A Model Trust Region Modification
of Inexact Newton's Method
for Nonlinear Two Point
Boundary Value Problems¹

by

Edward J. Dean

Technical Report 85-6, May 1985

¹A thesis submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy, Rice University.
RICE UNIVERSITY

A MODEL TRUST REGION MODIFICATION
OF INEXACT NEWTON'S METHOD
FOR NONLINEAR TWO POINT BOUNDARY VALUE PROBLEMS

by

EDWARD J. DEAN

A Thesis Submitted
In Partial Fulfillment Of The
Requirements For The Degree

DOCTOR OF PHILOSOPHY

Approved, Thesis Committee:

Richard A. Tapia,
Professor of Mathematical Sciences
Joint Chairman

John E. Dennis Jr.,
Professor of Mathematical Sciences
Joint Chairman

John W. Clark,
Professor of Electrical Engineering

Houston, Texas

May, 1985
ABSTRACT

A MODEL TRUST REGION MODIFICATION OF INEXACT NEWTON'S METHOD FOR NONLINEAR TWO POINT BOUNDARY VALUE PROBLEMS

by

EDWARD J. DEAN

The method of quasilinearization for the solution of nonlinear two point boundary value problems is Newton's method for a nonlinear differential operator equation. Semilocal convergence results, with the attendant error estimates, are available from the Kantorovich Theorem. Since the linear boundary value problem to be solved at each iteration must be discretized, it is natural to consider quasilinearization in the framework of Inexact Newton methods. Conditions on the size of the relative residual of the linear differential equation, given by an approximate solution, can then be specified to guarantee rapid local convergence. If initial value techniques are used to solve the linear boundary value problem then it is possible to implement a integration step selection scheme so that the residual criteria is satisfied by the approximate solution. The result is a sequence of approximate solutions to the linear boundary value problems that converge to the true solution of the nonlinear boundary value problem.

A model trust region approach to globalization can be extended to this infinite dimensional problem to allow convergence from an arbitrary initial point. The double dogleg implementation yields a globally convergent algorithm that is robust in solving difficult problems.
Acknowledgements

It is a pleasure to thank my advisors, Professor Richard Tapia and Professor John Dennis, for directing this thesis. I am grateful to Professor Tapia for the generous support, encouragement and inspiration he gave throughout my years at Rice. I also thank Professor Dennis for being so generous with his time and ideas. I thank Professor Francesco Zirilli and Dr. Trond Steihaug for their help. I'd also like to thank Teresa Parks for her generosity, understanding, and encouragement. For always being so helpful I thank Maria Rosa Cells and Mike Pearlman. I also thank Ms. Velva Powers for taking such good care of the students in the department. I would also like to thank Professor Richard Griego for his generous support during my undergraduate years.

This work was supported by DOE contract DE-AS05-82-ER13016

TABLE OF CONTENTS

Chapter 1  Introduction ....................................................  1

Chapter 2  Newton's Method and Quasilinearization .................... 12

Chapter 3  The Model Trust Region Modification ....................... 40

Chapter 4  Inexact Newton's Method .................................... 61

Chapter 5  Implementation and Numerical Results ..................... 75

Chapter 6  Concluding Remarks .......................................... 84

References ................................................................. 87
CHAPTER 1

Introduction

This thesis is concerned with the numerical solution of the nonlinear two point boundary value problem

\[ y'(t) = f(t, y(t)), \ t \in [a, b], \]  
\[ Ay(a) = \alpha, \ By(b) = \beta, \]  

where \( y : [a, b] \subset \mathbb{R} \rightarrow \mathbb{R}^n \), \( f : [a, b] \times \mathbb{R}^n \rightarrow \mathbb{R}^n \), and \( A \in \mathbb{R}^{(n-k) \times n} \), \( B \in \mathbb{R}^k \times n \) are matrices of full row rank, for some \( k \) with \( 1 \leq k \leq n-1 \). The formulation (1.1) describes a general explicit system of \( n \) nonlinear first order ordinary differential equations, for which \( n-k \) linear conditions are specified at one boundary point \( a \), and the remaining \( k \) conditions are specified at the other boundary point \( b \). This problem appears in many diverse areas of the physical sciences and engineering, including the study of chemical reactors, plasma physics, and certain types of fluid flow.

The assumption of separate boundary conditions is mild, and is the form most often encountered in applications. Most general linear boundary conditions, which are of the form

\[ Ay(a) + By(b) = \gamma, \]
the same. For many applications of interest, however, this is not the case (see, for example, Keller(1975)).

In this work, we will analyze modifications of quasilinearization to make it a more viable algorithm for solving nonlinear boundary value problems. The author's initial interest in the method stemmed from a desire to adapt an existing computer code for solving linear boundary value problems into one that would handle the nonlinear case. One advantage of quasilinearization is that it allows the use of existing methods to solve the linear problems at each iteration. It has also been noted (see Roberts and Shipman (1972, pp. 92-93)) that quasilinearization can be a more successful approach than shooting methods for numerically unstable problems. While a shooting method will have to integrate the original nonlinear equation at each iteration, quasilinearization allows the integration of linear equations that are often more stable. Shooting methods also have difficulty with problems that are extremely sensitive to the choice of missing initial conditions. Quasilinearization does not suffer from this drawback since the boundary conditions are satisfied at each iteration. Since quasilinearization is Newton's method on a nonlinear differential operator, we can expect rapid, i.e., q-quadratic, convergence if we start close enough to a nonsingular solution to the problem.

This also presents one of the most serious difficulties with quasilinearization cited in the literature: how to proceed from a poor initial profile. In the event the power of the boundary value problem does not have enough information about the qualitative behavior of the expected solution to generate a suitable initial guess, it is essential that the basic quasilinearization process be globalized. The globalization modification should allow the iteration to proceed from poor initial points into a region from which Newton's method will converge. It should also not hamper Newton's method when such a region is reached so that the local quadratic convergence rate is retained.

Guided by these concerns, we look to the successful globalization techniques developed in the past few years for the solution of nonlinear systems of equations in $R^n$. Specifically, we show how the model trust region approach to globalization can be adapted to the nonlinear boundary value problem. The model trust region modification, besides having an elegant justification, has proven to be a robust method for dealing with poor initial points. In this work, we implement the dogleg version of the technique to yield an essentially globally convergent algorithm.

The second major concern of this thesis is the effect on the convergence of quasilinearization of having only approximate solutions to the linear boundary value problems. Theoretical convergence results make the implicit assumption that each linear problem has been solved exactly. By necessity, however, all solutions to the infinite dimensional linear subproblems can only be discrete approximations. We employ the theory of Inexact Newton methods to determine how accurate the discrete solutions to the linear subproblems must be to ensure convergence to the true infinite dimensional solution of the nonlinear
boundary value problem. This is in contrast to discrete methods that can only claim convergence to a discrete approximate solution. Since the cost of solving each intermediate linear problem exactly is prohibitive, we extend to infinite dimensions the Inexact Newton framework to provide accuracy requirements on the linear problem that are sufficient to ensure local convergence. The theory also characterizes the accuracy needed to maintain desired convergence rates. These conditions are entirely general in the sense that no assumption is made as to how the approximate solution to the linear problem was obtained. The only requirement is that the approximate solution give a linear residual that is smaller than a prescribed tolerance. It is natural then to consider the Inexact Newton framework for the continuous limiting case of large nonlinear problems.

Using this approach, we develop a discretization scheme to generate discrete solutions to the linear problems that are designed to satisfy the residual tolerance requirements. We show how to control both the truncation error in integrating the linear differential equations as well as the interpolation error in the resulting discrete representation of the iterate. Both these errors are controlled by an appropriate step selection strategy for a single step integration method. The result is a sequence of discrete iterates that converge quadratically to the true solution of the nonlinear boundary value problem.

In Chapter 2, we develop the necessary machinery for a rigorous discussion of quasilinearization as a realization of Newton's method for a nonlinear differential operator. Implicit in the definition of the operator are the function spaces on which it acts. We choose these to be Sobolev spaces, which are not the traditional function spaces on which quasilinearization is usually analyzed. These were chosen to aid the analysis in later chapters. We develop a semilocal convergence theory using the powerful Kantorovich theorem. This will enable us to give sufficient conditions on the differential equation and boundary conditions to infer the existence of an isolated solution as well as the rapid convergence of the iterates generated by quasilinearization to this solution with the attendant sharp error estimates. We also show that by assuming mild regularity of the right hand side of the differential equation, the iterates that we generate in these function spaces are the $C^1$ iterates usually associated with quasilinearization. We also show the regularity of the operator is inherited from the the regularity of the differential equation. These results will also be used in later chapters. Finally, we derive a standard local convergence result for quasilinearization as a corollary to our main convergence theorem.

In Chapter 3, we present the globalization modification to Newton's method. We begin with a brief discussion of the descent methods used in solving nonlinear equations in $\mathbb{R}^n$ to motivate the choice of the model trust region modification for the nonlinear boundary value problem. The model trust region algorithm selects a step whose length and direction is determined by how well the linear boundary value problem is trusted to model the nonlinear problem. To determine the step, we approximately minimize a convex quadratic subject to a bound on the step length. This step bound gives the region in
which we trust the linear problem to adequately model the nonlinear one. The step selected can then range from a full Newton step, when the model is good, to a small fraction of the steepest descent step when the model is poor. The trust region is then updated according to a model fitting criteria. It should be noted that while the linear problem to be solved at each iteration is of the same form as that given by the Levenberg-Marquardt approach, how the linear problems change from iteration to iteration is considerably different. For the trust region algorithm, the linear problem is given by the Lagrange equation for the locally constrained problem. Therefore, the parameter in the Levenberg-Marquardt approach arises only indirectly as a Lagrange multiplier. There is no need then to update this parameter since it is implicitly updated as the step bound is updated. The model trust region philosophy for updating the step bound overcomes many of the difficulties usually associated with updating the Levenberg-Marquardt parameter. In particular, the choice of an initial step bound is better motivated than an initial choice of the Levenberg-Marquardt parameter (see Moré(1982)).

We decided to use the dogleg implementation of the model trust region algorithm for many reasons. First, the approximate solution to the locally constrained problem, given by the dogleg step, is sufficient to guarantee global convergence. More importantly, the cost of computing the gradient used in the dogleg step is only a fraction of the expense involved in solving the linear boundary value problem for the Newton step. This differs from the finite dimensional case, where the gradient is usually available at no added cost. Another important reason for choosing the dogleg, is that in the first few iterations it is usually the case, in difficult problems, that the modelling is so poor that only the gradient needs to be calculated. This avoids the cost of the much more expensive Newton step. Finally, an alternative to the dogleg, such as a Hebdon-Moré implementation, would require the solution of two or three linear integro-differential equations, at each iteration. The added cost does not seem justified in our case since the quadratic to be minimized is always convex.

To determine the dogleg step, we show how to derive an iterated integral form for the gradient of the norm squared residual of the nonlinear differential equation. Its computation requires the solutions to an initial value problem and a positive definite linear algebraic system.

Finally, we show how an existing global convergence result for the finite dimensional case will remain valid for the boundary value problem, with little modification.

To summarize Chapter 3, we develop an algorithm for the nonlinear boundary value problem with global convergence properties. The added cost is only a fraction of the cost involved in computing the Newton step. The method also provides a step, called the Cauchy step, when the Newton step cannot be calculated for some reason. One of these reasons is the possibility that the Newton step does not exist or is not unique when far from a solution. It may also happen that the linear integrator will encounter difficulties in calculating
the Newton step when it does exist. We can then fall back on taking a Cauchy step. This does have the effect of rapidly reducing the residual of the differential operator early in the iteration. We note there is no default step to fall back on if the globalization strategy is a damped Newton algorithm.

In Chapter 4, we present the Inexact Newton theory of local convergence for approximate solutions to the Newton equation. Any numerical solution to the Newton equation, in this case a linear two-point boundary value problem, can only be a discrete approximation. The Inexact Newton framework is sufficiently general to apply to all methods for approximately solving the linear problem. The theory gives sufficient conditions for convergence of the approximate iterates to the true solution of the nonlinear boundary value problem. These conditions require the size of the residual of the linear differential equation, given by the discrete solution, to be smaller than a specified tolerance. This tolerance can be explicitly specified before each linear problem is to be solved. The theory also characterizes superlinear and quasi-quadratic rates of convergence by specifying how fast the residual tolerances must decrease during the iteration.

Also in Chapter 4, we propose a method of approximately solving the linear boundary value problems, designed to satisfy the residual criteria. The method uses reduced superposition, otherwise known as the method of complementary functions, to transform the linear boundary value problem into a set of initial value problems. The initial value problems are integrated by a pair of Runge-Kutta-Fehlberg formulas. The approximate solutions of the initial value problems are defined to be cubic Hermite interpolants. These interpolate the output points generated by the integrator, with the derivatives at the output points defined to be the value of the right hand side of the linear differential equation evaluated at these points. The step selection strategy uses the difference of the two integration formulas to estimate the integration step size. The step size is chosen so that the residual of the linear differential equation, given by the Hermite cubic representation, is no larger than the specified tolerance. We note that the step selection formula we derive is different from the ones usually derived when the concern is to control the local truncation error in the integration. Based on numerical evidence that suggests that the step selection should be more conservative, our formula is the one usually implemented.

To summarize Chapter 4, we develop a method for approximately solving the linear boundary value problems that arise in quasilinearization. The method is designed to satisfy the residual criteria given by the Inexact Newton theory. A step selection strategy is proposed that allows the integrator to take steps that are as long as possible, for efficiency, while maintaining the desired residual tolerance. If this residual tolerance is maintained throughout the iteration, we are guaranteed local quasi-quadratic convergence of the discrete approximate solutions, of the linear problems, to the true solution of the nonlinear boundary value problem.
CHAPTER 2

Newton's Method and Quasilinearization

In this chapter we will investigate the method of quasilinearization for the solution of the nonlinear two-point boundary value problem:

\[ y'(t) = f(t, y(t)) \quad t \in [a, b], \]
\[ Ay(a) = \alpha, \quad By(b) = \beta, \]

where \( y : [a, b] \subset \mathbb{R} \rightarrow \mathbb{R}^n, f : [a, b] \times \mathbb{R}^n \rightarrow \mathbb{R}^n, \) and for \( 1 \leq k \leq n-1, A \in \mathbb{R}^{(n-k) \times n}, B \in \mathbb{R}^{k \times n} \) are of full rank.

This method, due to Bellman and Kalaba (1965), consists of solving a sequence of linear boundary value problems. Under appropriate conditions, it can be shown that the process is well defined and that the sequence of solutions to these linear problems converges to the solution of the original nonlinear problem. We can also interpret quasilinearization as Newton's method for the solution of a nonlinear differential operator equation (Roberts and Shipman, 1972). This viewpoint will be useful in later chapters when we consider modifications to the basic method.
The use of Newton’s method for the solution of nonlinear differential and integral equations is very widespread (Kantorovich and Akilov, 1964) (Collatz, 1966), (Bosarge and Falb, 1967), (Bryan, 1968), (Antosiewicz, 1968), (Tapia, 1969). Nearly all make use of some variant of the Kantorovich theorem to prove local convergence results. It is important, however, that the hypothesis and conclusions of the Kantorovich theorem be formulated in terms of the original problem. This is what we propose to do in this chapter.

Notation

The following notation will be used throughout this chapter:

\[ f_z = \frac{\partial f}{\partial y}, \quad f_{zz} = \frac{\partial^2 f}{\partial y^2}. \]

\[ \| \cdot \|_2 \] denotes the Euclidean norm in \( \mathbb{R}^n \).

\[ V = \{ y : y : [a,b] \subset \mathbb{R} \to \mathbb{R}^n \}. \]

\[ C[a,b] = \{ y \in V : y \text{ is continuous on } [a,b] \}. \]

\[ C^1[a,b] = \{ y \in V : y \text{ is continuously differentiable on } [a,b] \}. \]

The norms on \( C[a,b] \) and \( C^1[a,b] \) will be given by

\[ \| y \|_C = \max_{t = 1, \ldots, n} |y_i(t)| \quad \text{and} \quad \| y \|_C = \max_{i = 1, \ldots, n} |y_i(t)|, \]

respectively, where \( y_i \) is the \( i \)th component of \( y \).

\[ L^2[a,b] = \{ y \in V : \int_a^b \| y(t) \|^2 dt < \infty \}, \]

where the integral is taken in the sense of Lebesgue.

\[ H^1[a,b] = \{ y \in V : y \text{ is absolutely continuous and } y' \in L^2[a,b] \}. \]

\[ H^2[a,b] = H^1[a,b] \cap \{ y \in V : Ay(a) = 0, By(b) = 0 \}. \]

The inner products on \( L^2[a,b] \) and \( H^2[a,b] \) will be given by

\[ \langle y, z \rangle_{L^2} = \int_a^b y(t) z(t) dt \quad \text{and} \quad \langle y, z \rangle_{H^2} = \int_a^b y'(t) z'(t) dt + y(a) z(a), \]

respectively.

The norm \( \| \cdot \| \) on a matrix or other bounded linear operator will denote the corresponding induced operator norm.

\[ S(\hat{y}, r) = \{ y : \| \hat{y} - y \| < r \}, \] where the norm is implied by context.

Suppose \( y^* \) is a classical solution to (2.1). That is, \( y^* \in C^1[a,b] \), and solves problem (2.1). Quasilinearization proceeds as follows. We assume that we can find a function \( y_0 \in C^1[a,b] \) satisfying the boundary conditions in (2.1). To improve \( y_0 \) as an approximation to \( y^* \) we try to solve the following linear problem for \( y_1 \):

\[ \| y \|_C = \| y \|_C + \max_{i = 1, \ldots, n} |y_i(t)|, \]
\[ y'_1(t) = f_2(t,y_0(t))y_1(t) + [f(t,y_0(t)) - f_2(t,y_0(t))y_0(t)], \quad t \in [a,b], \tag{2.2} \]

\[ Ay_1(a) = \alpha, \quad By_1(b) = \beta. \]

which results from an affine approximation to \( f \) in (2.1) about the function \( y_0 \).

In Theorem 2.9 we give sufficient conditions for \( y_1 \) to be a better approximation than \( y_0 \) to \( y^* \). Continuing in this way, we can construct a sequence of functions \( \{y_k\} \) that become more accurate approximations to \( y^* \).

To improve the scaling of the linear problems, we will solve the linear equation for the step \( s_k = y_{k+1} - y_k \). Quasilinearization then proceeds as follows:

1) Find \( y_0 \in C^1[a,b] \) satisfying \( Ay_0(a) = \alpha, \quad By_0(b) = \beta. \) \hspace{1cm} (2.3)

2) For \( k = 0, 1, \ldots \), until convergence do

2.1) Solve

\[ s'_k(t) = f_2^2(t,y_k(t))s_k(t) + [f(t,y_k(t)) - f_2(t,y_k(t))]y'_k(t), \quad t \in [a,b], \]

\[ As_k(a) = 0, \quad Bs_k(b) = 0, \]

for \( s_k \in C^1[a,b] \).

2.2) Set \( y_{k+1} = y_k + s_k \).

In Chapter 4 we will specify the method used to solve each intermediate linear boundary value problem.

Roberts and Shipman (p.101,1972) give a type of local convergence proof, due to McGill and Kenneth (1963), for the above process. Under suitable

smoothness assumptions on \( f \), they use an implicit contraction mapping to show that the above process is well defined for any starting profile \( y_0 \). They also show the method to be \( q \)-quadratically convergent in the \( C[a,b] \) norm. They make the assumption that the interval \( [a,b] \) is sufficiently small. We prefer a different local result, using the Kantorovich theorem, which will include their result as a special case.

Without loss of generality we can consider only homogeneous boundary conditions. To see this, suppose our problem is of the form:

\[ z'(t) = g(t,z(t)), \quad t \in [a,b] \]

\[ Az(a) = \alpha, \quad Bz(b) = \beta. \]

Define \( y(t) = z(t) + l(t) \), where

\[ l(t) = \left[ (A^1 - B^1) \alpha + aB^1 \beta - bA^1 \alpha \right]/(b-a). \]

Here \( A^1 \) and \( B^1 \) are fixed right inverses for \( A \) and \( B \). Since our matrices are assumed to be of full rank, \( A^1 = A^T(AA^T)^{-1} \) and \( B^1 = B^T(BB^T)^{-1} \). We can consider, instead, the linearly translated problem:

\[ y'(t) = f(t,y(t)), \quad t \in [a,b] \]

\[ Ay(a) = 0, \quad By(b) = 0, \]

where \( f(t,y(t)) \equiv g(t,z(t)) + [A^1 \alpha - B^1 \beta]/(b-a) \).

We now introduce the operator that will be useful in reformulating the above in terms of Newton's method. Let \( P : H^1_0[a,b] \to L^1[a,b] \) be defined as the residual of the differential equation.
\[ P(y)(t) = y'(t) - f(t, y(t)), \quad t \in [a, b]. \]

We can now consider Newton's method for the solution of the problem:

\[ P(y) = 0, \quad y \in H^1_0[a, b]. \quad (2.4) \]

The choice of domain and range spaces for \( P \) was guided by several concerns. The structure of a Hilbert space will allow many of the results in the finite dimensional case to remain valid. In particular, under mild smoothness assumptions on \( f \), \( P \) will be shown to be Fréchet differentiable. Later modifications of the basic algorithm will require the gradient of a nonlinear functional whose domain is that of \( P \). By defining the functional on a Hilbert space, we are assured that the gradient, when it exists, is unique (see Tapia(1971)). It should be noted that convergence in \( H^1_0[a, b] \) is quite strong since it implies convergence of the functions in the \( C[a, b] \) norm as well as convergence of their derivative functions in \( L^2[a, b] \). Additionally, weak convergence in \( H^1_0[a, b] \) is still strong, since it still implies convergence of the functions in \( C[a, b] \). This will prove to be useful in the next chapter. Finally, from a computational point of view, actual calculations of norms will be much simpler than if we were considering a traditional domain space such as \( C^1[a, b] \).

By assuming sufficient regularity on \( f \), we will also show that a solution \( y^* \in H^1_0[a, b] \) will be a classical solution.

Roberts and Shipman (p.150,1972) have also given an alternate proof of the

\[ P(y) : C^1[a, b] \rightarrow C[a, b]. \]

Again, however, they make the assumption that the interval \([a, b]\) is sufficiently small. A corollary to our main theorem (2.9) will also give this result.

For the remainder of this chapter we will require the following smoothness assumption on \( f \).

**Assumption 2.1**

We assume there exists \( r > 0 \) such that in the closure \( \overline{D_0} \) of the tube

\[ D_0 = \{(t, y) : a \leq t \leq b, \| y_0(t) - y \|_2 < r \}, \]

\( f_2 \) exists and is continuous. Furthermore, we assume that in \( \overline{D_0} \), \( f_2 \) is Lipschitz continuous with respect to \( y \), uniformly in \( t \), i.e. there exists an \( L > 0 \) such that

\[ \| f_2(t, y) - f_2(t, z) \|_2 \leq L \| y - z \|_2, \text{ for all } (t, y), (t, z) \in \overline{D_0}. \]

We first prove the following critical lemma, which is a simple special case of the Sobolev Imbedding Theorem (Kantorovich and Akilov,1964,p. 379 ) and a vector-valued modification to a result in Tapia and Thompson(1978,p.152). It states that \( H^1_0[a, b] \) can be continuously imbedded in \( L^2[a, b] \).
Lemma 2.2

For all $s \in H_{\delta}^0[a,b]$, $\|s\|_{L^2} \leq M \|s\|_{H^1}$, where

$$M = \max \{ (b-a), (b-a)^\delta \}.$$  

Proof:

Let $s_i$ denote the $i^{th}$ component of $s$. By the absolute continuity of $s$, for all $t \in [a,b]$,

$$\|s(t)\|_2^2 = \sum_{i=1}^n \left( \int \frac{d}{ds} s_i'(r) \, dr + s_i(a) \right)^2 \leq \sum_{i=1}^n \left( \int \left| s_i'(r) \right| \, dr + |s_i(a)| \right)^2.$$  

By the Schwarz inequality in $L^2$,

$$\int s_i'(r) s_i(a) \, dr \leq (b-a)^\delta \|s'_i\|_{L^2}, \quad i = 1, \ldots, n.$$  

Applying the Cauchy Schwarz inequality in $R^n$, we have

$$\sum_{i=1}^n \|s_i'\|_{L^2} |s_i(a)| \leq \left( \sum_{i=1}^n \int s_i'(r)^2 \, dr \right)^{\frac{1}{2}} \left( \sum_{i=1}^n |s_i(a)|^2 \right)^{\frac{1}{2}} \quad \text{(2.7)}$$

Together, (2.5), (2.6), and (2.7) give

$$\|s(t)\|_2^2 \leq \left( (b-a)^\delta \|s'_i\|_{L^2} + \|s(a)\|_2 \right)^2,$$

for all $t \in [a,b]$. (2.8)

Integrating (2.8) from $a$ to $b$ yields

$$\|s\|_2^2 \leq \left( (b-a)^\delta \|s'_i\|_{L^2} + \|s(a)\|_2 \right)^2.$$  

which proves the lemma. 

Definition 2.3

Let $\Omega_0 \subset H_{\delta}^0[a,b]$ be the ball $S(y_0, \hat{r})$ where $\hat{r} = \frac{r}{\max \{1, (b-a)^\delta \}}$ and $r$ is given in Assumption 2.1. Note that by (2.8), $y \in \Omega_0$ gives $(t, y(t)) \in \overline{D_0}$, for all $t \in [a,b]$.

From the differentiability of $f$ and the above lemma we can now show the Fréchet differentiability of $P$.

Theorem 2.4

For all $y \in \Omega_0$, the Fréchet derivative of $P$ at $y$ exists and is given by

$$P'(y)(\cdot)(t) = \frac{d}{dt} f_2(t, y(t))(\cdot)(t) - f_2(t, y(t))(\cdot)(t).$$

Proof:

By the existence of $f_2(t, y(t))$, for all $(t, y(t)) \in D_0$, the Fréchet derivative of $P$ at a point $y \in \Omega_0$ in the direction $s \in H_{\delta}^0[a,b]$, $s \neq 0$, if it exists, will be given by
\[ P'(y)(s)(t) = \lim_{r \to 0^+} \left[ P(y + rs)f(t) - P(y)f(t) \right] = \lim_{r \to 0^+} \left[ (y'(t) + rs')(t) - (f(t,y(t))+rs(t)) - (y'(t)-f(t,y(t))) \right] = s'(t) - f_2(t,y(t))s(t). \]  \hspace{1cm} (2.9)

Clearly, \( P'(y) \) is linear. We show it is bounded. By the continuity of \( y \) on \([a,b] \) and \( f_2 \) on \( D_0 \), there exists a constant \( c_4 \) such that \( \|f_2(t,y(t))\|_2 \leq c_4 \) for all \( t \in [a,b] \). Together with (2.9), we have, for any \( s \in H_2^0 [a,b] \),

\[
\| P'(y)(s) \|_2^2 \leq \int_a^b \| s'(t) \|_2^2 \, dt + 2c_4 \int_a^b \| s'(t) \|_2 \| s(t) \|_2 \, dt + c_4^2 \int_a^b \| s(t) \|_2^2 \, dt.
\]

By the Schwarz inequality,

\[
\int_a^b \| s'(t) \|_2 \| s(t) \|_2 \, dt \leq \| s' \|_{L^2} \| s \|_{L^2},
\]

yielding

\[
\| P'(y)(s) \|_2^2 \leq \left( \| s' \|_{L^2}^2 + c_4 \| s \|_{L^2} \right)^2 \leq \left( \| s' \|_{L^2}^2 + c_4 M \| s \|_{H^1} \right)^2
\]

where \( M = \max(\| b-a \|, \| b-a \|) \). The second inequality in (2.10) follows from Lemma 2.2. Therefore,

\[
\frac{\| P'(y)(s) \|_{L^2}}{\| s \|_{L^2}} \leq \max\{1, c_4 M\}
\]

and, since \( s \in H_2^0 [a,b] \) was arbitrary, we have that \( P'(y) \) is bounded.

Since \( f_2 \in L_{p_2}(D_0) \), we can show (see Dennis and Schnabel, (1983), p.75), that for any \( s \in H_2^0 [a,b] \), \( s \neq 0 \), and for any \( t \in [a,b] \),

\[
\| f(t,y(t) + s(t)) - f(t,y(t)) - f_2(t,y(t))s(t) \|_2 \leq \frac{L}{2} \| s(t) \|_2^2
\]

Therefore,

\[
\| P(y+s) - P(y) - P'(y)(s) \|_2^2 \leq \frac{L^2}{4} \int_a^b \| s(t) \|_2^2 \, dt \leq \frac{L^2}{4} \| s \|_{H^1}^4 \leq L^2 M^4 \| s \|_{H^1}^4,
\]

which gives

\[
\lim_{\| s \|_{H^1}^2 \to 0} \frac{\| P(y+s) - P(y) - P'(y)(s) \|_{L^2}}{\| s \|_{H^1}^2} = 0,
\]

and proves the theorem. \( \Box \)

Newton's method for problem (2.4) can then be stated as:

1) Given \( y_0 \in H_0^2 [a,b] \) \hspace{1cm} (2.11)

2) For \( k = 0,1,2,\ldots \) until convergence do

2.1) Solve \( P'(y_k)s_k = -P(y_k) \), \( s_k \in H_0^2 [a,b] \)

2.2) Set \( y_{k+1} = y_k + s_k \).

If we can also show that \( a_k \in C^1[a,b] \), \( k = 0,1,2,\ldots \), then a comparison of (2.3) and (2.11) will give the equivalence of Newton's method for the operator \( P \) and the method of quasilinearization for problem (2.1). This additional regularity on the sequence \( \{ a_k \} \) follows immediately from the assumed regularity on \( f \):
Theorem 2.5

Let \( y_0 \in C^1[\alpha, \beta] \) and \( f \) satisfy the regularity conditions in Assumption 2.1. If the iterates generated by algorithm (2.11) stay in \( \Omega_0 \), then they satisfy \( s_k \in C^1[\alpha, \beta] \), \( k = 0, 1, \ldots \).

Proof:

For the case \( k = 0 \), if \( s_0 \in H^1_0[\alpha, \beta] \) solves

\[
P'(y_0)s_0 = -P(y_0)
\]
then

\[
\frac{d}{dt} \left( \int_a^b s_0'(t) - q(t) \right)^2 dt = 0
\]

where

\[
q(t) = f_2(t, y_0(t))s_0(t) + \frac{f(t, y_0(t)) - y_0'(t)}{2}.
\]

This then implies that \( s_0'(t) - q(t) = 0 \) almost everywhere \( (a.e.) \) in \( [\alpha, \beta] \). By the continuity of \( f_2, s_0, \) and \( y_0' \) on \( [\alpha, \beta] \)

\[
Q(t) = \int_a^t q(r) dr
\]
is defined for all \( t \in [\alpha, \beta] \) and differentiable \( a.e. \) in \( [\alpha, \beta] \). Therefore, we can write

\[
\frac{d}{dt}(s_0(t) - Q(t)) = 0, \quad a.e. \text{ in } [\alpha, \beta].
\]
Since \( s_0 \) and \( Q \) are absolutely continuous, we have (see Royden (1988), p.105)

\[
s_0(t) - Q(t) = c, \quad \text{for all } t \in [\alpha, \beta],
\]
where \( c \) is a constant. Hence

\[
s_0'(t) = Q'(t) = q(t) \quad \text{for all } t \in [\alpha, \beta],
\]
and since \( q \) is continuous, \( s_0 \in C^1[\alpha, \beta] \). It follows that \( y_1 \in C^1[\alpha, \beta] \) and the theorem holds by induction.

We now state the following version of the Kantorovich theorem. This theorem specifies sufficient conditions for the existence of an isolated solution to a nonlinear equation, and for Newton's method to be well defined and rapidly convergent to this solution. If such a solution to problem (2.4) exists, then an argument similar to Theorem 2.5 will allow us to conclude that it is also a classical solution to problem (2.1). The following version of the theorem is due to Gragg and Tapia (1974) and includes the authors' optimal error bounds.

Theorem (Kantorovich) 2.6

Let \( P : \Omega_0 \subset X \to Y \) be a mapping defined on an open convex subset \( \Omega_0 \) of a Banach space \( X \) into a Banach space \( Y \). Assume that \( P \) is Fréchet differentiable in \( \Omega_0 \) and for some \( x_0 \in \Omega_0 \), that \( P'(x_0)^{-1} \) exists. Assume also that
(i) \( \| P'(x_0)^{-1} \| \leq \beta \),

(ii) \( \| P'(x_0)^{-1}P(x_0) \| \leq \eta \),

(iii) \( \| P'(z) - P'(y) \| \leq K \| x-y \| \), for all \( x,y \in \Omega \),

with \( h = \beta K \eta \leq \frac{1}{2} \).

If \( S(x_0, r_0) \subset \Omega \), where \( r_0 = \frac{1 - \sqrt{1 - 2h}}{\beta K} \), then the Newton iterates
\[ x_{k+1} = x_k - P'(x_k)^{-1}P(x_k) \]
exist with \( x_k \in S(x_0, r_0) \subset \Omega \) for \( k \geq 0 \). Furthermore,
\[ z_k \to z^* \], with \( z^* \in S(x_0, r_0) \) and \( P(z^*) = 0 \). If \( h < \frac{1}{2} \) then \( z^* \) is the unique
solution to \( P(z) = 0 \) in \( S(x_0, r_1) \) where \( r_1 = \frac{1 + \sqrt{1 - 2h}}{\beta K} \). If \( h = \frac{1}{2} \), \( z^* \) is unique
in \( S(x_0, r_0) \).

Error bounds are given by
\[
\| z^* - z_k \| \leq \frac{2\sqrt{1 - 2h}}{h} \| z_{k+1} - x_0 \|
\]
and
\[
\left\| \frac{z_{k+1} - z_k}{1 + \sqrt{1 + 4\theta^2}/(1 + \theta^2)} \right\| \leq \| z^* - z_k \| \leq \theta^{k+1} \| z_{k+1} - z_k \|
\]
where
\[ \theta = \frac{1 - \sqrt{1 - 2h}}{1 + \sqrt{1 - 2h}} \leq 1. \]

In particular, the bounds for \( h = \frac{1}{2} \),
\[
\| z^* - z_k \| \leq 2^{k+1} \| z_1 - x_0 \|
\]
and
\[
2(\sqrt{2} - 1) \| z_{k+1} - z_k \| \leq \| z^* - z_k \| \leq \| z_{k+1} - z_k \|
\]
hold for \( h \leq \frac{1}{2} \). \( \square \)

In the finite dimensional case, determining whether the hypotheses of the Kantorovich theorem are satisfied for a particular problem is, generally, either impossible or more work than applying Newton's method. In our case, however, each iteration is very costly since it requires the solution of a linear boundary value problem. Since determining the existence of an isolated solution to a nonlinear boundary value problem is in general very difficult, it might be useful to try to calculate the Kantorovich constants. McCarthy and Tapia (1975) have been able to calculate these constants in the case when \( f \) is a polynomial in the dependent variable \( y \), thereby allowing the determination of the existence of a solution and application of the tight error bounds (2.13). Dennis (1971, p.457), under the Kantorovich hypothesis, gives the bounds
\[
\frac{1}{2} \| z_{k+1} - z_k \| \leq \| z_k - z^* \| \leq 2 \| z_{k+1} - z_k \|,
\]
which involve a single Newton step, and he shows
\[
\| z_{k+1} - z^* \| \leq \frac{\beta K}{\sqrt{1 - 2h}} \| z_k - z^* \|^2.
\]
Bounds of the type (2.14),(2.15) and (2.16) are useful when the Kantorovich constants cannot be calculated, while (2.17) gives the q-quadratic rate of Newton's method when \( h < \frac{1}{2} \).
We now apply the Kantorovich theorem to our specific boundary value problem. We determine the bounds \( n, K \), and \( \beta \) in terms of the function \( f \), the interval size \( b - a \), and the boundary condition matrices \( A \) and \( B \).

We will first establish conditions under which \( P'(y_0)^{-1} \) exists and derive an upper bound on its operator norm. To do so, consider the following linear boundary value problem

\[
\begin{align*}
\dot{s}_o(t) &= F(t)s_o(t) + g(t), \quad t \in [a, b], \\
As_o(a) &= 0, \quad Bs_o(b) = 0,
\end{align*}
\tag{2.18}
\]

where \( F(t) = f(t, y(t)) \) and \( g(t) = f(t, y_0(t)) - y_0'(t) \).

We will determine conditions under which a solution to (2.18) exists, is unique, and can be written in terms of a Green's function, i.e.

\[
\begin{align*}
s_o(t) &= \int_a^b G(t, \xi)g(\xi) \, d\xi
\end{align*}
\]

where \( G(t, \xi) \in R^{n \times n} \) for \( t, \xi \in [a, b] \). This is the same technique used in the Roberts and Shipman (1972) proof. We show it remains valid for the operator \( P : H^1_0[a, b] \rightarrow L^2[a, b] \).

The continuity of \( F \) ensures the unique \( C^1[a, b] \) solution of the initial value problem

\[
U'(t) = F(t)U(t), \quad t \in [a, b]
\tag{2.19}
\]

where \( U(t) \in R^{n \times n} \) for \( t \in [a, b] \), and \( I \) denotes the identity matrix. Since \( U(a) \) is nonsingular, \( U(t) \) is nonsingular for all \( t \in [a, b] \) (see Coddington and Levinson (1955), p.69), with \( U^{-1}(t) \) the \( C^1[a, b] \) solution to

\[
\begin{align*}
U^{-1}(t)f &= -U^{-1}(t)f, \quad t \in [a, b] \\
U^{-1}(a) &= I.
\end{align*}
\tag{2.20}
\]

Formally, the Green's function for problem (2.18) satisfies, for each \( \xi \in [a, b] \),

\[
\begin{align*}
\frac{d}{dt}G(t, \xi) &= F(t)G(t, \xi) + \delta(t - \xi), \quad t \in [a, b] \\
AG(a, \xi) &= 0, \quad BG(b, \xi) = 0,
\end{align*}
\tag{2.21}
\]

where \( \delta(t - \xi) \) is the matrix valued Dirac function. The variation-of-constants formula for the solution of (2.21) is then given by

\[
G(t, \xi) = U(t)C(\xi) + U(t)\int_a^t U^{-1}(r)\delta(r - \xi)\, dr
\tag{2.22}
\]

where \( C(\xi) \) is chosen to allow \( G(\cdot, \xi) \) to satisfy the boundary conditions. (2.22) and the boundary conditions in (2.21) then lead to

\[
G(t, \xi) = \begin{cases} 
U(t)C^{-1}(\xi), & t < \xi \\
U(t)[C + I]U^{-1}(\xi), & t > \xi
\end{cases}
\tag{2.23}
\]

where \( C \) solves

\[
\begin{bmatrix} A \\
[BU(b)] \end{bmatrix} C = - \begin{bmatrix} 0 \\
[BU(b)] \end{bmatrix}
\tag{2.24}
\]
Lemma 2.7

If \( \begin{bmatrix} A & B \end{bmatrix} \) is nonsingular then \( \Gamma_0 = P' \gamma_0 \) exists and is given by

\[
(\Gamma_0 \gamma)(t) = \int_a G(t, \xi) \gamma(\xi) \, d\xi \quad \text{for all } \gamma \in L^2[a, b].
\]

Proof:

We first show that \( \Gamma_0 \) is a right inverse for \( P' \gamma_0 \). Let \( \gamma \in L^2[a, b] \) and define \( \gamma_0 \) to be

\[
\gamma_0(t) = (\Gamma_0 \gamma)(t) = \int_a G(t, \xi) \gamma(\xi) \, d\xi
\]

\[
= U(t) \left[ \hat{C} + [ \left[ B \hat{U}^0(t) \right] \right] \int_a U^{-1}(\xi) \gamma(\xi) \, d\xi + U(t) \hat{C} \int_a U^{-1}(\xi) \gamma(\xi) \, d\xi
\]

Since \( U, U^{-1} \in C^1[a, b] \), we have both that \( U^{-1} \gamma \) is integrable and \( U \) is absolutely continuous. These then imply the absolute continuity of \( \gamma_0 \). Therefore, \( \gamma_0 \) is differentiable a.e. Differentiating (2.25) and recalling that \( U \) solves (2.19), we find

\[
\gamma_0'(t) = F(t) U(t) \left[ \hat{C} + [ \left[ B \hat{U}^0(t) \right] \right] \int_a U^{-1}(\xi) \gamma(\xi) \, d\xi + U(t) \hat{C} \int_a U^{-1}(\xi) \gamma(\xi) \, d\xi
\]

\[
+ F(t) U(t) \hat{C} \int_a U^{-1}(\xi) \gamma(\xi) \, d\xi - U(t) \hat{C} U(t) \gamma(t)
\]

\[
= F(t) \gamma_0(t) + \gamma(t), \quad \text{a.e.}
\]

Hence, \( P' \gamma_0 \gamma_0 = \gamma \).

Since \( \gamma \in L^2[a, b] \) and \( \gamma_0 \in C[a, b] \), we have that \( \gamma_0' \in L^2[a, b] \). Recalling that \( U(a) = I \) and \( \hat{C} \) solves (2.24), we find

\[
\begin{bmatrix} A \\ B \end{bmatrix} \gamma_0(a) + \begin{bmatrix} 0 \\ B \end{bmatrix} \gamma_0(b) = \begin{bmatrix} A \\ B \end{bmatrix} U(a) \hat{C} \int_a U^{-1}(\xi) \gamma(\xi) \, d\xi
\]

\[
+ \begin{bmatrix} 0 \\ B \end{bmatrix} U(b) \left[ \hat{C} + [ \left[ B \hat{U}^0(t) \right] \right] \int_a U^{-1}(\xi) \gamma(\xi) \, d\xi
\]

\[
= \begin{bmatrix} A \\ B \hat{U}^0(b) \end{bmatrix} \hat{C} + \begin{bmatrix} 0 \\ B \hat{U}^0(b) \end{bmatrix} \int_a U^{-1}(\xi) \gamma(\xi) \, d\xi
\]

\[
= 0,
\]

or, equivalently, \( A \gamma_0(a) = 0 \), \( B \gamma_0(b) = 0 \). Hence, \( \gamma_0 \in H^1_0[a, b] \).

Together, for all \( \gamma \in L^2[a, b] \),

\[
P' \gamma_0(\Gamma_0 \gamma) = P' \gamma_0 \gamma_0, \quad \gamma_0 \in H^1_0[a, b]
\]

\[
= \gamma,
\]

which shows that \( \Gamma_0 \) is a right inverse for \( P' \gamma_0 \).

We now show \( \Gamma_0 \) is a left inverse for \( P' \gamma_0 \). Let \( \gamma \in H^1_0[a, b] \). Clearly \( P' \gamma_0 \gamma = (s' - F \gamma) \in L^2[a, b] \). By the argument above, \( \Gamma_0 P' \gamma_0 \gamma \) exists and is in \( H^1_0[a, b] \), where for all \( t \in [a, b] \).
\[ \Gamma_0(P'(y_0)) = \int_a^b G(t, \xi) s'(\xi) - F(\xi) s(\xi) \, d\xi. \tag{2.27} \]

The first term in (2.27) can be written as
\[ \int_a^b G(t, \xi) s'(\xi) \, d\xi = U(t) \left[ \tilde{C} + I \right] \int_a^b U^{-1}(\xi) s'(\xi) \, d\xi \]
\[ + U(t) \tilde{C} \int_a^b U^{-1}(\xi) s'(\xi) \, d\xi. \tag{2.28} \]

Since \( U^{-1} \) and \( s' \) are integrable on \([a,b]\), integration by parts is justified for the indefinite integrals in (2.28) (see Riesz and Sz.Nagy (1956), p.54). Recalling that \( U^{-1} \) is the solution to (2.20), we find
\[ \int_a^b U^{-1}(\xi) s'(\xi) \, d\xi = [U^{-1}(t) s(t) - s(a)] + \int_a^b U^{-1}(\xi) F(\xi) s(\xi) \, d\xi \]
and
\[ \int_a^b U^{-1}(\xi) s'(\xi) \, d\xi = [U^{-1}(b) s(b) - U^{-1}(t) s(t)] + \int_a^b U^{-1}(\xi) F(\xi) s(\xi) \, d\xi. \]

Therefore, for all \( t \in [a,b] \),
\[ \int_a^b G(t, \xi) s'(\xi) \, d\xi = \int_a^b G(t, \xi) F(\xi) s(\xi) \, d\xi \tag{2.29} \]
\[ + U(t) \left[ \tilde{C} U^{-1}(b) s(b) + [I - \tilde{C}] s(a) \right]. \]

Using \( \tilde{C} = \begin{bmatrix} A \end{bmatrix}^{-1} \begin{bmatrix} 0 & 1 \end{bmatrix} \), \( As(a) = 0 \), \( B s(b) = 0 \) and noticing that
\[ I - \tilde{C} = \begin{bmatrix} A \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{2.29} \]

We note in the above proof that the existence and uniqueness of a solution to the linear boundary value problem (2.18) is equivalent to the existence and uniqueness, respectively, of a solution to the linear algebraic problem (2.24).

To establish a bound on \( \| P'(y_0)^{-1} \| \), let \( g \in L^2[0,b] \). Then, by the above Lemma 2.7,
\[ \| P'(y_0)^{-1} g \|_{L^2} \leq \frac{1}{\mu} \int_a^b G(t, \xi) g(\xi) \, d\xi + \int_a^b G(t, \xi) s(\xi) \, d\xi. \]

Since \( F(t) = f_2(t, y_0(t)) \) is continuous on \([a,b]\), there exists \( M_F > 0 \) such that \( \| F(t) \| \leq M_F \) for all \( t \in [a,b] \). It can also be shown (see Stoer and Bulirsch (1980), p.409) that
\[ \| U(t) - I \| \leq \exp \left( \int_a^t \| F(\tau) \| \, d\tau \right) - 1 \quad \text{for all} \quad t \in [a,b]. \tag{2.30} \]

A similar proof gives the identical bound on \( \| U^{-1}(t) - I \| \). If we let \( M_A = \| \begin{bmatrix} A \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \| \) and \( M_B = \| \begin{bmatrix} A \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \| \), then it follows
from (2.24), (2.25), and (2.30) that
\[ \| G(t, \xi) \| \leq e^{2(t \cdot \xi) M_{\max \{M_A, M_B\}}}, \text{ for all } t, \xi \in [a, b], = M_{G}. \]

Recalling (2.19), (2.23) and applying the triangle and Schwarz inequalities, we have the bound
\[
\begin{aligned}
\frac{d}{dt} \int_a^b G(t, \xi) \sigma(\xi)\, d\xi & = \int_a^b \frac{d}{dt} G(t, \xi) \sigma(\xi)\, d\xi + \sigma \| \xi \|^2 \\
& \leq \int_a^b \left| \frac{d}{dt} G(t, \xi) \sigma(\xi)\, d\xi \right| + 2 \| \sigma(\xi)\| \| \xi \|_2 \| \sigma(t) \|_2 \\
& + \| \sigma(t) \|_2^2 \, dt \\
& \leq \left[ M_{F} M_{G} (b-a)^2 + 4 \right] \| \sigma \|_2^2.
\end{aligned}
\]

Next, we show that \( P' \) is Lipschitz continuous.

**Lemma 2.8**

The operator \( P' \) is Lipschitz continuous in \( \Omega_0 \) with Lipschitz constant \( K = L \max\{b-a, (b-c)\} \), where \( L \) is the Lipschitz constant for \( f \) in \( D_0 \).

**Proof:**

The Lipschitz continuity of \( P' \) follows directly from that of \( f \). Let \( y, z \in \Omega_0 \) and \( s \in H^p_0[a, b], s \neq 0 \). Recalling that \( y, z \in \Omega_0 \) implies \( (t, y(t)), (t, z(t)) \in D_0 \), for all \( t \in [a, b] \), we have, using the Schwarz inequality,
\[
\begin{aligned}
\| (P'(y) - P'(z)) s \|_2^2 & = \int_a^b \| f(y(t)) s(t) - f(z(t)) s(t) \|_2^2 dt \\
& \leq L^2 \int_a^b \| y(t) - z(t) \|_2^2 \| s(t) \|_2^2 dt \\
& \leq L^2 \| y - z \|_2^2 \| s \|_2^2.
\end{aligned}
\]

Since \( s \) was arbitrary, Lemma 2.2 gives
\[
\| P'(y) - P'(z) \| = \sup_{s \in H^p_0[a, b]} \| (P'(y) - P'(z)) s \|_{L^2} \\
\leq \left[ (M_{F} M_{G} (b-a)^2 + 4) \right] \| s \|_{L^2},
\]

(2.31)
From Lemma 2.7, Lemma 2.8 and inequality (2.31) we can now restate the Kantorovich theorem for problem (2.4) as follows.

**Theorem 2.9**

Let \( y_0 \in H^1_0[a,b] \). Assume there is an \( r > 0 \) such that \( f_2 \) is continuous in the closure \( \overline{D_0} \) of \( D_0 = \{ (t,y) : a \leq t \leq b, \| y-y_0(t) \|_2 < r \} \). Furthermore, assume that \( f_2 \) is Lipschitz continuous, with respect to \( y \), uniformly in \( t \), in \( \overline{D_0} \) with Lipschitz constant \( L \). Let the matrices \( A \) and \( B \), defining the boundary conditions, be such that \( \begin{bmatrix} A \\ BU(b) \end{bmatrix} \) is nonsingular, where \( U \) solves the initial value problem (2.19). Let \( s_0 \in H^1_0[a,b] \) solve, in the \( L^2 \) sense, the linear boundary value problem (2.18). Let

\[
\eta = \| s_0 \|_{H^2_0},
\]

\[
K = L \max \{ (b-a)^2/(b-a), \}
\]

\[
\beta = \left( (M_{f}M_{C}(b-a)+1)^2 + (M_{C}(b-a))^2 \right)^{1/2},
\]

where \( h = \beta K \eta \leq 1/2 \) and \( r_0 \leq r \) where \( r_0 = \frac{1-\sqrt{1-2h}}{2K} \). Then Newton's method (2.11) converges in \( H^1_0[a,b] \), from \( y_0 \) to \( y^* \). Here, \( y^* \) is the unique solution to \( P(y) = 0 \), in the closed neighborhood \( \overline{S(y_0,r_0)} \) where \( r_0 = r_0/\max \{ 1, (b-a)^{2/3} \} \). If \( h < \frac{1}{2} \), then \( y^* \) is also unique in \( \overline{S(y_0,r_1)} \) where \( r_1 = r_1/\max \{ 1, (b-a)^{2/3} \} \) and \( r_1 = \min \{ r, \frac{1+\sqrt{1-2h}}{2K} \} \). Finally, the bounds given in (2.12)-(2.17) are valid in \( H^1_0[a,b] \).

As an immediate corollary, by the assumed regularity on \( f \), we have the following local convergence result to a solution of the problem of interest (2.1).

**Corollary 2.10**

Under the assumptions of Theorem 2.9, if \( y_0 \in C^1[a,b] \), then the method of quasilinearization (2.2) is well defined and converges in \( C^1[a,b] \) to a solution \( y^* \) of problem (2.1).

**Proof:**

We first show that \( y^* \) in Theorem 2.9 is continuously differentiable using an argument similar to the one in Theorem 2.5. Since \( P(y^*) = 0 \), we have \( y^{**}(t) = f(t,y^*(t)) \) a.e. in \([a,b] \). By the continuity of \( y^* \) and \( f \),

\[
F(t) = \int_{a}^{t} f(r,y^*(r)) \, dr
\]

is defined for all \( t \in [a,b] \) and differentiable a.e. in \([a,b] \). Hence,

\[
\frac{d}{dt}(y^*(t) - F(t)) = 0 \quad a.e. \quad [a,b],
\]

which yields, as in Theorem 2.5, that \( y^*(t) - F(t) \) is constant for all \( t \in [a,b] \).
Therefore,
\[ y''(t) = f(t, y'(t)) \quad \text{for all } t \in [a, b], \]
and the continuity of \( f \) gives \( y' \in C^1[a, b] \).

By Theorems 2.5 and 2.9, each linear problem in (2.3) can be uniquely solved for iterates \( y_k \in C^1[a, b] \). It follows that for each \( t \in [a, b] \),
\[
y_k'(t) - y_k''(t) = f_2(t, y_k(t)) y_k(t) + f(t, y_k(t)), \quad (2.32)
- f_2(t, y_k(t)) y_k(t) - f(t, y_k(t)).
\]
Since \( y_k \to y' \) in \( C[a, b] \) and \( f_2 \) is continuous on a compact set, the right hand side of (2.32) tends to zero, uniformly in \( t \), as \( k \) tends to infinity. Therefore, \( y_k \to y' \) in \( C^1[a, b] \).

We can now show that the McGill and Kenneth result follows from Theorem 2.9.

**Corollary 2.11**

Let \( f \) satisfy Assumption 2.1. If \( \begin{bmatrix} A \\ B \end{bmatrix} \) is nonsingular and \( b-a \) is sufficiently small, then for any \( M_y > 0 \), the method of quasilinearization (2.2) is well defined and converges in \( C^1[a, b] \) to a solution of (2.1) for any \( y_0 \in C^1[a, b] \) such that \( \| y_0 \|_{C^1[a, b]} \leq M_y \).

**Proof:**

We show that for \( b-a \) sufficiently small, we can make \( h \) in Theorem 2.9 less than \( \frac{1}{2} \). Since \( U(a) = I \), for \( b-a \) sufficiently small, we have that \( \begin{bmatrix} A \\ BU(b) \end{bmatrix} \) is nonsingular. This gives the unique existence of the solution \( e_0 \) to problem (2.18) and the bounds
\[
M_A \leq 2 \left\| \begin{bmatrix} A \\ B \end{bmatrix}^{-1} \begin{bmatrix} A \\ 0 \end{bmatrix} \right\|,
M_B \leq 2 \left\| \begin{bmatrix} A \\ B \end{bmatrix}^{-1} \begin{bmatrix} A \\ 0 \end{bmatrix} \right\|.
\]

Then, since \( M_F \) is a nondecreasing function of \( b \), we have that \( M_G, M_{G_i} \), and therefore, \( \beta \) are nondecreasing in \( b \). \( \eta \) is also a nondecreasing function of \( b \) since
\[
\eta \leq \beta \| P(y_0) \|_{L^2} \\
\leq \beta [2(b-a)(M_y + M_F)]^\beta.
\]

Finally, since \( L \), the Lipschitz constant for \( f_2 \), is nondecreasing in \( b \), we can make \( h = \beta K \eta < \frac{1}{2} \) for \( b-a \) sufficiently small. Therefore, the hypothesis of Theorem 2.9 is satisfied, and invoking Corollary 2.10, we have the desired result.

We note that in the Roberts and Shipman proof of the above corollary, they failed to specify a condition of the type
\[
\| y_0 \|_{C^1[a, b]} \leq M_y. \quad (2.33)
\]
Since \( y_0 \) can change as \( b \) decreases, it seems as if they fail to provide an upper bound on \( \eta \). McGill and Kenneth do stipulate a condition equivalent to (2.33). The condition is hardly restrictive since we can always take \( y_0 = 0 \) for any \( b \).

A result such as Corollary 2.11 is motivation for continuation methods that continue on the interval length. In the next chapter we prefer a different globalization strategy, one that generates iterates that we hope will enter a region in which the Kantorovich hypothesis is satisfied.

CHAPTER 3

The Model Trust Region Modification

In the previous chapter, a semilocal convergence theory based on the Kantorovich theorem for the quasilinearization process was presented. The distinction between local and semilocal is important; since for a semilocal analysis, no assumption is made as to the existence of a solution to the nonlinear problem. This is very satisfying when existence results are hard to come by, as in the case of nonlinear boundary value problems. A local result, on the other hand, does require that we assume something about the solution. For example, if we assume both the existence of a solution \( y^* \) of the problem \( P(y) = 0 \) and \( P'(y^*)^{-1} \), then it is not hard to show that the Kantorovich theorem gives the quadratic convergence of quasilinearization for any starting point \( y_0 \) close enough to \( y^* \). A separate proof of this result, not requiring the Kantorovich theorem, can also be obtained. (See Dennis (1971) for a further discussion on the distinction between local and semilocal results and their implications.)

Given a local or semilocal convergence result, the question then arises of how to choose the initial point \( y_0 \) to lie in the domain of attraction of Newton's method. It may happen that the poser of the boundary value problem may have sufficient information about the qualitative behavior of the solution, based
on physical considerations, say, to enable him to make a reasonable choice. How to choose \( y_0 \) in general, though, is still an open question. (See Keller (1976, p. 11).)

One popular approach is to imbed the original problem \( P(y) = 0 \) in a family of problems \( P(y, \lambda) = 0 \), depending on some parameter \( \lambda \). For a certain value of the parameter, \( \lambda_0 \), the problem \( P(y, \lambda_0) \) is usually easy to solve and an arbitrary initial guess \( y_0, \lambda_0 \) will usually suffice. For another value of the parameter, \( \lambda_* \), \( P(y, \lambda_*) = P(y) \). The idea is to then solve the family of problems \( P(y, \lambda) = 0 \) as \( \lambda \) varies from \( \lambda_0 \) to \( \lambda_* \), using the solution of the "neighboring" problem as the initial point of the next. There has been considerable interest in these continuation methods in recent years, (Watson, 1980), (Allgower and Georg, 1980). A major consideration is the method used to track the zero curve \( P(y^*, \lambda) \). By implicit differentiation, the curve can be parameterized by \( \lambda \) and shown to be the solution of an initial value problem. An A-stable integration technique is then used to track the curve. Since A-stable techniques are implicit, this necessitates the use of an iterative method to solve a nonlinear problem at each integration step. Usually some type of predictor-corrector method is used instead. Methods based on solving first order initial value problems, (Boggs, 1971), and second order initial value problems, (Zirilli, 1982), have been successful. Somewhat less successful has been the determination of the appropriate family of problems to solve. Naive imbeddings are almost always unsuccessful and the proper choice seems to be heavily problem dependent and quite an art. Recently, Watson and Scott (1985) have been successful in solving very difficult boundary value problems using a continuation method based on the Chow-Yorke algorithm. They solve the finite dimensional problem, however, that results from a discretization using B-spline collocation.

For quasilinearization, one popular continuation method is to continue on the interval of integration, (Roberts, Shipman and Roth, 1968), (Keller, 1976). This is reasonable, given Corollary 2.11. The approach here is to solve the original nonlinear boundary value problem over an increasing sequence of intervals \( [a, b_k] \), using the solution over the \( k^{th} \) interval as the initial point for the problem over the \( k+1^{st} \) interval, until, for some \( b_k \), \( b_k = b \). Roberts, Shipman and Roth report limited success with this approach. The most difficult problems are choosing the sequence of intervals and deciding how accurately to solve each intermediate nonlinear problem. The author is not aware of any work applying the curve tracking techniques directly to quasilinearization. The usual procedure is to first discretize the problem and then to apply the continuation technique to the resultant finite dimensional approximation.

Since the burden of handling the nonlinearity is usually placed on the finite dimensional problem, it is natural to ask if the globalization methods developed for these problems can be extended to the original nonlinear operator \( P \). In this chapter, we will consider one such method that has proven itself effective in globalizing Newton's method for the solution of nonlinear algebraic problems. It
is one of a class of iterative methods known as descent methods.

In a descent method, a merit function is associated with the nonlinear operator $P$. At each iteration, a step $s_k$ is selected as acceptable in providing convergence to the solution of (2.4), based on this merit function. A reasonable acceptance criteria is to require descent on the norm of the residual $P$. Let $\phi(y) = \|P(y)\|_2^2$ for all $y \in H^1_0[a,b]$. Since a solution of (2.4) is a global minimizer of $\phi$, we will accept a step in the iterative process if it will sufficiently decrease $\phi$. The major problem with this approach is that $\phi$ may have other local minima in addition to a solution $y^*$ of (2.4).

If $P$ is Fréchet differentiable in some domain, then so is $\phi$ with

$$
\phi'(y)(s) = \langle P'(y)(s), P(y) \rangle_{L^2}.
$$

Hence, if $P(y) \neq 0$, and $P(y)^{-1}$ exists, then the Newton step $s_N = -P'(y)^{-1}P(y)$ in (2.11) is a descent direction for $\phi$ since

$$
\phi'(y)(s_N) = -\|P(y)\|_2^2 < 0.
$$

Therefore, in step 2.2 of Algorithm (2.11), one could replace $y_{k+1} = y_k + s_k$ by $y_{k+1} = y_k + \lambda_k s_k$, where $\lambda_k \in \mathbb{R}$ can be chosen sufficiently small so that $\phi(y_{k+1}) < \phi(y_k)$.

This damped Newton scheme was implemented by Miele and Iyer (1970,1971) for two-point boundary value problems by choosing the largest $\lambda_k^{(m)} = \frac{1}{2^m}$ such that $\phi(y_k + \lambda_k^{(m)} s_k) < \phi(y_k)$. This approach does have the advantage of trying the full Newton step first so as to retain the quadratic convergence rate once the iterates enter a region in which the Kantotovich hypotheses are satisfied. Only requiring descent on $\phi$, however, is not sufficient to guarantee global convergence to a local minimizer of $\phi$.

P. Wolfe (Wolfe, 1960, 1971) gives sufficient conditions for a step in a descent direction to satisfy, that essentially ensure global convergence to a local minimizer. Two widely used conditions are given by

$$
\phi(y_k + \lambda_k s_k) \leq \phi(y_k) + \alpha \phi'(y_k)(\lambda_k s_k),
$$

and

$$
\phi'(y_k + \lambda_k s_k)(s_k) \geq \beta \phi'(y_k)(s_k),
$$

where $0 < \alpha < \beta < 1$. Condition (3.2) insures that sufficient descent is achieved on $\phi$, and (3.3) prevents the steps from becoming too small relative to the initial decrease in $\phi$, from $y_k$ in the direction $s_k$.

Wolfe showed that for any descent direction $s_k$, a $\lambda_k$ could be found to satisfy both (3.2) and (3.3). He also showed this was sufficient to guarantee that

$$
\phi'(y_k)\left(\frac{s_k}{\|s_k\|_{H^1_0[a,b]}}\right) \to 0.
$$

Since the domain of $P$ was chosen to be a Hilbert space, let us denote by $\nabla \phi(y_k) \in H^1_0[a,b]$ the Riesz reprenter of the bounded linear functional $\phi'(y_k)$, i.e. the gradient. If an algorithm can bound the step $s_k$ and the gradient $\nabla \phi(y_k)$ away from becoming orthogonal, then (3.4) implies that $\nabla \phi(y_k) \to 0$. 

Dennis and Moré (1974) showed that if the descent directions $s_k$ were chosen to approach the Newton step in both length and direction then, for some $k \geq 0$, $\lambda_k \equiv 1$ would satisfy (3.2) and (3.3), thereby eventually allowing the full step to be taken.

Most modern line search algorithms are based on these ideas. The most efficient are those that require the least number of the function evaluations in testing (3.2) and (3.3).

The author had previously implemented a linesearch modification to quasilinearization. The linesearch strategy used was the safeguarded cubic backtracking linesearch due to Dennis and Schnabel (1983). The method uses the function evaluations needed to test condition (3.2) to model $\phi(y_k + \lambda s_k)$ by a cubic polynomial. The minimizer of this cubic is chosen as the next candidate. The first such $\lambda_k$ to satisfy (3.2) becomes the final steplength parameter at iteration $k$. Since the candidate steplength parameters decrease no faster than a geometric sequence, it can be shown (Arndt, 1966), that the steplength chosen at iteration $k$ will not be too small. Therefore, condition (3.3) need not be tested, saving derivative evaluations. Finally, safeguards are implemented to protect the linesearch when the cubic model is poor.

Numerical experimentation did show that the linesearch modification allowed convergence from a poor initial point $y_0$. Through most of the iterations, however, only a small fraction of the Newton step could be taken. This resulted in very slow progress into a region where the full Newton steps were accepted. The problem is that the Newton step is not necessarily the optimal direction to take when far from a solution.

This leads us to consider an alternate globalization strategy, one that has proven effective in solving nonlinear problems in $\mathbb{R}^n$; namely, the model trust region modification to Newton's method. These methods are robust in the sense that they are not too dependent on the choice of parameters used in an implementation. They also have a very elegant justification.

Moré (1982) gives a history of model trust region methods and an overview of recent developments in their use for solving nonlinear problems in $\mathbb{R}^n$. He traces the methods back to the work of Levenberg (1944) and Marquardt (1963) for nonlinear least squares. Goldfeldt, Quandt and Trotter (1966) used the approach for the more general unconstrained minimization problem. (These early methods differ from modern versions in that they seek to control the Lagrange multiplier associated with the model trust region subproblem. Many of the difficulties with this approach can be overcome by modern implementations, which control, instead, the step bound in the subproblem.) Powell (1970a,1970b) used the method for nonlinear equations and unconstrained optimization with secant approximations for the derivative matrices. He restricted the solution to the subproblem to lie in the space spanned by the gradient and quasi-Newton steps; in what is known as the dogleg implementation. (See Moré (1982) for an account of the main developments in the area of model trust regions since these early works.)
To motivate the model trust region approach we again consider Newton’s method. Newton’s method may be viewed as providing a local affine model of $P$ about $y_k$, at each iteration $k$, i.e.

$$P(y_k+s_k) \approx P'(y_k)s_k + P(y_k).$$

Since the approximation is, in general, only good in some region about $y_k$, we consider solving the Newton equation

$$P'(y_k)s_k + P(y_k) = 0 \quad (3.5)$$

as well as possible, subject to the condition that we remain in the region in which we trust the affine model to be a valid approximation to $P$. This then leads us to consider solving, instead of (3.5), the following locally constrained problem

$$\text{minimize } \frac{1}{2} \| P'(y_k)s + P(y_k) \|_{L^2}^2, \quad (LCP)$$

subject to $\| s \|_{H^2}^2 \leq \Delta_k^2, \quad (3.6)$

for the step $s_k$. Here, $\Delta_k$ is the radius of the closed neighborhood of $y_k$ in which we trust the affine approximation. The solution of (LCP) is then accepted, or rejected, as the step $s_k$, based on a sufficient descent criteria for $\phi$. If rejected, the radius is decreased and problem (LCP) is solved with the new step bound.

If the solution to (LCP) is accepted as the step $s_k$, we update the iterate $y_{k+1} = y_k + s_k$ and, possibly, increase the trust radius. To be more specific, let

$$\psi(y_k+s) = \frac{1}{2} \| P'(y_k)s + P(y_k) \|_{L^2}^2$$

$$= \phi(y_k) - \langle \nabla \phi(y_k), s \rangle_{H^1} - \frac{1}{2} \langle s, B(y_k)s \rangle_{H^1},$$

where $\nabla \phi(y_k) = P'(y_k)^*P(y_k)$ and $B(y_k) = P'(y_k)^*P'(y_k)$. Here, $P'(y_k)^*$ is the Hilbert adjoint of $P'(y_k)$. Also, we define

$$\text{ared}_k = \phi(y_k) - \phi(y_k + s_k),$$

$$\text{pred}_k = \phi(y_k) - \psi(y_k + s_k),$$

$$\rho_k = \frac{\text{ared}_k}{\text{pred}_k}.$$

We could use condition (3.2) to determine whether to accept the solution to (LCP) as the new step. However, it has been suggested that the condition

$$\rho_k \geq \alpha \quad (3.7)$$

for some small $\alpha > 0$, is to be preferred since it also includes the curvature of $\phi$.

This can be seen by noticing that (3.2) can be written

$$\frac{\phi(y_k) - \phi(y_k + s_k)}{-\langle \nabla \phi(y_k), s_k \rangle_{H^1}} \geq \alpha,$$

and (3.7) leads to

$$\frac{\phi(y_k) - \phi(y_k + s_k)}{-\langle \nabla \phi(y_k), s_k \rangle_{H^1} - \frac{1}{2} \langle s_k, B(y_k)s_k \rangle_{H^1}} \geq \alpha.$$
will prevent the steps from becoming too small, too rapidly. This is similar to the backtracking argument referred to earlier. By decreasing the radius, we decrease the region in which we believe the affine model to be accurate.

If $\rho_k$ is close to one then the radius can be increased for the next iteration since the model $\psi$ provides a good fit to $\phi$. If $\rho_k$ happens to be greater than one, then negative curvature in $\phi$ is indicated. In this case, the radius should be increased to take advantage of this negative curvature. We summarize the above discussion by presenting the model trust region algorithm we will consider. The following algorithm is adapted from Moré and Sorensen (1981, 1982).

**Algorithm 3.1**

0) Set $0 < \eta_1 < \eta_2 < 1, \ 0 < \gamma_1 \leq \gamma_2 < 1 \leq \gamma_3$.

Set $\Delta > 0, \ 0 < \Delta_0 < \Delta$.

Find $y_0 \in H_{y} | s, k$. Calculate $\phi(y_0)$.

1) For $k=0, 1, \ldots$, until convergence do

1.1) Set $j=0$ and $\Delta_j = \Delta_0$.

1.2) Find an approximate solution $s_k^{(j)}$ to the problem

\[
\begin{align*}
\text{minimize} & \quad \psi(y_k + s) \\
\text{subject to} & \quad ||s||_H^2 \leq \Delta_j^2.
\end{align*}
\]

1.3) Calculate $\phi(y_k + s_k^{(j)})$ and $\phi(y_k + s_k^{(0)})$.

1.4) If $\frac{\text{ared}_k}{\text{pred}_k} < \eta_1$ then

1.4.1) Set $\gamma_1 \Delta_j \leq \Delta_{j+1} \leq \gamma_2 \Delta_j$, set $j=j+1$ and go to 1.2.

Else

1.4.2) Set $y_{k+1} = y_k + s_k^{(j)}$.

If $\frac{\text{ared}_k}{\text{pred}_k} > \eta_2$ then

1.4.2.1) Set $\Delta_j \leq \Delta_{j+1} \leq \min \{ \Delta, \gamma_3 \Delta_j \}$.

Else

1.4.2.2) Set $\Delta_{j+1} = \Delta_j$.

Set $k = k + 1.$
For the remainder of this chapter, we assume the following regularity condition on $f$ in (2.1).

**Assumption 3.2**

Let $D = [a,b] \times \mathbb{R}^n$. We assume that $f_2$ exists and is Lipschitz continuous, in $D$, with respect to its second variable $y$, uniformly in $t$, i.e., there exists a constant $L > 0$ such that $\|f_2(t,y) - f_2(t,z)\| \leq L \|y-z\|_2$ for all $(t,y),(t,z) \in D$.

We further make the assumption that $f_2$ is bounded in $D$, i.e., there exists a constant $M$ such that $\|f_2(t,y)\| \leq M$ for all $(t,y) \in D$.

This differs from the assumptions in Chapter 2 only in that we assume the regularity on the whole set $D$.

We will now consider the solution of the subproblem (LCP) used in step 1.2 of the algorithm. We first state the following theorem, giving sufficient conditions for the subproblem to have a unique solution.

**Theorem 3.3**

Under Assumption 3.2, problem (LCP) has a solution for all $y_k \in H^2_0[a,b]$. If, in addition, the matrix $\begin{bmatrix} A \\ BU_k(b) \end{bmatrix}$ is nonsingular, then the solution is unique.

Here, $U_k(b)$ is the solution to the initial value problem (2.19), evaluated at $b$, where $F(t) = f_2(t,y_k(t))$.

**Proof:**

By Theorem 2.4 and Assumption 3.2, $P'(y_k)$ exists as a bounded linear operator for any $y_k \in H^2_0[a,b]$. Hence, $\psi(y_k + s)$ is continuous in the variable $s$ for all $s \in H^2_0[a,b]$. $\psi(y_k + s)$ is easily seen to be twice Gâteaux differentiable with

$$\psi''(y_k + s)(\eta)(\eta) = \langle \eta, B(y_k)\eta \rangle_{H^2_0} = \|P'(y_k)\eta\|_{L^2} \geq 0,$$

which implies that $\psi(y_k + s)$ is convex. Clearly the constraint set (3.6) is norm closed, bounded, and convex. Since $\psi$ is defined on a Hilbert space, problem (LCP) has at least one solution (see Tapia and Thompson (1978, p.182), for example).

Since $\begin{bmatrix} A \\ BU_k(b) \end{bmatrix}$ is nonsingular, Lemma 2.7 gives that $P'(y_k)^{-1}$ exists as a bounded linear operator. This gives

$$\langle \eta, B(y_k)\eta \rangle_{H^2_0} = \|P'(y_k)\eta\|_{L^2} > 0,$$

for all $\eta \neq 0$. Hence, $\psi''(y_k + s)$ is positive definite and so the solution to (LCP) is unique.

Using Lagrange multiplier theory, it can be shown (Daniel, 1971, p.12) that a solution $s(\Delta_k)$ to problem (LCP) must solve
\[
\begin{align*}
\{ P'(y_k)'P'(y_k) + \lambda I \} s(\Delta_k) &= -P'(y_k)'P(y_k) \\
\lambda \{ \| s(\Delta_k) \|_{H_0^2}^2 - \Delta_k^2 \} &= 0,
\end{align*}
\]

for some \( \lambda \geq 0 \). Here, \( I \) denotes the identity operator in \( H_0^1[a,b] \). Because of the special form of this constrained problem, a complete characterization of problem (LCP) can be obtained from (3.8) (Sorensen, 1982b).

Hebden (1973), Moré (1981), Sorensen (1982b), Moré and Sorensen (1982), and Shultz, Schnabel, and Byrd (1985) have considered the numerical solution of problem (LCP) for minimization problems in \( \mathbb{R}^n \) in the general case when \( \psi \) is any quadratic with a symmetric Hessian. Their methods have to contend with the possibility of an indefinite Hessian. In our case, \( \psi \) is convex, so we can consider the dogleg approach (Powell, 1970a) in approximately solving subproblem (LCP). Specifically, we use the double dogleg method of Dennis and Mei (1979).

In this method, the curve \( s(\Delta) \) is approximated by a piecewise linear curve \( \tilde{s}(\tau) \). This curve uses the Newton step \( s_N = -P'(y_k)^{-1}P(y_k) \) (i.e. the solution to the linear two-point boundary value problem in (2.3)), as well as the Cauchy step

\[ s_c = -\frac{\| g_k \|_{H_0^2}^2}{\| P'(y_k)g_k \|_{L^2}^2} g_k, \]

where \( g_k = \nabla^2 \phi(y_k) = P'(y_k)'P(y_k) \). The Cauchy point is the minimizer of \( \psi \) in the steepest descent direction \(-\nabla \phi(y_k)\). A third point \( \gamma s_N, 0 < \gamma \leq 1 \), is also introduced. \( \tilde{s}(\tau) \) then is given by

\[ \tilde{s}(\tau) = \begin{cases} 
\tau s_c, & 0 \leq \tau \leq 1 \\
\gamma \tau - 1 s_N - s_c, & 1 < \tau \leq 1 + \gamma \\
(\tau - 1) s_N, & 1 + \gamma < \tau \leq 2.
\end{cases} \]

If \( P'(y_k)^{-1} \) exists, then this curve is well defined and runs, as \( \tau \) increases, from \( y_k \) to \( s_c \) to \( \gamma s_N \) to \( s_N \). It can also be shown (Dennis and Schnabel, 1983) that \( \psi(y_k + \tilde{s}(\tau)) \) is monotonically decreasing in \( \tau \), while \( \| \tilde{s}(\tau) \|_{H_0^2} \) is monotonically increasing. To find the solution to problem (LCP), we proceed as follows. If \( \| s_N \|_{H_0^2} \leq \Delta_k \), then the solution to (LCP) is given by \( s_N = \tilde{s}(2) \).

Otherwise, the solution is given by that point on the curve on the boundary of the trust region, i.e. \( \tilde{s}(\tau) \) such that \( \| \tilde{s}(\tau) \|_{H_0^2} = \Delta_k \).

In the case when \( s_N \) does not exist or is not unique, the default solution to (LCP) will be given by \(- \frac{\Delta_k}{\| y_k \|_{H_0^2}^2} g_k \).

This strategy has the advantage of taking the full Newton step when the trust region allows it. When the trust region is large, steps are permitted that are close to the Newton step in length and direction. When the trust region is small, but the Newton step is large, steps are chosen that are small and point more in the direction of steepest descent. The point \( \gamma s_N \) is introduced to give the step a bias in the Newton direction. Computational results (Dennis and Mei, 1979) show this is usually an improvement over the dogleg implementation.
In Chapter 2, the Newton step was defined as the solution of a linear two-point boundary value problem. To define the Cauchy step, it remains for us to derive the gradient $\nabla \phi(y_k)$. The following derivation follows the technique given in Tapia (1971, pp. 81-82).

Recalling that $\nabla \phi(y_k)$ is the Riesz representor of the linear functional $\phi'(y_k)$, (3.1) gives

$$<P'(y_k)\eta,P(y_k)>_{L^2} = <\nabla \phi(y_k)\eta>_{H^1_0}$$

$$= <\nabla \phi(y_k)\eta' + \nabla \phi(y_k)(a)^T \eta(a), \eta>_{L^2}$$

for all $\eta \in H^1_0[a,b]$. By (2.9),

$$<P'(y_k)\eta,P(y_k)>_{L^2} = <\eta',P(y_k)>_{L^2} - <\eta, \frac{d}{dt} P(y_k)>_{L^2}.$$  \hspace{1cm} (3.10)

Put $G(t) = \int_a^t \frac{d}{dt}(y_k(\tau))P(y_k)(\tau) \, d\tau$ and $H(t) = G(b) - G(t)$. Integration by parts yields

$$<\eta, \frac{d}{dt} P(y_k)>_{L^2} = -<\eta', G>_{L^2} + \eta(b)^T G(b).$$

Since $\eta$ is absolutely continuous, we can write

$$\eta(b) = \eta(a) + \int_a^b \eta' \, d\tau.$$

(3.10) then becomes

$$<P'(y_k)\eta,P(y_k)>_{L^2} = <\eta', P(y_k) - H>_{L^2} - \eta(a)^T G(b).$$  \hspace{1cm} (3.11)

Since $\eta$ satisfies the boundary conditions, we may write

$$0 = [A] \eta(a) + [B] \eta(b)$$

$$= [A] \eta(a) + \int_a^b [B] \eta'(\tau) \, d\tau.$$

Therefore, for all $c \in R^n$,

$$<\eta', [B]^T c>_{L^2} = -<\eta(a)^T [A]^T c>.$$  \hspace{1cm} (3.12)

Substituting (3.12) into (3.11), we have the identity

$$<P'(y_k)\eta,P(y_k)>_{L^2} = <\eta', P(y_k) - H + [B]^T c >_{L^2}$$

$$+ \eta(a)^T \left[ [A]^T c - H(a) \right].$$

Since $\nabla \phi(y_k)$ satisfies (3.9), (3.13) gives

$$\nabla \phi(y_k) = \int_a^b P(y_k)(\sigma) - H(\sigma) \, d\sigma + \left( (t-a) [B] + [A]^T \right) c - H(a).$$  \hspace{1cm} (3.14)

By requiring $\nabla \phi(y_k)$ to satisfy the boundary conditions, we determine $c$. That is,

$$[A] \nabla \phi(y_k)(a) + [B] \nabla \phi(y_k)(b) = 0$$

implies that $c$ must solve the $n \times n$ linear system

$$Mc = [A] H(a) - [B] \int_a^b P(y_k)(\sigma) - H(\sigma) \, d\sigma,$$  \hspace{1cm} (3.15)

where

$$M = [A] [A]^T + (b-a) [B] [B]^T.$$  \hspace{1cm} (3.16)

Under Assumption 3.2, $\phi$ is Fréchet differentiable for all $y \in H^1_0[a,b]$. Since $\phi$ is
defined on a Hilbert space, \( \nabla \phi(y_k) \) exists and is unique. Hence, (3.15) must have a unique solution.

To show directly that \( M \) is nonsingular, let \( x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^n \) with \( x_1 \in \mathbb{R}^{n-k} \), \( x_2 \in \mathbb{R}^k \), and put \( D = \begin{bmatrix} A \\ B \end{bmatrix} \begin{bmatrix} A & B \end{bmatrix}^T \). Clearly, \( M \) is positive semidefinite. Suppose \( x^T M x = 0 \). (3.16) then gives

\[
x^T D x + (b-a)x_2^T B^T x_2 = 0.
\]  
(3.17)

Since \( D \) is positive semidefinite, \( B \) is of full rank, and \( (b-a)>0 \), we must have \( x_2 = 0 \). From (3.17), \( x_1^T A A^T x_1 = 0 \). Since \( A \) is also of full rank, \( x_1 = 0 \). Therefore, \( M \) is positive definite.

In Chapter 5, we will discuss the numerical solution for \( \nabla \phi(y_k) \), via (3.14) and (3.15).

We will now show that under the regularity conditions given in Assumption 3.2, Algorithm 3.1 will generate a sequence of iterates such that \( \nabla \phi(y_k) \to 0 \). Powell(1970a) first showed that \( \liminf_{k \to \infty} \| \nabla \phi(y_k) \| = 0 \), for trust region methods that select steps that reduce the quadratic model \( \psi \) at least as much as the constrained steepest descent step. Thomas(1975), by requiring the \( \eta_1 \) in Algorithm 3.1 to be strictly greater than zero, showed that, in fact, \( \nabla \phi(y_k) \to 0 \).

A proof of this result, due to Schultz, Schnabel, and Byrd(1985), will remain valid in the present setting. To demonstrate this, we need to establish some inequalities that are critical to their proof.

Since the double dogleg step causes as much descent in \( \psi \) as the constrained steepest descent step, we have the inequality (Lemma 4.1 in Thomas(1975) or Lemma 4.1 in Schultz, Schnabel and Byrd(1985))

\[
pred_k \geq \frac{1}{2} \| \nabla \phi(y_k) \|_{H_k} \min \{ \Delta_k, \| P'(y_k) \|_{P'(y_k)} \}, \tag{3.18}
\]

We also need to show there exists a constant \( K>0 \) such that

\[
\left| \frac{\text{ared}_k}{\text{pred}_k} - 1 \right| \leq \frac{K \| s_k \|_{H_k}^2}{|\text{pred}_k|}. \tag{3.19}
\]

Under Assumption 3.2, Lemma 2.8 gives

\[
\| P'(y)-P'(x) \| \leq L \sigma \| y-x \|_{H_k}
\]

for all \( y, z \in H_0, [a, b] \), where \( L \) is the Lipschitz constant for \( f_2 \) and \( \sigma = \min \{ (b-a)^2, (b-a) \} \). From Theorem 2.4 we also have

\[
\| P'(y) \| \leq B, \text{ for all } y \in H_0, [a, b],
\]

where \( B = \max\{1, M \sqrt{\sigma} \} \) and \( M \) is the bound on the norm of \( f_2 \) given in Assumption 3.2. By the continuity of \( P' \) in \( H_0, [a, b] \) it follows that

\[
P(z) = P(y) + \int_0^1 P'(y + \xi(z-y)) (z-y) \, d\xi,
\]
and so

\[
\| P(z) - P(y) \|_{L^2} \leq B \| z-y \|_{H_k}, \text{ for all } y, z \in H_0, [a, b].
\]
Let $L_0 = \{ y \in H_0^1[a,b] : \phi(y) \leq \phi(y_0) \}$. Then for any $y \in L_0$, and $z, s \in H_0^1[a,b],$

$$|\phi'(y)(z) - \phi'(y)(s)| \leq |<P'(y)(z), P(y)>_{L^2} - <P'(y)(s), P(y)>_{L^2}|$$

$$+ |<P'(y)(z), P(y)>_{L^2} - <P'(y)(s), P(y)>_{L^2}|$$

$$\leq (L\sigma \| P(y_0) \|_{L^2} + B^2) \| y - z \|_{H_0^1} \| s \|_{H_0^1}.$$

Hence, we have the following Lipschitz condition for $L_0$:

$$\| \nabla \phi(y) - \nabla \phi(z) \|_{H_0^1} = \| \phi'(y) - \phi'(z) \|_{H_0^1} \leq \sup_{s \neq 0} \frac{|\phi'(y)(s) - \phi'(z)(s)|}{\| s \|_{H_0^1}}$$

$$\leq L \| y - z \|_{H_0^1}, \text{ for all } y \in L_0, s \in H_0^1[a,b],$$

with $L = L\sigma \| P(y_0) \|_{L^2} + B^2$.

Since $y_k \in L_0$ for all $k \geq 0$, the Lipschitz condition (3.20) for $L_0$ implies

$$|\phi(y_k + s_k) - \phi(y_k) - <\nabla \phi(y_k), s_k>_{H_0^1}| \leq \int_0^1 <\nabla \phi(y_k + \xi s_k) - \nabla \phi(y_k), s_k>_{H_0^1} \ d\xi$$

$$\leq \frac{L}{2} \| s_k \|_{H_0^1}^2.$$

Finally, since $<s_k, B(y_k)s_k>_{H_0^1} = \| P'(y_k)s_k \|_{L^2} \leq B^2 \| s_k \|_{H_0^1}^2$ we have

$$|\text{pred}_k - \text{pred}_{k+1}| \leq K \| s_k \|_{H_0^1}^2,$$

where $K = \frac{L}{2} + B^2$ and we see that (3.19) holds.

We may now state the global convergence result.

**Theorem 3.4**
CHAPTER 4

Inexact Newton’s Method

In the previous two chapters, Newton’s method and a globalization modification were presented for the nonlinear two point boundary value problem. In practice, the problem must be discretized for its numerical solution. This chapter will be concerned with the discretization technique and its effect on the convergence of the iterative method.

Dembo, Eisenstat, and Steihaug (1982) give conditions for an approximate solution to the Newton equation (3.5) to satisfy in order to guarantee local convergence. Since any technique for solving the infinite dimensional linear problem can only yield an inexact solution, the Inexact Newton framework is an appropriate setting to analyze the technique. Their criterion in deciding how accurately to solve the linear equation (3.5) is in terms of the relative residual. If \( s_k \) is an approximate solution to (3.5), let

\[
r_k = P'(y_k)s_k + P(y_k),
\]

and define the relative residual as

\[
\frac{\|r_k\|}{\|P'(y_k)\|}.
\]

To define the Inexact Newton method, let \( P : B_1 \rightarrow B_2 \) be a mapping between Banach spaces. Assume there exists \( y^* \in B_1 \) such that \( P(y^*) = 0 \). Furthermore, assume that \( P \) is continuously Fréchet differentiable in a neighborhood of \( y^* \) and that \( P'(y^*)^{-1} \) exists. We define the Inexact Newton sequence by

\[
y_{k+1} = y_k + s_k
\]

where

\[
\rho_k = \frac{\|r_k\|}{\|P'(y_k)\|} \leq \eta_k.
\]

Dembo, Eisenstat, and Steihaug (1982) established the following local convergence result for the Inexact Newton iteration (4.1),(4.2).

**Theorem 4.1**

If \( \eta_k \) is chosen so that \( \eta_k \leq \tilde{\eta} < t < 1 \), then

1) There exists \( \epsilon > 0 \) such that if \( \|y_0 - y^*\| < \epsilon \) then \( y_k \rightarrow y^* \) in \( B_1 \) with

\[
\|P'(y^*)(y_{k+1} - y^*)\| \leq t \|P'(y^*)(y_k - y^*)\|.
\]

If we assume that \( y_k \rightarrow y^* \), we have

2) \( \rho_k \rightarrow 0 \) if and only if the convergence rate is \( q \)-superlinear, i.e.,

\[
\lim_{k \rightarrow \infty} \frac{\|y_{k+1} - y^*\|}{\|y_k - y^*\|} = 0.
\]

3) If \( P' \) is Lipschitz continuous, then \( \rho_k = O(\|P(y_k)\|) \) if and only if the convergence is \( q \)-quadratic, i.e.,
\[ \lim_{k \to \infty} \frac{\|y_{k+1} - y\|}{\|y_k - y\|^2} < \infty. \]

Krasnoselskii and Rutickii (1961) first proved a Kantorovich type theorem for Inexact Newton's method with error bounds on the relative errors. Dennis and Moré (1974) first characterized \( q \)-superlinear convergence as requiring the approximate step \( s_k \) to approach the Newton step in both length and direction. Altman (1980) proved a Kantorovich type theorem for a damped Inexact Newton method using the relative residual criterion (4.2).

We would like to use Theorem 4.1 to determine a discretization to approximately solve the linear boundary value problem given by the Newton equation. Dembo, Eisenstat, and Steihaug (1982) have noted that Theorem 4.1 is constructive in that it specifies how we may choose the forcing sequence \( \eta_k \). It is sufficient that we choose

\[ \eta_k = \min \left\{ \frac{1}{2} c_1 \|P(y_k)\| \right\} \quad (4.3) \]

for some constant \( c_1 > 0 \) to retain the \( q \)-quadratic convergence of Newton's method.

In our case, we propose a method to discretize and approximately solve

\[ P'(y_k)s_k + P(y_k) = 0, \quad s_k \in H^1_0[a,b], \]

or, equivalently,

\[ s_k'(t) = f_d(t,y_k(t))s_k(t) + (f(t,y_k(t)) - y_k'(t)), \quad t \in [a,b], \quad (4.4) \]

\[ A_{s_k}(a) = 0, \quad B_{s_k}(b) = 0, \]

so that

\[ \|P'(y_k)s_k + P(y_k)\|_{L^\infty} \leq \eta_k \|P(y_k)\|_{L^\infty}, \quad (4.5) \]

where \( \eta_k \) is given in (4.3).

To do so, we consider solving (4.4) by the method of reduced superposition (see Scott and Watts (1977a)). This will reduce the problem to solving a set of initial value problems. To solve the initial value problems, we use a Runge-Kutta-Fehlberg pair of integration formulas to control the stepsize, so that (4.5) is satisfied with a minimal amount of work. Finally, the method of quasilinearization and the variable step integration method require that we have available the value of \( y_k(t) \) at each \( t \in [a,b] \) in order to solve (4.4) for \( y_{k+1} = y_k + s_k \).

The method of reduced superposition proceeds as follows. From the theory of linear ordinary differential equations, it is known (see, for example, Coddington and Levinson, 1955) that any solution to (4.4) may be written as

\[ s_k(t) = \sum_{i=1}^{n} \gamma_i u_i(t) + v(t), \quad t \in [a,b], \]

where the \( u_i \) are linearly independent solutions to the homogeneous problems

\[ u_i(t) = F(t) u_i(t), \quad i = 1, \ldots, n, \]

and \( v \) is a particular solution to the inhomogeneous problem

\[ v'(t) = F(t) v(t) + g(t), \]

where

\[ F(t) = f_d(t,y_k(t)) \]

\[ g(t) = f(t,y_k(t)) - y_k'(t). \]
If we choose \( u_1(a), \ldots, u_k(a) \) to be \( k \) linearly independent vectors in the null space of \( A \), and choose \( v \) to also lie in the null space of \( A \), we find that the initial condition \( A\gamma(a) = 0 \) implies

\[
A\gamma(a) = \sum_{i=k+1}^{n} \gamma_i A u_i(a) = 0,
\]
which gives \( \gamma_i = 0 \) for \( i = k+1, \ldots, n \). Since we require \( B\gamma(b) = 0 \), we solve the \( k \times k \) linear system

\[
BU(b)\gamma = -Bu(b) \quad (4.6)
\]
for the coefficients \( \gamma_i \). Here, \( U(t) \) is the matrix \( [u_1(t), \ldots, u_k(t)] \) and \( \gamma \) the vector of coefficients \( \gamma = [\gamma_1, \ldots, \gamma_k]^T \).

Hence, the solution to (4.4) is determined by solving the set of initial value problems

\[
U'(t) = F(t)U(t), \quad t \in [a,b], \quad AU(a) = 0, \quad (4.7)
\]
and

\[
v'(t) = F(t)v(t) + g(t), \quad t \in [a,b], \quad Av(a) = 0. \quad (4.8)
\]
To satisfy the final boundary condition we solve the linear system (4.6) and set

\[
s_i(t) = U(t)\gamma + v(t), \quad t \in [a,b]. \quad (4.9)
\]

Again, it can be argued, as in Chapter 2, that the existence and uniqueness of a solution to (4.4) is determined by the existence and uniqueness of a solution to (4.6), since the initial value problems (4.7) and (4.8) have unique solutions whenever \( F \) is continuous in \( t \in [a,b] \).

We shall rewrite (4.7) and (4.8) as the system of initial value problems

\[
u'(t) = k(t, u(t)), \quad t \in [a,b]
\]
\[
u(a) = u_0.
\]
To solve (4.10), we consider a single step integration method

\[
u_{n+1} = u_n + h_n \psi(h_n, \xi_n, \xi_n, u_n)
\]
\[
u_0 = u(a),
\]
where \( t_0 = a \), \( t_{n+1} = t_n + h_n \) for \( n = 0, \ldots, N \) and \( t_{N+1} = b \).

Once (4.10) is integrated, we choose the Hermite cubic interpolant \( \tilde{u}(t) \) as the approximate solution to (4.10), where we require the interpolation conditions

\[
\tilde{u}(t_n) = u_n,
\]
\[
\tilde{u}'(t_n) = k(t_n, u_n) = u_n',
\]
for \( n = 0, 1, \ldots, N+1 \).

A Runge-Kutta formula of order \( r \) will be used in (4.11) while a second formula, of order \( r+1 \), will be used to choose the step size \( h_n \). The pair of formulas considered are those given by Fehlberg (1970). The \( r+1 \)st order formula may be had with a single extra evaluation of \( k \).

Let \( r(t) = \tilde{u}'(t) - k(t, \tilde{u}(t)) \), \( t \in [a,b] \). The step selection strategy will try to ensure that

\[
\| r \|_{\infty} < \epsilon
\]
for a prescribed residual tolerance \( \epsilon \). To do so, we consider a function \( z(t) \) that
has the properties of interpolating the data points \( \{ u_n \} \) and giving a small residual for problem (4.10). We then show that by choosing the step size \( h_n \) correctly, \( r \) will be close to the residual given by \( x \).

Gear (1971, pp. 64-65) shows how such a function \( z \) may be defined. If \( k \) in (4.10) has \( r+2 \) continuous derivatives, then we may write the local truncation error for \( \psi \) in (4.11) as

\[
h \psi(h; t, u(t)) - \left( u(t+h) - u(t) \right) = h^{r+1} \psi(t, u(t)) + O(h^{r+2}).
\]

On each interval \([t_n, t_{n+1}]\), let \( u(t_n + \tau, u_n) \) be the solution to the local initial value problem

\[
\frac{d}{d\tau} u(t_n + \tau, u_n) = k(t_n + \tau, u(t_n + \tau, u_n))
\]

\[
u(t_n + \tau, u_n) = u_n.
\]

The local error for \( u(t_n + \tau, u_n) \) is then given by

\[
u_n + h \psi(t_n, u_n) - u(t_n + \tau, u_n) = h^{r+1} \left[ \phi(t_n, u_n) + \tau \tilde{T}(t_n, u_n) \right].
\]

Let

\[
T(t_n, u_n) = \phi(t_n, u_n) + \tau \tilde{T}(t_n, u_n),
\]

and define \( z_n \) on \([t_n, t_{n+1}]\) by

\[
z_n(t_n + \tau) = h^{r} T(t_n, u_n) + u(t_n + \tau, u_n).
\]

For \( t \in [a, b] \) we will define \( z \) by \( z(t) = z_n(t_n + \tau) \) where \( t = t_n + \tau \). Notice that \( z \) interpolates the data points \( \{ u_n \} \) since

\[
z(t_n) = z_n(t_n) = u(t_n; t_n, u_n) = u_n
\]

and

\[
z(t_{n+1}) = z_n(t_{n} + h_n)
\]

\[
= h_n^{r+1} T(h_n; t_n, u_n) + u(t_n + h_n; t_n, u_n)
\]

\[
= u_n + h_n \psi(h_n; t_n, u_n)
\]

\[
= u_{n+1}.
\]

Gear (1971, pp. 64-65) shows that on each \([t_n, t_{n+1}]\), \( z \) gives a small residual, i.e.,

\[
\| z_n' (t_n + \tau) - k(t_n + \tau, z_n(t_n + \tau)) \| = h_n^{r} \| \psi(t_n, u_n) \| + O(h_n^{r+1}).
\]

(4.13)

Since \( z \) is defined piecewise on \([a, b]\), \( z' \) has jump discontinuities at the points \( \{ t_n \} \). We would like to compare \( r \) to the residual given by \( z \). We define \( \tilde{z}_n(t) \) for \( t \in [t_n, t_{n+1}] \) to be the Hermite cubic that interpolates \( z(t_n) = u_n \), \( z(t_{n+1}) = u_{n+1} \), \( z'(t_n) \) and \( z'(t_{n+1}) \). Here, \( z'(t_n) \) and \( z'(t_{n+1}) \) are to be understood as the right and left derivatives of \( z \) at \( t_n \) and \( t_{n+1} \), respectively.

We will first estimate \( \| \tilde{u}(t) - \tilde{z}_n(t) \| \) on \([t_n, t_{n+1}]\). Since both \( \tilde{u} \) and \( \tilde{z}_n \) are Hermite cubic interpolants on \([t_n, t_{n+1}]\), we can write

\[
\tilde{u}(t) - \tilde{z}_n(t) = h_n \left[ a_n(t) (\tilde{u} - \tilde{z}_n)(t_n) + b_n(t)(\tilde{u} - \tilde{z}_n)'(t_n) \right]
\]

(4.14)

where

\[
a_n(t) = z_n(t) - 2z_n(t)^2 + z_n(t)^3
\]

\[
b_n(t) = -z_n(t)^2 + z_n(t)^3
\]

and
\[ z_n(t) = \frac{(t-t_n)}{h_n} . \]

(4.14) is just the Hermite interpolation formula for \( \tilde{u} - \tilde{z}_n \) since \( \tilde{u}(t_n) = z_n(t_n) = u_n \) and \( \tilde{u}(t_{n+1}) = \tilde{z}_n(t_{n+1}) = u_{n+1} \). From (4.12) we find that

\[
z_n'(t_n) = h_n^{-1} T(\tau; t_n, u_n) + h_n^{-1} \frac{d}{dt} T(\tau; t_n, u_n)
+ k(t_n + \tau u(t_n + \tau t_n, u_n)).
\]

where \( T(\tau; t_n, u_n) \) is continuously differentiable if \( k \) has \( r+3 \) continuous derivatives. Hence,

\[ z_n'(t_n) = h_n^{-1} \phi(t_n, u_n) + u_n' , \]

and so

\[
u_n'(t_n) - \tilde{z}_n'(t_n) = u_n' - z_n'(t_n) = h_n^{-1} \phi(t_n, u_n) . \tag{4.15}
\]

We also have

\[ z_n'(t_{n+1}) = z_n'(t_n + h_n)
= h_n^{-1} T(h_n; t_n, u_n) + h_n^{-1} \frac{d}{dt} T(h_n; t_n, u_n)
+ k(t_n + h_n u(t_n + h_n, t_n, u_n)). \]

Now, \( k \) is Lipschitz continuous in its second variable since \( F \) is continuous on

\([a,b] \). If we denote the Lipschitz constant by \( L \) we have

\[
\| \tilde{z}_n'(t_{n+1}) - u_{n+1}' \|
\leq h_n^{-1} \| \phi(t_n, u_n) \|
+ L \| u_{n+1}' - u(t_n + h_n, t_n, u_n) \|
+ ch_n^{-1}
\]

for some constant \( c > 0 \) independent of \( h_n \). Recalling that

\[ u_{n+1}' - u(t_n + h_n, t_n, u_n) = h_n^{-1} T(h_n; t_n, u_n) \]

we find that

\[
\| \tilde{z}_n'(t_{n+1}) - u_{n+1}' \| = h_n^{-1} \| \phi(t_n, u_n) \| + O(h_n^{-1}) . \tag{4.16}
\]

Since \( a_n \) and \( b_n \) are polynomials on \([t_n, t_{n+1}]\) they can be bounded in magnitude, so (4.14),(4.15), and (4.16) give

\[
\| \tilde{u}(t) - \tilde{z}_n(t) \| = O(h_n^{-1}) \tag{4.17}
\]

for all \( t \in [t_n, t_{n+1}] \).

Also, \( a_n'(t) = \frac{1}{h_n} q_1(t) \) and \( b_n'(t) = \frac{1}{h_n} q_2(t) \) where \( q_1 \) and \( q_2 \) are quadratics with bounds \( |q_1(t)| \leq 1, \) and \( |q_2(t)| \leq 1 \) for all \( t \in [t_n, t_{n+1}] \). Therefore, (4.14),(4.15), and (4.16) also give

\[
\| \tilde{u}'(t) - \tilde{z}_n'(t) \| = 2 h_n^{-1} \| \phi(t_n, u_n) \| + O(h_n^{-1}) , \tag{4.18}
\]

for all \( t \in [t_n, t_{n+1}] \). Finally, on \([t_n, t_{n+1}]\),

\[
\| \tilde{u}'(t) - k(t, \tilde{u}(t)) \| \leq \| \tilde{u}'(t) - \tilde{z}_n'(t) \|
+ \| \tilde{z}_n'(t) - z_n'(t) \|
+ \| z_n'(t) - k(t, z_n(t)) \|
+ L \| z_n(t) - z_n(t) \| + L \| z_n(t) - \tilde{u}(t) \| \tag{4.19}
\]

The errors in the Hermite interpolation of \( z_n \) are (see, for example, Stoer and Bulirsch (1980, p. 51))

\[
\| \tilde{z}_n'(t) - z_n'(t) \| = O(h_n^2) \tag{4.20}
\]

and
If we choose the order of the Runge-Kutta formula in (4.11) to be \( r=2 \), then (4.17),(4.18),(4.20),(4.21), and (4.19) along with (4.13) give
\[
\| \tilde{u}'(t) - k(t, \tilde{u}(t)) \| = 2h_n^2 \| \phi(t_n, u_n) \| + O(h_n^3), \tag{4.22}
\]
for all \( t \in [t_n, t_{n+1}] \).

Given a prescribed residual tolerance \( \epsilon > 0 \), we show how to use (4.22) to choose the step sizes \( h_n, n=1, 2, \ldots, N \) (given an initial step size \( h_0 \)) so as to satisfy
\[
\| \tilde{u}'(t) - k(t, \tilde{u}(t)) \| < \epsilon
\]
for all \( t \in [a, b] \). Let \( r_n(t) = \tilde{u}'(t) - k(t, \tilde{u}(t)) \) for \( t \in [t_n, t_{n+1}] \). Suppose for the step size \( h_n \) we have \( \| r_n \|_\infty < \epsilon \). By (4.22) we then have
\[
2h_n^2 \| \phi(t_n, u_n) \| < \epsilon + O(h_n^3). \tag{4.23}
\]
Recalling (4.11), we obtained \( u_{n+1} \) from the second order formula
\[
u_{n+1} = u_n + h_n \psi(t_n, u_n). \tag{4.24}
\]
Suppose we calculate an alternate approximation \( \hat{u}_{n+1} = u_n + h_n \hat{\psi}(t_n, u_n) \),
using a third order formula \( \hat{\psi} \). The well known strategy of using the difference \( \hat{u}_{n+1} - u_{n+1} \) to estimate the local errors can also be used to estimate the maximum size of the residual on each interval \([t_n, t_{n+1}]\).

It is well known (Gear, 1971, p.62) that the local errors may be expressed as
\[
\| u_{n+1} - u(t_{n+1}; t_n, u_n) \| = h_n^2 \| \phi(t_n, u_n) \| + O(h_n^4) \tag{4.25}
\]
\[
\| \hat{u}_{n+1} - u(t_{n+1}; t_n, u_n) \| = h_n^4 \| \phi(t_n, u_n) \| + O(h_n^5). \tag{4.26}
\]
Subtracting (4.24) from (4.23) we get
\[
\| u_{n+1} - \hat{u}_{n+1} \| = h_n^3 \| \phi(t_n, u_n) \| + O(h_n^4). \tag{4.27}
\]
Hence
\[
\| r_n \|_\infty = \frac{2}{h_n} \| u_{n+1} - \hat{u}_{n+1} \| + O(h_n^3). \tag{4.28}
\]
Similarly, on the \( n+1 \)st interval \([t_{n+1}, t_{n+2}]\), we have
\[
\| r_{n+1} \|_\infty = \frac{2}{h_{n+1}} \| u_{n+2} - \hat{u}_{n+2} \| + O(h_{n+1}^3). \tag{4.29}
\]
Now, to first order
\[
\phi(t_{n+1}, u_{n+1}) \approx \phi(t_n, u_n) + \partial_t \phi(t_n, u_n) h_n + \partial_u \phi(t_n, u_n) (u_{n+1} - u_n). \tag{4.30}
\]
However, \( u_{n+1} - u_n = h_n \psi(t_n, u_n) \) and \( \psi \) is bounded in \( h \) since \( \psi \to k(t_n, u_n) \) as \( h_n \to 0 \). Therefore,
\[
\| \phi(t_{n+1}, u_{n+1}) \| = \| \phi(t_n, u_n) \| + O(h_n). \tag{4.31}
\]
The condition \( \| r_{n+1} \|_\infty < \epsilon \) requires
\[
2h_{n+1}^2 \| \phi(t_{n+1}, u_{n+1}) \| < \epsilon + O(h_{n+1}^3), \tag{4.32}
\]
or, from (4.26)
\[
2h_{n+1}^2 \| \phi(t_n, u_n) \| < \epsilon + O(h_{n+1}^3) + O(h_n h_{n+1}^2), \tag{4.33}
\]
or, from (4.22) and (4.25)
\[
\frac{2h_{n+1}^2}{h_n^3} \| u_{n+1} - \bar{u}_{n+1} \| < \epsilon + O(h_{n+1}^3) + O(h_n h_{n+1}^2).
\] (4.27)

From (4.27) we can estimate the new stepsize by
\[
h_{n+1} = h_n \sqrt{\frac{\epsilon h_n}{2 \| u_{n+1} - \bar{u}_{n+1} \|}}.
\] (4.28)

It is interesting to note that the step size selection for controlling the local error is usually derived as
\[
h_{n+1} = h_n \sqrt{\frac{\epsilon}{\| u_{n+1} - \bar{u}_{n+1} \|}}.
\]

Many authors (see, for example, Stoer and Bulirsch (1980, p.42)), however, choose (4.28), up to a factor, based on numerical evidence that suggests the step selection should be more conservative.

By choosing the step size by (4.28), we hope to maintain \( \| r_n \|_\infty < \epsilon \) for \( n = 0, 1, \ldots, N \), or
\[
\| r(t) \| < \epsilon
\] (4.29)
for all \( t \in [a, b] \).

If (4.29) holds, then by (4.8), (4.7), (4.8), and (4.9) we have
\[
\| g_k' - F x_k - g \|_\infty \leq \left( \sum_{i=1}^{k} |\gamma_i| + 1 \right) \epsilon,
\]
and (4.5) will be satisfied if we choose \( \epsilon \) by
\[
\epsilon^2 = \frac{\eta_k \| P(y_k) \|_{L^2}}{(b-a)(\sum_{i=1}^{k} |\gamma_i| + 1)^2}.
\] (4.30)

The one difficulty with (4.30) is that the \( \gamma_i, i = 1, \ldots, k \) are not determined until after the initial value problems are solved. One suggestion is to choose these in (4.30) to be those calculated at the previous iteration.
CHAPTER 5

Implementation and Numerical Results

In this chapter, we briefly discuss the numerical implementation of the Model Trust Region modification of an inexact Newton method for the solution of nonlinear two-point boundary value problems. We will present some of the main implementation details and display preliminary numerical results for a set of test problems.

The computer code was written in FORTRAN, running on the VAX 11/780 under the VMS operating system. The LINPACK library was used for all matrix computations except for those needed in the linear boundary value solver. The IMSL routine DCADRE, an adaptive quadrature routine, was used for evaluating norms and inner products. The optimization package MINPACK was used to help generate some of the test problems.

To solve the linear boundary value problem, we used a modified version of the code SUPORT. This linear solver, developed by Scott and Watts (see Scott and Watts(1977a)), is a robust implementation of the reduced superposition algorithm. The code monitors the columns of the matrix valued function that solves the initial value problem given in (4.7). It is essential that the columns remain numerically linearly independent as the integration proceeds. This is required so that the linear algebraic system in (4.6) is well conditioned. When the code detects that the columns are becoming too dependent, it halts the integration. It then orthonormalizes the columns of the matrix and continues the integration of the initial value problem with these orthonormal columns providing new initial conditions. When the integration is complete, the code then pieces together the orthonormalized solutions. (See Scott and Watts (1977a) for further details). This code was chosen because it is capable of handling notoriously difficult linear boundary value problems.

The integrator used in the modified SUPORT code is a modification of the code RKF45, developed by Shampine and Watts (1979). This code utilizes a pair of fourth and fifth order Runge-Kutta-Fehlberg formulas for variable step selection. We replaced these formulas with third and fourth order Runge-Kutta-Fehlberg formulas. These differ from the second and third order pair suggested in Chapter 4. We found that integrating with a second order formula produced an unacceptable number of output points. We also found that the residual tolerance given in (4.30) was too conservative and that we could still satisfy (4.5) with these higher order formulas. We also incorporated the step selection strategy described Chapter 4.

To calculate the gradient given by (3.14) and (3.15), we solve the following initial value problems:

\[ z'(t) = f(t, y_k(t)) \cdot P(y_k)(t), \quad t \in [a, b], \]  

(5.1)
$v'(t) = P(y_b)(t) + z(t)$

$z(a) = 0, \quad v(a) = 0.$

The gradient is then given by

$$\nabla \phi(y_b)(t) = v(t) + \begin{bmatrix} A \\ l(t)B \end{bmatrix}^T e - l(t)z(b)$$

where $l(t) = 1 + (t-a)$ and $c$ solves the linear equation

$$Mc = \begin{bmatrix} A \\ l(b)B \end{bmatrix} z(b) - \begin{bmatrix} 0 \\ B \end{bmatrix} v(b).$$

Here, $M$ is the symmetric positive definite matrix given in (3.16). The integration of (5.1) used the original RKF45 code with stringent error tolerances.

This was to try to get an accurate approximation to the gradient.

The code was tested on the following set of problems. In each case, we include the analytic solution if it is known. For those problems without a known analytic solution, we first solved an initial value problem very accurately and then reposed it as a boundary value problem.

This first problem is mildly nonlinear and is widely used as a test problem.

**Problem Exp**

$$y'' = e^y$$

$$y(0) = 0 = y(1).$$

The solution is

$$y''(t) = -\ln 2 + 2 \ln[c \sec(c(t-(1/2))/2)],$$

where $c$ solves

$$c \sec\left(\frac{c}{4}\right) = \sqrt{2}.$$}

The following problem is extremely difficult for large values of the parameter $\lambda$.

The solution has a singularity just outside the interval of integration. This singularity gets closer to the endpoint $b$ as the value of the parameter increases.

We've chosen the value of the parameter large enough to cause most boundary value solvers difficulty.

**Problem Troesch**

$$y'' = \lambda \sinh(\lambda y), \quad \lambda = 10$$

$$y(0) = 0, \quad y(1) = 1.$$

$$y'(t) = \frac{2}{\lambda} \text{arc} \text{sin} h \left[ \frac{2+c}{2} \text{sc} \left( \lambda t, \frac{2-c}{4} \right) \right]$$

where

$$c = y'(0)^2 - 2$$

and $\text{sc}(x,k^2)$ is the Jacobian elliptic function with modulus $k$.

The next two problems are polynomials in the dependent variable.
Problem Poly 1

\[ y'' = (y')^3 - 2yy' + y^4 \]

\[ y(0) = 1, \quad y(1) = y_{int}(1). \]

Problem Poly 2

\[ y''' = -\frac{2}{9} y^2 \]

\[ y(1) = 1, \quad y(1.5) = (1.5)^{2/3} \]

\[ y'(1) = t^{2/3}. \]

The next problem was derived from a common test problem for nonlinear equation solvers in R^n.

Problem Powell

\[ y_1' = 10y_1y_2 - 1 \]

\[ y_2 = e^{-y_1} + e^{-y_2} - 1.0001 \]

\[ y_1(0) = .1, \quad y_2(.1) = y_{int}(.1). \]

The following two problems are due to Miele and Iyer (1971).

Problem Miele 1

\[ x''' = \frac{1}{6} x'' y y' \]

\[ y' = \frac{1}{2} x'(y')^3 \]

\[ x(0) = 1, \quad x(1) = 16 \]

\[ y(0) = 1, \quad y(1) = \frac{1}{2}, \quad y'(0) = -1. \]

Problem Miele 2

\[ z' = 13y \]

\[ y' = 13x \]

\[ z' = -20.15xz + 1.3y^2 - 13u^2 + 2.6y + 13 \]

\[ u' = 13w \]

\[ w' = -20.15xw + 14.3yu + 2.6u - 2.6 \]

\[ x(0) = 0, \quad y(0) = 0, \quad u(0) = 0 \]

\[ y(1) = 0, \quad u(1) = 1. \]

We now present the results of a comparison of quasi-linearization to the dogleg implementation of the Model Trust Region algorithm. Both methods specified, in advance, the error tolerance the linear boundary value solver was to achieve. Neither made use of the Inexact Newton framework. The error
tolerance given was $10^{-4}$. All problems, except Troesch, were given, as their starting point, a linear function satisfying the boundary conditions. The Troesch problem was given a piecewise linear function, on a nonuniform mesh, satisfying the boundary conditions. Neither method was able to succeed on this problem. Before discussing the reasons for the failures, we present a summary of the results for the two methods on the test problems.

The approximate solutions of the linear boundary value problems were stored as Hermite cubics interpolating the function values returned by the integrator. The derivatives on the mesh points were given by evaluating the right hand side of the differential equation at the function values. In the table below, we indicate whether the method converged to a solution of the problem and the number of iterations required. Also listed is the maximum relative error observed when evaluating the Hermite cubic solution on a uniform mesh.

<table>
<thead>
<tr>
<th>Problem</th>
<th>no. of iterations</th>
<th>max. rel. err.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Newton</td>
<td>Dogleg</td>
</tr>
<tr>
<td>Exp</td>
<td>n.c.</td>
<td>2</td>
</tr>
<tr>
<td>Troesch</td>
<td>n.c.</td>
<td>2</td>
</tr>
<tr>
<td>Poly1</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Poly2</td>
<td>n.c.</td>
<td>7</td>
</tr>
<tr>
<td>Powell</td>
<td>n.c.</td>
<td>7</td>
</tr>
<tr>
<td>Miele1</td>
<td>n.c.</td>
<td>73</td>
</tr>
<tr>
<td>Miele2</td>
<td>n.c.</td>
<td>8</td>
</tr>
</tbody>
</table>

Note: $(n) = 10^n$ and n.c. $=$ no convergence

We notice that quasilinearization failed to converge for most of the problems. The reason for the failures was that the linear solver could not produce a Newton step within the maximum number of output points allowed. In the dogleg method, when the linear solver failed to find a Newton step the Cauchy step was taken. Eventually a region was entered where the linear problem could be solved. Eventually more of the Newton step could be taken and, in most cases, the full Newton step was taken in the last couple of iterations. The dogleg failed on the Troesch problem by exceeding the maximum number of iterations allowed. All steps taken by the dogleg were Cauchy steps; giving sure but slow reduction on the norm of the residual. We do note that at the beginning of the iteration the decrease in the norm of the residual was very rapid.

The next test is a comparison of the dogleg implementation of the Model Trust Region algorithm and a damped Newton algorithm using the safeguarded cubic linesearch described in Chapter 3. Both methods use the Inexact Newton algorithm described in Chapter 4 for the approximate solution of the Newton step. In this test, the maximum relative error is given only for those problems with an analytic solution. For the remainder of the problems we give the norm of the residual at the calculated solution.
We notice that on the easiest problems, *Exp* and *Poly1*, the two globalization strategies performed similarly. However, the dogleg algorithm performed poorly on the *Miele2* problem compared to both the linesearch algorithm and the dogleg algorithm without the Inexact Newton modification. We do notice that the dogleg algorithm with the Inexact Newton modification was the only algorithm able to solve the difficult *Troesch* problem. The failure of the linesearch algorithm, on the *Troesch* problem, was again due to the trouble the linear integrator was having in solving the Newton equation.

It is difficult to make any generalizations concerning the performances of the various algorithms based on such limited numerical results. It does seem clear that the dogleg algorithm, especially when coupled with the Model Trust Region modification, is the most robust of the algorithms. It does give the Cauchy step to fall back on when the linear solver develops problems. In the first test we did see that this was sufficient to eventually allow the iterates to enter a region where the full Newton step could be taken.

### CHAPTER 6

**Concluding Remarks**

In this work, we have considered an iterative process for obtaining an approximate solution to an infinite dimensional nonlinear problem. Virtually all such processes involve the operations of linearization and discretization. In Chapter 2, we analyzed Newton’s method, the most common form of linearization. The linear model, given by the Newton equation, is only a locally valid model. Hence, trying to predict a zero of the nonlinear operator from the zero of the linear operator is only locally valid. In order to globalize the method, we adopt the model trust region philosophy, which has been so successful in finite dimensions. Algorithms based on the model trust region philosophy explicitly recognize the limited validity of the linear model. The nonlinear operator these algorithms use to predict a zero of the original nonlinear operator varies with the region in which the Newton model is valid. However, since the nonlinear operator equations that give rise to these models are themselves expensive to solve, an approximation to the model then seems warranted. This is the motivation behind the dogleg implementation of the model trust region algorithm. In Chapter 3, we show how to implement this technique for the solution of the nonlinear two-point boundary value problem.
This yields a globally convergent modification of the basic quasilinearization method. This globalization method was compared to one which has been used in the past, the damped Newton algorithm. Numerically, we found evidence that the dogleg algorithm was better able to handle difficult problems. This algorithm seems to be less efficient in solving the easier problems. This will probably change with a more refined implementation. Such refinements would include the step doubling technique discussed in Dennis and Schnabel (1983,p.145). This would allow the acceptance of a step larger than the trust region radius if the model criteria indicated such a step was acceptable. This would save the expense of calculating some of the gradients, which are relatively expensive. Another refinement discussed in Dennis and Schnabel (1983,p.144) is the reduction of the trust region radius by using a one dimensional quadratic model of the operator in the direction of the latest step. The minimizer of this quadratic would be used to predict the new trust region size. We note the Algorithm 3.1 is general enough to accommodate such refinements.

The other major issue considered in this work is the convergence analysis of discrete approximations to the infinite dimensional iterates. The Inexact Newton framework is a natural setting in which to regard infinite dimensional nonlinear problems, since the iterates can only be solved for inexactly in the first place. We extend the framework to infinite dimensions and develop an inexact method for solving the linear boundary value problems at each iteration. This method allows control over both the discretization of the space and the discretization of the integration. With this control, we can satisfy the accuracy requirements given by the Inexact Newton theory to generate a sequence of discrete approximations to the solutions of the linear problems. This theory then gives the rapid local convergence of the discrete approximations to the infinite dimensional solution of the nonlinear boundary value problem.

Preliminary numerical results indicate that this can be a robust influence in solving difficult problems. In the case of the dogleg algorithm, using an inexact method for the solution of the Newton step, we see that the linear problem is usually easier to solve than if a strict error tolerance is imposed. The dogleg step can then contain a portion of the inexact Newton step. This seems to be more beneficial than having to take the default Cauchy step when the linear problem could not be solved to within the strict error tolerance.

Finally, we note that the two main ideas developed in this thesis have a proven record in solving nonlinear equations in finite dimensions. It would be surprising if they could not contribute to solving other nonlinear problems. This work has been an attempt to see how these ideas might be introduced for nonlinear two-point boundary value problems.
References


