-1}(t^T s) \leq \phi(s) \quad \text{and} \quad \theta(t) = \lim_{j\to\infty} \theta_j = 0 \ ,$$

so that $t \in N \cap S'$ by definition of ϕ in (25) and θ in (1.48)

As an immediate consequence of Lemma 2.2 we note that the union

$$W \equiv \cup \{\hat{W}(s) \mid s \in N \cap S'\}$$

is a starlike domain of convergence too. Moreover the set of intermediate points

$$\mathcal{D} \equiv \bigcup_{j=o}^{\infty} g^j(W)$$

is by (23) a domain of contraction with

$$\sup \{\|g(x)-x^*\|/\|x-x^*\| \mid x \in \mathcal{D}\} \leq \frac{k+\frac{1}{4}}{k+1} < 1 \ .$$

The domain $\mathcal{D}$ is not necessarily open but it contains the starlike domain W, which includes all regular directions in $N \cap S$.

According to (24) the limiting direction and unique tangent $t \in N$ of any Newton sequence from within $\hat{W}(s)$ is regular too so, that all but finitely many iterates must belong to the starlike domain of convergence $\hat{W}(t) \subseteq W$. This suggests a certain numerical stability of the iteration

as an occasional numerical error in the calculation of the next point is unlikely to lead immediately outside the domain of convergence W. However if the errors rotate the directions t_j persistently towards an irregular direction in N the convergence pattern may break down. This can be the case even if the relative error in calculating the steps is uniformly bounded by some arbitrarily small $\varepsilon > 0$.

A sequence $\{x_j\}_{j\geq 1} \subset \mathbb{R}^n$ will be called an *approximate Newton sequence of relative accuracy* ε if

$$\|x_{j+1}-g(x_j)\| \,/\, \|x_{j+1}-x_j\| \leq \varepsilon < 1 \quad \text{for all} \quad j \geq 1 .$$

Using the triangular inequality we obtain from (11)

$$\begin{aligned} \|x_{j+1}-x^*-g_1(x_j-x^*)\| &\leq \varepsilon\|x_{j+1}-x_j\| + d(\rho_j/\nu_j)^2 \\ &\leq \varepsilon\rho_{j+1} + (\varepsilon + d\rho_j/\nu_j^2)\rho_j . \end{aligned} \tag{38}$$

Now we can replace (20) by

$$\|x_{j+1}-x^*-\frac{k}{k+1}(x_j-x^*)\| \leq \varepsilon\rho_{j+1} + \left[\varepsilon + \frac{d}{\nu_j^2}\rho_j + \frac{kc}{(k+1)\nu_j}\sin\theta_j\right]\rho_j \tag{39}$$

which implies

$$\rho_{j+1}(1-\varepsilon) \leq \left[\frac{k}{k+1} + \varepsilon + \frac{d}{\nu_j^2}\rho_j + \frac{kc}{(k+1)\nu_j}\sin\theta_j\right]\rho_j \tag{40}$$

and furthermore

$$\sin\Delta\psi_j \leq \frac{k+1}{k}\left[\varepsilon\left(1+\frac{\rho_{j+1}}{\rho_j}\right) + \frac{d}{\nu_j^2}\rho_j + \frac{kc}{(k+1)\nu_j}\sin\theta_j\right] , \tag{41}$$

where $\Delta\psi_j$ is again the angle between t_j and t_{j+1}.

Whenever

$$\rho_{j+1} \leq 2\rho_j t_j^T t_{j+1} \tag{42}$$

we have

$$\|x_{j+1} - x_j\| \leq \rho_j ,$$

so that by (38)

$$\|x_{j+1} - x^* - g_1(x_j - x^*)\| \leq (\varepsilon + d\rho_j/\nu_j^2)\rho_j ,$$

which implies

$$\sin\theta_{j+1} \leq (\varepsilon + d\rho_j/\nu_j^2)\rho_j/\rho_{j+1} , \tag{43}$$

$$\|x_{j+1} - x^* - \frac{k}{k+1}(x_j - x^*)\| \leq \left[\varepsilon + \frac{d}{\nu_j^2}\rho_j + \frac{kc}{(k+1)\nu_j}\sin\theta_j\right]\rho_j$$

and finally

$$\left|\frac{\rho_{j+1}}{\rho_j} - \frac{k}{k+1}\right| \leq \left[\varepsilon + \frac{d}{\nu_j^2}\rho_j + \frac{kc}{(k+1)\nu_j}\sin\theta_j\right] . \tag{44}$$

Whereas (43) suggests that the angles θ_j between the directions t_j and N can be uniformly bounded we see from (41) that the t_j's may rotate at each step through an angle greater than $\sin^{-1}(\varepsilon)$ within the nullspace N. Therefore we have to make the assumption that x^* is a *strongly regular singularity* in that it is balanced and all directions in N are regular, i.e.

$$N \cap \pi_0^{-1}(0) = \{0\} . \tag{45}$$

In the case $m = \dim(N) = 1$ this equation is equivalent to the regularity condition $\pi_0(N) \neq \{0\}$. For $m>1$ strong regularity is a rather restrictive condition which can only be satisfied if $p = km$ is even. By its definition π_0 is homogeneous of degree p so that for any $t \in S \cap N$

$$\pi_0(-t) = (-1)^p \pi_0(t) .$$

If $m > 1$ and p is odd there is a continuous path of directions in $S \cap N$ connecting $-t$ and t along which π_0 must vanish by the mean-value theorem at some $s \in S \cap N$. The condition (45) is equivalent to the assumption that $\bar{B}(z)$ is nonsingular for all nonzero $z \in N$ which was originally used by Decker and Kelley in [9]. Under the assumption of strong regularity we can obtain a version of Lemma 2.2 which applies to approximate rather than exact Newton sequences.

LEMMA 2.3 *Linear Convergence of Approximate Newton Sequence*

Let $f \in C^{k+1,1}(\mathbb{R}^n,\mathbb{R}^n)$ have a strongly regular singularity of order k at $x^* \in f^{-1}(0)$. Then there are positive constants $\hat{\theta}$ and $\hat{\rho}$ such that any approximate Newton sequence $\{x_j = x^* + \rho_j t_j\}_{j\geq 1}$ of relative accuracy $\varepsilon \leq \hat{\varepsilon} \equiv \dfrac{\sin\hat{\theta}}{4(k+1)}$ that starts at some x_1 in the starlike domain

$$V \equiv \{x^*+\rho t \mid t\in S\,,\ \theta(t)<\hat{\theta}\,,\ 0<\rho<\hat{\rho}\}$$

stays inside V and converges linearly to x^* with

$$\limsup_{j\to\infty} \sin\theta_j \leq 4\varepsilon$$

and

$$\limsup_{j\to\infty} \left| \frac{\rho_{j+1}}{\rho_j} - \frac{k}{k+1} \right| \leq \varepsilon\left[1 + ((k+1)\sin\hat{\theta})^{-1}\right].$$

Proof.

By assumption of strong regularity we have $S \cap N \cap \pi_0^{-1}(0) = \emptyset$, so that with the convention $\min(\emptyset) = 90^\circ$

$$\bar{\theta} \equiv \frac{1}{2}\min\{\theta(t) \mid t\in S\cap\pi_0^{-1}(0)\} \in (0,45^\circ]\,.$$

and because $\nu^{-1}(0) = S \cap \pi_0^{-1}(0) = \bar{r}^{-1}(0)$

$$\hat{r} \equiv \min\{\bar{r}(t) \mid t\in S\,,\ \theta(t) \leq \bar{\theta}\} > 0\,,$$

$$\hat{\nu} \equiv \min\{\nu(t) \mid t\in S\,,\ \theta(t) \leq \bar{\theta}\} > 0\,.$$

Now we define recursively

$$\sin \hat{\theta} \equiv \min \left\{ \sin \bar{\theta} \, , \hat{\nu}/(4kc) \right\} \leq \frac{1}{2}$$

and

$$\hat{\rho} \equiv \min \left\{ \hat{r} \, , \frac{\hat{\nu}^2}{4d(k+1)} \sin \hat{\theta} \right\} \leq \frac{\hat{\nu}^2}{8(k+1)d} \, ,$$

so that

$$\hat{\varepsilon} \equiv \frac{\sin \hat{\theta}}{4(k+1)} \leq \frac{1}{8(k+1)} \, .$$

Firstly we show for arbitrary $x_j \in V$ that any x_{j+1} satisfying

$$\|x_{j+1} - g(x_j)\| \leq \hat{\varepsilon} \|x_{j+1} - x_j\|$$

belongs also to V .

By definition of $\hat{\nu}$, $\hat{\theta}$ and $\hat{\rho}$ we have for all $x_j = x^* + \rho_j t_j \in V$ with $\nu_j \equiv \nu(t_j)$

$$\frac{d}{\nu_j^2} \rho_j + \frac{kc}{(k+1)\nu_j} \sin \theta_j \leq \frac{3}{8(k+1)}$$

so that with (40) by definition of $\hat{\varepsilon}$

$$\frac{\rho_{j+1}}{\rho_j} \leq \frac{k+\frac{1}{2}}{(k+1)(1-\varepsilon)} \leq \frac{k+\frac{1}{2}}{k+7/8} < 1 \, . \tag{46}$$

Thus we obtain from (41) for the angle between t_j and t_{j+1}

$$\sin \Delta\psi_j \leq \frac{k+1}{k} \left[2\varepsilon + \frac{3}{8(k+1)} \right] \leq \frac{5}{8k} \leq \frac{5}{8} \, ,$$

which implies for $\Delta\psi_j \in [0^\circ, 90^\circ]$

$$t_j^T t_{j+1} = \cos \Delta\psi_j \geq \frac{3}{4} \, ,$$

so that the condition (42) is satisfied. Hence we can apply (44) to derive

$$\frac{k-\frac{1}{2}}{k+1} \leq \frac{\rho_{j+1}}{\rho_j} \leq \frac{k+\frac{1}{2}}{k+1} \, , \tag{47}$$

so that by (43)

$$\sin\theta_{j+1} < \left(\frac{k+1}{k-\frac{1}{2}}\right)\frac{\sin\hat{\theta}}{2(k+1)} = \frac{\sin\hat{\theta}}{2k-1} \le \sin\hat{\theta}\ .$$

Therefore all iterates $\{x_j\}$ must belong to V and since by (47)

$$\rho_{j+1} \le \rho_1\left(\frac{k+\frac{1}{2}}{k+1}\right)^j < \hat{\rho}\left(\frac{k+\frac{1}{2}}{k+1}\right)^j$$

the approximate Newton sequence converges linearly to x^*. Applying again (43) we have with (47)

$$\limsup_{j\to\infty} \sin\theta_j \le \varepsilon \limsup_{j\to\infty} (\rho_j/\rho_{j+1}) < 4\varepsilon$$

and similarly by (44)

$$\limsup_{j\to\infty} \left|\frac{\rho_{j+1}}{\rho_j} - \frac{k}{k+1}\right| \le \varepsilon\left(1 + \frac{4kc}{(k+1)\hat{\vee}}\right),$$

which is equivalent to the last assertion by definition of $\hat{\theta}$. ////

Lemma 2.3 establishes a remarkable numerical stability of the Newton iteration in the neighbourhood of strongly regular singularities even though the assumption that a sequence of Newton steps can be calculated with uniformly bounded relative error is certainly optimistic. However we can realistically expect that the error occurring in the calculation of the steps has a comparatively small component orthogonal to the nullspace N of the Jacobian at x^* and a main component parallel to it. Even if the latter is rather large (e.g. 25% of total step length) the iterates would remain inside V since the bound (43) on the angle between the t_j's and N is largely unaffected.

As a consequence of Lemma 2.3 we note that V is a domain of contraction not only with respect to g but to any approximate iteration function $\tilde{g} : \mathbb{R}^n - \delta^{-1}(0) \to \mathbb{R}^n$ that satisfies

$$\|\tilde{g}(y) - g(y)\|/\|\tilde{g}(y) - y\| \le \hat{\varepsilon} \quad \text{for all} \quad y \in V\ .$$

The ratio of contraction is bounded by (47) so that

$$\sup \{\|\tilde{g}(y) - x^*\|/\|y - x^*\| \mid y \in V\} \leq (k+\tfrac{1}{2})/(k+1) .$$

3. First Step Analysis and Main Result

If $N = \mathbb{R}^n$, i.e. in the case of pure, regular singularities, the two lemmas of Section 2 are quite strong since then the starlike domain of convergence W has density 1 at x^*, and the domain of contraction V contains a deleted spherical neighbourhood of x^*. However the assumptions of Lemma 2.3 are unlikely to be satisfied if $m = n > 1$ since we have already noted that strong singularity is a rather restrictive condition whenever $m > 1$. In general we can expect that $m = \dim(N)$ is small compared to n so that the directions included in W or V represent only a small fraction of the full unit sphere S in $\mathbb{R}^n$. Fortunately we can show that for most directions $t_0 \in S$ the first step from some point $x = x^* + \rho_0 t_0 \in R'$ leads into W or V provided ρ_0 is sufficiently small and the calculation of the step is sufficiently accurate.

THEOREM 2.4 *Starlike Domain of Convergence R at Regular Singularity*

Let $f \in C^{k+1,1}(\mathbb{R}^n, \mathbb{R}^n)$ have a regular singularity of order k at x^*. Then

(i) There is a nonnegative continuous function $r : S \to \mathbb{R}$ such that the Newton iteration converges linearly to x^* with Q-factor $k/(k+1)$ from any initial point in the starlike domain

$$R \equiv \{x = x^* + \rho t \mid t \in S , 0 < \rho < r(t)\} .$$

(ii) The domain R has density 1 at x^* since the closed set of excluded directions

$$r^{-1}(0) = \{t \in S \mid R \cap \{x^* + \rho t\}_{\rho>0} = \emptyset\}$$

is given by the intersection of S with the solution set $\pi^{-1}(0)$ of the nontrivial homogeneous polynomial

$$\pi(z) \equiv \pi_0(\pi_0(z)g_1(z))\pi_0(z) \ .$$

(iii) For any $t \in r^{-1}(0)$ that is not necessarily excluded from all starlike domains of convergence to x^* either of the following conditions must be satisfied

- $\pi_0(z)$ or $\hat{\pi}(z) \equiv \pi_0(\pi_0(z)g_1(z))$ attains a local extremum at t .
- $g_1(t) = 0$ and $\dfrac{d^{\ell+1}}{d\mu^{\ell+1}} f(x^*+\mu t)\Big|_{\mu=0} = 0$

where $\ell \in [1,\bar{k}]$ is the smallest index for which $\nabla^{\ell+1} f(x^*)$ is nontrivial.

(iv) For Newton's method to have a spherical domain of convergence about x^* it is sufficient that $\pi_0^{-1}(0) = \{0\}$ and necessary that π_0 is either nonnegative or nonpositive on $\mathbb{R}^n$ (assuming $n>1$).

Proof. Without loss of generality we assume $x^* = 0$.

(i) Since $\hat{\rho}(s)$ and $\sin \hat{\phi}(s)$ are bounded the function

$$r(t) \equiv \min\left\{\bar{r}(t) \ , \ \frac{\nu^2(t)\hat{\rho}(g_1(t)/\|g_1(t)\|)}{dr_b+c\nu(t)+\nu^2(t)} \ , \ \frac{\|g_1(t)\|\nu^2(t)\sin\hat{\phi}(g_1(t)/\|g_1(t)\|)}{2d}\right\}$$

is well defined and continuous on S . Now we derive from (18) and (21) that for any $x_0 = x^* + \rho_0 t_0 \in R$

$$\sin \psi_1(t_0) < \sin \hat{\phi}(g_1(t_0)/\|g_1(t_0)\|) \quad \text{and} \quad \rho_1 < \hat{\rho}(g_1(t_0)/\|g_1(t_0)\|) \ ,$$

so that $x_1 \in \hat{W}(g_1(t_0)/\|g_1(t_0)\|)$ which implies the assertion by Lemma 2.2.

(ii) Inspecting the individual terms in the definition of $r(t)$ we see that $r(t) = 0$ iff

$$t \in \bar{r}^{-1}(0) = \nu^{-1}(0) \subseteq \pi_0^{-1}(0) \quad \text{or otherwise} \quad g_1(t) \in \pi_0^{-1}(0) \ .$$

It follows directly from the expression for $g_1(z)$ in (12) that the m nontrivial components of $\pi_0(z)g_1(z)$ are homogeneous polynomials of degree $p+1$ in z so that $\pi(z)$ is a homogeneous polynomial of degree $(p+2)p$ in $z \in \mathbb{R}^n$. Clearly $t \in S$ belongs to the solution set $\pi^{-1}(0)$ of π iff

$$\pi_0(t) = 0 \quad \text{or otherwise} \quad \pi_0(g_1(t)) = 0$$

which shows that $r^{-1}(0) = \pi^{-1}(0) \cap S$. For any $z \in N$ we have $g_1(z)\pi_0(z) = z\pi_0(z)k/(k+1)$ and consequently $\pi(z) = \pi_0^{p+1}(z)[k/(k+1)]^p$ so that neither $\pi_0(z)$ nor $\pi(z)$ can vanish identically as by assumption $\pi_0(s) = \det(\bar{B}(s)) \neq 0$ for some $s \in N$.

(iii) We know from Lemma 1.5 (iii) that $t \in \pi_0^{-1}(0)$ must be tangential to $\delta^{-1}(0)$ and therefore by (1.15) necessarily excluded from any starlike domain of invertibility unless π_0 attains an extremum at t. Now suppose $\pi(z)$ does not attain a local extremum at some $t \in S \cap \hat{\pi}^{-1}(0) - \pi_0^{-1}(0)$, that is included in a starlike domain of convergence A with boundary function a. Since A is open and π_0 has the same sign in a sufficiently small neighbourhood of t in S, there must be sequences $t_j^- \to t$ and $t_j^+ \to t$ of included directions such that

$$\frac{\hat{\pi}(t_j^-)}{\pi_0(t_j^-)^{km}} = \pi_0(g(t_j^-)) < 0 < \pi_0(g(t_j^+)) = \frac{\hat{\pi}(t_j^+)}{\pi_0(t_j^+)^{km}} \ .$$

Since a is lower semicontinuous and positive at t the Newton steps from $y_j^- \equiv \mu_j t_j^-$ and $y_j^+ \equiv \mu_j t_j^+$ to z_j^- and z_j^+ respectively are well defined for μ_j smaller than some $\bar{\mu}_j$ with $\bar{\mu}_j \to a(t)$. Combining (1.11) with (11) we obtain

$$\delta(z_j^-) = \mu_j^{km}\pi_0(g_1(t_j^-)) + o(\mu_j^{km})$$

so that $\mu_j < \bar{\mu}_j$ may be chosen sufficiently small such that

$$\delta(z_j^-) < 0 < \delta(z_j^+) .$$

Since $\nabla f^{-1}(x)f(x)$ is continuous on a domain of invertibility there must be multipliers $\alpha_j \in (0,1)$ such that the Newton step from each

$$y_j \equiv \mu_j(\alpha_j t_j^- + (1-\alpha_j)t_j^+) \in A$$

leads to a point $z_j \in \delta^{-1}(0)$. By assumption A is a domain of convergence to $x^* = 0$, so that we must have $z_j = 0$ for all j . Since the y_j/μ_j are convex combinations of the t_j^- and t_j^+ we find $s_j \equiv y_j/\|y_j\| \to t$, so that t is tangential to the set of points from which Newton's method converges in one step. Writing $y_j = \tau_j s_j$ we derive from $z_j = 0$

$$0 = \nabla f(y_j)y_j - f(y_j) = \left[\tau_j \nabla f(\tau_j s_j) - \int_0^{\tau_j} \nabla f(\mu s_j)d\mu\right]s_j$$

$$= \frac{\ell\tau_j^{\ell+1}}{(\ell+1)!}\nabla^{p+1}f(x^*)s_j^{\ell+1} + O(\tau_j^{\ell+2}) .$$

Here we have used the fact that $\nabla^i f(x^*) = 0$ for $i = 2..\ell$. After division by $\tau_j^{\ell+1}$ we obtain in the limit

$$\frac{d^{\ell+1}}{d\mu^{\ell+1}} f(x^*+\mu t)\Big|_{\mu=0} = \lim_{j\to\infty} \nabla^{\ell+1}f(x^*)s_j^{\ell+1} = 0 .$$

Because of (1.47) a similar argument applied to the identity $P\nabla f(y_j)y_j = Pf(y_j)$ shows that

$$k!\,P\nabla^{k+1}f(x^*)t^{k+1} = \begin{pmatrix} \bar{B}(t) & , & \bar{C}^T(t) \\ 0 & , & 0 \end{pmatrix} t = 0$$

which implies $g_1(t) = 0$ as $\bar{B}(t)$ is nonsingular.

(iv) If $\pi_0^{-1}(0) = \{0\}$ the set $\nu^{-1}(0) = \pi_0^{-1}(0) \cap S$ is empty, so that $\tilde{\nu} \equiv \min \{\nu(t) \mid t \epsilon S\} > 0$ and furthermore

$$\tilde{r} \equiv \min \left\{ \frac{\tilde{\nu}^2}{2(k+1)d} , \min\{r(t) \mid t\epsilon S , \|g_1(t)\| \geq 1\} \right\} > 0 .$$

Now consider any point $x_j = \rho_j t_j$ with $\rho_j < \tilde{r}$. If $\|g_1(t_j)\| \geq 1$ then $\rho_j < \tilde{r} \leq r(t_j)$ so that convergence is guaranteed by (i) . If $\|g_1(t_j)\| \leq 1$ we obtain from (11)

$$\rho_{j+1} < \left(\frac{k}{k+1} + \frac{d\rho_j}{\nu_j^2} \right) \rho_j \leq \frac{2k+1}{2k+2} \rho_j .$$

Thus we must have in any case at least Q-linear convergence to x^* .

If π_0 attains positive and negative values there must be a $t \in S \cap \pi_0^{-1}(0)$ that is neither minimiser nor maximiser and therefore necessarily excluded by Lemma 1.5 (iii). ////

The fact that R and consequently the full domain of convergence X_0 discussed in Section 1.5 have density 1 at a regular singularity is probably the most important result of this thesis. Whenever the equivalent regularity conditions

$$\det(\bar{B}(t)) \neq 0 \iff \|\nabla f^{-1}(x^*+\rho t)\| = O(\rho^{-k})$$

are satisfied for at least one $t \in N \cap S$ then the probability that Newton's method converges linearly to x^* from a given point x_0 in the ball B_ρ is $1 - o(\rho^0)$.

Since nontrivial homogeneous polynomials are unbounded and all their stationary points must have zero value it is quite likely that they have no local extreme besides the origin. If this is the case and the set

$$T \equiv \left\{ t\epsilon S' \mid g_1(t) = 0 , \nabla^{\ell+1} f(x^*) t^{\ell+1} = 0 \right\} \subseteq S$$

is empty then the set of directions excluded from R is minimal so that the boundary function of the maximal starlike domain of convergence to x^* differs from $r(t)$ only in size but not sign. In the case of pure singularities we have $m = n$ and $P = I$ so that T reduces to

$$T \equiv \{t \in S' \mid \nabla^{k+1} f(x^*) t^{k+1} = 0\} \tag{48}$$

which must be empty if the isolation condition (1.47) is satisfied. For any $t \in r^{-1}(0)$ that satisfies either of the two conditions in Theorem 6.1 (iii) the question whether it is necessarily excluded can only be decided on the basis of $(2+k)$-th and higher derivative information.

To illustrate the result we consider the following examples in two dimensions. After suitable nonsingular affine transformations any function $f \in C^3(\mathbb{R}^2, \mathbb{R}^2)$ with a Jacobian of rank 1 at a first order singularity can be written in the form

$$f\begin{pmatrix}\xi\\ \zeta\end{pmatrix} = \begin{pmatrix} \frac{\tau}{2}\xi^2 + \frac{\varepsilon}{2}\zeta^2 \\ \zeta + \frac{\alpha}{2}\xi^2 + \beta\xi\zeta + \frac{\gamma}{2}\zeta^2 \end{pmatrix} + O(\left\|\begin{matrix}\xi\\ \zeta\end{matrix}\right\|^3) ,$$

so that

$$\nabla f\begin{pmatrix}\xi\\ \zeta\end{pmatrix} = \begin{pmatrix} \tau\xi & , & \varepsilon\zeta \\ \alpha\xi + \beta\zeta & , & 1 + \beta\xi + \gamma\zeta \end{pmatrix} + O(\left\|\begin{matrix}\xi\\ \zeta\end{matrix}\right\|^2) .$$

If $\tau = 0$ the problem is irregular as B vanishes identically. Otherwise we can use linear transformations to obtain $\tau = 1$, $\alpha = 0$ and $\varepsilon \in \{-1, 0, +1\}$. Thus we find

$$\pi_0(\xi,\zeta) = \xi , \; g_1(\xi,\zeta) = \tfrac{1}{2}((\xi^2+\varepsilon\zeta^2)/\xi, 0)^T , \; \hat{\pi}(\xi,\zeta) = \tfrac{1}{2}(\xi^2+\varepsilon\zeta^2) .$$

Consequently the set of excluded directions is given by

$$S \cap \pi^{-1}(0) = \begin{cases} \{\pm(0,1)^T\} & \text{if} \quad \varepsilon \in \{0,1\} \\ \{\pm(0,1)^T,\ (\pm 1,\pm 1)^T/\sqrt{2}\} & \text{if} \quad \varepsilon = -1 \end{cases} .$$

Since π_0 attains positive and negative values in the neighbourhood of $(0,1)^T$ the ζ-axis is necessarily excluded from any starlike domain of invertibility, so that $S \cap \pi^{-1}(0)$ is minimal if $\varepsilon \geq 0$. If $\varepsilon = -1$ the two straight lines $\{\xi = \pm\zeta\}$ are mapped by g_1 into the origin but $\hat{\pi}$ attains positive and negative values in their neighbourhood. Since $p = k = 1$ we have with $\tau = 1$ and $\alpha = 0$

$$\frac{d^2}{d\mu^2} f\begin{pmatrix}\mu\\ \mu\end{pmatrix}\Big|_{\mu=0} = \begin{pmatrix}0\\ 2\beta+\gamma\end{pmatrix} \quad \text{and} \quad \frac{d^2}{d\mu^2} f\begin{pmatrix}\mu\\ -\mu\end{pmatrix}\Big|_{\mu=0} = \begin{pmatrix}0\\ -2\beta+\gamma\end{pmatrix}$$

so that by Theorem 2.4 (iii) the directions $\{(\pm 1,\pm 1)/\sqrt{2}\}$ are necessarily excluded whenever $|\gamma| \neq |2\beta|$.

Secondly we consider the case where the Jacobian of $f \in C^3(\mathbb{R}^2,\mathbb{R}^2)$ vanishes at a regular first order singularity. After suitable affine transformations we have with $\varepsilon \in \{-1,0,1\}$

$$f\begin{pmatrix}\xi\\ \zeta\end{pmatrix} = \begin{pmatrix} \frac{1}{2}\xi^2 + \frac{\varepsilon}{2}\zeta^2 \\ \alpha\xi\zeta + \frac{\beta}{2}\zeta^2 \end{pmatrix} + o(\left\|\begin{matrix}\xi\\ \zeta\end{matrix}\right\|^3) ,$$

so that

$$\nabla f\begin{pmatrix}\xi\\ \zeta\end{pmatrix} = \begin{pmatrix} \xi & , & \varepsilon\zeta \\ \alpha\zeta & , & \alpha\xi + \beta\zeta \end{pmatrix} + o(\left\|\begin{matrix}\xi\\ \zeta\end{matrix}\right\|^2) .$$

We need only consider

$$\pi_0(\xi,\zeta) = \alpha\xi^2 + \beta\xi\zeta - \alpha\varepsilon\zeta^2 = (\xi,\zeta)\begin{pmatrix} \alpha & , & \frac{1}{2}\beta \\ \frac{1}{2}\beta & , & -\alpha\varepsilon \end{pmatrix}\begin{pmatrix}\xi\\ \zeta\end{pmatrix} .$$

Depending on whether the determinant $\det(\nabla^2\pi_0) = -(\alpha^2\varepsilon + \frac{1}{4}\beta^2)$ is positive, negative or zero there is a spherical domain of convergence, a minimal set of two necessarily excluded straight lines or one not necessarily excluded straight line respectively. The last case is particularly interesting as we have for $\alpha = 1$ and $\beta = 0 = \varepsilon$

$$f\begin{pmatrix}\xi\\ \zeta\end{pmatrix} = \begin{pmatrix} \frac{1}{2}\xi^2 \\ \xi\zeta \end{pmatrix} + o(\|\begin{matrix}\xi\\ \zeta\end{matrix}\|^3) ,$$

with

$$\nabla f\begin{pmatrix}\xi\\ \zeta\end{pmatrix} = \begin{pmatrix} \xi & , & 0 \\ \zeta & , & \xi \end{pmatrix} + o(\|\begin{matrix}\xi\\ \zeta\end{matrix}\|^2) .$$

This example has already been considered briefly in Section 2.1. If the higher order terms are zero all points on the ζ-axis are solutions of f and Newton's method converges from all other points linearly to the particular solution $x^* = 0$. Even though x^* is not an isolated solution we obtain from Theorem 2.4 a starlike domain of convergence to x^* with only the ζ-axis excluded. In contrast the result is not applicable to the other solution points which are first order singularities of degree zero with $m = 1$.

If the higher order terms are of the form $(-\frac{1}{4}\zeta^4 , 0)^T$ then both

$$f\begin{pmatrix}\xi\\ \zeta\end{pmatrix} = \begin{pmatrix} \frac{1}{2}\xi^2 - \frac{1}{4}\zeta^4 \\ \xi\zeta \end{pmatrix}$$

and the determinant of the Jacobian

$$\det(\nabla f) = \det \begin{pmatrix} \xi & , & -\zeta^3 \\ \zeta & , & \xi \end{pmatrix} = \xi^2 + \zeta^4$$

vanish only at the origin $x^* = 0$. The Newton iteration is given by

$$\begin{pmatrix} \xi_{j+1} \\ \zeta_{j+1} \end{pmatrix} = \frac{1}{2}\begin{pmatrix} \xi_j \\ \zeta_j \end{pmatrix} + \frac{1}{4}\begin{pmatrix} -\xi_j \\ \zeta_j \end{pmatrix} \frac{\zeta_j^4}{\xi_j^2+\zeta_j^4}$$

which yields linear convergence to x^* from all points on the ζ-axis with ratio $\frac{3}{4}$ and from all others with ratio $\frac{1}{2}$. Consequently f has a spherical domain of convergence, which does not follow from Theorem 2.4 (iv) as the condition $\pi_0^{-1}(0) = \{0\}$ is not met. If on the other hand f is of the form

$$f\begin{pmatrix}\xi\\ \zeta\end{pmatrix} = \begin{pmatrix} \frac{1}{2}\xi^2 + \frac{1}{4}\zeta^4 \\ \xi\zeta \end{pmatrix}$$

then the determinant $\det(\xi) = \xi^2 - \zeta^4$ vanishes on the parabolas $\{\xi = \pm\zeta^2\}$ so that the ζ-axis must be necessarily excluded.

At strongly regular singularities we can obtain a version of Theorem 2.4 that applies to approximate Newton sequences but restricts the initial points to starlike domains, which have in general a density less than 1 at x^*.

THEOREM 2.5 *Domains of Convergence for Approximate Newton Sequences*

Let $f \in C^{k+1,1}(\mathbb{R}^n,\mathbb{R}^n)$ have strongly regular singularity of order k at x^*. Then we have with $\hat{\varepsilon}$ and $\sin\hat{\theta}$ as in Lemma 2.3.

(i) There is a constant $\bar{\varepsilon} \le \hat{\varepsilon}$ and a family of nonnegative continuous functions $\{r_\varepsilon\}_{\varepsilon\in[0,\bar{\varepsilon}]}$ from S to $\mathbb{R}$ such that any approximate Newton sequence of relative accuracy $\varepsilon \in [0,\bar{\varepsilon}]$ converges Q-linearly to x^* if the initial point belongs to the nonempty starlike domain

$$R_\varepsilon \equiv \{x^*+\rho t \mid t\in S,\ 0<\rho<r_\varepsilon(t)\}\ .$$

(ii) The closed set of excluded directions is given by

$$r_\varepsilon^{-1}(0) = \left\{t\in S' \mid \|g_1(t)\| \le \varepsilon[(1-\varepsilon)\sin\hat{\theta}-\varepsilon]^{-1}\right\} \cup \bar{r}^{-1}(0)\ .$$

(iii) If the set

$$C \equiv \{t \in S \mid P\nabla^{k+1} f(x^*) t^{k+1} = 0\} \subseteq g_1^{-1}(0) \cup \bar{r}^{-1}(0) \tag{49}$$

is empty there is an $\tilde{\varepsilon} \in (0,\bar{\varepsilon})$ such that for all $\varepsilon \in [0,\tilde{\varepsilon})$

$$r_\varepsilon^{-1}(0) = S \cap \pi_0^{-1}(0) = \bar{r}^{-1}(0) ,$$

which implies that R_ε includes all regular directions and has therefore density 1 at x^* .

Proof. Starting from some $x_0 = x^* + \rho_0 t_0 \in R'$ we derive from (38) for any approximate first iterate $x_1 = x^* + \rho_1 t_1$

$$(1-\varepsilon)\rho_1 \leq (\|g_1(t_0)\| + \varepsilon + d\rho_0/\nu^2(t_0))\rho_0 \tag{50}$$

Provided $g_1(t_0)$ is nonzero the minimal angle θ_1 between t_1 and N is not greater than the angle ψ_1 between t_1 and $g_1(t_0) \in N$ so that by (38)

$$\begin{aligned} \sin\theta_1 \leq \sin\psi_1 &\equiv \min_{\lambda \in \mathbb{R}} \left\| \lambda t_1 - g_1(t_0)/\|g_1(t_0)\| \right\| \\ &\leq \frac{\varepsilon}{\|g_1(t_0)\|\rho_0}\rho_1 + \frac{\varepsilon + d\rho_0/\nu^2(t_0)}{\|g_1(t_0)\|} \\ &\leq \frac{1}{1-\varepsilon}\left[\varepsilon\left(1 + \frac{1}{\|g_1(t_0)\|}\right) + \frac{d\rho_0}{\|g_1(t_0)\|\nu^2(t_0)}\right] , \end{aligned} \tag{51}$$

where the last inequality follows by (50).

Whereas the condition $\rho_1 < \hat{\rho}$ can be met by sufficiently small ρ_0 whenever t_0 is regular we see from (51) that the condition $\sin\theta_1 < \sin\hat{\theta}$ can only be satisfied if

$$\frac{\varepsilon}{1-\varepsilon} < \frac{\sin\hat{\theta}}{1 + \|g_1(t_0)\|^{-1}} ,$$

which is for $\varepsilon \in (0,1)$ equivalent to

$$\varepsilon < \frac{\sin\hat{\theta}}{1+\sin\hat{\theta}+\|g_1(t_0)\|^{-1}} \quad \frac{\sin\hat{\theta}}{1+\sin\hat{\theta}} \ .$$

Taking the supremum over the initial directions we obtain the upper bound

$$\bar{\varepsilon} \equiv \min\left\{\hat{\varepsilon}_1, \frac{\sin\hat{\theta}}{1+\sin\hat{\theta}+\min\{\|g_1(t_0)\| \mid t_0 \in S'\}^{-1}}\right\}$$

which is well defined and positive as g_1 cannot vanish identically by Theorem 1.6.

Abbreviating

$$\eta_\varepsilon(t) \equiv (1-\varepsilon)\|g_1(t)\|\sin\hat{\theta} - \varepsilon(1+\|g_1(t)\|)$$

we can now define the boundary function

$$r_\varepsilon(t) \equiv \max\left\{0, \min\{\bar{r}(t), \frac{2}{3}\hat{\rho}\|g_1(t)\|^{-1}, \eta_\varepsilon(t)\nu^2(t)/d\}\right\}$$

which implies (ii) as

$$r_\varepsilon(t) = 0 \quad \text{iff} \quad \pi_0(t) = 0 \quad \text{or otherwise} \quad \eta_\varepsilon(t) \le 0 \ .$$

It follows from (51) with the fourth inequality implied in the definition of r_ε that $\sin\theta_1 < \sin\hat{\theta}$ whenever $x_0 \in R_\varepsilon \subseteq R$. Furthermore we derive from the same inequality

$$\|g_1(t_0)\| + \varepsilon + d\rho_0/\nu^2(t_0) \le \frac{3}{2}(1-\varepsilon)\|g_1(t_0)\| \ ,$$

so that by (50) and the third inequality implied in the definition of r_ε also $\rho_1 < \hat{\rho}$. Thus the first step of any approximate Newton sequence of relative accuracy $\varepsilon \in [0,\bar{\varepsilon})$ from within R_ε leads into V which was constructed as a domain of linear convergence in Lemma 2.3. The inclusion (49) holds by the second part of (7). Multiplying the same equation from the left by $\hat{A}_1(t)$ we find that for all $s \in S'$

$$\|g_1(s)\| \geq \alpha \equiv \min\left\{ \frac{k}{k+1} \|P\nabla f^{k+1}(x^*)t^{k+1}\|/\|A_1(t)\| \mid t\epsilon S\right\},$$

where the minimum on the RHS exists since the ratio of the two norms is continuous on the compact domain S. If the set C in (49) is empty α is positive and we can define

$$\tilde{\varepsilon} \equiv \min\left\{\hat{\varepsilon}, \sin\hat{\theta}/(1+\sin\hat{\theta}+\alpha^{-1})\right\},$$

so that for all $\varepsilon \in [0,\tilde{\varepsilon})$ and $t \in S'$

$$\|g_1(t)\| > \varepsilon[(1-\varepsilon)\sin\hat{\theta} - \varepsilon]^{-1}$$

which implies (iii) by (ii). ///

Considering Theorem 2.5 we note that the regular directions that are excluded from a given starlike domain R_ε are those for which $g_1(t_0)$ is comparatively small. This means for an approximate Newton step from $x_0 = x^* + \rho_0 t_0$ to $x_1 = x^* + \rho_1 t_1$ that the ratio ρ_1/ρ_0 is rather small but the minimal angle θ_1 between N and t_1 may be greater than $\hat{\theta}$. Even though x_1 can belong to the singular set or be otherwise unfavourable there is a fair chance that the next step leads into V and then to convergence. In the case of pure singularities we have $g_1^{-1}(0) = T$ as defined in (48) so that (iii) applies if the isolation condition (1.47) is satisfied. In general we can expect a comparatively stable numerical convergence of Newton's method at strongly regular singularities, including in particular all those with $m = 1$ that satisfy the isolation condition.

CHAPTER 3

MODIFICATION OF NEWTON'S METHOD AT SINGULARITIES

1. The Numerical Difficulty of Singular Problems

At first glance it might be thought that the singularity of the Jacobian at a solution point represents a merely technical difficulty for Newton's method, which could be overcome by suitably chosen alternative methods. In fact singular problems are inherently more difficult to solve than nonsingular ones, and even Q-linear convergence in a reasonably stable fashion is quite an achievement. To see this we consider an arbitrary iteration of the form

$$x_{j+1} = h(x_j, f(x_j)) \ , \tag{1}$$

where $h : \mathbb{R}^{2n} \to \mathbb{R}^n$ satisfies the identity

$$h(x,0) = x \quad \text{for all} \quad x \in \mathbb{R}^n \tag{2}$$

and has a Jacobian $\nabla_f h$ with respect to f such that

$$H(x) \equiv -\nabla_f h(x,0) \quad \text{is continuous in} \quad x \quad \text{at} \quad x^* \ . \tag{3}$$

The iteration function h may involve values of arbitrarily high derivatives of f at x and several intermediate or previous points and could even be designed or selected in view of the particular problem function f at hand.

The two conditions (2) and (3) hold in particular for iterations of the form

$$x_{j+1} = x_j - H(x_j)f(x_j) \tag{4}$$

provided $H(x)$ is continuous in x. In order to enhance the global convergence properties of Newton's method a bounded matrix H is often used as a substitute for the inverse Jacobian whenever ∇f is singular or nearly singular. Usually such modifications are only meant to apply at a finite number of intermediate points before the unmodified Newton iteration converges superlinearly to a nonsingular solution. For examples of such modified Newton methods see [18], [19] and [20]. D. Gay advocated in [21] to treat singular or nearly singular problems by defining H on the basis of the singular value decomposition of ∇f as a continuous approximation to the inverse Jacobian. Like any method of the form (1) for which (2) and (3) are satisfied this approach is not viable in the exactly singular case.

Expanding $h(x_j,f(x_j))$ at $(x_j,0)$ and $f \in C^1(\mathbb{R}^n,\mathbb{R}^n)$ at a singular solution $x^* \in f^{-1}(0)$ we obtain from (1)

$$\begin{aligned} x_{j+1} &= h(x_j,0) - H(x_j)f(x_j) + o(\|f(x_j)\|) \\ &= x_j - H(x^*)\nabla f(x^*)(x_j-x^*) + o(\rho_j) , \end{aligned}$$

so that

$$x_{j+1} - x^* = A\,(x_j-x^*) + o(\rho_j) , \tag{5}$$

where $\rho_j = \|x_j-x^*\|$ as before and

$$A \equiv I - H(x^*)\nabla f(x^*) .$$

Thus we have a perturbed linear difference equation and according to a remark on page 193 on [10] it is "essentially" necessary for linear convergence that the spectral radius of A, i.e. the modulus of its largest eigenvalue is less than 1. Apparently most results have been

developed under this condition which is clearly violated in our case since $Aw = w$ for all w in the nullspace N of $\nabla f(x^*)$. With $M \subset \mathbb{R}^n$ the range of $H(x^*)\nabla f(x^*)$ and Q the orthogonal projection onto the orthogonal complement $M^\perp$ we derive from (5)

$$Q\,(x_{j+1}-x^*) = Q\,(x_j-x^*) + o(\rho_j)\ , \tag{6}$$

which implies that h has at x^* no spherical domain of contraction in the sense of Section 1.5. This does not preclude the existence of a Q-linearly converging sequence $\{x_j\}_{j\geq 0} \subset \mathbb{R}^n - \{x^*\}$ with

$$\rho_{j+1}/\rho_j \leq \gamma \in [0,1) \quad \text{for sufficiently large } j\ .$$

Dividing (6) by ρ_j we obtain for the angle $\theta_j = \theta(t_j)$ between $t_j = (x_j-x^*)/\rho_j$ and $M^\perp$.

$$\|Qt_{j+1}\| = \cos\theta_{j+1} \geq \gamma^{-1}\cos\theta_j + o(\rho_j^0)$$

so that in the limit

$$\gamma \limsup_{j\to\infty} \cos\theta_{j+1} \geq \limsup_{j\to\infty} \cos\theta_j\ ,$$

which requires because $\gamma < 1$

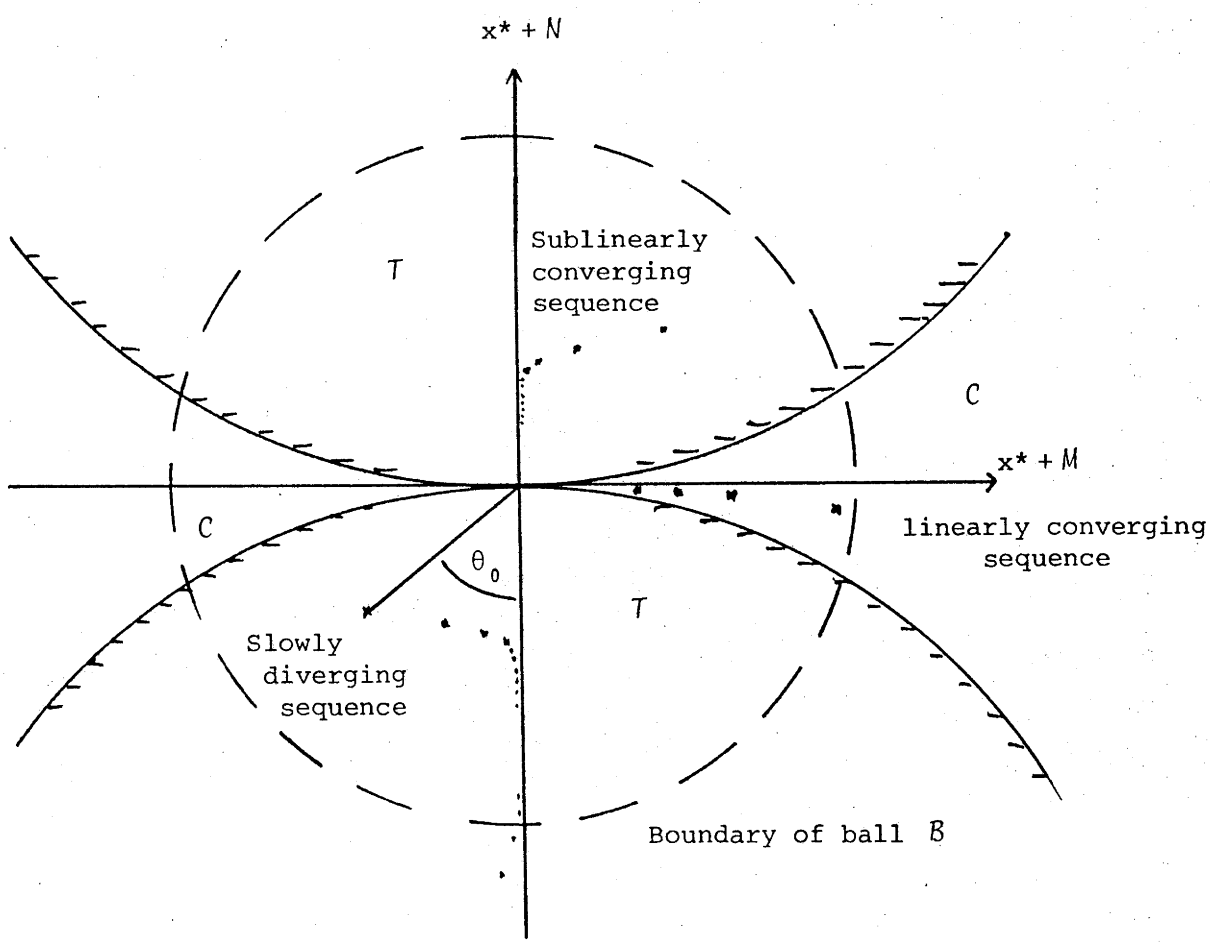
$$\lim_{j\to\infty} \theta_j = 90^\circ\ .$$

This means that the component of $x_j - x^*$ orthogonal to M must become infinitely small compared to the main component parallel to M. In practice rounding errors will prevent $\cos\theta_j$ from becoming arbitrarily small, which destroys the Q-linear rate of convergence in a theoretical sense. However, unless A has eigenvalues of modulus larger than 1, these errors need not be magnified and the main component parallel to M may still be reduced Q-linearly until the solution has been approximated with satisfactory accuracy. Nevertheless it seems obvious that Q-linear

convergence will only occur from a comparatively small set of initial points x_0, namely those for which θ_0 is close to 90°. If A is symmetric with $\|A\| = 1$ and $M^\perp = N = P(\mathbb{R}^n)$ it can be easily seen that

$$\lim_{j\to\infty} A^j = Q = P ,$$

which is always the case for the iteration (4) with H based on the singular value decomposition of the Jacobian. Then it can be expected that all initial points, for which $\cos\theta_0$ is not negligible, are projected into the affine set $x^* + N$ during the first few steps. Subsequently the steps are of $o(\rho_j)$ and may converge sublinearly to x^* or lead out of the ball $\mathcal{B}$ in which the expansion (5) is valid. In the case $n = 2$ and $\operatorname{rank}(A) = 1 = \operatorname{rank}(\nabla f(x^*))$ we have the following situation:



Depending on the higher order terms, iteration sequences from within T can theoretically inch along $x^* + N$ out of the ball B, then skirt around its boundary and finally converge through C linearly to x^*. Since a large number of intermediate steps must be expected, the overall convergence of such iterations would probably be unacceptably slow. Moreover if such essentially circular iteration sequences did exist from points arbitrarily close to x^* the whole method (1) would be highly unstable, as rounding errors could lead to repeated return trips to the boundary of B.

The situation shown in the figure is indeed typical for the general case, and according to the following result we can even rule out R-linear convergence from most initial points that are close to x^*.

THEOREM 3.1 *Sublinear Convergence of Continuous Methods*

Let the iteration function h satisfy (2), (3) and be x^* a singular solution of $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ with $\operatorname{rank}(\nabla f(x^*)) = n-m \leq n-1$. Then there exist a ball B about x^*, a starlike domain T centred at x^* and a subspace M of dimension $n-m' \leq n-m$ such that

(i) Any iteration sequence $\{x_j\}_{j\geq 0}$ from some $x_0 \in T$ that converges to x^* without ever leaving B does so R-sublinearly in that

$$\limsup_{j\to\infty} \|x_j - x^*\|^{1/j} = 1$$

(ii) The starlike domain T has the set of excluded directions $S \cap M$ and therefore density 1 at x^*.

Proof.

Firstly we reduce A by real similarity transformations to a suitable block diagonal form. Let M be the subspace of vectors $w \in \mathbb{R}^n$ for which

$$\lim_{j\to\infty} A^j w = 0 \ .$$

Since $w \in M$ implies $Aw \in M$, and $M \cap N = \emptyset$, we find that M is an $n-m' \leq n-m$ dimensional invariant subspace with respect to A. After a suitable orthogonal transformation we can assume without loss of generality

$$M = \{0\}^{m'} \times \mathbb{R}^{n-m'} \ , \tag{7}$$

so that A takes the block triangular form

$$A = \begin{pmatrix} T & O \\ S & M \end{pmatrix} .$$

By definition of M the $(n-m') \times (n-m')$ matrix M must satisfy $\lim M^j = 0$ which is according to Theorem 4 in Chapter 1 of [2] equivalent to the condition that the spectral radius of M is less than 1. On the other hand we can show by contradiction that all eigenvalues of T have modulus greater or equal to 1.

Suppose T has a pair of complex conjugate eigenvalues $\lambda, \bar{\lambda}$ of modulus $|\lambda| < 1$ with corresponding normalized eigenvectors $u, \bar{u} \in \mathbb{C}^{m'}$. Considering the sequence

$$\begin{pmatrix} y_j \\ z_j \end{pmatrix} = \begin{pmatrix} T & , & O \\ S & , & M \end{pmatrix}^j \begin{pmatrix} u+\bar{u} \\ 0 \end{pmatrix}$$

we find for some norm with $\gamma \equiv \|M\| < 1$, which exists by 2.2.8 in [10].

$$\|y_j\| = \|\lambda^j u + \bar{\lambda}^j \bar{u}\| \leq |\lambda|^j \|u+\bar{u}\| \ ,$$

and consequently

$$\|z_{j+1}\| \leq \|u+\bar{u}\| \ \|S\| \ |\lambda|^j + \gamma \|z_j\| \leq \|u+\bar{u}\| \ \|S\| / (1-\gamma) \ .$$

This gives in the limit

$$\limsup_{j\to\infty} \|z_{j+1}\| \le \gamma \limsup_{j\to\infty} \|z_j\| < \infty \quad ,$$

which implies because $\gamma < 1$

$$\lim_{j\to\infty} \|z_j\| = 0 = \lim_{j\to\infty} \|y_j\| \ .$$

Thus the real vector $(u^T+u^{-T},0)^T \in \mathbb{R}^n$ should belong to M which contradicts (7). Consequently the spectral radius of the inverse T^{-1} is less than or equal to 1 .

The matrix equation

$$UT - MU = S \tag{8}$$

represents a square linear system in the $(n-m')m'$ entries of the matrix U . According to Theorem 2.3.15 in [22] each eigenvalue of the homogeneous part on the LHS is the difference between one eigenvalue of T and one of M so that none of them can vanish. Hence (8) has a unique solution U for arbitrary S and we obtain a similarity transformation of A to the block diagonal form

$$A = \begin{pmatrix} T & , & O \\ O & , & M \end{pmatrix} = \begin{pmatrix} I & , & O \\ -U & , & I \end{pmatrix} \begin{pmatrix} T & , & O \\ S & , & M \end{pmatrix} \begin{pmatrix} I & , & O \\ U & , & I \end{pmatrix} \ .$$

As stated on page 183 in [6] there are positive constants α , β such that

$$\|T^{-q}\| \le \alpha q^{m'-1} \quad \text{and} \quad \|M^q\| \le \beta\gamma^q q^{n-m'-1} \ , \tag{9}$$

where the exponents of q allow for the worst possibility that each T and M have only one Jordan block.

Now let

$$\{x_j=x^*+(y_0^T,z_j^T)^T\}_{j\ge 0} \quad \text{with} \quad \{y_j\} \subset \mathbb{R}^{m'} \quad \text{and} \quad \{z_j\} \subset \mathbb{R}^{n-m'}$$

be the sequence of iterates generated by (1) from some initial point

$x_0 = x^* + (y_0^T, z_0^T)^T$. It can be easily verified by induction that for all $q > 0$

$$x_{j+q} - x^* = A^q(x_j - x^*) + o(\|x_j - x^*\|) \ .$$

Hence there is for each q a sequence of constants $\varepsilon_\ell^{(q)} \to 0$ such that $(\|y_j\| + \|z_j\|) \le \ell^{-1}$ implies

$$\left\| \begin{matrix} y_{j+q} - T^q y_j \\ z_{j+q} - M^q z_j \end{matrix} \right\| \le \varepsilon_\ell^{(q)} (\|y_j\| + \|z_j\|) \quad \text{for all} \quad q \ge 1 \ . \tag{10}$$

Now we find with (9) for the angle θ_j between $x_j - x^*$ and $M^\perp$.

$$\begin{aligned} \tan \theta_{j+q} = \frac{\|z_{j+q}\|}{\|y_{j+q}\|} &\le \frac{\|M^q z_j\| + \varepsilon_\ell^{(q)} \cdot (\|y_j\| + \|z_j\|)}{\|T^q y_j\| - \varepsilon_\ell^{(q)} \cdot (\|y_j\| + \|z_j\|)} \\ &\le \frac{\beta\gamma^q q^{n-m'-1} \tan \theta_j + \varepsilon_\ell^{(q)} (1 + \tan \theta_j)}{\alpha^{-1} q^{1-m'} - \varepsilon_\ell^{(q)} (1 + \tan \theta_j)} \ . \end{aligned} \tag{11}$$

Here we have used the fact that $\|T^{-q}\|^{-1} \ge \alpha^{-1} q^{1-m'}$ is the smallest singular value of T^q . For each integer $i \ge 1$ we can choose firstly q_i and then ℓ_i such that

$$i\alpha\beta\gamma^{q_i} q_i^{n-2} \le \frac{1}{3} \quad \text{and} \quad (i+1)\varepsilon_{\ell_i}^{(q_i)} \alpha q_i^{m'-1} \le \frac{1}{3} \ . \tag{12}$$

Let $\mathcal{B}$ be the ball with radius ℓ_1^{-1} about x^*, and consider for all $i > 0$ the starlike domains

$$\mathcal{T}_i \equiv \{x^* + (y^T, z^T)^T \mid \|z\|/\|y\| < i \ , \ 0 < \|z\|^2 + \|y\|^2 < \ell_i^{-2}\} \ . \tag{13}$$

For $x_0 \in \mathcal{T}_i$ we derive from (11) with (12) that $\tan \theta_{q_i} < 1$ so that the q_i-th iterate x_{q_i} lies either inside $\mathcal{T}_1$ or outside $\mathcal{B}$. Provided the full iteration sequence $\{x_j\}_{j \ge 0}$ from $x_0 \in \mathcal{T}_i$ remains inside $\mathcal{B}$ and converges to x^* , the subsequence

$$\{\tilde{x}_j = x^* + (\tilde{y}_j^T, \tilde{z}_j^T)^T \equiv x_{q_i + j \cdot q_1}\}_{j \geq 0}$$

must remain in T_1 and there is an index sequence $\tilde{\ell}_j \to \infty$ such that

$$\tilde{\rho}_j \equiv \|\tilde{x}_j - x^*\| \leq \tilde{\ell}_j^{-1} \quad \text{for all} \quad j$$

and

$$\lim_{j \to \infty} \varepsilon_{\tilde{\ell}_j}^{(q)} = 0 \quad \text{for all} \quad q \, . \tag{14}$$

Now we derive from (11) with $q = q_1$ and (12) for the angle $\tilde{\theta}_j$ between $t_j \equiv (\tilde{x}_j - x^*)/\tilde{\rho}_j$ and $M^{\perp}$

$$\limsup_{j \to \infty} \tan \tilde{\theta}_{j+1} \leq \frac{1}{3} \limsup_{j \to \infty} \tan \tilde{\theta}_j \leq \frac{1}{3} \, ,$$

which requires

$$\lim_{j \to \infty} \|\tilde{z}_j\| / \|\tilde{y}_j\| = \lim_{j \to \infty} \tan \tilde{\theta}_j = 0 \, .$$

With $\tilde{q}$ any integer and $q \equiv \tilde{q} \cdot q_1$ we obtain from (10), (9) and (14)

$$\liminf_{j \to \infty} \frac{\tilde{\rho}_{\tilde{q} \cdot (j+1)}}{\tilde{\rho}_{\tilde{q} \cdot j}} = \liminf_{j \to \infty} \left[\frac{\|\tilde{y}_{\tilde{q} \cdot (j+1)}\|^2 (1 + \tan^2 \tilde{\theta}_{\tilde{q} \cdot (j+1)})}{\|\tilde{y}_{\tilde{q} \cdot j}\|^2 (1 + \tan^2 \tilde{\theta}_{\tilde{q} \cdot j})} \right]^{\frac{1}{2}}$$

$$= \liminf_{j \to \infty} \frac{\|\tilde{y}_{\tilde{q} \cdot (j+1)}\|}{\|\tilde{y}_{\tilde{q} \cdot j}\|} \geq \liminf_{j \to \infty} \left[\frac{\|T^{q} \tilde{y}_{\tilde{q} \cdot j}\|}{\|\tilde{y}_{\tilde{q} \cdot j}\|} - \varepsilon_{\tilde{\ell}_{\tilde{q} \cdot j}}^{(q)} (1 + \tan \tilde{\theta}_{\tilde{q} \cdot j}) \right]$$

$$\geq \alpha^{-1} q^{1-m'} = \alpha^{-1} (\tilde{q} \cdot q_1)^{1-m'} \, .$$

It can be easily shown by contradiction that the linear root factor R_1 must satisfy

$$R_1\{\tilde{x}_j\} \equiv \limsup_{j \to \infty} (\tilde{\rho}_j)^{1/j} \geq \left[\alpha^{-1} (\tilde{q} \cdot q_1)^{1-m'} \right]^{1/\tilde{q}} \, .$$

Since $\tilde{q}$ may be chosen arbitrarily large we find for the full sequence $\{x_j\}_{j\geq 0}$

$$R_1\{x_j\} \geq \limsup_{j\to\infty} [\tilde{\rho}_j]^{(j\cdot q_1)^{-1}} \geq \lim_{q\to\infty} \left[\alpha^{-1} q^{1-m'}\right]^{1/q} = 1 \quad .$$

Thus we have shown that any iteration sequence from within the union

$$T = \bigcup_{i=1}^{\infty} T_i \quad ,$$

that remains inside the ball B, can only converge R-sublinearly to x^*. Inspecting (13) we note that T has the set of excluded directions

$$S \cap M = \{(0, z^T)^T \mid \|z\|=1\} \quad .$$

If we transform A back into its original, general form, the ball B is mapped into an ellipsoid $\tilde{B}$, M into a subspace $\tilde{M}$ of the same dimension $< n$ and T into a starlike domain $\tilde{T}$ with the set of excluded directions $S \cap \tilde{M}$. By Lemma 1.4 (i) $\tilde{T}$ has the density $\tau^*(\tilde{T}) = 1$ at x^*. Since R-factors are norm invariant all statements apply to the original problem with B replaced by some ball $\bar{B} \subset \tilde{B}$. ////

In view of Theorem 3.1 it is clear that either condition (2) or (3) must be violated if linear convergence is to be restored. Even if H is merely bounded or does not exist at all the dilemma is essentially unchanged as long as h is Lipschitz continuous in f. This is so because any "sensible" scheme will use the linear information provided by the Jacobian to drive the iterates into the proximity of $x^* + N$ so that subsequent steps are of $o(\rho) = O(\|f\|)$. which allows only sublinear convergence. If ∇f is Lipschitz continuous and x belongs exactly to $x^* + N$ then a step which is not $o(\rho)$ can only be achieved if $\|H(x)\|^{-1} = O(\rho)$, which is by Lemma 1.5 (iv) the case for Newton's method with $H = \nabla f^{-1}$. Thus we can conclude that unless (2) is violated the

numerical difficulties observed for Newton's method are essentially inevitable and cannot be overcome by multipoint methods or the use of higher derivatives.

Given the difficulties discussed above the performance of Newton's method at regular singularities is surprisingly good. On the basis of a detailed examination of the unmodified iteration carried out in Section 3.2 we develop methods to accelerate the convergence by variation of the stepsize or by extrapolation, in Sections 3.3 and 4.1 respectively.

The condition (2) means that any point at which f vanishes is acceptable as a fixed point of the iteration which seems a natural property of any nonlinear equation solver. However in certain applications it may be known in advance that the solution is singular in which case we can *border* the system by additional conditions e.g. $\det(\nabla f) = 0$. On the basis of LU and QR decompositions of the Jacobian, this approach is developed in Section 4.2 for singular and underdetermined systems of nonlinear equations. Test calculations with all discussed methods on a family of problems in three variables are reported in the Tables 1-10 of the Appendix.

2. Asymptotic Behaviour of Newton's Method at Regular Singularities

In Lemma 2.2 and Theorem 2.4 we were mainly concerned with the proof of convergence from within R as such. Analyzing the final convergence behaviour of the unmodified method at regular singularities more closely, we obtain the following result.

LEMMA 3.2 *Convergence Behaviour of Regular Newton Sequences*

Let $f \in C^{k+1,1}(\mathbb{R}^n,\mathbb{R}^n)$ have a regular singularity of order k at x^*. Then any Newton sequence $\{x_j=x^*+\rho_j t_j\}_{j\geq 0}$ that is *regular* in that it is

not disjoint from the starlike domain R, as defined in Theorem 2.4, exhibits the following asymptotic properties

$$\rho_j \to 0 \ , \ t_j \to t \in S' \cap N \tag{15}$$

$$\rho_{j+1}/\rho_j = k/(k+1) + O(\rho_j) \tag{16}$$

$$\cos^{-1}(t^T t_j) = \begin{cases} O(\rho_j^2) & \text{if } m = 1 = k \\ O(\rho_j) & \text{otherwise} \end{cases} \tag{17}$$

$$\sin \theta_j = \begin{cases} O(\rho_j^2) & \text{if } k = 1 \\ O(\rho_j) & \text{otherwise} \end{cases} \tag{18}$$

$$\kappa_j \equiv \frac{\|x_{j+1}-x_j\|}{\|x_j-x_{j-1}\|} = \frac{k}{k+1} + O(\rho_j) \tag{19}$$

$$\omega_j \equiv \cos^{-1}\left[\frac{(x_{j+1}-x_j)^T(x_j-x_{j-1})}{\|x_{j+1}-x_j\| \ \|x_j-x_{j-1}\|}\right] = \begin{cases} O(\rho_j^2) & \text{if } m = 1 = k \\ O(\rho_j) & \text{otherwise} \end{cases} \tag{20}$$

$$f(x_j)/\rho_j^2 = \frac{1}{2}\nabla^2 f(x^*)t^2/k^2 + O(\rho_j) \tag{21}$$

$$\frac{\sigma(x_{j+1})}{\sigma(x_j)} + O(\rho_j) = \left(\frac{k}{k+1}\right)^k \in \left[\frac{1}{e}, \frac{1}{2}\right] \tag{22}$$

$$\delta(x_{j+1})/\delta(x_j) = [k/(k+1)]^{km} + O(\rho_j) \tag{23}$$

Proof.

The two limits in (15) were already established in Lemma 2.2 and all other assertions can be derived from its proof as follows. Equation (2.34) gives the lower part of (18) and implies with (2.21) assertion (16). The

lower part of (17) is an immediate consequence of (2.37). Since $\sin \theta_{j-1} = O(\rho_{j-1})$ we observe in (2.14) for the case $k=1$ that the component of $x_j - x^*$ orthogonal to N is $O(\rho_{j-1}^3)$, which implies with (16) the upper part of (18). If $k=l=m$ the limiting tangent t must span N so that (17) is equivalent to (18) as $\cos^{-1}(t^T t_j) = \theta_j$. As a consequence of (17) we have

$$t_j^T t_{j-1} = 1 - O(\rho_j^2) ,$$

so that

$$\|x_j - x^* - (x_{j-1} - x^*)\|/\rho_j = \|t_j - t_{j-1}\rho_{j-1}/\rho_j\|$$

$$= \left[1 - 2t_j^T t_{j-1}\rho_{j-1}/\rho_j + (\rho_{j-1}/\rho_j)^2\right]^{\frac{1}{2}} = \left[(1-\rho_{j-1}/\rho_j)^2 + O(\rho_j)\right]^{\frac{1}{2}}$$

$$= 1 - \rho_{j-1}/\rho_j + O(\rho_j^2) = \frac{1}{k} + O(\rho_j) , \qquad (24)$$

where we have used (16) to obtain the last equality. Applying the above result for j and $j+1$ we find

$$\frac{\|x_{j+1} - x_j\|}{\|x_j - x_{j-1}\|} = \frac{\rho_{j+1}/k + O(\rho_{j+1}^2)}{\rho_j/k + O(\rho_j^2)} = \frac{k}{k+1} + O(\rho_j) ,$$

which proves (19).

Applying the triangular inequality in S twice we find

$$\omega_j \le \cos^{-1}\left[-\frac{(x_{j+1}^T - x_j^T)t_j}{\|x_{j+1} - x_j\|}\right] + \cos^{-1}\left[-\frac{(x_j - x_{j-1})^T t_{j-1}}{\|x_j - x_{j-1}\|}\right] + \cos^{-1}(t_j^T t_{j-1}) . \qquad (25)$$

By (17) we have $t_j^T t_{j-1} = 1 - O(\rho_j^{2\ell})$ with $\ell = 1$ or $\ell = 2$. Then we find

$$\frac{-(x_j - x_{j-1})^T t_{j-1}}{\|x_j - x_{j-1}\|} = \frac{\rho_{j-1} - \rho_j + O(\rho_j^{2\ell+1})}{\sqrt{\rho_j^2 + \rho_{j-1}^2 - 2\rho_j\rho_{j-1} + O(\rho_j^{2\ell+2})}} = 1 - O(\rho_j^{2\ell}) ,$$

which can be applied for j and $j+1$ to obtain with (17) from (25) the upper and lower part of (20) for $\ell=2$ and $\ell=1$ respectively.

For future reference we prove the more general result

$$f(x_j(\lambda)) = \frac{\rho_j^2}{2}\left[\frac{\lambda^2}{(k+1)^2} + \frac{1-\lambda}{k^2}\right]\nabla^2 f(x^*)t^2 + O(\rho_j^3) \ , \tag{26}$$

where

$$\begin{aligned} x_j(\lambda) &\equiv x_j + \lambda(x_{j+1}-x_j) \\ &= x^* + \lambda(x_{j+1}-x^*) + (1-\lambda)(x_j-x^*) \\ &= x^* + \rho_j(\lambda g_1(t_j) + (1-\lambda)t_j) + \rho_j^2 \lambda g_2(t_j) + O(\rho_j^3) \ . \end{aligned} \tag{27}$$

The last equality holds by Theorem 1.6 (i). From the Taylor expansion of f at x^* we derive with $g_1 \in N$

$$\begin{aligned} f(x_j(\lambda)) &= \rho_j(1-\lambda)\nabla f(x^*)t_j + \rho_j^2 \lambda \nabla f(x^*) g_2(t_j) \\ &\quad + \frac{1}{2}\rho_j^2 \nabla^2 f(x^*)(\lambda g_1(t_j) + (1-\lambda)t_j)^2 + O(\rho_j^3) \ . \end{aligned}$$

Substituting

$$\rho_j t_j = \rho_{j-1} g_1(t_{j-1}) + \rho_{j-1}^2 g_2(t_{j-1}) + O(\rho_j^3)$$

and using

$$t_j + O(\rho_j) = t = t_{j-1} + O(\rho_j) \ ,$$

we obtain with (16) and $g_1(t) = t\, k/(k+1)$

$$\begin{aligned} f(x_j(\lambda)) &= \rho_j^2(\lambda + (1-\lambda)(k+1)^2/k^2)\nabla f(x^*) g_2(t) \\ &\quad + \frac{1}{2}\rho_j^2(1-\lambda/(k+1))^2 \nabla^2 f(x^*)t^2 + O(\rho_j^3) \ . \end{aligned}$$

The $(2+\Delta m)$-th "row" of the linear system (1.22) reads

$$\nabla^2 f(x^*)t g_1(t) + \nabla f(x^*) g_2(t) = \frac{1}{2}\nabla^2 f(x^*)t^2 \ , \tag{28}$$

which allows the elimination of $\nabla f(x^*) g_2(t)$ and gives after some elementary manipulations (26). With the elementary inequality $e > (1+1/k)^k$ assertion

(22) follows directly from Lemma 21(ii). The last equation (23) is a consequence of (1.9) and (2.4). ////

Lemma 3.2 shows that the regular Newton sequences converge to a regular singularity in a very structured way. We notice in particular that the residual $f(x_j)$ becomes colinear to $\nabla^2 f(x^*)t^2$ and its length declines linearly with a ratio of $k/(k+1)$, so that any given vector norm of f is reduced at each step unless $\nabla^2 f(x^*)t$ vanishes completely. Excluding the latter possibility we derive from (26) that the ratio between the actual gain $\|f(x_j)\| - \|f(x_{j+1})\|$ and the linearly expected reduction

$$- \frac{\partial}{\partial \lambda} \|f(x_j+\lambda(x_{j+1}-x_j))\| \Big|_{\lambda=0} = \|f(x_j)\|$$

is given by

$$1 - \|f(x_{j+1})\|/\|f(x_j)\| = \frac{2k+1}{(k+1)^2} + O(\rho_j) ,$$

where $\|\cdot\|$ may be any elliptic norm. Consequently the usual line search conditions of *stabilised* Newton methods (e.g. Goldstein test [23]) will always be met by the full Newton step during the final approach to a regular singularity. However it was found in [24] that such modifications can slow down the iteration considerably before the final pattern has been established.

If a Newton sequence does not converge superlinearly as usually expected the first noticeable sign is obviously that the ratio κ_j between consecutive stepsizes fails to become arbitrarily small. Provided the sequence converges at all the limiting point x^* must be a singular solution of x^* . Naturally it is important to determine the type of singularity by interpreting the unmodified Newton iterations before any convergence accelerating procedures may be applied.

Whereas by (19) at regular singularities

$$\kappa_j/(1-\kappa_j) \to k \ ,$$

this is not necessarily true in other cases. For instance we find for the system $f \equiv (\frac{1}{2}\xi^2, \frac{1}{3}\zeta^3)^T$ whose unique unbalanced singularity $x^* = 0$ is of order $k = 1$, that all Newton sequences satisfy asymptotically

$$\kappa_j/(1-\kappa_j) \to 2 \tag{29}$$

and

$$\omega_j \to 0 \ , \ f(x_j) = (0, \tfrac{1}{3}\rho_j^3)^T + o(\rho_j^3) \ , \tag{30}$$

$$\sigma(x_{j+1})/\sigma(x_j) \to 4/9 \ , \ \delta(x_{j+1})/\delta(x_j) \to 2/9 \ . \tag{31}$$

Whereas by (29) and (30) the singularity could be of second order and regular, the two limits in (31) cannot be matched with (22) and (23) for any k and m as $2/9$ is not an integral power of $4/9$.

Even though a proper determination of numerical rank [25] requires the singular value decomposition of the Jacobian, one can get some indication as to the dimension of the nullspace N from the LU decomposition and may use the smallest diagonal element in U as an estimate for σ . If the small elements in the diagonal of U decline at different rates or oscillate, the problem has most likely an irregular singularity to which the analysis of this thesis does not apply, even though extrapolation of the kind described in Section 3.4 would work for the unbalanced problem mentioned above. If on the other hand ω_j tends to zero, $\kappa_j/(1-\kappa_j)$ comes close to an integer k and σ and δ decline with rates that are compatible in the sense of (22), (23), then we can be reasonably sure to deal with a regular singularity for which the modifications developed in Sections 3.3 and 4.1 are designed.

3. Variation of the Stepsize at Regular Singularities

According to Lemma 3.2 regular Newton sequences approach x^* along a unique tangent $t \in S' \cap N$ roughly reducing the distance to x^* by a factor of $k/(k+1)$ at each step. Now it seems promising to accelerate the convergence by taking a step $(k+1)$ times the Newton correction or its projection into some approximation $\tilde{N}$ to the nullspace N. For the scalar case where necessarily $m=n=1$ this idea is rather old and has been shown to restore the quadratic rate of convergence of the unmodified method at nonsingular solutions by Schroeder [26] and several other authors. Unfortunately this situation is very atypical for the general multi-dimensional case in which Rall [3] and more recently Reddien [7] have discussed the properties of such *corrected* Newton steps. Rall's paper suggested that the multidimensional case could be treated successfully in essentially the same way as the scalar case. Unfortunately his analysis contains a flaw which amounts to the omission of certain cross terms and was first detected by Cavanagh. Reddien found in test calculations [7] that the corrected Newton step from some point $x_j = g(x_{j-1})$ leads usually to a point $x_{j+1}^{(1)}$ much closer to the solution x^* than $x_{j+1} = g(x_j)$ but that the subsequent Newton step from $x_{j+1}^{(1)}$ tended to be disadvantageous. In our framework this means that $x_{j+1}^{(1)}$ can lie outside W and may even belong to the singular set $\delta^{-1}(0)$. If $x_{j+1}^{(1)}$ is an element of R the next normal Newton step leads back into W, but it may be large enough to offset the original gain in the step from x_j to x_{j+1}. Whereas this situation seems typical in the general case, we find that at strongly regular first order singularities convergence of order $2^{1/3}$ can be obtained by taking two normal Newton steps after each corrected Newton step of double length. This result holds only if certain cubic terms do not vanish

and does therefore not apply to Reddien's test function which involves no cubic terms at all.

Considering a corrected Newton step from some point $x_j = x^* + \rho_j t_j \in W$ in the neighbourhood of a regular singularity with arbitrary k and m, we obtain from (27) with $\lambda = k+1$

$$x_{j+1}^{(1)} = x^* + \rho_j \left[(k+1) g_1(t_j) - k t_j \right] + (k+1) \rho_j^2 g_2(t_j) + O(\rho_j^3) \quad . \tag{32}$$

Now we look for conditions under which

$$\| x_{j+1}^{(1)} - x^* \| = O(\rho_j^2) \quad \text{and} \quad x_{j+1}^{(1)} \in W \ , \tag{33}$$

so that the next Newton step whether normal or corrected is well defined and does not increase the distance to x^*. We know from (2.12) that

$$(k+1) g_1(\bar{t}) - k\bar{t} = 0 \quad \text{for} \quad \bar{t} \in N \cap S \ ,$$

but $g_1(t)$ can be rather large if the minimal angle $\theta(t) = \cos^{-1}(t^T \bar{t})$ between t and some $\bar{t} \in N$ is not small. Imposing the condition $\sin \theta_j \equiv \sin \theta(t_j) = O(\rho_j)$ we find that the first requirement in (33) is satisfied and that the term

$$(k+1) \rho_j^2 g_2(t_j) = (k+1) \rho_j^2 g_2(\bar{t}_j) + O(\rho_j^3) \ , \ \bar{t}_j \in N$$

is now leading in (32). In order to show that $x_{j+1}^{(1)}$ belongs to W we have to bound the angle between $x_{j+1}^{(1)} - x^*$ and some regular direction in N, which seems only possible if

$$\pi_0(\bar{t}) \neq 0 \quad \text{and} \quad 0 \neq g_2(\bar{t}) \in N \quad \text{for all} \quad \bar{t} \in N \cap S \ . \tag{34}$$

In other words the singularity must be strongly regular and the then well defined vector $g_2(\bar{t})$ must be a nonzero element of N for all $\bar{t} \in N \cap S$. By (28) and (2.15) we have for $\bar{t} \in N \cap S$

$$\nabla f(x^*) g_2(\bar{t}) = \left(\frac{1}{2} - \frac{k}{k+1}\right) \nabla^2 f(x^*) \bar{t}^2 \ , \tag{35}$$

so that the condition (34) can only be met at a regular first order singularity, except for rather special cases where $\nabla^2 f(x^*)\bar{t}^2$ vanishes for all $\bar{t} \in N \cap S$. Excluding the latter possibility we must have $k=1$ and $\Delta m=0$ so that with (2.15) according to the second "row" of the linear system (2.6) for $t \in N \cap S'$

$$(\nabla f + P\nabla^2 f(x^*)t) g_2(t) = \frac{1}{2} P\nabla^3 f(x^*) t^3 \ . \tag{36}$$

Consequently we have to assume that the RHS does not vanish for any $t \in N \cap S$ in order to ensure (34), which leads to the following result.

THEOREM 3.3 *Second Order Three-Point Method*

Let $f \in C^{3,1}(\mathbb{R}^n, \mathbb{R}^n)$ have a strongly regular first order singularity at x^* . If

$$P\nabla^3 f t^3 \neq 0 \quad \text{for all} \quad t \in N \cap S \tag{37}$$

then there exists a constant $\bar{\rho}$ such that the three point iteration

$$y_{j+1} = 2g(g(g(y_j))) - g(g(y_j)) \tag{38}$$

converges Q-quadratically to x^* with

$$\theta\left(\frac{y_{j+1} - x^*}{\|y_{j+1} - x^*\|}\right) = O(\|y_j - x^*\|) \tag{39}$$

from all initial points in the starlike domain

$$\hat{V} \equiv \{x^* + \rho t \mid t \in S, \theta(t) < \hat{\theta}, 0 < \rho < \bar{\rho}\} \subseteq V \ ,$$

where V , $\theta(t)$ and $\hat{\theta}$ are defined as in Lemma 2.3.

Proof.

By definition of $\hat{\theta}$ in Lemma 2.3 we have

$$\min \{|\pi_0(t)| \mid t\in S, \theta(t)\le\hat{\theta}\} > 0$$

so that there are constants c_1, $c_2 \equiv c/\hat{\nu}$ and c_3 such that by Theorem 1.6 for all $x = x^*+\rho t \in V$

$$\begin{aligned} \|g(x)-x^*-\rho g_1(t)-\rho^2 g_2(t)\| &\le c_1\rho^3 , \\ \|2g_1(t)-t\| &\le c_2 \sin\theta(t) \end{aligned} \tag{40}$$

and

$$\|g_2(t)-g_2(\bar{t})\| \le c_3 \sin\theta(t) ,$$

where $\bar{t} \in N$ with $\cos\theta(t) = t^T\bar{t}$.

Because of the assumption (37) it follows from (36) that $g_2(\bar{t})$ cannot vanish for any $\bar{t} \in N \cap S$, so that there are constants c_4 and c_5 such that

$$0 < c_4 \le \|g_2(\bar{t})\| \le c_5 \quad \text{for all} \quad \bar{t} \in N \cap S . \tag{41}$$

Abbreviating

$$c_6 \equiv (c_2+2\hat{\rho}c_3)(c_3^2\sin\hat{\theta}+3c_1+\hat{\rho}c_3c_1) ,$$

we can now define

$$\bar{\rho} \equiv \min\left\{\hat{\rho} , \frac{c_4\sin\hat{\theta}}{16c_6} , \frac{1}{6c_5}\right\} . \tag{42}$$

According to equation (2.47) in the proof of Lemma 2.3 V is a domain of contraction so that for any

$$x_0 \equiv y_j \equiv x^*+\rho_0 t_0 \in \hat{V} \subseteq V$$

$$x_{i+1} = g(x_i) = x^*+\rho_i t_i \in \hat{V} , \frac{1}{4} \le \rho_{i+1}/\rho_i \le \frac{3}{4} \quad \text{for} \quad i = 0,1 .$$

According to Theorem 1.6 (iii) and equation (35) the vectors $g_1(t)$ and $g_2(\bar{t})$ always belong to N so that by (39)

$$\rho_1 \sin\theta_1 \le \rho_0^2 c_3 \sin\hat{\theta} + \rho_0^3 c_1 ,$$
$$\rho_2 \sin\theta_2 \le \rho_1^2 c_3 \sin\theta_1 + \rho_1^3 c_1$$
$$\le \rho_0^3 (c_3^2 \sin\hat{\theta} + \rho_0 c_3 c_1 + c_1) .$$

Now we obtain from (40) for the corrected step from x_2 to $x_3 \equiv y_{j+1} \equiv 2g(x_2) - x_2$ that

$$\|x_3 - x^* - 2\rho_2^2 g_2(\bar{t}_2)\|$$
$$\le \rho_2 c_2 \sin\theta_2 + 2\rho_2^2 c_3 \sin\theta_2 + 2\rho_2^3 c_1 \qquad (43)$$
$$\le \rho_0^3 (c_2 + 2\rho_0 c_3)(c_3^2 \sin\hat{\theta} + \rho_0 c_3 c_1 + c_1) + \rho_0^3 c_1 \le \rho_0^3 c_6 ,$$

where we have used $c_2 \ge 1$ to obtain the last inequality. Hence we find by (41) for $\rho_3 \equiv \|x_3 - x^*\|$

$$c_4 \rho_0^2/8 - \rho_0^3 c_6 \le \rho_3 \le 2c_5 \rho_0^2 + \rho_0^3 c_6 .$$

Since $\rho_0 < \bar{\rho}$ as defined by (42) we have

$$\rho_0 c_6 \le c_4/16 \le c_5/16 ,$$

so that

$$\rho_0 c_4/16 \le \rho_3/\rho_0 \le 3\rho_0 c_5 \le \frac{1}{2} ,$$

and furthermore by (43) with $g_2(\bar{t}_0) \in N$

$$\sin\theta_3 \le \rho_0^3 c_6 (\rho_0^2 c_4/16)^{-1} < \sin\hat{\theta} . \qquad (44)$$

Consequently $y_{j+1} = x_3$ belongs to $\hat{V}$ with

$$\|y_{j+1} - x^*\| \le 3c_5 \|y_j - x^*\|^2 \le \frac{1}{2} \|y_j - x^*\| ,$$

so that the sequence $\{y_j\}$ converges Q-quadratically to x^*. Equation (39) follows from the first inequality in (44). ////

It was found in practical calculations that the three-point method may still converge quite rapidly even if a regular first order singularity is not strongly regular. As we can see in Table 5 each fully corrected Newton step causes a shift of direction within N which can theoretically lead to a point $x^* + \rho t$ with t irregular. However since by assumption of regularity almost all directions in N are regular this is unlikely to occur, and as $g_1(t) = tk/(k+1)$ for all $t \in N \cap S'$ the next step can be favourable even if t is nearly irregular, i.e. $|\pi_0(t)|$ small.

The assumption that the singularity be of first order is essential because otherwise any fully corrected Newton step is likely to lead to a point outside W. This can be observed in Table 8 where only every fifth Newton step is corrected, which nevertheless destroys any prospect of convergence.

As we can see in Table 1 and Table 5 the two point method

$$y_{j+1} = 2g(g(y_j)) - g(y_j) \tag{45}$$

converges like the three point method quite rapidly to regular first order singularities. This observation could not be supported theoretically because the ratio between consecutive angles $\theta_j = \theta((y_j-x^*)/\|y_j-x^*\|)$ is bounded but not in general less than 1. However it can be shown on the basis of (39) that any combination of one three-point step (38) with $q-1$ two-point steps (45), or equivalently one normal Newton step with q two-point steps, yields a $(2q+1)$-point method which converges from within some starlike domain $\hat{R}_q \subseteq R$ with density 1. If the solution x^* is in fact nonsingular we have for each two-point step (45)

$$y_{j+1} - x^* = 2[g(g(y_j))-x^*] - (g(y_j)-x^*)$$

$$= 2\,O(\|g(y_j)-x^*\|^2) + O(\|y_j-x^*\|^2) = O(\|y_j-x^*\|^2)\ ,$$

so that the (2q+1)-point iteration does still converge provided it comes sufficiently close to x^*. The efficiency of the (2q+1)-point method in the sense of Brent [27] is given by

$$\frac{\log(\text{R-order})}{\text{evaluations of } f \text{ and } \nabla f} = \begin{cases} \frac{1+q}{1+2q} \log 2 & \text{if } \det(\nabla f(x^*)) \neq 0 \\ \\ \frac{q}{1+2q} \log 2 & \text{if } \det(\nabla f(x^*)) = 0 \ , \end{cases} \tag{46}$$

where x^* is assumed to be a strongly regular first order singularity in the second case. Here we have relaxed the usual definition of the R-order of an iterative process at a solution point x^* [10] to mean the minimal R-order of all iterations from within some starlike domain of density 1.

Now it would theoretically be the best strategy to start with the unmodified Newton iteration (q=0) until the convergence pattern described in Lemma 3.2 is observable, then to increase q gradually by taking more and more two-point steps and finally to revert to the unmodified method when the rounding errors become significant or the solution turns out to be only nearly singular. Unfortunately there is no simple criterion to decide whether any of the domains $\hat{R}_p$ has been reached and the working hypothesis that the singularity is strongly regular can never be verified. At each iteration point y_j we can calculate the angle ω_j between the Newton correction $g(y_j) - y_j$ and the previous step $y_j - y_{j-1}$. If ω_j is sufficiently small we may select a two-point step of the form (45) and otherwise a normal Newton step must be taken. The challenge to implement this kind of "line search" in a computer routine could not be met in this thesis.

On our test problem the three-point method (q=1) and the two-point method (q=∞) converge with similar speeds to a first order singularity with one or two dimensional nullspace [Tables 1,5] and a nearly singular solution

[Table 4]. In all three cases the three-point method exhibits a more regular convergence behaviour, with $\theta_j \to 0$ as ensured by (39), than the two-point method, which takes intermittently steps away from the solution. Even though this problem might be overcome by a judicious choice of $q < \infty$, it seems doubtful whether q should ever be raised above 1.

In the nearly singular case both multipoint methods are faster than Newton's method during the initial phase of the iteration, which is listed in Table 4. Once the residual $\|f\|$ is of the same magnitude as the smallest singular value of the Jacobian at the solution, Newton's method is naturally superior, so that a final switch back to $q = 0$ would be advantageous in both singular and nearly singular cases.

Whenever the assumption of Theorem 3.3 are not satisfied we face the dilemma that any fully corrected Newton step may lead to a point outside W, which was observed in Table 8 for a five point method at a strongly regular third order singularity. Several authors, e.g. Reddien [7] and Keller [28], suggested to determine from the singular value, eigenvalue, or simply some triangular decomposition of $\nabla f(x_j)$ an approximation $N(x)$ to N and then to project the Newton correction $g(x_j) - x_j$ into $N(x)$ before multiplying it by $k+1$. This idea is based on the observation that after several normal Newton steps $x_j - x^*$ belongs essentially to N, i.e. $\theta_j = \theta(t_j)$ is small. If this is so we can assume that

$$x_j = g(g(x_{j-2})) \quad \text{with} \quad x_{j-2} \in R ,$$

so that $x_{j-1} = g(x_{j-2})$ belongs to $\hat{W}(s)$ for some $s \in S' \cap N$. Then we have by (2.34) $\theta(t_j) = O(\rho_j)$ and because of (2.14) with (2.32)

$$\theta\left(\frac{g^i(x_j)-x^*}{\|g^i(x_j)-x^*\|}\right) = \begin{cases} O(\rho_j^2) & \text{if } k=1 \\ O(\rho_j) & \text{otherwise,} \end{cases} \quad \text{for } i=1,2 \ .$$

Using again (2.32) one can easily show that this implies

$$\theta\left(\frac{g(x_j)-x_j}{\|g(x_j)-x_0\|}\right) = O(\rho_j) \ , \tag{47}$$

and furthermore if $k=1$

$$\theta\left(\frac{g(x_{j+1})-x_{j+1}}{\|g(x_{j+1})-x_{j+1}\|}\right) = O(\rho_j^2) \tag{48}$$

where $x_{j+1} = g(x_j)$ as usual.

Hence the angle between the Newton correction evaluated at x_j and the nullspace N is $O(\rho_j)$ and the corresponding angle at the next iterate x_{j+1} is only $O(\rho_j^2)$ if $k=1$.

As will be shown in Section 4.2 the approximate nullspaces $N(x)$ derived from matrix decompositions of ∇f are spanned by vectors $\{v_j(x)\}_{j=1..m}$ which are Lipschitz continuously differentiable on some neighbourhood U of x^* provided this is true for ∇f . The ranges of the Jacobians $\{\nabla v_j(x^*)\}_{j=1..m}$ are in general not contained in $N = N(x^*)$ so that for some $j \in [1,m]$ and a suitably scaled vector $z \in \mathbb{R}^n$

$$s \equiv \nabla v_j(x^*)z \in S - N \ . \tag{49}$$

Without loss of generality we can assume that v_j is normalised such that

$$v_j(x) = v_j(x)/\|v_j(x)\| \in S \quad \text{for all } x \in U \ .$$

Since the columns of the matrix

$$V^* \equiv (v_1(x^*), v_2(x^*), \ldots v_m(x^*)) \in \mathbb{R}^{n\times m}$$

span N we find for the minimal angle θ_λ between $v_j(x^*+\lambda z)$ and N

$$\sin \theta_\lambda = \min_{y \in \mathbb{R}^m} \|V^*y - v_j(x^* + \lambda z)\|$$

$$= \min_{y \in \mathbb{R}^m} \|V^*y - v_j(x^*) - \lambda s\| - O(\lambda^2)$$

$$= \lambda \sin \theta(s) - O(\lambda^2) ,$$

where $\theta(s)$ is the minimal angle between s and N which is by (49) nonzero. Hence we conclude that the angle

$$\bar{\theta}(x) \equiv \max_{t \in S \cap N(x)} \theta(t) = \max_{t \in S \cap N(x)} \min_{s \in S \cap N} \cos^{-1}(t^T s)$$

is by differentiability of the $\{v_j\}_{j=1..m}$ of $O(\rho)$ but in general not smaller.

Consequently the projection of the Newton correction $g(x_j) - x_j$ into $N(x_j)$ is by comparison with (47) unlikely to reduce the angle with N significantly if the singularity is regular. If it is furthermore of first order the angle between N and the Newton correction at the next iterate x_{j+1} is by (48) much smaller than what we can possibly ensure by any kind of projection. Rather than expending any computing time for the approximation of N by $N(x)$ we prefer to take one or more normal Newton steps between any two corrected Newton steps, which was already shown to be successful in the case of regular singularities with $k = 1$. The unmodified Newton iteration functions approximately as a power method for the calculation of the eigenvectors of the homogeneous vector function $g_1 : \mathbb{R}^n - \pi_0^{-1}(0) \to N$. Since by (2.15)

$$g_1(t) = \lambda t \text{ for } t \in S' , \lambda \in \mathbb{R} \iff t \in N \cap S' , \lambda = k/(k+1) ,$$

this process which is based on values of ∇f and f generates Newton corrections which are at least as "close" to N as any approximation $N(x)$ that is based on the current Jacobian alone.

In the case of higher order singularities we have already noted that a fully corrected Newton step can never be taken even after arbitrarily many normal steps. Thus we consider partially corrected Newton steps of the form

$$x_{j+1} = x_j - \lambda_j \nabla f^{-1}(x_j) f(x_j)$$

$$= x^* + \rho_j \begin{bmatrix} (1 - \frac{\lambda_j}{k+1})\ I\ , & -\frac{k\lambda_j}{k+1} \bar{B}^{-1}(t_j)\bar{C}^T(t_j) \\ 0\ , & (1-\lambda_j)\ I \end{bmatrix} t_j + O(\rho_j^2)$$

with $\lambda_j \in (1,k+1)$, so that the leading term on the RHS has a nonzero component in N provided this is true for t_j. Abbreviating

$$\Delta\lambda_j \equiv (1 - \frac{\lambda_j}{k+1}) \in (0, \frac{k}{k+1}) \tag{50}$$

we derive from Lemma 2.1 (iii) the inequalities

$$\sin \theta_{j+1} \leq \left[(\lambda_j - 1) \sin \theta_j + \frac{\tilde{d}}{\nu_j^2} \rho_j\right] \rho_j / \rho_{j+1} \tag{51}$$

and

$$\|x_{j+1} - x^* - \Delta\lambda_j (x_j - x^*)\| \leq \eta_j \rho_j \equiv \left[\frac{k\lambda_j c}{(k+1)\nu_j} \sin \theta_j + \frac{\tilde{d}}{\nu_j^2} \rho_j\right] \rho_j\ , \tag{52}$$

where $\nu_j \equiv \nu(t_j)$ as before *and* $\tilde{d} \equiv d(k+1)$.

As immediate consequences of (52) we obtain

$$|\rho_{j+1}/\rho_j - \Delta\lambda_j| \leq \eta_j \tag{53}$$

and for the angle $\Delta\psi_j$ between t_j and t_{j+1}

$$\sin \Delta\psi_j \leq \eta_j / \Delta\lambda_j\ . \tag{54}$$

Now suppose we want to choose $\lambda_j = \lambda$ for some constant $\lambda > 1$. Then it follows from (51) and (53) that the size of the θ_j can only be controlled if

$$\frac{|\lambda-1|}{\Delta\lambda} = \frac{\lambda-1}{1-\lambda/(k+1)} < 1 \ . \tag{55}$$

This condition is sufficient for the existence of starlike domains of convergence as constructed in the following theorem.

THEOREM 3.4 *Partially Corrected One Point Method*

Let $f \in C^{k+1,1}(\mathbb{R}^n,\mathbb{R}^n)$ have a regular singularity of order k at $x^* \in f^{-1}(0)$. Then there exist for any fixed $\lambda \in (1,1+k/(k+2))$ and all regular directions $s \in N \cap S'$, two positive constants $\hat{\phi}_\lambda(s)$ and $\hat{\rho}_\lambda(s)$ such that the partially corrected Newton iteration

$$x_{j+1} = x_j - \lambda \nabla f^{-1}(x_j) f(x_j)$$

converges to x^* from all initial points in the starlike domain

$$\hat{W}_\lambda(s) \equiv \{x^*+\rho t \mid t\in S \ , \cos^{-1}(t^T s) < \hat{\phi}_\lambda(s) \ , \ 0<\rho<\hat{\rho}_\lambda(s)\}$$

and satisfies asymptotically

$$\frac{\rho_{j+1}}{\rho_j} \to (1-\frac{\lambda}{k+1}) > \frac{k}{k+2} \quad \text{and} \quad t_j \to t \in S' \cap N \ .$$

Proof.

The proof of this result is omitted because it is based on the same idea as the proofs of Lemma 2.2 and the next Theorem 3.5. ////

It seems doubtful whether the partially corrected one-point method considered in Theorem 3.4 represents a real improvement over the unmodified iteration. The simplicity and structure of the latter is lost and the reduction of the linear Q-factor from $k/k+1$ to $\Delta\lambda > k/(k+2)$ is only a small gain especially for $k \geq 3$. It was found in practical calculations [Tables 5,8] that the partially corrected one-point method, though faster than Newton's method, was not competitive with other modifications.

Instead of imposing the condition (55) we can control the size of the θ_j by taking a normal Newton step after each partially corrected step, which leads to the following result.

THEOREM 3.5 *Partially Corrected Two-Point Method*

Let $f \in C^{k+1,1}(\mathbb{R}^n,\mathbb{R}^n)$ have a regular singularity of order k at $x^* \in f^{-1}(0)$. Then there are, for any fixed multiplier $\lambda \in (1,k+1)$ and all regular directions s in N, two positive constants $\tilde{\phi}_\lambda(s)$ and $\tilde{\rho}_\lambda(s)$ such that the two-point iteration

$$y_{i+1} = \lambda\, g(g(y_i)) - (\lambda-1)g(y_i)$$

converges to x^* from all initial points in the starlike domain

$$\tilde{W}_\lambda(s) \equiv \{x^*+\rho t \mid t\in S\,,\ \cos^{-1}(s^T t) < \tilde{\phi}_\lambda(s)\ ,\ 0<\rho<\tilde{\rho}_\lambda(s)\}\ ,$$

with

$$\frac{\|y_{i+1}-x^*\|}{\|y_i-x^*\|} \le 3\left(1-\frac{\lambda}{k+1}\right) \quad \text{for all} \quad i \ge 0 \tag{56}$$

and in the limit

$$\frac{\|y_{i+1}-x^*\|}{\|y_i-x^*\|} \to \frac{k}{(k+1)}\left(1-\frac{\lambda}{k+1}\right)\ , \tag{57}$$

$$\frac{y_i-x^*}{\|y_i-x^*\|} \to t \in N \cap S'\ . \tag{58}$$

Proof.

Including the intermediate points $g(y_j)$ we obtain the sequence $\{x_j\}_{j\ge 0}$ with

$$x_{2i} \equiv y_i \quad \text{and} \quad x_{2i+1} \equiv g(y_i) \quad \text{for} \quad i \ge 0\ .$$

Let ψ_j and the constants ϕ, $\hat{\nu}$ and $\hat{r}$ be defined for fixed $s \in N \cap S'$ as in Lemma 2.2. Our aim is to choose $\tilde{\phi}_\lambda < \phi$ and $\tilde{\rho}_\lambda < \hat{r}$ such that

for all j

$$\rho_j < \tilde{\rho}_\lambda \ , \ \theta_j < \tilde{\phi}_\lambda \quad \text{and} \quad \psi_j < \phi \tag{59}$$

whenever $\rho_0 < \tilde{\rho}_\lambda$ and $\psi_1 = \cos^{-1}(s^T t_0) < \tilde{\phi}_\lambda$.

To this end we impose several conditions on $\tilde{\phi}_\lambda$ and $\tilde{\rho}_\lambda$ the first of which is the inequality

$$\frac{kc}{\hat{\nu}} \sin \tilde{\phi}_\lambda + \frac{\tilde{d}}{\hat{\nu}^2} \tilde{\rho}_\lambda < \hat{\eta} \equiv \frac{1}{2} \min\left\{ \frac{1}{k+1} , \Delta\lambda \right\} \tag{60}$$

where $\Delta\lambda$ is given by (50) .

It follows from (53) for the normal Newton step $(\lambda_{2i} = 1)$ from some point $x_j = x_{2i}$ satisfying (59) that by definition of $\hat{\eta}$ in (60)

$$\frac{1}{4} \le \frac{k}{k+1} - \hat{\eta} \le \frac{\rho_{2i+1}}{\rho_{2i}} \le \frac{k}{k+1} + \hat{\eta} \le \min\left\{ \frac{k+\frac{1}{2}}{k+1} , \frac{k}{k+1} + \frac{1}{2} \Delta\lambda \right\} , \tag{61}$$

and consequently by (51)

$$\sin \theta_{2i+1} \le 4\tilde{d} \ \rho_{2i} / \hat{\nu}^2 \ . \tag{62}$$

Provided $\theta_{2i+1} \le \tilde{\phi}_\lambda$ and $\psi_{2i+1} < \phi$ which will be ensured later, we obtain from (53) for the partially corrected Newton step to x_{2i+2}

$$\frac{1}{2} \Delta\lambda \le \lambda - \hat{\eta} \le \frac{\rho_{2i+2}}{\rho_{2i+1}} \le \hat{\eta} + \Delta\lambda \le \min\left\{ \frac{k+\frac{1}{2}}{k+1} , \frac{3}{2} \Delta\lambda \right\} , \tag{63}$$

so that by (51) with $\lambda_{2i+1} = \lambda$ and (62)

$$\sin \theta_{2i+2} \le \frac{d}{\hat{\nu}^2} (4k\rho_{2i} + \rho_{2i+1}) \frac{2}{\Delta\lambda} \le \frac{10k\tilde{d}}{\hat{\nu}^2 \Delta\lambda} \rho_{2i} \ . \tag{64}$$

Imposing the condition

$$\tilde{\rho}_\lambda < \frac{\hat{\nu}^2 \Delta\lambda}{10k\tilde{d}} \sin \tilde{\phi}_\lambda \tag{65}$$

we can ensure that θ_{2i+1} and θ_{2i+2} are less than $\tilde{\phi}_\lambda$ provided (59) holds for $j = 2i$. In order to bound ψ_j we note that by (54) with η_j

as implicitly defined in (52).

$$\sum_{j=0}^{2i} \sin \Delta\psi_j \le \sum_{\ell=0}^{i} (\eta_{2\ell}+\eta_{2\ell+1}/\Delta\lambda) \ . \tag{66}$$

As long as (59) holds we know from (62), (63) and (64) that for some constant $\tilde{c} > 0$

$$\eta_{2\ell+2} + \eta_{2\ell+1}/\Delta\lambda \le \tilde{c}\rho_{2\ell} \ , \tag{67}$$

which gives by (61), (63) and (54)

$$\begin{aligned} \sin \psi_j &\le \sin \psi_0 + \sin \Delta\psi_0 + 2\tilde{c}\rho_0(k+1)^2 \\ &\le \left(1+\frac{kc}{\hat{\nu}}\right) \sin \tilde{\phi}_\lambda + \left(\frac{\tilde{d}}{\hat{\nu}^2}+2\tilde{c}(k+1)\right)\tilde{\rho}_\lambda . \end{aligned}$$

By a suitable choice of first $\tilde{\phi}_\lambda < \phi$ and then $\tilde{\rho}_\lambda < \hat{r}$ we can ensure that the RHS is less than $\sin \phi$ and that the conditions (60) and (65) are simultaneously satisfied. Then (59) must hold for all $j \ge 0$ so that by (61) and (63)

$$\begin{aligned} \frac{\rho_{2i+2}}{\rho_{2i}} &\le \min\left\{\frac{k+\frac{1}{2}}{k+1} \ , \ \frac{k}{k+1} + \frac{1}{2}\Delta\lambda\right\} \min\left\{\frac{k+\frac{1}{2}}{k+1} \ , \ \frac{3}{2}\Delta\lambda\right\} \\ &= \min\left\{\left(\frac{k+\frac{1}{2}}{k+1}\right)^2 \ , \ 3\Delta\lambda\left(\frac{\frac{1}{2}k}{k+1}+\frac{1}{4}\Delta\lambda\right)\right\} \end{aligned}$$

which implies (56) as $\Delta\lambda \le 1$. Therefore the ρ_j and consequently by (62) and (64) the η_j decline Q-linearly so that by (53) with $\lambda_{2i} = 1$ and $\lambda_{2i+1} = \lambda$

$$\frac{\rho_{2i+2}}{\rho_{2i}} = \left(\frac{k}{k+1}+O(\rho_{2i})\right)\left(\Delta\lambda+O(\rho_{2i})\right) = \frac{k\Delta\lambda}{k+1} + O(\rho_{2i})$$

which proves (57). The $t_j \in S$ form by (66) and (67) a Cauchy sequence whose limit t is because of (59), (62) and (64) a regular direction in N .

////

As an immediate consequence of Theorem 3.5 we obtain the following corollary.

COROLLARY 3.6 *Q-Superlinearly Converging Iteration*

Let $f \in C^{k+1,1}(\mathbb{R}^n,\mathbb{R}^n)$ have a regular singularity of order k at x^*. Then there exists for every initial point y_0 in $\mathcal{R}$ as defined in Theorem 2.4, a nondecreasing sequence of multipliers

$$\lambda_j \to k+1$$

such that the iteration

$$y_{j+1} = \lambda_j g(g(y_j)) - (\lambda_j - 1) g(y_j) \tag{68}$$

converges Q-superlinearly to x^* in that

$$\|y_{j+1} - x^*\| / \|y_j - x^*\| \to 0 .$$

Proof.

Initially we may choose constantly $\lambda_j = 1$. According to (58) the effectively unmodified Newton iteration approaches x^* along a unique, regular tangent, say $\hat{t}_1 \in N$. After finitely many steps the process must reach a point in $\tilde{W}_{k+\frac{1}{2}}(\hat{t}_1)$ as defined in Theorem 3.5. Then we can change over to the two point iteration with $\lambda_j \equiv k + \frac{1}{2}$. The new sequence has again by (58) a regular tangent, say $\hat{t}_2$ and must reach after finitely many steps the domain $\tilde{W}_{k+3/4}(\hat{t}_2)$. Then we may reset λ_j to $k + \frac{3}{4}$ and repeat the readjustments such that in the limit $\lambda_j \to k+1$ and consequently $\Delta\lambda_j \to 0$, which ensures Q-superlinear convergence by (56). ////

If the solution x^* is in fact nonsingular we have

$$y_{j+1} - x^* = \lambda_j (g(g(y_j)) - x^*) - (\lambda_j - 1)(g(y_j) - x^*) = O(\|y_j - x^*\|^2) ,$$

so that the iteration converges for any multiplier sequence

$\{\lambda_j\}_{j\geq 0} \subset (1,k+1)$ Q-quadratically to x^*, provided it comes sufficiently close to x^* .

As in the case of the (2q+1)-point method discussed at the beginning of this section, we again face the problem that no simple criterion is available to decide whether any of the starlike domains $\hat{W}_\lambda(s)$ has been reached. If λ_j is increased too rapidly the iteration may not converge at all and if it is increased too conservatively the convergence will be initially slow and when λ_j comes finally close to $k+1$ the theoretical benefit might be completely foiled by rounding errors. As we can see in Table 8 for the case $k=2$ and $m=1$ the partially corrected two-point method with fixed $\lambda = 2.8$ converges quite nicely with Q-factor ≈ 0.044, once the convergence pattern has been established. Since none of the other methods discussed in this thesis except extrapolation, is applicable at higher order singularities, the development of practical criteria for the choice of the λ_j in a two-point iteration of the form (68) would be a considerable achievement.

Unless the singularity x^* is pure (n=m) fast convergence of a sequence $y_j \to x^*$ implies by Lemma 2.1 (ii) rapid deterioration of the conditioning of the Jacobians $\{\nabla f(y_j)\}$. The partially or fully corrected Newton steps are only advantageous as long as they can be calculated with a high relative accurace since otherwise they may lead to a point outside W . Therefore the variations of stepsize discussed in this Section should mainly be applied during the intermediate stages of an iteration especially if the singularity is not strongly regular.

CHAPTER 4

EXTRAPOLATION AND BORDERING

1. Extrapolation at Regular Singularities

According to equation (3.22) in Lemma 3.2 the conditioning of the Jacobian at a regular Newton sequence deteriorates by a factor between $\frac{1}{e}$ and $\frac{1}{2}$ at each step no matter how high the order of the singularity. This "cautious" approach to the singularity should enable the unmodified method to "squeeze" the maximal accuracy out of the routines for the evaluation of f, the Jacobian ∇f and the subsequent solution of a linear system in ∇f. We know from Lemma 2.3 and Lemma 3.2 that the convergence of regular Newton sequences is reasonably robust and very structured, so that it seems promising to extrapolate the location of x^* without abandoning the unmodified Newton iteration.

For the scalar case $n = 1$ several authors, e.g. Ostrowski [29] and King [30] have developed extrapolation procedures to speed up the convergence of Newton's method to both singular and nonsingular solutions. Like most acceleration techniques for slowly converging scalar sequences [31] (e.g. Aitken's δ^2-process [1] or the ε-algorithm [32]) these methods involve divisions by function value differences or derivatives. Therefore they are not directly applicable to vector sequences and could be computationally expensive if division by derivatives would generalise to multiplication by inverse Jacobians. Another feature of these methods is that the extrapolated point serves as initial point for a new cycle of the respective scheme, as for instance in King's fourth order three-point method. In contrast we will never actually "take" the step to the extrapolated point which, though probably a good estimate for x^*, is of dubious value as a starting

point for subsequent steps.

Now let $\{x_j\}_{j\geq 0}$ be a regular Newton sequence in the sense of Lemma 3.2. Abbreviating $\kappa \equiv k/(k+1)$ we obtain in agreement with (3.32) as first stage of the extrapolation the sequence

$$x_{j+1}^{(1)} = \frac{x_{j+1} - \kappa x_j}{1 - \kappa} = x^* + O(\rho_i^2) \quad \text{for} \quad j \geq 0 \,. \tag{1}$$

Substituting κ by the approximation

$$\kappa_j \cos \omega_j = \frac{(x_{j+1}-x_j)^T (x_j - x_{j-1})}{\|x_j - x_{j-1}\|^2} = \kappa + O(\rho_j)$$

with κ_j and ω_j as defined in Lemma 3.2 we obtain the formula

$$\tilde{x}_{j+1}^{(1)} = \frac{\|x_j - x_{j-1}\|^2 x_{j+1} - (x_{j+1}-x_j)^T (x_j - x_{j-1}) x_{j-1}}{\|x_j - x_{j-1}\|^2 - (x_{j+1}-x_j)^T (x_j - x_{j-1})} \,,$$

which reduces in the scalar case $n=1$ to Aitken's δ^2-process. In what follows the form (1) will be preferred as it allows the interpretation as Richardson's deferred approach to the limit applied to an assumed expansion

$$x_j = x^* + v_1 \kappa^j + v_2 \kappa^{2j} + \ldots \,, \quad \text{for} \quad j \geq 0 \,. \tag{2}$$

Since κ is always positive we can write $\kappa^j = h_j^2$ so that (2) looks exactly like the h^2-error expansion of a central difference scheme for the solution of differential equations [33]. It should be noted that in contrast to this classical case the expansion (1), if it exists at all, depends not only on the problem as such but also on the particular Newton sequence, so that the vectors v_ℓ are in fact functions of the initial point x_0. Truncating (2) after q terms, we define the extrapolants $x_j^{(q)}$ as unique solutions of the linear system

$$x_{j-\ell} = x_j^{(q)} + \hat{v}_1 \kappa^{q-\ell} + \hat{v}_2 \kappa^{2(q-\ell)} \ldots \hat{v}_q \kappa^{q(q-\ell)} \quad \text{for} \quad \ell=0,\ldots q \ ,$$

which implies in particular $x_j^{(0)} = x_j$ for all $j \geq 0$. Eliminating the unknowns $\hat{v}_\ell$, we derive from Lagrange's extrapolation formula

$$x_j^{(q)} = \sum_{\ell=0}^{q} x_{j-\ell} \frac{(-1)^\ell \kappa^{\frac{1}{2}(\ell+1)\ell}}{\prod_{i=1}^{\ell} (1-\kappa^i) \prod_{i=1}^{q-\ell} (1-\kappa^i)} \ . \tag{3}$$

The $x_j^{(q)}$ can be calculated recursively by linear iterative extrapolation of the form

$$x_j^{(q+1)} = \frac{x_j^{(q)} - \kappa^{q+1} x_{j-1}^{(q)}}{1 - \kappa^{q+1}} \ , \tag{4}$$

which is a special case of a formula given by Bulirsch [34] and effectively eliminates the leading $\kappa^{(q+1)j}$ term in the expansion of $x_j^{(q)}$. As we will see later this is only true if $x_j^{(q)}$ is considered as a function of x_{j-2q} rather than x_{j-q}, so that we have the extrapolation triangle

$$\begin{array}{lllll}
x_0 & & & & \\
x_1 & \ddots & & & \\
x_2 & , & x_2^{(1)} & & \\
x_3 & , & x_3^{(1)} & \ddots & \\
\vdots & & \vdots & \ddots & \ddots \\
x_{2j} & , & x_{2j}^{(1)} & , \ x_{2j}^{(2)} & , \ldots x_{2j}^{(j)}
\end{array} \tag{5}$$

first column : regular Newton sequence

(1+q)-th column : q-th stage extrapolants

The extrapolants have been indexed such that the subscripts indicate the number of function and Jacobian evaluations required for their calculation. In practice the extrapolation should only be started when the unmodified Newton iteration exhibits the convergence pattern described in Lemma 3.2.

Without deciding the question whether (2) exists or not we show that

$$x_j^{(q)} - x^* = O\left(\rho_{j-2q}^{q+1}\right) = O\left(\kappa^{j(1+q)}\right) ,$$

provided f is sufficiently often differentiable.

In order to establish this result we consider for some fixed regular direction $s \in S' \cap N$ the starlike domains $\bar{W} \equiv \bar{W}(s)$ and $\hat{W} \equiv \hat{W}(s) \subset \bar{W}$ as defined in Lemma 2.2. Let

$$\bar{U} \equiv \{t \in S \mid \cos^{-1}(t^T s) < \phi(s) , \theta(t) < \hat{\phi}(s)\}$$

and

$$\hat{U} \equiv \{t \in S \mid \cos^{-1}(t^T s) < \hat{\phi}(s)\} \subset \bar{U}$$

be the sets of those directions that are included in $\bar{W}$ and $\hat{W}$ respectively For any $t \in \bar{U}$ we derive from (2.12) with (2.28)

$$\|g_1(t) - \kappa t\| \leq \kappa \subset \hat{\nu}^{-1} \sin \hat{\phi} \leq \chi/(k+1) ,$$

so that

$$\|g_1(t)\| \geq (k-\chi)/(k+1)$$

and

$$\sin[\cos^{-1}(t^T g_1(t)/\|g_1(t)\|)] \leq \chi = \frac{1}{4} \sin \phi .$$

Thus we have by the triangular inequality in S

$$\cos^{-1}(s^T g_1(t)/\|g_1(t)\|) \leq 5\ \phi/4 ,$$

which implies by definition of ϕ in (2.25) that

$$\alpha \equiv \inf\{|\pi(t)| \mid t \in \bar{U}\} > 0 , \tag{6}$$

where π is the homogeneous polynomial of degree $p(p+2)$ defined in Theorem 2.4 (ii). Abbreviating $\hat{k} \equiv \Delta p - \Delta m$ we obtain from (1.21) for all $x \equiv x^* + \rho t \equiv x^* + z \in \bar{W}$

$$g(x) = x^* + \sum_{i=1}^{\hat{k}} \frac{u_i(z)[\pi_0(\pi_0(z)g_1(z))]^{i+\Delta m}}{\pi(z)^{i+\Delta m}} + O(\rho^{\hat{k}+1}) \tag{7}$$

where the remainder on the RHS is by (6) uniform in $t \in \bar{U}$. By (2.5) and (2.12) the components of the vector function $\pi_0(z)g_1(z)$ are homogeneous polynomials in z, so that the components of each term in the expansion (7) are homogeneous rational functions of the form

$$\eta/\pi^{\ell} : \mathbb{R}^n - \pi^{-1}(0) \to \mathbb{R} \ . \tag{8}$$

For each integer i let H_i be the set of all scalar functions of the form (8) with $\ell \in \mathbb{N}$ and $\eta \in C^\infty(\mathbb{R}^n)$ a homogeneous polynomial of degree $i + p(p+2)\ell$. Then we have for any nonzero $y \in \mathbb{R}^n$ and $\lambda \in \mathbb{R} - \{0\}$

$$\eta(\lambda y)/\pi(\lambda y)^{\ell} = \lambda^i \eta(y)/\pi(y) \ ,$$

so that all elements of H_i have the degree of homogeneity i. It can be easily seen that

$$h, \tilde{h} \in H_i \ , \ \lambda \in \mathbb{R} \ \Rightarrow \ h + \lambda\tilde{h} \in H_i \ , \tag{9}$$

$$h \in H_i \ , \ \tilde{h} \in H_{\tilde{i}} \quad \Rightarrow \ h \cdot \tilde{h} \in H_{i+\tilde{i}} \tag{10}$$

and

$$i \neq \tilde{i} \qquad \Leftrightarrow \ H_{\tilde{i}} \cap H_i = \{0\} \ .$$

Therefore the sets H_i form linear subspaces of $C^\infty(\mathbb{R}^n - \pi^{-1}(0))$ and their direct sum

$$H \equiv \bigcup_{i=1}^{\infty} \{H_{-i} + H_{1-i} \ldots + H_0 + \ldots H_{i-1} + H_i\}$$

is a subalgebra of $C^\infty(\mathbb{R}^n - \pi^{-1}(0))$, i.e. H contains all sums and products of its elements. H consists of all rational functions of the form

$$h = \tau/\pi^{\ell} : \mathbb{R}^n - \pi^{-1}(0) \to \mathbb{R} ,$$

where $\ell \in \mathbb{N}$ and $\tau \in C^{\infty}(\mathbb{R}^n)$ any polynomial. Ordering the terms in τ according to their degree we obtain the unique decomposition

$$h = (\eta_0+\eta_1+\ldots\eta_q)/\pi^{\ell}$$

$$= \sum_{i=i_0}^{i_0+q} h_i \ , \ h_i \in H_i \ , \ h_{i_0} \neq 0 \neq h_{i_0+q} \ , \tag{11}$$

where the $\{\eta_j\}_{j\in[0,q]}$ are homogeneous polynomials of degree $i_0 + (p+2)p\ell + j$. We are mainly interested in vector functions

$$h \in H^n \equiv \underbrace{H \times H \times \ldots \times H}_{n}$$

for which the decomposition (11) exists with $h_i \in H_i^n$. The smallest index i_0 for which $h_{i_0} \neq 0$ will be called the *order* ord(h) of h . For our purposes the following properties of the elements in H^n are important.

LEMMA 4.1 *Polynomials over Powers of* π

Let $\bar{U}$, $\bar{W}$, $\{H_i^n\}$ and H^n be the sets defined above and $g_1 \in H_1^n$ the leading term in the expansion (7) of the Newtonian iteration function g . Then

(i) For any $h \in H_i$ and $j \in \mathbb{N}$ the entries of the derivative tensor $\nabla^j h$ belong to H_{i-j} .

(ii) The restriction of any $h \in H^n$ to $\bar{U}$ is bounded so that for all $x^* + z = x^* + \rho t \in \bar{W}$

$$h(z) = O(\rho^{ord(h)}) .$$

(iii) For any $h \in H_i$ the composition $h \circ g_1$ belongs also to H_i .

(iv) For any vector function $h \in H^n$ with $i_0 \equiv \text{ord}(h) > 0$ there is a vector function $\hat{h} \in H^n$ with

$$\text{ord}(\hat{h}) \geq \text{ord}(h) + 1$$

such that for all $x^* + \rho t = x^* + z \in \hat{W}$

$$h(g(x)-x^*) = h_{i_0}(g_1(z)) + \hat{h}(z) + O(\rho^{\hat{k}+1}) \quad , \tag{12}$$

where the remainder on the RHS is uniform in $t \in \hat{U}$.

Proof.

(i) The partial derivative of $h = \eta/\pi^{\ell}$ with respect to some variable ξ is given by

$$\frac{\partial h}{\partial \xi} = \left(\pi \frac{\partial \eta}{\partial \xi} - \ell\eta \frac{\partial \pi}{\partial \xi}\right) / \pi^{\ell+1} \quad .$$

Each term in the denominator polynomial has the same degree

$$\deg\left(\pi \frac{\partial \eta}{\partial \xi} - \ell\eta \frac{\partial \pi}{\partial \xi}\right) = \deg(\eta) + \deg(\pi) - 1$$

so that $\partial h/\partial \xi$ is homogeneous of degree

$$\deg(\eta) - \ell \deg(\pi) - 1 = \deg(h) - 1 \ .$$

Thus each component of the gradient ∇h belongs to H_{i-1} and we can obtain assertions (i) by induction on j .

(ii) Without loss of generality we can assume $h = \eta/\pi^{\ell} \in H_{i_0}$ so that by (6)

$$|h(z)| = |\eta(z)/\pi(z)^{\ell}| = \rho^{i_0}|\eta(t)/\pi(t)^{\ell}|$$

$$\leq \rho^{i_0}\alpha^{-\ell}\max\{|\eta(t)| \cdot \Big| \ t \in \bar{U}\} \ .$$

(iii) For all $z \in \mathbb{R}^n$ we have with $h = \eta/\pi^{\ell}$

$$h \circ g_1(z) = h(g_1(z)) = h(\pi_0(z)g_1(z))/\pi_0(z)^i$$

$$= \frac{\eta(\pi_0(z)g_1(z))}{\pi_0(z)^i \pi(\pi_0(z)g_1(z))^{\ell}} \quad . \tag{13}$$

Furthermore by definition of π

$$\pi_0(\pi_0(z)g_1(z)) = \pi(z)/\pi_0(z) \tag{14}$$

and because of (2.15)

$$g_1(\pi_0(z)g_1(z)) = \kappa\pi_0(z)g_1(z) \ ,$$

so that

$$\pi(\pi_0(z)g_1(z)) = (\pi(z)/\pi_0(z))\pi_0((\pi(z)/\pi_0(z))\kappa\pi_0(z)g_1(z))$$

$$= \kappa^p \pi(z)^{p+1} \pi_0(g_1(z))/\pi_0(z)$$

$$= \kappa^p \pi(z)^{p+1} \pi_0(\pi_0(z)g_1(z))/\pi_0(z)^{p+1}$$

$$= \kappa^p [\pi(z)/\pi_0(z)]^{p+2} \quad .$$

Thus we can rewrite (13) as

$$h(g_1(z)) = \frac{\eta(\pi_0(z)g_1(z))/\kappa^{p\ell}}{\pi(z)^{\ell(p+2)} \pi_0(z)^{i-\ell(p+2)}} \quad .$$

If $i \le \ell(p+2)$ then $h \circ g_1$ is already in the form (8). Otherwise we multiply both denominator and numerator by $\pi_0(\pi_0(z)g_1(z))^{i-\ell(p+2)}$ which makes the denominator by (14) a power of π. Hence $h \circ g_1$ belongs to H and since for all nonzero $z \in \mathbb{R}^n$ and $\lambda \in \mathbb{R} - \{0\}$

$$h(g_1(\lambda z)) = h(\lambda g_1(z)) = \lambda^i h(g_1(z))$$

the composition $h \circ g_1$ must be an element of H_i .

(iv) Because of (2.32) we have $g(x) - x^* = O(\rho)$ which allows us to ignore the higher order terms of h, so that without loss of generality

$$h = \sum_{i=i_0}^{\hat{k}} h_i \quad , \quad h_i \in H_i^n \quad , \quad h_{i_0} \neq 0 \ .$$

For each $i \in [i_0, \hat{k}]$ we have the Taylor expansion

$$\begin{aligned} h_i(g(x)-x^*) &= h_i(g_1(z)) + \sum_{j=1}^{\hat{k}-i} \frac{1}{j!} \nabla^j h_i(g_1(z))(g(x)-x^*-g_1(z))^j \\ &+ \frac{1}{(\hat{k}-i+1)!} \nabla^{\hat{k}-i+1} h_i(y_i)(g(x)-x^*-g_1(z))^{\hat{k}-i+1} \ , \end{aligned} \tag{15}$$

where for some mean value $\alpha_i \in (0,1)$

$$y_i = \alpha_i(g(x)-x^*) + (1-\alpha_i)g_1(z) \ .$$

Because of (i), (iii), (10) and $g_q \in H_q^n$ the vector functions

$$w_{ij}(z) \equiv \nabla^j h_i(g_1(z)) \left(\sum_{q=2}^{\hat{k}} g_q(z) \right)^j$$

belong to H^n and have the order

$$\mathrm{ord}(w_{ij}) \geq i - j + 2j = i + j \ .$$

Because of (7) and (ii) the lowest order term in the discrepancy between $w_{ij}(z)$ and

$$\nabla^j h_i(g_1(z))(g(x)-x^*-g_1(z))^j$$

is given by

$$\begin{aligned} &\nabla^j h_i(g_1(z))(g(x)-x^*-g_1(z))^{j-1} O(\rho^{\hat{k}+1}) \\ &= \rho^{i-j+2(j-1)} h_i(g_1(t))(g_2(t)+O(\rho))^{j-1} O(\rho^{\hat{k}+1}) = O(\rho^{\hat{k}+1}) \end{aligned}$$

since $i \geq 1 \leq j$.

Thus (12) holds with

$$\hat{h} \equiv \sum_{i=i_0+1}^{\hat{k}} h_i \circ g_1 + \sum_{i=i_0}^{\hat{k}} \sum_{j=1}^{\hat{k}-i} w_{ij}/j! \ ,$$

provided it can be shown that the remainder in (15) is $O(\rho^{\hat{k}+1})$ for each $i \in [i_0,\hat{k}]$. Since $x \in \hat{\mathcal{W}}$ we derive from Lemma 2.1 (iii)

$$\begin{aligned} \|y_i/\rho - \kappa t\| &= \|\alpha_i(g(x)-x^*)/\rho + (1-\alpha_i)g_1(t) - \kappa t\| \\ &\le \alpha_i \|(g(x)-x^*)/\rho - g_1(t)\| + \|g_1(t) - \kappa t\| \\ &\le d\hat{\rho}/\hat{\nu}^2 + \kappa\, c\, \hat{\nu}^{-1} \sin\hat{\phi} \le [4(k+1)]^{-1} \sin\phi \ , \end{aligned}$$

where the last inequality follows from the definition of $\hat{\phi}$ and $\hat{\rho}$ in (2.28) and (2.29) respectively.

Therefore we have

$$\left(\frac{k-\frac{1}{4}}{k+1}\right) \le \frac{\|y_i\|}{\rho} \le \left(\frac{k+\frac{1}{4}}{k+1}\right) , \tag{16}$$

and the angle $\Delta\psi$ between y_i and t satisfies

$$\sin\Delta\psi = \min_{\lambda\in\mathbb{R}} \|t-\lambda y_i\| \le \frac{1}{4}\sin\phi \ ,$$

so that with $\hat{\phi} \le \phi/4$

$$\cos^{-1}(s^T y_i/\|y_i\|) \le \cos^{-1}(s^T t) + \cos^{-1}(t^T y_i/\|y_i\|) \le \frac{1}{2}\phi \ .$$

Consequently $s_i \equiv y_i/\|y_i\|$ belongs to $\bar{U}$ and we obtain from (i) and (ii) with (16)

$$\begin{aligned} &\nabla^{\hat{k}-i+1} h_i(y_i)(g(x)-x^*-g_1(z))^{\hat{k}-i+1} \\ &= \|y_i\|^{2i-\hat{k}-1}\rho^{2(\hat{k}-i+1)} h_i(s_i)(g_2(t)+O(\rho))^{\hat{k}-i+1} \\ &= O(\rho^{\hat{k}+1}) \ , \end{aligned}$$

which completes the proof. ////

After these preparations we can now prove the main extrapolation result.

THEOREM 4.2 *Extrapolation at Regular-Singularities*

Let $f \in C^{\bar{k}+1,1}(\mathbb{R}^n,\mathbb{R}^n)$ have a regular singularity of order k at $x^* \in f^{-1}(0)$ with rank $(\nabla f(x^*)) = n-m < n$. Suppose the Newton sequence $\{x_j=x_j^{(0)}\}_{j\geq 0}$ converges regularly to x^* with the unique tangent $s \in N \cap S^1$. Then there exist $\hat{k} \geq \bar{k}-(m+1)(k-1)$ functions $\{h^{(q)}\}_{q\in[0,\hat{k}-1]} \subset H^n$ with $\mathrm{ord}(h^{(q)}) \geq q+1$ such that the sequences of extrapolants $\{x_j^{(q)}\}_{j\geq q}$ defined by (3) satisfy for $j \geq 2q$

$$x_j^{(q)} = x^* + h^{(q)}(x_{j-2q}-x^*) + O(\|x_{j-2q}-x^*\|^{\hat{k}+1}) \tag{17}$$

$$= x^* + \rho_{j-2q}^{q+1}\, h_{q+1}^{(q)}(s) + O(\rho_{j-2q}^{q+2}) , \tag{18}$$

where $h_{q+1}^{(q)} \in H_{q+1}^n$ is the "leading term" in the representation (11) of $h^{(q)}$. Consequently each sequence $\{x_j^{(q)}\}_{j\geq q}$ with $q \in [0,\hat{k}-1]$ converges linearly to x^* such that

$$\lim_{j\to\infty} \frac{\|x_{j+1}^{(q)}-x^*\|}{\|x_j^{(q)}-x^*\|} = \left(\frac{k}{k+1}\right)^{q+1} \quad \text{if} \quad h_{q+1}^{(q)}(s) \neq 0 \tag{19}$$

and

$$\lim_{j\to\infty} \|x_j^{(q)}-x^*\|^{1/j} \leq \left(\frac{k}{k+1}\right)^{q+1} \quad \text{otherwise.} \tag{20}$$

Proof.

According to (3.15) all but finitely many of the iterates $\{x_j\}$ belong to the set

$$\mathcal{D} \equiv \bigcap_{j=0}^{\infty} g^{-j}(\hat{W}) \quad , \quad g(\mathcal{D}) \subset \mathcal{D} \subseteq \hat{W}$$

of all points from which Newton's method converges to x^* without ever leaving $\hat{W}$. By (2.32) we have for any $x = x^*+\rho t \in \mathcal{D}$ and all $j \geq 1$

$$\left(\frac{k-\frac{1}{4}}{k+1}\right)^j \le \frac{\|g^j(x)-x^*\|}{\rho} \le \left(\frac{k+\frac{1}{4}}{k+1}\right)^j < 1 \ . \tag{21}$$

On the domain $\mathcal{D}$ we define recursively the vector functions

$$g^{(q+1)} \equiv \frac{g^{(q)} \circ g \circ g - \kappa^{q+1} g^{(q)} \circ g}{1 - \kappa^{q+1}} \quad \text{for} \quad q \ge 0 \tag{22}$$

with

$$g^{(0)}(x) = x \qquad \text{for all} \quad x \in \mathcal{D} \ .$$

Because of (4) it can be easily checked by induction on $q \ge 0$ that

$$x_j^{(q)} = g^{(q)}(x_{j-2q}) \quad \text{for all} \quad j \ge 2q \ . \tag{23}$$

On the basis of Lemma 3.7 (iv) we show by induction on q that there are $h^{(q)} \in H^n$ with $\operatorname{ord}(h^{(q)}) \ge q+1$ such that for all $x = x^*+z = x^*+\rho t \in \mathcal{D}$

$$g^{(q)}(x) = x^* + h^{(q)}(z) + O(\rho^{\hat{k}+1}) \ , \tag{24}$$

which is true for $q = 0$ with $h^{(0)} = g^{(0)} \in H_1$. Suppose (24) holds for some $q \ge 0$. By Lemma 4.1 (iv) there is a function $\hat{h}^{(q)} \in H^n$ with $\operatorname{ord}(\hat{h}^{(q)}) \ge q+2$ such that for all $x = x^*+\rho t = x^*+z \in \hat{W}$

$$h^{(q)}(g(x)-x^*) = h^{(q)}_{q+1}(g_1(z)) + \hat{h}^{(q)}(z) + O(\rho^{\hat{k}+1}) \ .$$

Applying the same result again we obtain a function $\tilde{h}^{(q)} \in H^n$ with $\operatorname{ord}(h^{(q)}) \ge q+2$ such that for all $x = x^*+\rho t = x^*+z \in \hat{W}$

$$h^{(q)}_{q+1}(g_1(g(x)-x^*)) + \hat{h}^{(q)}(g(x)-x^*)$$

$$= h^{(q)}_{q+1}(g_1(g_1(z))) + \tilde{h}^{(q)}(z) + O(\rho^{\hat{k}+1}) \ .$$

Since by homogeneity of $h^{(q)}_{q+1}$ and (2.15)

$$h^{(q)}_{q+1}(g_1(g_1(z))) = \kappa^{q+1} h^{(q)}_{q+1}(g_1(z)) \ ,$$

we derive from the definition (22) with (21) for all $x = x^*+\rho t = x^*+z \in \mathcal{D}$

$$(1-\kappa^{q+1})g^{(q+1)}(x) = g^{(q)}(g(g(x))) - \kappa^{q+1}g^{(q)}(g(x))$$

$$= x^* + h^{(q)}(g(g(x))-x^*) - \kappa^{q+1}[x^*+h^{(q)}(g(x)-x^*)] + O(\rho^{\hat{k}+1})$$

$$= (1-\kappa^{q+1})x^* + h^{(q)}_{q+1}(g_1(g(x)-x^*)) + \hat{h}^{(q)}(g(x)-x^*)$$

$$- \kappa^{q+1}[h^{(q)}_{q+1}(g_1(z)) + \hat{h}^{(q)}(z)] + O(\rho^{\hat{k}+1})$$

$$\equiv (1-\kappa^{q+1})x^* + \tilde{h}^{(q)}(z) - \kappa^{q+1}\hat{h}^{(q)}(z) + O(\rho^{\hat{k}+1}) .$$

Thus (24) holds for $q+1$ with

$$h^{(q+1)} = \frac{\tilde{h}^{(q)}-\kappa^{q+1}\hat{h}^{(q)}}{1-\kappa^{q+1}} \in H^n , \quad \mathrm{ord}(h^{(q+1)}) \geq q+2 .$$

Therefore (24) holds for all $q \geq 0$ which implies (17) by (23)

Since each $h^{(q)}_{q+1} \in H^n_{q+1}$ is differentiable in some neighbourhood of $s \in \hat{U}$ it must be locally Lipschitz continuous so that by (3.17) and Lemma 4.1 (ii) with $\mathrm{ord}(h^{(q)}-h^{(q)}_{q+1}) \geq q+2$

$$x^{(q)}_j - x^* = \rho^{q+1}_{j-2q}\left[h^{(q)}_{q+1}(t_{j-2q}) + O(\rho_{j-2q})\right]$$

$$= \rho^{q+1}_{j-2q}\left[h^{(q)}_{q+1}(s) + O(\rho_{j-2q})\right] .$$

For any fixed $q \in [0,\hat{k}-1]$ this implies (18) and (3.16) if $h^{(q)}_{q+1}(s) \neq 0$. Otherwise there must be a constant τ_q such that

$$\|x^{(q)}_j - x^*\| \leq \tau_q \rho^{q+2}_{j-2q} \quad \text{for all} \quad j \geq 2q ,$$

which implies

$$\lim_{j\to\infty} \|x^{(q)}_j - x^*\|^{1/j} \leq \left[\lim_{j\to\infty} \rho^{1/j}_{j-2q}\right]^{q+2} \leq \left[\lim_{j\to\infty} \frac{\rho_{j-2q+1}}{\rho_{j-2q}}\right]^{q+2} = \left(\frac{k}{k+1}\right)^{q+2} ,$$

where the last inequality holds by 9.3.1 in [10]. ////

As we can see in Table 2 the leading terms $h^{(q)}_{q+1}$ are in general nontrivial, since the first four extrapolation columns (q=1,2,3,4) converge linearly with Q-factors $1/4$, $1/8$, $1/16$ and $1/32$ as predicted by (19). With the notable exceptions of the column $q = 1$ and of course $q = 0$ the singular and nonsingular variables converge with the same Q-factor, so that the ranges of the vector functions $h^{(q)}_{q+1}$ are in general not contained in N. Since by (18)

$$f^{(q)}_j \equiv f(x^{(q)}_j) = \nabla f(x^*)(x^{(q)}_j - x^*) + O(\|x^{(q)}_j - x^*\|^2)$$

$$= \rho^{q+1}_{j-2q}\left[\nabla f(x^*) h^{(q)}_{q+1}(s) + O(\rho_{j-2q})\right] ,$$

the residuals $\{f^{(q)}_j\}_{j\geq 2q}$ decline essentially colinearly at the same rate as the discrepancies $\{x^{(q)}_j - x^*\}_{j\geq 2q}$. Hence we can gauge the progress of each column towards x^* by evaluating the residual $f^{(q)}_j$ even though it provides no reliable measure of the distance to x^*, since the angle between $h^{(q)}_{q+1}(s)$ and N can be large or small and may even be zero. By (3.26) we have with $\lambda = 1$

$$f^{(0)}_{j+1} = f(x_j(1)) = \frac{\rho_j^2}{2(k+1)^2} \nabla^2 f(x^*) s^2 + O(\rho_j^3)$$

and with $\lambda = k + 1$

$$f^{(1)}_{j+1} = f(x_j(k+1)) = \frac{\rho_j^2 (k-1)}{2k} \nabla^2 f(x^*) s^2 + O(\rho_j^3) ,$$

so that

$$\lim_{j\to\infty} \|f^{(1)}_{j+1}\| / \|f^{(0)}_{j+1}\| = (k^3+k^2-k-1)/k . \qquad (25)$$

At first order singularities $f^{(1)}_{j+1}$ is $O(\rho^3_j) = O(\rho^3_{j-1})$ which implies that the leading term $h^{(1)}_2(s)$ must be an element of N. This is indeed the case as we can see in Table 2 that the nonsingular components of both

$x_j^{(0)}$ and $x_j^{(1)}$ converge with Q-factor $1/8$. Whereas for $k = 1$ the first extrapolation column achieves a considerable reduction of the residual, this is not the case at higher order singularities as

$$(k^3+k^2-k-1)/k > k^2 \quad \text{for} \quad k > 1 .$$

Nevertheless we see in Table 9 that the extrapolants $x_j^{(1)}$ are from the third step onwards consistently closer to x^* than the $x_j^{(0)}$, which must hold in the limit by (19).

Especially in cases where the main objective is the reduction of the residual, the fact that the quality of the extrapolants can be tested by evaluation of f is certainly a great advantage compared to most other applications of extrapolation processes. Another important difference is that the cost of obtaining x_{j+1} from x_j is constant in j, whereas for instance in the case of differential equations each refinement of the discretization increases the computational requirements considerably. Finally we note that in contrast to most other applications the errors of subsequent iterates x_j are not mutually independent. Strictly speaking any error that occurs in the step from x_j to x_{j+1} moves the iteration onto another Newton sequence with a different expansion (2) should that exist at all.

Suppose the Newton iterates $\{x_{j-\ell}\}_{\ell \in [0,2q]}$ have been calculated in finite precision arithmetic from x_{j-2q}. Neglecting the error that occurs in the extrapolation process itself we obtain from (3) by the triangular inequality

$$\varepsilon_j^{(q)} \equiv \|x_j^{(q)} - g^{(q)}(x_{j-2q})\| \leq \sum_{\ell=0}^{q} \varepsilon_{j-\ell} \alpha_\ell(k,q) , \tag{26}$$

where

$$\varepsilon_{j-\ell} \equiv \|x_{j-\ell} - g^{2q-\ell}(x_{j-2q})\| \quad \text{for} \quad \ell \in [0,2q]$$

and

$$\alpha_\ell(k,q) \equiv \frac{\kappa^{\frac{1}{2}(\ell+1)\ell}}{\prod_{i=1}^{\ell}(1-\kappa^i)\prod_{i=1}^{q-\ell}(1-\kappa^i)} \quad \text{for} \quad \ell \in [0,q] .$$

As will be shown later we can assume that the errors ε_j grow geometrically in j such that for some $\gamma \in (0,1]$ and $\bar{\varepsilon}_j > 0$

$$\varepsilon_{j-\ell} \le \gamma^\ell \bar{\varepsilon}_j \quad \text{for} \quad \ell \in [0,q] . \tag{27}$$

Substituting this inequality into (26) we find that

$$\varepsilon_j^{(q)}/\bar{\varepsilon}_j \le \hat{\alpha}_\gamma(k,q) \equiv \sum_{\ell=0}^{q} \gamma^\ell \alpha_\ell(k,q) \tag{28}$$

$$= \left[\sum_{\ell=0}^{q} \gamma^\ell \kappa^{\frac{1}{2}(\ell+1)\ell} \cdot \frac{(1-\kappa^q)(1-\kappa^{q-1})\ldots(1-\kappa^{q-\ell+1})}{(1-\kappa)(1-\kappa^2)\cdot(1-\kappa^\ell)}\right] \Big/ \prod_{i=1}^{q}(1-\kappa^i) .$$

By Theorem 348 in [35] the $\hat{\alpha}_\gamma$ have the product form

$$\hat{\alpha}_\gamma(k,q) = \frac{(1+\gamma\kappa)(1+\gamma\kappa^2)\ldots(1+\gamma\kappa^q)}{(1-\kappa)(1-\kappa^2)\ldots\ldots(1-\kappa^q)} \le \hat{\alpha}_1(k,q) , \tag{29}$$

which can be checked by induction on q. Since $\kappa = k/(k+1)$ the $\hat{\alpha}_\gamma(k,q)$ are strictly increasing in k,γ and q. According to a much more general result by Laurent [36] the $\hat{\alpha}_1(k,q)$ are bounded in q for fixed k so that there are limits

$$\bar{\alpha}_\gamma(k) = \lim_{q\to\infty} \hat{\alpha}_\gamma(k,q) < \infty . \tag{30}$$

The $\bar{\alpha}_\gamma(k)$ grow exponentially in k as shown in the following Lemma.

LEMMA 4.3 *Growth of Error with Order of Singularity*

The bounds $\bar{\alpha}_\gamma(k)$ defined by (18) and (29) satisfy with $\pi = 3.1415...$, for all $k \in \mathbb{N}$

$$|\ell n\ \kappa|\ell n\ \bar{\alpha}_\gamma(k) + (k+1)^{-1} < \lim_{k\to\infty} k^{-1}\ell n\ \bar{\alpha}_\gamma(k) = \gamma\tau_0(\gamma) + \pi^2/6\ , \quad (31)$$

where

$$\tau_0(\gamma) = \int_0^1 (1+\gamma w)^{-1}|\ell n\ w|dw \in [\pi^2/12\ ,\ 1-\gamma(1-\pi^2/12)]\ . \quad (32)$$

Proof.

Since the ratio $(1+\gamma\kappa^y)/(1-\kappa^y)$ is strictly decreasing in $y > 0$ we can bracket $\ell n\ \bar{\alpha}_\gamma(k)$ by the following integrals

$$\ell n\left(\frac{1+\gamma\kappa}{1-\kappa}\right) + \int_2^\infty \ell n\left(\frac{1+\gamma\kappa^y}{1-\kappa^y}\right) dy$$

$$\le \ell n(\bar{\alpha}_\gamma(k)) = \sum_{q=1}^{\infty}\left[\ell n(1+\gamma\kappa^q) - \ell n(1-\kappa^q)\right] \quad (33)$$

$$\le \ell n\left(\frac{1+\gamma\kappa}{1-\kappa}\right) + \int_1^\infty \ell n\left(\frac{1+\gamma\kappa^y}{1-\kappa^y}\right) dy\ .$$

For $\beta \in [-1,1]$ and $i \ge 0$ we derive with $w = \kappa^y$

$$\int_i^\infty \ell n(1+\beta\kappa^y)dy = |\ell n\ \kappa|^{-1}\int_0^{\kappa^i} w^{-1}\ell n(1+\beta w)dw$$

$$= |\ell n\ \kappa|^{-1}\left[\ell n\ w\ \ell n(1+\beta w)\Big|_0^{\kappa^i} + \beta\int_0^{\kappa^i}(1+\beta w)^{-1}|\ell n\ w|dw\right] \quad (34)$$

$$= -\ i\ \ell n(1+\beta\kappa^i) + \beta|\ell n\ \kappa|^{-1}\tau_i(\beta)\ ,$$

where

$$\tau_i(\beta) \equiv \int_0^{\kappa^i} |\ell n\ w|(1+\beta w)^{-1}dw \le \tau_0(\beta) \le \tau_0(-1)\ .$$

Substituting (34) for $\beta = \gamma, -1$ and $i = 1,2$ into (33) we find after multiplication by $|\ln \kappa|$

$$|\ln \kappa|\ \ln\left[\frac{(1+\gamma\kappa)(1-\kappa^2)^2}{(1-\kappa)(1+\gamma\kappa^2)^2}\right] + \gamma\tau_2(\gamma) + \tau_2(-1)$$

$$\leq |\ln \kappa|\ \ln(\bar{\alpha}_\gamma(k)) \leq \gamma\tau_1(\gamma) + \tau_1(-1) \leq \gamma\tau_0(\gamma) + \tau_0(-1) \ . \tag{35}$$

Whereas $\tau_1(\gamma) = \tau_0(\gamma) - O(|1-\kappa|^2)$, we find with $|\ln w| \geq 1-w$

$$\tau_1(-1) = \tau_0(-1) - \int_\kappa^1 |\ln w|(1-w)^{-1}dw \leq \tau_0(-1) - (1-\kappa) \ .$$

It can be easily seen that $\tau_0(\gamma)$ is decreasing and convex in $\gamma \in [0,1]$, so that

$$\tau_0(1) \leq \tau_0(\gamma) \leq (1-\gamma)\tau_0(0) + \gamma\tau_0(1) \ ,$$

which implies (32) and the first part of (31) since

$$\tau_0(0) = 1 \ , \ \tau_0(1) = \pi^2/12 \quad \text{and} \quad \tau_0(-1) = \pi^2/6 \ ,$$

as stated on page 352 in [37] .

Since obviously

$$\lim_{k\to\infty} \kappa = 1 \quad \text{and} \quad \lim_{k\to\infty} k|\ln \kappa| = 1 \ ,$$

we have

$$\lim_{k\to\infty} \tau_i(\beta) = \tau_0(\beta) \quad \text{for} \quad i=1,2 \quad \text{and} \quad \beta = \gamma,-1 \ ,$$

and furthermore

$$\lim_{k\to\infty} |\ln \kappa|\ \ln\left[\frac{(1+\gamma\kappa)\ (1-\kappa^2)^2}{(1+\gamma\kappa^2)^2(1-\kappa)}\right]$$

$$= \lim_{k\to\infty} \frac{1}{k} \ln\left[\frac{2}{1+\gamma}\ (1-\kappa^2)\right] = \lim_{k\to\infty} \frac{1}{k} \ln\left(\frac{2k+1}{(k+1)^2}\right) = 0 \ .$$

Substituting these limits into (35) we obtain the equality in (31) which completes the proof. ////

Since $|\ell n\ \kappa|^{-1} \approx k$ we must expect by (31)

$$\bar{\alpha}_\gamma(k) \approx e^{k[\pi^2/6+\gamma-(k+1)^{-1}]}\ , \tag{36}$$

and in particular for $\gamma = 1$ by (32)

$$\bar{\alpha}_1(k) \approx e^{k[\pi^2/4-(k+1)^{-1}]}\ . \tag{37}$$

These approximations are quite good even for small k which are of course the only ones of any practical interest. Before listing some exact values of $\hat{\alpha}_\gamma(k,q)$ and $\bar{\alpha}_\gamma(k)$ we have to choose a suitable constant $\gamma \in (0,1]$.

Since for $m<n$ the condition number of the Jacobian $\nabla f(x_{j-\ell})$ is by Lemma 2.1 (ii) proportional to $\rho_{j-\ell}^{-k}$, and the stepsize $\|x_{j-\ell+1} - x_{j-\ell}\|$ according to (3.24) approximately $\rho_{j-\ell}/(k+1)$, we can expect that for some constant η

$$\Delta\varepsilon_{j-\ell} \equiv \|x_{j-\ell+1}-g(x_{j-\ell})\| \approx \eta\rho_{j-\ell}^{1-k} \approx \Delta\varepsilon_j\ \kappa^{(k-1)\ell} \tag{38}$$

where we have used (3.16) to obtain the last "equality". Assuming that the inverse image $g^{-q}(x_{j-q})$ contains some element $\tilde{x}_{j-2q} \in \hat{W}(s)$ with $\|\tilde{x}_{j-2q}-x^*\| \approx \rho_{j-2q}$ we may consider the iterate x_{j-q} to be exact. Differentiating (1.21) we find with (2.12) and Lemma 4.1 (i) for $x = x^*+\rho t = x^*+z \in \hat{W}$

$$\nabla g(x) = \nabla g_1(z) + O(\rho) = \frac{k}{k+1}\begin{pmatrix} I\ , & \bar{B}^{-1}(t)\bar{C}^T(t) \\ 0\ , & 0 \end{pmatrix} + O(\rho)\ .$$

All eigenvalues of the matrix on the RHS are zero or κ so that we can expect that the errors satisfy approximately the recurrence relation

$$\varepsilon_{j-\ell+1} \approx \kappa\varepsilon_{j-\ell} + \Delta\varepsilon_{j-\ell}\ ,$$

which leads by (38) to the estimate

$$\varepsilon_{j-\ell} \approx \Delta\varepsilon_j \left(\frac{1-\kappa^{k(q-\ell)}}{1-\kappa^k}\right) \kappa^{(k-1)(\ell+1)}$$

$$\approx \frac{\Delta\varepsilon_j \kappa^{k-1}}{1-\kappa^k} \kappa^{(k-1)\ell} .$$

By comparison with (27) and (38) we conclude that with $\gamma \equiv \kappa^{k-1} \in (1/e,1]$

$$\varepsilon_j^{(q)} \approx \frac{\Delta\varepsilon_j \kappa^{k-1}}{1-\kappa^k} \hat{\alpha}_\gamma(k,q) \approx \frac{\eta\kappa^{k-1}}{1-\kappa^k} \hat{\alpha}_\gamma(k,q)\rho_j^{1-k} \tag{39}$$

In our context this estimate is certainly more appropriate than the assumption that all vectors $\{x_{j-\ell}\}_{\ell\in[0,q]}$ are computed with errors of the same magnitude, which is usually made in the case of differential equations. The $\hat{\alpha}_\gamma(k,q)$ are always smaller than the $\hat{\alpha}_1(k,q)$ and we have according to (36) and (37) in the limit $q \to \infty$

$$\bar{\alpha}_\gamma(k)/\bar{\alpha}_1(k) \approx e^{k(\gamma-\pi^2/12)} \approx .63^k$$

since $\kappa^{k-1} \approx 1/e$ for sufficiently large k. The values of $\hat{\alpha}_\gamma(k,q)$, $\hat{\alpha}_1(k,q)$ and their bounds $\bar{\alpha}_\gamma(k)$ and $\bar{\alpha}_1(k)$ are listed in the following table for $k \in [1,4]$ and $q \in [1,7]$.

	k = 1	k = 2	k = 3	k = 4
$\hat{\alpha}_\gamma(k,1)/\hat{\alpha}_1(k,1)$	3.0/3.0	4.3/5.0	5.7/7.0	7.1/9.0
$\hat{\alpha}_\gamma(k,2)/\hat{\alpha}_1(k,2)$	5.0/5.0	10./13.	17.1/25.	26./41.
$\hat{\alpha}_\gamma(k,3)/\hat{\alpha}_1(k,3)$	6.4/6.4	17./24.	37./62.	67./127.
$\hat{\alpha}_\gamma(k,4)/\hat{\alpha}_1(k,4)$	7.3/7.3	24./36.	63./118.	138./303.
$\hat{\alpha}_\gamma(k,5)/\hat{\alpha}_1(k,5)$	7.8/7.8	30./47.	94./192.	239./599
$\hat{\alpha}_\gamma(k,6)/\hat{\alpha}_1(k,6)$	8.0/8.0	35./56.	126./275.	368./1020.
$\hat{\alpha}_\gamma(k,7)/\hat{\alpha}_1(k,7)$	8.1/8.1	39./62.	156./360.	515./1750.
:	:	:	:	:
$\bar{\alpha}_\gamma(k)/\bar{\alpha}_1(k)$	8.3/8.3	47./79.	294./803.	1890./8450.

As with all extrapolation procedures we face the difficult question up to which stage the extrapolation should be carried out. Since the smallest singular value of ∇f declines by (3.22) with a Q-factor of κ^k it would seem desirable to choose $\bar{q} \geq k$ such that the (exact) extrapolants $x_j^{(\bar{q})}$ converge faster than the conditioning of the Jacobian deteriorates. According to the table above the diagonal elements $\hat{\alpha}_\gamma(k,k)$ equal roughly the arithmetic mean of $\hat{\alpha}_\gamma(k,1)$ and $\bar{\alpha}_\gamma(k)$ so that the error of $x_j^{(k)}$ may still be acceptable if extrapolation works at all. For any $q \geq 0$ we have by (38) and (39) with (3.16)

$$\varepsilon_{j+1}^{(q)} / \varepsilon_j^{(q)} \approx \kappa^{1-k} ,$$

so that the error of all extrapolants including the Newton iterates $\{x_j\}$ grows unbounded whenever the order k of the singularity x^* is greater than 1 .

Combining (18) with (39) we obtain for the computed extrapolants $x_j^{(q)}$

$$\|x_j^{(q)} - x^*\| \approx \rho_j^{q+1} \frac{h_{q+1}^{(q)}(s)}{\kappa^{2(q+1)q}} + \rho_j^{1-k} \frac{\eta \kappa^{k-1}}{1-\kappa^k} \hat{\alpha}_\gamma(k,q) ,$$

where we have used (3.16) to replace ρ_{j-2q} by $\rho_j \kappa^{-2q}$. The constant η was introduced in (38) and can be expected to be of order 10^{-t} if the calculation is performed in t-digit floating point arithmetic. The leading coefficients $h_{q+1}^{(q)}$ depend on the terms g_i in the expansion (1.21) of g and thus by (1.22) on the higher derivatives of f . Therefore we can make no general statement about their magnitude which could grow very rapidly with q . If this is so the extrapolants may not ever come closer to x^* than the Newton iterates $\{x_j\}$ themselves before the structure of the iteration is destroyed by rounding errors. On the other hand some extrapolants can approximate the solution up to the desired accuracy long before the conditioning of the Jacobian becomes critical. This can be

observed in Table 2 where the best extrapolants of the 8-th row approximate the solution with single precision accuracy while the smallest pivot in the LU decomposition of ∇f is .004. Since the Jacobian has for $x \approx x^*$ a spectral norm close to 1 its conditioning at that stage of the iteration does still allow a fairly accurate calculation of the Newton steps. Five steps later the solution is already approximated with double precision accuracy, but the conditioning of the Jacobian has only deteriorated by a factor of 32.

The best extrapolants $x_j^{(q)}$ in each row of the Tables 2, 6, 9 and 10 satisfy approximately

$$q \approx \frac{1}{2} j \quad \text{if} \quad k = 1 \quad \text{and} \quad q \approx \frac{1}{3} j \quad \text{if} \quad k = 2 \ .$$

Whereas the first observation can be interpreted as a direct consequence of Theorem 4.2, the second one seems to indicate some underlying influence of k on the structure of the extrapolation table, which is not apparent from our analysis. Under the somewhat ideal conditions of our test calculations, extrapolation up to at least the sixth stage is quite successful in all four cases.

Compared to the factorisation of the Jacobian which requires $n^3/3 + O(n^2)$ arithmetic operations at each step, the computational effort for the update of the extrapolation table (5) according to (6) is almost negligible even for a large number of columns $\bar{q} < n$. Since we only have to keep the current row of extrapolants the storage requirement is $\bar{q}n$ words. Thus it seems worthwhile to set up an extrapolation table whenever the Newton iteration looks like converging to a regular singularity and then to test the quality of the extrapolants by periodic evaluation of the residuals $f_j^{(q)} = f(x_j^{(q)})$.

2. Bordering of Underdetermined or Singular Systems

In this final section we consider the numerical treatment of problems for which the Jacobian is known to have a nontrivial nullspace at a

solution of interest. Therefore we allow the nonlinear system to be underdetermined so that for some $n' \geq 0$

$$f \in C^1(\mathbb{R}^{n+n'}, \mathbb{R}^n).$$

In many applications the n' additional variables are *control parameters* whose physical interpretation is different from that of the remaining n *state variables*. Typically the dependence of f on the control parameters is rather straightforward so that the corresponding partial derivatives of f have a simple mathematical form and some may be constant or even zero. Since there is otherwise no intrinsic mathematical difference between the control parameters and state variables we will avoid the distinction for the sake of notational simplicity.

In the neighbourhood of any point at which the Jacobian ∇f has full row rank n, the solution set $f^{-1}(0) \subset \mathbb{R}^{n+n'}$ forms according to the implicit function theorem a differentiable n'- dimensional manifold. Of particular importance is the case $n' = 1$, in which the solutions form smooth curves as long as the condition $\text{rank } (\nabla f) = n$ is satisfied. Numerical methods for tracing such curves have been developed by several authors, e.g. [38] and [39] and there is no real difficulty until the procedure approaches a singular point, i.e. a solution at which the Jacobian has rank $(n-1)$ or less. Such points are of particular interest because in their neighbourhood $f^{-1}(0)$ no longer forms a manifold but may have a rather involved structure. In the case $n' = 1$ solution curves may end, branch or intersect with one or several other curves.

Now suppose we want to locate numerically a solution $x^* \in f^{-1}(0)$ at which $\text{rank } (\nabla f(x^*)) = n-m$, where the number of variables $n+n'$ may or may not exceed the number of equations n. As a consequence of Theorem 3.1

linear or faster convergence without discontinuities in the iteration function can only be obtained if we find at least $m+n'$ equations in addition to the obvious condition $f = 0$ such that the Jacobian of the resulting overdetermined system has full column rank $n+n'$. Since ∇f has rank $(n-m)$ iff its m smallest singular values vanish it may seem natural to impose the condition

$$\sigma_i(x) = 0 \quad \text{for} \quad i = 1...m \ .$$

Except for the case $n' = 0$, $m = 1$ this approach is not feasible because singular values are only differentiable when they are properly separated. Otherwise the individual elements of a singular value cluster can hardly be identified and the associated singular vectors may rotate very rapidly in the corresponding invariant subspace [40]. In any case we need $(m+n')m$ rather than m equations to ensure that $\text{rank}\,(\nabla f) = n-m$. To see this we assume that the leading $(n-m) \times (n-m)$ submatrix $B(x)$ is nonsingular in some neighbourhood U of x^* so that

$$\nabla f = \begin{pmatrix} B & , & C^T \\ D & , & E \end{pmatrix} = \begin{pmatrix} I_{n-m} & , & O \\ DB^{-1} & , & I_m \end{pmatrix} \begin{pmatrix} B & , & C^T \\ O & , & \Sigma^T \end{pmatrix} , \tag{40}$$

where the $m \times (m+n')$ matrix Σ^T is given by

$$\Sigma^T(x) = E(x) - D(x)B^{-1}(x)C^T(x) \ .$$

Clearly we have for all $x \in U$

$$\text{rank}(\nabla f(x)) = n-m \iff \Sigma(x) = 0 \ ,$$

so that in particular $\Sigma(x^*) = 0$. The entries of the matrix $\det(B(x))\Sigma(x)$ represent the determinants of all $(n-m+1) \times (n-m+1)$ submatrices of $\nabla f(x)$ that include $B(x)$. Each one of them involves an element of $E(x)$ that does not enter into any other determinant, so that the overdetermined system

$$f(x) = 0 \quad , \quad \Sigma(x) = 0$$

consists of $n+m(m+n')$ equations which are independent in that any subsystem may have a solution in U that does not solve the full system.

Provided all leading principal submatrices of $B(x)$ are nonsingular it has an LU factorization [11] and we can rewrite (40) as

$$\nabla f = \begin{pmatrix} L & , & O \\ DU^{-1} & , & I_m \end{pmatrix} \begin{pmatrix} U & , & L^{-1}C^T \\ O & , & \Sigma^T \end{pmatrix} , \tag{41}$$

which represents a partly completed LU decomposition of ∇f without pivoting. In the immediate neighbourhood of x^* all entries of $\Sigma(x)$ will be rather small, so that any attempt to complete the factorization would lead to large rounding errors.

Since we intend to solve an overdetermined system of some form it seems natural to consider a corresponding partly completed QR factorization

$$\nabla f = \begin{pmatrix} Q_{11} & , & Q_{12} \\ Q_{21} & , & Q_{22} \end{pmatrix} \begin{pmatrix} R_1 & , & R_2 \\ O & , & \hat{\Sigma}^T \end{pmatrix} , \tag{42}$$

where $R_1 \in \mathbb{R}^{(n-m)\times(n-m)}$ is upper triangular and $Q_{22} \in \mathbb{R}^{m\times m}$ forms together with Q_{12} , Q_{21} and Q_{11} an orthogonal $n\times n$ matrix. Whereas Q_{11} and Q_{12} are uniquely determined as

$$\begin{pmatrix} Q_{11} \\ Q_{21} \end{pmatrix} = \nabla f \begin{pmatrix} R_1^{-1} \\ O \end{pmatrix} ,$$

the matrices Q_{12} and Q_{22} depend on the particular triangularisation method employed. Because of the required orthogonality we derive from (42)

$$R_1^{-T}(I_{n-m},0)\nabla f^T\begin{pmatrix}Q_{12}\\ Q_{22}\end{pmatrix} = Q_{11}^T Q_{12} + Q_{21}^T O_{22} = 0 ,$$

so that the m column vectors of $(Q_{12}^T,Q_{22}^T)^T$ must form an orthonormal basis of the nullspace M of the $(n-m)\times n$ matrix $(I_{n-m},0)\nabla f^T$. In fact it can be easily seen that any orthonormal basis of M can be used to form (Q_{12}^T,Q_{22}^T) such that (42) holds for a suitable $\hat{\Sigma}$ with everything else unchanged.

In order to develop a more general framework which allows for arbitrary pivoting in the LU and QR decomposition we derive from (41) and (42) the equations

$$\nabla f^T\begin{pmatrix}-B^{-T}D^T\\ I_m\end{pmatrix} = \begin{pmatrix}O\\ I_{m+n'}\end{pmatrix}\Sigma \tag{43}$$

and

$$\nabla f^T\begin{pmatrix}Q_{12}\\ Q_{22}\end{pmatrix} = \begin{pmatrix}O\\ I_{m+n'}\end{pmatrix}\hat{\Sigma} . \tag{44}$$

The special structure of the matrix $(O,I_{m+n'})^T$ on both right hand sides is related to our assumption that $B(x)$ is nonsingular for $x \in U$, so that the $m+n'$ columns of $(O,I_{m+n'})^T$ together with the n columns of ∇f^T span the full space $\mathbb{R}^{n+n'}$.

Now consider any matrix function

$$W \in C^1(U,\mathbb{R}^{(n+n')\times(m+n')}) \tag{45}$$

such that for all x in the neighbourhood U of x^*

$$\operatorname{rank}(W(x)) = m+n' \quad \text{and} \quad \operatorname{rank}(\nabla f^T(x),W(x)) = n+n' . \tag{46}$$

Replacing $(O,I_{m+n'})^T$ by W we obtain instead of (43) and (44) the more general matrix equation

$$\nabla f^T(x)U(x) = W(x)\Sigma(x) \ , \tag{47}$$

where the "unknowns" $U(x)$ and $\Sigma(x)$ are $n\times m$ and $(m+n')\times m$ matrices respectively. Multiplying from the right by the generalised inverse $W^\dagger$ of W we find

$$W^\dagger(x)\nabla f^T(x)U(x) = \Sigma(x) \tag{48}$$

and

$$F(x)U(x) \equiv (I_{n+n'},-W(x)W^\dagger(x))\nabla f^T(x)U(x) = 0 \ . \tag{49}$$

Hence $U(x)$ can be any matrix whose range is contained in the nullspace $M(x)$ of $F(x) \in \mathbb{R}^{(n+n')\times n}$ and the corresponding $\Sigma(x)$ is then determined by (48)

To show that the dimension of $M(x)$ is m for all $x \in U$ we note that all nonzero vectors of the form $(I_{n+n'},-W^{\dagger T}(x)W^T(x))y$ with $y \in \mathbb{R}^{n+n'}$ are orthogonal to the columns of W and therefore by (46) cannot be orthogonal to the columns of $\nabla f^T(x)$. Consequently the nullspace of the transpose $F^T(x)$ is identical to the $m+n'$-dimensional range of $W^{\dagger T}(x)$ so that $\dim(M(x)) = m$ for all $x \in U$. In order to obtain for all $x \in U$ an essentially unique solution $U(x)$ of (49) with full column rank m we impose m^2 additional conditions in the form

$$N(x,U) = (N_{ij}(x,U))_{\substack{j=1..m \\ i=1..m}} = I_m \ , \tag{50}$$

where

$$N \in C^1(U\times\mathbb{R}^{n\times m},\mathbb{R}^{m\times m}) \ . \tag{51}$$

For instance we can require that some $m \times m$ submatrix of U equals the identity by setting

$$N(x,U) \equiv MU \ , \tag{52}$$

where the constant $m \times n$ matrix M is a column permutation of (O, I_m). This includes the case (43) but allows for arbitrary row exchanges in the Jacobian ∇f during the LU decomposition. Another natural condition would be $U^T(x)U(x) = I_m$ which represents because of its symmetry only $m(m+1)/2$ equations. In addition we can require that some row permutation of $U(x)$ is lower trapezoidal (i.e. has no nonzero elements above the diagonal) so that

$$N_{ij}(x,U) \equiv \begin{cases} e_{p_i}^T u_j & \text{if } i < j \\ u_i^T u_j & \text{if } i \geq j \ , \end{cases} \tag{53}$$

where e_i is the i-th Cartesian vector in $\mathbb{R}^n$ and $\{p_i\}_{i=1..m}$ a subset of m indices in $[1,n]$. In the case of the QR decomposition of ∇f, by a sequence of elementary reflectors

$$Q_i \equiv (I - 2\|q_i\|^{-2} q_i q_i^T) \quad \text{for } i=1..n-m$$

with

$$\begin{pmatrix} \hat{L} \\ S \end{pmatrix} \equiv (q_1, q_2, .. q_{n-m}) \in \mathbb{R}^{n \times (n-m)} \quad \text{lower trapezoidal}$$

the matrix $U \equiv (Q_{12}^T, Q_{22}^T)^T$ satisfies the condition

$$N(x,U) \equiv (-S(x)\hat{L}^{-1}(x), I)U = I_m \ , \tag{54}$$

where $S(x)$ and $\hat{L}(x)$ denote the matrices S and $\hat{L}$ computed from the Jacobian $\nabla f(x)$. The identity (54) can be derived from the fact that, for all y in the nullspace of $(\hat{L}^T, S^T)$ which is spanned by the rows of $(-S\hat{L}^{-1}, I)$, we must have

$$Q_1 Q_2 \cdot\cdot Q_{n-m} y = y \ .$$

Equation (54) is only important in so far as it allows us to treat the QR decomposition (42) in the general framework and has otherwise no apparent value.

The normalisation condition (50) will be called *nondegenerate* at $(x,U) \in N^{-1}(I_m)$ if for any $A \in \mathbb{R}^{m\times m}$

$$\frac{d}{d\lambda} N(x,U(I+\lambda A))\Big|_{\lambda=0} = 0 \Leftrightarrow A = 0 \ , \tag{55}$$

which requires in particular $\mathrm{rank}(U) = m$ as otherwise $UA = 0$ for some $A \neq 0$. For any N of the form

$$N(x,U) = M(x)U$$

we have

$$\frac{d}{d\lambda} N(x,U(I+\lambda A)) = M(x)UA = A \ ,$$

so that (55) is automatically satisfied. This applies for the examples (52) and (54).

In the case of N as defined by (53) the nondegeneracy condition (55) is satisfied if and only if all "diagonal" elements $\{e_{p_i}^T u_i\}_{i=1..m}$ are nonzero. If some diagonal element $e_{p_i}^T u_i$ is zero we can choose A as a rotation in the plane spanned by the n-vectors u_i and u_{i+1} without disturbing the orthogonality nor the (permuted) lower trapezoidal structure. Thus $N(x,U(I+\lambda A))$ is constant and the normalisation condition must be degenerate. Conversely the LHS of (55) requires that $U(I+\lambda A)$ be permuted lower trapezoidal for all λ, which implies because all diagonal elements are nonzero, that A is lower triangular as can be easily checked by contradiction. Differentiating the orthogonality condition in (53) we find

$$\frac{d}{d\lambda}\,(I+\lambda A^T)U^TU(I+\lambda A)\Big|_{\lambda=0} = A + A^T = 0\ ,$$

which implies $A = 0$ as no nontrivial matrix can be triangular and antisymmetric. Nondegeneracy of N ensures uniqueness and differentiability of U as shown in the following lemma.

LEMMA 4.4 *Uniqueness and Differentiability of* U *and* Σ

In some open set $\mathcal{U} \subseteq \mathbb{R}^{n+n'}$ let

$$N \in C^{1,1}(\mathcal{U}\times\mathbb{R}^{n\times m},\mathbb{R}^{m\times m}) \quad \text{and} \quad W \in C^{1,1}(\mathcal{U},\mathbb{R}^{(n+n')\times(m+n')})$$

be defined such that (46) holds at all $x \in \mathcal{U}$. If N is nondegenerate at $(x_0,U_0) \in N^{-1}(I_m)$ with $F(x_0)U_0 = 0$ then there are unique differentiable matrix functions

$$U \in C^{1,1}(\mathcal{U}',\mathbb{R}^{n\times m}) \quad \text{and} \quad \Sigma \in C^{1,1}(\mathcal{U}',\mathbb{R}^{(m+n')\times m})$$

defined on some neighbourhood $\mathcal{U}' \subseteq \mathcal{U}$ of x_0 such that $U(x_0) = U_0$ and for all $x \in \mathcal{U}'$

$$\nabla f^T(x)U(x) = W(x)\Sigma(x)\ ,\ N(x,U(x)) = I\ , \tag{56}$$

and

$$\text{defect}(\nabla f(x)) = n\text{-rank}(\nabla f(x)) = m\text{-rank}(\Sigma(x)) = \text{defect}(\Sigma(x))\ . \tag{57}$$

Proof.

The two matrix equations in (56) have $(n+n')m + m^2 = nm + (n'+m)m$ entries which equals the number of elements in U and Σ. At x_0 the system (56) has the solution pair U_0 and $\Sigma_0 = W^{\dagger}(x_0)\nabla f^T(x_0)U_0$. Now let $U' \in \mathbb{R}^{n\times m}$ and $\Sigma' \in \mathbb{R}^{(n'+m)\times m}$ be any matrix pair such that

$$\frac{d}{d\lambda}\left[\nabla f(x_0)(U_0 + \lambda U') - W(x_0)(\Sigma_0+\lambda\Sigma')\right]_{\lambda=0} = \nabla f(x_0)U' - W(x_0)\Sigma' = 0$$

and

$$\frac{d}{d\lambda} N(x_0, U_0+\lambda U') \Big|_{\lambda=0} = 0 \ .$$

Because of the first condition the columns of U' belong to the nullspace $M(x_0)$ which is spanned by the columns of U_0 so that $U' = U_0 A$ for some $A \in \mathbb{R}^{m\times n}$. It follows immediately from the nondegeneracy assumption (55) that $U' = 0$ and consequently $\Sigma' = 0$ so that the Jacobian of the full system (56) has no nonzero null "vector". Thus the implicit function theorem ensures the existence of unique solutions $U(x)$ and $\Sigma(x)$ whose derivatives can be explicitly given in terms of the derivatives of ∇f, W and N and are therefore locally Lipschitz continuous. To show (57) we note that $\nabla f(x)$ must have an $n'+m'$ dimensional nullspace if $\text{defect}(\nabla f^T(x)) = m' \leq m$. Multiplying the first equation (56) from the left by any $(n'+m') \times (n+n')$ matrix $A^T(x)$ whose rows span the nullspace of $\nabla f(x)$ we find $A^T(x)W(x)\Sigma(x) = 0$. Since each row of $A^T(x)$ is orthogonal to the columns of $\nabla f^T(x)$ no linear combination $y^T A^T(x)$ with $y \in \mathbb{R}^{n'+m'}$ can by (46) be orthogonal to the linearly independent columns of $W(x)$. Consequently the nullspace of $\Sigma^T(x)$ must contain the $n'+m'$ dimensional range of $W^T(x)A(x)$ so that $m'' \equiv \text{defect}(\Sigma(x)) \geq m'$. Conversely multiplication of (56) from the right, by any $m\times m''$ matrix $A(x)$, whose columns form a basis of the m'' dimensional nullspace of $\Sigma(x)$, yields $\nabla f^T(x)U(x)A(x) = 0$, so that the m'' dimensional range of $U(x)A(x)$ must be contained in the nullspace of $\nabla f^T(x)$ which implies $m' = m''$ and thus (57). ////

According to (57) the matrix $\Sigma(x)$ indicates to what extent the Jacobian $\nabla f^T(x)$ is still nontrivial on the subspace $M(x)$ where it is comparatively weak. We may view the columns of $U(x)$ as generalised left singular vectors of $\nabla f(x)$ and the entries of $\Sigma(x)$ as generalised singular values. If the columns of $W(x)$ form a differentiable basis of

the invariant subspace spanned by the right singular vectors associated with the smallest m singular values of $\nabla f(x)$, then the columns of $U(x)$ are linear combinations of the corresponding left singular vectors which are in general not differentiable as noted before. Besides being differentiable the columns of U and the entries of Σ have the distinct advantage that they can be calculated in finitely many arithmetic operations, provided this is true for the matrix W and $N(x,U) = I$ can be satsified by a finite transformation.

Now we return to the original problem of locating a solution $x^* \in U \cap f^{-1}(0)$ where $\text{rank}(\nabla f) = n-m$. Provided W and N satisfy the assumption of Lemma 4.4 we can apply the Gauss-Newton method to the overdetermined system

$$f(x) = 0 \ , \ \Sigma(x) = 0 \ . \tag{58}$$

In order to determine the conditions under which the Jacobian of this system is nonsingular at x^*, we consider a prospective nullvector $y \in \mathbb{R}^{n+n'}$, which must clearly belong to the nullspace N of $\nabla f(x^*)$. Denoting directional differentiation with respect to y by a subscript "y" we obtain from (56) with $\Sigma(x^*) = 0$

$$\nabla f_y^T(x^*)U(x^*) + \nabla f^T(x^*)U_y(x^*) = W(x^*)\Sigma_y(x^*) \ . \tag{59}$$

Let $\tilde{P}$ and P be the orthogonal projections onto the nullspaces of $\nabla f(x^*)$ and its transpose respectively. Because of (46) the matrix $\tilde{P}W(x^*)$ has full column rank so that

$$\Sigma_y(x^*) = 0 \iff \tilde{P}\nabla f_y^T(x^*)U(x^*) = 0 \ .$$

Since $U(x^*)$ spans the nullspace of $\nabla f^T(x^*)$ we find by transposing the RHS with $\nabla f_y = \nabla^2 f \cdot y$

$$\Sigma_y(x^*) = 0 \iff P(\nabla^2 f(x^*)y)\tilde{P} = 0 \ .$$

Thus we conclude that the Jacobian of (58) has full column rank $n+n'$ iff

$$P\nabla^2 f(x^*)y\Big|_N \neq 0 \quad \text{for all} \quad y \in N - \{0\} \ . \tag{60}$$

This condition is independent of the particular choice of W and certainly not very strong. If $n' = 1 = m$ the nullspaces of $\nabla f(x^*)$ and its transposed are spanned by vectors $u^* \in \mathbb{R}^n$ and v_1^*, $v_2^* \in \mathbb{R}^{n+1}$ respectively. Then (60) reduces to the condition that the symmetric 2×2 matrix

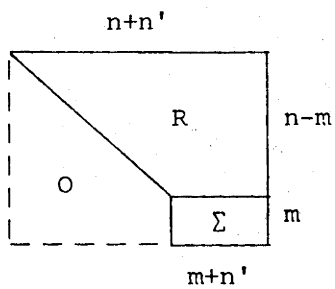
$$A \equiv \begin{pmatrix} u^{*T}\nabla^2 f(x^*)v_1^*v_1^* & , & u^{*T}\nabla^2 f(x^*)v_2^*v_1^* \\ u^{*T}\nabla^2 f(x^*)v_1^*v_2^* & , & u^{*T}\nabla^2 f(x^*)v_2^*v_2^* \end{pmatrix}$$

be nonsingular. If $\det(A) < 0$ (the Crandall-Rabinowitz or transversality condition [41]) there are two smooth solution curves that intersect at x^*. If $\det(A) > 0$, x^* is isolated in $f^{-1}(0)$ and if $\det(A) = 0$ it is most likely to be a cusp point, but $f^{-1}(0)$ may have an even more complicated structure in the neighbourhood of x^*.

The numerical solution of (58) by the exact Gauss-Newton method would involve the exact derivatives of Σ which depend by (59) on the second derivative tensor $\nabla^2 f$ and U. Even though the entries of ∇f may have a simple mathematical structure, e.g. if ∇f is essentially a discretisation matrix, the explicit evaluation of their derivatives would require a lot of additional coding if not computing time. Therefore it is much more practical to approximate the gradients of the $(m+n')m$ entries in Σ by successive updates according to Broyden's method [ ]. Even though the theory of quasi-Newton methods has apparently not yet been extended to overdetermined systems, there seems little doubt that the

analytical tools provided by Powell [43], Dennis and Moré [44] others can be used to establish superlinear convergence of *quasi-Gauss-Newton* methods based on the Broyden update of the rectangular Jacobian or parts of it. The test calculations reported in Tables 3, 7 with $n = 3$, $n' = 0$ and $m = 1$ or $m = 2$ show clear evidence of superlinear convergence, which is more reliable than that of any other method discussed in this thesis. One routine employed for these calculations is based on the QR decomposition by elementary reflectors.

Firstly the Jacobian is reduced by an orthogonal transformation, which simultaneously transforms the residual f to $\bar{f}$, to some column permutation of the form



(61)

At the initial point this process is carried out with full column pivoting such that the first $n-m$ diagonal elements in (61) have the largest possible moduli in nonincreasing order. From then on the same pivoting pattern must be applied as long as the modulus of the $(n-m)$-th diagonal element is clearly larger than the Euclidean norm of any column in the remaining rectangular matrix Σ. Otherwise the method must be restarted with a different pivoting pattern. It is important that none of the $n-m$ diagonal elements changes its sign, which would cause discontinuities in Σ. This may happen in Stewarts Algorithm 3.6 [11], which does however suit our requirements if the sign of the diagonal element σ is determined by the sign of the column component v_i with the largest modulus rather than v_1. Provided these precautions are taken $\Sigma(x)$, whose elements

are rational functions in the entries of $f(x)$, is clearly differentiable in some neighbourhood U of x^*. Now let $\sigma \in C^{1,1}(U,\mathbb{R}^{m(m+n')})$ be a vector function whose components are the elements of Σ in some fixed ordering.

Before the first step the $(m+n')m\times(n+n')$ matrix $G \approx \nabla\sigma$ is initialised as

$$G_0 \equiv \begin{pmatrix} 0 & , & I_{n+m'} \\ 0 & , & 0 \end{pmatrix}$$

such that the $(n+(m+n')m)\times(n+n')$ matrix

$$J = \begin{array}{|c|c|} \hline & R \\ \hline O & \Sigma \\ \hline G & \\ \hline \end{array} \qquad (62)$$

has full column rank $n+n'$ for $G = G_0$. After each step from $x - \tilde{s}$ to x the previous version $\tilde{G}$ of G is updated to $\hat{G}$ according to the "good" Broyden formula

$$\hat{G} = G + [\sigma(x)-\sigma(x-\tilde{s}) - \tilde{G}\tilde{s}] \frac{\tilde{s}^T}{\|\tilde{s}\|^2} .$$

The next correction $\hat{s}$ is determined as the solution of the linear least squares problem

$$\underset{s\in\mathbb{R}^{n+n'}}{\text{Min}} \left\| Js - \begin{pmatrix} \bar{f} \\ \sigma \end{pmatrix} \right\|_2^2 . \qquad (63)$$

To compute $\hat{s}$ we use $n+n'$ elementary reflectors $(I-2u_iu_i^T)$ to bring J into upper triangular form and apply them simultaneously to the RHS $(\bar{f}^T,\sigma^T)^T$. From the resulting triangular system $\hat{s}$ can be obtained by

back substitution, provided J has full column rank which can be expected if the condition (60) is satisfied and x_0 is sufficiently close to x^*. Since J is already in the form (62) the first $n-m$ vectors $u_i \in \mathbb{R}^{n+m(m+n')}$ have only $1 + (m+n')m$ nonzero elements.

It can be easily seen that, provided $(1+n')(m+1)$ is small compared to n, the computational expense at each step, including the Broyden update of G, is dominated by the orthogonal transformation of ∇f into the form (61), which requires approximately $\frac{2}{3} n^3$ arithmetic operations. This number can be halved if ∇f is brought into the form (61) by Gaussian elimination with complete pivoting. This simplication, which will be referred to as LU-bordering, hardly affects the speed of convergence [Table 3] even though the solutions of (63) minimise the residual

$$\begin{pmatrix} \nabla f \\ G \end{pmatrix} s - \begin{pmatrix} f \\ \sigma \end{pmatrix} \in \mathbb{R}^{n+m(m+n')} \tag{64}$$

no longer with respect to the Euclidean norm, but some other ellipsoidal norm which varies differentiably in $x \in \mathcal{U}$. Here $\mathcal{U}$ is a neighbourhood of x^* in which Gaussian elimination with some fixed pivoting pattern yields $n-m$ pivots, that are clearly separated from the elements of the remaining rectangular matrix Σ.

The bordering approach developed in the final sections appears to be the most reliable and accurate way to solve systems that are known to be singular. If one cannot be sure that the solution is exactly singular or does not know the dimension of the nullspace $m+n'$, the components of the residual (64) can be weighted by varying multipliers which may either emphasise the reduction of $\|f\|$ or enforce the singularity of ∇f with a nullspace of a certain dimension. In view of Theorem 3.1 it seems doubtful whether a weighting strategy can be designed that automatically ensures local superlinear convergence to any nonsingular or singular solution for which (60) is satisfied.

DISCUSSION AND CONCLUSION

In a theoretical sense the variety of structurally different singularities seems vast and is probably beyond a comprehensive mathematical description.

In practice we can expect that most singularities are of first order (k=1) with one-dimensional nullspace (m=1) in which case the conditions of isolation (1.47), regularity (2.16) and strong regularity (2.45), are equivalent to the assumption (4.60). If these are satisfied the slow linear convergence of Newton's method can be considerably accelerated at little cost either by the three-point method [Table 1] or extrapolation [Table 2]. Whenever the solution is required with high accuracy bordering based on the QR or LU decomposition of the Jacobian should be employed during the final stages of the computation. Since after some initial steps, the iterates are essentially confined to the one-dimensional nullspace N of $\nabla f(x^*)$, it can be conjectured that the quasi-Gauss-Newton method with Broyden update of the gradient $\nabla\sigma \approx \nabla\det(\nabla f)$ has the Q-order $\frac{1}{2}(1+\sqrt{5})$ of the secant method in one variable. As is well known [44] quasi-Newton updates yield usually poor approximations of the Jacobian if consecutive steps are essentially confined to a subspace of $\mathbb{R}^n$. Nevertheless it seems just possible that the case of a regular first order singularity with $m = 1$ could still be treated successfully when ∇f itself is not explicitly available. Any such scheme would necessarily involve two levels of differencing along the direction t that spans N, which amounts to quadratic interpolation and is therefore somewhat risky [45].

At first order singularities with $m > 1$, the condition (4.60) is considerably weaker than the regularity assumption (2.16) which in turn

is implied by the rather restrictive condition (2.45) of strong regularity. Even if the latter is not satisfied we can expect on the basis of numerical experience that the Newton method itself and the three-point method converge in a reasonably stable fashion. On our test problem the three-point method [Table 5] is considerably faster than the bordering scheme [Table 7] which makes initially little progress towards x^* until the derivatives of Σ have been approximated to some accuracy. This may have been caused by the fact that the four additional equations $\Sigma = 0$ dominated the condition $f = 0$ which effectively represents only one equation as $\text{rank}(\nabla f(x^*)) = 1$. Instead the singularity condition $\Sigma = 0$ should probably be phased in gradually as its approximated Jacobian becomes more accurate and $\|f\|$ sufficiently small. As in the other three cases extrapolation [Table 6] works surprisingly well and the best extrapolants are consistently closer to x^* than the iterates of the three-point method. It is remarkable that in all four extrapolation Tables 2, 6, 9 and 10 the extrapolants $\|x_j^{(q)}\|$ which are closest to x^* in the Euclidean norm are mostly identical to those that have the minimal residual $f_j^{(q)}$ in the same norm. Therefore we can generally expect that the extrapolant $x_j^{(q)}$ with the smallest residual $\|f_j^{(q)}\|$ in the last computed row is the best approximation to x^*.

At higher order singularities the condition (4.60) cannot be satisfied so that the Jacobian of the overdetermined system $f = 0$, $\Sigma = 0$ has at x^* linearly dependent columns. Then one might attempt a second level of bordering, but this approach seems only feasible if second derivatives of f can be explicitly calculated. Otherwise we are left with extrapolation [Tables 9,10] and the partially corrected two-point method [Table 8]. Whereas extrapolation requires only the choice of k,

the selection of suitable multipliers λ_j for the steps (3.68) is a difficult question which requires further investigation.

Singularities that are irregular and do not satisfy condition (4.60) cannot be treated by any of the methods discussed in this thesis. As shown in [24] such cases can arise in the context of minimisation problems and lead to very irregular behaviour of the Newton iteration even if the singularity is balanced and thus of first degree.

Apart from computational considerations there are many unresolved theoretical questions arising from the analyses in Chapter 1. In the unbalanced case the linear system (1.22) cannot be solved explicitly in the block triangular form (2.6). Then we have no way to determine the vector functions g_i and the crucial degree $\hat{i}$ explicitly. This may of course be possible by some other method, which would be of great benefit for the classification of singularities.

The analysis of regular singularities in Chapter 2 seems quite satisfactory and yields the important result that convergence of Newton's method is almost sure if the initial point is sufficiently close to the solution. The concepts of starlike domains and their density together with the rational expansions developed in Sections 1.3 and 4.1 may be useful for the analysis of other iterative methods that do not necessarily have spherical domains of convergence at certain solution points.

APPENDIX

TEST CALCULATIONS

All calculations were carried out on the following system of three equations in the variables $x = (\xi,\zeta,\eta)^T \in \mathbb{R}^3$.

$$f(x) \equiv \begin{pmatrix} (2-k)(-.4\xi^2-1.5\zeta^2+.3\eta^2+.5\xi\zeta-2\zeta\eta) \\ -.8\xi^3-3\zeta^3+.6\eta^3+.5\xi^2\zeta-2\zeta\eta^2 \\ (2-m)\xi+.5\zeta^3+.4\xi^2\zeta-\zeta\eta^2 \\ +[1-(m-1)(k-1)](.35\xi^2+\zeta^2+.75\eta^2+.4\xi\zeta+2\xi\eta-\zeta\eta) \\ \eta+.4\xi^2+.5\eta^2-2\zeta\xi-2.5\xi\eta+2\zeta\eta \end{pmatrix} = 0$$

The parameters $k=1,2$ and $m=1,2$ enter into f such that k gives the order of the singularity $x^* = 0 \in f^{-1}(0)$ and m the dimension of the nullspace N of $\nabla f(x^*)$. In all four cases f is in normal form at x^*, and we have

$$\bar{B}(x) = -.8\xi + .5\zeta \qquad \text{if} \quad k=1=m ,$$

$$\bar{B}(x) = -2.4\xi^2 + \xi\zeta \qquad \text{if} \quad k=2 , m=1 ,$$

$$\bar{B}(x) = \begin{pmatrix} -.9\xi+.5\zeta & , -3\zeta+.5\xi-2\eta \\ .7\xi+.4\zeta+2\eta & , 2\zeta+.4\xi-\eta \end{pmatrix} \qquad \text{if} \quad k=1 , m=2$$

and

$$\bar{B}(x) = \begin{pmatrix} -2.4\xi^2+\xi\zeta & , -9\zeta^2+.5\xi^2-2\eta^2 \\ .8\xi\zeta & , 1.5\zeta^2+.4\xi^2-\eta^2 \end{pmatrix} \qquad \text{if} \quad k=2=m .$$

With e_1 and e_2 the first two Cartesian base vectors in $\mathbb{R}^3$ we find

$$\det(\bar{B}(e_1)) = -.8 \qquad \text{if} \quad k=1=m \ ,$$

$$\det(\bar{B}(e_1)) = -2.4 \qquad \text{if} \quad k=2 \ , m=1$$

$$\det(\bar{B}(e_1)) = -.67 \quad \text{and} \quad \det(\bar{B}(e_2)) = 2.2 \qquad \text{if} \quad k=1 \ , m=2 \ ,$$

$$\det(\bar{B}(e_1)) = -.96 \quad \text{and} \quad \det(\bar{B}(e_2)) = 0 \qquad \text{if} \quad k=2=m \ .$$

Since e_1 or e_1 and e_2 span the nullspace of $\nabla f(x^*)$ if $m=1$ or $m=2$ respectively, the singularity $x^* = 0$ is regular in all four cases. Whereas for $m=1$ regularity implies strong regularity, the singularity x^* is not strongly regular in the two cases with $m=2$ as $\pi_0 = \det(\bar{B})$ vanishes for some $t \in N \cap S$.

The calculation reported in Table 4 was performed on the nearly singular system

$$f(x) + \xi \cdot 10^{-6} = 0 \quad \text{with} \quad k=1=m \ .$$

In all calculations, except for those based on bordering (Table 3 and 7), the Jacobian was reduced to triangular form by Gaussian elimination with complete pivoting. The smallest pivot is listed as σ . The factor by which the Newton correction is multiplied for the step *from* the current point is listed as λ . At each step in the bordering calculations the residual of the linear least squares solution of (4.63) is listed as μ . All iterations were started at the initial point $x_0 = (.1,.1,.1)^T$, from which an unmodified Newton step was taken to the first point listed in the Tables.

The calculations were performed on a UNIVAC 1110 in double precision (i.e. 16 digit) floating point arithmetic.

Table 1 , k = l = m

	Newton		three-point		two-point	
	.10+01	-.44-01	.10+01	-.44-01	.20+01	-.44-01
1	-.23-01	.49-01	-.23-01	.49-01	-.23-01	.49-01
	.64-01	-.31-01	.64-01	-.31-01	.64-01	-.31-01
	.10+01	.35-01	.20+01	.35-01	.10+01	.11+00
2	.36-01	.89-02	.36-01	.89-02	.85-01	-.31-01
	.96-02	.74-03	.96-02	.74-03	.39-01	.32-01
	.10+01	.16-01	.10+01	-.22-02	.20+01	.62-01
3	.14-01	.21-03	-.61-02	-.85-02	.62-01	.61-02
	.39-03	-.33-03	.85-02	-.14-02	.77-02	-.27-02
	.10+01	.84-02	.10+01	.20-01	.10+01	.56-02
4	.69-02	-.67-05	.17-01	.23-03	.12-02	-.63-02
	.45-04	.22-06	.45-03	-.39-03	.65-02	.20-02
	.10+01	.42-02	.20+01	.10-01	.20+01	-.33-01
5	.34-02	-.20-06	.82-02	-.97-05	-.24-01	.24-03
	.12-04	-.17-06	.63-04	.25-06	.12-02	.44-03
	.10+01	.21-02	.10+01	.92-04	.10+01	.14-02
6	.17-02	-.27-07	.78-04	.91-05	.98-03	-.24-03
	.30-05	-.29-07	.91-05	-.83-06	.43-03	-.36-03
	.10+01	.11-02	.10+01	.47-04	.20+01	.38-03
7	.86-03	-.34-08	.38-04	.96-10	.30-03	-.44-06
	.76-06	-.37-08	.13-08	-.83-09	.20-05	.19-05
	.10+01	.53-03	.20+01	.24-04	.10+01	-.12-06
8	.43-03	-.43-09	.19-04	-.76-13	.13-06	.45-06
	.19-06	-.48-09	.37-09	.10-14	.20-05	-.19-05
	.10+01	.27-03	.10+01	.56-09	.20+01	.20-04
9	.21-03	-.54-10	.45-09	.67-13	.16-04	.79-10
	.48-07	-.61-10	.68-13	-.12-13	.28-09	-.76-10
	.10+01	.13-03	.10+01	.28-09	.10+01	.44-09
10	.11-03	-.67-11	.22-09	-.71-23	.31-09	-.79-10
	.12-07	-.76-11	.52-19	-.39-22	.11-09	.76-10
	.10+01	.67-04	.20+01	.14-09	.20+01	.21-09
11	.54-04	-.84-12	.11-09	-.76-29	.17-09	.49-19
	.30-08	-.96-12	.13-19	-.87-29	.66-19	-.12-19
	.10+01	.33-04	.10+01	.19-19	.10+01	.73-19
12	.27-04	-.10-12	.16-19	.57-29	.34-19	-.49-19
	.74-09	-.12-12	.86-29	.65-29	.50-19	.12-19

first, third and fifth column : λ = step multiplier; σ = smallest pivot; $\|f\|$ = residual norm

second, fourth and sixth column : ξ (singular); ζ (nonsingular); η (nonsingular)

Table 2 , k = l = m , Extrapolation

First column: { j = number of steps; σ = smallest pivot; ‖f‖ = residual norm }

Top of (2+q)-th column Q-factor = $1/2^{q+1}$

(2+q)-th column: { $\xi^{(q)}$ (singular); $\zeta^{(q)}$ (nonsingular); $\eta^{(q)}$ (nonsingular) }

$$f^{(q)} \approx \begin{pmatrix} 0 \\ \zeta^{(q)} \\ \eta^{(q)} \end{pmatrix} \text{ for } q > 0$$

best extrapolant in each row: $\boxed{x^{(q)}}$

0	.10+00							
.30+00	.10+00							
.16+00	.10+00	1/4						
1	-.44-01	-.19+00						
-.23-01	.49-01	-.28-02						
.64-01	-.31-01	-.16+00	1/8					
2	.35-01	.11+00	.21+00					
.36-01	.89-02	-.31-01	-.40-01					
.96-02	.74-03	.32-01	.97-01	1/16				
3	.16-01	-.22-02	-.41-01	-.77-01				
.14-01	.21-03	-.85-02	-.11-02	.45-02				
.39-03	-.33-03	-.14-02	-.13-01	-.28-01	1/32			
4	.84-02	.41-03	.13-02	.73-02	.13-01			
.69-02	-.67-05	-.23-03	.25-02	.30-02	.29-02			
.45-04	.22-06	.33-03	.91-03	.28-02	.49-02	1/64		
5	.42-02	.65-04	-.49-04	-.24-03	-.74-03	-.12-02		
.34-02	-.20-06	.63-05	.84-04	-.27-03	-.49-03	-.60-03		
.12-04	-.17-06	-.56-06	-.11-03	-.26-03	-.46-03	-.64-03	1/128	
6	.21-02	.18-04	.18-05	.91-05	.26-04	.50-04	.70-04	
.17-02	-.27-07	.14-06	-.19-05	-.14-04	.26-05	.18-04	.28-04	
.30-05	-.29-07	.12-06	.34-06	.16-04	.34-04	.50-04	.61-04	1/256
7	.11-02	.45-05	.10-06	-.15-06	-.76-06	-.16-05	-.24-05	-.30-05
.86-03	-.34-08	.20-07	-.21-07	.25-06	.12-05	.12-05	.89-06	.68-06
.76-06	-.37-08	.21-07	-.10-07	-.60-07	-.11-05	-.23-05	-.31-05	-.36-05
8	.53-03	.11-05	.13-07	.74-09	.11-07	.36-07	.62-07	.81-07
.43-03	-.43-09	.26-08	-.33-08	-.69-09	-.17-07	-.57-07	-.76-07	-.84-07
.19-06	-.48-09	.28-08	-.34-08	-.25-08	.14-08	.38-07	.76-07	.10-06
9	.27-03	.29-06	.17-08	.30-10	-.18-10	-.36-09	-.93-09	-.14-08
.21-03	-.54-10	.32-09	-.43-09	-.15-10	.30-10	.59-09	.15-08	.21-08
.48-07	-.61-10	.36-09	-.45-09	-.27-10	.14-09	.94-10	-.52-09	-.11-08
10	.13-03	.71-07	.21-09	.15-11	-.37-12	.18-12	.59-11	.13-10
.11-03	-.67-11	.40-10	-.53-10	-.35-12	.64-12	-.30-12	-.97-11	-.22-10
.12-07	-.76-11	.45-10	-.59-10	-.28-11	-.11-11	-.55-11	-.71-11	-.31-11
11	.67-04	.18-07	.26-10	.94-13	-.39-15	.12-13	.90-14	-.37-13
.54-04	-.84-12	.50-11	-.67-11	-.20-13	.18-14	-.19-13	-.14-13	.62-13
.30-08	-.96-12	.57-11	-.75-11	-.18-12	-.80-14	.28-13	.12-12	.17-12
12	.33-04	.45-08	.33-11	.58-14	-.59-16	-.48-16	-.23-15	-.31-15
.27-04	-.10-12	.63-12	-.84-12	-.11-14	.13-15	.77-16	.38-15	.49-15
.74-09	-.12-12	.72-12	-.95-12	-.12-13	-.29-15	-.43-16	-.49-15	-.14-14
13	.17-04	.11-08	.41-12	.36-15	-.20-17	-.16-18	.60-18	.25-17
.13-04	-.13-13	.79-13	-.10-12	-.68-16	.43-17	.26-18	-.97-18	-.39-17
.19-09	-.15-13	.90-13	-.12-12	-.73-15	-.94-17	-.29-18	.39-18	.43-17
14	.84-05	.28-09	.52-13	.23-16	-.65-19	-.30-20	-.50-21	-.52-20
.67-05	-.16-14	.98-14	-.13-13	-.41-17	.14-18	.47-20	.76-21	.84-20
.47-10	-.19-14	.11-13	-.15-13	-.46-16	-.30-18	-.39-20	.70-21	-.24-20
15	.42-05	.70-10	.65-14	.14-17	-.20-20	-.30-23	.45-22	.49-22
.33-05	-.20-15	.12-14	-.16-14	-.25-18	.45-20	.77-22	.26-23	-.34-23
.12-10	-.23-15	.14-14	-.19-14	-.29-17	-.94-20	-.61-22	-.54-24	-.61-23

Table 3 , k = l = m

	QR-Bordering			LU-Bordering		
	.15-02	-.44-01	.10+01	.16-02	-.44-01	.11+01
1	-.23-01	.49-01	.37+00	-.23-01	.49-01	.41+00
	.64-01	-.31-01	.94+00	.64-01	-.31-01	.10+01
	.14-03	-.32-01	.95+00	.87-04	-.34-01	.11+01
2	-.22-01	.41-02	.71+00	-.23-01	.38-02	.77+00
	.49-02	-.22-02	.73+00	.48-02	-.24-02	.81+00
	.29-04	-.79-02	.71+00	.59-04	-.11-01	.70+00
3	-.60-02	.30-03	.75+00	-.86-02	.25-03	.83+00
	.41-03	.22-03	.70+00	.36-03	.14-03	.77+00
	.43-06	.11-02	.81+00	.76-06	.14-02	.79+00
4	.88-03	.23-04	.74+00	.12-02	.53-04	.83+00
	.48-04	.41-04	.70+00	.89-04	.71-04	.77+00
	.11-08	.51-04	.77+00	.21-08	.68-04	.75+00
5	.41-04	.46-06	.74+00	.55-04	.86-06	.82+00
	.54-06	.27-06	.70+00	.93-06	.36-06	.77+00
	.11-11	-.17-05	.79+00	.57-11	-.37-05	.79+00
6	-.13-05	.10-08	.74+00	-.30-05	.19-08	.83+00
	.14-08	.10-08	.70+00	.26-08	.18-08	.77+00
	.67-16	.13-07	.80+00	.59-15	.39-07	.80+00
7	.10-07	.98-12	.74+00	.31-07	.50-11	.83+00
	.15-11	.11-11	.70+00	.76-11	.57-11	.77+00
	.45-22	.11-10	.80+00	.43-21	.33-10	.80+00
8	.84-11	.58-16	.74+00	.26-10	.52-15	.83+00
	.89-16	.67-16	.70+00	.79-15	.59-15	.77+00
	.38-30	-.97-15	.80+00	.13-28	-.57-14	.80+00
9	-.78-15	.39-22	.74+00	-.45-14	.38-21	.83+00
	.59-22	.45-22	.70+00	.58-21	.43-21	.77+00
	.12-37	.71-20	.80+00	.19-36	.13-18	.80+00
10	.56-20	.33-30	.74+00	.11-18	.11-28	.83+00
	.50-30	.38-30	.70+00	.17-28	.13-28	.77+00
	.00	.37-27	.80+00	.00	.14-25	.80+00
11	.30-27	-.10-37	.74+00	.11-25	-.23-37	.83+00
	.15-37	-.10-37	.70+00	.39-37	-.31-37	.77+00

first and fourth column : μ = least squares residual; σ = only element of Σ; $\|f\|$ = residual norm

second and fifth column : ξ (singular); ζ (nonsingular); η (nonsingular)

third and sixth column : $\approx \partial\sigma/\partial\xi$; $\approx \partial\sigma/\partial\zeta$; $\approx \partial\sigma/\partial\eta$

Table 4 , nearly singular problem with $k = l = m$.

	Newton		three-point		two-point	
	.10+01	-.44-01	.10+01	-.44-01	.20+01	-.44-01
1	-.23-01	.49-01	-.23-01	.49-01	-.23-01	.49-01
	.64-01	-.31-01	.64-01	-.31-01	.64-01	-.31-01
	.10+01	.35-01	.20+01	.35-01	.10+01	.11+00
2	.36-01	.89-02	.36-01	.89-02	.85-01	-.31-01
	.96-02	.74-03	.96-02	.74-03	.39-01	.32-01
	.10+01	.16-01	.10+01	-.22-02	.20+01	.62-01
3	.14-01	.21-03	-.61-02	-.85-02	.62-01	.61-02
	.39-03	-.33-03	.85-02	-.14-02	.77-02	-.27-02
	.10+01	.84-02	.10+01	.20-01	.10+01	.56-02
4	.69-02	-.67-05	.17-01	.23-03	.12-02	-.63-02
	.45-04	.23-06	.45-03	-.39-03	.65-02	.20-02
	.10+01	.42-02	.20+01	.10-01	.20+01	-.33-01
5	.34-02	-.20-06	.82-02	-.97-05	-.24-01	.24-03
	.12-04	-.17-06	.63-04	.26-06	.12-02	.44-03
	.10+01	.21-02	.10+01	.91-04	.10+01	.14-02
6	.17-02	-.25-07	.78-04	.91-05	.98-03	-.24-03
	.30-05	-.27-07	.91-05	-.83-06	.43-03	-.36-03
	.10+01	.11-02	.10+01	.46-04	.20+01	.38-03
7	.86-03	-.25-08	.38-04	.14-09	.30-03	-.44-06
	.76-06	-.27-08	.13-08	-.78-09	.20-05	.19-05
	.10+01	.53-03	.20+01	.22-04	.10+01	-.14-05
8	.43-03	.38-10	.19-04	.20-10	.13-06	.45-06
	.19-06	.54-10	.37-09	.22-10	.20-05	-.19-05
	.10+01	.27-03	.10+01	-.12-05	.20+01	.23-04
9	.21-03	.18-09	.53-07	-.97-12	.19-04	.12-09
	.48-07	.21-09	.10-11	-.12-11	.40-09	-.69-10
	.10+01	.13-03	.10+01	.11-04	.10+01	-.12-05
10	.11-03	.11-09	.94-05	.92-11	.52-07	-.96-10
	.12-07	.12-09	.91-10	.11-10	.13-09	.91-10
	.10+01	.66-04	.20+01	.47-05	.20+01	.11-04
11	.54-04	.57-10	.48-05	.41-11	.96-05	.94-11
	.30-08	.65-10	.22-10	.47-11	.94-10	.11-10
	.10+01	.32-04	.10+01	-.99-06	.10+01	-.11-05
12	.27-04	.28-10	.21-06	-.86-12	.10-06	-.98-12
	.74-09	.32-10	.99-12	-.99-12	.10-11	-.11-11
	.10+01	.16-04	.10+01	.19-05	.20+01	.48-05
13	.13-04	.14-10	.25-05	.16-11	.48-05	.42-11
	.19-09	.15-10	.54-11	.19-11	.23-10	.48-11
	.10+01	.72-05	.20+01	.55-06	.10+01	-.99-06
14	.67-05	.63-11	.14-05	.48-12	.21-06	-.87-12
	.46-10	.72-11	.11-11	.55-12	.10-11	-.99-12
	.10+01	.31-05	.10+01	-.38-06	.20+01	.19-05
15	.34-05	.27-11	.69-06	-.34-12	.25-05	.17-11
	.11-10	.31-11	.54-12	-.38-12	.56-11	.19-11

Entries as in Table 1.

Table 5 , k = 1 , m = 2

	three-point		two-point		one-point	
	.10+01	.70-01	.20+01	.70-01	.13+01	.70-01
1	.17+00	.70-01	.17+00	.70-01	.17+00	.70-01
	.11-01	.93-02	.11-01	.93-02	.11-01	.93-02
	.20+01	.34-01	.10+01	-.16-02	.13+01	.23-01
2	.80-01	.39-01	-.13-02	.70-02	.54-01	.29-01
	.34-02	.23-03	.89-02	-.88-02	.39-02	-.25-02
	.10+01	-.73-03	.20+01	-.83-01	.13+01	.90-02
3	-.67-03	.89-03	-.12+00	-.51-01	.22-01	.98-02
	.21-03	-.21-03	.90-02	-.27-03	.64-03	.76-03
	.10+01	-.97-03	.10+01	.30-02	.13+01	.30-02
4	-.67-03	.40-03	.18-02	-.35-03	.71-02	.38-02
	.43-06	-.14-05	.20-03	.20-03	.25-03	-.23-03
	.20+01	-.49-03	.20+01	.19-02	.13+01	.11-02
5	-.33-03	.20-03	.12-02	-.48-03	.28-02	.12-02
	.30-06	-.80-08	.12-05	-.34-05	.65-04	.68-04
	.10+01	-.31-07	.10+01	-.13-04	.13+01	.38-03
6	.59-07	.12-06	-.27-07	.72-05	.92-03	.46-03
	.79-08	.79-08	.35-05	.35-05	.21-04	-.20-04
	.10+01	-.35-07	.20+01	.19-02	.13+01	.14-03
7	-.42-09	.59-07	.11-02	-.35-03	.35-03	.16-03
	.66-14	-.43-14	.32-05	.75-07	.61-05	.61-05
	.20+01	-.18-07	.10+01	-.35-06	.13+01	.48-04
8	-.21-09	.30-07	.14-05	.27-05	.12-03	.57-04
	.20-14	-.11-18	.64-07	-.64-07	.18-05	-.18-05
	.10+01	.12-12	.20+01	-.16-07	.13+01	.18-04
9	.28-14	.16-14	.97-06	.14-05	.43-04	.19-04
	.12-18	.12-18	.34-11	.69-12	.55-06	.55-06
	.10+01	.59-13	.10+01	-.61-11	.13+01	.60-05
10	.14-14	.78-15	-.58-11	.13-11	.15-04	.69-05
	.23-26	.13-29	.69-12	-.69-12	.16-06	-.16-06
	.20+01	.29-13	.20+01	-.39-11	.13+01	.22-05
11	.72-15	.39-15	-.26-11	.11-11	.52-05	.24-05
	.56-27	-.69-31	.55-23	-.16-22	.49-07	.49-07
	.10+01	.11-19	.10+01	.13-18	.13+01	.75-06
12	-.17-22	-.19-19	.68-19	-.64-19	.18-05	.85-06
	.68-31	.68-31	.16-22	.16-22	.15-07	-.15-07
	.10+01	.18-16	.20+01	.32-19	.13+01	.27-06
13	-.23-18	.23-18	.14-19	-.22-19	.64-06	.29-06
	.21-33	-.81-36	.13-37	.11-37	.44-08	.44-08

Entries as in Table 1 with ζ now a singular variable.
The one-point method is partially corrected with $\lambda = 1.3 < 1 + k/(k+2)$.

First column: { j = number of steps; σ = smallest pivot; ‖f‖ = residual norm

Top of (2+q)-th column Q-factor = $1/2^{q+1}$

(2+q)-th column: { $\xi^{(q)}$ (singular); $\zeta^{(q)}$ (singular); $\eta^{(q)}$ (nonsingular)

$f^{(q)} \approx \begin{pmatrix} 0 \\ 0 \\ \eta^{(q)} \end{pmatrix}$ for $q > 0$

best extrapolant in each row $\boxed{x^{(q)}}$

0	.10+00							
.43+00	.10+00							
.95-01	.10+00	1/4						
1	.70-01	.39-01						
.17+00	.70-01	.41-01						
.11-01	.93-02	-.81-01	1/8					
2	.34-01	-.16-02	-.15-01					
.80-01	.39-01	.70-02	-.43-02					
.34-02	.23-03	-.88-02	.15-01	1/16				
3	.17-01	-.73-03	-.44-03	.17-02				
.40-01	.20-01	.89-03	-.12-02	-.71-03				
.89-03	.11-04	-.21-03	.27-02	.86-03	1/32			
4	.82-02	-.21-03	-.41-04	.17-04	-.94-04			
.20-01	.10-01	.22-03	.38-06	.17-03	.22-03			
.23-03	.89-06	-.89-05	.58-04	-.31-03	-.39-03	1/64		
5	.41-02	-.59-04	-.79-05	-.32-05	-.46-05	-.17-05		
.10-01	.50-02	.58-04	.31-05	.35-05	-.74-05	-.15-04		
.57-04	.78-07	-.73-06	.20-05	-.60-05	.15-04	.28-04	1/128	$\boxed{x^{(q)}}$
6	.20-02	-.16-04	-.11-05	-.16-06	.46-07	.19-06	.22-06	
.50-02	.25-02	.15-04	.47-06	.95-07	-.13-06	.10-06	.34-06	
.14-04	.72-08	-.63-07	.16-06	-.10-06	.29-06	-.17-06	-.61-06	1/256
7	.10-02	-.40-05	-.15-06	-.12-07	-.26-08	-.42-08	-.73-08	-.92-08
.25-02	.13-02	.38-05	.65-07	.70-08	.11-08	.53-08	.38-08	.11-08
.36-05	.74-09	-.58-08	.13-07	-.77-08	-.14-08	-.11-07	-.85-08	-.37-08
8	.51-03	-.10-05	-.20-07	-.90-09	-.14-09	-.57-10	.87-11	.66-10
.13-02	.63-03	.94-06	.85-08	.50-09	.71-10	.37-10	-.47-10	-.77-10
.89-06	.82-10	-.58-09	.12-08	-.59-09	-.11-09	-.70-10	.10-09	.17-09
9	.25-03	-.26-06	-.25-08	-.61-10	-.52-11	-.94-12	-.55-13	-.12-12
.63-03	.32-03	.24-06	.11-08	.34-10	.27-11	.46-12	-.13-12	.24-12
.22-06	.95-11	-.63-10	.11-09	-.41-10	-.43-11	-.81-12	.29-12	-.51-12
10	.13-03	-.64-07	-.32-09	-.40-11	-.18-12	-.18-13	-.37-14	-.33-14
.31-03	.16-03	.59-07	.14-09	.22-11	.92-13	.90-14	.18-14	.29-14
.56-07	.11-11	-.72-11	.11-10	-.27-11	-.15-12	-.16-13	-.34-14	-.57-14
11	.63-04	-.16-07	-.40-10	-.26-12	-.59-14	-.33-15	-.46-16	-.17-16
.16-03	.79-04	.15-07	.17-10	.14-12	.30-14	.16-15	.23-16	.84-17
.14-07	.14-12	-.86-12	.13-11	-.17-12	-.50-14	-.29-15	-.45-16	-.18-16
12	.32-04	-.40-08	-.50-11	-.16-13	-.19-15	-.51-17	.52-19	.41-18
.78-04	.40-04	.37-08	.22-11	.89-14	.97-16	.30-17	.43-18	.25-18
.35-08	.17-13	-.11-12	.15-12	-.11-13	-.16-15	-.50-17	-.43-18	-.79-19
13	.16-04	-.10-08	-.63-12	-.10-14	-.67-17	-.74-18	-.67-18	-.68-18
.39-04	.20-04	.93-09	.27-12	.56-15	.28-17	-.24-18	-.29-18	-.30-18
.87-09	.22-14	-.13-13	.18-13	-.69-15	-.51-17	-.82-19	-.39-20	-.60-21
14	.79-05	-.25-09	-.79-13	-.64-16	-.10-17	-.82-18	-.82-18	-.82-18
.20-04	.99-05	.23-09	.34-13	.35-16	-.26-18	-.36-18	-.36-18	-.36-18
.22-09	.27-15	-.16-14	.22-14	-.43-16	-.16-18	-.13-20	-.15-22	.15-22
15	.40-05	-.63-10	-.99-14	-.59-17	-.20-17	-.20-17	-.20-17	-.20-17
.98-05	.50-05	.58-10	.43-14	.14-17	-.85-18	-.87-18	-.88-18	-.88-18
.54-10	.34-16	-.20-15	.27-15	-.27-17	-.50-20	-.43-22	-.23-22	-.23-22

Table 7 , k = 1 , m = 2 , QR-Bordering

	.24+00	.70-01	-.84+00	.18+00	-.29-02	-.90+00
1	.34+00	.70-01	-.82+00	.17+00	-.28-02	-.88+00
	.11-01	.93-02	-.25+01	.53+00	-.86-02	-.27+01
	.78-01	-.24-01	-.52+00	.85+00	-.13+01	-.72+00
2	.98-01	.26-01	-.66+00	.50+00	-.60+00	-.79+00
	.97-02	-.10-01	-.24+01	.67+00	-.27+00	-.27+01
	.10-01	-.43-01	-.46+00	.80+00	-.17+01	-.14+01
3	.60-01	-.99-02	-.55+00	.40+00	-.15+01	-.21+01
	.28-02	-.22-02	-.25+01	.69+00	-.67-01	-.24+01
	.11-01	.14-01	-.44+00	.47+00	-.78+00	.73-02
4	.18+00	-.59-01	-.57+00	.69+00	-.23+01	-.33+01
	.13-01	.12-01	-.25+01	.61+00	.17+00	-.20+01
	.40-02	.40-02	-.40+00	.45+00	-.92+00	-.95-01
5	.38-01	-.12-01	-.74+00	.78+00	-.17+01	-.28+01
	.61-02	.62-02	-.24+01	.60+00	.94-01	-.21+01
	.14-02	.34-02	-.41+00	.44+00	-.89+00	-.88-01
6	.12-01	.32-02	-.69+00	.85+00	-.23+01	-.30+01
	.24-02	-.23-02	-.25+01	.56+00	.43+00	-.20+01
	.31-03	-.21-02	-.67+00	.77+00	-.57+00	.36+00
7	.43-02	.10-02	-.79+00	.98+00	-.22+01	-.28+01
	.50-03	-.50-03	-.24+01	.45+00	.32+00	-.21+01
	.16-05	.42-03	-.78+00	.90+00	-.37+00	.56+00
8	.10-02	-.26-03	-.74+00	.91+00	-.23+01	-.29+01
	.12-03	.12-03	-.24+01	.48+00	.36+00	-.21+01
	.56-06	-.11-04	-.79+00	.91+00	-.39+00	.57+00
9	.15-04	-.14-05	-.73+00	.90+00	-.23+01	-.29+01
	.15-05	.15-05	-.24+01	.48+00	.36+00	-.21+01
	.37-07	-.10-05	-.72+00	.82+00	-.46+00	.48+00
10	.17-05	.34-06	-.72+00	.89+00	-.23+01	-.29+01
	.18-06	-.18-06	-.24+01	.50+00	.37+00	-.21+01
	.68-09	.20-07	-.73+00	.84+00	-.39+00	.53+00
11	.97-07	-.28-07	-.72+00	.88+00	-.23+01	-.30+01
	.13-07	.13-07	-.24+01	.50+00	.38+00	-.21+01
	.65-10	-.34-08	-.77+00	.89+00	-.40+00	.55+00
12	.43-08	.92-10	-.67+00	.82+00	-.23+01	-.30+01
	.20-09	-.20-09	-.25+01	.53+00	.38+00	-.21+01
	.17-12	-.38-09	-.68+00	.78+00	-.37+00	.50+00
13	.47-09	.21-11	-.67+00	.82+00	-.23+01	-.30+01
	.19-10	-.19-10	-.24+01	.52+00	.38+00	-.21+01
	.53-16	-.10-11	-.68+00	.78+00	-.37+00	.50+00
14	.13-11	.19-14	-.67+00	.82+00	-.23+01	-.30+01
	.49-13	-.49-13	-.24+01	.52+00	.38+00	-.21+01
	.64-18	-.60-15	-.68+00	.78+00	-.37+00	.50+00
15	.76-15	-.40-16	-.67+00	.82+00	-.23+01	-.30+01
	.12-16	-.12-16	-.24+01	.52+00	.38+00	-.21+01

First two columns as in Table 3 with second element in first column now Frobenius norm of Σ . Other four columns = approximate gradients of the four elements of Σ .

Table 8 , k = 2 , m = 1

	two-point		one-point		five-point	
	.28+01	-.48-01	.14+01	-.48-01	.10+01	-.48-01
1	.17-02	.50-01	.17-02	.50-01	.17-02	.50-01
	.67-01	-.32-01	.67-01	-.32-01	.67-01	-.32-01
	.10+01	-.14+01	.14+01	-.72+00	.10+01	-.53+00
2	.28+01	-.19+00	.14+01	-.69-01	.74+00	-.35-01
	.24+01	-.14-01	.35+00	-.23-01	.15+00	-.26-01
	.28+01	-.92+00	.14+01	-.38+00	.10+01	-.35+00
3	.17+01	-.19+00	.38+00	-.68-01	.32+00	-.37-01
	.69+00	.24-01	.48-01	.73-03	.40-01	-.94-02
	.10+01	-.61-01	.14+01	-.20+00	.30+01	-.23+00
4	.99-02	-.21-01	.11+00	-.70-02	.14+00	-.19-01
	.26-01	-.17-01	.17-01	-.14-01	.12-01	-.79-02
	.28+01	-.35-01	.14+01	-.11+00	.10+01	-.25-02
5	.30-02	.10-02	.29-01	-.41-02	-.14-04	.14-01
	.19-02	-.15-02	.31-02	-.70-03	.14-01	.22-03
	.10+01	-.19-02	.14+01	-.58-01	.10+01	.12+01
6	.13-04	-.22-02	.83-02	-.63-03	.10+01	-.41-02
	.31-02	.22-02	.88-03	-.15-02	.15+01	.35-01
	.28+01	.75-02	.14+01	-.31-01	.10+01	.72+00
7	.14-03	-.97-05	.23-02	-.30-03	-.19+00	.31+00
	.42-04	.18-04	.35-03	-.19-04	.54+00	-.29+00
	.10+01	.50-03	.14+01	-.17-01	.10+01	.40+00
8	.61-06	-.62-06	.67-03	-.49-04	-.24+00	.98-01
	.54-04	-.54-04	.79-04	-.16-03	.12+00	-.14+00
	.28+01	.34-03	.14+01	-.89-02	.30+01	.25+00
9	.27-06	-.45-07	.19-03	-.25-04	.16+00	.84-02
	.36-07	-.91-08	.43-04	.12-04	.18-01	-.46-01
	.10+01	.22-04	.14+01	-.47-02	.10+01	.44-02
10	.12-08	.45-07	.54-04	-.32-05	.29-02	-.36-01
	.52-07	-.25-07	.11-04	-.19-04	.59-01	.53-01
	.28+01	.15-04	.14+01	-.25-02	.10+01	.36+00
11	.53-09	-.58-10	.15-04	-.23-05	.32+00	-.36-01
	.29-10	-.67-10	.60-05	.34-05	.89-01	.23-01
	.10+01	.99-06	.14+01	-.13-02	.10+01	.23+00
12	.24-11	.33-10	.43-05	-.12-06	.13+00	.19-01
	.50-10	.37-10	.19-05	-.25-05	.27-01	-.37-01
	.28+01	.66-06	.14+01	-.72-03	.10+01	.16+00
13	.11-11	-.11-12	.12-05	-.25-06	.59-01	-.32-02
	.58-13	-.13-12	.89-06	.68-06	.41-02	-.14-01
	.10+01	.44-07	.14+01	-.38-03	.30+01	.11+00
14	.47-14	.64-13	.35-06	.14-07	.26-01	-.31-02
	.98-13	.73-13	.31-06	-.37-06	.28-02	-.34-02

Entries as in Table 1. Two-point method partially corrected with $\lambda = 2.8 < k+1$. One-point method partially corrected with $\lambda = 1.4 < 1+k/(k+2)$. Five-point method fully corrected with $\lambda = k+1$.

First column: { j = number of steps; σ = smallest pivot; ‖f‖ = residual norm }

Top of (2+q)-th column Q-factor = $(2/3)^{q+1}$

(2+q)-th column: { $\xi^{(q)}$ (singular); $\zeta^{(q)}$ (nonsingular); $\eta^{(q)}$ (nonsingular) }

$f^{(q)} \approx \begin{pmatrix} 0 \\ \zeta^{(q)} \\ \eta^{(q)} \end{pmatrix}$ for $q > 0$

best extrapolant in each row $\boxed{x^{(q)}}$

0	.10+00							
.63-01	.10+00							
.16+00	.10+00	$\frac{4}{9}$						
1	-.48-01	-.35+00						
.17-02	.50-01	-.50-01						
.67-01	-.32-01	-.30+00	$\frac{8}{27}$					
2	-.53+00	-.15+01	-.24+01					
.74+00	-.35-01	-.21+00	-.33+00					
.15+00	-.26-01	-.12-01	.22+00	$\frac{16}{81}$				
3	-.35+00	.75-02	.12+01	.27+01				
.32+00	-.37-01	-.40-01	.92-01	.27+00				
.40-01	-.94-02	.23-01	.51-01	-.18-01	$\frac{32}{243}$			
4	-.23+00	-.27-02	-.11-01	-.52+00	-.13+01			
.14+00	-.19-01	.17-01	.63-01	.51-01	-.35-02			
.12-01	-.79-02	-.49-02	-.27-01	-.60-01	-.71-01	$\frac{64}{729}$		
5	-.16+00	-.25-02	-.23-02	.13-02	.13+00	.35+00		
.62-01	-.79-02	.14-01	.12-01	-.94-02	-.24-01	-.27-01		
.42-02	-.52-02	.22-03	.43-02	.18-01	.37-01	.53-01	$\frac{128}{2187}$	
6	-.10+00	-.12-02	-.86-04	.86-03	.76-03	-.19-01	-.54-01	
.27-01	-.33-02	.60-02	-.52-03	-.58-02	-.49-02	-.20-02	.49-03	
.16-02	-.27-02	.22-02	.38-02	.36-02	.11-03	-.55-02	-.11-01	$\frac{256}{6561}$
7	-.70-01	-.47-03	.74-04	.14-03	-.35-04	-.16-03	.16-02	.51-02
.12-01	-.14-02	.24-02	-.56-03	-.58-03	.70-03	.15-02	.19-02	.20-02
.65-03	-.13-02	.15-02	.96-03	-.24-03	-.12-02	-.14-02	-.97-03	-.34-03
8	-.47-01	-.20-03	.21-04	-.10-05	-.36-04	-.36-04	-.25-04	-.13-03
.53-02	-.60-03	.97-03	-.14-03	.40-04	.19-03	.11-03	-.23-04	-.14-03
.28-03	-.61-03	.79-03	.21-03	-.11-03	-.80-04	.86-04	.23-03	.30-03
9	-.31-01	-.86-04	.52-05	-.16-05	-.17-05	.35-05	.73-05	.93-05
.24-02	-.26-03	.41-03	-.32-04	.13-04	.62-05	-.22-04	-.35-04	-.36-04
.12-03	-.28-03	.38-03	.54-04	-.10-04	.14-04	.29-04	.23-04	.11-04
10	-.21-01	-.37-04	.15-05	-.88-07	.29-06	.59-06	.32-06	-.12-06
.10-02	-.12-03	.18-03	-.85-05	.12-05	-.16-05	-.28-05	-.98-06	.11-05
.55-04	-.13-03	.18-03	.16-04	-.40-06	.20-05	.18-06	-.26-05	-.42-05
11	-.14-01	-.16-04	.44-06	.11-08	.23-07	-.17-07	-.76-07	-.10-06
.46-03	-.51-04	.78-04	-.24-05	.14-06	-.13-06	.98-07	.38-06	.46-06
.25-04	-.56-04	.82-04	.46-05	-.63-07	.19-07	-.29-06	-.33-06	-.19-06
12	-.93-02	-.72-05	.13-06	.69-10	-.19-09	-.37-08	-.24-08	.21-08
.21-03	-.23-04	.34-04	-.70-06	.26-07	-.12-08	.19-07	.11-07	-.12-07
.11-04	-.25-04	.37-04	.14-05	-.19-07	-.81-08	-.12-07	.14-07	.36-07
13	-.62-02	-.32-05	.38-07	-.31-10	-.56-10	-.35-10	.32-09	.49-09
.91-04	-.10-04	.15-04	-.21-06	.51-08	.10-10	.20-09	-.16-08	-.23-08
.50-05	-.11-04	.17-04	.40-06	-.43-08	-.63-09	.51-09	.17-08	.96-09
14	-.41-02	-.14-05	.11-07	-.70-11	-.11-11	.72-11	.11-10	-.78-11
.41-04	-.45-05	.67-05	-.60-07	.98-09	-.27-10	-.33-10	-.55-10	.40-10
.22-05	-.50-05	.75-05	.12-06	-.88-09	-.47-10	.41-10	-.38-11	-.11-09
15	-.27-02	-.62-06	.34-08	-.12-11	.21-12	.41-12	-.25-12	-.96-12
.18-04	-.20-05	.30-05	-.18-07	.19-09	-.51-11	-.18-11	.12-11	.47-11
.99-06	-.22-05	.33-05	.35-07	-.18-09	-.49-11	.16-11	-.22-11	-.21-11

First column: { j = number of steps; σ = smallest pivot; ‖f‖ = residual norm

Top of (2+q)-th column Q-factor = $2/3^{q+1}$

(2+q)-th column: { $\zeta^{(q)}$ (singular); $\zeta^{(q)}$ (singular); $\eta^{(q)}$ (nonsingular)

$f^{(q)} \approx \begin{pmatrix} 0 \\ 0 \\ \eta^{(q)} \end{pmatrix}$ for q > 0

best extrapolant in each row $x^{(q)}$

0	.10+00							
.32-02	.10+00							
.84-01	.10+00	4/9						
1	-.33+00	-.12+01						
.28-01	-.21-01	-.26+00						
.20+00	-.13+00	-.60+00	8/27					
2	-.23+00	-.17-01	.93+00					
.20-01	.41-04	.42-01	.29+00					
.20-01	-.25-01	.19+00	.82+00	16/81				
3	-.15+00	-.25-03	.13-01	-.37+00				
.91-02	.26-04	-.29-05	-.34-01	-.17+00				
.29-02	-.73-02	.28-01	-.10+00	-.49+00	32/243			
4	-.10+00	-.10-04	.18-03	-.52-02	.85-01			
.41-02	.17-04	-.25-06	.19-05	.14-01	.59-01			
.95-03	-.29-02	.60-02	-.11-01	.27-01	.15+00	64/729		
5	-.67-01	-.12-05	.62-05	-.67-04	.12-02	-.12-01		
.18-02	.12-04	-.48-07	.12-06	-.61-06	-.35-02	-.13-01		
.40-03	-.13-02	.19-02	-.13-02	.30-02	-.29-02	-.27-01	128/2187	
6	-.45-01	-.23-06	.55-06	-.18-05	.14-04	-.17-03	.93-03	
.81-03	.77-05	-.14-07	.14-07	-.29-07	.11-06	.53-03	.18-02	
.18-03	-.58-03	.82-03	-.65-04	.45-03	-.17-03	.24-03	.28-02	256/6561
7	-.30-01	-.47-07	.95-07	-.96-07	.33-06	-.18-05	.14-04	-.43-04
.36-03	.51-05	-.42-08	.33-08	-.12-08	.57-08	-.11-07	-.51-04	-.17-03
.81-04	-.26-03	.38-03	.16-04	.50-04	-.49-04	-.31-04	-.57-04	-.24-03
8	-.20-01	-.10-07	.20-07	-.11-07	.94-08	-.40-07	.13-06	-.74-06
.16-03	.34-05	-.13-08	.10-08	.62-10	.37-09	-.43-09	.57-09	.32-05
.37-04	-.12-03	.17-03	.74-05	.37-05	-.78-05	-.15-05	.13-05	.49-05
9	-.13-01	-.21-08	.43-08	-.24-08	-.16-09	-.16-08	.21-08	-.57-08
.71-04	.23-05	-.39-09	.32-09	.28-10	.20-10	-.33-10	.48-11	-.31-10
.17-04	-.53-04	.77-04	.23-05	.15-06	-.73-06	.34-06	.52-06	.47-06
10	-.89-02	-.42-09	.89-09	-.53-09	-.76-10	-.62-10	.86-10	-.38-10
.32-04	.15-05	-.12-09	.10-09	.62-11	.72-12	-.22-11	.82-12	.57-12
.76-05	-.23-04	.35-04	.68-06	-.48-08	-.43-07	.61-07	.35-07	.49-08
11	-.59-02	-.86-10	.18-09	-.11-09	-.12-10	-.18-11	.40-11	-.11-11
.14-04	.10-05	-.35-10	.31-10	.12-11	.16-13	-.91-13	.11-12	.67-13
.34-05	-.10-04	.16-04	.20-06	-.22-08	-.15-08	.47-08	-.74-09	-.29-08
12	-.39-02	-.18-10	.37-10	-.24-10	-.16-11	-.47-13	.12-12	-.12-12
.62-05	.67-06	-.10-10	.92-11	.24-12	.73-16	-.23-14	.62-14	-.31-15
.15-05	-.47-05	.69-05	.59-07	-.45-09	-.25-10	.21-09	-.23-09	-.20-09
13	-.26-02	-.36-11	.75-11	-.49-11	-.21-12	-.21-14	.22-14	-.51-14
.28-05	.45-06	-.31-11	.28-11	.48-13	-.14-16	-.28-16	.19-15	-.18-15
.68-06	-.21-05	.31-05	.17-07	-.89-10	.14-11	.54-11	-.14-10	-.62-12
14	-.18-02	-.76-12	.15-11	-.99-12	-.27-13	-.16-15	.24-16	-.11-15
.12-05	.30-06	-.93-12	.83-12	.95-14	-.12-17	.74-18	.35-17	-.82-17
.31-06	-.92-06	.14-05	.51-08	-.17-10	.25-12	.74-13	-.44-12	.39-12
15	-.12-02	-.16-12	.32-12	-.20-12	-.36-14	-.14-16	.59-19	-.15-17
.55-06	.20-06	-.28-12	.25-12	.19-14	-.79-19	.97-19	.35-19	-.18-18
.14-06	-.41-06	.61-06	.15-08	-.34-11	.32-13	-.92-15	-.82-14	.19-13

INDEX

of frequently used symbols and expressions some of which may have a different meaning within certain sections.

Integers

n = number of variables and equations 3

m = dimension of nullspace N of $\nabla f(x^*)$ 3

p = order of determinant function 6

k = order of singularity 21

$\hat{i}$ = degree of singularity 21

Scalar and Vector Functions

δ = determinant function 3

π_0 = leading term in expansion of δ 6

$\bar{r}$ = boundary function of R' 9

$\tau^*(A)$ = upper outer density of set A at x^* 15

$\sigma(x)$ = smallest singular value of $\nabla f(x)$ 18

g = Newtonian iteration function 20

g_1 = leading term in expansion of g 21

$\theta(t)$ = minimal angle between t and N 45

$\nu(t)$ = smallest singular value of $\bar{B}(t)$ 50

$\phi(t)$ = $\frac{1}{2}$ minimal angle between t and $S \cap \pi_0^{-1}(0)$ 59

r = boundary function of R 68

π = homogeneous polynomial 69

Matrices and Matrix Functions

P = orthogonal projection onto nullspace of $\nabla f^T(x^*)$ 5

B, C, D, E = submatrices of Jacobian in normal form 5

G = reduced Jacobian 5

$\bar{B}$ = leading term in expansion of B 48

$\bar{C}$ = leading term in expansion of C 48

Sets and Spaces

$\delta^{-1}(0)$ = singular set 2

N = nullspace of $\nabla f(x^*)$ 3

$B_{\bar{\rho}}$ = ball with radius $\bar{\rho}$ about x^* 5

R' : starlike domain of invertibility 9

X_0 = full domain of convergence 29

$W \subset X_0$: starlike domain of convergence 62

$R \subset X_0$: starlike domain of convergence 68

Variables

- singular 5
- nonsingular 5

Equations

- singular 5
- nonsingular 5

Directions

- tangential 10
- excluded 11
- included 14
- irregular 17
- regular 17

Domains

- starlike 10
- of invertibility 14
- of convergence 30
- of bounded conv. 34
- of contraction 40

Singularities

- balanced 48
- regular 54
- strongly regular 64

Newton sequences

- approximate 63
- regular 90

BIBLIOGRAPHY

[1] J.F. Traub, *Iterative Methods for the Solution of Equations*, Prentice-Hall, 1964.

[2] E. Isaacson and H.B. Keller, *Analysis of Numerical Methods*, John Wiley & Sons, 1966.

[3] L.B. Rall, "Convergence of the Newton process to multiple solutions", *Num. Math.* Vol. 9, pp.23-37, 1966.

[4] R.C. Cavanagh, *Difference Equations and Iterative Methods*, Ph.D. Diss., University of Maryland, College, Park, Maryland, 1970.

[5] G.W. Reddien, "On Newton's method for singular problems", *SINUM*, Vol. 15, pp.993-996, 1978.

[6] A.S. Householder, *The Theory of Matrices in Numerical Analysis*, Blaisdell, 1964.

[7] G.W. Reddien, "Newton's method and high order singularities", *Computers and Mathematics with Applications*, Vol. 5, pp.79-86, 1979.

[8] D.W. Decker and C.T. Kelley, "Newton's method at singular points I", *SINUM*, Vol. 17, pp.66-70, 1980.

[9] D.W. Decker and C.T. Kelley, "Newton's method at singular points II", to appear.

[10] J.M. Ortega and W.C. Rheinbold, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, 1970.

[11] G.W. Stewart, *Introduction to Matrix Computation*, Academic Press, 1973.

[12] P. Deuflehard and G. Heindl, "Affine invariant theorem for Newton's method and extensions to related methods", *SINUM*, Vol. 16, pp.1-10, 1979.

[13] A.A. Goldstein, *Constructive Real Analysis*, Harper & Row, 1967.

[14] R. Narasimhan, *Analysis of Real and Complex Manifolds*, North-Holland, 1968.

[15] H.L. Royden, *Real Analysis*, Collier Macmillan, 1968.

[16] M.E. Munroe, *Introduction to Measure and Integration*, Addison-Wesley, 1953.

[17] H. Federer, *Geometric Measure Theory*, Springer Verlag, 1969.

[18] K. Levenberg, "A method for the solution of certain non-linear problems in least squares", *Quart. Appl. Math.*, Vol. 2, pp.164-168, 1944.

[19] R. Fletcher, "Generalized inverses for nonlinear equations and optimization", in *Numerical Methods of Nonlinear Algebraic Equations*, P.H. Rabinowitz, Editor, Gordon and Breach, 1970.

[20] D.C. Sorenson, *Updating the Symmetric Indefinite Factorization with Applications in a Modified Newton's Method*, ANL-77-49, Argonne National Laboratory, Illinois, 1977.

[21] D. Gay, "Modifying singular values: existence of solutions of systems of nonlinear equations having a possibly singular Jacobian matrix", *Math. of Comp.*, Vol. 31, pp.962-973, 1977.

[22] S. Barnett and C. Storey, *Matrix Methods in Stability Theory*, Nelson, 1970.

[23] A.A. Goldstein, "On Newton's method", *Numer. Math.*, Vol. 7, pp.391-393. 1965.

[24] A. Griewank and M.R. Osborne, "On Newton's method for singular problems", submitted for publication in *SINUM*.

[25] G.H. Golub and W. Kahan, "Calculating the singular values and pseudo-inverse of a matrix", *SINUM*, Vol. 2, pp.202-224.

[26] E. Schroeder, "Ueber unendlich viele Algorithmen zur Auflösung der Gleichungen", *Math. Ann.*, Vol. 2, pp.317-365, 1870.

[27] R.P. Brent, "Some efficient algorithms for solving systems of nonlinear equations", *SINUM*, Vol. 10, pp.327-343, 1973.

[28] H.B. Keller, private communication.

[29] A.M. Ostrowski, *Solution of Equations and Systems of Equations*, Academic Press, 1966.

[30] R.F. King, "A family of fourth order methods for nonlinear equations", *SINUM*, Vol. 10, pp.876-879, 1973.

[31] D.A. Smith and W.T. Ford, "Acceleration of linear and logarithmic convergence", *SINUM*, Vol. 16, 1979.

[32] D. Shanks, "Nonlinear transformations of divergent and slowly convergent sequences", *J. Math. Phy.*, Vol. 34, pp.1-42, 1955.

[33] D.C. Joyce, "Survey of extrapolation processes in numerical analysis", *SIAM Review*, Vol. 13, pp.435-489, 1971.

[34] R. Bulirsch, "Bemerkungen zur Romberg Integration", *Num. Math.*, Vol. 6, pp.6-16, 1964.

[35] G.H. Hardy and E.M. Wright, *An Introduction to the Theory of Numbers*, Oxford, at the Clarendon Press, 1959.

[36] P.J. Laurent, "Une theoreme de convergence pour le procede d'extrapolation de Richardson", *Comptes Rendus, Acad. Sci. Paris*, Ser A-B, 256, 1435-1437.

[37] I.N. Bronstein and K.A. Semendjajev, *Taschenbuch der Mathematik*, B.G. Teubner Verlagsgesellschaft, Leipzig, 1968.

[38] H.B. Keller, "Numerical solution of Bifurcation and nonlinear eigenvalue problems", in *Applications of Bifurcation Theory*, P.H. Rabinowitz, editor, Academic Press, 1976.

[39] W.C. Rheinboldt, "Numerical continuation methods for finite element applications" *U.S.-German Symposium on Finite Elements*, J. Bathe, editor, MIT Press, 1979.

[40] C. Davis, "The rotation of eigenvectors by a perturbation II", *J. Math. Anal. and Appl.*, Vol. 11, pp.20-27, 1965.

[41] M.G. Crandall and P.H. Rabinowitz, "Bifurcation from simple eigenvalues",*J. Funct. Anal.*, Vol. 8, pp.321-340, 1971.

[42] C.G. Broyden, "A class of methods for solving nonlinear simultaneous equations", *Math. of Comp.*, Vol. 19, pp.577-593, 1965.

[43] M.J.D. Powell, "On the convergence of the variable metric algorithm", *J. Init. Math. Appl.*, Vol. 7, pp.21-36, 1971.

[44] J.E. Dennis, Jr. and J.J. Moré, "A characterization of superlinear convergence and its application to Quasi-Newton methods, *Math. of Comp.*, Vol. 28, pp.549-560, 1974.

[45] S.M. Robinson, "Quadratic interpolation is risky", *SINUM*, Vol. 16, pp.377-379, 1979.

[46] A.O. Griewank and M.R. Osborne, "Newton's method for singular problems when the dimension of the nullspace is > 1 ", to appear in *SINUM*.

[47] A.O. Griewank, "Starlike domains of convergence for Newton's method at singularities", TR-CS-80-02, Dept. of Comp. Science, ANU, to appear in *Numerische Mathematik*.