Historical Development of the BFGS Secant Method 
and Its Characterization Properties 
by 
Joanna Maria Papakonstantinou
A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree
Doctorate of Philosophy

Approved, Thesis Committee:

Richard A. Tapia, Chairman
University Professor, Maxfield-Oshman Chair in
Engineering, Rice University

Keith D. Cooper
Professor of Electrical and Computer Engineering
and Computer Science, Rice University

Illya Hicks
Associate Professor of Computational and
Applied Mathematics, Rice University

Wotao Yin
Assistant Professor of Computational and
Applied Mathematics, Rice University

Yin Zhang
Professor of Computational and Applied
Mathematics, Rice University

Houston, Texas
April, 2009
Abstract

Historical Development of the BFGS Secant Method and Its Characterization Properties

by

Joanna Maria Papakonstantinou

The BFGS secant method is the preferred secant method for finite-dimensional unconstrained optimization. The first part of this research consists of recounting the historical development of secant methods in general and the BFGS secant method in particular. Many people believe that the secant method arose from Newton’s method using finite difference approximations to the derivative. We compile historical evidence revealing that a special case of the secant method predated Newton’s method by more than 3000 years. We trace the evolution of secant methods from 18th-century B.C. Babylonian clay tablets and the Egyptian Rhind Papyrus. Modifications to Newton’s method yielding secant methods are discussed and methods we believe influenced and led to the construction of the BFGS secant method are explored.

In the second part of our research, we examine the construction of several rank-two secant update classes that had not received much recognition in the literature. Our study of the underlying mathematical principles and characterizations inherent in the updates classes led to theorems and their proofs concerning secant updates. One class of symmetric rank-two updates that we investigate is the Dennis class. We demonstrate how it can be
derived from the general rank-one update formula in a purely algebraic manner not utilizing Powell’s method of iterated projections as Dennis did it. The literature abounds with update classes; we show how some are related and show containment when possible. We derive the general formula that could be used to represent all symmetric rank-two secant updates. From this, particular parameter choices yielding well-known updates and update classes are presented. We include two derivations of the Davidon class and prove that it is a maximal class. We detail known characterization properties of the BFGS secant method and describe new characterizations of several secant update classes known to contain the BFGS update. Included is a formal proof of the conjecture made by Schnabel in his 1977 Ph.D. thesis that the BFGS update is in some asymptotic sense the average of the DFP update and the Greenstadt update.
Contents

Abstract ii

List of Figures x

List of Tables xi

1 Introduction 1

1.1 Basic Algorithms in 1-D 3

1.2 Basic Algorithms in n-D 6

1.3 Convergence Behavior 8

2 Historical Development of the 1-D Secant Method 10

2.1 Original Formulation of the Secant Method in 1-D 11

2.1.1 The Rules of False Position in Early Texts 11

2.1.2 The Rule of Single False Position 12

2.1.2.1 Example problem solved using the Rule of Single False Position 14
2.1.3 The Rule of Double False Position ........................................ 16
  2.1.3.1 Example problem: a linear equation ................................ 18
  2.1.3.2 Example problem: a system of linear equations ............... 19
  2.1.3.3 Example problem: a system of equations involving higher
           orders ........................................................................... 20
  2.1.4 Rule of Double False Position in Other Texts ....................... 21
  2.1.5 Introduction of the term Regula Falsi ................................. 23
  2.1.6 Modification of the Rule of Double False Position applied
       non-iteratively to Quadratic Equations in One Unknown ............ 24
  2.1.7 Use of the Rule of Double False Position as an Iterative Process-
       The Secant Method ........................................................... 26
  2.2 Newton’s Geometric Approach to the Secant Method ................. 26
  2.3 The Naming and First Convergence Rate Proof of the Secant Method in 1-D 27
  2.4 Confusion and Inconsistencies in Naming Methods .................... 28
    2.4.1 Initial Misuse and Confusion of the Name Regula Falsi ........... 28
    2.4.2 Evolution of the Names: Rule of Double False Position and
       Regula Falsi ....................................................................... 29
    2.4.3 Continued Confusion of Names and Understanding ................. 29

3 Extension of the 1-D Secant Method to n-D .................................. 36
  3.1 General Position ..................................................................... 37
  3.2 Linear Interpolation Methods .................................................. 38
3.2.1 The Ortega and Rheinboldt General Framework . . . . . . . . . . . . 39
3.2.2 Wolfe’s Method . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41
3.3 Discretized Newton Methods . . . . . . . . . . . . . . . . . . . . . . . . 42
3.3.1 \((n + 1)\)-point Secant Method . . . . . . . . . . . . . . . . . . . . 43
3.3.1.1 Sequential \((n + 1)\)-point Secant Method . . . . . . . . . . . 45
3.3.2 Sequential 2-point Secant Method . . . . . . . . . . . . . . . . . . . 45
3.4 Properties of \((n + 1)\)- and 2-point Secant Methods . . . . . . . . . . 46
3.4.1 Convergence Properties of Generalized Secant Methods . . . . . . 47
3.5 Remarks . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 49

4 Development of Secant Methods . . . . . . . . . . . . . . . . . . . . . . . . 51
4.1 Nonlinear Equations . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 52
4.1.1 Inverse versus Direct Updating . . . . . . . . . . . . . . . . . . . . . 53
4.1.2 Good Broyden . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 54
4.1.3 Bad Broyden . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 55
4.1.4 Symmetric Rank One (SR1) . . . . . . . . . . . . . . . . . . . . . . . 56
4.1.5 Powell Symmetric Broyden (PSB) . . . . . . . . . . . . . . . . . . . 56
4.1.6 Least Change Problem . . . . . . . . . . . . . . . . . . . . . . . . . . 58
4.2 Unconstrained Optimization . . . . . . . . . . . . . . . . . . . . . . . . . 60
4.2.1 Davidon, Fletcher and Powell (DFP) . . . . . . . . . . . . . . . . . . 61
4.2.1.1 Conditioning Problem . . . . . . . . . . . . . . . . . . . . . . . . 64
4.2.1.2 Restarting . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 64
4.2.2  Broyden, Fletcher, Goldfarb, Shanno (BFGS) .... 65
  4.2.2.1  Fletcher ........................................... 65
  4.2.2.2  Broyden ........................................... 66
  4.2.2.3  Shanno ............................................. 69
  4.2.2.4  Goldfarb .......................................... 70
4.2.3  Weighted Least Change Problem .................. 72
4.2.4  Limited Memory BFGS (L-BFGS) .................. 74
4.2.5  Broyden Class of Secant Updates ................. 77
4.3  Equality Constrained Optimization .................. 80
  4.3.1  Successive Quadratic Programming (SQP) ........ 81
  4.3.2  Powell’s Damped BFGS Algorithm (PDA) ........ 83
  4.3.3  SQP Augmented Lagrangian BFGS Secant Method . 84
  4.3.4  Tapia’s BFGS Structured Augmented Lagrangian Secant Algorithm (SALSA) ............... 85

5  Convergence Theory 87
  5.1  Conjugacy ................................................. 87
  5.2  Convexity .................................................. 88
  5.3  Finite Termination ....................................... 89
  5.4  Bounded Deterioration ................................... 90
  5.5  Superlinear Convergence .................................. 91
    5.5.1  Broyden, Dennis and Moré Characterizations .... 92
5.5.2 Broyden, Dennis, and Moré Results .............................. 93
5.5.3 Dennis and Moré Characterizations ............................... 94
5.5.4 Dennis and Moré Results ........................................... 96
   5.5.4.1 Byrd, Tapia, Zhang Result ..................................... 97
5.5.5 Powell’s Characterization .......................................... 97
5.5.6 Powell’s Results .................................................... 98
5.5.7 Powell’s Observation ............................................... 101
5.5.8 Byrd, Nocedal and Yuan ........................................... 102
5.5.9 Byrd and Nocedal .................................................. 104
5.5.10 Zhang and Tewarson .............................................. 104

6 Known Characterizations of the BFGS Secant Method 106
   6.1 Nazareth’s Step Length Result ...................................... 106
   6.2 Finite Termination .................................................. 108
   6.3 Updates to Cholesky Factors ....................................... 108
      6.3.1 Least Change Updates to Cholesky Factors ..................... 110
      6.3.2 Weighted Least Change Updates to Cholesky Factors ............ 112
         6.3.2.1 Least Change Cholesky (LCC) Updates ....................... 112
   6.4 Least Change in Byrd-Nocedal Measure ............................ 114
   6.5 Yabe, Martinez, Tapia Least Change in Weighted Byrd-Nocedal Measure . 116
      6.5.1 Sizing .......................................................... 116
      6.5.2 Selective Sizing ................................................ 119
6.5.3 Sizing and Shifting ......................................................... 121
6.5.4 Weighted $\psi$-Optimal Values for the Sized-Broyden Class .... 122
6.5.4.1 Solutions ............................................................. 124

7 Secant Update Classes and Some of their Properties ................. 125
7.1 Dennis Class ................................................................ 126
7.1.1 Derivation of the Dennis Class using Iterated Projections .... 128
7.1.2 Algebraic Derivation of the Dennis Class ....................... 131
7.1.3 Extension of the Dennis Class ...................................... 132
7.2 $u$-class ..................................................................... 134
7.3 $d$-class ..................................................................... 135
7.4 Schnabel’s Observation .................................................. 136
7.5 Schnabel’s Conjecture .................................................... 137
7.6 Tapia’s Discovery ......................................................... 138
7.7 Huang Class ............................................................... 140
7.8 General Form for Symmetric Rank-2 Secant Updates ............ 142
7.8.1 Extended Dennis-Davidon Class ................................. 145

Bibliography .................................................................... 147
# List of Figures

1.1 Newton’s Method .................................................. 3
1.2 The Secant Method .................................................. 5
2.1 Papyrus Problem 26 in hieratic notation ...................... 14
List of Tables

2.1 Description of Problem 26 .................................................. 15
2.2 Evolution of the naming of the Rule of Double False Position .... 30
2.3 Different names for Regula Falsi, Modified Regula Falsi, Secant Method . 31
Chapter 1

Introduction

The first part of our research consists of recounting the historical development of secant methods in general and the BFGS (Broyden, Fletcher, Goldfarb, Shanno) secant method in particular. The motivation for this work is based on the fact the BFGS secant method is the preferred secant method for unconstrained optimization. We trace the evolution of secant methods from a method (most commonly referred to as the Rule of Double False Position but also referred to as Regula Falsi) that can be found in 18th-century B.C. Babylonian clay tablets and the Egyptian Rhind Papyrus. This method can be viewed as the secant method in one dimension (1-D) applied to a linear equation; hence, we believe that this should be considered the origin of the secant method.

Throughout the years, there has been widespread confusion concerning the origins and the terminology used to refer to the secant method and the Regula Falsi method. It is interesting that confusion still exists today. To remove the existing confusion, we determine the
origins of these methods and clarify the terminology. Modifications to Newton's method that yield secant methods are discussed and methods that we believe influenced and led to the construction of the BFGS secant method are explored. We detail known uniqueness properties and characterizations of the BFGS secant method and study various classes of secant updates that are known to contain the BFGS update.

In the second part of our research, we examine the construction of several rank-two secant update classes that have not received much recognition in the literature. Our study of the underlying mathematical principles and characterizations inherent in these update classes lead us to several theorems concerning classes of secant updates.

This thesis is organized in the following manner. In Chapter 2, we explain that the origin of the secant method in 1-D dates back to 18th century B.C. We present the obvious extension of the 1-D secant method to higher dimensions with an explanation of why it fails in Chapter 3. In Chapter 4, we describe various secant methods with a focus on the BFGS method. Convergence theory and convergence results are outlined in Chapter 5. Several interesting characterizations and uniqueness properties of the BFGS secant method are detailed in Chapter 6. Finally, in Chapter 7, we examine the construction of several rank-two secant update classes that have not received much recognition in the literature and present new results.
1.1 Basic Algorithms in 1-D

Three numerical algorithms which have played important roles in the contemporary numerical methods literature for solving a nonlinear equation are Newton’s method, the secant method, and the Regula Falsi method, as we call them today. In this section, we present these algorithms in one dimension, since this is where they were born and where confusion originated.

There are many iterative methods for approximating a solution, \( x^* \), of \( f(x) = 0 \) for \( f : \mathbb{R} \to \mathbb{R} \). Newton’s method in 1-D uses a succession of zeros of tangent lines to better approximate a zero of a function. In Figure 1.1, \( f(x) \) represents the nonlinear function whose zero we are trying to find.

![Figure 1.1: Newton’s method uses the tangent line passing through the points \((x_0, f(x_0))\) and \((x_1, f(x_1))\).](image)

The iteration

\[
x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}
\]  

(1.1)
is Newton’s method (in 1-D), and \( x_k \) represents the \( k \)th approximation to the solution.

If the derivative in the Newton iteration (1.1) is replaced with the difference quotient

\[
\frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}},
\]

which can be viewed as an approximation to \( f'(x_k) \), the resulting iteration

\[
x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k)
\]
is the secant method (in 1-D) and can also be written as

\[
x_{k+1} = \frac{x_{k-1} f(x_k) - x_k f(x_{k-1})}{f(x_k) - f(x_{k-1})}.
\]

In contrast to Newton’s method which uses a succession of zeros of tangent lines, the secant method in 1-D uses a succession of zeros of secant lines to better approximate a zero of a function. In Figure 1.2, \( f(x) \) represents the nonlinear function whose zero we are trying to find.

If instead of always using the two most recently computed iterates in the secant method (1.3), one of the initial estimates is held fixed for all subsequent iterations while the other is updated, the resulting iteration

\[
x_{k+1} = \frac{\bar{x} f(x_k) - x_k f(\bar{x})}{f(x_k) - f(\bar{x})}
\]
is the Regula Falsi method where \( \bar{x} \) represents the initial estimate that remains fixed. We choose to call this method the Regula Falsi method to follow Booth [9], who in 1955 seems to be the first to describe this method and refer to it as Regula Falsi.\(^1\) However, as we

\(^1\)In §2.4.3, we explain that we found earlier references, for example by Whittaker and Robinson [125] in 1944, however, these descriptions were ambiguous so we chose Booth’s naming since his description was more explicit and descriptive.
Figure 1.2: The secant method uses the secant line passing through the points \((x_0, f(x_0))\) and \((x_1, f(x_1))\).

discuss in §2.4.3, this method, post Booth, has also been referred to by other names.

Another difference between the secant method and the Regula Falsi method, is that in the Regula Falsi method, the two initial estimates, \(x_0\) and \(x_1\), are chosen such that \(f(x_0)\) and \(f(x_1)\) are of opposite signs \((f(x_0)f(x_1) < 0)\), i.e., the initial estimates bracket a zero of the nonlinear function. However, it is important to mention that in the Regula Falsi method, a zero does not necessarily remain bracketed by successive iterates at each step and, in some instances, the method fails. Modifications of the Regula Falsi method have been made that ensure that a zero remains bracketed at each step.\(^2\) For example, in what we choose to call the Modified Regula Falsi method, the interval endpoints are changed, instead of holding one fixed, to ensure that at each step, the new interval contains a zero of \(f(x)\), i.e., the \(x\)-value corresponding to the function value that has opposite sign as the

\(^2\)Although we are not sure who was the first to present modifications of the Regula Falsi method, early examples can be found in Willers’ 1948 book [126] and in Householder’s 1953 book [70].
current function value is always retained, not just in the first step as in the Regula Falsi method. This Modified Regula Falsi method is the method that many current texts, and in particular popular internet sites, call the Regula Falsi method. We discuss the naming confusion and present examples of naming inconsistencies in §2.4.3.

1.2 Basic Algorithms in \( n \)-D

In \( \mathbb{R}^n \), Newton’s method is a tool that allows us to approximate the solution of a square nonlinear system of equations by solving a sequence of square linear systems. Let \( F : \mathbb{R}^n \rightarrow \mathbb{R}^n \) be a differentiable function and consider the square nonlinear system of equations

\[
F(x) = 0. \tag{1.4}
\]

The iteration

\[
x_{k+1} = x_k - F'(x_k)^{-1}F(x_k) \tag{1.5}
\]

is Newton’s method, where again \( x_k \) represents the \( k \)th approximation to the solution of (1.4). Recall that in the case when solving the system of nonlinear equations (1.4), the derivative \( F'(x_k) \) in the Newton iteration (1.5) represents the Jacobian of \( F \) at \( x \). Newton’s method is theoretically attractive, but it may be difficult to use in practice for various reasons including the need to calculate the derivative.

\[^{3}\text{To learn more about modifications of the Regula Falsi method that ensure that each new interval contains a zero, see Bronson [11].}\]
The general quasi-Newton iteration is given by
\[ x_{k+1} = x_k - B_k^{-1}F(x_k) \] (1.6)
where \( B_k \) is viewed as an approximation to \( F'(x_k) \) and we call \(-B_k^{-1}F(x_k)\) the correction. If we introduce a step-length parameter, \( \alpha_k \), into the general quasi-Newton iteration (1.6), we obtain the iteration
\[ x_{k+1} = x_k - \alpha_k B_k^{-1}F(x_k), \] (1.7)
where, in this case, \( p_k = -B_k^{-1}F(x_k) \) is called the search direction and \( \alpha_k p_k \) is called the correction. The use of \( \alpha_k \) is referred to as step-length control, and by an exact line search, we mean that the 1-D minimization problem
\[ \min_{\alpha_k > 0} \phi(x_k + \alpha_k p_k) \] (1.8)
has been solved exactly for \( \alpha_k \) where \( \phi \) is a merit function associated with the iteration method. Popular choices for nonlinear equations is \( \phi(x) = F(x)^T F(x) \) and in unconstrained optimization \( \phi = f(x) \).

We call an iterative method of the form (1.6) a secant method if it satisfies the secant equation:
\[ B_{k+1}(x_{k+1} - x_k) = F(x_{k+1}) - F(x_k), \] (1.9)
more commonly written in the literature as
\[ B_{k+1}s_k = y_k, \] (1.10)
where
\[ y_k = F(x_{k+1}) - F(x_k), \]
and the displacement is
\[ s_k = x_{k+1} - x_k. \]

If the method satisfies the secant equation, then it reduces to the secant method in 1-D.

Clearly, Newton’s method is not a secant method.

### 1.3 Convergence Behavior

There are many criteria by which we can evaluate an iterative procedure, for example, the length of time taken to calculate a solution or the amount of computer storage space used in the computation. The convergence behavior of an algorithm is an important measure of performance. In this section, we outline different convergence behaviors. In particular, we emphasize the notion of superlinear convergence as many of the methods we discuss in this thesis demonstrate this behavior. We only briefly discuss convergence behaviors here because they will be referred to in the upcoming chapters. Moreover, we devote Chapter 5 to detailing the important convergence theory and presenting some convergence results of the basic secant methods that we discuss throughout the thesis.

If a solution \( x^* \) exists, then we define \textit{local convergence} by saying that there exists a neighborhood of \( x^* \), such that for all initial vectors in the neighborhood, the iterates generated by the algorithm in question are well-defined and converge to \( x^* \). This means
that when our initial point \(x_0\) is sufficiently close to \(x^*\) then \(\lim_{k \to \infty} x_k = x^*\). _Global convergence_ asserts convergence to a solution from any starting point.

Let \(\{x_k\}\) be a sequence in \(\mathbb{R}^n\) that converges to \(x^*\). We say the convergence is _linear_ if there is a constant \(M \in (0, 1)\) and a choice of norm such that

\[
\frac{||x_{k+1} - x^*||}{||x_k - x^*||} \leq M
\]

for all \(k\) sufficiently large. This indicates that eventually the error, the distance from the solution measured in this norm, decreases at each iteration by at least a constant factor \(M\).

We say the convergence is _superlinear_ if

\[
\frac{||x_{k+1} - x^*||}{||x_k - x^*||} \leq r_k
\]

holds for some sequence \(\{r_k\}\) which converges to zero, i.e., \(\lim_{k \to \infty} \frac{||x_{k+1} - x^*||}{||x_k - x^*||} = 0\). Observe that in finite dimensions, superlinear convergence is norm independent, while linear convergence is not. We say the convergence has order, or rate, of at least \(p\) if

\[
\frac{||x_{k+1} - x^*||}{||x_k - x^*||^p} \leq M
\]

for all \(k\) sufficiently large where \(M\) is a positive constant, not necessarily less than 1 and \(p = 1\). In the case that \(p = 2\), we use the term _quadratic_ and in the case \(p = \frac{1 + \sqrt{5}}{2}\), we use the term _golden mean_. Observe that order is norm independent and order greater than one implies superlinear convergence. However, superlinear convergence does not necessarily imply a rate greater than one.
Chapter 2

Historical Development of the 1-D Secant Method

Many believe that the secant method arose from Newton’s method using a finite difference approximation to the derivative. However, historical evidence reveals that a special case of the secant method (most commonly referred to as the Rule of Double False Position) predated Newton’s method by more than 3000 years and can be found in 18th-century B.C. Babylonian clay tablets and the Egyptian Rhind Papyrus.

It is the purpose of this chapter to present the historical development of the secant method in 1-D. We describe the Rule of Double False Position and present examples found in multiple texts throughout the centuries which demonstrate how the Rule of Double False Position was used to solve a variety of problems, such as, how to obtain the exact solution to a $2 \times 2$ system of linear equations and how the Rule of Double False Position was extended...
(non-iteratively) to quadratics to obtain only an approximate solution. Throughout the years, there has been much confusion concerning the origins and the terminology used to refer to the secant method and the Regula Falsi method. It is interesting that confusion still exists today. In an effort to remove the existing confusion, we determine the origins of these methods and clarify the terminology.

2.1 Original Formulation of the Secant Method in 1-D

2.1.1 The Rules of False Position in Early Texts

In ancient times, mathematics was used as a tool to answer questions that arose in daily life. The earliest evidence of these tools (eventually referred to as the Rules of False Position) was found in Egyptian papyri and Babylonian clay tablets from the 18th century B.C. The Babylonian civilization flourished at about the same time as Pharaonic Egypt, but there seems to be little formal evidence that either nation influenced the other’s mathematics but they clearly must have. The Rules of False Position were always written rhetorically rather than in symbolic language, whose use in mathematics at the time was unknown, and were often presented within the context of a real-life situation. The Egyptians and Babylonians did not know algebra, indeed it did not exist at that time, nor did they have the notion of an equation or work with a general rule. There is no evidence of the use of a procedure, instead, each problem used specific numbers with the solution given as a set of instructions. Hence, problems that would be considered trivial today posed a high degree of difficulty in
ancient times.

The most important mathematical text from ancient Egypt is the Ahmes Papyrus, written by the scribe Ahmes in about 1659 B.C. and derived from material dated approximately 2000-1800 B.C. [27]. It was named the *Rhind Mathematical Papyrus* after the Scottish Egyptologist and antiquarian A. Henry Rhind who purchased it from a shop in Luxor while traveling in Egypt\(^4\) and brought it back to England in 1858 where it has resided since it was donated by Rhind’s estate to the British Museum in 1864 [27] [89]. The *Rhind Mathematical Papyrus*, written in hieratic notation (see Figure 2.1), is a two-sided document containing a collection of 87 real-life word problems with solutions on one side and tables to aid in computation on the other. The examples cover a wide range of mathematical ideas needed for a scribe to fulfill his duties. Thus we deduce that this treatise was used in the training of scribes.

2.1.2 The Rule of Single False Position

Problems 24-34 of the *Rhind Mathematical Papyrus* are examples of problems in one unknown of the first degree which can be represented using contemporary algebraic notation as finding a number \(x\) such that

\[
a_1x + \ldots + a_nx = c. \tag{2.1}
\]

\(^4\)Many of the wealthy British travelled to resort towns in Egypt during the winters. At the time of this visit to Egypt, Rhind was in poor health.
Of course, today we would simplify such problems to

\[ ax = c \tag{2.2} \]

where \( a = a_1 + a_2 + \ldots + a_n \). From a current mathematical point of view, this problem is simple to solve if we sum the \( a_i \)s in (2.1) and divide \( c \) by \( a \) to get the solution \( x = \frac{c}{a} \).

However, the people of the time did not perform algebraic simplifications. Moreover, since \( a \) in equation (2.2) was usually a fraction in real-life problems, they utilized a method that avoided the possibility of dividing by a fraction.\(^5\)

The first step of the method they used to solve for \( x \) in the rhetorical analog of the linear equation (2.1) was to choose a (probably false, yet, not so arbitrary) guess of the solution.\(^6\) Suppose (falsely) the solution is \( x = x_0 \), then we get \( c_0 \) where

\[ ax_0 = c_0 \neq c. \]

Now, multiply \( c \) by \( \frac{x_0}{c_0} \) to get the solution

\[ x = \frac{(c)(x_0)}{c_0} = (c) \left( \frac{1}{a} \right). \tag{2.3} \]

Hence, we obtain the solution without determining \( a \) or dividing by \( a \). This method was later called Simple False Position \([27]\), Process of Supposition, or most commonly, the Rule of Single False Position \([75]\) and avoids the need to explicitly determine \( a \) in the reduced equation. In essence, this method was a way of using an initial guess to obtain the

\(^5\)Egyptians could perform division but preferred to avoid dividing by a fraction as it was more laborious than dividing by integers or performing multiplication \([27]\).

\(^6\)The initial guess was not so arbitrary. Instead, the initial guess was chosen with the aim to operate with whole numbers since calculation with fractions could present difficulties \([27]\).
solution to a specific problem and was not a general rule for solving other problems of the same kind. It is unclear how they got the solution (2.3) from the initial guess $x_0$ and the corresponding $c_0$. In an attempt to provide a reasonable explanation, we show that from (2.2) and $ax_0 = c_0$, we can write the proportion

$$\frac{x}{x_0} = \frac{c}{c_0}$$

from which we can obtain the solution (2.3).

### 2.1.2.1 Example problem solved using the Rule of Single False Position

Figure 2.1 illustrates Problem 26 from the Egyptian *Rhind Mathematical Papyrus* written in hieratic notation. To demonstrate how difficult the notation and the technique of calculation was at that time, we transcribe Problem 26 from hieroglyphic notation into algebraic notation and describe the enumerated steps to solve the problem using the Rule of Single False Position (see Table 2.1).
Problem 26 of the *Rhind Mathematical Papyrus* using hieroglyphic notation (Chabert [27].)

<table>
<thead>
<tr>
<th>Transcription of hieroglyphics</th>
<th>Description using algebraic notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. A quantity, $\frac{1}{4}$ of it added to it, becomes 15</td>
<td>$x + \frac{1}{4}x = 15$</td>
</tr>
<tr>
<td>2. Operate on 4; make thou $\frac{1}{4}$ of them, namely 1; The total is 5.</td>
<td>Guess $x = 4$: $4 + 1 = 5$.</td>
</tr>
</tbody>
</table>
| 3. Operate on 5 for the finding of 15 \[1 \quad 5 \]
   \[2 \quad 10\] There becomes 3. | Divide: $\frac{15}{5} = 3$ |
| 4. Multiply: 3 times 4. \[1 \quad 3 \]
   \[2 \quad 6, \]
   \[4 \quad 12\] There becomes 12. | Multiply wrong answer ($x = 4$) by 3: $3 \times 4 = 12$. |
| 5. 1 12, \[\frac{1}{4} \quad 3\] Total 15 | $12 + \frac{1}{4}(12) = 15$ |
| 6. The quantity is 12. $\frac{1}{4}$ of it is 3; the total is 15. | Thus, $x = 12$ |

Table 2.1: Description of Problem 26 of the *Rhind Mathematical Papyrus*. 
2.1.3 The Rule of Double False Position

Since the Babylonians and the Egyptians already had the Rule of Single False Position to solve for $x$ in $ax = c$, they quite naturally tried to apply it to other real-life word problems which we would today represent using algebraic notation as finding a number $x$ such that

$$ax + b = c,$$  \hspace{1cm} (2.4)

where $b \neq 0$. Having no knowledge of algebra at the time, people did not know how to move terms from one side of an equation to the other [75]. Furthermore, they considered (2.2) and (2.4) to represent two different mathematical phenomena.

The first step of the method they used to solve for $x$ in the rhetorical analog of the linear equation (2.4) was to choose two guesses of the solution (probably false solutions). There were no restrictions on the initial guesses. Suppose the first guess of the solution is $x = x_0$, then we get the corresponding error $c_0$ where

$$ax_0 + b - c = c_0.$$  \hspace{1cm} (2.5)

Suppose the second guess of the solution is $x = x_1$, then we get the corresponding error $c_1$ where

$$ax_1 + b - c = c_1.$$  \hspace{1cm} (2.6)

They gave instructions\(^7\) of how to obtain the solution using the relation

$$x = \frac{x_0c_1 - x_1c_0}{c_1 - c_0}. \hspace{1cm} (2.7)$$

\(^7\)Instructions resemble: ‘Multiply the second error, $c_1$, by the first guess, $x_0$, and multiply the first error, $c_0$ by the second guess, $x_1$. Subtract whichever product is smaller from the larger and divide this result by the difference of the smaller error subtracted from the larger error.’
This method for solving for \( x \) is now most commonly referred to as the Rule of Double False Position [75].

We justify algebraically how they obtained the solution (2.7) from the initial guesses, \( x_0 \) and \( x_1 \), and the corresponding errors, \( c_0 \) and \( c_1 \). Subtract (2.6) from (2.5) and solve for \( a \) to obtain

\[
a = \frac{c_0 - c_1}{x_0 - x_1}.
\]

(2.8)

Add (2.6) and (2.5), use (2.8), and solve for \( c - b \) to obtain

\[
c - b = \frac{x_1c_0 - x_0c_1}{x_0 - x_1}.
\]

(2.9)

Then, from (2.4) we have

\[
x = \frac{c - b}{a},
\]

and following the substitution of \( a \) using (2.8), and the substitution of \( c - b \) using (2.9), we can obtain the solution (2.7). If for the two arbitrary initial guesses, \( x_0 \) and \( x_1 \), we write \( c_0 = f(x_0) \) and \( c_1 = f(x_1) \), where \( f(x) = ax + b - c \), then we obtain the solution

\[
x = \frac{x_0f(x_1) - x_1f(x_0)}{f(x_1) - f(x_0)}
\]

(2.10)

which can also be viewed as the first step of the secant method (1.3) applied to a nonlinear equation \( f(x) = 0 \). It is most interesting that the secant method applied to a linear equation

\[8\]We explain in §2.1.4 that this rule was first given an English name, rule of two (false) positions, by Chuquet in 1484.

\[9\]In 1978, Smeur [113] described how Frisius’ in 1540 solved \( 1 \frac{1}{2} x^2 = 200 \) using the “Rule of Double False.” Smeur explained that \( x \) is calculated from \( x = \frac{x_1^2f_2 - x_2^2f_1}{f_2 - f_1} \) which lends itself to the notation we use in (2.10).
converges in one step; hence, it is correct to say that the Rule of Double False Position is
the secant method applied to a linear equation.

2.1.3.1 Example problem: a linear equation

The Egyptian papyri and Babylonian clay tablets contained rhetorical examples that rep-resent linear equations solved using the Rule of Double False Position. While many of the
problems dealt with the sale and distribution of properties, inheritance, or for the purpose
of portion control and the prediction of production, some of the problems presented seem
inconsequential in comparison. For example, consider the following problem\textsuperscript{10}:

> When asking someone his age he answers: if my age were doubled and added
to this $\frac{1}{2}$, $\frac{1}{3}$, and $\frac{1}{4}$ part of my age and 6 years, then all together should equal
80. How old is he.

This example can be written using algebraic notation as the linear equation

$$2x + \frac{1}{2}x + \frac{1}{3}x + \frac{1}{4}x + 6 = 80.$$ 

To solve this problem using the Rule of Double False Position, first let $x_0 = 36$, then
$c_0 = 2x_0 + \frac{1}{2}x_0 + \frac{1}{3}x_0 + \frac{1}{4}x_0 + 6 - 80 = 37$.\textsuperscript{11} Next, let $x_1 = 16$, then $c_1 = 2x_1 + \frac{1}{2}x_1 + \frac{1}{3}x_1 + \frac{1}{4}x_1 + 6 - 80 = -24\frac{2}{3}$. Thus, the solution to this problem is $x = \frac{x_0c_1 - x_1c_0}{c_1 - c_0} = 24$.

\textsuperscript{10}Taken from p.67 of Smeur [113] but originally appeared on p.186 v. of J. van der Scheure’s 1611 edition
of his 1600 Arithmetica, oft Rekenconst, Haarlem.

\textsuperscript{11}At that time, they used the initial guess, $x_0$, to calculate $2x_0 + \frac{1}{2}x_0 + \frac{1}{3}x_0 + \frac{1}{4}x_0 + 6 = 117$. Then, they
evaluated $117 - 80$ to determine the corresponding error $c_0 = 37$. 
2.1.3.2 Example problem: a system of linear equations

Both the Egyptian papyri and Babylonian clay tablets contained rhetorical examples that represent systems of two linear equations in two unknowns that are solved using the Rule of Double False Position. The fact that similar problems were solved using the same method in different civilizations provides evidence that these problems reflect the problems of that time. Although the literature suggests that each civilization independently invented the same method to solve these problems, we are of the opinion that traders carried stories hence, transferring information (such as the explanation of this process used to answer the problems that arose) along trade lines between Egypt and Babylonia. Consider the following problem\textsuperscript{12}:

\begin{quote}
Let a 1 m\text{"} of good field cost 3 hundred; and 7 m\text{"} of poor field cost 5 hundred.

Now 1 q\text{"}ing field is bought together, the price is 1 myriad. Of the good and poor fields, how much is there each?
\end{quote}

This example can be written using algebraic notation as the system of linear equations

\begin{align*}
g + p &= 100 \text{ (m\text{"})} \\
300g + \frac{500}{7}p &= 10000 \text{ (coins)}
\end{align*}

where \( g \) and \( p \) represent the areas (in m\text{"}) of the good and poor fields respectively.\textsuperscript{13} To solve this problem using the Rule of Double False Position, first let \( g_0 = 20 \), then \( p_0 = 80 \), and \( c_0 = 300g_0 + \frac{500}{7}p_0 - 10000 = 1714\frac{2}{3} \). Next, let \( g_1 = 10 \), then \( p_1 = 90 \), and

\textsuperscript{12}Taken from p.37 of Lun [77].

\textsuperscript{13}q\text{"}ing and m\text{"} are units of area measure such that 1 q\text{"}ing = 100 m\text{"}. 
\[ c_1 = 300g_1 + \frac{500}{7}p_1 - 10000 = -571 \frac{3}{7}. \] Thus, the solution to this problem is \[ g = \frac{9001 - g_1c_0}{c_1 - c_0} = 12 \frac{1}{2} \] and \[ p = 87 \frac{1}{2}. \] They cleverly eliminated one variable, in turn, reducing the system to a linear equation of the form (2.4). As a result, they were able to use the Rule of Double False Position to obtain the exact solution to the system of linear equations.

2.1.3.3 Example problem: a system of equations involving higher orders

There is evidence that the Egyptians and Babylonians each extended the Rule of Double False Position to quadratics but neither used the rule in an iterative manner. They performed only one step and were aware that when the problem was more complicated (quadratic), the solution they obtained using the Rule of Double False Position was only approximate.

Consider the problem\(^{14}\):

**Divide 40 into two numbers so that the sum of both squares is 850.**

This example can be written using algebraic notation as the system of two equations (one of which is second order) with two unknowns

\[ \begin{align*}
x + y & = 40 \\
x^2 + y^2 & = 850.
\end{align*} \]

To solve this problem using the Rule of Double False Position, first let \( x_0 = 30 \), then \( y_0 = 10 \), and \( c_0 = x_0^2 + y_0^2 - 850 = 150 \). Next, let \( x_1 = 20 \), then \( y_1 = 20 \), and \( c_1 = x_1^2 + y_1^2 - 850 = -50 \). Thus, the answer to this problem obtained using the Rule of

\(^{14}\)Taken from p.71 of Smeur [113] but originally appeared on p.186 v. of J. van der Scheure’s 1611 edition of his 1600 *Arithmetica, oft Rekenconst, Haarlem*. This example can also be found on p.7 of Ma [77].
Double False Position is \( x = \frac{x_0c_1 - x_1c_0}{c_1 - c_0} = \frac{45}{2} = 22.5 \) and \( y = \frac{35}{2} \). However, we see that 
\[ x^2 + y^2 = \left(\frac{45}{2}\right)^2 + \left(\frac{35}{2}\right)^2 = 813\frac{1}{2} \neq 850, \]
thus the solution is only approximate. Therefore, the application of the Rule of Double False Position to a quadratic can be viewed as taking one step of the secant method on the given quadratic. To justify this algebraically, first simplify the above system to the quadratic equation (with one unknown)
\[ f(x) = x^2 - 40x + 375. \]
Let \( x_0 = 30 \), so \( f(x_0) = 75 \). Now, let \( x_1 = 20 \), so \( f(x_1) = -25 \). After performing one step of the secant method, the approximate solution obtained is 
\[ x = \frac{x_0f(x_1) - x_1f(x_0)}{f(x_1) - f(x_0)} = 22.5, \]
which is the same approximate solution obtained from using the Rule of Double False Position.

### 2.1.4 Rule of Double False Position in Other Texts

The use of the Rule of Double False Position appeared in the texts of many civilizations in the centuries following those found in Egypt and Babylonia. For example, the earliest surviving Chinese mathematics text, Jiǔ zhāng suànsú (Computational Prescriptions in Nine Chapters) [27], a.k.a. The Nine Chapters on the Mathematical Art [75], dates back to the Hán Dynasty around 200 B.C. and represents the collective efforts of many scholars over several centuries. It contains 246 problems in nine chapters with each chapter containing practical problems connected with everyday life, their solutions, and brief descriptions of the methods used to solve them.\(^{15}\) In Chapter 7 (the title, in English, translates to “Excess

\(^{15}\)The purpose of The Nine Chapters on the Mathematical Art was similar to that of the Egyptian Rhind Mathematical Papyrus – to serve as a practical handbook with problems that the ruling officials of the state
and Deficit”), twenty problems were solved using yìng bèi zuǐ shù which literally means ‘too much and not enough’ and can be recognized as the Rule of Double False Position [27]. This is the first evidence of the Rule of Double False Position being considered a general rule to be used on particular problems and given a name.

In the 9th century, Arab mathematician Abu Jafar Mohammad ibn-Mūsa al-Khwārizmī wrote two influential books which were translated into Latin in the 12th century and circulated throughout Europe [6]. Also in the 9th century, Abū Kāmil wrote Kitāb fil-jabr w’al muqābalah, ‘Book of Algebra’ (a commentary on, and elaboration of, Al-Khwārizmī’s work) which was entirely devoted to Hīrāb al Khāta’ayn which literally means ‘rule of the two errors’ and can be recognized as the Rule of Double False Position [27].

The first evidence of the use of the Rule of Double False Position in 12th-century India is found in an anonymous Latin book, Liber Augmenti et Diminutionis which literally means ‘Book of Increase and Decrease.’ This Latin book was translated from Arabic and presented the rule Hisab al Khāta’ayn, (in Latin translates to ‘Regula Augmenti et Diminutionis,’ and in English translates to ‘Rule of Increase and Decrease’) to solve a linear equation.

In 1202, Leonardo Pisano or “Fibonacci,” wrote Liber Abaci, ‘Book of the Abacus,’ which contained fifteen chapters dealing with arithmetic and algebra including a mixture were likely to encounter [20].

---

16 Only John of Seville’s Latin translation from the beginning of the 12th century of Al-Khwārizmī’s second book, Algoritmi de numero indorum, ‘Calculation within Indian Numerals,’ still exists [20]. Latin was the lingua franca of the scientific world [118].

17 In 12th-century India, problems were posed simply for the pleasure of solving them instead of for utilitarian function (unlike the texts that previously contained mathematics).
of Indian arithmetic methods and Arab algebraic methods.\textsuperscript{18} In Chapter 13 of \textit{Liber Abaci}, Fibonacci described the Arabic rule, Elchataym (which can be recognized as the Rule of Double False Position) and referred to it as the Augmented and Diminished Method. He applied this rule to rhetorical problems that represent linear equations and 29 of these problems were reproduced with little or no change from the Kamil’s Arabic ‘Book of Algebra’ \textsuperscript{20}, \textsuperscript{112}.

In 1484, Chuquet completed a three-part mathematical manuscript\textsuperscript{19} entitled “Triparty.” In the final section of the first part of his “Triparty,” Chuquet described the Rule of Double False Position which he called “the rule of two false positions.” We believe that this marks the inception of the English naming that evolved into the name Rule of Double False Position.

\subsection{2.1.5 Introduction of the term \textit{Regula Falsi}}

In the 16th century, Latin names and terms were introduced to describe existing mathematical methods. In 1527, Bienewitz [1], a.k.a. Petrus Apianus, introduced the term “\textit{Regula Falsi},”\textsuperscript{20} his Latin translation of the Rule of Double False Position, and defined it as a method that “learns to produce truth from two lies” \textsuperscript{78}.\textsuperscript{21} Bienewitz explained that the

\textsuperscript{18}Fibonacci wrote \textit{Liber Abaci} after returning from extensive travel about the Mediterranean, visiting Egypt, Syria, Greece, Sicily, and Provence, to receive a solid mathematical foundation. \textit{Liber Abaci} was revised in 1228, circulated in manuscript form until it was printed in Italy in 1857, and was not translated into English until 2002 [20].

\textsuperscript{19}We have not been able to locate a copy of Chuquet’s original 1484 manuscript. However, we did obtain a study [43], published in 1985, that includes an extensive translation of Chuquet’s mathematical manuscript.

\textsuperscript{20}The term “\textit{Regula Falsi}” literally translates to “rule of falseness.”

\textsuperscript{21}This is secondhand information from Maas. I recently obtained a digital copy of a scan of Bienewitz’ German text and still need to determine for myself what it contains.
term ‘false’ is used because the solution is produced from two ‘false’ initial estimates and not because the method is wrong or false [1].

The fact that Bienewitz introduced the term Regula Falsi to refer to the method already known as the Rule of Double False Position (and named “Rule of Two False Positions” by Chuquet in 1484) arguably explains how the Rule of Double False Position acquired the name the Regula Falsi method. From this point on, the Rule of Double False Position was referred to as not only the Rule of Double False Position but and also as Regula Falsi.\textsuperscript{22} This marks the start of the naming confusion involving Regula Falsi which we elaborate on in §2.4.

\subsection{Modification of the Rule of Double False Position applied non-iteratively to Quadratic Equations in One Unknown}

Recall that the Rule of Double False Position was originally defined for linear equations but was also used, in a non-iterative manner, to obtain approximate solutions to quadratic equations. In 1540, Frisius [48] claimed that he was the first to apply the Rule of Double False Position (which he called Regula Falsi) to quadratic equations of the form $ax^2 = b$.\textsuperscript{23} Frisius’ application of (his slightly modified version of) the Rule of Double False Position

\textsuperscript{22}The term “Regula Falsi” came into use far before the Regula Falsi method, as defined in Chapter 1, was developed (in the 1950s) and it was used to describe the Rule of Double False Position.

\textsuperscript{23}German mathematician Christoff Rudolf, in 1525, wrote the first German algebra book \textit{Die coss}, which means “the variable,” where the the Rules of Coss (where $ax^2 = b$ represents the second Rule of Coss, $ax^3 = b$ represents the third Rule of Coss, etc.) are presented and the modern symbol for the square root is introduced. Gemma Frisius made his claim in response to Rudolf’s comment that it was impossible to solve the second, third and fourth Rules of Coss using Rules of False which consist of the Rules of Single and Double False Position. Frisius solved the third and fourth Rules of Coss using a modification of the Rule of Single False Position.
was an exercise performed strictly out of theoretical interest since, by this time, algebra was known and practiced. Consider the following problem\textsuperscript{24}:

*From a rectangle of 200 square yards the length is one and a half times the width. What are the length and the width.*

This example can be written using algebraic notation as the system of two equations:

\[
\begin{align*}
l \times w &= 200 \\
l &= \frac{1}{2}w
\end{align*}
\]

where \(l\) and \(w\) represent the length and width of the rectangle respectively. Frisius was aware that he could have used direct substitution to simplify this system to \(1\frac{1}{2}w^2 = 200\), a quadratic equation in one variable, and solve for \(w^2\), however, he wanted to demonstrate that it was possible to solve the problem using the Rule of Double False Position. To do this, he first let \(w_0 = 4\), then the length is 6, the area is 24 and \(c_0 = \frac{1}{2}w_0^2 - 200 = -176\).

Next, he let \(w_1 = 20\), then the length is 30, the area is 600 and \(c_1 = \frac{1}{2}w_1^2 - 200 = 400\).

At this step, Frisius modified the Rule of Double False Position. Instead of solving for \(w\) as \(\frac{w_0c_1 - w_1c_0}{c_1 - c_0}\), he evaluated \(w^2\) as \(\frac{w_0^2c_1 - w_1^2c_0}{c_1 - c_0}\) and took the square root of this result.\textsuperscript{25} Frisius calculated the width to be \(11\frac{27}{55}\) and the length to be \(15\frac{77}{100}\), however, \(11\frac{27}{55} \times 15\frac{77}{100} \approx 181 \neq 200\). Frisius knew, as he stated, that the Rule of False is correct only for linear equations. He demonstrated that he realized that the solution he attained from using his modified Rule of Double False Position on his example problem described above was only approximate.

\textsuperscript{24}Taken from p.72 of Smeur [113] but originally appeared in Frisius [48].

\textsuperscript{25}At the time, knowledge of square and cube roots was known and root tables existed for quick reference.
when he stated that it is impossible to get the exact answer using this method.

2.1.7 Use of the Rule of Double False Position as an Iterative Process

- The Secant Method

In 1545, Cardano [26], in his *Artis Magnae*, demonstrated that the Rule of Double False Position (calling it “De Regula Liberae Positionis,” which literally translates to ‘(Concerning) the rule of free position’) could be used as an iterative procedure. He described the rule as an iterative process where multiple steps must be performed in order to improve the approximation [25]. He solved quadratic and cubic equations using the rule and included explanations of how he solved the problems using the rule with elaborate geometric illustrations [6]. We now have the secant method for a nonlinear equation. Cardano called it “De Regula Liberae Positionis.” Of course, this awkward name never achieved acceptance in the literature.

2.2 Newton’s Geometric Approach to the Secant Method

Newton kept a notebook of his scientific and mathematical ideas. Whiteside’s collection of these unpublished notes, entitled *Newton’s Waste Book* [124], includes in Volume I an illustration of his interpretation of a technique Newton used to approximate a zero

---

26 We were able to scan the microfiche of the original text (which is in Latin) and make a copy.
27 Newton studied the writings of John Wallis and the Dutch, as well as examined existing algorithms (such as that of of Viète and William Oughtred) to find the root of a polynomial. Newton had also made a thorough study of Descartes’ *La Géométrie* and from it took his algebraic symbolism and the use of algebraic equations to describe geometrical curves [69].
28 Volume I covers the period 1664-1669.
of a nonlinear equation. Newton’s technique (which is not referred to by any name) is equivalent to taking one step of the secant method in 1-D. Ypma [129], using the phrasing in Newton’s text and Whiteside’s annotation of it, illustrated the similar triangles underlying Newton’s geometric approach for approximating a zero of a function, and iterated the method.  


Although Newton picked both initial estimates to be on the same side of a zero, both Whiteside and Ypma’s description of Newton’s technique allowed for the initial estimates to be on either side of the zero.

2.3 The Naming and First Convergence Rate Proof of the Secant Method in 1-D

Thomas Fincke [42] introduced the word “secant” in his 1583 treatise on geometry [6]. In 1958, T.A. Jeeves [74] seems to be the first to use the term “secant method” to refer to the secant method (which Jeeves explained is ‘the secant modification of Newton’s method’), however, initially subsequent mathematical texts did not perpetuate the use of the name “secant method.” None of the earlier works Jeeves referenced utilized this name. Jeeves only included a footnote referencing the previous work of Wegstein who, in this 1958 paper [122], presented the secant method (which he referred to as a ‘modified form of Newton’s method’) and explained that this method was contained implicitly in Willers’ 1948 book [126]. In that book, Willers presented three methods - the method of false position, the


---

29 Whiteside did not iterate Newton’s technique.
method of iteration, and the Newton method of approximation - for determining the roots of any function of one variable, or of several functions with the corresponding number of variables for the latter two methods. In his method of false position, Willers did not specify whether or not one of the initial estimates remains fixed so we cannot assume that it does.\(^{30}\)

As a result, the method that Willers called the Method of False Position could be either our Regula Falsi or our Modified Regula Falsi method described in Chapter 1.

Jeeves also presented the first proof that we can locate of the golden mean convergence rate of the secant method (in 1-D). He proved that at each iteration of the secant method, the increase in the number of significant digits is \(\frac{1}{2}(1 + \sqrt{5}) \approx 1.62\) (the golden mean) times the previous increase. This is one of the convergence results discussed in Chapter 3.

### 2.4 Confusion and Inconsistencies in Naming Methods

#### 2.4.1 Initial Misuse and Confusion of the Name Regula Falsi

In 1955, Booth [9] was arguably the first to describe the Regula Falsi method for a nonlinear equation as defined in Chapter 1 and call it by that name, however, he stated that it was also known as the Rule of Double False Position. Although we know the two methods to not be the same, Booth must have believed them to be equivalent. This is further demonstrated by the fact that he explained that the Regula Falsi method is an ancient method also known as the Method of Double False Position. The fact that Booth utilized the term Rule of

\(^{30}\)Willers explained that this method was frequently used in business arithmetic, for example, for the determination of the effective interest on a loan from tables for the cash value of a bond.
Double False Position (a name which followed from Chuquet) and the pre-existing term (Regula Falsi), that Bienewitz introduced in 1527, to describe a new different method, greatly contributed to the confusion in the use of the names Rule of Double False Position and Regula Falsi in the mathematical texts of the following decades.

2.4.2 Evolution of the Names: Rule of Double False Position and Regula Falsi

The secant method evolved from the Rule of Double False Position which predated Newton’s method by over 3000 years. Even though the Rule of Double False Position dates back to the 18th century B.C., it was not thought of as a general rule or a method at that time, and therefore, was not given a specific name. It was not until 200 B.C., in China, that it was considered a general rule and given a name - yíng bù zú shu. Since then, it has been given different names (see Table 2.2) and has most commonly been referred to as the Rule of Double False Position since the 11th century A.D.

2.4.3 Continued Confusion of Names and Understanding

The naming confusion continued and grew to include other methods. Table 2.3 represents some of the different names that have been used to describe the Regula Falsi method, the Modified Regula Falsi method and the secant method. The blank spaces in the table should be interpreted as either the method was not presented or the method was not referred to by a specific name by that particular author(s). We use the remainder of this chapter to remark
Table 2.2: Evolution of the naming of the Rule of Double False Position

<table>
<thead>
<tr>
<th>Country</th>
<th>Century</th>
<th>Rule Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egypt</td>
<td>18th B.C.</td>
<td>-</td>
</tr>
<tr>
<td>Babylonia</td>
<td>18th B.C.</td>
<td>-</td>
</tr>
<tr>
<td>China</td>
<td>2nd B.C.</td>
<td>yíng bù zú (too much and not enough)</td>
</tr>
<tr>
<td>Arab</td>
<td>9th A.D</td>
<td>hisab al-Khataa'yn (rule of two errors)</td>
</tr>
<tr>
<td>Europe</td>
<td>11th A.D</td>
<td>elchataym (two errors)</td>
</tr>
<tr>
<td>Africa</td>
<td>13th A.D.</td>
<td>method of scales</td>
</tr>
<tr>
<td>Europe</td>
<td>15th,16th A.D.</td>
<td>rule of two false positions/regula falsi/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rule of double false position/regula positionum</td>
</tr>
<tr>
<td>America</td>
<td>20th A.D.</td>
<td>rule of double false position/method of false position/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>regula falsi/secant method</td>
</tr>
</tbody>
</table>

As we mentioned earlier, Booth, in 1955, was the first we found to present a description of the Regula Falsi method consistent with our description in Chapter 1. He explained that it could either be called Regula Falsi or Rule of False Position.\(^{31}\) In some earlier texts, for example Whittaker and Robinson [125], Willers [126], and Householder,[70], the methods were not rigorously presented. Primarily, no specific mention was made of whether or not one of the initial estimates was to remain fixed.\(^{32}\) As a result, it is difficult to determine if they are describing the Regula Falsi or the Modified Regula Falsi method.

\(^{31}\)Booth explained the ancient Egyptians invented the method of false position for the solution of nonlinear algebraic equations. However, the historical evidence we found reveals that it was first applied to linear equations.

\(^{32}\)In these descriptions, only the first step was described in detail. If graphs were included, only the first step was depicted to demonstrate how the choice of the initial estimates bracket a zero. Furthermore, only examples of convex or concave functions were presented. All of this considered, it is not clear if the method intended to retain one of the initial estimates in subsequent iterations.
<table>
<thead>
<tr>
<th>Date</th>
<th>Authors</th>
<th>Regula Falsi (R.F.)</th>
<th>Modified R.F.</th>
<th>Secant Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1944</td>
<td>Whittaker and Robinson</td>
<td>rule of false position</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1948</td>
<td>Willers</td>
<td></td>
<td>method of false position</td>
<td></td>
</tr>
<tr>
<td>1953</td>
<td>Householder</td>
<td></td>
<td>regula falsi</td>
<td></td>
</tr>
<tr>
<td>1955</td>
<td>Booth</td>
<td>regula falsi/ rule of false position</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1958</td>
<td>Wegstein</td>
<td></td>
<td></td>
<td>modified form of Newton’s method</td>
</tr>
<tr>
<td>1958</td>
<td>Jeeves</td>
<td></td>
<td></td>
<td>secant method</td>
</tr>
<tr>
<td>1960</td>
<td>Ostrowski</td>
<td>regula falsi</td>
<td>iteration with successive adjacent points</td>
<td></td>
</tr>
<tr>
<td>1961</td>
<td>Stanton</td>
<td></td>
<td>regula falsi</td>
<td></td>
</tr>
<tr>
<td>1962</td>
<td>Hochstrasser</td>
<td>rule of false position</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1964</td>
<td>Traub</td>
<td></td>
<td>regula falsi</td>
<td>secant iteration function (I.F.)</td>
</tr>
<tr>
<td>1964</td>
<td>Henrici</td>
<td></td>
<td></td>
<td>regula falsi</td>
</tr>
<tr>
<td>1964</td>
<td>Fröberg</td>
<td></td>
<td></td>
<td>regula falsi</td>
</tr>
<tr>
<td>1966</td>
<td>Isaacson and Keller</td>
<td>classical regula falsi method</td>
<td>method of false position</td>
<td></td>
</tr>
<tr>
<td>1970</td>
<td>Ortega and Rheinboldt</td>
<td>regula falsi</td>
<td></td>
<td>secant method</td>
</tr>
<tr>
<td>1972</td>
<td>Blum</td>
<td></td>
<td>method/rule of false position/ regula falsi</td>
<td></td>
</tr>
<tr>
<td>1974</td>
<td>Dahlquist and Björck</td>
<td></td>
<td>regula falsi</td>
<td></td>
</tr>
<tr>
<td>1975</td>
<td>Smeur</td>
<td>rule of false</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1977</td>
<td>Gellert, Hellwich, Küstner and Kästner</td>
<td>fixed point method</td>
<td>secant method</td>
<td>method of false position</td>
</tr>
<tr>
<td>1978</td>
<td>Atkinson</td>
<td></td>
<td>regula falsi</td>
<td></td>
</tr>
<tr>
<td>1981</td>
<td>Gill, Murray and Wright</td>
<td>regula falsi/ method of false position</td>
<td></td>
<td>secant method/ method of linear interpolation</td>
</tr>
<tr>
<td>2008</td>
<td>Dahlquist and Björck</td>
<td>false-position method/ regula falsi</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3: Some examples of the inconsistencies in naming the Regula Falsi Method, the Modified Regula Falsi Method, and the Secant Method
In 1960, Ostrowski [92] was the first that we found to distinguish between the Regula Falsi method, as defined in Chapter 1, and the secant method when he explained that the secant method uses the two last points instead of constantly using one of the initial points. However, he referred to the secant method as “iteration with successive adjacent points” and explained that Regula Falsi could also be called Method/Rule of False Position. In 1961, Stanton [115] was the first we found to explicitly describe our Modified Regula Falsi method which he referred to as the Regula Falsi method. He stated that there are various modifications of the Regula Falsi method but he does not present any modifications.

The inconsistency in naming the methods continued throughout the 1960s. For example, in a 1962 book edited by Todd [119], Hochstrasser described the Regula Falsi method consistent with our description in Chapter 1, but called it Rule of False Position which is essentially the same. Then, he explained that “Instead of keeping $\overline{x}$ fixed, one can move it during the computation - for example, by always using the latest two points given by the iteration.” He described the secant method but did not call it by any name.\(^{33}\) In 1964, Traub [121] explained that the secant method (which he called the Secant Iteration Function (I.F.)) always used ‘the latest two approximants.’ He further explained that the Secant I.F. is closely related to the Regula Falsi method which he defined as ‘a method that keeps two approximants which bracket the root’ (therefore, describing our Modified Regula Falsi method).

The first instance we found of the term “Regula Falsi” being used to describe the secant

\(^{33}\) Hochstrasser referenced Ostrowski’s earlier work, in which Ostrowski described the Regula Falsi and secant methods but unlike Hochstrasser, Ostrowski gave the secant method a descriptive name - “iteration with successive adjacent points.”
method was by Henrici [67] in 1964. Another instance occurred in the same year when Fröberg [49] presented the secant method and explained it was a ‘variable secant’ formula known as Regula Falsi.34

Two years later, in 1966, Isaacson and Keller [73] used the term “Classical Regula Falsi Method” to describe the Regula Falsi method defined in Chapter 1. Their use of the term “Classical” more than likely implies that they believed that this method was defined long ago. However, they referred to the secant method as the Method of False Position which interestingly this name was used for a method that was defined and named prior to the Regula Falsi method.

An interesting inconsistency that may have greatly promoted the naming confusion was found in subsequent texts of Dahlquist and Björck. In their 1974 book [29], they distinguished between the secant method and the Modified Regula Falsi method (which they referred to as Regula Falsi).35 Yet, in their 2008 book they described our Regula Falsi method (not the Modified Regula Falsi method) and stated that it was the Latin translation of the false-position method.36 Thus, they used the term Regula Falsi in two subsequent texts to describe two different methods.

The first naming convention that we found to be consistent with both the descriptions of the Regula Falsi and secant methods presented in Chapter 1 was by Ortega and Rheinboldt

34This is an English translation of the 1962 Swedish edition Lärobok i numerisk analys.
35This is an English translation and extension of the 1969 Swedish edition.
36They explained that the Regula Falsi method was a very old method that originated in 5th-century Indian texts and was used in medieval Arabic mathematics. They further explained that it got its name from the Italian mathematician Fibonacci. The only truth in these statements is that this method was used in Arabic texts. As for the other statements, we have already explained why they are false.
[91] in 1970. We claim that their usage of the names is more accurate in reflecting how the methods were historically introduced. Therefore, we view this as an attempt to clarify the existing confusion. Unfortunately, this naming convention was not perpetuated because in 1972, Blum [8] described the secant method and said it was called Method (or Rule) of False Position or Regula Falsi.


Moreover, references to the actual Rule of Double False Position became a part of the naming confusion. In 1978, Smeur [113] described the Rule of Double False Position, explained that it is called Regula Falsi or Rule of False, and stated that the rule is only correct for linear equations.38 In 1991, Hämmerlin and Hoffman [65] stated that the Regula Falsi method was one step of the secant method and that the secant method was a result of iterating the Regula Falsi method. It seems to be implicit that they understood that the Rule of Double False Position (which they called Regula Falsi - the name originally introduced by Bienewitz) was used for linear equations, in turn, implying that the secant method was used for nonlinear equations and was iterated.

---

37Atkinson presented this method as an improvement to the bisection method.
38Smeur stated that this rule could be found in the 1537 Dutch book by G.V. Hoecke. Like Bienewitz, Hoecke and Smeur both use the term Regula Falsi to refer to the Rule of Double False Position.
We showed that the terms Regula Falsi and Rule of False Position have been used interchangeably to describe the Regula Falsi method, the Modified Regula Falsi method, as well as the secant method. In addition, we presented some of the many inconsistencies in the naming of the Regula Falsi method, the Modified Regula Falsi method and even the secant method. To make sense of terms, we offer the following descriptions. A “false position” is an initial approximation to the solution. “Regula Falsi” or “Rule of False Position” is any method that uses linear interpolation based on two false positions to obtain a new approximation to the solution. We must admit that in recent years, primary sources have been overlooked. As a result, contemporary usage is as follows. The Modified Regula Falsi method described in Chapter 1 is now what current mathematics texts and popular websites call the Regula Falsi method and it is presented in sections on bracketing methods since it is a natural extension of the method of bisection (which is a bracketing method). Furthermore, the Regula Falsi method described in Chapter 1 is ignored in current mathematical texts. However, today everyone calls the secant method the secant method.
Chapter 3

Extension of the 1-D Secant Method to \(n\)-D

We continue the historical development of the secant method with a discussion of how it was extended to an \(n\)-dimensional setting. In 1970, Ortega and Rheinboldt [91] described a general framework underlying the construction of a basic secant approximation which was built to present the theory introduced independently by Bittner [7] and Wolfe [127] in 1959.\(^{39}\) In this chapter, we describe the linear interpolation and discretized Newton formulations of the secant method in \(n\) dimensions (\(n\)-D) and give examples of each formulation. We present the general framework of Ortega and Rheinboldt and show that their iterative method is a secant method in the sense that the secant equation (1.9) is satisfied. In addition to discussing some properties of the generalized secant methods, we address the potential

\(^{39}\)Ortega and Rheinboldt credit the modern generalization of the secant method to \(n\)-D to Heinrich and his unpublished lectures (circa 1955).
poor behavior of higher-dimensional secant methods.

3.1 General Position

To aid in our discussion of $n$-D secant methods, we define the term general position that Ortega and Rheinboldt utilized (in §7.2 of [91]) to describe a concept previously used, but not formally defined, by Bittner [7]. This definition is followed by a corresponding proposition that enlightens us with respect to the significance of points being in general position. First, we introduce some notation. The $i$th component of the point $x \in \mathbb{R}^n$ is denoted

$$(x)_i \quad \text{for} \quad i = 1, \ldots, n.$$ 

In the case where we are interested in $m$ points at the $k$th iteration, we denote them as

$$\{x_{k,1}, \ldots, x_{k,m}\}.$$

**Definition 3.1.1.** Any $n + 1$ points $x_0, \ldots, x_n$ in $\mathbb{R}^n$ are said to be in general position if the points $x_0 - x_k$, for $k = 1, \ldots, n$, are linearly independent.

**Proposition 3.1.2.** Let $x_0, \ldots, x_n$ be any $n + 1$ points in $\mathbb{R}^n$. Then the following statements are equivalent:

(a) $x_0, \ldots, x_n$ are in general position.

(b) For any $k$, $0 \leq k \leq n$, the points $x_k - x_m$, $m = 0, \ldots, n$, $k \neq m$, are linearly independent.
(c) The $(n+1) \times (n+1)$ matrix, $(e, X^T)$, where $e^T = (1, \ldots, 1)$ and $X = (x_0, \ldots, x_n)$ is nonsingular.

(d) For any $y \in \mathbb{R}^n$, there exist scalars $\alpha_0, \ldots, \alpha_n$ with $\sum_{m=0}^{n} \alpha_m = 1$ such that $y = \sum_{m=0}^{n} \alpha_m x_m$.

For a proof, see Ortega and Rheinboldt [91]. The geometric interpretation of general position given by Ortega and Rheinboldt [91] is that the points $x_0, \ldots, x_n$ are in general position if they do not lie in an affine subspace of dimension less than $n$. For example, for $n = 2$, the points $x_0, x_1, x_2$ are in general position if they are not colinear, that is, if they do not lie on a line in $\mathbb{R}^2$.

3.2 Linear Interpolation Methods

The secant method in $n$-D is used to approximate the solution of $F(x) = 0$ for $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$. Recall that a function $L : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is called an affine function if it is a translate of a linear function, i.e., if it has the form

$$L(x) = Bx + a \quad \forall x \in \mathbb{R}^n$$

(3.1)

where $B \in \mathbb{R}^{n \times n}$ and $a \in \mathbb{R}^n$. The 1-D secant method can be generalized to higher dimensions using the linear interpolation idea. In this formulation, the function $F$ is approximated by the affine function (3.1) that interpolates $F$ at the $n + 1$ given points, i.e., $L(x_k) = F(x_k)$ for $k = 0, \ldots, n$. Then, the linear system, $L(x_k) = 0$, is solved to obtain
3.2.1 The Ortega and Rheinboldt General Framework

In this section, we describe how the 1-D secant method may be generalized to $n$-D using the linear interpolation idea. We follow, in a direct manner, Ortega and Rheinboldt’s presentation on page 192 of [91].\(^4\) Begin with $n+1$ points, $x_0, \ldots, x_n$, and the corresponding $F(x_0), \ldots, F(x_n)$. Solve the linear system with $n$ right-hand sides

$$
\begin{bmatrix}
1 & (x_0)_1 & (x_0)_2 & \cdots & (x_0)_n \\
1 & (x_1)_1 & (x_1)_2 & \cdots & (x_1)_n \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & (x_n)_1 & (x_n)_2 & \cdots & (x_n)_n
\end{bmatrix}
\begin{bmatrix}
a_1 \\
a_2 \\
\vdots \\
B \end{bmatrix} =
\begin{bmatrix}
F(x_0)_1 & \cdots & F(x_0)_n \\
F(x_1)_1 & \cdots & F(x_1)_n \\
\vdots & \vdots & \vdots \\
F(x_n)_1 & \cdots & F(x_n)_n
\end{bmatrix}
$$

(3.2)

to obtain $B \in \mathbb{R}^{n \times n}$ and $a \in \mathbb{R}^n$ that satisfy

$$
Bx_k + a = F(x_k), \text{ for } k = 0, \ldots, n. \quad \text{(3.3)}
$$

Using (3.1) and (3.3), clearly we can see that the conditions

$$
L(x_k) = Bx_k + a = F(x_k) \quad \text{for } k = 0, 1, \ldots, n
$$

imply

$$
B(x_k - x_{k-1}) = F(x_k) - F(x_{k-1}) \quad \text{for } k = 1, \ldots, n. \quad \text{(3.4)}
$$

Next, solve for $x$ in

$$
Bx + a = 0, \quad \text{(3.5)}
$$

\(^4\)According to Ortega and Rheinboldt, the idea of replacing the function by a linear interpolating function to extend the 1-D secant method to higher dimensions dates back to Gauss [50], [51] in the 2-D case.

\(^5\)In 1959, Bittner [7] presented a description of the $(n+1)$-point method.
that is, solve for $x$ in the linear system

$$L(x) = 0.$$  \hfill (3.6)

If the $x_0, \ldots, x_n$ are in general position, then the affine function (3.1) is unique. If the $F(x_0), \ldots, F(x_n)$ are in general position, then $B$ is invertible. Accordingly, the solution to the linear system (3.6) is well-defined and this point

$$x^s = -B^{-1}a$$  \hfill (3.7)

is what Ortega and Rheinboldt call a basic secant approximation with respect to $x_0, \ldots, x_n$.

From (3.3), we can write the iteration

$$x_{k+1} = -B_k^{-1}a_k = B_k^{-1}(B_kx_k - F(x_k)) = x_k - B_k^{-1}F(x_k)$$  \hfill (3.8)

from which a basic secant approximation is obtained. From (3.4), we see that $B$ not only satisfies the secant equation, but it also satisfies the previous $n$ secant equations.

It turns out that there is no need to compute the interpolating function (3.1) explicitly. In the next section, we outline an alternative formulation which shows that a basic approximation (3.7) can be obtained by solving one linear system, namely (3.9), instead of solving for $a$ and $B$ in (3.2), and then solving for $x$ in (3.5) as described above.
3.2.2 Wolfe’s Method

In 1959, Wolfe [127] suggested and implemented an interpolation formulation of the \((n + 1)\)-point secant method for simultaneous nonlinear equations that he viewed as a generalization of the 1-D secant method to \(n\)-D. We present Wolfe’s method in the context of Ortega and Rheinboldt’s general framework and show that Wolfe’s basic secant approximation is clearly the same as the Ortega-Rheinboldt basic secant approximation (3.7).

Begin with \(n + 1\) points, \(x_0, \ldots, x_n\), and the corresponding \(F(x_0), \ldots, F(x_n)\); each of the two sets in general position. Solve the \((n + 1) \times (n + 1)\) linear system

\[
\begin{bmatrix}
1 & 1 & \cdots & 1 \\
F(x_0) & F(x_1) & \cdots & F(x_n)
\end{bmatrix}
\begin{bmatrix}
z_0 \\
z_1 \\
\vdots \\
z_n
\end{bmatrix} = \begin{bmatrix}
1 \\
0 \\
\vdots \\
0
\end{bmatrix}
\] (3.9)

for \(z = (z_0, \ldots, z_n)^T\) which is the unique solution since the \(F(x_k)\) are in general position. Note that 0 in the right-hand side of (3.9) is viewed in \(\mathbb{R}^n\). Using (3.3), we can write

\[
0 = \sum_{k=0}^{n} z_k F(x_k)
= \sum_{k=0}^{n} z_k (Bx_k + a)
= B \left( \sum_{k=0}^{n} z_k x_k \right) + a.
\]

Since \(x^s\) is the unique solution of \(Bx + a = 0\), the point that Wolfe calls a basic secant approximation satisfies

\[
x^s = \sum_{k=0}^{n} z_k x_k,
\] (3.10)
and is clearly the same as the Ortega-Rheinboldt basic secant approximation (3.7).

Observe that in Wolfe’s interpolation method, a basic secant approximation is obtained by solving only one linear system of equations - namely (3.9). This is followed by the calculation of the linear combination of the vectors $x_0, \ldots, x_n$ by means of (3.10).

However, the $(n + 1)$-point method can fail due to the loss of general positioning of the $x_j$. We elaborate on this point in §3.4 and §3.5 and then discuss the convergence properties of the $(n + 1)$-point method and other generalized secant methods in §3.4.1.

### 3.3 Discretized Newton Methods

The 1-D secant method can equivalently be generalized to higher dimensions as a discretized Newton method instead of using the linear interpolation idea. Recall that to obtain the 1-D secant method

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k),$$

the derivative $f'(x_k)$ in the 1-D Newton iteration (1.1) can be replaced with the difference quotient (1.2).

In the $n$-D discretized Newton formulation, to avoid explicitly computing the derivative, $F'(x_k)$, in the $n$-D Newton iteration

$$x_{k+1} = x_k - F'(x_k)^{-1} F(x_k),$$

---

42At each iteration, Wolfe replaced the $x_j$ for which $F(x_j)^T F(x_j)$ is maximal.
the Jacobian $F'(x_k)$ is replaced with the matrix of difference quotients, $J(x_k, H_k)$:

$$J(x_k, H_k) = (F(x_k + H_k e_1) - F(x_k), \ldots, F(x_k + H_k e_n) - F(x_k))H_k^{-1}$$

(3.12)

that approximates the Jacobian, $F'(x_k)$. Clearly, the matrix of difference quotients (3.12) is the $n$-D counterpart to the 1-D difference quotient (1.2) which can be written as

$$f(x_k) - f(x_{k-1})$$

$$h(x_k)$$

where we write $h_k = x_k - x_{k-1}$, to make the comparison more evident. In (3.12), $e_i$ for $1 \leq i \leq n$ represent the natural basis vectors for $\mathbb{R}^n$ and the matrix $H$ at the $k$th iteration

$$H_k = (x_{k,1} - x_{k,0}, \ldots, x_{k,n} - x_{k,0}) \equiv (h_{k,1}, \ldots, h_{k,n})$$

for $h_{k,i} \in \mathbb{R}^n$ (3.13)

is constructed from chosen auxiliary points, $\{x_{k,1}, \ldots, x_{k,n}\}$. Substituting the matrix of difference quotients (3.12) for $F'(x_k)$ in the $n$-D Newton iterations (3.11) yields the $n$-D discretized Newton iteration

$$x_{k+1} = x_k - J(x_k, H_k)^{-1}F(x_k)$$

(3.14)

for $k = 0, 1, \ldots$, from which a basic secant approximation is obtained: different choices of the auxiliary points determine different iterative methods. In the upcoming sections, we present examples of $n$-D discretized Newton methods, for example, $(n+1)$-point and 2-point methods.

### 3.3.1 $(n+1)$-point Secant Method

In this section, we outline how the discretized Newton formulation of the $(n+1)$-point secant method is used to obtain a secant approximation. We follow Ortega and Rhein-
boldt’s presentation on page 194 of [91]. Begin with $x_0, \ldots, x_n$ and the corresponding $F(x_0), \ldots, F(x_n)$; each of the two sets in general position. Use the $n$-D discretized Newton iteration (3.14) where the auxiliary points used to construct the $H$ matrix are chosen to be any $n$ points from the set of previously computed iterates $x_0, \ldots, x_k$ where $k \geq n$. Since $F(x_{k,i}) = F(x_{k,0} + He_i)$, it follows from (3.4) and (3.12) that

$$B_k H_k = \left( F(x_{k,0} + H_k e_1) - F(x_{k,0}), \ldots, F(x_{k,0} + H_k e_n) - F(x_{k,0}) \right)$$

$$= \left( F(x_{k,0} + h_{k,1}) - F(x_{k,0}), \ldots, F(x_{k,0} + h_{k,n}) - F(x_{k,0}) \right)$$

$$= \left( F(x_{k,1}) - F(x_{k,0}), \ldots, F(x_{k,n}) - F(x_{k,0}) \right).$$

(3.15)

Due to the general position of the $x_j$, we know $H$ is nonsingular. It follows from (3.12) and (3.15) that $B_k = J(x_k, H_k)$. Accordingly, since the $F(x_j)$ are in general position, the matrix of difference quotients, $J(x_k, H_k)$, is nonsingular. Also, observe that the matrix of difference quotients (3.12) satisfies the secant equation (1.9) because when we replace $s_k$ in the secant equation with $H_k$ and recall $B_k = J(x_k, H_k)$, then, it follows directly that

$$J(x_k, H_k) H_k = \left( F(x_k + H_k e_1) - F(x_k), \ldots, F(x_k + H_k e_n) - F(x_k) \right) H_k^{-1} H_k$$

$$= \left( F(x_k + H_k e_1) - F(x_k), \ldots, F(x_k + H_k e_n) - F(x_k) \right).$$

Using (3.8) and the fact that $B_k = J(x_k, H_k)$, we can write the iteration (3.14) from which a basic secant approximation is obtained. Clearly, this iteration reduces to the secant method in 1-D.

In the discretized Newton method described above, as in Wolfe’s interpolation method described in §3.2.2, a basic secant approximation is obtained by solving only one linear
system of equations followed by the calculation of a linear combination of the vectors $x_0, \ldots, x_n$.

3.3.1.1 Sequential $(n+1)$-point Secant Method

In the sequential $(n+1)$-point secant method, the $n$-D discretized Newton iteration (3.14) is used, and the auxiliary points that construct the $H$ matrix are not just any $n$ points from the set of most recently computed iterates, but specifically the $n$ previously computed iterates: 
\[
\{x_{k-1}, \ldots, x_{k-n}\}.
\]

3.3.2 Sequential 2-point Secant Method

In the sequential 2-point secant method, the $n$-D discretized Newton iteration (3.14) is used and the auxiliary points that construct the $H$ matrix depend on the two most recently computed iterates. Various authors explored 2-point secant methods. For example, Korganoff [76], in 1961, worked with a sequential 2-point method that uses the auxiliary points given by
\[
x_{k,j} = x_{k,0} + (x_{k-1,j} - x_{k,j})e_j \quad \text{for} \quad j = 1, \ldots, n.
\]
(3.16)

Clearly, $h_j = (x_{k-1,j} - x_{k,j})e_j$ for $j = 0, \ldots, n$ which demonstrates that the iteration (3.14) with the auxiliary points (3.16) reduces to the secant method in 1-D.

Ortega and Rheinboldt discuss another example of a sequential 2-point method that
uses the auxiliary points given by
\[ x_{k,j} = x_{k,0} + \sum_{i=1}^{j} (x_{k-1,i} - x_{k,i})e_i \text{ for } j = 1, \ldots, n. \] (3.17)

### 3.4 Properties of \((n + 1)\)- and 2-point Secant Methods

In general, secant methods require \(n + 1\) function evaluations at each step - namely at the points \(x_{k,0}, x_{k,1}, \ldots, x_{k,n}\). In particular, this is true for the 2-point method that uses the auxiliary points (3.16). However, in certain cases, the particular choice of the auxiliary points may permit fewer function evaluations, for example, the sequential 2-point method that uses the auxiliary points (3.17). In this case, \(x_n = x_{k-1}\) and since \(F(x_{k-1})\) is available from the previous stage, only \(n\) new function evaluations are required. However, the \((n+1)\)-point method requires an even fewer number of function evaluations than either of these 2-point methods.

At the first step of the sequential \((n + 1)\)-point secant method, we have \(x_0, \ldots, x_n\) and calculate \(F(x_0), \ldots, F(x_n)\). At subsequent steps, \(F(x_{k-1}), \ldots, F(x_{k-n})\) are already available, thus the sequential \((n + 1)\)-point secant method requires the computation of only one new function evaluation per step – namely, \(F(x_k)\). While this requirement of only one function evaluation per step is attractive, it comes at the cost of storing \(n + 1\) points and their corresponding function values at each step. Another disadvantage is that the \((n + 1)\)-point method is prone to unstable behavior. For example, the sequential \((n + 1)\)-point secant method can fail due to the fact that the points \(x_k - x_m\), for \(k \neq m\), that are in general position become numerically dependent, that is, they approximately lose their
general positioning (and, in particular, this is the case in higher dimensions). This is not the case for the 2-point methods. However, there are other reasons why 2-points methods were not pursued; this is discussed in the next section.

### 3.4.1 Convergence Properties of Generalized Secant Methods

In 1959, Wolfe [127] stated that his \((n + 1)\)-point method exhibited golden mean convergence for a variety of 2-D problems that he solved. Even though Wolfe did not provide convergence analysis, he believed his method to have the golden mean convergence that Jeeves [74] had demonstrated for the secant method in 1-D. We discuss the convergence theory later in this section.

For the remainder of our discussion of convergence behavior, we make the following assumptions for the nonlinear equation problem \(F(x) = 0\) with \(F : \mathbb{R}^n \to \mathbb{R}^n\):

1. There exists \(x^*\) such that \(F(x^*) = 0\). (3.18)
2. \(F\) is continuously differentiable in an open convex set \(D\) containing \(x^*\). (3.19)
3. \(F'(x^*)\) is nonsingular. (3.20)

We also include the following restatement of a definition given by Ortega and Rheinboldt [91] as it will be relevant in the upcoming discussion.

---

\(^{43}\)Ortega and Rheinboldt [91] present a 2-D example for which the \((n + 1)\)-point method fails.

\(^{44}\)Wolfe wrote a FORTRAN II program to test his procedure. He stated that for his 2-D example problems, only one of which he included in his 1959 paper, the error at a given step is proportional to the product of the errors at the two previous steps leading to a convergence order of \(\frac{1}{2}(\sqrt{5} + 1)\).
Definition 3.4.1. For given $\sigma > 0$ and $H$ given by (3.13), set

$$K(\sigma) = \left\{ H \in \mathbb{R}^{n \times n} \mid h_i \neq 0, i = 1, \ldots, n; \left| \det \left( \frac{h_1}{\|h_1\|}, \ldots, \frac{h_n}{\|h_n\|} \right) \right| \geq \sigma \right\}. \quad (3.21)$$

Then a collection of matrices $Q \in \mathbb{R}^{n \times n}$ is said to be uniformly nonsingular if $Q \subset K(\sigma)$ for some $\sigma > 0$.

Remark 3.4.2. All matrices in $K(\sigma)$ are nonsingular.

In 1959, Bittner [7], under the basic assumptions (3.18) - (3.20), and the further assumption that $\sigma > 0$ is chosen so that $K(\sigma)$ given by (3.21) is not empty, was the first to state that the convergence of the $(n + 1)$-point secant method is superlinear.

In 1964, Tornheim [120] proved that if $F$ satisfies the same conditions that Bittner assumed, and additionally, $F'$ satisfies the Lipschitz condition

$$\|F'(x) - F'(x^*)\| \leq \gamma \|x - x^*\| \quad \forall \ x \in D \text{ and } \gamma \geq 0, \quad (3.22)$$

where $D$ is an open convex set containing $x^*$, then the order of convergence of the sequential $(n + 1)$-point secant method is at least the value of the largest root of $r^{n+1} - r^n - 1 = 0$ where $n$ is the dimension of the space we are working in.\footnote{This result was also proved shortly after this by Barnes [4] in 1965 and Robinson [103] in 1966.} This means that at least golden mean convergence is achieved for $n = 1$. Observe that as $n$ increases, the value of the largest root decreases.

Tornheim [120] studied Wolfe’s 2-D example and stated that it complied with his convergence theory. Thus, Wolfe’s claim that golden mean convergence is achieved is consistent with Tornheim’s theory. While the order of convergence for $n = 2$ should be at
least 1.45, according to Tornheim’s theory, Wolfe said it was golden mean. However, we believe it is difficult to really ascertain the difference between whether it was 1.45 or 1.62 in Wolfe’s example.

In 1961, Schmidt [104] was the first to state that the convergence of the sequential 2-point secant method is superlinear. He showed that this is true when the sequential 2-point secant method is used to solve nonlinear equations in Banach spaces. In 1963, he concluded [105] that in addition, if $F'$ additionally satisfies the Lipschitz condition (3.22), then the sequential 2-point secant method is guaranteed to achieve golden mean convergence. Unlike the sequential $(n + 1)$-point secant method, this is true for any $n$.

### 3.5 Remarks

The 1-D secant method does not suffer from bad conditioning. Nonetheless, we showed, in this chapter, how its generalization to higher dimensions can become less effective. For example, in the $(n + 1)$-point methods that we described, as the dimension increases, the vectors $x_0 - x_k$ for $k = 1, \ldots, n$ that are in general position effectively lose general position, i.e., they become effectively numerically linearly dependent. As a result, the system we need to solve becomes ill-conditioned which makes it difficult to solve. Consequently, the algorithm becomes numerically unstable due to solving nearly singular systems. Furthermore, there was no serious implementation work on the 2-point methods. Therefore, it is in our considered opinion that for these reasons, historically, these methods have not been pursued. In addition, new exciting secant methods were being introduced that are nu-
merically effective in higher dimensions. These methods are the topic of the next chapter.
Chapter 4

Development of Secant Methods

In 1959, Davidon [30] introduced a new method for minimization using the secant idea and in 1965, Broyden [12] presented a new method for solving systems of equations using the secant idea. These iterative methods were unlike any others in use at the time due to the novel procedures used to approximate the Jacobian [36]. The work of Davidon, Broyden, as well as Fletcher and Powell, marks the birth of a new class of secant methods: these methods being the topic of this chapter. We explain how these methods have also been referred to by some as quasi-Newton or variable-metric methods but eventually the naming convention was to call them secant methods. The evolution of secant methods is traced from the perspective of solving nonlinear equations, unconstrained, and equality constrained optimization problems. We detail the development of the BFGS secant method and discuss the contributions of its developers. Throughout the chapter, concepts that aid in our understanding and appreciation of the methods are included.
4.1 Nonlinear Equations

Consider the nonlinear equations problem:

\[
\text{Given } \quad F : \mathbb{R}^n \rightarrow \mathbb{R}^n \quad (4.1)
\]

Find \(x^* \in \mathbb{R}^n\) such that \(F(x^*) = 0\).

By a secant method for the nonlinear equations problem (4.1), we mean the iterative procedure (1.6) where the nonsingular matrix \(B_k\) is updated at each iteration and the update satisfies the secant equation (1.10). From this point on, we drop the iteration index \(k\), when no confusion can occur, and adopt the notation \(x_+\) to represent \(x_{k+1}\), \(B_+\) to represent the \(B_{k+1}\) update, and so forth.\(^{46}\)

If we substitute the general formula for a rank-one update

\[
B_+ = B + uv^T \quad (4.2)
\]

where \(u, v \in \mathbb{R}^n\), into the secant equation (1.10), we obtain

\[
y = B_+ s = (B + uv^T)s = Bs + uv^Ts.
\]

From this, we get

\[
u = \frac{y - Bs}{v^Ts}.
\]

\(^{46}\)This notation is due to Davidon [30].
which when substituted into the rank-one update formula (4.2) yields

\[
B_+ = B + \frac{(y - Bs)v^T}{v^Ts}
\]  

(4.3)

for any \(v \in \mathbb{R}^n\) as long as \(v^Ts \neq 0\). This formula represents the class of rank-one updates that satisfy the secant equation. In upcoming sections, we present some choices of \(v\) that yield popular rank-one update methods.

It is important to mention here that updates can also be viewed from an inverse point of view, that is, instead of updating \(B\) we can let \(H = B^{-1}\) and update \(H\) where \(H_+\) must satisfy the inverse secant equation

\[
H_+y = s.
\]  

(4.4)

In the literature, \(B_+\) is often referred to as a direct update and \(H_+\) as an inverse update. It follows directly, that we can rewrite the class of rank-one direct updates (4.3) as the rank-one inverse update class

\[
H_+ = H + \frac{(s - Hy)d^T}{d^Ty}
\]  

(4.5)

where update (4.3) and update (4.5) are inverses if \(d = H^Tv\). This forces the question of why it would be of any interest to perform inverse updating instead of direct updating, the subject of the next section.

### 4.1.1 Inverse versus Direct Updating

We take a moment to mention the motivation behind performing inverse updating instead of direct updating. The computational cost of direct updating, when used in the secant method
is $O(n^3)$. The $O(n^3)$ work can be explained by the need to solve a linear $n \times n$ system at each iteration. In an effort to reduce the computational cost to $O(n^2)$, inverse updating was pursued. In this case, the inverse matrix itself is updated at each iteration. Consequently, at each iteration, all that is required to solve the linear $n \times n$ system for the search direction is taking the product of a matrix and a vector which reduces the computational cost to $O(n^2)$.

One of the researchers who spent considerable effort in the early days working on computational optimization problems was Walter Murray. In private communication (March 30, 2006), Murray stated that “the generally held view at the Math Division of the National Physical Laboratory (NPL) [when he worked there in the late 60s and early 70s] was that there was never any need to use an inverse.” This resulted in an interest to research methods that update the factors of $B$ allowing one to use direct updating in $O(n^2)$ work. We present one updating matrix factorization method in §6.3 and we refer the interested reader to Gill, Murray and Wright [58] to learn more about updating matrix factorization methods.

4.1.2 Good Broyden

In 1965, Broyden [12] presented two new methods for approximating Jacobian matrices which were intended for use to solve the nonlinear equations problem (4.1). These methods have subsequently been referred to as Good Broyden and Bad Broyden as a result of their good and bad numerical properties respectively.

Let $v = s$ in the general rank-one update class (4.3), with $s \neq 0$, to obtain Broyden’s
rank-one update formula

\[ B_+ = B + \frac{(y - B s)s^T}{s^T s} \]  \hspace{1cm} (4.6)

for square, non-singular \( B \), which is often called \textit{good Broyden}. If one knows \( B^{-1} \), a straightforward application of the Sherman-Morrison-Woodbury formula to the good Broyden update (4.6) yields the formulation of Broyden’s method expressed as

\[ B_+^{-1} = B^{-1} + \frac{(s - B^{-1} y)s^T B^{-1}}{s^T B^{-1} y} \]  \hspace{1cm} (4.7)

which is equivalent to (4.6). This formulation produces values of \( B_+^{-1} \) which allows us to update \( B^{-1} \) instead of \( B \), but makes it difficult to detect ill-conditioning in \( B_+ \). In addition, it is possible for the denominator in (4.7) to become zero.

The good Broyden update does not preserve symmetry and positive-definiteness: that is, even if \( B \) is symmetric, \( B_+ \) will not be unless \( y - Bs \) is a multiple of \( s \) [38]. Even though the good Broyden update was not designed to preserve symmetry, we will see later, when we consider optimization, that it is desirable for an update to possess this property since the Hessian is symmetric.

\subsection*{4.1.3 Bad Broyden}

In an attempt to generate values of \( B_+^{-1} \) directly instead of applying the Sherman-Morrison-Woodbury formula to the good Broyden update, Broyden interchanged \( B \) with \( B^{-1} \) and \( s \) with \( y \) in the good Broyden update (4.6) to get his second update formula

\[ B_+^{-1} = B^{-1} + \frac{(s - B^{-1} y)y^T}{y^T y}, \]
often called \textit{bad Broyden}. The bad Broyden update is attractive as it produces values of $B^{-1}$ directly. In addition, this formulation circumvents the problem of a zero denominator which could possibly occur in formulation (4.7). However, in practice, the bad Broyden update has been considerably less successful than the good Broyden update [38], hence the use of the terminology good and bad.

### 4.1.4 Symmetric Rank One (SR1)

If we let $v = y - Bs$ in the general rank-one update class (4.3), then we obtain the symmetric rank-one (SR1) update

$$B_+ = B + \frac{(y - Bs)(y - Bs)^T}{(y - Bs)^T s}$$

(4.8)

that Davidon presented in the Appendix of his 1959 paper [30]. While there is no convergence theory, SR1 works well often, but not always. One drawback of SR1 updating is that there is no guarantee that $(y - Bs)^T s \neq 0$, which would then cause the denominator in (4.8) to vanish. However, if $(y - Bs)^T s \neq 0$, then the SR1 update satisfies the secant equation and maintains the symmetry of the matrix but it does not guarantee that the updated matrix maintains positive-definiteness - a quality that we will see is desired in some applications.

### 4.1.5 Powell Symmetric Broyden (PSB)

In 1970, Powell [96], in an attempt to derive a symmetric secant update for solving symmetric systems, introduced the method of iterated projections, and used it to symmetrize the good Broyden update (4.6). We only briefly describe Powell’s application of his method
of iterated projections to good Broyden because this topic is revisited in more detail in Chapter 7.

Powell began with the good Broyden rank-one update (4.6) which satisfies the secant equation but, in general, is not symmetric. To symmetrize $B_+$, Powell considered
\[ C_+ = \frac{B_+ + B_+^T}{2}. \] (4.9)
Recall the Frobenius norm:
\[ \|C\|_F^2 = \sum_{ij} c_{ij}^2 = \text{trace}(CC^T). \] (4.10)
Powell’s construction of $C_+$ (4.9) from $B_+$ can be viewed as making the Frobenius norm projection of $B_+$ onto the subspace of symmetric matrices. At this point, the update (4.9) is symmetric but no longer satisfies the secant equation so Powell continued this process of iterated projections to generate the infinite sequence of updated matrices \( \{C_k\} \) where
\[ B_{k+1} = C_k + \frac{(y - C_k s)s^T}{s^T s}, \]
\[ C_{k+1} = \frac{B_{k+1} + B_{k+1}^T}{2}, \quad k = 1, 2, \ldots \]
The limit of the sequence \( \{C_k\} \) is the rank-two update
\[ B_+ = B + \frac{(y - Bs)s^T + s(y - Bs)^T}{s^T s} - \frac{s^T(y - Bs)ss^T}{(s^T s)^2} \] (4.11)
which is referred to as the Powell Symmetric Broyden (PSB) update [38].\(^{47}\) Like the SR1 method, the PSB method does not guarantee that the updated matrix maintains positive-definiteness. However, the PSB method does not fail in the same way as the SR1 method, yet in practice, the PSB method is not numerically effective.

\(^{47}\)Powell stated that this limit was the result of straightforward algebra but did not include a formal proof of how he obtained this limit.
While, Powell [96], in 1970, introduced the process of iterated projections and applied it to Broyden’s rank-one update, it was Dennis [33] who, in 1972, applied Powell’s method of iterated projections to the rank-one inverse update class (4.5) to develop what we refer to as the **Dennis class**:

\[
H_+ = H + \frac{(s - Hy)d^T + d(s - Hy)^T}{d^Ty} - \frac{y^T(s - Hy)dd^T}{(d^Ty)^2}
\]  

(4.12)

also written in the form of a direct update class:

\[
B_+ = B + \frac{(y - Bs)v^T + v(y - Bs)^T}{v^Ts} - \frac{s^T(y - Bs)vv^T}{(v^Ts)^2}.
\]  

(4.13)

Following Dennis and Walker [39], we call \(v\) in the Dennis direct update class (4.13) the *scale* of the update formula. The Dennis class has not received much attention in the literature, nor has it been credited to John Dennis. In Chapter 7, we examine the Dennis class in more detail and discuss the different choices of \(v\) that yield well-known updates since most of the well-known secant updates are members of the Dennis class. In addition, since Dennis did not construct a proof of how he used Powell’s method of iterated projections to derive the Dennis class, we construct one in §7.1.1.

### 4.1.6 Least Change Problem

We take a moment to present the property of least change. This idea, as we explain in §4.2.2.4, is due to Greenstadt [62]. We mention the least change property briefly here because many of the successful updates satisfy this property even though this concept is not what motivated the development of updating methods (with the exception of one instance...
of the update presented in §4.2.2.4).

The least change secant update strategy determines the best choice of available approximating matrices that preserves information gained from prior iterations by minimizing some reasonable measure of the change to the current approximation of the Jacobian matrix subject to the new approximation satisfying the secant equation. We call, for a given matrix norm, the constrained minimization problem

$$\min_{B_{+}} \| B_{+} - B \| \quad \text{subject to} \quad B_{+} s = y, \quad (4.14)$$

the least change problem, and we call $B_{+}$ determined according to this criterion a least-change secant update of $B$. One norm which is useful in the least change problem (4.14) is the Frobenius norm (4.10). Another useful norm is the weighted Frobenius norm (4.31) which we discuss in §4.2.3.

Seemingly, the unique solution to the least change problem in the Frobenius norm

$$\min_{B_{+}} \| B_{+} - B \|_F \quad \text{subject to} \quad B_{+} s = y, \quad (4.15)$$

should yield a quite suitable update formula to use in solving the nonlinear equation problem (1.4) when the Jacobian has no special structure which $B$ should possess. In 1977, Dennis, and Moré [36], presented a proof of how the good Broyden rank-one update (4.6) is a least change update to $B$. Accordingly, the bad Broyden update is a least change secant update of $B^{-1}$, i.e., it is the solution to the constrained minimization problem

$$\min_{B_{+}^{-1}} \| B_{+}^{-1} - B^{-1} \|_F \quad \text{subject to} \quad B_{+}^{-1} y = s.$$

\[48\] In private communication (February 11, 2009) John Dennis explained that “Jorge (Moré) was responsible for the elegant short proof that Broyden’s method is a least change method.”
Furthermore, Dennis, and Moré [36] state that the PSB update (4.11) is a least change symmetric secant update of $B$, i.e., it is the unique solution to the constrained minimization problem (4.15) with the added constraint that $B$ be symmetric. In the upcoming sections, we discuss more secant updates that are solutions to the least change problem in the Frobenius norm or the weighted Frobenius norm (4.31). While it is interesting that many of the successful updates are least change updates, satisfying the least change property is not enough to qualify an update as good. We now return to our study of the development of secant methods.

4.2 Unconstrained Optimization

Consider the unconstrained minimization problem

$$\min_x f(x)$$

(4.16)

where $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable. We may solve the minimization problem (4.16) by viewing it as a nonlinear equation problem where the nonlinear equation is obtained by setting $\nabla f(x) = 0$. By a secant method for the optimization problem (4.16), we mean the iterative procedure (1.7) applied to the nonlinear equation problem $\nabla f(x) = 0$ where the nonsingular matrix $B$ is updated at each iteration, for example, using updates obtained from different choices of $v$ in the Dennis class of updates (4.13). In §4.2.3 and §7.1, we discuss choices of scale $v$ that yield popular updates.
4.2.1 Davidon, Fletcher and Powell (DFP)

In 1959, Davidon [30] introduced a new method for minimizing nonlinear functions using the secant idea [62]. Davidon’s idea for updating $H$ marked the emergence of a class of secant methods in optimization which Davidon called variable-metric methods. At that time, Davidon used the term variable-metric because he viewed his method as Steepest Descent in a weighted inner product, i.e., in a varying metric. Others used the term quasi-Newton to refer to methods in which the derivative (in Newton’s method) is approximated. Today, most people call these updating methods secant methods.

In 1963, Fletcher and Powell [47] simplified and reformulated the method that Davidon originally proposed in 1959. The modified method has become known as the DFP (Davidon, Fletcher and Powell) method. We briefly compare and contrast the differences between Davidon’s original formulation and the DFP method. For starters, the manner in which the two algorithms were presented is very different. Davidon used a hunting metaphor to name the five parts of his algorithm\(^{49}\) and he described the iterative procedure using symbols for memory locations rather than successive values of a variable; e.g., he wrote $x + 3 \rightarrow x$ instead of $x_{k+1} = x_k + 3$. Fletcher and Powell’s presentation read more like the standard straightforward technical paper without the use of Davidon’s hunting metaphor, however, they described the method using Dirac bra-ket notation [40].\(^{50}\) This

---

\(^{49}\)Davidon chose the hunting metaphor with ‘tongue in cheek’ as he expected his paper would be read mostly by his friends who knew he opposed killing for a sport. Davidon’s use of this hunting metaphor was not well-received by the general public.

\(^{50}\)In the Dirac bra-ket notation notation applied to real vectors, the column vector $[x_1, x_2, \ldots, x_n]^T$ is written as $|x\rangle$. The row vector with these same elements is denoted by $\langle x|$. The scalar product of $\langle x|$ and $|y\rangle$ is written $\langle x|y\rangle$. It then follows, say for matrix $H$, that $H|x\rangle$ is a column vector, $\langle x|H$ is a row vector and $\langle x|H|y\rangle$ is a scalar.
is interesting because this notation was not common in this field at the time. Aside from
the difference in the aesthetic presentation, the underlying process of the DFP method was
embedded in Davidon’s original formulation.

Fletcher and Powell presented their procedure which was intended for use to minimize
a quadratic function. In their procedure, the initial inverse Hessian approximation could
be chosen to be any symmetric, positive-definite matrix, but was conveniently chosen to
be the identity matrix. In contrast, Davidon chose his initial approximation to be any sym-
metric positive semi-definite matrix and did not intend for his method to be used only with
quadratic functions. As a result, Davidon included options at each iteration to improve the
speed of convergence of his algorithm for nonquadratic functions. These options are the
key differences between Davidon’s original formulation and the DFP method. In the DFP
method, the search directions are conjugate at each step. We elaborate on the concept of
conjugacy in §5.1. When determining in which direction the step should be taken, Davidon
compares the improvement expected by taking a step in a conjugate direction to that which
is made by taking a step made by cubic interpolation - the better of the two options is used.
The purpose for allowing for this option was to improve the speed of convergence when the
function is not quadratic. However, the fact that the minimizer is often being sought along
more than one direction in a single iteration makes Davidon’s algorithm more complicated.

Fletcher and Powell point out that Davidon utilized an unsatisfactory procedure for
terminating his process which, in some instances, resulted in a poor inverse Hessian esti-
mation which in turn forced the procedure to make slow progress, consequently causing
the procedure to not converge in $n$ iterations or causing $H$ to not always converge to the inverse Hessian even for quadratics. Fletcher and Powell omitted Davidon’s options and explained two obvious and useful ways of terminating the procedure. One was to stop when the predicted absolute distance from the minimum is less than a prescribed amount and the other was to finish when every component of the search direction is less than a prescribed accuracy. In addition, they introduced two safeguards. The first was to perform at least $n$ (the number of variables) iterations, and the second was to test to make sure that all of the search directions were conjugate. Finally, Fletcher and Powell provided a theoretical basis justifying the manner in which $H$ was modified. They proved that for a strictly convex quadratic function the procedure terminates in at most $n$ iterations and the sequence of inverse Hessian approximations converges to the inverse Hessian matrix evaluated at the minimizer. They also proved that $H_+$ remains positive definite.

In the DFP method, the iteration (1.7) is utilized and the initial inverse Hessian approximation can be chosen to be any positive-definite symmetric matrix. At each iteration, $H$ is updated using the formula

$$H_+ = H - \frac{Hyy^TH}{y^THy} + \frac{ss^T}{s^Ty}.$$  \hspace{1cm} (4.17)

We can instead update $B$ using the DFP direct update formula which is most commonly written in what is called the product form\textsuperscript{51}

$$B_+ = \left(I - \frac{sy^T}{y^Ts}\right)^TB\left(I - \frac{sy^T}{y^Ts}\right) + \frac{yy^T}{y^Ts}. \hspace{1cm} (4.18)$$

\textsuperscript{51}Greenstadt [63] explained that while at the University of Dundee in 1971, he thought of the possibility of using product updates of the form $H_+ = (I + D)^TH(I + D)$ for inverse Hessians. This appears to be the advent of the product form.
The DFP method, generates symmetric updates that not only satisfy the secant equation, but are also positive-definite under mild conditions.

4.2.1.1 Conditioning Problem

As noted by Bard [3], Broyden [14], and Pearson [94], among others, the DFP method produces updates which may have increasingly small eigenvalues. Moreover, the DFP method suffers numerically when the eigenvalues of the approximating matrices become too small. The numerical problem of the eigenvalues of successive updating matrices tending to zero, causing the updating matrices to approach singularity, is often referred to as the conditioning problem.

4.2.1.2 Restarting

In 1968, Bard [3] explained that the conditioning problem could lead to the failure of the algorithm or premature termination depending on the particular stopping criterion. Bard credits McCormick [79] for noting that restarting (also known as resetting or reinitializing) the matrix every now and then to a positive-definite matrix improved the method’s performance. That same year, Pearson [94] observed that restarting $H$ may enable the method to behave better which in turn will aid in more effectively attaining the solution. As a result, a standard strategy to deal with the conditioning problem was to restart the iteration, generally after every $n$ iterations, by setting $H$ to a given positive-definite matrix, often the identity matrix [3], [14], [80]. Other remedies to deal with the conditioning problem of the DFP method led to the development of the BFGS secant method which is the focus of the
4.2.2 Broyden, Fletcher, Goldfarb, Shanno (BFGS)

In 1970, Broyden [15], Fletcher [45], Goldfarb [59], and Shanno [108] independently developed the BFGS (Broyden, Fletcher, Goldfarb, Shanno) rank-two secant update formula.\(^{52}\) In the next four sections, we describe how each of the developers arrived at the BFGS update formula which, if used as a direct update, is most commonly written as

\[
B_+ = B - \frac{Bss^TB}{s^TBs} + \frac{yy^T}{y^Ts}.
\]  (4.19)

but can also be written as an inverse update in product form

\[
H_+ = \left( I - \frac{ys^T}{y^Ts} \right)^T H \left( I - \frac{ys^T}{y^Ts} \right) + \frac{ss^T}{y^Ts}.
\]  (4.20)

The BFGS method not only produces symmetric positive-definite secant updates (as does the DFP method) but it also corrects the conditioning problem associated with the DFP method.

4.2.2.1 Fletcher

In 1970, Fletcher [45] made the interchange \(B \leftrightarrow H\) and \(s \leftrightarrow y\) in the DFP inverse secant update formula (4.17) to obtain the BFGS direct secant update formula (4.19). This interchange of variables, which makes the DFP and the BFGS updating formulas duals of

\(^{52}\)An interesting aside that Dave Shanno shared with me in private communication (September, 2006) was that he knows of a photograph that contains the four developers of the BFGS method and he thinks it is the only one. This photograph was taken June, 1981 outside the bar at Trinity Hall College, Cambridge.
each other, was a simple way to produce successive update matrices whose eigenvalues do not tend to zero. However, if the eigenvalues of the DFP updates are excessively small, then the interchange of variables may make the eigenvalues of the BFGS update excessively large: we address this issue in §6.5.3. For this reason, Fletcher (and others) mentioned the need to bound the updates from above and not just be concerned with their tendency to become singular. Fletcher mentioned that he had “heard very recently from C.G. Broyden in a private communication, that he has also come across this formula in a different way...” In the next section, we describe how Broyden [15] arrived at the BFGS update formula.

4.2.2.2 Broyden

In 1970, Broyden wrote a two-part paper [14], [15] while at the Computing Centre at the University of Essex. In Part 1 [14], he presented his class of updates (4.21) which is known as the Broyden class and is more commonly written as (4.37).\textsuperscript{53} We elaborate on this class in §4.2.5 but we present the original formulation of the Broyden class here as this is the formula Broyden used to obtain the BFGS inverse update formula.

Broyden considered minimizing a quadratic function of the form

$$F(x) = \frac{1}{2} x^T A x - b^T x + c$$

where $A$ is a symmetric, positive-definite matrix, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$. He presented his\textsuperscript{53}

\textsuperscript{53}Broyden [13] first posed his class of update formulae in 1967. While the BFGS update was implicit in this class, it was not identified as a member of this class until 1970.
class of updates using the following formula

\[ H_+ = H - H y w^T + p t q^T \]  \hspace{1cm} (4.21)

where

\[ p = -H f \]
\[ y = f_+ - f, \]
\[ q^T = \alpha p^T - \beta y^T H, \]
\[ w^T = \gamma y^T H + \beta t p^T, \]
\[ \alpha = \frac{1 + \beta y^T H y}{p^T y}, \]
\[ \gamma = \frac{1 - \beta t p^T y}{y^T H y}, \]
\[ t = -\frac{p^T A e}{p^T A p}, \]

\( e = x - x^*, \) and \( A(e) = A(x - x^*) = Ax - b \) (where \( x^* \) is the solution and \( x \) is the current approximation), and \( \beta \) is an arbitrary scalar parameter. He analyzed this class of updates to determine how, for quadratic functions, the choice of \( \beta \) affects convergence.\(^{54}\)

Broyden knew that when \( \beta = 0 \) in his class of updates (4.21), the DFP update was obtained. However, he wanted to find a value of \( \beta \) that would address the conditioning problem of the DFP method described in §4.2.1.1. To accomplish this, he first let \( B \) be the positive-definite matrix that satisfies the equation \( B^2 = A, \) and defined \( z = B e \) and

\(^{54}\)Broyden investigated how the successive errors depend upon the initial choice of the iteration matrix, in particular, if it was the identity or a good approximation to the inverse Hessian. He obtained error bounds on \( \|e_i\| \) for \( i = 1, 2, \ldots \) and was searching for a choice of \( \beta \) “with a view to the greatest possible reduction of the vector error norms.” To learn more on Broyden’s error analysis, see [14], [15].
\( K = BH B \), which he used to write

\[
t = \frac{z^T K z}{z^T K z^2}.
\]

(4.22)

He then showed that \( K \) is updated according to the equation

\[
K_+ = K + \begin{bmatrix} q & q^+ \end{bmatrix} \begin{bmatrix} \omega & \xi \\ \xi & \eta \end{bmatrix} \begin{bmatrix} q^T \\ q^+ \end{bmatrix}
\]

where

\[
\omega = 1 - q^T K q
\]
\[
\xi = -q^T K q^+
\]

with uniquely determined orthonormal vectors \( q \) and \( q^+ \), and arbitrary parameter \( \eta \) that depends on \( \beta \). To simplify notation, he denoted \( z^T K^j z \) by \( \theta_j \) for \( j = 1, 2, 3, \) and \( 4 \), and rewrote (4.22) as

\[
\theta_2 t = \theta_1.
\]

(4.23)

Next, he obtained \( \eta = k(\beta t^2 \theta_2 - 1) \) where \( k = \left( \frac{\theta_4}{\theta_3} - \frac{\theta_3}{\theta_2} \right) \) and showed that, in general, \( \eta \) increases with \( \beta \). Then, he set \( \eta \) equal to zero and explained that this follows from (4.23) and that in order to achieve this, \( \beta \) must be chosen to satisfy \( \beta t z^T K z = 1 \) which becomes

\[
\beta = -\frac{1}{t y^T},
\]

and is equivalent to \( \beta = \frac{1}{t y^T} \).\(^{55}\) Finally, by substituting this value of \( \beta \) into the update formula (4.21), Broyden obtained the BFGS inverse update (4.35) which he represented as

\[
H_+ = H + \frac{1}{y^T} (p p^T - p y^T H - H p y^T)
\]

(4.24)

\(^{55}\)Broyden proved that this choice of \( \beta \) yielded the only member of his class of updates for which a certain matrix error norm is reduced strictly monotonically when minimizing quadratic functions.
where

\[ \rho = t + \frac{y^T H y}{p^T y}. \]

### 4.2.2.3 Shanno

Also in 1970, Shanno [108] developed an update class of approximating matrices that he used to address the conditioning problem of the DFP method. Shanno considered the shifted inverse Hessian approximation matrix \( H \) given by

\[ \hat{H} = H + t \frac{ss^T}{s^T y} \]

where \( t \) is a scaling parameter. By taking the composition of (4.25) and the SR1 inverse update formula

\[ H_+ = H + \frac{(s - Hy)(s - Hy)^T}{(s - H y)^T y} \]

he obtained the following class of rank-two matrices

\[ H_+ = H + t \frac{ss^T}{s^T y} + \frac{[(1 - t)s - Hy][(1 - t)s - Hy]^T}{(1 - t)s - Hy]^T y} \]

where different values of \( t \) yield different update formulae. Observe that the choice \( t = 0 \) yields the SR1 update formula and the choice \( t = 1 \) yields the DFP update formula. Shanno showed that all of the eigenvalues of \( H_+ \) as defined by (4.26) increase monotonically with \( t \). In an effort to solve the conditioning problem by maximizing the smallest eigenvalue of \( H_+ \), he was able to find a closed-form representation of \( H_+ \) for \( t = \infty \). The following is a restatement of a theorem that shows how Shanno arrived at the BFGS inverse update formula. For the proof, see [108].
Theorem 4.2.1. (Shanno, 1970) Let $H_+$ be defined by (4.26). Then

$$
\lim_{t \to \infty} H_+ = H + (s - rHy)(s - rHy)^T \frac{(s - rHy)^Ty}{(s - rHy)^Ty} + (r - 1) \frac{Hy^THy}{y^THy} \tag{4.27}
$$

where

$$r = \frac{s^Ty}{s^Ty + y^THy}.$$

If we rewrite (4.27) as an update formula, it represents the BFGS inverse update formula. Shanno [108] acknowledged that “the same representation for $H_+$ has been derived by Goldfarb [59] from other considerations.” In the next section, we discuss how Goldfarb independently arrived at the BFGS update formula.

4.2.2.4 Goldfarb

In 1970, Greenstadt [62] presented an optimization problem whose solution yields a symmetric inverse update formula that minimizes the change in inverse Hessian approximations in the weighted Frobenius norm (4.31).\footnote{\textit{In 2000, Greenstadt [63] wrote an article on his recollections of the events surrounding the variable-metric events and dedicated it to Davidon.}} In an attempt to calculate the updated inverse Hessian approximation $H_+$ from current inverse Hessian approximation $H$ that limits the size of the difference $E = H_+ - H$, Greenstadt’s idea was to solve for the correction term $E$ that minimizes the norm

$$trace(WEWE^T) \tag{4.28}$$

subject to the conditions

$$E^T = E \tag{4.29}$$
and

$$Ey = s - Hy$$  

(4.30)

where $W$ is a positive-definite symmetric matrix.\(^{57}\)

The norm chosen by Greenstadt (4.28), is basically the weighted Frobenius norm

$$\|E\|_{F,W} = \|WEW\|_F$$  

(4.31)

for a symmetric nonsingular weighting matrix $W$. Condition (4.29) ensures that $H_+$ will be symmetric as long as the initial inverse Hessian approximation $H$ is chosen to be symmetric. Condition (4.30) ensures that $H_+$ satisfies the secant equation (4.4). All of this considered, we credit Greenstadt for being the first to pose the concept of least change. Accordingly, $H_+$ determined according to this criterion is a weighted least-change secant update of $H$. Greenstadt’s solution to his constrained minimization problem is given by the formula

$$E = \frac{1}{y^TW^{-1}y} \left[ sy^TW^{-1} + W^{-1}ys^T - Hyy^TW^{-1} - W^{-1}yy^TH - \frac{y^T(Hs - Hyy^TW^{-1}yy^TW^{-1})}{y^TW^{-1}y} \right].$$

Due to the freedom in the choice of $W$, different choices of $W$ yield different update formulae for $H_+$.

Greenstadt presented two choices for the weighting matrix $W$: $W = I$ and $W = H^{-1}$.

---

\(^{57}\)Greenstadt [63] explained that in the course of a discussion he had with Robert Mertz (a colleague at IBM), as they were wondering why the Davidon method was so effective, Mertz said something like “There ought to be a way of finding a ‘best’ DFP method.” Greenstadt interpreted ‘best’ to be the ‘smallest’ possible correction, to encourage stability. Greenstadt credits Mertz’ suggestion (that he look for the “best” $H$-correction) as what started him on the ‘least change’ path.
The choice $W = I$ yields the update formula

\[ H_+ = H + \frac{1}{y^T y} \left\{ sy^T + ys^T - Hyy^T - yy^T H - \frac{1}{y^T y} \left[ (y^T s) - (y^T H y) \right] yy^T \right\}. \]

The choice $W = H^{-1}$ yields what has become known as the Greenstadt update formula

\[ H_+ = H + \frac{1}{y^T H y} \left\{ sy^T H + Hys^T - \left[ 1 + \left( \frac{y^T s}{y^T H y} \right) \right] Hyy^T \right\} \]

(4.32)

and can be represented in direct form by choosing $v = Bs$ in the direct Dennis class (4.13).

It was Goldfarb [59], a referee of Greenstadt’s paper, who pointed out that another update (subsequently called the BFGS update) could be derived by using the weighting matrix $W = H_{+1}$ in Greenstadt’s formula (4.32).\(^{58}\) Goldfarb also showed the DFP update could be derived using the same process.

In private communication (March 31, 2006), Donald Goldfarb stated that the entire Broyden class of updates, including the rank-one update, could be derived using Greenstadt’s formulas. It is of interest to note that both Goldfarb’s and Greenstadt’s articles appeared in the same volume of the Mathematics of Computation journal in 1970.

### 4.2.3 Weighted Least Change Problem

Recall that for $f \in C^2$, the Hessian matrix of $f$ at $x$ is symmetric. Furthermore, if the Hessian is invertible, then it is also positive-definite at a local minimizer.\(^{59}\) All of this considered, it seems advantageous to maintain as much of this structure as possible and to

\(^{58}\)Greenstadt [63] explained that he suggested to Goldfarb, “who made so many brilliant suggestions” that “he write his own paper and explain them.”

\(^{59}\)This follows from the second-order necessary conditions. We know that if $x^*$ is a local minimizer of $f$ and $\nabla^2 f(x)$ is continuous in an open neighborhood of $x^*$, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semi-definite. In the case that the Hessian is invertible, then we know $\nabla^2 f(x^*)$ must be positive-definite.
use a least change secant update that incorporates these special properties into the approximations. As we saw in the previous section, Greenstadt was not only the first to pose the least change strategy but was the first to derive a weighted least change update as the solution to such a constrained optimization problem. So although both the DFP and the BFGS updates are symmetric weighted least change secant updates, only one construction of the BFGS update was derived this way.

As we mentioned in §4.1.6, Dennis and Moré [36], in 1977, showed that the good Broyden, the bad Broyden and the PSB updates were least change updates. They also presented the following theorems describing how members of the Dennis class (4.13), for particular choices of scale $v$ and weighting matrix $W$, are weighted least change secant updates.

**Theorem 4.2.2.** Let $B, W \in \mathbb{R}^{n \times n}$ be symmetric, $W$ nonsingular and let $s, y \in \mathbb{R}^n$ be such that $s \neq 0$ and $y \neq Bs$. Then the unique solution to

$$
\min_{B_+} \|B_+ - B\|_{F,W} \quad \text{subject to} \quad B_+s = y, \text{ and } B_+ \text{ symmetric} \quad (4.33)
$$

is the member of the Dennis direct update class (4.13) with $v = W^{-2} s$.

If $W = I$, then the solution to the optimization problem (4.33) is the PSB update which is obtained by setting $v = s$ in the Dennis class (4.13).

If $B$ is positive definite and $W = B^{-\frac{1}{2}}$, then the solution to (4.33) is the Greenstadt update which is obtained by setting $v = Bs$ in the Dennis class (4.13).

If $B$ is positive definite, $W = B_+^{-\frac{1}{2}}$ and $s^T y > 0$, then the unique solution to (4.33) is the DFP update which is positive definite and is obtained by setting $v = y$ in the Dennis class.
Theorem 4.2.3. Let $H, W \in \mathbb{R}^{n \times n}$ be symmetric, $W$ nonsingular, and let $s, y \in \mathbb{R}^n$ be such that $s \neq 0$ and $s \neq Hy$. Then the unique solution to

$$
\min_{H_+} \|H_+ - H\|_{F,W} \quad \text{subject to} \quad H_+ y = s, \text{ and } H_+ \text{ symmetric}
$$

(4.34)

is the member of the Dennis inverse update class (4.12) with $d = W^{-2}y$.

If $H$ is positive definite, $W = H_+^{-\frac{1}{2}}$ and $s^T y > 0$, then the unique solution to (4.34) is the BFGS update which is positive definite and is obtained by setting $d = s$ in the Dennis inverse update class (4.12).

4.2.4 Limited Memory BFGS (L-BFGS)

Motivated by the desire to reduce storage and possibly improve the behavior of conjugate gradient (CG) methods, several algorithms that combine CG steps and quasi-Newton updates were proposed. These methods originated with the work of Perry [95] and Shanno [110], and were subsequently developed and analyzed by Buckley [17], [18] and Nocedal [84], [85], among others. We choose to describe Nocedal’s contributions as it was a method that Nocedal and Liu presented in 1989 which has become quite popular.60

In 1978, Nocedal, in his Ph.D. thesis [84], proposed methods that were similar to Buckley’s. These methods utilized a form of the CG search direction for most iterations but

---

60In the same year, Gilbert and Lemaréchal [55] presented an almost identical implementation of the L-BFGS method. Nocedal and Liu explain that while their work for their paper was in progress in 1988, they became aware that Gilbert and Lemaréchal were performing experiments similar to theirs. Gilbert and Lemaréchal, in their paper, acknowledge the earlier works of Nocedal [85], as well as the work of Nocedal and Liu [87] and others.
periodically used the BFGS method. In private communication (November, 9, 2008), Jorge Nocedal explained that it was “after finishing my thesis that I realized that we needed to get CG out of the way – it was just complicating things.” Consequently, in 1980, Nocedal [85] was the first to introduce a method for reducing storage requirements using strictly the BFGS method. This method, which he called SQN (special quasi-Newton), was an adaptation of the BFGS method to large problems that reduced the amount of storage required by the original BFGS method. In 1989, Nocedal and Liu [87] presented the final form of this method which is referred to as the limited memory BFGS (L-BFGS) method and which others have referred to as LMQN (limited memory quasi-Newton).

In our discussion of the L-BFGS method, we use the product form (4.20) of the BFGS method which can be written as

\[ H_+ = V^T H V + \rho s s^T \]  

(4.35)

where \( V = I - \rho y s^T \) and \( \rho = \frac{1}{y^T s} \). Observe that \( H_+ \) is obtained by updating \( H \) using what is often referred to as the correction pair \((s, y)\).

At the start of the L-BFGS method, the user provides an initial point \( x_0 \) and a sparse symmetric positive-definite matrix to use as the initial inverse Hessian approximation \( H_0 \). In addition, the user specifies a number \( \ell \) which can represent the number of correction pairs that are to be kept. During the first \( \ell \) iterations, the updates are generated using (4.35). That is, at the first step, we have

\[ H_1 = V_0^T H_0 V_0 + \rho_0 s_0 s_0^T. \]
At the second step, we have

\[
H_2 = V_1^T H_1 V_1 + \rho_1 s_1 s_1^T \quad (4.36)
\]

\[
= V_1^T [V_0^T H_0 V_0 + \rho_0 s_0 s_0^T] V_1 + \rho_0 s_0 s_0^T
\]

\[
= V_1^T V_0^T H_0 V_0 V_1 + \rho_0 V_1^T s_0 s_0^T V_1 + \rho_0 s_0 s_0^T.
\]

Notice that instead of storing the full current inverse Hessian approximation, one stores the initial matrix \(H_0\) and the set of correction pairs separately which define \(H_+\) implicitly.

For iterations greater than \(\ell\), instead of keeping all of the \(s\) and \(y\) from the past iterations, one stores only the past \(\ell\) pairs of \((s, y)\) and after each iteration, the set of pairs is refreshed, that is, the oldest pair is deleted and the newly generated pair is added. Hence, the \(\ell\) most recent correction pairs are always kept and the new inverse Hessian approximation is obtained using information from the \(\ell\) previous iterations. Deleting a correction pair is equivalent to setting the oldest \(V = I\) and the oldest \(\rho s s^T = 0\). For example, in our second step (4.36), we would set \(V_0 = I\) and \(\rho_0 s_0 s_0^T = 0\) to obtain

\[
H_2 = V_1^T H_0 V_1 + \rho_1 s_1 s_1^T.
\]

Therefore, after the \(\ell\)-th iteration, we have

\[
H_{k+1} = (V_k^T \ldots V_{k-\ell}^T)H_0(V_{k-\ell} \ldots V_k)
\]

\[
+ \rho_{k-\ell}(V_k^T \ldots V_{k-\ell}^T)s_{k-\ell}s_{k-\ell}^T(V_{k-\ell} \ldots V_{\ell})
\]

\[
\vdots
\]

\[
+ \rho_k s_k s_k^T.
\]
The L-BFGS method produces positive-definite matrices $H_+$ that are defined implicitly. Furthermore, for strictly convex quadratic functions, the secant equation is satisfied in the past $\ell$ directions.

### 4.2.5 Broyden Class of Secant Updates

In §4.2.2.2, we mentioned how Broyden [14], in 1967, presented the one-parameter family of rank-two secant updates that is referred to as the Broyden class. This class of formulae, which was originally presented in the somewhat different form (4.21), is more commonly written as

$$
B_+ = B - \frac{B_{ss^TB}}{s^TBs} + \frac{yy^Ty}{y^Ts} + \theta(s^TBs) \left[ \frac{y}{y^Ts} - \frac{Bs}{s^TBs} \right] \left[ \frac{y}{y^Ts} - \frac{Bs}{s^TBs} \right]^T
$$

(4.37)

with scalar parameter $\theta$. The last term in the Broyden class of direct updates (4.37) is a rank-one correction which decreases the eigenvalues of $B_+$ when $\theta$ is negative [61]. However, as $\theta$ decreases, $B_+$ eventually becomes singular and then indefinite. The degenerate value of $\theta$ that causes $B_+$ to be singular is

$$
\theta^C = \frac{1}{1 - \mu}
$$

(4.38)

where

$$
\mu = \frac{(y^TB^{-1}y)(s^TBs)}{(y^Ts)^2}.
$$

For symmetric, positive-definite $B$, all direct updates that are members of the Broyden class (4.37) are symmetric positive-definite if $\theta > \theta^C$, and if $s^Ty > 0$. The following is
a restatement of a theorem Dixon [41] proved which describes a property of the Broyden class.

**Theorem 4.2.4.** (Dixon, 1972) Let \( f \) be a differentiable function to be minimized. Consider any two members of the Broyden class (4.37) and assume each begins with the same initial point \( x_0 \) and the same initial inverse Hessian approximation \( H_0 \). Then, the successive iterates generated by these two members of the Broyden class are identical provided an exact line search (1.8) is used and the degenerate \( \theta \) value, \( \theta^C \) (4.38), is never used.

This is quite a surprising and amazing result and it also serves as the underlying theory for proving a uniqueness property of the BFGS that we discuss in Chapter 6. The following are well-known choices of the parameter \( \theta \) in the Broyden class:

\[
\text{DFP } \theta = 1 \\
\text{BFGS } \theta = 0 \\
\text{Convex Class } \theta \in [0, 1] \\
\text{Preconvex Class } \theta \in (\theta^C, 0).
\]

The Broyden convex class (4.39), also known as the *restricted class*, consists of updates in the Broyden class (4.37) obtained by restricting \( \theta \) to \([0, 1]\). Accordingly, both the DFP and the BFGS updates are members of this subset of the Broyden class of updates; the convex class.

In 1970, Fletcher [45] noticed that other formulae could be generated by taking a linear combination of the DFP inverse update \( H^{\text{DFP}}_+ \) (4.17) and the BFGS inverse update \( H^{\text{BFGS}}_+ \)
(4.20) of the following form

$$H_+ = (1 - \theta)H_+^{\text{DFP}} + \theta H_+^{\text{BFGS}}$$  \hspace{1cm} (4.41)

for any $\theta$. Fletcher explained that this class of formulae is related directly to the Broyden class (4.21) based on the parameter $\beta$, through the relationship $\theta = \beta s^T y$. Additionally, he pointed out the new result that (4.41) can be rearranged as

$$H_+ = H_+^{\text{DFP}} + \theta \nu \nu^T$$  \hspace{1cm} (4.42)

where

$$\nu = \sqrt{y^T H y} \left[ \frac{s}{s^T y} - \frac{H y}{y^T H y} \right].$$

It follows directly from the duality of the BFGS and DFP methods and (4.42), that we can rewrite the Broyden class (4.37) as

$$B_+ = B_+^{\text{BFGS}} + \theta \nu \nu^T$$  \hspace{1cm} (4.43)

where

$$\nu = \sqrt{s^T B s} \left[ \frac{y}{y^T s} - \frac{B s}{s^T B s} \right].$$

Fletcher investigated what he calls the “convex class of formulae” which, as he explains, is a result of taking $\theta$ as a convex combination of 0 and 1, i.e., taking (4.41) with $\theta \in [0, 1]$. Since any formula obtained from Fletcher’s “convex class of formulae” is a member of what eventually became known as the Broyden convex class, we credit Fletcher with the identification of the Broyden convex class.
Zhang and Teawarson were aware that the BFGS method was considered the most effective update method from the Broyden class. In 1988, they [131] investigated updates in the Broyden class (4.37) that use negative values of $\theta$ to determine if there was an update method more effective than the BFGS method. They referred to the set of updates obtained by restricting $\theta$ to $(\theta^C, 0)$ as the preconvex class (4.40), thus we acknowledge Zhang and Tewarson for being the first to define the preconvex class. They suggested that updating formulae should not be confined only to the convex class because the preconvex class may contain more efficient methods in terms of function evaluations. Zhang and Tewarson studied methods that used a fixed negative value of $\theta$ as well as methods that used a varying negative value of $\theta$ that changes at each iteration and claimed that it seemed plausible that the “best” formula in Broyden’s class should have a varying parameter. A number of formulae with varying parameters have been proposed. We mention examples of such formulae in §6.5.3 and we direct the interested reader to Zhang and Tewarson [131] for further detail.

4.3 Equality Constrained Optimization

Consider the equality constrained minimization problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

subject to $g(x) = 0$  

(4.44)
where we assume $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ for $(m \leq n)$ are twice continuously differentiable even though this is not always required. The Lagrangian function associated with problem (4.44) is the function

$$l(x, \lambda) = f(x) + \lambda^T g(x), \quad (4.45)$$

where $\lambda \in \mathbb{R}^m$ is called the vector of Lagrange multipliers or simply the Lagrange multiplier. In an attempt to solve the first-order necessary conditions ($\nabla l(x^*, \lambda^*) = 0$), we consider the following secant method.

### 4.3.1 Successive Quadratic Programming (SQP)

By a successive quadratic programming (SQP) quasi-Newton method for the constrained optimization problem (4.44), we mean the iterative process

$$
x_+ = x + s \\
\lambda_+ = \lambda + \Delta \lambda \\
B_{l+} = B(x, s, \lambda_+, B_l),
$$

(4.46)

where $B$ is an update function and $s$ and $\Delta \lambda$ are respectively the solution and the multiplier vector associated with the solution of the quadratic programming problem

$$
\min_x \nabla_x l(x, \lambda)^T s + \frac{1}{2} s^T B_l s \\
\text{subject to} \quad \nabla g(x)^T s + g(x) = 0.
$$

(4.47)

In (4.47), $\nabla_x l(x, \lambda)$ is the gradient (with respect to $x$) of the Lagrangian (4.45) evaluated at the current iterate $(x, \lambda)$, and $B_l$ is intended to be an approximation to $\nabla^2_l l(x, \lambda)$, the
Hessian of the Lagrangian at \((x, \lambda)\). We denote by \(x^*\), a solution of the constrained minimization problem (4.44) with associated multiplier vector \(\lambda^*\) satisfying \(\nabla_x l(x^*, \lambda^*) = 0\).

We call (4.46)-(4.47) an \textit{SQP Lagrangian secant method} if the Lagrangian update of \(B^l\) satisfies the \textit{Lagrangian secant equation}

\[
B^l_+ s = y_l, \quad (4.48)
\]

where

\[
y_l = \nabla_x l(x_+, \lambda_+) - \nabla_x l(x, \lambda_+) \quad (4.49)
\]

and \(x, x_+, s, \) and \(\lambda_+\) are as in (4.46)-(4.47) [117]. Secant update formulas from unconstrained optimization, for example, those obtained from the Dennis class (4.13) can be used with the SQP secant method (4.46)-(4.47). Accordingly, the extension of the BFGS secant method (4.19) can be made from unconstrained optimization (problem (4.16)) to constrained optimization (problem (4.44)) by employing the SQP framework. The deficiency of the SQP Lagrangian BFGS method is that \(\nabla^2_x l(x^*, \lambda^*)\) is not guaranteed to be positive-definite under the standard assumptions:

\[(i) \quad f \text{ and } g \text{ have second derivatives which are Lipschitz continuous} \quad (4.50)\]

in an open neighborhood \(D\) of the local solution \(x^*\)

\[(ii) \quad \nabla^2 l(x^*, \lambda^*) \text{ is nonsingular.} \quad (4.51)\]
It is well known that the second assumption (4.51) is equivalent to the two assumptions:

(iii) $\nabla g(x^*)$ has full rank

(iv) $\nabla^2 l(x^*, \lambda^*)$ is positive-definite on $S(x^*) = \{ \eta : \nabla g(x^*)^T \eta = 0, \eta \neq 0 \}$.

In unconstrained optimization, positive-definiteness is a reasonable assumption, however, in constrained optimization the Hessian remains symmetric, but in general, it will not be positive-definite. In the upcoming sections, we discuss alternative formulations of the SQP Lagrangian BFGS secant method that circumvent this lack of positive-definiteness.

### 4.3.2 Powell’s Damped BFGS Algorithm (PDA)

In 1978, Powell [99] proposed a modification to the SQP Lagrangian BFGS secant method that compensates for the lack of positive-definiteness in the Hessian of the Lagrangian at the solution. Despite the fact that the true Hessian of the Lagrangian may not be positive definite at a solution, Powell chose to maintain a positive-definite matrix by modifying $y_l$ (4.49) whenever necessary and instead used

$$ y_l^P = \theta y_l + (1 - \theta) B_l s, $$

where

$$ \theta = \begin{cases} 
1 & \text{if } y_l^T s \geq \epsilon s^T B_l s, \\
\frac{(1-\epsilon)s^T B_l s}{s^T B_l s - y_l^T s} & \text{otherwise}
\end{cases} $$

and $\epsilon$ is a small positive constant, e.g., $\epsilon = 0.2$. The update $B_l^+$ is then obtained as a BFGS secant update using $y_l^P$ instead of $y_l$. This modification guarantees that $y_l^P s > 0$,
and consequently, allows the BFGS secant update to maintain positive-definiteness (even far from the solution) but at the expense of the Lagrangian secant equation (4.48). Powell’s modified SQP Lagrangian BFGS secant method, known as Powell’s damped BFGS algorithm (PDA), works reasonably well, but theory has not been developed for it and it is quite doubtful that it retains superlinear convergence.

### 4.3.3 SQP Augmented Lagrangian BFGS Secant Method

Another alternative formulation of the SQP Lagrangian BFGS secant method considered was to replace the Lagrangian (4.45) with the augmented Lagrangian function

\[ L(x, \lambda \rho) = l(x) + \frac{\rho}{2} g(x)^T g(x), \quad \rho \geq 0, \]

associated with problem (4.44) where \( \rho \) is an augmentation parameter (see Han [66] and Tapia [116]). Then, \( y_t \) (4.49) is replaced with

\[ y_L = \nabla_x L(x_+, \lambda_+, \rho) - \nabla_x L(x, \lambda_+, \rho). \quad (4.52) \]

A fundamental issue in using the augmented Lagrangian in a secant algorithm is the choice of the augmentation parameter \( \rho \). It is well-known that for any augmentation parameter \( \rho \) greater than a threshold value \( \bar{\rho} \), the Hessian of the augmented Lagrangian at a local solution of problem (4.44), under the standard assumptions (4.50)-(4.51), is positive definite and we can guarantee that near the solution \( y_L^T s > 0 \) for \( \rho \) sufficiently large [24].

Though theoretically attractive, this SQP augmented Lagrangian BFGS secant method has some disadvantages. First, a priori knowledge of the threshold value \( \bar{\rho} \) for a given
problem is generally unavailable. Second, the attempt to use large $\rho$ seems to present numerical problems [24] (see the examples given by Tapia [116] and Nocedal and Overton [88]). In addition, $y_L$ given by (4.52) has the disadvantage that at some iterations it may not be possible to choose $\rho$ sufficiently large so that $y_L^T s$ is positive (even though it must be near the solution).

4.3.4 Tapia’s BFGS Structured Augmented Lagrangian Secant Algorithm (SALSA)

In 1988, Tapia [117] proposed an algorithm for nonlinear equality constrained optimization which circumvents the lack of positive-definiteness in the Hessian of the Lagrangian. He derived the Structured Augmented Lagrangian Secant Algorithm, later known as SALSA, by considering SQP augmented Lagrangian secant methods and taking advantage of the structure present in the Hessian of the augmented Lagrangian function for problem (4.44) as it displays significant structure in that there is a clear separation between first- and second-order information. That is, Tapia introduced structure in $y_L$ but not in $B_L$ and then used a new method to choose the Lagrangian augmentation parameter $\rho$ that does not require prior knowledge of the true Hessian.

In SALSA, $y_L$ (4.49) is replaced with

$$y_S = y_L + \rho A^T s$$

where $A$ is the matrix whose columns are $\nabla g_1, \nabla g_2, \ldots, \nabla g_m$, and instead of sacrificing the Lagrangian secant equation (4.48), the approximate Hessian of the structured augmented
Lagrangian update of $B^L$ satisfies the \textit{structured augmented Lagrangian secant equation}

\[ B^L_+ s = y_S. \]

For $\rho$ large enough, the local positivity of $y_S^T s$ is guaranteed and consequently, $B^L$ remains positive definite. Even globally, $y_S^T s$ can be made positive by increasing $\rho$ as long as $A^T_+ s \neq 0$. Tapia [117] demonstrated that the BFGS version of SALSA is locally superlinearly convergent.

In 1988, Byrd, Tapia and Zhang [23] introduced a new reliable method for choosing the Lagrangian augmentation parameter in Tapia’s BFGS SALSA that does not require prior knowledge of the true Hessian. They performed considerable numerical experiments with SALSA and compared it to PDA. To learn more about their strategy for choosing $\rho$, and a corresponding back-up strategy, see their papers [23] [24]. It is important to mention that while SALSA for equality constrained optimization has an attractive theory, SALSA has its shortcomings. For example, numerical experimentation shows that one of the technical assumptions holds only in a local manner.
Chapter 5

Convergence Theory

In this chapter, we collect convergence theory. While this presentation is far from exhaustive, we give an overview of some of the important properties and results that pertain to basic secant methods with a focus on the BFGS secant method. We discuss the well-known Dennis-Moré characterization of superlinear convergence and detail some convergence results by Powell [111], Broyden, Dennis and Moré [16], Byrd, Nocedal and Yuan [22], Zhang and Tewarson [131] and others. Many of the topics outlined in this chapter are revisited in later chapters. To avoid excessive redundancy, we refer the reader to the corresponding later sections for further detail.

5.1 Conjugacy

We begin the chapter by discussing the concepts of orthogonality and conjugacy. As we will see in upcoming sections, these concepts are particularly important in some algorithms
for unconstrained optimization. Recall that two vectors $a$ and $b$, are said to be orthogonal, or conjugate, if $a^Tb = 0$.

**Definition 5.1.1.** Let $H$ be an $n \times n$ symmetric, positive-definite matrix. The vectors $a$ and $b$ are called $H$-conjugate, or simply conjugate if $H$ is understood, if $a^T H b = 0$.

Consider any symmetric positive-definite matrix $H$. We can interpret conjugacy to be orthogonality with respect to an $H$-weighted inner product, i.e., $\langle a, b \rangle_H = a^T H b$. Orthogonality, $a^T b = 0$, does not imply $H$-conjugacy, in general, unless $H = I$.

### 5.2 Convexity

We take a moment to distinguish among the different types of convexity as it aids in the understanding of the discussions in this chapter and there is some confusion in the literature. The following definition is from Ortega and Rheinboldt [91]. Note, in this chapter, $\| \cdot \|$ will denote the $l_2$ vector norm or its induced operator norm.

**Definition 5.2.1.** A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex on an open convex set $D$, if for all $x, y \in D$ and $0 < \delta < 1$

$$f(\delta x + (1 - \delta)y) \leq \delta f(x) + (1 - \delta)f(y).$$

(5.1)

The function $f$ is strictly convex on $D$, if strict inequality holds in (5.1) whenever $x \neq y$.

The function $f$ is uniformly convex on $D$ if there is a constant $c > 0$ such that, for all
$x, y \in D \text{ and } 0 < \delta < 1$

\[ \delta f(x) + (1 - \delta)f(y) - f(\delta x + (1 - \delta)y) \geq c\delta(1 - \delta)\|x - y\|^2. \]

It is clear that uniform convexity implies strict convexity, which in turn, implies convexity. In addition, due to the equivalence of norms in $\mathbb{R}^n$, if a function is uniformly convex in one norm, it is uniformly convex in all norms. A function uniformly convex on $\mathbb{R}^n$ always has a unique minimizer. This is not the case for convexity or strict convexity.

Some authors refer to uniformly convex as strongly convex. A quadratic function in $\mathbb{R}^n$ is convex if and only if the Hessian matrix is positive semi-definite and strictly convex if and only if the Hessian matrix is positive definite. Moreover, for a quadratic function, strict convexity and uniform convexity are equivalent. To see this, one need only consider the Rayleigh quotient.

### 5.3 Finite Termination

A given algorithm is said to satisfy the finite termination property if it finds the minimizer of a quadratic function, if it exists, in a finite number of steps.\footnote{In the literature, the term ‘quadratic termination’ is often used to refer to an algorithm that finds the exact minimizer of a strictly convex quadratic function after a finite number of steps. One example of this usage can be found in Broyden, Dennis and Moré [16] from 1973.} A particular way to obtain finite termination is to invoke the concept of conjugate directions. In a neighborhood of a strict local minimizer, we expect that a strictly quadratic function approximates the function reasonably well. This coupled with the fact that we can minimize a strictly convex quadratic function in at most $n$ steps, if we search along conjugate directions of the Hessian matrix,
makes the notion of conjugacy very useful for optimizing both quadratic and nonquadratic functions.

In 1967, Broyden [13] was the first to point out that the DFP formula is only one of a family of formulae (the Broyden class) that achieve finite termination. He showed that every member of the Broyden class \((\theta \geq 0)\) terminates after at most \(n\) iterations with the exact minimizer for any strictly convex quadratic function, given an exact line search is used.

**Remark 5.3.1.** In 1976, Moré and Trangenstein [81] showed that for the affine function \(f(x) = Ax - b\), with nonsingular \(A \in \mathbb{R}^{n \times n}\) and \(b \in \mathbb{R}^n\), Broyden’s method, in a slightly modified form, is globally and superlinearly convergent to the unique solution.

In 1979, Gay [53] proved the following convergence result for Broyden’s method when applied to the same linear function.

**Theorem 5.3.2.** (Gay, 1979) If \(f(x) = Ax - b\) for \(b \in \mathbb{R}^n\) and nonsingular \(A \in \mathbb{R}^{n \times n}\), then the Broyden method given by (4.6), with steplength \(\alpha = 1\), converges in at most \(2n\) steps.

### 5.4 Bounded Deterioration

In 1971, Dennis’ [32] study of error bounds for the Jacobian in Broyden’s method led him to discover the bounded deterioration property. The following definition of bounded deterioration can be found in [33].
**Definition 5.4.1.** (Dennis, 1972) Any secant method of the form

\[ x_{k+1} = x_k - B_k^{-1}F(x_k) \]  \hspace{1cm} (5.2)

that satisfies

\[ \|B_k - F'(x_k)\| \leq \delta + \gamma \sum_{j=1}^{k} \|x_{j-1} - x_j\| \]

for \( k = 1, 2, \ldots \) and for some \( \delta, \gamma > 0 \), is said to be of bounded deterioration.

We mention the bounded deterioration property because it served as the fundamental idea behind the convergence theory that Broyden, Dennis, and Moré [16] and Dennis and Moré [35], [36] developed soon after bounded deterioration was introduced. In the next section, we discuss this convergence theory.

### 5.5 Superlinear Convergence

The original technique for proving that a quasi-Newton method achieves superlinear convergence was to show that if \( \{x_k\} \) converges to \( x^* \), then \( \{B_k\} \) converges to \( F'(x^*) \). It is well-known that this consistency condition is sufficient but as we will explain shortly, is not necessary for superlinear convergence. Until the early 1970s, all of the practical methods for which there were published proofs of superlinearity satisfied the property of consistency; it seemed to be effectively necessary. In the upcoming sections, we present what we believe are among the most important characterizations of superlinear convergence.
5.5.1 Broyden, Dennis and Moré Characterizations

In 1973, Broyden, Dennis, and Moré [16] presented a local convergence analysis for several well-known secant methods when used without line searches. Despite the fact that it is known that the sequence of approximate Jacobians \( \{B_k\} \) does not necessarily converge to the true Jacobian at the solution, the methods they considered generate superlinearly convergent sequences \( \{x_k\} \). Before proceeding to the characterization of local and superlinear convergence, we restate the following theorem from which they obtained convergence results.

**Theorem 5.5.1.** (Broyden, Dennis and Moré, 1973) Let \( F : \mathbb{R}^n \rightarrow \mathbb{R}^n \) satisfy the following assumptions:

(a) There exists \( x^* \) such that \( F(x^*) = 0 \).

(b) \( F \) is continuously differentiable in an open convex set \( D \) containing \( x^* \).

(c) \( F'(x^*) \) is nonsingular.

Suppose that for some \( D \) containing \( x^* \), the Lipschitz condition

\[
\| F'(x) - F'(x^*) \| \leq \gamma \| x - x^* \| \quad \text{for} \quad \gamma \geq 0, \quad \text{and} \quad \forall x \in D,
\]  

(5.3)

holds. Consider any secant method obtained from the general rank-one update class (4.3), from the Dennis direct update class (4.13) with symmetric \( B \), or from the Dennis inverse update class (4.12) with symmetric \( H \). Assume this secant method uses steplength \( \alpha = 1 \). Let \( N' = N_1' \times N_2' \) where \( N_1' \subset D \) and \( N_2' \) only contains nonsingular matrices so that
$s = -B^{-1}F(x)$ is well-defined. Let the vector $v$ in the secant method obtained from (4.3) or (4.13) be chosen so that for all $(x, B)$ in a neighborhood of $N'$ of $(x^*, F'(x^*))$,

$$
\frac{\|Mv - M^{-1}s\|}{\|M^{-1}s\|} \leq \mu_1 \|s\|^p, \quad s \neq 0
$$

(5.4)

for some constant $\mu_1 \geq 0$, and some nonsingular, symmetric matrix $M \in \mathbb{R}^{n \times n}$ both independent of $v$ and $s$. Then this method is locally, and superlinearly convergent at $x^*$.

Let the vector $d$ in the secant method obtained from (4.12) be chosen so that for all $(x, H)$ in a neighborhood of $N'$ of $(x^*, F'(x^*))$,

$$
\frac{\|Md - M^{-1}y\|}{\|M^{-1}y\|} \leq \mu_2 \|y\|^p, \quad y \neq 0
$$

(5.5)

for some constant $\mu_2 \geq 0$, and some nonsingular, symmetric matrix $M \in \mathbb{R}^{n \times n}$ both independent of $d$ and $s$. Then this method is locally, and superlinearly convergent at $x^*$.

### 5.5.2 Broyden, Dennis, and Moré Results

The convergence results in the following theorem are consequences of Theorem 5.5.1.

**Theorem 5.5.2.** (Broyden, Dennis, and Moré, 1973) Let $F : \mathbb{R}^n \to \mathbb{R}^n$ satisfy the assumptions of Theorem 5.5.1. Consider the inequalities (5.4) and (5.5) where $p = 1$.

The Broyden update (4.6) satisfies inequality (5.4) with $v = s$, $M = I$ and where $\mu_1 = 0$.

The PSB update (4.11) satisfies inequality (5.4) with $v = s$, $M = I$ and where $\mu_1 = 0$.

The Greenstadt update (4.32) satisfies inequality (5.5) with $d = y$, $M = I$ where $\mu_2 = 0$.

Suppose, in addition, $F'(x^*)$ is positive-definite.

The DFP direct update (4.18) satisfies inequality (5.4) with $v = y$, where $M^{-1}$ is the
positive-definite square root of the Hessian of $f$ at $x^*$ and a suitable value of $\mu_1$.

The BFGS inverse update (4.20) satisfies inequality (5.5) with $d = s$ where $M$ is the positive-definite square root of the Hessian of $f$ at $x^*$ and a suitable value of $\mu_2$.

Thus, the Broyden, the PSB, the Greenstadt, the DFP direct, and the BFGS inverse updates are locally and superlinearly convergent at $x^*$.

5.5.3 Dennis and Moré Characterizations

In 1970, Ortega and Rheinboldt [91], in their study of general rates of convergence, presented the following definition of superlinear convergence.

**Definition 5.5.3.** If $\{x_k\} \subset \mathbb{R}^n$ converges to $x^*$, then $\{x_k\}$ converges superlinearly to $x^*$ if either $x_k = x^*$ for all sufficiently large $k$, or $x_k \neq x^*$ for $k \geq k_0$ and

$$
\lim_{k \to +\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0.
$$

**Remark 5.5.4.** Since in finite dimensions norms are equivalent, the notion of superlinear convergences is norm independent.

In 1974, Dennis and Moré [35] presented the following well-known characterization of superlinear convergence. For a proof, see Dennis and Moré [35].

**Theorem 5.5.5.** *(Dennis and Moré, 1974)* Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be differentiable in the open convex set $D \subset \mathbb{R}^n$, and assume that for some $x^* \in D$, $F'$ is continuous at $x^*$ and $F'(x^*)$ is nonsingular. Let $\{B_k\} \subset \mathbb{R}^{n \times n}$ be a sequence of nonsingular matrices. Suppose that for
some $x_0 \in D$, the sequence (5.2) for $k = 0, 1, \ldots$ remains in $D$ and converges to $x^*$. Then
\{x_k\} converges superlinearly to $x^*$ if and only if
\[
\lim_{k \to +\infty} \frac{\|B_k - F'(x^*)(x_{k+1} - x_k)\|}{\|x_{k+1} - x_k\|} = 0.
\] (5.6)
In the literature, (5.6) is often referred to as the Dennis and Moré Condition.

When Dennis and Moré [35] state that the sequence $\{x_k\}$ generated by the secant iteration (5.2) is well-defined and converges to $x^*$, it is an implicit assumption that $x_k \neq x^*$ for all sufficiently large $k$.

Although the characterization described in Theorem 5.5.5 was given for secant methods for the nonlinear equation problem, clearly the characterization immediately carries over to unconstrained optimization by working with the nonlinear equation (gradient equal to zero) that results from the first-order necessary conditions. Theorem 5.5.5 implies that consistency is not a necessary condition for superlinearity. However, consistency readily implies the Dennis and Moré condition (5.6).

We end this section by highlighting two additional results given by Dennis and Moré [35]. For their proofs, see [35]. The first result we present is an interesting by-product of superlinear convergence. If $\{x_k\} \subset \mathbb{R}^n$ converges superlinearly to $x^*$, then
\[
\lim_{k \to +\infty} \frac{\|x_{k+1} - x_k\|}{\|x_k - x^*\|} = 1.
\]
This justifies the very commonly used computational technique of estimating $\|x_k - x^*\|$ with $\|x_{k+1} - x_k\|$, which is appropriate if we have superlinear convergence.

The next result of Dennis and Moré [35] has to do with the iteration $x_{k+1} = x_k + \alpha_k p_k$,
where $p_k$ is a search direction and $\alpha_k$ is a steplength obtained by an exact line search. They
state that this iteration is superlinearly convergent if and only if

\[ a_k p_k = p_k^N + o(\|p_k^N\|) \]

where \( p_k^N \) is the Newton step. This condition states that a method is superlinearly convergent if and only if the step it produces approximates that of Newton’s method asymptotically.

### 5.5.4 Dennis and Moré Results

In this section, we highlight two convergence results for the DFP and the BFGS methods that Dennis and Moré [36] proved in 1977. The first is an interesting result that makes no assumptions on the Hessian approximations. It states that if the iterates generated by the BFGS or the DFP methods satisfy

\[ \sum_{k=0}^{\infty} \|x_k - x^*\| < \infty, \]

then the rate of convergence is superlinear.\(^{62}\)

The following is a restatement of a theorem regarding a global convergent result for the DFP and the BFGS methods.

**Theorem 5.5.6.** (Dennis and Moré, 1977) Let \( f : \mathbb{R}^n \to \mathbb{R} \) be the strictly convex quadratic function given by \( f(x) = \frac{1}{2} x^T A x - x^T b + c \) for some symmetric, positive-definite \( A \in \mathbb{R}^{n \times n} \), \( b \in \mathbb{R}^n \) and \( c \in \mathbb{R} \). Then for steplength \( \alpha = 1 \), the DFP and the BFGS methods

---

\(^{62}\)These results have been extended to the Broyden convex class by Stachurski [114] in 1981, by Griewank and Toint [64] in 1982 and others.
converge globally and superlinearly to $A^{-1}b$ assuming that the approximating matrices 
\{B_k\} for the DFP method, or \{H_k\} for the BFGS method, are well-defined, positive definite 
and that the corresponding \{\|H_k\|\}, or \{\|B_k\|\} is bounded.\(^\text{63}\)

In the next section, we state another result for the BFGS method that was proven later by 
Byrd, Tapia and Zhang.

\subsection*{5.5.4.1 Byrd, Tapia, Zhang Result}

In 1989, Byrd, Tapia and Zhang [23] proved the following theorem.

\textbf{Theorem 5.5.7.} (Byrd, Tapia, and Zhang, 1989) If $x^*$ is a local minimizer of the function 
$f(x)$ such that $\nabla^2 f(x^*)$ is nonsingular and the sequence \{x_k\} generated by the BFGS 
method with steplength $\alpha = 1$ converges to $x^*$, then the convergence is superlinear.

This result is utilized in Chapter 7.

\subsection*{5.5.5 Powell’s Characterization}

In 1969, Pearson [93] expressed the determinant of an inverse Hessian approximation up-
date $H_+$ for the DFP update in closed-form as

$$
det(H_+) = det(H) \frac{y^T s}{y^T H y}.
$$

In 1971, Powell [97] presented an expression for the trace of $H_+$ for the DFP update which 
can equivalently be written for $B_+$ in the BFGS update using (4.19). He expressed the trace

\(^{63}\)To prove that \{\|B_k\|\} for the DFP method and \{\|H_k\|\} for the BFGS method are bounded, the tech-
niques of Goldfarb found in Fletcher’s 1969 book [44] or the 1971 techniques of Powell [97] could be utilized.
of $B_+$ in terms of the trace of $B$ as

$$\text{trace}(B_+) = \text{trace}(B) - \frac{\|Bs\|^2}{s^T Bs} + \frac{\|y\|^2}{y^T s}$$

where $\| \cdot \|$ represents the standard 2-norm.

Recall that the trace of a symmetric matrix is the sum of its eigenvalues. Therefore, the trace of a positive-definite matrix is an upper bound on the greatest eigenvalue and the inverse of the trace is a lower bound on the least eigenvalue of the inverse of the matrix. In the next section, we present some convergence results that arose out of Powell’s study of the trace of $B_k$.

### 5.5.6 Powell’s Results

Powell showed that the global convergence of the BFGS method could be studied by measuring the trace and determinant of $B$. Before proceeding to his main theorem, we present other convergence results that led to it. A proof of the following theorem can be found in [97].

**Theorem 5.5.8.** (Powell, 1971) *Let $f$ have continuous second derivatives. Suppose there exists a constant $\epsilon > 0$ that is a lower bound for the eigenvalues of $\nabla^2 f(x)$ uniformly in $x$, then the sequence of points $\{x_k\}$ for $k = 0, 1, \ldots$ generated by the DFP method with exact line search converges to the global minimizer $x^*$.*

This theorem by Powell requires that the function be uniformly convex as is implied by his eigenvalue assumption. Hence, a unique global minimizer always exists. The following
results require only that the function be convex. In 1971, Powell [97] showed that if $f$ is convex, then for any positive-definite initial approximating matrix $B_0$ and any initial point $x_0$, the BFGS method gives

$$\lim \inf_k \| \nabla f_k \| = 0.$$  

We now present Powell’s global convergence theorem for the BFGS update where the steplength $\alpha$ is chosen by an inexact line search satisfying the two conditions

$$f(x_k + \alpha_k p_k) \leq f(x_k) + \eta \alpha_k \nabla f(x_k)^T p_k$$  \hspace{1cm} (5.7)
$$\nabla f(x_k + \alpha_k p_k)^T p_k \geq \zeta \nabla f(x_k)^T p_k$$  \hspace{1cm} (5.8)

where $0 < \eta < \frac{1}{2}$ and $\zeta < \eta < 1$. In the literature, (5.7) - (5.8) are known as the Wolfe conditions. In §5.5.10, we discuss how Zhang and Tewarson extended Powell’s theorem for the BFGS update to the Broyden preconvex class of updates.

**Theorem 5.5.9.** (Powell, 1976) Given $x_0$, let $f$ be a convex function, such that the set

$$\{ x : f(x) \leq f(x_0) \}$$  \hspace{1cm} (5.9)

is bounded and such that $f$ has continuous second derivatives in this set. Let $B_0$ be any positive-definite matrix. Then the BFGS method with steplength chosen to satisfy the Wolfe conditions generates a sequence $\{x_k\}$ for $k = 0, 1, \ldots$ such that $f(x_k)$ for $k = 0, 1, \ldots$ converges to a minimum of $f$.

**Remark 5.5.10.** Assumption (5.9) implies that $f$ has at least one minimizer. If the sequence $\{x_k\}$ generated by the BFGS method as described in Theorem 5.5.9 converges, and
converges to a solution, and if the Hessian matrix is positive definite, and the Lipschitz condition is satisfied, then the rate of convergence of the sequence is superlinear.

The following is a restatement of a theorem given by Powell [97] which is one of the first important superlinear convergence results.

**Theorem 5.5.11.** (Powell, 1971) Let $f$ be a $C^2$ nonlinear function and let $\nabla^2 f$ satisfy the Lipschitz condition (5.3) on all of $\mathbb{R}^n$. Moreover, let there exist some positive lower bound, independent of $x$, on the spectrum of $\nabla^2 f(x)$. Then the DFP method with exact line search converges superlinearly to $x^*$, the global minimizer of $f$, from any $x_0$ and for any positive-definite $H_0$.

Powell’s assumption that there exists some positive lower bound, independent of $x$, on the spectrum of $\nabla^2 f(x)$ implies that $f$ is uniformly convex and, therefore, has a unique global minimizer. Following Dixon’s Theorem 4.2.4, Powell [111] in 1976, stated that Theorem 5.5.11 holds for every member of the Broyden class and in particular for the BFGS method.\textsuperscript{64} To learn more about studies of the convergence behavior of secant methods with various line searches, we refer the interested reader to Buhmann and Fletcher [19], Nocedal [86], and Powell [98], in addition to those references cited in this section.

\textsuperscript{64}This result was also proved by Werner [123], in 1978, for a large class of line search methods.
5.5.7 Powell’s Observation

We are of the opinion that Powell’s 1986 paper [100] sheds light on some of the attributes of the BFGS method that explain its good performance. In 1986, Powell gained much insight into the global behavior of the BFGS and the DFP methods by focusing on a narrower class of problems. He considered the strictly convex quadratic objective function of two variables given by

\[ f(u, v) = \frac{1}{2}(u^2 + v^2) \]

and studied the DFP and BFGS methods with steplengths \( \alpha = 1 \). He analyzed the behavior of the DFP and BFGS methods for different choices of the initial point \( x_0 \) and the initial approximation \( B_0 \).

The analysis of the two-dimensional quadratic showed the following. If one eigenvalue of \( B_0 \) is large, it seems that the advantages of the BFGS method over the DFP method are greater if the other eigenvalue is small. Powell explained that the BFGS method is better at correcting small eigenvalues of \( B \) than large ones. In fact, an eigenvalue of \( B \) that is too small is easy to correct. Powell showed that the DFP formula can be highly inefficient at correcting large eigenvalues of \( B \) which are problematic and this is better dealt with by the BFGS method. This may be the main reason in general for the observed superiority of the BFGS formula. However, the BFGS method can take several iterations to correct a large eigenvalue of \( B \) and that during this calculation there may be little change in the objective function.

---

According to Byrd, “The biggest insight into what makes the BFGS better than the DFP method is in Powell’s 1986 paper [100].”
Some authors feel that the ability to correct large eigenvalues of $B$ can be improved by using negative values for the parameter $\theta$ in the Broyden class. We first mentioned in §4.2.5 that Zhang and Tewarson investigated this preconvex class. In §5.5.10, we elaborate on some of the results of their investigations.

5.5.8 Byrd, Nocedal and Yuan

In 1987, Byrd, Nocedal, and Yuan [22] extended the analysis of Powell described in the previous section to the Broyden convex class except for the DFP method. They established the following determinant relation for the Broyden convex class

$$\det(B_+) \geq \det(B) \frac{y^T s}{s^T B s}. \quad (5.10)$$

Using this, they proved global and superlinear convergence on uniformly convex problems for all members of the Broyden convex class using an exact line search, except for the DFP method, i.e., for $\theta \in [0, 1)$. These results are stated in the next two theorems.

**Theorem 5.5.12.** (Byrd, Nocedal and Yuan, 1987) Let $f \in C^2$ and let $x_0$ be an initial point for which $f$ satisfies the assumption that there is an open convex set $D$. Assume there exist constants $m, M > 0$ such that

$$m \|z\|^2 \leq z^T \nabla^2 f(x) z \leq M \|z\|^2 \quad (5.11)$$

for all $z \in \mathbb{R}^n$ and all $x \in D$. Consider any member of the Broyden class with $\theta \in [0, 1)$ with the steplength $\alpha$ chosen by an inexact line search that satisfies the Wolfe conditions
(5.7) - (5.8). Then for any positive-definite $B_0$, this secant method generates iterates which converge to $x^*$. 

The compound inequality (5.11) implies $f$ is uniformly convex. The following theorem describes under what conditions Byrd, Nocedal and Yuan were able to prove superlinear convergence.

**Theorem 5.5.13.** (Byrd, Nocedal and Yuan, 1987) Let $f \in C^2$ and $D$ be an open convex set. Assume that any algorithm generated by the Broyden class with $\theta \in [0,1)$ is implemented so that $\alpha$ is chosen by a backtracking strategy. Then, if there exist constants $m, M > 0$ such that (5.11) holds for all $z \in \mathbb{R}^n$ and all $x \in D$ and the Hessian at $x^*$ satisfies

$$\|\nabla^2 f(x) - \nabla^2 f(x^*)\| \leq \gamma \|x - x^*\|^p \quad \text{for} \quad p, \gamma > 0, \quad (5.12)$$

and for all $x$ in a neighborhood of $x^*$ and additionally, if $B_0$ is any positive-definite matrix, then the sequence $\{x_k\}$ converges superlinearly to the unique minimizer $x^*$.

In particular, Byrd, Nocedal and Yuan showed that if the objective function is convex, then the sequence generated by any method obtained from the Broyden class with $\theta \in [0,1)$ satisfies $\lim \inf_k \|\nabla f_k\| = 0$. Moreover, when they assumed that the objective function was uniformly convex and the procedure included a backtracking strategy, superlinear convergence is achieved.\(^{67}\)

\(^{66}\)A backtracking strategy first tries $\alpha = 1$ to determine if it satisfies the steplength criterion, e.g., the Wolfe conditions. If it does not, then a strategy is used to diminish the steplength. For example, $\alpha$ is chosen to satisfy the Wolfe conditions, in particular, the sufficient decrease condition (5.7)

\(^{67}\)This follows from the results of Dennis and Moré [35] and Griewank and Toint [64]. However, to apply them, they also required backtracking and that the Hessian matrix be Holder continuous (5.12) at $x^*$. 
In addition, the convergence analysis given in Byrd, Nocedal and Yuan’s paper shows that the BFGS method has a property that enables it to rapidly correct large eigenvalues and also shows that this property is diminished as $\theta$ is increased in $[0, 1)$. This complements Powell’s observation we presented in §5.5.7.

5.5.9 Byrd and Nocedal

In 1989, Byrd and Nocedal [21] showed that it was easier to work simultaneously with the trace and determinant relations of $B$. They introduced the measure function $\psi : \mathbb{R}^{n\times n} \to \mathbb{R}$ defined by

$$
\psi(B) = \text{trace}(B) - \ln(\det(B)),
$$

(5.13)

for any positive-definite matrix $B$.\textsuperscript{68} In §6.4 we elaborate on their work and explain how they simplified existing proofs by working with their measure function (5.13). This, in turn led to new characterization properties of the BFGS method which we present in §6.4 (see Theorem 6.4.1).

5.5.10 Zhang and Tewarson

As we mentioned earlier, Zhang and Tewarson [131], in 1988, performed numerical tests on updates from the Broyden class including updates obtained from using negative values of $\theta$. In this section, we outline their convergence results. We first restate a theorem in

\textsuperscript{68}While Powell’s results were generally accepted, many researchers found it difficult to understand the intuition behind them. In private communication (March 16, 2009), Richard Byrd explained that his development of the measure function with Nocedal was a result of their studies of Powell’s work in an attempt to uncover the intuition.
which Zhang and Tewarson [131] extended Powell’s global convergence Theorem 5.5.9 to the preconvex class.

**Theorem 5.5.14.** *(Zhang and Tewarson, 1988)* Let the conditions stated in Theorem 5.5.9 hold. Let $B_+$ be defined as in the Broyden class and $\theta^C$ as in (4.38). If, for $k$ large,

$$(1 - \nu)\theta^C \leq \theta < 0,$$

where $\nu$ is a constant satisfying $0 < \nu < 1$, then the corresponding secant method defined as in the Broyden class (4.37) with inexact line search generates a sequence $\{x_k\}$ for $k = 0, 1, \ldots$ such that $f(x_k)$ for $k = 0, 1, \ldots$ converges to, or terminates at, the minimizer of $f$.

Zhang and Teawarson [131] showed that for convex objective functions, updates from the preconvex class (4.40) have the same global convergence property as the BFGS update and may also possess a superlinear convergence rate if $\theta \leq 0$ is suitably chosen at each iteration. Their numerical tests with fixed negative values of $\theta_k$ gave results that show a moderate but consistent improvement over the BFGS method. However, they concluded that fixed-value update formulae in the preconvex class are not as recommended as varying-parameter algorithms.
Chapter 6

Known Characterizations of the BFGS Secant Method

In this chapter, we present several characterization properties of the BFGS secant method. These properties could shed some light on factors responsible for the empirical observation that the BFGS method is the most effective secant method for solving finite-dimensional unconstrained optimization problems.

6.1 Nazareth’s Step Length Result

The Conjugate Gradient (CG) method was developed in 1952 by Hestenes and Stiefel [68] to solve symmetric positive-definite linear systems of equations. It uses exact line search (readily obtained in closed form) and takes its first step in the direction of the negative gradient, i.e., in the steepest descent direction. In 1968, Myers [82] discovered a relation-
ship between the DFP and the CG methods. The following is a restatement of her theorem.

**Theorem 6.1.1.** (Myers, 1968) Consider minimizing a strictly convex quadratic function \( f \). Also consider the DFP and the CG methods and assume they each begin with the same initial point and the DFP method begins with the identity as the initial Hessian approximation. Then, the search direction vectors generated by the CG method and the DFP method are positive scalar multiples of each other; hence their iterates are the same when an exact line search is also used in the DFP method.

Recall Dixon’s [41] theorem (Theorem 4.2.4) that we mentioned in §4.2.5 which basically stated that the successive iterates generated by any two members of the Broyden class are identical, when minimizing a general differentiable function, provided an exact line search (1.8) is used and the degenerate \( \theta \) value, \( \theta^C \) (4.38), is never used. In 1979, Nazareth [83] made the following observation which yields a uniqueness property of the BFGS method.

**Theorem 6.1.2.** (Nazareth, 1979) Consider minimizing a strictly convex quadratic function \( f \). Also consider the CG method and a member of the Broyden class (\( \theta \neq \theta^C \)) and assume that they each begin with the same initial point \( x_0 \) and the member of the Broyden class uses the identity as the initial inverse Hessian approximation \( H_0 \). Then, the search direction vectors generated by the CG method and this member of the Broyden class are non-negative scalar multiples of each other; hence, with an exact line search, also for the member of the Broyden class, the iterates are the same. The BFGS method with steplength \( \alpha_k = 1 \) achieves an exact line search and, therefore, generates the same search direction and iterate as the CG method. Moreover, the BFGS method is the unique member of the Broyden class.
which generates the CG step with steplength $\alpha_k = 1$.

### 6.2 Finite Termination

Recall the definition of finite termination from §5.3 which states that a method is said to possess the finite termination property if it finds the minimizer of a quadratic function, if it exists, in a finite number of steps. Also, recall that Broyden [13] proved that every member of the Broyden class ($\theta \geq 0$) terminates after at most $n$ iterations with the exact minimizer of any strictly convex quadratic function given that an exact line search is used. From Nazareth we know that the BFGS method with $\alpha = 1$ gives an exact line search. Thus, it seems reasonable to expect that the BFGS method is the only member of the Broyden class that achieves finite termination with steplength $\alpha = 1$.

### 6.3 Updates to Cholesky Factors

The BFGS method preserves positive-definiteness, however, there is always a possibility that in the actual computation of $B_+$, positive-definiteness may be lost due to numerical rounding error. One strategy to deal with this difficulty is to work with a factorized positive-definite $B$. Methods that utilized this idea originated in the work of Gill and Murray [57] in 1972, and were subsequently developed and analyzed by Goldfarb [60] and Gill, Golub, Murray, Saunders [56], among others. In this section, we describe the contributions of Dennis and Schnabel [37] and Zhang and Tewarson [130] as they gave way to interesting
characterizations of the BFGS method.\textsuperscript{69}

Consider the Hessian approximation $B$ with the Cholesky factorization

$$B^C = LL^T$$  \hspace{1cm} (6.1)

where $L$ is a lower-triangular matrix. Now, instead of explicitly updating $B$ to obtain $B_+$, the Cholesky factor $L$ in (6.1) is stored and updated at each step via a simple additive update of $L$

$$L \rightarrow J_+ = L + D$$

where $D$ is the correction.\textsuperscript{70} The new approximation $B_+$ is expressed in like manner

$$B^C_+ = J_+ J_+^T,$$  \hspace{1cm} (6.2)

and the secant equation is written as

$$J_+ J_+^T s = y$$  \hspace{1cm} (6.3)

and reduces to

$$LD^T s + DL^T s + DD^T s = y - LL^T$$  \hspace{1cm} (6.4)

which is nonlinear in $D$. To address this nonlinearity, Dennis and Schnabel [37], in 1981, suggested “linearizing” the secant equation (6.4). Their use of the term “linearize” was

\textsuperscript{69}One of the earliest papers devoted to modifying matrix factorization is Bennett’s 1965 paper [5] in which he investigated $B = LDU^T$ where $L$ is a lower triangular matrix with unit elements on the diagonal, $U$ is an upper triangular matrix with unit elements on the diagonal, and $D$ is a diagonal matrix, however, Bennett’s method was not numerically stable. The subsequent work of Gill and Murray explored the matrix factorization $B = LDL^T$ in the context of secant methods.

\textsuperscript{70} $J_+$ is used instead of $L_+$ because the update of the factor $L$ is not a lower triangular matrix. To learn more about how to obtain $L_+$ from $J_+$ see [37].
unfortunate because it misled people into thinking that the nonlinear term was just dropped which, of course, would mean that the secant equation would no longer be satisfied.\textsuperscript{71}

However what they meant when they said “linearize” the secant equation was to replace the nonlinear secant equation by two coupled linear ones. To accomplish this, they defined the variable $u$ by

$$u = J^T_+ s \quad \text{(6.5)}$$

in which case (6.3) would become

$$J_+ u = y. \quad \text{(6.6)}$$

The following is a restatement of a theorem in which Dennis and Schnabel explain the conditions under which $J_+$ exists. For a proof, see [37].

**Theorem 6.3.1.** (Dennis and Schnabel, 1981) Let $s, y \in \mathbb{R}^n$ and $s \neq 0$. Then there exists a nonsingular $J_+ \in \mathbb{R}^{n \times n}$ such that $J_+ J^T_+ s = y$, if and only if $s^T y > 0$.

### 6.3.1 Least Change Updates to Cholesky Factors

We showed in Chapter 4, that many updates (in particular, the BFGS update) could be expressed as rank-one updates to the factor $L$ (see Brodlie et al. [10] and Davidon [31]). However, due to the nonlinearity of the secant equation (6.4), it was not until much later that updates could be shown to be a least-change update to the Cholesky factor $L$ itself.

Using the decomposition of the secant equation described earlier by (6.5) - (6.6), Dennis Greenstadt [63] reminisced on how he incorrectly interpreted Dennis’ suggestion to linearize (6.4) to mean dropping the quadratic term.
and Schnabel [37] were able to present a new derivation of the BFGS update which demonstrates how the BFGS update is a least change update to the Cholesky factor. We describe their procedure for determining (6.2).

Assume the Hessian approximation $B$ has the Cholesky factorization (6.1). The $J_+$ which is nearest to $L$ in the Frobenius norm and satisfies (6.6) is

$$ J_+ = L + \frac{(y - Lu)u^T}{u^T u}. \quad (6.7) $$

The $u$ that satisfies (6.5) is

$$ u = L^T s + u \frac{(y - Lu)^T s}{u^T u}, \quad (6.8) $$

which can be satisfied only if

$$ u = \alpha L^T s \quad (6.9) $$

for some scalar $\alpha$. Substitute (6.9) into (6.8) and simplify to obtain

$$ \alpha^2 = \frac{y^T s}{s^T L L^T s} = \frac{y^T s}{s^T B^C s}. $$

Choose the positive square root for $u$ given by

$$ u = \left( \sqrt{\frac{y^T s}{s^T B^C s}} \right) L^T s $$

and substitute it in (6.7) to obtain

$$ J_+ = L + \left[ y - \left( \sqrt{\frac{y^T s}{s^T B^C s}} \right) B^C (L^T s)^T \right] \left( \sqrt{\frac{y^T s}{s^T B^C s}} \right) s^T B^C s. \quad (6.10) $$

Form $B^C_+ = J_+ J_+^T$ using (6.10) to obtain the BFGS secant update in terms of its Cholesky factors.
6.3.2 Weighted Least Change Updates to Cholesky Factors

Dennis and Schnabel [37] presented another derivation of the BFGS secant method. Again, to determine (6.2), first assume you have the Cholesky factorization (6.1). However, this time, choose $J_+$ to solve

$$\min_{J_+} \|W_L (J_+ - L) W_R\|_F \quad \text{subject to} \quad J_+ u = y$$

(6.11)

for nonsingular weighting matrices $W_L, W_R \in \mathbb{R}^{n \times n}$ and then solve for the $u$ that satisfies (6.5). Solving (6.11) with $W_L = W_R = I$ yields the BFGS update. In fact, the BFGS results from any choice of $W_L$ and $W_R$ for which $L^T s$ is an eigenvector of $M = W_R^{-T} W_R^{-1}$.

To obtain the BFGS update formula, solve (6.11) for either the positive or negative square root of (6.10) and substitute this $J_+$ into (6.2).

6.3.2.1 Least Change Cholesky (LCC) Updates

In 1987, Zhang and Tewarson [130] derived a class of least-change updates to Cholesky factors of $B$ known as least change Cholesky (LCC) updates using the same underlying idea as Dennis and Schnabel. We discuss their work that led them to discover a uniqueness property of the BFGS update which we present at the end of this section. To aid in our discussion, we introduce the following notation utilized by Zhang and Tewarson:

1. $Q(u, v) = \{ X : X \in \mathbb{R}^{n \times n}, X u = v \}$ \quad $(u, v \in \mathbb{R}^n)$.

2. $Q^T(u, v) = \{ X : X^T \in Q(u, v) \}$. 
Given symmetric positive-definite weighting matrix \( M \in \mathbb{R}^{n \times n} \) and \( s, y \in \mathbb{R}^n \) with \( s^T > 0 \), they considered the constrained optimization problem

\[
\min \{ \| J_+ - L \|_{F,M} : J_+ J_+^T \in Q(s, y) \}.
\]

(6.12)

To address the nonlinearity of the constraint in (6.12), Zhang and Tewarson considered the following two subproblems:

1. For any given \( u \in \mathbb{R}^n \) with \( u^T u = y^T s > 0 \), find the solution \( D(u) \) to the subproblem

\[
\min \{ \| D \|_{F,M} : L + D \in Q^T(s, u) \cap Q(u, y) \}.
\]

(6.13)

2. Find a global solution \( u^* \) to the problem

\[
\min \{ \| D(u) \|_{F,M} : u \in \mathbb{R}^n, u^T u = y^T s \}
\]

(6.14)

and let \( J_+ = L + D(u^*) \).

The following is a restatement of a theorem in which Zhang and Tewarson [130] describe how solving the two subproblems above is equivalent to solving the constrained optimization problem (6.12). For a proof, see [130].

**Theorem 6.3.2.** (Zhang and Tewarson, 1987) Problem (6.12) is equivalent to problem (6.14), with \( D(u) \) being the unique solution to problem (6.13). That is, if \( J_+ \) solves (6.12), then \( u = J_+^T s \) solves (6.14), and if \( u = u^* \) solves (6.14), then \( J_+ = L + D(u^*) \) solves (6.12).

Zhang and Tewarson identified an essential difference between \( B_+ \) obtained as the solution to the least change problem (4.33) and \( B_+^C = J_+ J_+^T \) where \( J_+ = L + D(u^*) \) solves
In general, $B_+ - B$ is of rank two for any nonsingular symmetric weighting matrix [36], but Zhang and Tewarson showed that, in general, $B^C_+ - B$ is of rank four. The following is a restatement of a theorem which is a result of their investigations. For a proof, see [130].

**Theorem 6.3.3.** (Zhang and Tewarson, 1987) Let $M$ satisfy

$$Ms = (s^T Ms)r$$

where

$$r = \frac{y - \alpha LL^T s}{y^T s - \alpha s^T LL^T s} \quad \text{with} \quad \alpha = \left( \frac{y^T s}{s^T LL^T s} \right)^{\frac{1}{2}}$$

and let $u^* = \alpha L^T s$. Then the rank-one LCC update $D(u^*)$ is independent of $M$ and is the correction to the Cholesky factor that yields the BFGS update.

From this, they discovered the following uniqueness property of the BFGS update.

**Remark 6.3.4.** The BFGS update is the unique update that can be obtained from a rank-one least change update to the Cholesky factor $L$ in the weighted Frobenius norm and independent of the choice of the weighting matrix.

### 6.4 Least Change in Byrd-Nocedal Measure

We first mentioned in §5.5.9 how Byrd and Nocedal [21] introduced the measure function (5.13), in 1989, to study the convergence of quasi-Newton methods. They followed the framework that Dennis and Moré [36] employed to prove the superlinear convergence
of quasi-Newton methods. However, they used the $\psi$-measure function instead of the weighted Frobenius norm to measure the change of the Hessian approximations. This means that we now have the least change property in a non-norm measure. As a consequence, Byrd and Nocedal were able to modify existing convergence proofs previously given by Powell [101] and Dennis and Moré [36] that concluded that the rate of convergence of the BFGS method is superlinear if steplength $\alpha_k = 1$ for all sufficiently large $k$. In private communication (March 16, 2009), Richard Byrd explained that the most surprising result of their investigations was that by scaling $B$ and looking at $\psi(G^{-1}B) = \text{trace}(G^{-1}B) - \ln(\det(G^{-1}B))$ where $G$ is the Hessian at the solution, they were able to prove superlinear convergence of the BFGS method.

Recall that Goldfarb [59] showed that the correction in the BFGS formula solves a minimization problem in the weighted Frobenius norm $\|E\|_W^2 = \text{trace}(EWEW)$. In 1991, Fletcher [46] showed that the BFGS formula solves an optimization problem with respect to the Byrd-Nocedal measure function.\footnote{Although Byrd and Nocedal proved this result, it was Fletcher who first published it.} The following is a restatement of Fletcher’s theorem. For a proof, see Fletcher [46].

**Theorem 6.4.1.** (Fletcher, 1991) If $H$ is symmetric, positive definite and $s^Ty > 0$, then the unique solution to the optimization problem

$$\min_B \psi(H^{1/2}BH^{1/2}) \quad \text{subject to} \quad Bs = y \quad \text{and} \quad B \text{ symmetric}$$

is the matrix $B_+$ given by the BFGS update formula (4.19).

In §6.5.4, we state two characterizations of the BFGS method that Yabe, Martinez and Tapia
proved using the Byrd-Nocedal $\psi$-measure function, $\psi(D^{\frac{1}{2}}BD^{\frac{1}{2}})$, with symmetric positive-definite weighting matrix $D$.

6.5 Yabe, Martinez, Tapia Least Change in Weighted Byrd-Nocedal Measure

The BFGS method works well, but when it is not effective, it can be due to the fact that it produces updates with large eigenvalues, especially in higher dimensions, as it was, in fact, designed to do in order to correct the conditioning problem of the DFP method. In particular, see the derivation by Shanno in §4.2.2.3. To overcome this difficulty, two remedies have been considered: sizing and shifting. In the upcoming sections, we describe the size and shift approach and outline some results that Yabe, Martinez and Tapia [128] obtained in 2004, when they combined both remedies and applied them to a two-parameter family they called the sized Broyden class (6.18), (6.19).

6.5.1 Sizing

In 1968, Bard used the term scaling to refer to multiplying a matrix by a scalar. In 1981, Dennis, Gay and Welsch [34] decided to call this process sizing instead of scaling. To obtain a better Hessian approximation, we can size the initial approximation, $B$, by multi-

---

[73] John Dennis, in private communication (March 13, 2009), explained that the term scaling had been used by some to mean multiplying the objective function by a constant. In this case, a rather direct calculation shows that the Newton search direction is independent of this constant. Hence, he decided to use the term sizing when only the Hessian is multiplied by a constant.
plying it by a scalar factor, $\gamma$, before the secant update is made. Then, we update $\gamma B$, rather than $B$ to obtain $B_+$. The following propositions and definitions introduced by Contreras and Tapia [28], in 1993, will aid in our understanding of the effects of sizing.

**Definition 6.5.1.** The convex spectrum of a matrix $B$, denoted $\text{conspectrum}(B)$, corresponds to the convex hull of the eigenvalues of $B$.

When $B$ is symmetric, the convex spectrum of $B$ is an interval of the reals, and thus, we refer to it as the *interval spectrum* of $B$.

**Definition 6.5.2.** The scalar $\gamma$ sizes $B \in \mathbb{R}^{n \times n}$ relative to $A \in \mathbb{R}^{n \times n}$ if

$$\text{conspectrum}(\gamma B) \cap \text{conspectrum}(A) \neq \emptyset.$$

**Proposition 6.5.3.** Let $A, B \in \mathbb{R}^{n \times n}$ be symmetric matrices. Then the scalar $\gamma$ sizes $B$ relative to $A$ if and only if there exists $u, v \in \mathbb{R}^n$ such that

$$\gamma = \frac{u^T A u}{u^T v} \frac{v^T v}{v^T B v}.$$

The proof, which is a direct consequence of the Rayleigh quotient, can be found in Contreras and Tapia [28].

**Definition 6.5.4.** Consider a $C^2$ function $f : \mathbb{R}^n \to \mathbb{R}$. Also consider a secant method for minimizing $f$ with current iterate $x$, current Hessian approximation $B$, and current step $s$.

We say that the scalar $\gamma$ sizes $B \in \mathbb{R}^{n \times n}$ relative to the Hessian of $f$ at $x$ if there exists $\pi \in \mathbb{R}^n$ such that

$$\text{conspectrum}(\gamma B) \cap \text{conspectrum}(\nabla^2 f(\pi)) \neq \emptyset$$

for some $\pi \in B(x, \|s\|)$.
Sizing a matrix shifts its spectrum. In particular, sizing the Hessian approximation relative to the true Hessian overlaps the spectra of the two matrices.\(^{74}\) In 1974, Oren and Luenberger [90] introduced the Oren-Luenberger (O-L) sizing factor:

\[
\gamma_{OL} = \frac{y^T s}{s^T B s}
\]  

(6.15)

and suggested sizing \(B\) with \(\gamma_{OL}\) in the BFGS direct update (4.19) at each iteration prior to updating.

In 1993, Contreras and Tapia [28] observed that \(\gamma_{OL}\) has the property that the interval spectrum of \(\gamma_{OL} B\) overlaps the interval spectrum of \(\nabla^2 f(x + \theta s)\) for some \(\theta \in (0, 1)\), i.e., sizes \(B\) relative to the Hessian of \(f\). A proof was not included in the original text so, for the sake of completeness, we offer a proof of this observation.

**Theorem 6.5.5.** Consider a \(C^2\) function \(f : \mathbb{R}^n \to \mathbb{R}\) with \(x, s \in \mathbb{R}^n\) and \(B \in \mathbb{R}^{n \times n}\). Assume \(B\) is nonsingular. Let \(s = -B^{-1} \nabla f(x)\) and \(y = \nabla f(x + s) - \nabla f(x)\). Then \(\gamma_{OL}\) sizes \(B\) relative to \(\nabla^2 f(x + \theta s)\) for some \(\theta \in (0, 1)\).

**Proof.** We begin by sizing the Hessian approximation, \(B\), with the Oren-Luenberger sizing factor \(\gamma_{OL}\) so that

\[
\gamma_{OL} B = \frac{y^T s}{s^T B s} B.
\]

For \(s_k \neq 0\), we multiply \(\frac{y^T s}{s^T B s} B\) on the left by \(s^T\), on the right by \(s\), and divide by \(s^T s\). This simplifies to

\[
\frac{y^T s}{s^T s}.
\]

\(^{74}\)As early as 1981, Dennis, Gay, and Welch [34] hoped that sizing would cause the spectrum of the approximate Hessian update to overlap the spectrum of the current Hessian approximation.
Using the Mean-Value Theorem, we obtain

\[ y^T s = (\nabla f(x + s) - \nabla f(x))^T s \]
\[ = (\nabla^2 f(x + \theta s)s)^T s \]
\[ = s^T \nabla^2 f(x + \theta s)^T s \]
\[ = s^T \nabla^2 f(x + \theta s)s \]

for some \( \theta \in (0, 1) \). This leads to

\[ \frac{s^T \left( \frac{y^T B s}{s^T s} \right)}{s^T s} = \frac{s^T \nabla^2 f(x + \theta s)s}{s^T Bs} \]

for some \( \theta \in (0, 1) \). Therefore, we have shown that \( \gamma_{OL} \) sizes \( B \) relative to \( \nabla^2 f(x + \theta s) \).

\[ \blacksquare \]

### 6.5.2 Selective Sizing

In 1978, Shanno and Phua [109] suggested that the BFGS update should be sized only at the first iteration as opposed to every iteration based on numerical evidence that showed that, in general, sizing only the initial iteration was superior to sizing at every iteration. In 1993, Contreras and Tapia [28] studied sizing strategies for both the BFGS and the DFP updates - sizing at every iteration, sizing only at the first iteration, selectively sizing, or never sizing - to investigate how sizing affected the performance of the DFP and the BFGS methods. They observed that a direct consequence of the secant equation is \( \frac{y^T s}{s^T B s} = 1 \). This, in turn, implies that in any secant method, all Hessian approximations (except for the initial
approximation) are already automatically sized, adding credibility to the Shanno-Phua proposal. However, as we will soon explain, Contreras and Tapia discovered a selective sizing strategy that was superior to other sizing strategies.

Contreras and Tapia [28] realized that sizing the BFGS with $\gamma_{OL}$ at each iteration is not good. In fact, $\gamma_{OL}$ tends to size the BFGS update too much, especially for large dimensional problems. Recall that sizing is performed in hopes of significantly overlapping the respective spectra (of the Hessian approximation and the Hessian). Of course, sizing with $\gamma_{OL}$ does not ensure that the spectrum of $B$ entirely overlaps the spectrum of $\nabla^2 f(x + \theta s)$ (see Theorem 6.5.5). In fact, it could happen that the overlap of the respective spectra is not much more than a single point.

In an effort to better overlap the spectrum of the approximate Hessian with the spectrum of the true Hessian, Contreras and Tapia [28] introduced the Centered Oren-Luenberger sizing factor (C-O-L factor)

$$\gamma_{COL} = \frac{y^T s_-}{s^T s_-} + \frac{y^T s}{s^T s} \quad (6.16)$$

where $s_-$ and $y_-$ are calculated at the previous iteration. The idea behind $\gamma_{COL}$ was to attempt to match the center of the interval spectrum of $B$ with the center of the interval spectrum of $\nabla^2 f(x)$. Since these centers are unknown, Contreras and Tapia used an average of two points (the current point and the previous point) in the interval spectrum to serve as an approximation to its center.

Since sizing the BFGS update at each iteration proved to be unsatisfactory, Contreras and Tapia [28] suggested selectively sizing the update to yield the best results for the BFGS
method. They suggested sizing the Hessian approximation in the first iteration of the BFGS (or the DFP) method with $\gamma_{OL}$ (6.15) and then selectively sizing the BFGS update at other iterations with $\gamma_{COL}$ (6.16). Essentially, they felt that $\gamma_{OL}$ sized the BFGS update too much and suggested $\gamma_{COL}$ because, for $\gamma_{OL} < 1$, it follows that $\gamma_{OL} < \gamma_{COL} < 1$. Hence, they viewed $\gamma_{COL}$ as a damping or softening of $\gamma_{OL}$. They proposed, for small positive constants $\epsilon_1$ and $\epsilon_2$, if $\gamma_{COL} \leq 1 - \epsilon_1$, then $B_k$ should be sized using $\max(\epsilon_2, \gamma_{COL})$. Using $\epsilon_2$ prevents inadvertently creating a near singular matrix as a result of sizing with an excessively small sizing factor. In summary, this means that if the sizing factor is close to 1, do not size, or if it is too small (close to 0), then size with $\epsilon_2$. However, if the sizing factor is safely in between 0 and 1, then size with $\gamma_{COL}$. It is interesting that by selectively sizing the DFP method, Contreras and Tapia [28] were able to make it competitive with the BFGS method (selectively sized, or not). Yet, in any case, they found that the DFP method is not as robust as the BFGS method.

6.5.3 Sizing and Shifting

The BFGS direct update (4.19) may have large eigenvalues (especially in higher dimensions). However, sizing the update may be problematic because sizing the BFGS update may lead to an excessively small sizing factor and near singularity of the sized matrix. To compensate for near singularity, Yabe, Martinez and Tapia [128], in 2004, suggested to size and then shift the sized matrix $\gamma B$, by adding a matrix $M$ to the sized matrix. They
introduced an update of the form

\[ B_+ = BFGS(\gamma B + M, s, y) \]  

(6.17)

where \( \gamma \) is a sizing factor. Note, for any \( \gamma \) and \( M \), the update (6.17) satisfies the secant equation. Yabe, Martinez and Tapia chose \( M = \theta vv^T \), where

\[ v = \sqrt{s^T B s} \left( \frac{y}{y^T s} - \frac{B s}{s^T B s} \right), \]

to obtain what is referred to as the sized Broyden class

\[ B_+ = BFGS(\gamma B + \theta vv^T, s, y) \]  

(6.18)

\[ = BFGS(\gamma B, s, y) + \theta vv^T. \]  

(6.19)

This form of the sized Broyden class (6.19) allows the sizing and shifting to be performed independently of each other allowing us to separately control the sizing parameter and the shifting term and, hence, only size the matrix \( B \). A significant feature is that this choice of \( M \) allows the shift to be performed after the sized matrix is updated. In this case, shifting, also known as switching, refers to updating the approximate Hessian by a different member of the Broyden class, that is, switching from the BFGS update, where \( \theta = 0 \), to an alternate member of the Broyden class, where \( \theta \neq 0 \). In the next section, we discuss the size and shift approach that Yabe, Martinez and Tapia [128] used to determine the best parameters \( \gamma \) and \( \theta \) in the sized Broyden class (6.18).

### 6.5.4 Weighted \( \psi \)-Optimal Values for the Sized-Broyden Class

Yabe, Martinez and Tapia [128] concluded that it is beneficial to follow a sizing of the BFGS update with a shifting of the BFGS update. To determine the optimal values of
the sizing factor $\gamma$ and the shift $\theta$ in the sized Broyden class (6.18), they considered the following three optimization problems:

1. Given the sizing factor $\gamma^*$, find the shift $\theta^*$ as solution of

$$\min_{\theta} \psi \left( D^{-\frac{1}{2}} B_+ D^{-\frac{1}{2}} \right)$$

   (6.20)

2. Given the shift $\theta^*$, find the sizing factor $\gamma^*$ as solution of

$$\min_{\gamma} \psi \left( D^{-\frac{1}{2}} B_+ D^{-\frac{1}{2}} \right)$$

   (6.21)

3. Find the sizing factor $\gamma^*$ and the shift $\theta^*$ as solution of

$$\min_{\gamma, \theta} \psi \left( D^{-\frac{1}{2}} B_+ D^{-\frac{1}{2}} \right)$$

   (6.22)

and studied three choices for the symmetric positive-definite weighting matrix $D$: $D = I$, the identity, $D = B$, the approximate Hessian, and $D = \nabla^2 f(x)$, the exact Hessian.

The $\psi$-measure is globally and uniquely minimized by the identity matrix. Therefore, there is a bias towards parameters that force $(D^{-\frac{1}{2}} B_+ D^{-\frac{1}{2}})$ to approximate the identity. If $(D^{-\frac{1}{2}} B_+ D^{-\frac{1}{2}}) = I$, then $D = B_+$. Hence, the choice $D = B$ yields the member of the sized Broyden class closest to $B$, a least change update in the Byrd-Nocedal measure. The choice $D = \nabla^2 f(x)$ yields the member of the sized Broyden class closest to the Hessian, an update closest to Newton’s method in the Byrd-Nocedal measure. Finally, the choice $D = I$ yields the member of the sized Broyden class closest to the identity, an update closest to steepest descent in the Byrd-Nocedal measure.
6.5.4.1 Solutions

After lengthy algebraic calculations, Yabe, Martinez and Tapia [128] solved the three minimization problems (6.20), (6.21), (6.22) in closed form using each of the different choices of weighting matrices $D$. They reached the following conclusions. The BFGS update solved the three optimization problems uniquely with $D = B$. Setting $\theta = 0$, which means we are using the BFGS update, yields the sizing factor $\gamma = 1$ which means no sizing is optimal. Similarly, if we set $\gamma = 1$, which means we are not going to size, then the optimal solution is the BFGS update. Therefore, Yabe, Martinez and Tapia recognized that the BFGS update does not like to be sized and that BFGS update with no sizing is optimal in this sense. However, the choice $D = B$ was not their best choice. Instead, they found that the most effective choice numerically was $D = I$. In this case, when the sizing factor dictated that we should size, we infer that there is a bad match of information between $B$ and $\nabla^2 f(x)$. Thus, the choice $D = I$ prevents this faulty information from further contaminating the update.
Chapter 7

Secant Update Classes and Some of their Properties

Throughout the previous chapters we included significant historical development of the methods we discussed. This pattern is continued as we trace the evolution of several rank-two secant update classes that have appeared in the literature but have not been given much recognition. Since the literature abounds with update classes, we want to show some relationships among them and show containment when possible. Some known and some new characterizations of these update classes are presented. Included is a conjecture made by Schnabel in his 1977 Ph.D. thesis, which we prove, as well as, some interesting findings that Tapia presented in a 1984 unpublished paper entitled “On Averaging and Representation Properties of the BFGS and Related Secant Updates.”

Recall, in Chapter 4, we explained how a broad class of rank-one secant updates (4.3)
was obtained from the general rank-one update formula (4.2) and we stated some choices of the parameter $v$ that yielded popular rank-one update methods. In this chapter, we construct a proof of how Dennis [33] applied Powell’s method of iterated projections to the general rank-one update formula to generate the Dennis class of symmetric rank-two updates. In addition, we present an alternative derivation of the Dennis class. Different choices of the parameter $v$ in the Dennis class that yield popular rank-two update methods are given. Furthermore, we derive an extension to the Dennis class which turns out to be the known Davidon class. We present the derivation of a general formula which can be used to represent any symmetric rank-two update. Parameter choices that yield different update classes are given. The chapter concludes with the derivation of what we call the extended Dennis-Davidon class.

### 7.1 Dennis Class

We first mentioned in Chapter 4 how Dennis [33], in 1972, applied Powell’s method of iterated projections to members of the rank-one inverse update class (4.5) to develop the Dennis inverse update class (4.12), more commonly written in the following form of the direct update class

$$B^DC_+ = B + \frac{(y - Bs)v^T + v(y - Bs)^T}{v^Ts} - \frac{s^T(y - Bs)vv^T}{(v^Ts)^2} \quad (7.1)$$
where \( y, s, v \in \mathbb{R}^n \) and \( v^T s \neq 0 \) and \( B \in \mathbb{R}^{n \times n} \) is symmetric. The following are well-known choices of the parameter \( v \):

\[
\begin{align*}
\text{SR1} & \quad v = y - Bs \\
\text{PSB} & \quad v = s \\
\text{DFP} & \quad v = y \\
\text{BFGS} & \quad v = y + \left[ \frac{y^T s}{s^T Bs} \right]^\frac{1}{2} Bs \\
\text{Greenstadt} & \quad v = Bs.
\end{align*}
\]

In 1983, Dennis and Schnabel [107] stated that all of the members of the Broyden class (4.37) for which the rank-two correction \( B_+ - B \) has one negative and one positive eigenvalue can equivalently be represented as members of the Dennis class (7.1) where \( v \) is a linear combination of \( y \) and \( Bs \).\(^{75}\)

Dennis did not include a proof of how he obtained the Dennis class by applying Powell’s method of iterated projections to the general rank-one update formula so we construct one in the next section.\(^{76}\) We follow this proof with an alternative derivation of the Dennis class.

\(^{75}\)This follows from the 1973 work of Brodlie, Gourlay, and Greenstadt [10] and the subsequent 1976 work of Gay [52]; both of which we discuss in §7.3.

\(^{76}\)In Chapter 4, we stated that Powell also did not provide a formal proof of how he derived the PSB method by applying his method of iterated projections to the good Broyden method.
7.1.1 Derivation of the Dennis Class using Iterated Projections

The following is a restatement of a theorem due to Dennis. For the sake of completeness we construct a proof.

**Theorem 7.1.1.** (Dennis, 1972) Consider the rank-one secant update

\[ B_1 = B_0 + \frac{(y - B_0s)v^T}{v^Ts} \]  

(7.2)

where \( y, s, v \in \mathbb{R}^n, v^Ts \neq 0 \), and \( B_0 \in \mathbb{R}^{n \times n} \) is symmetric. Also consider the symmetrized form of \( B_1 \)

\[ C_1 = \frac{B_1 + B_1^T}{2}. \]  

(7.3)

Using (7.2) and (7.3), generate the sequence \( \{C_k\} \) where

\[ B_{k+1} = C_k + \frac{(y - C_ks)v^T}{v^Ts}, \]
\[ C_{k+1} = \frac{B_{k+1} + B_{k+1}^T}{2} \quad k = 1, 2, \ldots \]  

(7.4)

The limit of the sequence \( \{C_k\} \) is

\[ B_\infty = B_0 + \frac{(y - B_0s)v^T + v(y - B_0s)^T}{v^Ts} - \frac{s^T(y - B_0s)v^Tv}{(v^Ts)^2}. \]  

(7.5)

**Remark 7.1.2.** If we rewrite (7.5) as an update formula it represents, for different choices of \( v \), all members of the Dennis class of secant updates (7.1).

**Proof.** Consider the general rank-one secant update formula (7.2) and its symmetrized form (7.3). From (7.4), we can write

\[ C_{k+1} = \frac{(C_k + (y - C_ks)d^T) + (C_k + (y - C_ks)d^T)^T}{2} \]  

(7.6)
where \( d = \frac{v}{s^T v} \). Define
\[
\overline{B}_+ = B + (y - Bs)d^T + ((y - Bs)d^T)^T - s^T(y - Bs)dd^T.
\]  
(7.7)

We begin by demonstrating that the update given by (7.7) is a fixed point of the recursion formula (7.6). To prove this, we first notice that (7.6) holds with both \( C_{k+1} \) and \( C_k \) replaced by \( \overline{B}_+ \) if and only if
\[
\overline{B}_+ = B + \frac{(y - \overline{B}_+ s)d^T + ((y - \overline{B}_+ s)d^T)^T}{2}.
\]  
(7.8)

Noting that \( \overline{B}_+ \) is symmetric, our proof reduces to showing that
\[
(y - \overline{B}_+ s)d^T + ((y - \overline{B}_+ s)d^T)^T = 0.
\]

But this is an immediate consequence of the fact that
\[
\overline{B}_+ s = \left[ B + (y - Bs)d^T + ((y - Bs)d^T)^T - s^T(y - Bs)dd^T \right] s
\]
\[
= Bs + (y - Bs) + d(y - Bs)^T s - s^T(y - Bs)d
\]
\[
= y + d(y^T s) - d(s^TBs) - (s^Ty)d + (s^TBs)d
\]
\[
= y.
\]

Now we need to prove
\[
C_k \to \overline{B}_+.
\]  
(7.9)

To do this, first let \( E_k = C_k - \overline{B}_+ \) for \( k = 1, 2, \ldots \) Then, using the fact that (7.7) is a fixed point of (7.6) we can subtract (7.8) from (7.6) to obtain
\[
E_{k+1} = \frac{(E_k - E_k sd^T) + (E_k - E_k sd^T)^T}{2}
\]
\[
= \frac{E_k(I - sd^T) + (I - sd^T)^T E_k}{2}.
\]  
(7.10)
Let $P = (I - sd^T)$. Repeated substitution of $E_{k-j}$ for $j = 0, 1, \ldots$ in the recursion formula (7.10) yields

$$E_k = \frac{1}{2^k} \sum_{j=0}^{k} \binom{k}{j} (P^T)^j E_0 P^{k-j}.$$ 

But $P^j = P$ and $(P^T)^j = P^T$ for $j = 1, 2, \ldots$, and, therefore, the above sum reduces to

$$E_k = \frac{1}{2^k} [E_0 P + P^T E_0] + \frac{2^k - 2}{2^k} P^T E_0 P.$$ 

Consequently, $E_k \to P^T E_0 P$ and (7.9) follows immediately if we can show $P^T E_0 P = 0$. To this end, observe that $Ps = 0$, and, therefore, $P^T E_0 Ps = 0$. If we let $w \in \{v\}^\perp$ where the $\perp$ denotes the orthogonal complement of the subspace $\{v\}$, then we have

$$P^T E_0 P w = P^T E_0 w$$

$$= -P^T [(y - C_0 s)d^T + ((y - C_0 s)d^T)^T - s^T(y - C_0 s)dd^T]w$$

$$= -P^T [(y - C_0 s)d^T + d(y^T - s^T C_0) - s^T(y - C_0 s)dd^T]w$$

$$= -P^T [0 + d(y^T - s^T C_0)w - 0]$$

$$= -P^T d(y^T - s^T C_0)w$$

$$= 0,$$

where the last equality follows from the fact that $P^T d = 0$ since $d = \frac{v}{s^Tv}$. Therefore,

$$P^T E_0 P|_{\{v\}^\perp} = 0.$$ 

Since we know $s \notin \{v\}^\perp$ and $P^T E_0 Ps = 0$ then it follows that $P^T E_0 P = 0$ and the proof is complete.  

$\blacksquare$
7.1.2 Algebraic Derivation of the Dennis Class

Given Powell’s construction of the PSB update from the Broyden update, Dennis took the timely approach of applying Powell’s method of iterated projections to derive the Dennis class. In this section, we show how the Dennis class can be derived from the rank-one secant update formula in a purely algebraic manner and need not utilize the method of iterated projections.

Consider the general rank-one secant update formula (7.2) where \( y, s, v \in \mathbb{R}^n \) and \( v^T s \neq 0 \), and \( B \in \mathbb{R}^{n \times n} \) is symmetric. Also consider the symmetrized form of \( B_+ \),

\[
B_+ = \frac{B_+ + B_+^T}{2}
\]

\[
= B + \frac{(y - Bs)v^T}{v^T s} + \frac{v(y - Bs)^T}{v^T s}.
\] (7.11)

Let us consider the task of adding to this as simple a rank-one matrix as possible, call it \( R \), so that \( B_+ + R \) remains rank-two, symmetric, and satisfies the secant equation. For \( B_+ \) to satisfy the secant equation, \( R \) must satisfy

\[
Rs = -\frac{(y - Bs)^T s}{v^T s} v.
\] (7.12)

Observe from (7.12), that we can view \( Rs \) as \( \gamma v \). Thus, we can choose \( R \) to be of the form

\[
R = v[\alpha(y - Bs) + \beta v]^T,
\] (7.13)

and the rank will not be increased. However, symmetry dictates we must choose \( \alpha = 0 \), which implies

\[
R = \beta vv^T.
\] (7.14)
It follows directly that

\[ \beta = -\frac{(y - Bs)^T s}{(v^T s)^2}. \quad (7.15) \]

Substituting (7.15) in (7.14) yields a rank-one matrix \( R \) which when added to (7.11) yields the Dennis class of rank-two symmetric secant updates (7.1).

### 7.1.3 Extension of the Dennis Class

Now we are motivated to consider the interesting task of adding a rank-one matrix, call it \( M \), to the Dennis class (7.1) so that \( B_{DC}^T + M \) remains rank-two, symmetric, and satisfies the secant equation. The reason we consider a rank-one matrix is because we expect to be able to solve this problem in simple closed form. The following theorem details the result of this challenge.

**Theorem 7.1.3.** Consider the Dennis class (7.1) where \( B \in \mathbb{R}^{n \times n} \) is symmetric, \( y, s, v \in \mathbb{R}^n \) and \( v^T s \neq 0 \). The only rank-one matrix that can be added to the Dennis class such that the formula remains rank-two, symmetric, and satisfies the secant equation is of the form

\[ M = \theta mm^T \text{ where } m = \frac{y - Bs}{(y - Bs)^T s} - \frac{v}{v^T s} \text{ and } \theta \in \mathbb{R}. \]

The result is given by

\[ B_{EDC}^+ = B_{DC}^+ + M \]

\[ = B + \left( \frac{y - Bs}{v^T s} v^T s + \frac{v(y - Bs)^T}{v^T s} - \frac{s^T(y - Bs)vv^T}{(v^T s)^2} \right) \]

\[ + \theta \left[ \frac{y - Bs}{s^T(y - Bs)} - \frac{v}{v^T s} \right] \left[ \frac{y - Bs}{s^T(y - Bs)} - \frac{v}{v^T s} \right]^T. \quad (7.16) \]
Proof. Consider

\[
B_{+}^{E_{DC}} = B_{+}^{DC} + M \\
= B + \frac{(y - Bs)v^T}{v^T s} + \frac{v(y - Bs)^T}{v^T s} + \theta_{mm^T}.
\]

To not increase the rank, \( m \) must be a linear combination of \( v \) and \( y - Bs \), i.e., \( m = \alpha(y - Bs) + \beta v \). To satisfy the secant equation, \( B_{+}^{E_{DC}} \) must satisfy

\[
B_{+}^{E_{DC}} s = (B_{+}^{DC} + M) s = y
\]

which implies \( Ms = 0 \), that is, \( M \) must have \( s \) in its nullspace. If \( s \) is in the nullspace of \( M = \theta_{mm^T} \), it follows immediately that \( s \) must be orthogonal to \( \alpha(y - Bs) + \beta v \), which leads to

\[
0 = s^T[\alpha(y - Bs) + \beta v]
\\
= \alpha(s^T y) - \alpha(s^T Bs) + \beta(s^T v).
\]

(7.17)

If we let \( \alpha = 0 \), we see that \( \beta(s^T v) = 0 \). Yet, we know \( s^T v \neq 0 \), hence \( \beta = 0 \). Also observe that if we let \( \beta = 0 \), then \( \alpha s^T (y - Bs) = 0 \). However, we know \( s^T (y - Bs) \neq 0 \), hence \( \alpha = 0 \). In either of these cases, we obtain \( M = 0 \). In the nontrivial case \( \alpha, \beta \neq 0 \), we can solve (7.17) for \( \beta \):

\[
\beta = -\frac{\alpha s^T (y - Bs)}{s^T v}
\]

for any \( \alpha \in \mathbb{R} \). Substituting this \( \beta \) in \( m = \alpha(y - Bs) + \beta v \) allows us to construct \( M = \theta_{mm^T} \):

\[
M = \theta \left[ \frac{(y - Bs)}{s^T (Bs - y) - v^T s} \right] \left[ \frac{(y - Bs)}{s^T (Bs - y) - v^T s} \right]^T
\]

(7.18)
where $\theta$ absorbs the constant $\alpha^2$. Adding the rank-one matrix (7.18) to the Dennis class yields (7.16).

We were motivated to call the class (7.16) the ‘extended Dennis class’ until we realized that this is exactly the class of updates that Davidon [31] introduced in 1975.\textsuperscript{77} In addition, Schnabel [106], in 1977, presented this class and referred to it as the $v$-class. For the remainder of this chapter, we choose to refer to (7.16) as the Davidon class since we believe it was Davidon who first introduced this class. It is interesting that extending the Dennis class by adding a rank-one matrix so as to not increase the rank and preserve symmetry and the secant equation leads uniquely to the Davidon class. Clearly, if we set $\theta = 0$ in the Davidon class (7.17), it reduces to the Dennis class (7.1).

### 7.2 $u$-class

In 1977, Schnabel [106] demonstrated that a portion of the Davidon class (7.16) can be represented in what he calls the $u$-form defined by

\begin{align*}
B^U_t &= B + \frac{(y - Bs)u^T + u(y - Bs)^T}{u^Ts} - \frac{(y - Bs)^T s(uu^T)}{(u^Ts)^2} \\
u &= v + \sigma(y - Bs).
\end{align*}

We choose to call (7.19) the $u$-class, and recognize it as the Dennis class (4.13) with the parameter $v$ replaced by $u = v + \sigma(y - Bs)$.

\textsuperscript{77}To learn more about Davidon’s treatment of this class, we refer the interested reader to [31].
We briefly mention a special case of the \( u \)-class because it will be referred to in upcoming sections. For \( \omega > 0 \), consider the special \( u \)-form

\[
B_{+}^{SU} = B + \frac{(y - Bs)u^T + u(y - Bs)^T}{u^Ts} - \frac{(y - Bs)^Ts(uu^T)}{(u^Ts)^2}
\] (7.20)

\[u = y + \sigma(y - Bs),\]

\[\sigma = \frac{-y^Ts(1 \pm \omega \frac{1}{2})}{(y - Bs)^Ts}.\] (7.21)

We refer to this class as the special \( u \)-class.

### 7.3 \( d \)-class

In 1973, Brodlie, Gourlay, and Greenstadt [10] showed that the symmetric rank-two correction matrix \( B_+ - B \) may be expressed as the difference \( ww^T - zz^T \) of rank-one matrices where \( w, z \neq 0 \) are linearly independent, if and only if, \( B_+ - B \) is indefinite.\(^{78}\) Updates which can be represented in this \( d \)-form

\[
B_+ = B + ww^T - zz^T
\] (7.22)

and satisfy the secant equation are members of what we call the \( d \)-class.

Brodlie, Gourlay, and Greenstadt [10] showed that, in the case that \( y^Ts \neq s^T Bs \), the Broyden Class can be written as

\[
B_+ = B + \frac{(y - Bs)(y - Bs)^T}{(y - Bs)^Ts} - \xi \frac{uu^T}{(y - Bs)^Ts}
\] (7.23)

\(^{78}\)In 1976, Gay added to the results of Brodlie, Gourlay, and Greenstadt and showed that correction matrices that are indefinite possess additional characteristics, for example the rank-two correction \( B_+ - B \) has one positive and one negative eigenvalue. As a result, these correction matrices appear to be best in some sense. To learn more, see [52].
where \( \xi = (1 - \theta)y^Ts + \theta s^T B s \). In addition, if \( \xi > 0 \) then (7.23) can be written in the \( d \)-form (7.22). If \( \xi = 0 \), then (7.23) reduces to the SR1 update (4.8). In the case that \( y^Ts = s^T B s \), Brodlie, Gourlay, and Greenstadt observed that any member of the Broyden class can be written in the \( d \)-form.

### 7.4 Schnabel’s Observation

We now present some of Schnabel’s results from his Ph.D. thesis [106]. Schnabel let \( S \) be the set of \( \theta \) for which the Broyden class can be written in the \( d \)-form (7.22). Using this \( S \), defined as \( S = \{ \theta : 1 + \theta \frac{(y-Bs)^T s}{s^T B s} > 0 \} \), he proved that there exists a one-to-one and onto mapping from members of the Broyden class for which \( \theta \in S \), to members of the special \( u \)-class (7.20) for which \( \omega \in (0, \infty) \), with \( B_+ = B_+^{SU} \) for \( \omega = \frac{s^T B s}{(s^T B s) + \theta (y-Bs)^T s} \). This means that each member of the Brodyen class which can be written in the \( d \)-form (7.22) can also be written in the special \( u \)-form (7.20). In particular, they can be written in the \( u \)-form (7.22) with the following two choices of \( \sigma \):

\[
\sigma = \frac{y^Ts}{(y-Bs)^T s} \left[ -1 \pm \left( \frac{s^T B s}{s^T B s + \theta (y-Bs)^T s} \right)^{\frac{1}{2}} \right]
\]

and therefore have two distinct representations in the \( u \)-class (7.22). Letting \( b = y^Ts \) and \( c = s^T B s \), the following are well-known choices of the parameter \( u \):

- **DFP** \( u = y \) or \( y - \frac{2b}{b-c}(y-Bs) \)
- **BFGS** \( u = y + \frac{-b \pm \sqrt{bc}}{b-c}(y-Bs) \)
In the next section, we prove a conjecture that Schnabel presented regarding the BFGS secant update as a result of this observation.

### 7.5 Schnabel’s Conjecture

Schnabel [106] stated that when a secant method converges superlinearly, we should expect

$$\frac{y^Ts}{s^TBs} \to 1.$$  

This follows directly from the 1974 analysis of Dennis and Moré [35], which shows that we can expect the values $\frac{y^Ts - s^TBs}{y^Ts}$ at successive iterations to approach zero. Accordingly, asymptotically only one of each of the two choices of the parameter $u$ in the convex class presented in the previous section yields a numerically stable update, in the sense that you do not get division by zero asymptotically. Letting $b = y^Ts$ and $c = s^TBs$, the choices of parameter $u$ that yield numerically stable DFP and BFGS updates are given by

$$\text{DFP} \quad u = y$$

$$\text{BFGS} \quad u = y + \frac{-b + \sqrt{bc}}{b - c} (y - Bs).$$  \hfill (7.24)

From this, Schnabel conjectured the following.

**Conjecture 7.5.1.** (Schnabel, 1977) If the BFGS method converges, then its scale (7.24) converges to the average of the scales of the DFP update ($u = y$) and the Greenstadt update ($u = Bs$).

We prove Schnabel’s conjecture.
Proof. Begin with the numerically stable scale of the BFGS update (7.24) which we can write as

\[ u = y - \frac{\sqrt{y^T s}}{\sqrt{y^T s + s^T B s}} (y - Bs). \]

From Byrd, Tapia and Zhang [23], we know that if the BFGS method (with steplength \( \alpha = 1 \)) converges, then the convergence is superlinear. Contreras and Tapia [28] showed that if the secant method in question converges superlinearly, then we have

\[ \frac{y^T s}{s^T B s} \to 1. \]

Combining the results of Byrd, Tapia and Zhang with those of Contreras and Tapia, we obtain

\[ u = y - \frac{\sqrt{y^T s}}{\sqrt{y^T s + s^T B s}} (y - Bs) \to \frac{1}{2} (y + Bs). \]

This demonstrates that the scale of the BFGS update can be viewed as (asymptotically) the average of the scales of the DFP update \( u = y \) and the Greenstadt update \( v = Bs \).

In the next section, we describe how Tapia was able to prove a stronger result about the BFGS update using complete updates and not just the scales of updates.

### 7.6 Tapia’s Discovery

In 1984, Tapia proved many interesting properties of the BFGS update in an unpublished paper entitled “On Averaging and Representation Properties of the BFGS and Related Se-
cant Updates.” He considered the kernel

\[ A(v) = \frac{vv^T}{v^Ts} \]

for \( A : \mathbb{R}^n \to \mathbb{R}^{n \times n} \) and \( v^Ts \neq 0 \). Using this, Tapia was able to show

\[
A(y) - A(Bs) = A'[\theta y + (1 - \theta) Bs](y - Bs) \quad \text{for some } \theta \in (0, 1)
\]

\[
= \int_0^1 A'[\theta y + (1 - \theta) Bs](y - Bs) d\theta
\]

\[
= \sum_{k=1}^{\infty} \frac{1}{k!} A^k(Bs)(y - Bs, \ldots, y - Bs).
\]

Note that \( A' \) represents the first derivative of \( A \) and \( A^k \) represents the \( k \)th derivative of \( A \).

These tools allowed Tapia to represent the BFGS updates as a member of each of the \( u \)-, \( d \)-, and Davidon classes.

Tapia began with the known fact that the BFGS could be represented as a member of the \( d \)-class to write

\[
BFGS = A(y) - A(Bs) = \frac{yy^T}{y^Ts} - \frac{s^TBB^Ts}{s^Ts Bs}
\]

for \( s^T Bs, y^T s \neq 0 \): this representation of the BFGS was known. Using Taylor’s Theorem, Tapia wrote the BFGS update as the following member of the Davidon class (7.16)

\[
\sum_{k=1}^{\infty} \frac{1}{k!} A^k(Bs)(y - Bs, \ldots, y - Bs).
\]

(7.25)

He used the Mean Value Theorem (MVT) to write the BFGS update as a member of the \( u \)-class (7.19):

\[
A(y) - A(Bs) = A'[\theta y + (1 - \theta) Bs](y - Bs)
\]

(7.26)
for some $\theta$ where $0 < \theta < 1$. We know that the MVT does not necessarily hold for non-scalar valued functions; however, Tapia’s lengthy calculations demonstrated that the MVT does hold for the BFGS update for $\theta = \frac{- (B_s)^T s \pm \sqrt{(B_s)^T s y T s}}{(y - B_s)^T s}$ which has meaning if and only if $y^T s > 0$. Tapia then turned to the integral form of the MVT to obtain

$$ A(y) - A(B_s) = \int_0^1 A'(\theta y + (1 - \theta) B_s) (y - B_s) d\theta. \quad (7.27) $$

Using this, he calculated the infinite average of all the updates along the line $\theta y + (1 - \theta) B_s$ where $y$ and $B_s$ are the extreme points of this line. Tapia proved that the BFGS is the average of all updates between the extreme points: $\theta = 0$ (which yields the DFP update) and $\theta = 1$ (which yields the Greenstadt update). He accomplished this by working with complete updates and not the scales of updates as Schnabel did.

### 7.7 Huang Class

As we have shown, several authors introduced general classes of updating formulae. Another example is Huang [71], who in 1970, made every effort to keep the class as general as possible to the point of not requiring the updates to be symmetric. Huang considered updates in the general form

$$ H_+ = H + \rho \left[ \frac{s (C_1 s + C_2 H^T y)^T y}{(C_1 s + C_2 H^T y)^T y} \right] - \frac{H y (K_1 s + K_2 H^T y)^T}{(K_1 s + K_2 H^T y)^T} \quad (7.28) $$

with $H \in \mathbb{R}^{n \times n}$ where $\rho, C_1, C_2, K_1, K_2$ are constants, and with the restriction that $K_1$ and $K_2$ must not vanish simultaneously. While the choice of the constants in the Huang class (7.28) were arbitrary, and different choices yield different updates, Huang did intend for
the constants $C_1, C_2, K_1, K_2$ to remain the same from iteration to iteration. The following are the particular choices of $C_1, C_2, K_1, K_2$ for well-known updates pertaining to the case $\rho = 1$:

DFP $\quad C_1 = 1, \quad C_2 = 0, \quad K_1 = 0, \quad K_2 = 1$

SR1 $\quad C_1 = 1, \quad C_2 = -1, \quad K_1 = 1, \quad K_2 = -1$.

Allowing the constant $C_1$ to change between iterations leads to

BFGS $\quad C_1 = -1 - \frac{y^T H y}{s^T y}, \quad C_2 = -1, \quad K_1 = 1, \quad K_2 = 0$.

In 2002, Hull [72] showed that requiring the Huang class (7.28) to produce symmetric updates yields the Broyden class. In particular, it yields

$$H_+ = \left[ H_+^{\text{BFGS}} + \gamma \left( \frac{y^T H y}{s^T y} \right) H_+^{\text{DFP}} \right] \frac{1}{1 + \gamma \left( \frac{y^T H y}{s^T y} \right)}$$

(7.29)

where $\gamma = \frac{K_1}{K_2}$. This formula (7.29), for arbitrary values of $\gamma$, can be recognized as Fletcher’s [45] parameterization of the Broyden class of updates (4.37), i.e., the linear combination of the BFGS inverse update $H_+^{\text{BFGS}}$ and the DFP inverse update $H_+^{\text{DFP}}$ (4.41) where

$$\theta = \frac{1}{1 + \gamma \left( \frac{y^T H y}{s^T y} \right)}.$$

This demonstrates that the Huang class is a more general class than the Broyden class; it subsumes the Broyden class.
7.8 General Form for Symmetric Rank-2 Secant Updates

Our study of multiple update classes motivated us to write the general formula that could be used to represent all symmetric rank-two secant updates. In this section, we derive the formula (7.30) found in Brodlie, Gourlay, and Greenstadt [10], and discuss particular parameter choices that yield well-known updates and update classes.79

Theorem 7.8.1. Any symmetric rank-two update matrix can be written in the form

\[ B_+ = B + \alpha uu^T + \beta (uv^T + vu^T) + \delta vv^T \] (7.30)

for specific values of \( \alpha, \beta, \delta \in \mathbb{R} \) and \( u, v \in \mathbb{R}^n \) where \( B \in \mathbb{R}^{n\times n} \) is symmetric and \( u \) is not a scalar multiple of \( v \).

Proof. Consider the singular-value decomposition (SVD) of a symmetric rank-two matrix \( B \) given by

\[ B = \sigma_1 L_1 L_1^T + \sigma_2 L_2 L_2^T \] (7.31)

where \( \sigma_1, \sigma_2 \) represent the singular values of \( B \), and \( L_1, L_2^T \) represent the left- and right-singular vectors of \( B \) respectively, and \( L_1^T L_2 = 0 \). Our challenge is to find a formula for all rank-two update matrices \( B_+ \) such that the \( \text{Range}(B) = \text{span}\{u, v\} \) where \( u, v \in \mathbb{R}^n \).

Towards this end, it follows directly that the \( \text{span}\{u, v\} = \text{span}\{L_1, L_2\} \), which leads us to consider

\[ L_1 = \phi_1 u + \phi_2 v \quad L_2 = \phi_3 u + \phi_4 v \]

79In their 1973 paper, Brodlie, Gourlay, and Greenstadt were concerned with the conditions in which \( H_+ \) in formula (7.30) could be written in product form.
for \( \phi_i \in \mathbb{R} \). Now observe that substituting \( L_1 \) and \( L_2 \) in the SVD of \( B \) (7.31) gives us

\[
B = \sigma_1 (\phi_1 u + \phi_2 v)(\phi_1 u + \phi_2 v)^T + \sigma_2 (\phi_3 u + \phi_4 v)(\phi_3 u + \phi_4 v)^T
\]

\[
= [\sigma_1 \phi_1^2 + \sigma_2 \phi_3^2] uu^T + [\sigma_1 \phi_1 \phi_2 + \sigma_2 \phi_3 \phi_4] (uv^T + vu^T) + [\sigma_1 \phi_2^2 + \sigma_2 \phi_4^2] vv^T
\]

which can be written using the symmetric rank-two update matrix (7.30) with \( \alpha = \sigma_1 \phi_1^2 + \sigma_2 \phi_3^2 \), \( \beta = \sigma_1 \phi_1 \phi_2 + \sigma_2 \phi_3 \phi_4 \), and \( \delta = \sigma_1 \phi_2^2 + \sigma_2 \phi_4^2 \). This completes the proof.

**Remark 7.8.2.** For the symmetric rank-two update matrix (7.30) to be a secant update, it must also satisfy the secant equation \( B + s = y \), that is, the update must satisfy the condition

\[
\alpha u^T s + \beta (v^T s u + u^T s v) + \delta v^T s v = y - Bs
\]

(7.32)

where \( u^T s, v^T s \neq 0 \).

An obvious way to obtain secant updates from (7.30) is to let \( u = y - Bs \) and choose \( \alpha, \beta, \delta \) so that

\[
\alpha u^T s + \beta v^T s = 1
\]

\[
\beta u^T s + \delta v^T s = 0.
\]

(7.33)

The choice \( u = y - Bs \) is a convenient way to ensure that the update satisfies the secant equation. In the remainder of this section, we present parameter choices that yield the Davidon class, the Dennis class, and the BFGS secant update. We first state parameter choices that yield the Davidon class given \( u = y - Bs \).

**Corollary 7.8.3.** If we let \( \alpha = \frac{\theta}{(v^T s)^2}, \beta = \frac{1}{v^T s} - \frac{\theta}{u^T s u^T s}, \delta = \frac{-u^T s}{(v^T s)^2} + \frac{\theta}{u^T s} \) with \( u = y - Bs \) in the symmetric rank-two update formula (7.30) then we obtain the Davidon class (7.16).
We now state that there are two different ways to obtain the Dennis class from the symmetric rank-two update formula (7.30).

**Corollary 7.8.4.** If we let \( \alpha = 0 \), \( \beta = \frac{1}{v^T s} \), \( \delta = -\frac{v^T s}{(v^T s)^2} \) with \( u = y - Bs \) in the symmetric rank-two update formula (7.30), then we obtain the Dennis class (7.1).

**Corollary 7.8.5.** If we let \( \alpha = \delta = 0 \) and \( \beta = 1 \), and set \( u = \frac{y - Bs}{v^T s} - \frac{s^T (y - Bs)v}{2(v^T s)^2} \) in the symmetric rank two update formula (7.30), then we obtain the Dennis class (7.1).

We end the section by discussing parameter choices that yield the BFGS secant update.

In an effort to produce compact notation, i.e., “minimize clutter,” a moment’s reflection on equation (7.32) should motivate the thoughtful reader to suggest the choice \( \beta = 0 \). Once we choose \( \beta = 0 \) in (7.30), then for the update to satisfy the secant equation, the condition

\[
y - Bs = \alpha (u^T s) u + \delta (v^T s) v
\]

must be satisfied. The obvious choice here is \( u = y \) and \( v = Bs \). Moreover, this choice maintains a sense of balance, since we expect \( y^T s \) and \( s^T Bs \) to be of the same magnitude.

Our reasoning has led us to the BFGS update in a most direct manner. Hence, we feel that the BFGS update is a natural update and would be postulated by most readers once they were exposed to the general update formula (7.30). We now give a formal statement of our insight.

**Corollary 7.8.6.** If we let \( u = y \) and \( v = Bs \), \( \alpha = \frac{1}{v^T s} \), \( \beta = 0 \), and \( \delta = -\frac{1}{s^T Bs} \), in the symmetric rank-two update formula (7.30), then we obtain the BFGS secant update (4.19).
7.8.1 Extended Dennis-Davidon Class

We end this chapter by deriving what we call the Extended Dennis-Davidon Class, i.e., we set \( u = y - Bs \) in (7.30) and solve for the parameters so that the secant equation is satisfied. Recall from Remark 7.8.2, that for the symmetric rank-two update matrix (7.30) to satisfy the secant equation, it must satisfy the condition (7.32) where \( u^T s, v^T s \neq 0 \). If we set \( u = y - Bs \) in (7.32), we obtain

\[
\alpha (y - Bs)^T s(y - Bs) + \beta (v^T s(y - Bs) + (y - Bs)^T sv) + \delta v^T s v = y - Bs. \quad (7.34)
\]

If \( v \) and \( y - Bs \) are linearly independent, we can solve (7.34) for \( \alpha \) and \( \delta \) in terms of \( \beta \) to obtain

\[
\alpha = \frac{1 - \beta v^T s}{(y - Bs)^T s},
\delta = -\frac{\beta (y - Bs)^T s}{v^T s}
\]

for any \( \beta \in \mathbb{R} \). Substituting these values for \( \alpha \) and \( \delta \) in the symmetric rank-two update matrix (7.30) yields the extended Dennis-Davidon class

\[
B_+ = B + \frac{1 - \beta v^T s}{(y - Bs)^T s}(y - Bs)(y - Bs)^T + \beta [(y - Bs)v^T + v(y - Bs)^T] - \frac{\beta(y - Bs)^T s}{v^T s} vv^T.
\]

(7.35)

Remark 7.8.7. What we see interesting about the Dennis class and the Davidon class is that they use the choice \( u = y - Bs \) in the general formula. Moreover, we have the following interesting and surprising result.

Theorem 7.8.8. The Davidon class is a maximal class for the choice \( u = y - Bs \). That
is, any member of the general secant update class which satisfies the secant equation and uses the choice \( u = y - Bs \) must be a member of the Davidon class.

Proof. Consider the update formula (7.30) with the choice \( u = y - Bs \). The conditions (7.32), assuming \( y - Bs \) and \( v \) are linearly independent, allows us to write \( \alpha \) and \( \delta \) in terms of \( \beta \). This leads to what we might call the extended Dennis-Davidon class (7.35). However, equating coefficients in (7.35) with those in the Davidon class (7.16) demonstrates the relationship

\[
\beta = \frac{1}{v^T s} - \frac{\theta}{(y - Bs)^T sv^T s}.
\]

Hence, the Davidon class (7.16) and our so-called extended Dennis-Davidon class (7.35) are merely reparameterizations of the same class. ■
Bibliography


[100] ———, How Bad are the BFGS and DFP Methods When the Objective Function is Quadratic, Mathematical Programming, 34 (1986), pp. 34–47.


