ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF NATURAL SCIENCES

DEPARTMENT OF MATHEMATICS

NEWTON'S METHOD FOR SOLVING OPTIMIZATION PROBLEMS

A SEMINAR PRESENTED TO THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABABA UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN MATHEMATICS

BY: Simon Derkee

Advisor: Prof. Dr.rer.nat.habil.R.Deumlich

June 2010

Acknowledgement

First of all I thank to my God, because preparing and collecting this seminar report would n't have been possible without his help. And next, I am heartily thankful to my advisor, Prof.dr.rer.nat.habil.R.Deumlich, whose encouragement, guidance, and support from the initial to the final enabled me to develop an understanding of the subject and finally to prepare and organize this paper successfully. Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of the paper.

Simon Derke

Preface

Mainly this paper considers three nonlinear problems; which are solving systems of nonli near equations, unconstrained minimization of nonlinear functional and nonlinear constrained Minimization. In order to solve such problems many numerical methods have been developed. But in this paper we consider only the so called Newton's method.

The structure of a system of nonlinear equations problem is of the following form and is considered in chapter two. Given $F : \mathbf{R}^n \to \mathbf{R}^m$ be continuously differentiable, then find $\mathbf{x}^* \in \mathbf{R}^n$ such that $F(\mathbf{x}^*) = \mathbf{0}$.

The unconstrained minimization problem of nonlinear functional is given in the following way. And it is contained in chapter three. Given $f : \mathbf{R}^n \to \mathbf{R}$, be twice continuously differentiable, then find $\mathbf{x}^* \in \mathbf{R}^n$ such that $f(\mathbf{x}^*) \le f(\mathbf{x})$ for every $\mathbf{x} \in \mathbf{R}^n$.

Another optimization problem which we consider in this paper is nonlinear constrained minimization, which is treated in chapter four. This problem is of the following type. Given $f : \mathbf{R}^n \to \mathbf{R}$, be twice continuously differentiable nonlinear functional, and $g_i : \mathbf{R}^n \to \mathbf{R}$, $i = 1,2,3,...,m$, $h_j : \mathbf{R}^n \to \mathbf{R}$, $j = 1,2,3,...,k$ are all twice continuously differentiable nonlinear functionals. And S is a subset of $\mathbf{R}^n$ which contains all points satisfying the constraints. Then the problem is

$f(\mathbf{x}) \to \min$, $\mathbf{x} \in S$.

The first chapter discusses about some algorithms for nonlinear problems in one variable. Even though it is not the heart of this paper, it helps us to observe easily basic philosophy of the multivariable nonlinear algorithms which we consider in three subsequent chapters.

The last chapter discusses about constrained minimization problems. In this chapter we consider only the Karush Kuhn Tucker conditions as a necessary condition for a constrained minimization problem and we will not discuss about further properties of the constraints, for instance constraint qualification and so on. We are only going to apply the Newton's method and to produce the candidate for the solution.

Finally it is very important in this paper to understand the following remarks when we are solving the problems using the announced Newton method.

1) The question of existence and uniqueness does a given problem have a solution and is it unique? Is beyond the capabilities one can expect of algorithms that solve nonlinear problem. The methods that we develop can only find one approximate root of a nonlinear system of equations and one approximate local minimum of a nonlinear functional.
2) Finite precision arithmetic has effect on the algorithms to solve our problems. To see the details of this fact refer [4].

Contents Pages

Introduction

In almost all fields of analysis and also in other fields of mathematics in particular in applied mathematics we have the situation to solve the system of equations of n-real variables. To have a solution or all solutions of a system of equations belongs to the important questions of the mathematics.

This is so, because a lot of problems can be reduced to solving a system of equations. An important example for this situation can be found in optimization theory, where we can transfer an optimization problem by means of Karush-Kuhn-Tucker conditions in a problem to solve a system of equations. In the case of a linear system of equations there are some algorithms (for instance the Gauss algorithm or other iteration algorithms) to solve it. Also in this case if the number of variables is large, then such systems can also be difficult to solve.

In the case of the system nonlinear equations the situation becomes very complicated and in some cases it can be hop less to solve it, if we don't suppose anything. There for in the case of nonlinearity we will at least assume in this paper that the involved functions are partially continuously differentiable on an open set $D \subset \mathbf{R}^n$. But even then, it remains a complicate problem to find a solution. For some special classes of functions the announced Newton's method is working in a reasonable way.

Newton's method is one of the most powerful and well known numerical methods for solving a root finding problem. The method is to approximate roots of functions. I.e. solutions of the form $f(x) = 0$. Not only is the method easy to comprehend; it is a very efficient way to find the solution of the equation.

The idea is that: one starts with an initial guess which reasonably close to the true root, then the function is approximated by its tangent line (which can be computed by using the tools of calculus) and one computes the x-intercept of this tangent line. This x-intercept will typically be a better approximation to the function's root than the original guess.

1. Newton's method for a Nonlinear Equations and Unconstrained Minimization in One Variable

1.1. Solving Nonlinear Equations in one Variable

Finding the zeros of a given function f , that is the argument ε for which $f(\varepsilon) = 0$, is a classical problem. To solve such a problem we have different numerical methods. From these methods we are going to see the one which is considered in this paper, which is called the Newton's method. The Newton method is an iterative method of solving non linear equations.

1.1.1. Newton's Method

Let **R** denotes the set of all real numbers and $a, b \in \mathbf{R}$. Suppose $f : [a,b] \to \mathbf{R}$ be a real valued function, which is differentiable on the interval $[a,b] \subseteq \mathbf{R}$. The formula for converging on the root can be easily derived. Suppose we have a current approximation x_0, then we can derive the formula for a better approximation x_{n+1} as follows. If our initial or current estimate of the answer is x_0, we get a better approximation or estimate x_1, by drawing a tangent line to $f(x)$ at $(x_0, f(x_0))$ and finding the point x_1, where this line crosses the x-axis. The following figure clarifies this procedure.

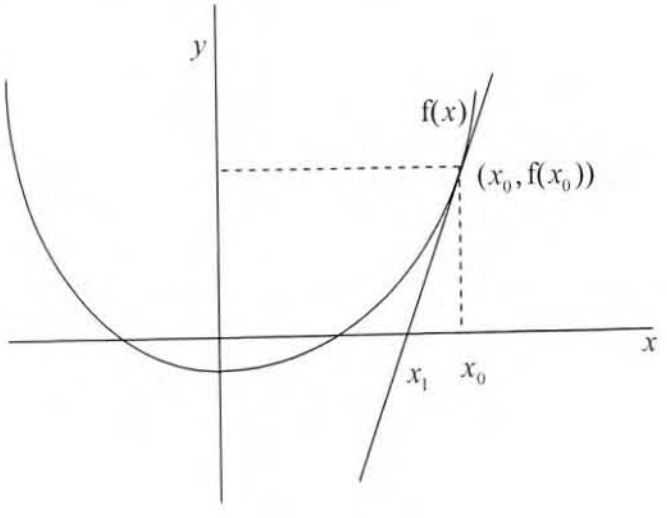

Figure (1.1.1)

Since $x_1 = x_0 - \Delta x$ and $f'(x_0) = \frac{\Delta y}{\Delta x} = \frac{f(x_0)}{\Delta x}$. From this we get

$f'(x_0) = \dfrac{f(x_0)}{x_0 - x_1}$, which gives us

$x_0 - x_1 = \dfrac{f(x_0)}{f'(x_0)}$, where $f'(x_0) \neq 0$. Finally we have

$x_1 = x_0 - \dfrac{f(x_0)}{f'(x_0)}$.

From the graph above also we can see that x_1 is better approximation than x_0. And if this is not close enough to the real solution, use the method again on x_1 to get a better approximation x_2.

$x_2 = x_1 - \dfrac{f(x_1)}{f'(x_1)}$, where $f'(x_1) \neq 0$. And similarly for arbitrary n,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{1.1.1}$$

Iterate the method until the required accuracy is obtained. The method just we have developed is called ***Newton Raphson*** method or simply Newton's method. What we have done is at each iteration we have constructed a local model (which is affine) of our function f and solved for the root of the model. Let

$$M_c(x) := f(x_c) + f'(x_c)(x - x_c). \tag{1.1.2}$$

The model (1.1.2) is a unique line with function value $f(x_c)$ and slope $f'(x_c)$ at the point x_c. And we will refer to the above model (1.1.2) as an affine model.

Generalized Newton's Method

If ε is a root of $f(x) = 0$ with multiplicity p, then the iteration formula corresponds to (1.1.1) becomes,

$$x_{n+1} = x_n - p\frac{f(x_n)}{f'(x_n)}. \tag{1.1.3}$$

This means that $1/p\, f'(x)$ is the slope of the straight line passing through (x_n, y_n) and intersecting the x-axis at $(x_{n+1}, 0)$. Equation (1.1.3) is called the generalized Newton's formula and reduces to (1.1.1) for $p = 1$. Since ε is also a root of $f(x) = 0$ with multiplicity p. It follows that ε is also a root of $f'(x) = 0$ with multiplicity $p-1$, of $f''(x) = 0$ with multiplicity $p-2$, and so on. Hence expressions

$$x_0 = x_0 - p\frac{f(x_0)}{f'(x_0)},\ x_0 = x_0 - (p-1)\frac{f'(x_0)}{f''(x_0)},\ x_0 = x_0 - (p-2)\frac{f''(x_0)}{f'''(x_0)}$$

must have the same value if there is a root with multiplicity p, provided that the initial approximation x_0 is chosen sufficiently close to the real root.

Unlike the present case, the geometric derivation of such models becomes less manageable in multivariable problems. Thus we require other type of derivations. From basic calculus theorem, we have

$$f(x) = f(x_c) + \int_{x_c}^{x} f'(z)dz \tag{1.1.4}$$

And it seems reasonable to approximate the indefinite integral in (1.1.3) by

$$\int_{x_c}^{x} f'(z)dz \approx f'(x_c)(x - x_c), \text{ hence}$$

$$f(x) \approx f(x_c) + f'(x_c)(x - x_c) =: M_c(x). \tag{1.1.5}$$

Denotation:

1. In this paper the local model at x_k, for finding an approximate root of non linear equations and an approximate local minimum of a non linear functional are denoted by M_k and m_k, respectively.
2. $\mathbf{R}$ and $\mathbf{R}_+$ represents the set of all real numbers and all nonnegative real numbers, respectively.
3. $\mathbf{R}^n$ and $\mathbf{R}_+^n$ represents the set of all n-tuples of real numbers and nonnegativereal numbers, respectively. Where as $\mathbf{N}$ denotes the set of all natural numbers.

Remark: For a lot of functions this iterative method works well, but not for all functions. For instance it doesn't work for any function that doesn't have a root. Even for those functions does work for, you might choose a bad initial point x_0, for instance $f'(x_0)$ might be zero.

Definition 1.1.1: Let $x_n \in \mathbf{R}$, $n = 0,1,2,3,...$, and also $x \in \mathbf{R}$. Then the sequence $\{x_n\}$ is said to be

a. Convergence and converges to $x \in \mathbf{R}$, if given any $\varepsilon > 0$, there exits a number $n_0 \in \mathbf{N}$ such that $|x_n - x| < \varepsilon$, $\forall n \geq n_0$. Equivalently, we can write it as $x_n \to x$ as $n \to \infty$.
b. Converges linearly to $x \in \mathbf{R}$, if and only if $x_n \to x$ as $n \to \infty$ and there exists a constant $c \in [0,1]$ and an integer $n_0 \geq 0$ such that $\forall n \geq n_0$, $|x_{n+1} - x| \leq c|x_n - x|$.

c. If for some sequence $\{c_k\}$, which converges to zero, $x_n \to x$ and $|x_{k+1} - x| \le c_k |x_k - x|$, then $\{x_k\}$ is said to converge super linearly to x.

d. If there is a constant $c \ge 0$ and an integer $k_0 \ge 0$ such that $\{x_k\}$ converges to x and for all $k \ge k_0$, $|x_{k+1} - x| \le c|x_k - x|^2$, then $\{x_k\}$ is said to converge quadratically to x.

Example 1: Show that

(a) The sequence $\{1 + 2^{-k}\}$ converges linearly to 1,

(b) The sequence $\{1 + 2^{-2^k}\}$ converges quadratically to 1.

For (a) $x_k = 1 + 2^{-k}$ and $x_{k+1} = 1 + 2^{-k+1}$. From this we get

$|x_{k+1} - 1| = |1 + 2^{-k-1} - 1| = \frac{1}{2} 2^{-k} = \frac{1}{2}|2^{-k} - 1| = \frac{1}{2}|x_k - 1|$. Thus we have

$|x_{k+1} - x| \le \frac{1}{2}|x_k - x|$, with the constant $c = \frac{1}{2}$.

There for the given sequence converges linearly to 1.

(b) $x_k = 1 + 2^{-2^k}$ and $x_{k+1} = 1 + 2^{-2^{k+1}}$. From this we get

$|x_{k+1} - 1| = |1 + 2^{-2^{k+1}} - 1| = (2^{-2^{k+1}}) = (2^{-2^k})^2 = (1 + 2^{-2^k} - 1)^2 = |x_k - 1|^2$.

i.e. $|x_{k+1} - x| \le 1|x_k - x|^2$, with the constant $c = 1$.

There for the given sequence converges quadratically to 1.

Definition 1.1.2: Let $f : D \to \mathbf{R}$ be real valued function from an open interval $D \subset \mathbf{R}$, then,

(a) f is said to satisfy a Lipchitz condition at $a \in D$ if and only if there are constants δ and γ such that $x \in D$ and $|x - a| < \delta$ implies $|f(x) - f(a)| < \gamma |x - a|$.

(b) f is said to satisfy a Lipschitz condition on D if and only if there is a constant γ such that for every $x, y \in D$, $|f(x) - f(y)| < \gamma |x - y|$. The constant γ is called a L-ipchitz constant. If f satisfies a Lipchitz condition at a or on D, then f is said to be Lipchitz continuous at a or on D. This can be is written as $f \in Lip_\gamma(D)$.

Lemma 1.1.1: For an open interval $D \subset \mathbf{R}$, let $f : D \to \mathbf{R}$ be a function and $f \in Lip_\gamma(D)$. Then for any $x, y \in D$,

$$|f(y) - f(x) - f'(x)(y - x)| \le \frac{\gamma}{2}|y - x|^2. \qquad (1.16)$$

Proof: From basic calculus we have

$$f(y)-f(x)=\int_x^y f'(z)dz \text{ or equivalently}$$

$$f(y)-f(x)-f'(x)(y-x)=\int_x^y [f'(z)-f'(x)]dz\,. \tag{1.1.7}$$

Making change of variables, $z=x+t(y-x)$, $dz=(y-x)dt$ for $0\le t\le 1$, then (1.1.7) becomes,

$$f(y)-f(x)=\int_0^1 [f'(x+t(y-x))-f'(x)](y-x)dt \text{ or}$$

$$\left|f(y)-f(x)-f'(x)(y-x)\right|=\left|\int_0^1 [f'(x+t(y-x))-f'(x)](y-x)dt\right|$$

$$\le \int_0^1 \left|f'(x+t(y-x))-f'(x)(y-x)\right|dt \le \int_0^1 \left|f'(x+t(y-x))-f'(x)\right|\left|y-x\right|dt$$

$$\le \int_0^1 \left|t(y-x)\right|\left|y-x\right|dt \quad \text{(By Lipschitz continuity)}$$

$$\le \frac{\gamma\left|y-x\right|^2}{2}\,.//$$

Remark: The lemma says that if $f\in Lip_\gamma(D)$. Then we can obtain a bound on how close to $f(y)$ is the value at y of the local model. See the following figure.Look at the following figure to understand the above lemm geometricaly.

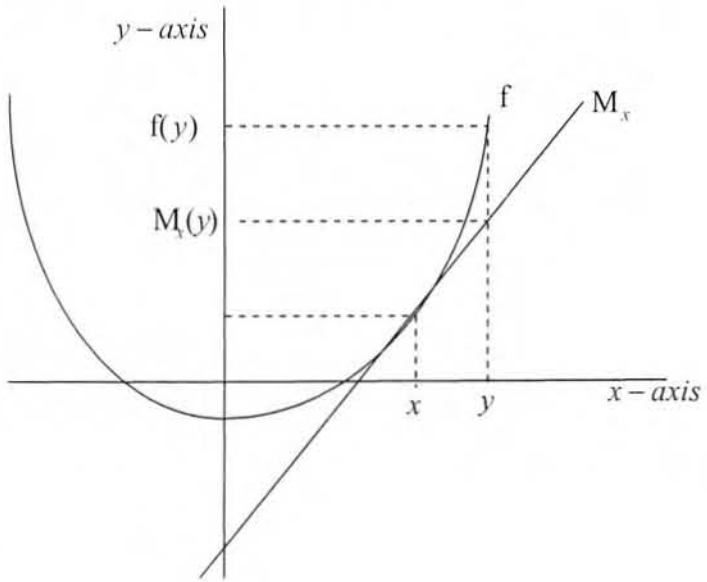

Figure (1.1.2)

The Figure above shows that how close $f(y)$ to $M_x(y)$ at the point y.

Theorem1.1.2: Let $f: D \to \mathbf{R}$ be differentiable on an open interval $D \subset \mathbf{R}$ and $f \in Lip_\gamma(D)$. Assume that for some $\mu > 0$, $|f'(x)| \geq \mu$ for every $x \in D$. If $f(x) = 0$ has a solution $x^* \in D$, then there is some $\eta > 0$ such that the following is true. If $|x_0 - x^*| < \eta$, then the sequence (x_k), generated by $x_{k+1} = x_k - \dfrac{f(x_j)}{f'(x_k)}$, for $k = 0,1,2,...$ exists and converges to x^*. Further more

$$\text{for } k = 0,1,2,..., \quad |x_{k+1} - x^*| \leq \frac{\gamma}{2\mu}|x_k - x^*|^2. \tag{1.1.8}$$

Proof: Let $\tau \in (0,1)$ and η be the radius of the largest open interval around x^*, that is contained in D, and define $\eta := \min\{\eta, \dfrac{2\mu\tau}{\gamma}\}$. We will show by induction that (1.1.6) holds for $k = 0,1,2,...$, and $|x_k - x^*| \leq \tau|x_0 - x^*| < \eta$.

In the proof we will show that, at each iteration the new error $|x_k - x^*|$ is bounded by a constant times the error the affine model makes in approximating x_0 to x^*, which is by lemma (1.1.1) $\dfrac{\gamma|x_k - x^*|^2}{2}$.

Now for $k = 0$,

$$x_1 - x^* = x_0 - \frac{f(x_0)}{f'(x_0)} - x^* = x_0 - x^* - \frac{f(x_0) - f(x^*)}{f'(x_0)}$$

$$= \frac{1}{f'(x_0)}[f(x^*) - f(x_0) - (x^* - x_0)f'(x^*)].$$

But the term in bracket is equal to $f(x^*) - M_0(x^*)$, which is the error at x^* in the local affine model at $x_0 = x_c$. Thus from lemma (1.1.1),

$$|x_1 - x^*| = \left|\frac{1}{f'(x_0)}[f(x^*) - f(x_0) - (x^* - x_0)f'(x^*)\right| = \left|\frac{1}{f'(x_0)}[f(x^*) - M_0(x^*)]\right|$$

$$\leq \frac{\gamma|x^* - x^*|^2}{2|f'(x_0)|}.$$

But we have also by assumption, $|f'(x)| \geq \mu$ thus we get $|x_1 - x^*| \leq \dfrac{\gamma}{2\mu}|x_0 - x^*|^2$. This is (1.1.6) for $k = 0$.

Also from the given, we have that $|f'(x)| \geq \mu$. This implies

$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$ is welldefined.

More over,

$$|x_1 - x^*| \le \frac{\gamma}{2\mu}|x_0 - x^*|^2 = \frac{\gamma}{2\mu}|x_0 - x^*||x_0 - x^*| \le \frac{\gamma}{2\mu}(\frac{2\mu}{\gamma})|x_0 - x^*| = |x_0 - x^*| < \eta .$$

This completes the proof for $k = 0$.

Assume that the theorem also holds for $k = 0,1,2,...,n$ i.e.

$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$ $|x_{k+1} - x^*| \le \frac{\gamma}{2\mu}|x_0 - x^*|^2$ also

$|x_{k+1} - x_*| < \eta$. Then we want to show that it is also true for $k = n+1$. Now

$$x_{n+2} - x^* = [x_{n+1} - \frac{f(x_{n+1})}{f'(x_{n+1})}] - x^* .$$

But by induction assumption, we have that x_{n+1} exists as $f'(x_{n+1}) \neq 0$, thus

$$x_{n+2} = [x_{n+1} - \frac{f(x_{n+1})}{f'(n+1)}] - x^* = x_{n+1} - x^* - \frac{f(x_{n+1})}{f'(x_{n+1})}$$ is well defined.

Hence we have,

$x_{n+2} - x^* = \frac{1}{f'(x_{n+1})}[f(x^*) - f(x_{n+1}) - f'(x_{n+1})(x^* - x_{n+1})]$, And from this we get

$$|x_{n+2} - x^*| = \left|\frac{1}{f'(x_{n+1})}\left[f(x^*) - f(x_{n+1}) - f'(x_{n+1})(x^* - x_{n+1})\right]\right|$$

$$= \frac{1}{|f'(x_{n+1})|}|f(x^*) - M_{n+1}(x^*)| \le \frac{\gamma}{2\mu}|x_{n+1} - x^*|^2 .$$ I.e.

$$|x_{n+2} - x^*| \le \frac{\gamma}{2\mu}|x_{n+1} - x^*|^2 .$$

Again from induction assumption we have,

$|x_{n+1} - x^*| \le \frac{\gamma}{2\mu}|x_0 - x^*|^2$. Also from the first part we have,

$$|x_{n+2} - x^*| \le \frac{\gamma}{2\mu}|x_{n+1} - x^*|^2 \le (\frac{\gamma}{2\mu})^3|x_0 - x^*|^2|x_0 - x^*|^2 \le (\frac{\gamma}{2\mu})^3|x_0 - x^*|(\tau(\frac{2\mu}{\gamma}))^3 < \eta$$

since $|x_0 - x^*| \le \tau(\frac{2\mu}{\gamma})$. This implies $|x_0 - x^*|^2 \le \tau^2(\frac{2\mu}{\gamma})^2$ and also $\tau \in (0,1)$. There for we have $|x_{n+2} - x^*| \le \eta$. This completes the proof. //

Remark:

1. The condition in theorem (1.1.2) is that $f'(x)$ has a none zero lower bound in D is simply means that $f'(x)$ must be nonzero for Newton's method to conve-rges quadratically.

2. From the proof at each iteration, we have that

   $$\left|x_{k+1}-x^*\right| \le \frac{\gamma}{2\mu}\left|x_k-x^*\right|^2, \forall k = 0,1,2,\ldots,.$$

   This implies that the Newton method converges quadratically for a good starting point.

Example: In this example we want to apply the Newton's method to the functions f and g defined by $f(x) = x^2 - 1, x \in \mathbf{R}$ and $g(x) = x^2 - 2x + 1, x \in \mathbf{R}$ with the same starting point $x_0 = 2$. The following table is calculated by means of (Program 1) for the first five iterations.

Iterations for $f(x) = x^2 - 1$		Iterations for $g(x) = x^2 - 2x + 1$	
k	x_k	k	x_k
0	2	0	2
1	1.25	1	1.5
2	1.025	2	1.25
3	1.0003048780388	3	1.125
4	1.00000046411	4	1.065
5	1.0	5	1.03125

Table (1.1.1)

From the table above we can se that the solution is obtained by 5th iteration step. This is due to the starting point is good enough.

Remarks:

1. At each iteration x_k the Newton's method needs that the value $f'(x_k)$ must be different from zero.
2. The method converges for only a good starting point and may not converge at all if $\left|x_0 - x^*\right|$ is large.
3. In some cases f' at x_0, may not exist. In this case it is better to replace this model by the secant line that goes through the points $(x_c, f(x_c))$ and $(x_c + h, f(x_c + h))$, for some nearby point $x_c + h_c$ of x_c. The slope of this line is $a_c = \frac{f(x_c + h_c) - f(x_c)}{x_c + h_c - x_c}, = \frac{f(x_c + h_c)}{h_c}, h_c \neq 0$ and so that the model we obtain is, $M_c(x) := f(x_c) + a_c(x - x_c)$. If h_c is chosen to be small number, a_c is called a finite difference approximation to $f'(x_c)$. To see the details refer [7].
4. If the starting point is not good, the method may not converge at all. Such iterative methods are called locally convergent methods. On the other hand those iterative that guarantee convergence from any starting point are called globally convergent methods or global methods. And Newton's method, finite

difference method, secant methods etc are local methods and backtracking method, bisection method etc are all global methods.

5 If our starting point is close to the true solution, then the method converges very fast to the real root.

1.2. Minimization of a Nonlinear Functional of One Variable

Definition 1.2.1:

a) A function f is said to have a local minimum at x_0, if there is an open interval $S(x_0,\varepsilon)$ of x_0 such that $f(x_0) \le f(x), \forall x \in S(x_0,\varepsilon)$. Similarly f is said to have a local maximum at x_0, if there is an open interval $S(x_0,\varepsilon)$ of x_0 such that $f(x_0) \ge f(x), \forall x \in S(x_0,\varepsilon)$. We say that f has a local extremum at x_0, if f has either local minimum or local maximum at x_0.

b) Let I be an interval in which f is defined. Then we say that f has an absolute maximum at the point x_0 in I, if $f(x) \le f(x_0)$, $\forall x$ in I. And we say that f has an absolute minimum at x_0, if $f(x_0) \le f(x)$ $\forall x$ in I. We say that f has an absolute extremum at x_0, if f has either an absolute maximum or absolute minimum at x_0.

The following two theorems give the necessary and sufficient conditions for some point to be the minimum of the function f of one variable on some open interval.

Theorem 1.2.1: Let D be an open interval in **R**, $f \in C^{(1)}(D)$ and $z \in D$. if $f'(z) \neq 0$. Then for any s with $f'(z)s < 0$, there is a constant $\lambda > 0$ for which $f(z+ts) < f(z)$ for every $t \in (0,\lambda)$.

Remark: Theorem (1.2.1) says that if $f'(z) \neq 0$, then z cannot be a local minimum point of f or equivalently local minimum point of a continuously differentiable function must be obtained at the point where $f'(z) = 0$. That is solving $f'(z) = 0$ is necessary for finding a minimum point of the function f, but not sufficient.

Theorem 1.2.2: let D be an open interval in **R** and $f \in C^{(2)}(D)$ and let $x^* \in D$ for which $f'(x^*) = 0$ and $f''(x^*) > 0$. Then there is some open subinterval D' of D for which $x^* \in D'$ and $f(x^*) < f(x), \forall x \in D$.

Remark: Theorem (1.2.2) gives the second order sufficient condition for a point $x^* \in D$ to be a local minimum point of a function f of one variable.

1.2.1. Newton's method

When we are in need of solving the problem $\min f(x)$, $x \in D$ the necessary condition is $f'(x) = 0$, for a continuously differentiable function f on D .Thus the Newton step becomes,

$$x_N = x_c - \frac{f'(x_c)}{f''(x_c)}, \text{ for } f''(x_c) \neq 0. \tag{1.2.1}$$

Now let the model,

$$m_c(x) := f(x_c) + f'(x_c)(x - x_c) + \frac{1}{2} f''(x_c)(x - x_c)^2 \tag{1.2.2}$$

a quadratic model, and (1.2.1) is derived by making affine model of $f'(x)$ around x_c.
It is equivalent to having made a quadratic model of $f(x)$ around x_c and setting x_N to the critical point of this model. The critical point of the quadratic model is found by solving the problem $m_c'(x) = 0$ for x. Where,

$m_c(x) = f(x_c) + f'(x_c)(x - x_c) + \frac{1}{2} f''(x_c)(x - x_c)^2$, and we have

$m_c'(x) = f'(x_c) + f''(x_c)(x - x_c)$. From this we have,

$m_c'(x) = 0$ implies $f'(x_c) + f''(x_c)(x - x_c) = 0$.

The last equation can be rewritten as

$f''(x_c)(x - x_c) = -f'(x_c)$ .This gives us

$x - x_c = \frac{f'(x_c)}{f''(x_c)}$, hence we have the Newton step becomes,

$$x_N = x_c - \frac{f'(x_c)}{f''(x_c)}. \tag{1.2.3}$$

Remarks:

1. Formula (1.2.1) shows we use the Newton method for minimization only if the derivatives $f'(x)$ and $f''(x)$ exists. But sometimes those derivatives may not be analytically available.In such cases we can use finite difference approximations to the derivatives. To see the details refer [7].
2. Newton's method doesn't specify whether the given solution maximizes or minimizes the problem, i.e. the solution may approach toward a maximizer or a minimizer or a saddle point of f. Equivalently each step simply goes to the critical point of the current local quadratic model and one can be sure by looking at the value to satisfy $f'' > 0$.

The details of minimization using Newton's method will be discussed in chapter three.

2. Newton's Method for a System of Nonlinear Equations

This chapter discusses the local algorithm (Newton's method) for a system of nonlinear equations. In the first section we summarize some topics in Linear Algebra and multivariable calculus which are needed to implement and analyze Newton's method for a system of nonlinear equations. In the second section we describe the Newton's method for a system of nonlinear equations. Finally we will try to see some examples of the method applied to system of nonlinear equations using computer programs.

2.1. Introduction

2.1.1. Vector and Matrix Norms

Denotations:

1. In this paper we denote the set of all (nxm)-matrices by $\mathbf{R}^{nxm}$. And for $\mathbf{A} \in \mathbf{R}^{nxm}$, $\mathbf{A}=(\mathbf{a}_{ij})$, i=1,2,3,…,n, j=1,2,3,…,m and we denote the identity matrix by $\mathbf{I}$.
2. $\nabla \mathrm{f}(\mathbf{x})$ be the gradient vector of a real valued function of several variables.
3. $\dfrac{\partial \mathrm{f}(\mathbf{x})}{\partial x_i}$ be the partial derivative of a real valued function of several variables with respect to the component x_i.

Definition 2.1.1.1: Let $\mathbf{X}$ be arbitrary vector space. A nonnegative real valued function $\|\cdot\| : \mathbf{X} \to \mathbf{R}$, is called a norm if the following holds.

a) $\|\mathbf{x}\| \geq 0$ and $\|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$ for any vector $\mathbf{x}$,

b) $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$, For α is a scalar and $\mathbf{x}$ is a vector,

c) $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ for all vectors $\mathbf{x}, \mathbf{y}$.

The most commonly used vector norms in $\mathbf{R}^n$ are the following.

1. Absolute value norm, $l_1(n)$ norm, $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$
2. Euclidean norm, $l_2(n)$ norm, $\|\mathbf{x}\|_2 = (\sum_{i=1}^{n} |x_i|)^{1/2}$
3. Maximum norm, $l_\infty(n)$ norm, $\|\mathbf{x}\|_\infty = \max_{i=1}^{n} |x_i|$

Now using one of these norms we can define a notion of convergence of a sequence as follows.

Definition 2.1.1.2: A sequence, $\{\mathbf{x}_n\}$ of elements in $\mathbf{R}^n$, is said to be convergent sequence and converges to an element $\mathbf{x}$ in $\mathbf{R}^n$ with regard to the norm, $\|.\|$ if and only if $\|\mathbf{x}_n - \mathbf{x}\| \to 0$, as $n \to \infty$.

Whether a particular algorithm converges, in practice might depend on what norm its stopping criteria. Fortunately in $\mathbf{R}^n$ we have no such problems. See (Theorem 2.1.1).

Definition 2.1.1.3: Two norms $\|.\|_1$ and $\|.\|_2$ are said to be equivalent in a normed space X if and only there exist real numbers α_1 and α_2 such that $\alpha_1 \|.\|_1 \le \|.\|_2 \le \alpha_2 \|.\|_1$, where $0 \le \alpha_1 \le \alpha_2$.

The following theorem is very important in this section which is given without proof, and the proof. The proof is given in [3].

Theorem 2.1.1.1: All norms on a finite dimensional normed space are equivalent.

Corollary 2.1.1.2: All norms in a normed space $\mathbf{R}^n$ are equivalent.

Proof: Since the space $\mathbf{R}^n$ together with one of the norms defined above, is a normed space and also it is of finite dimension, by (Theorem 2.1.1), all of the norms defined in $\mathbf{R}^n$ are equivalent.

The following are the most commonly used matrix norms.

1. Frobenius norm, which is just the $l^2(n)$ norm of an (nxm)-matrix $\mathbf{A} = (\mathbf{a}_{ij})$ is given a follows. $\|\mathbf{A}\|_F := \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |\mathbf{a}_{ij}|^2 \right)^{\frac{1}{2}}$,where $\mathbf{A} \in \mathbf{R}^{nxm}$.
2. Column sum $\|\mathbf{A}\|_1 := \max_{j=1}^{m} (\sum_{i=1}^{n} |\mathbf{a}_{ij}|)$.
3. Row sum or maximum norm $\|\mathbf{A}\|_\infty := \max_{i=1}^{n} (\sum_{j=1}^{m} |\mathbf{a}_{ij}|)$.
4. Let $\mathbf{A} \in \mathbf{R}^{nxm}$ and let $\{\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda m\}$ be the eigen values of the (nxn)-matrix $\mathbf{A}^T\mathbf{A}$, Then the Hilbert or spectral norm is defined as, $\|\mathbf{A}\|_2 := \max_{i=1}^{m} \{|\lambda_i|\}$.

Now consider the matrix $\mathbf{A}$ as a linear operator .Let $\mathbf{A} : \mathbf{R}^n \to \mathbf{R}^n$ be a linear map that maps $\mathbf{x} \to \mathbf{Ax}$ .This is a continuous linear map. Then we define the norm of this operator as follows.

$$\|\mathbf{A}\| = \sup_{\mathbf{x}\neq 0} \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|}, \mathbf{x} \in \mathbf{R}^n.$$

This is called operator norm induced by the vector norm. Practically it is not necessary at all to use the same norms on $\mathbf{Ax}$ and $\mathbf{x}$, but we have no need of such generality.

Example: let

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 2 & 1 \end{pmatrix}.$$

Then using the given formula above we can easily check that the Frobenius norm is 3, absolute row sum or maximum norm is 3 and also the absolute column sum is 3.

Theorem 2.1.1.4: Let $\|.\|$ be any norm on $\mathbf{R}^{nxn}$ and $\|\mathbf{I}\| = 1$. Let $\mathbf{E} \in \mathbf{R}^{nxn}$.

a) If $\|\mathbf{E}\| < 1$ then, $(\mathbf{I}-\mathbf{E})^{-1}$ exists and $\|(\mathbf{I}-\mathbf{E})^{-1}\| \leq \frac{1}{1-\|\mathbf{E}\|}$. (2.1.1.1)

b) If A is nonsingular and $\|\mathbf{A}^{-1}(\mathbf{B}-\mathbf{A})\| < 1$, then $\mathbf{B}$ is also nonsingular and

$$\|\mathbf{B}^{-1}\| \leq \frac{\|\mathbf{A}^{-1}\|}{1-\|\mathbf{A}^{-1}(\mathbf{B}-\mathbf{A})\|}. \quad (2.1.1.2)$$

Proof: (a) considers the following homogeneous system of linear equations the given fact, $\|\mathbf{E}\| < 1$

$(\mathbf{I}-\mathbf{E})(\mathbf{x}) = \mathbf{Ix} - \mathbf{Ex} = \mathbf{x} - \mathbf{Ex} = \mathbf{0}$, then we have,

$\|(\mathbf{I}-\mathbf{E})(\mathbf{x})\| = \|\mathbf{x}-\mathbf{Ex}\| \geq \|\mathbf{x}\| - \|\mathbf{Ex}\| \geq \|\mathbf{x}\| - \|\mathbf{E}\|\|\mathbf{x}\| = \|\mathbf{x}\|(1-\|\mathbf{E}\|)$.

And using the fact $\|\mathbf{E}\| < 1$,

We get $1-\|\mathbf{E}\| > 0$. This implies $\|\mathbf{x}-\mathbf{Ex}\| \neq o$, $\forall \mathbf{x} \neq \mathbf{o}$. From this we get

$(\mathbf{I}-\mathbf{E})\mathbf{x} = \mathbf{0}$, has only the trivial solution $\mathbf{x} = \mathbf{0}$.

This implies $\mathbf{I}-\mathbf{E}$ is nonsingular.

Hence we have $(\mathbf{I}-\mathbf{E})^{-1}$ exists. Let us set $\mathbf{K} := (\mathbf{I}-\mathbf{E})^{-1}$, then we have

$$1 = \|\mathbf{I}\| = \|(\mathbf{I}-\mathbf{E})(\mathbf{I}-\mathbf{E})^{-1}\| = \|(\mathbf{I}-\mathbf{E})\mathbf{K}\| = \|\mathbf{K}-\mathbf{EK}\| \geq \|\mathbf{K}\| - \|\mathbf{E}\|\|\mathbf{K}\| = \|\mathbf{K}\|(1-\|\mathbf{E}\|).$$

i.e. $1 \geq \|\mathbf{K}\|(1-\|\mathbf{E}\|)$. This gives us

$\|\mathbf{K}\| \leq \frac{1}{1-\|\mathbf{E}\|}$ i.e. $\|(I-\mathbf{E})^{-1}\| \leq \frac{1}{1-\|\mathbf{E}\|}$. This is (2.1.1.1).

To prove **(b)** let $\mathbf{A}$ be nonsingular matrix. Then we can write

$$\mathbf{B} = \mathbf{A} - (\mathbf{A} - \mathbf{B}) = \mathbf{A}(\mathbf{I} - \mathbf{A}^{-1}(\mathbf{A} - \mathbf{B})) \ .$$

This is a product of two invertible matrices and so it is also invertible. More over we have

$\mathbf{B}^{-1} = (\mathbf{I} - \mathbf{A}^{-1}(\mathbf{A} - \mathbf{B}))^{-1}\mathbf{A}^{-1}$. From this we get

$\left\|\mathbf{B}^{-1}\right\| = \left\|(\mathbf{I} - \mathbf{A}^{-1}(\mathbf{A} - \mathbf{B}))^{-1}\mathbf{A}^{-1}\right\| \le \left\|(\mathbf{I} - \mathbf{A}^{-1}(\mathbf{A} - \mathbf{B}))^{-1}\right\|\left\|\mathbf{A}^{-1}\right\|$. But by(a), we have

$$\left\|(\mathbf{I} - \mathbf{A}^{-1}(\mathbf{A} - \mathbf{B}))^{-1}\mathbf{A}^{-1}\right\| \le \frac{1}{1 - \left\|\mathbf{A}^{-1}(\mathbf{A} - \mathbf{B})\right\|} \text{ and also } \left\|\mathbf{A}^{-1}(\mathbf{A} - \mathbf{B})\right\| = \left\|\mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\right\|$$

These things together give us

$$\left\|\mathbf{B}^{-1}\right\| \le \left\|(\mathbf{I} - \mathbf{A}^{-1}(\mathbf{A} - \mathbf{B}))^{-1}\right\|\left\|\mathbf{A}^{-1}\right\| \le \frac{\left\|\mathbf{A}^{-1}\right\|}{1 - \left\|\mathbf{A}^{-1}(\mathbf{A} - \mathbf{B})\right\|} \text{ . I.e.}$$

$$\left\|\mathbf{B}^{-1}\right\| \le \frac{\left\|\mathbf{A}^{-1}\right\|}{1 - \left\|\mathbf{A}^{-1}(\mathbf{A} - \mathbf{B})\right\|} \ .$$

This is completes the proof. //

Remark: Theorem (2.1.4) says that, matrix inversion is continuous in norm. Further more it gives us a relation between the norms of the inverses of nearby matrices.

2.1.2. Solving System of Linear Equations-Matrix Factorization

Definition 2.1.2.1: Let $\mathbf{A} \in \mathbf{R}^{n\times n}$ be a matrix, then the eigen values and eigen vectors of matrix $\mathbf{A}$ are real or complex scalar λ and an n-dimensional vector $\mathbf{v}$, (which is a nonzero), respectively such that $\mathbf{Av} = \lambda\mathbf{v}$.

Definition 2.1.2.2: Let $\mathbf{A} \in \mathbf{R}^{n\times n}$ be a real and symmetric matrix, then $\mathbf{A}$ is said to be positive definite if $\mathbf{x}^{\mathrm{T}}\mathbf{Ax} > 0, \forall \mathbf{x} \in \mathbf{R}^{n} \setminus \{\mathbf{0}\}$.

The details of eigen values and positive definiteness will be discussed in section 3.1.2 of Chapter 3.

Even though a system of linear equations is not the objective of this paper, multidimens ional nonlinear algorithms almost always requires the simultaneous solution of at least one of n-equations with n-unknowns,

$$\mathbf{Ax} = \mathbf{b}, \mathbf{A} \in \mathbf{R}^{n\times n}, \mathbf{x}, \mathbf{b} \in \mathbf{R}^{n} \qquad (2.1.2.1)$$

at each iteration usually to find the Newton point by solving model problem or some modification of it.

Matrix Factorization

It is a common practice to write $\mathbf{A}^{-1}\mathbf{b}$ for economy of notion in mathematical formula. For example $\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{A}_k^{-1} F(\mathbf{x}_k)$ in the Newton's iteration for the equation $F(\mathbf{x}) = \mathbf{0}$.

But it does not mean that we actually compute $\mathbf{A}^{-1}$ and its product with $\mathbf{b}$.It is more effective to calculate $\mathbf{A}^{-1}\mathbf{b}$ by solving the linear system by matrix factorization method s.

Matrix factorization techniques are based on decomposing the (nxn)-matrix, $\mathbf{A}$ into $\mathbf{A} = \mathbf{A}_1\mathbf{A}_2\mathbf{A}_3...\mathbf{A}_m$,where each $\mathbf{A}_i$ is also the (nxn)-matrix of the form (2.1.2.1). This is easy to compute. If $\mathbf{A}$ is nearly singular then the problem (2.1.2.1) is not numerically posed. Thus it is important to know this before we decide to proceed with the solution of $\mathbf{Ax} = \mathbf{b}$, then we peel off the factors to solve in order:

$\mathbf{A}_1\mathbf{b}_1 = \mathbf{b}$, $\mathbf{A}_2\mathbf{b}_2 = \mathbf{b}_1$, $\mathbf{A}_3\mathbf{b}_3 = \mathbf{b}_2, \ldots$, $\mathbf{A}_m\mathbf{b}_m = \mathbf{b}_{m-1}$ and $\mathbf{x} = \mathbf{b}_m$ is the desired solution.

To verify this note that each

$\mathbf{b}_i = \mathbf{A}_i\mathbf{A}_{i-1}\mathbf{A}_{i-2}...\mathbf{A}_1\mathbf{b}$ so, $\mathbf{x} = \mathbf{A}_m^{-1}\mathbf{A}_{m-1}^{-1}\mathbf{A}_{m-2}^{-1}...\mathbf{A}_1^{-1} = \mathbf{A}^{-1}\mathbf{b}$.

There are different methods of decomposing a symmetrical positive definite matrix, $\mathbf{A}$. The most efficient algorithm is the Cholesky decomposition. Let $\mathbf{A}$ be a (nxn)-symmetrical and positive definite matrix. Then the Cholesky decomposition is $\mathbf{A} = \mathbf{LL}^T$, where $\mathbf{L}$ is (nxn)-Lower triangular matrix. Such a matrix $\mathbf{L}$ is simply found from the equation $\mathbf{A} = \mathbf{LL}^T$.

For the existence of such lower triangular matrix and the details of the decomposition refer [6].

Example 1: Solve the given system of linear equations using Cholesky decomposition.

$$2x + y = 3$$
$$x + 2y = 3$$

It is clear that the given matrix is nonsingular and also it is positive definite. Thus we can apply the Cholesky decomposition. To do this we have

$\mathbf{L} = \begin{pmatrix} L_{11} & 0 \\ L_{21} & l_{22} \end{pmatrix}$ and $\mathbf{A} = \mathbf{LL}^T$. Thus ew have

$$\begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix}\begin{pmatrix} L_{11} & L_{21} \\ 0 & L_{22} \end{pmatrix}=\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

Solving for L_{11}, L_{21} and L_{22}, we get $L_{11}=\sqrt{2}$, $L_{21}=\sqrt{2}/2$ and $L_{22}=\sqrt{3}/\sqrt{2}$. And the required Cholesky decomposition is $\mathbf{A}=\mathbf{L}\mathbf{L}^{\mathrm{T}}$. Now $\mathbf{A}\mathbf{x}=\mathbf{b}$. This can be rewritten as $(\mathbf{L}\mathbf{L}^{\mathrm{T}})\mathbf{x}=\mathbf{b}$ or $\mathbf{L}(\mathbf{L}^{\mathrm{T}}\mathbf{x})=\mathbf{b}$. From this we get $\mathbf{b}=(3,3)^{\mathrm{T}}$, $\mathbf{b}_1=(\sqrt{2}x+\sqrt{3}/2\,y, \sqrt{3}/\sqrt{2}\,y)^{\mathrm{T}}$, and $\mathbf{b}_2=(1,1)^{\mathrm{T}}$. Now by setting

$\mathbf{A}_1 := \mathbf{L}$ and $\mathbf{A}_2 := \mathbf{L}^{\mathrm{T}}$, we have the following system of linear equations.
$\mathbf{A}_1\mathbf{b}_1=\mathbf{b}$, $\mathbf{A}_2\mathbf{b}_2=\mathbf{b}_1$. And finally we get as a solution $(x,y)^{\mathrm{T}}=(1,1)^{\mathrm{T}}$.

Example 2: Solve the following system using matrix decomposition.

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 8 & 22 \\ 3 & 22 & 82 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix}=\begin{pmatrix} 5 \\ 6 \\ -10 \end{pmatrix}$$

The given coefficient matrix is nonsingular and hence we can proceed the process of decomposition. Moreover it is symmetrical and positive definite matrix. There for by Cholesky decomposion, we get $\mathbf{A}=\mathbf{L}\mathbf{L}^{\mathrm{T}}$ where

$$\mathbf{L}=\begin{pmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ 3 & 8 & 3 \end{pmatrix}$$ and the required solution is, $\mathbf{x}=(5,-2,-3)^{\mathrm{T}}$.

2.1.3. Errors in Solving Linear Equations

An iteration of a nonlinear algorithm will use the solution s of a linear system $\mathbf{A}_c\mathbf{s}=-\mathrm{F}(\mathbf{x}_c)$ to determine the step or direction to the next approximate solution $\mathbf{x}_+$. There for it is important to know how much the computed step may be affected by the finite precision arithmetic. Also since $\mathbf{A}_c$ and $\mathrm{F}(\mathbf{x}_c)$ are sometimes approximations to the quantities one really wants to use, one is interested in how sensitive the computed step is to changes in the data $\mathbf{A}_c$ and $\mathrm{F}(\mathbf{x}_c)$. To see this consider the following two systems of linear equations

$$\mathbf{A}_1\mathbf{x} = \begin{pmatrix} 8 & -5 \\ 4 & 10 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \mathbf{b}_1 = \begin{pmatrix} 3 \\ 14 \end{pmatrix} \text{ and } \mathbf{A}_2\mathbf{x} = \begin{pmatrix} .66 & 3.24 \\ 1.99 & 10.01 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \mathbf{b}_2 = \begin{pmatrix} 4 \\ 12 \end{pmatrix}.$$

For these two systems of linear equations we can check that both have the solution $\mathbf{x} = (1\ ,\ 1)^T$. If we change $\mathbf{b}_1 \to \mathbf{b}_1 - (0.04, 0.06)^T$, the new solution to the first system becomes $(0.993, 0.9968)^T$. And if we change $\mathbf{b}_2 \to \mathbf{b}_2 - (0.04, 0.06)^T$, by the same quantity. The new solution of the second system becomes $(6,\ 0)^T$.

From this fact we can observe that, small change in the data of the first system makes vary small change in its solution and small change in the data of the second system makes vary large change in its solution, i.e. the second system is very sensitive to changes in its data.

Linear systems whose solutions are very sensitive to changes in their data are called ill conditioned. And similarly nonlinear systems whose solutions are sensitive to small changes in their data are called ill conditioned.

Lemma 2.1.3.1: Let $\mathbf{A}$ is nonsingular (nxn) – matrix .Then the relative change in the solution of the system of linear equations, $\mathbf{Ax} = \mathbf{b}$ is bounded by the relative change in the data multiplied by $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|$. More precisely,

a) If we change $\mathbf{b}$ by $\Delta\mathbf{b}$, then the new solution can be written as $\mathbf{x}^* + \Delta\mathbf{x}$, where $\mathbf{x}^*$ is the solution of $\mathbf{Ax} = \mathbf{b}$, and we have the inequality $\dfrac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}^*\|} \le \|\mathbf{A}\|\|\mathbf{A}^{-1}\|\dfrac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}$.

b) Similarly, if we change $\mathbf{A}$ by $\Delta\mathbf{A}$, then the new solution can be written as $\mathbf{x}^* + \Delta\mathbf{x}$, and we have the inequality $\dfrac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}^*\|} \le \dfrac{\|\mathbf{A}\|\|\mathbf{A}^{-1}\|}{1 - \|\mathbf{A}^{-1}\Delta\mathbf{A}\|}\dfrac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}$.

Proof: (a) Consider a system of linear equations with $\mathbf{A}$ is nonsingular, $\mathbf{Ax} = \mathbf{b}$ and let $\mathbf{x}^*$ be the solution of the system. If we change $\mathbf{b}$ by $\Delta\mathbf{b}$ then the new solution can be written as $\mathbf{x}^* + \Delta\mathbf{x}$. Where

$\mathbf{A}(\mathbf{x}^* + \Delta\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b}$ i.e. $\mathbf{Ax}^* + \mathbf{A}\Delta\mathbf{x} = \mathbf{b} + \Delta\mathbf{b}$. From this we get

$\mathbf{A}\Delta\mathbf{x} = \Delta\mathbf{b}$ or $\Delta\mathbf{x} = \mathbf{A}^{-1}\Delta\mathbf{b}$. (2.1.3.1)

Then for any vector norm $\|.\|$ and induced matrix norm $\|.\|$, we have

$\|\mathbf{A}^{-1}\| = \sup_{\mathbf{x}\neq 0}\dfrac{\|\mathbf{A}^{-1}\mathbf{x}\|}{\|\mathbf{x}\|}$. This gives us $\|\mathbf{A}^{-1}\| \ge \dfrac{\|\mathbf{A}^{-1}\mathbf{x}\|}{\|\mathbf{x}\|}, \forall \mathbf{x} \neq \mathbf{0}$ i.e.

$\|\mathbf{A}^{-1}\|\|\mathbf{x}\| \ge \|\mathbf{A}^{-1}\mathbf{x}\|$. Now letting $\mathbf{x}$ to be $\Delta\mathbf{b}$, we get

$\|\mathbf{A}^{-1}\|\|\Delta\mathbf{b}\| \ge \|\mathbf{A}^{-1}\Delta\mathbf{b}\|$. (2.1.3.2)

We have also, $\|\mathbf{Ax}\| = \|\mathbf{b}\|$ and by (2.1.3.2), $\|\mathbf{Ax}\| \le \|\mathbf{A}\|\|\mathbf{x}\|$. But from (2.1.3.1) and (2.1.3.2) we get ,

$$\|\Delta\mathbf{x}\| \le \|\mathbf{A}^{-1}\|\|\Delta\mathbf{b}\| . \qquad (2.1.3.3)$$

From $\|\mathbf{A}\|\|\mathbf{x}^*\| \ge \|\mathbf{b}\|$, we get

$$\frac{1}{\|\mathbf{x}^*\|} \le \frac{\|\mathbf{A}\|}{\|\mathbf{b}\|} \mathbf{x}, \mathbf{b} \ne \mathbf{0} . \qquad (2.1.3.4)$$

Now multiplying the left hand side of (2.1.3.4) by $\|\Delta\mathbf{x}\|$ and the right hand side by

$\|\mathbf{A}^{-1}\|\|\Delta\mathbf{b}\|$, we get the inequality

$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}^*\|} \le \|\mathbf{A}^{-1}\|\|\Delta\mathbf{b}\| \frac{\|\mathbf{A}\|}{\|\mathbf{b}\|} = \|\mathbf{A}\|\|\mathbf{A}^{-1}\| \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}$. Which completes (a).

Now to prove (b), let $\mathbf{A}$ is changed by $\Delta\mathbf{A}$, and we have the original system $\mathbf{Ax} = \mathbf{b}$, with $\mathbf{A}$ is nonsingular. Then the new solution can be written as $\mathbf{x}^* + \Delta\mathbf{x}$, where $\mathbf{x}^*$ is the solution of the original system $\mathbf{Ax} = \mathbf{b}$.This can be written as,

$(\mathbf{A} + \Delta\mathbf{A})(\mathbf{x}^* + \Delta\mathbf{x}) = \mathbf{b}$. This implies $\mathbf{x}^* + \Delta\mathbf{x} = (\mathbf{A} + \Delta\mathbf{A})^{-1}\mathbf{b}$ .From this we get

$\Delta\mathbf{x} = (\mathbf{A} + \Delta\mathbf{A})^{-1}\mathbf{b} - \mathbf{x}^* = (\mathbf{A} + \Delta\mathbf{A})^{-1}\mathbf{b} - \mathbf{A}^{-1}\mathbf{b}$ or

$\Delta\mathbf{x} = [(\mathbf{A} + \Delta\mathbf{A})^{-1} - \mathbf{A}^{-1}]\mathbf{b}$.

But also $(\mathbf{A} + \Delta\mathbf{A})^{-1} - \mathbf{A}^{-1} = [(\mathbf{I} + \mathbf{A}^{-1}\Delta\mathbf{A})^{-1} - \mathbf{I}]\mathbf{A}^{-1}$. Hence

$\Delta\mathbf{x} = [(\mathbf{I} + \mathbf{A}^{-1}\Delta\mathbf{A})^{-1} - \mathbf{I}]\mathbf{A}^{-1}\mathbf{b}$. From this we have

$$\begin{aligned}
\|\Delta\mathbf{x}\| = \|[(\mathbf{I} + \mathbf{A}^{-1}\Delta\mathbf{A})^{-1} - \mathbf{I}]\mathbf{A}^{-1}\mathbf{b}\| &\le \|[\mathbf{I} - (\mathbf{I} + \mathbf{A}^{-1}\Delta\mathbf{A})^{-1}]\|\|\mathbf{A}^{-1}\mathbf{b}\| \\
&= \|(\mathbf{I} + \mathbf{A}^{-1}\Delta\mathbf{A})^{-1}[(\mathbf{I} + \mathbf{A}^{-1}\Delta\mathbf{A}) - \mathbf{I}]\|\|\mathbf{A}^{-1}\mathbf{b}\| \\
&= \|(\mathbf{I} + \mathbf{A}^{-1}\Delta\mathbf{A})^{-1}(\mathbf{A}^{-1}\Delta\mathbf{A})\|\|\mathbf{A}^{-1}\mathbf{b}\| \\
&\le \|(\mathbf{I} + \mathbf{A}^{-1}\Delta\mathbf{A})^{-1}\|\|\mathbf{A}^{-1}\Delta\mathbf{A}\|\|\mathbf{A}^{-1}\mathbf{b}\| \\
&\le \frac{1}{1 - \|\mathbf{A}^{-1}\Delta\mathbf{A}\|}\|\mathbf{A}^{-1}\Delta\mathbf{A}\|\|\mathbf{A}^{-1}\mathbf{b}\| .
\end{aligned}$$

Assuming

$\|\mathbf{A}^{-1}\Delta\mathbf{A}\| < 1$, we get $\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}^*\|} \le \frac{\|\mathbf{A}\|\|\mathbf{A}^{-1}\|}{1 - \|\mathbf{A}^{-1}\Delta\mathbf{A}\|}\left[\frac{\|\Delta\mathbf{A}\|\|\mathbf{A}^{-1}\mathbf{b}\|}{\|\mathbf{A}\|\|\mathbf{x}^*\|}\right]$. Then finally we get,

$$\frac{\|\Delta \mathbf{x}\|}{\|\mathbf{x}^*\|} \le \frac{\|\mathbf{A}\|\|\mathbf{A}^{-1}\|}{1-\|\mathbf{A}^{-1}\Delta\mathbf{A}\|}\left[\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}\right]$$ .This completes the proof. //

Note: The term $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ in the inequalities is called the condition number and it may be denoted by $K_p(\mathbf{A})$ when using the corresponding induced $l_p(n)$ norm. For the condition number $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|$,

a. if $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ is small then, small relative changes in the data **A** or **b** produces only small change in the solution.
b. If $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ is large, then the small relative changes in the data **A** or **b** produces large relative change in the solution, and the system of equations becomes ill-conditioned.
c. If the condition number is nearly unity, then the system is well conditioned.

Let $\{\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda n\}$ be the eigen values of the (nxn)-matrix, $\mathbf{A}^T\mathbf{A}$. And $\lambda := \max_{i=1}^{n}\{|\lambda_i|\}$ and $\mu := \min_{i=1}^{n}\{|\lambda_i|\}$. Then the spectral norm is given as follows.

$K_p(\mathbf{A}) = \|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2 = \sqrt{\lambda/\mu}$. If **A** is real and symmetrical we have

$K_p(\mathbf{A}) = \|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2 = \lambda/\mu$.

Example: Find the condition number of the following system and distinguish whether the system is ill conditioned or not.

$$\begin{pmatrix} 2.1 & 1.8 \\ 6.2 & 5.3 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2.1 \\ 6.2 \end{pmatrix}$$

Using the spectral norm the eigen values of $\mathbf{A}^T\mathbf{A}$ are $\lambda_1 = 74.179987787$ and $\lambda_2 = .000012132$. There for the condition number is

$K_p(\mathbf{A}) = \|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2 = \sqrt{\lambda/\mu} = 2472.73$. Here we can see that the condition number is very large and hence the given system is highly ill conditioned and very sensi tive to round-off errors.

We will look out for ill conditioned linear systems in extracting information from our local model, because however simple and accurate our model is , the solution of a local model that is sensitive to small changes is of limited uses as an approximation to the solution of the nonlinear system.

2.1.4. First Order Derivatives of $F:\mathbf{R}^n \to \mathbf{R}^m$: Gradient and Jacobeans

Definition 2.1.4.1

1. Let $f:\mathbf{R}^n \to \mathbf{R}$ be a continuous functional then
   a) f is said to be continuously partially differentiable at $\mathbf{x} \in \mathbf{R}^n$, if the partial derivatives $\frac{\partial f(\mathbf{x})}{\partial x_i}, i = 1,2,3,...,n$ exists and are continuous for all i.

   b) The n-vector $\nabla f(\mathbf{x}) = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, \frac{\partial f(\mathbf{x})}{\partial x_3}, ..., \frac{\partial f(\mathbf{x})}{\partial x_n} \right)^T$ (2.1.4.1)
   is called the gradient of the functional f at $\mathbf{x}$.

   c) The functional f is said to be continuously partially differentiable in an open region, $D \subset \mathbf{R}^n$ denoted by $f \in C^{(1)}(D)$ if it is continuously differentiable at each point of D.
2. Let $\mathbf{f}:\mathbf{R}^n \to \mathbf{R}$ be continuously differentiable in an open region $D \subset \mathbf{R}^n$, then for each $\mathbf{x} \in D$ and any p such that $\mathbf{0} \neq \mathbf{p} \in \mathbf{R}^n$. Then the directional derivative of f at x in the direction of $\mathbf{p}$ is defined by

$$\lim_{t \to 0} \frac{f(\mathbf{x}+t\mathbf{p}) - \mathbf{f}(\mathbf{x})}{t} =: \frac{\partial f(\mathbf{x})}{\partial \mathbf{p}}. \tag{2.1.4.2}$$

Using the inner product and the gradient vector the directional derivative can be written as,

$$\left. \frac{d\, f(\mathbf{x}+t\mathbf{p})}{dt} \right|_{t=0} = \frac{\partial f(\mathbf{x})}{\partial \mathbf{p}} =: \nabla f(\mathbf{x})^T \mathbf{p}. \tag{2.1.4.3}$$

It is also useful to observe a slight extension to (2.1.4.3), replacing the point $\mathbf{x}$ with the point $\mathbf{x}+t\mathbf{p}$. Where $f:\mathbf{R}^n \to \mathbf{R}$ and $\mathbf{f}$ has continuous first order partial derivatives at $\mathbf{x}$ with D is an open subset of $\mathbf{R}^n$ containing $\mathbf{x}$. It follows that

$\frac{d\, f(\mathbf{x}+t\mathbf{p})}{dt} = \nabla f(\mathbf{x}+t\mathbf{p})^T \mathbf{p}$, $0 \le t \le 1$, provided that the segment between $\mathbf{x}$ joining $\mathbf{x}+t\mathbf{p}$ lies in D.

Note: If the gradient of f is a constant vector, then f is said to be linear function of $\mathbf{x}$, in this case f is of the form $f(\mathbf{x}) = \mathbf{c}^T\mathbf{x} + \alpha$ where $\mathbf{c}$ is a fixed vector and $\nabla f(\mathbf{x}) = \mathbf{c}$.

Definition 2.1.4.2: Let $F:\mathbf{R}^n \to \mathbf{R}^m$ be continuous functional. Then

a) F is said to be continuously partially differentiable at $\mathbf{x} \in \mathbf{R}^n$, if each component function $F_i:\mathbf{R}^n \to \mathbf{R}, i = 1,2,3,...,m$ is continuously differentiable at $\mathbf{x}$, where $F(\mathbf{x}) = (F_1(\mathbf{x}), F_2(\mathbf{x}),..., F_m(\mathbf{x}))^T$.

b) The matrix $\left(\frac{\partial F_i(\mathbf{x})}{\partial x_j}\right)_{ij}, i = 1,2,...,m \quad j = 1,2,...,n$ is called the Jacobian of F at $\mathbf{x}$ which is an $(mxn)-$matrix. And the common notations for Jacobians of F at $\mathbf{x}$ are $J(\mathbf{x}), F'(\mathbf{x}), \nabla F(\mathbf{x})^T$.

c) F is said to be continuously differentiable in an open set $D \subset \mathbf{R}^n$, denoted by $F \in C^{(1)}(D)$, if F is continuously differentiable at each point in D.

Example: Let $F:\mathbf{R}^2 \to \mathbf{R}^2$ be defined by

$F(\mathbf{z}) = \begin{pmatrix} x+2y \\ x^2 - y \end{pmatrix}$ where $\mathbf{z} = (x,y)^T$ then, obviously, J is a function defined on $\mathbf{R}^n$ with values in $\mathbf{R}^{mxn}$ i.e. $J:\mathbf{R}^n \to \mathbf{R}^{mxn}$ which is $J(\mathbf{z}) = \begin{pmatrix} 1 & 2x \\ 2 & -1 \end{pmatrix}$.

Definition 2.1.4.3: Let $m,n > 0$ and $G:\mathbf{R}^n \to \mathbf{R}^{mxn}, \mathbf{x} \in \mathbf{R}^n$. Let $\|.\|_1$ be a norm on $\mathbf{R}^n$ and $\|.\|_2$ be a norm on $\mathbf{R}^{mxn}$. Then G is said to be Lipchitz continuous at $\mathbf{x}$, if there exists an open set $D \subset \mathbf{R}^n$, $\mathbf{x} \in D$ and a constant γ such that

$$\|G(\mathbf{y}) - G(\mathbf{x})\| \le \gamma \|\mathbf{y} - \mathbf{x}\|, \forall \mathbf{y} \in D. \qquad (2.1.4.4)$$

The constant γ is called Lipchitz constant for G at $\mathbf{x}$. For any specific D containing $\mathbf{x}$ for which (2.1.4.4) holds, G is said to be Lipchitz continuous at $\mathbf{x}$ in the neighborhood D. If (2.1.4.4) holds for every $\mathbf{x} \in D$, we have then $\|G(\mathbf{y}) - G(\mathbf{x})\| \le \gamma \|\mathbf{y} - \mathbf{x}\|, \forall \mathbf{x},\mathbf{y} \in D$ and we say that G is Lipchitz continuous on D, and is denoted by $G \in Lip_\gamma(D)$.

Note: The value γ depends on the norms in (2.1.4.4) but the existence of γ doesn't.

Theorem 2.1.4.1(Directional Derivative)

Let $D \subset \mathbf{R}^n$ be an open set and suppose that $f:D \to \mathbf{R}$ is continuously differentable. Then for each $\mathbf{x} \in D$, and each nonzero point $\mathbf{p} \in \mathbf{R}^n$, the function $f:D \to \mathbf{R}$ has a directional derivative at a point $\mathbf{x}$ in the direction of $\mathbf{p}$ that is given by the formula

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{p}} = \sum_{i=1}^{n} p_i \left(\frac{\partial f}{\partial x_i}(\mathbf{x})\right) = \nabla f(\mathbf{x})^T \mathbf{p}. \qquad (2.1.4.5)$$

Remark: Theorem 2.1.4.1 says that if a function $f: D \to \mathbf{R}$ is continuously partially differentiable, then for each $\mathbf{x} \in D$ and each nonzero point $\mathbf{p} \in \mathbf{R}^n$, the directional derivative at $\mathbf{x}$ in the direction of $\mathbf{p}$ exists.

Lemma 2.1.4.2: Let D be an open subset of $\mathbf{R}^n$, that contains the point $\mathbf{x}$ and suppose that the function $f: D \to \mathbf{R}$ be continuously differentiable.
Choose a positive numbe γ such that the open ball $B_\gamma(\mathbf{x})$ is contained in D.
Then for $\|\mathbf{h}\| < r$ and, $|t| < 1$

$$\frac{d\, f(\mathbf{x}+t\mathbf{h})}{dt} = \nabla f(\mathbf{x}+t\mathbf{h})^T \mathbf{h}. \qquad (2.1.4.6)$$

Proof: Let I be an open interval of real numbers containing the points 0 and 1, such that $\mathbf{x}+t\mathbf{h} \in D$ if $t \in I$. Now define $\phi: I \to \mathbf{R}$ by $\phi(t) = f(\mathbf{x}+t\mathbf{h})$. Since f is differentiable ϕ is also differentiable and more over from the theorem of directional derivative we get,

$$\phi'(t) = \frac{d\, f(\mathbf{x}+t\mathbf{h})}{dt} = \nabla f(\mathbf{x}+t\mathbf{h})^T \mathbf{h}\ . //$$

Now observe that in the theorem (2.1.4.1)above, for each $t \in I$ such that $\mathbf{x}+t\mathbf{h} \in D$, $f: D \to \mathbf{R}$ be continuously differentiable implies ϕ is also continuously differentiable on I. We have for each

$t \in I, \mathbf{x}+t\mathbf{h} \in B_r(\mathbf{x})$, $\phi'(t) = \frac{d\, f(\mathbf{x}+t\mathbf{h})}{dt} = \nabla f(\mathbf{x}+t\mathbf{h})^T \mathbf{h}$ .If the interval I contains the point 0 and 1.Integrate both sides of the last equation with respect to t from 0 to 1.

$\phi(1) - \phi(0) = \int_0^1 \nabla f(\mathbf{x}+t\mathbf{h})^T \mathbf{h}\, dt$. This is similar with

$$f(\mathbf{x}+t\mathbf{h}) - f(\mathbf{x}) = \int_0^1 \nabla f(\mathbf{x}+t\mathbf{h})^T \mathbf{h}\, dt\ .$$

Now formally if we set $\mathbf{z} := \mathbf{x}+t\mathbf{h}$, $d\mathbf{z} := \mathbf{h}dt$ for $t = 1$ implies $\mathbf{z} = \mathbf{x}+\mathbf{h}$, and $t = 0$ implies $\mathbf{z} = \mathbf{x}$. Then (2.1.4.6) may be rewritten as

$$f(\mathbf{x}+t\mathbf{h}) - f(\mathbf{x}) = \int_0^1 \nabla f(\mathbf{x}+t\mathbf{h})^T \mathbf{h}\, dt =: \int_{\mathbf{x}}^{\mathbf{x}+\mathbf{h}} \nabla f(\mathbf{z}) d\mathbf{z}\ . \qquad (2.1.4.7)$$

Theorem 2.1.4.3 :(Mean Value Theorem)
Let D be an open subset of $\mathbf{R}^n$ and suppose that the function $f: D \to \mathbf{R}$ is continuously differentiable. If the segment joining the point $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$ lies in D, then there is a n umber $\theta, 0 < \theta < 1$ such that

$$f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) = \nabla f(\mathbf{x}+\theta\mathbf{h})^T \mathbf{h}\ . \qquad (2.1.4.8)$$

Theorem 2.1.4.3 is a mean value theorem for a real valued function, and the theorem for general mapping has different structure, which is given below.

Theorem 2.1.4.4 :(Mean Value Theorem for general mapping)

Let D be an open subset of $\mathbf{R}^n$ and suppose that the mapping $F: D \to \mathbf{R}^m$ is continuously differentiable. Suppose that the points $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$ are in D and that the segment joining the points also lies in D .Then there are real numbers $\theta_1, \theta_2, \theta_3, ..., \theta_m$ in the open interval (0.1) such that

$F_i(\mathbf{x}+\mathbf{h}) - F_i(\mathbf{x}) = \nabla F_i(\mathbf{x}+\theta_i \mathbf{h})^T \mathbf{h}$, for each $1 \le i \le m$. I.e.

$F(\mathbf{x}+\mathbf{h}) - F(\mathbf{x}) = \mathbf{A}\mathbf{h}$, where $\mathbf{A}$ is $(mxn)-$matrix with i^{th} row $\nabla F_i(\mathbf{x}+\theta_i \mathbf{h})$.

A natural question is whether in (2.1.4.9) we can choose all $\theta_i s'$ equal. In general it is not possible.

Lemma 2.1.4.5: let $f: \mathbf{R}^n \to \mathbf{R}$ be continuously differentiable in an open convex set $D \subseteq \mathbf{R}^n$.

a) ThenFor $\mathbf{x} \in D$ and any $\mathbf{0} \neq \mathbf{p} \in \mathbf{R}^n$, the directional derivative of f at $\mathbf{x}$ in the d irection of $\mathbf{p}$ exists and equals $\nabla f(\mathbf{x})^T \mathbf{p}$.

b) For any $\mathbf{x}, \mathbf{x}+\mathbf{p} \in D$,

$$f(\mathbf{x}+\mathbf{p}) - f(\mathbf{x}) = \int_0^1 \nabla f(\mathbf{x}+t\mathbf{p})^T \mathbf{p}\, dt =: \int_{\mathbf{x}}^{\mathbf{x}+\mathbf{p}} \nabla f(\mathbf{z})\, d\mathbf{z}. \qquad (2.1.4.9).$$

c) For any $\mathbf{x}, \mathbf{x}+\mathbf{p} \in D$ there exists $\mathbf{z} \in (\mathbf{x}, \mathbf{x}+\mathbf{p})$ such that

$$f(\mathbf{x}+\mathbf{p}) = f(\mathbf{x}) + \nabla f(\mathbf{z})^T \mathbf{p} \ . \qquad (2.1.4.10)$$

Remark: In the lemma 2.1.4.5(a) is the theorem of directional derivative, (b) comes from directly from lemma 2.1.4.2 and (c) is the mean value theorem. The detail of the proof of this lemma is given in [1].

Lemma 2.1.4.6: Let $F: \mathbf{R}^n \to \mathbf{R}^m$, be continuously differentiable in an open convex set $D \subset \mathbf{R}^n$ .For any $\mathbf{x}, \mathbf{x}+\mathbf{p} \in D$,

$$F(\mathbf{x}+\mathbf{p}) - F(\mathbf{x}) = \int_0^1 J(\mathbf{x}+t\mathbf{p})\mathbf{p}\, dt =: \int_{\mathbf{x}}^{\mathbf{x}+\mathbf{p}} F'(\mathbf{z})\, d\mathbf{z}. \qquad (2.1.4.11)$$

Remark: Unlike the case of real valued functions, in general there may not exist $\mathbf{z} \in \mathbf{R}^n$ such that $F(\mathbf{x}+\mathbf{p}) = F(\mathbf{x}) + J(\mathbf{z})\mathbf{p}$ .The proof of the above Lemma is given in [1].

In solving affine model (local) for a function $F: \mathbf{R}^n \to \mathbf{R}^m$ the Taylor series approach is quit unsatisfactory, because the mean value theorem cannot be applied in such a way

that we wish. It was in anticipation of this fact we used another approach in chapter one to drive our affine model.

Lemma 2.1.4.7: Let $D \subset \mathbf{R}^n$, be an open and convex set, $\mathbf{0} \neq \mathbf{p} \in \mathbf{R}^n$, $\mathbf{x} \in D$ and $\mathbf{x}+\mathbf{p} \in D$. Let $G: D \to \mathbf{R}^{n\times m}$ be integrable on the closed segment $[\mathbf{x}, \mathbf{x}+\mathbf{p}]$ then for any arbitrary norm, it is true that

$$\left\| \int_0^1 G(\mathbf{x}+t\mathbf{p})\mathbf{p}\,dt \right\| \leq \int_0^1 \left\| G(\mathbf{x}+t\mathbf{p})\mathbf{p} \right\| dt . \qquad (2.1.4.12)$$

2.2. Newton's Method for Solving Nonlinear System of Equations

In this section we try to see one of the methods to solve the following problem. For arbitrary but fixed given $n \in \mathbf{N}$, let $F: \mathbf{R}^n \to \mathbf{R}^n$ is continuously differentiable function. Then find $\mathbf{x} \in \mathbf{R}^n$ such that,

$$F(\mathbf{x}) = \mathbf{0} . \qquad (2.2.1)$$

Derivation of Newton's Algorithm

Newton's method for problem (2.2.1) is derived by finding the root of an affine approximation to F at the current iteration $\mathbf{x}_c$. This approximation is created by using the same techniques as for the one variable problem.

Let $\mathbf{x}_c \in \mathbf{R}^n$ be a fixed vector to be a first approximation (estimation) for the solution of (2.2.1) and let $\mathbf{p} \in \mathbf{R}^n$ be a variable vector with "small" $\|\mathbf{p}\|$ then by the lemma (2.1.4.5) we have,

$$\begin{aligned} F(\mathbf{x}_c+\mathbf{p}) &= F(\mathbf{x}_0) + \int_0^1 F'(\mathbf{x}_c+t\mathbf{p})\mathbf{p}\,dt \\ &= F(\mathbf{x}_c) + \int_{\mathbf{x}_c}^{\mathbf{x}_c+\mathbf{p}} F'(\mathbf{z})\,d\mathbf{z} . \end{aligned} \qquad (2.2.2)$$

We approximate the integral in (2.2.2) by the linear term $J(\mathbf{x}_c)\mathbf{p}$ to get the affine approximation to F at a perturbation $\mathbf{p}$ of $\mathbf{x}_c$. If the perturbation $\mathbf{p}$ is "small" then

$\int_{\mathbf{x}_c}^{\mathbf{x}_c+\mathbf{p}} F'(z)\,d\mathbf{z} = \int_0^1 F'(\mathbf{x}_c+t\mathbf{p})\mathbf{p}\,dt \approx F'(\mathbf{x}_c)$. There for

$F(\mathbf{x}_c+\mathbf{p}) \approx F(\mathbf{x}_0) + F'(\mathbf{x}_c)\mathbf{p}$.

The approximating term is obviously affine for $\mathbf{p}$ and we denote it by

$$M_c(\mathbf{x}_c + \mathbf{p}) := F(\mathbf{x}_c) + J(\mathbf{x}_c)\mathbf{p} \ . \tag{2.2.3}$$

Now we consider the following two systems of equations:

$$F(\mathbf{x}_c + \mathbf{p}) = \mathbf{0} . \tag{2.2.4}$$

$$M_c(\mathbf{x}_c + \mathbf{p}) = \mathbf{0} . \tag{2.2.5}$$

Equation (2.2.4) is a nonlinear system and (2.2.5) is a linear system of equations. By c-ontinuity of $\mathbf{F}$ and $J(\mathbf{x})$, the solution of (2.2.5) can be expected to be 'close 'to the solution of (2.2.4) for a "small" $\mathbf{p}$.

Next we solve the step $\mathbf{s}_N$ that makes, $M_c(\mathbf{x}_c + \mathbf{s}_N) = \mathbf{0}$, giving the Newton iteration for (2.2.1) i.e. to find

$$\mathbf{x}_N := \mathbf{x}_c + \mathbf{s}_N \text{ s.t } J(\mathbf{x}_c)\mathbf{s}_N = -F(\mathbf{x}_c) \text{ and } \mathbf{x}_N = \mathbf{x}_c + \mathbf{s}_N . \tag{2.2.6}$$

The new point '$\mathbf{x}^+$' can be considered as an approximated vector for the solution of (2.2.4).

An equivalent way to view this procedure is that we are finding a simultaneous zero of the affine models of n-functions of F given by,
$(M_c)_i(\mathbf{x}_c + \mathbf{s}_N) = f_i(\mathbf{x}_c) + \nabla f_i(\mathbf{x}_c)\mathbf{s}_N, i = 1,2,3,...,n$. All these things are generalized in the following well-known Algorithm, the so called Newton's Algorithm for a system of nonlinear equations.

Theorem 2.2.1: (Newton's Algorithm for a system of nonlinear equations) Let $F : \mathbf{R}^n \to \mathbf{R}^n$ be continuously differentiable and $\mathbf{x}_0 \in \mathbf{R}^n$, be a fixed vector. Then the sequence $\{\mathbf{x}_k\}$ is generated, where for each iteration $k \in \mathbf{N}_0 = \{0\} \cup \mathbf{N}$, $J(\mathbf{x}_k)s_k = -F(\mathbf{x}_k)$ and $\mathbf{x}_{k+1} := \mathbf{x}_k + \mathbf{s}_k$.

Example: Let

$$F(\mathbf{z}) = \begin{pmatrix} x + y - 3 \\ x^2 + y^2 - 9 \end{pmatrix}, \ \mathbf{x}_0 = (1,5)^T$$

For this system we can immediately calculate all solutions of the equations $F(\mathbf{x}) = 0$, namely $(0,3)^T$ and $(3,0)^T$. Now we want to apply Newton's method here. We have the Jacobian is $J(\mathbf{z}) = \begin{pmatrix} 1 & 1 \\ 2x & 2y \end{pmatrix}$ and the initial point given is $x_0 = (1,5)^T$.

The first two iteration of the Newton's method are

$J(\mathbf{x}_0)\mathbf{s}_0 = -F(\mathbf{x}_0)$. Solving this equation we get

$\mathbf{s}_0 = (-13/8, -11/8)^T$ and $\mathbf{x}_1 = \mathbf{x}_0 + \mathbf{s}_0 = (-0.625, 3.625)^T$. We have again

$\mathbf{J}(\mathbf{x}_1)\mathbf{s}_1 = -F(\mathbf{x}_1)$. Then we get and we get

$\mathbf{s}_1 = (145/272, -145/272)^T$, then we have the next iteration is

$\mathbf{x}_2 = \mathbf{x}_1 + \mathbf{s}_1 = (-0.92, 3.092)^T$.

The roots of F are $(0,3)^T$ and $(3,0)^T$.From the above calculation $\mathbf{x}_2$ is quite close to the solution $(0,3)^T$. The following table is calculated by means of (program1) for the first seven steps.

k	$\mathbf{s}_k$	$\mathbf{x}_k$
1	(-1.625, -1.375)	(1., 5)
2	(0.533088, -0.533088)	(-0.625, 3.625)
3	(0.0892584, -0.0892584)	(-0.0919118, 3.09191)
4	(0.002651, -0.002651)	(0.00265334 , 3.00265)
5	($2.3426\text{x}10^{-6}$, $-2.3426\text{x}10^{-6}$)	($-2.3426\text{x}10^{-6}$, 3.)
6	($1.82965\text{x}10^{-12}$, $-1.82965\text{x}10^{-12}$)	($-1.82953\text{x}10^{-12}$, 3.)
7	(0., 0.)	(0., 3.)

Table (2.2.1a)

For the above example the Newton's method seems to be working well. This is the main advantage of Newton's method when x_0 is close enough to the solution $\mathbf{x}^*$ and $J(\mathbf{x}^*)$ is non singular.

Example 2: Let

$$F(\mathbf{z}) = \begin{pmatrix} e^x - 1 \\ e^y - 1 \end{pmatrix} \text{ and } \mathbf{x}_0 = (-1,-1)^T$$

Here $\mathbf{x}^* = (0,0)^T$ is the solution of the system. Now for this problem, we can easily see that, the first five Newton's iterations, the 5^{th}-iteration, which is very close to the real solution.The table is calculated by means of (program 1).

K	s_k	x_k
0	(1.71828, 1.71828)	(-1, -1)
1	(-0.512411, -0.512411)	(0.718282, 0.718282)
2	(-0.186062, -0.186062)	(0.205871, 0.205871)
3	(-0.0196142, -0.0196142)	(0.0198091, 0.0198091)
4	(-0.000194892, -0.000194892)	(0.000194911, 0.000194911)
5	($-1.89939\text{x}10^{-8}$, $-1.89939\text{x}10^{-8}$)	($1.89939\text{x}10^{-8}$, $1.89939\text{x}10^{-8}$)

Table (2.2.1b)

In this example at the fifth iteration we have a solution which is very close to the real root $(0, 0)^T$ for the starting point $(-1, -1)^T$. Next let us see that what will happen if we change the initial point to $(-10, -10)^T$. See the following table.

In the following table the first five iterations for the given initial point is far from the real solution. This is due to the initial point is far from the real solution.

k	s_k	x_k
0	(22025.5, 22025.5)	(-10, -10)
1	(-1.00000000000, -1.00000000000)	(22015.5, 22015.5)
2	(-1.00000000000, -1.00000000000)	(22015.5, 22015.5)
3	(-1.00000000000, -1.00000000000)	(22015.5, 22015.5)
4	(-1.00000000000, -1.00000000000)	(22015.5, 22015.5)
5	(-1.00000000000, -1.00000000000)	(22015.5, 22015.5)

Table (2.2.1c)

From these examples, we can see that when the initial point $\mathbf{x}_0$ is close enough to the real root, then the Newton method converges quadratically to the truth solution and if the initial point $\mathbf{x}_0$ is not good, and then the sequence $\{\mathbf{x}_k\}$ doesn't converge to the solution. That is the choice of initial point in the Newton method plays an important role. The major advantages and disadvantages of the Newton's method are the following.

Advantages

1. If the starting point $\mathbf{x}_0$ is good enough and the Jacobean is nonsingular, then the sequence $\{\mathbf{x}_k\}$ converges quadratically to $\mathbf{x}^*$.
2. Gives exact solution in one iteration if the function, F is affine.

Disadvantages

1. Not globally convergent for many problems.
2. Requires $J(\mathbf{x})$ at each iteration.
3. Each iteration requires the solution of a system of linear equations that may be singular or ill conditioned.

We will always want to use Newton's method at each final iteration of any nonlinear algorithm to take the advantage of its fast local convergence, but it will have to be modified in order to converge globally.

Local Convergence of Newton's Method

There are different approaches to prove that the sequence $\{\mathbf{x}_k\}$ generated by the algorithm (2.2.1) converges quadratically to $\mathbf{x}^*$. In this paper we consider one important approach. First we are going to prove the following lemma.

Lemma 2.2.2.1: Let $D \subseteq \mathbf{R}^n$ be an open convex setand $\mathbf{x} \in D$ be fixed. $F: \mathbf{R}^n \to \mathbf{R}^n$, be a Lipchitz continuous at $\mathbf{x}$ with respect to D. Then

$$\|F(\mathbf{x}+\mathbf{p}) - F(\mathbf{x}) - J(\mathbf{x})\mathbf{p}\| \le \frac{\gamma}{2}\|\mathbf{p}\|^2, \forall \mathbf{p} \in \mathbf{R}^n \text{ such that } \mathbf{x}+\mathbf{p} \in D. \qquad (2.2.7)$$

Proof: By lemma 2.1.4.6,

$$F(\mathbf{x}+\mathbf{p}) - F(\mathbf{x}) - J(\mathbf{x})\mathbf{p} = \int_0^1 J(\mathbf{x}+t\mathbf{p})\mathbf{p}\,dt - J(\mathbf{x})\mathbf{p}$$

$$= \int_0^1 [J(\mathbf{x}+t\mathbf{p})\mathbf{p} - J(\mathbf{x})\mathbf{p}]dt.$$

From this we get, $\|F(\mathbf{x}+\mathbf{p})F(\mathbf{x}) - J(\mathbf{x})\mathbf{p}\| = \left\| \int_0^1 [J(\mathbf{x}+t\mathbf{p})\mathbf{p} - J(\mathbf{x})\mathbf{p}]dt \right\|$

$$\le \int_0^1 \|J(\mathbf{x}+t\mathbf{p})\mathbf{p} - J(\mathbf{x})\mathbf{p}\|\,dt \le \int_0^1 \gamma \|\mathbf{x}+t\mathbf{p}-\mathbf{x}\|\|\mathbf{p}\|\,dt$$

$$= \int_0^1 \gamma t \|\mathbf{p}\|\|\mathbf{p}\|\,dt = \int_0^1 \gamma \|\mathbf{p}\|^2\,dt = \frac{\gamma}{2}\|\mathbf{p}\|^2.$$

This completes the proof. //

Remark: lemma (2.2.1) provides a useful upper bound of the errors of affine model (2.2.3) with respect to the equation $F(\mathbf{x}) = \mathbf{0}$. Now we are ready to prove the following theorem, which is the core of Newton method.

Theorem 2.2.2.2: Let $D \subset \mathbf{R}^n$ be an open and convex set, $F: \mathbf{R}^n \to \mathbf{R}^n$ be continuously differentiable on D .Suppose that there is $\mathbf{x}^* \in D$ such that

1. $F(\mathbf{x}^*) = \mathbf{0}$ and $J(\mathbf{x}^*)^{-1}$ exists
2. There is $\beta > 0$ such that $\|J(\mathbf{x}^*)^{-1}\| \le \beta$,
3. There are $r, \gamma > 0$ such that $J \in Lip_\gamma\left(B(x^*, r)\right)$ where $B(\mathbf{x}^*, r) \subseteq D$ is an open ball. Then there exists $\varepsilon > 0$ such that for each initial point $\mathbf{x}_0 \in B(\mathbf{x}^*, r)$,

a) The sequence $\{\mathbf{x}_k\}$ generated by $\mathbf{x}_{k+1} = \mathbf{x}_k - J(\mathbf{x}_k)^{-1}F(\mathbf{x}_k)$ where $k = 0,1,2,3,...$ exists

b) Converges to $\mathbf{x}^*$

c) Satisfies the inequality $\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \le \beta\gamma\|\mathbf{x}_k - \mathbf{x}^*\|^2$, $\forall k = 0,1,2,...$

Proof: Let $\mathbf{x}^* \in D$, choose $\varepsilon := \min\{r, \frac{1}{2\beta\gamma}\}$ and we will show that this number is the convenient number in the theorem. Let $\mathbf{x}_0 \in B(\mathbf{x}^*, r)$, then $\|\mathbf{x}^* - \mathbf{x}_0\| < \varepsilon \le \frac{1}{2\beta\gamma}$. Now first we show that $J(\mathbf{x}_0)^{-1}$ exists. To see this, we have that J is Lipchitz continuous

nction at x^* with constant γ. Thus we have

$$\left\|J(\mathbf{x}^*)^{-1}(J(\mathbf{x}_0)-J(\mathbf{x}^*))\right\| \le \left\|J(\mathbf{x}^*)^{-1}\right\|\left\|J(\mathbf{x}_0)-J(\mathbf{x}^*)\right\|$$
$$\le \beta\gamma\left\|\mathbf{x}_0-\mathbf{x}^*\right\| \le \beta\gamma\varepsilon \le \beta\gamma\frac{1}{2\beta\gamma}=\frac{1}{2}<1.$$

hat is we have $\left\|J(\mathbf{x}^*)^{-1}(\mathbf{J}(\mathbf{x}_0)-J(\mathbf{x}^*))\right\|<1$. By theorem (2.1.14) we get that the verse of $J(\mathbf{x}_0)$ exists. More over we have the following relation

$$\left\|J(\mathbf{x}_0)^{-1}\right\| \le \frac{\left\|J(\mathbf{x}^*)^{-1}\right\|}{1-\left\|J(\mathbf{x}^*)^{-1}(J(\mathbf{x}_0)-J(\mathbf{x}^*))\right\|}.$$

here for we have that, $\mathbf{x}_1=\mathbf{x}_0-J(\mathbf{x}_0)^{-1}F(\mathbf{x}_0)$. Hence $\mathbf{x}_1$ exists and well defined .This (a) for k=0.We have with $F(\mathbf{x}^*)=\mathbf{0}$,

$$\mathbf{x}_1-\mathbf{x}^* = \mathbf{x}_0-\mathbf{x}^*-J(\mathbf{x}_0)^{-1}F(\mathbf{x}_0)=\mathbf{x}_0-\mathbf{x}^*-J(\mathbf{x}_0)^{-1}(F(\mathbf{x}_0)-\mathbf{F}(\mathbf{x}^*)$$
$$= J(\mathbf{x}_0)^{-1}\left(F(\mathbf{x}_0)-F(\mathbf{x}^*)-J(\mathbf{x}_0)(\mathbf{x}^*-\mathbf{x}_0)\right). \text{ There for,}$$
$$\left\|\mathbf{x}_1-\mathbf{x}^*\right\| \le \left\|J(\mathbf{x}_0)^{-1}\right\|\left\|F(\mathbf{x}_0)-F(\mathbf{x}^*)-J(\mathbf{x}_0)(\mathbf{x}^*-\mathbf{x}_0)\right\|$$
$$\le 2\beta[\frac{\gamma}{2}\left\|\mathbf{x}^*-\mathbf{x}_0\right\|^2]=\beta\gamma\left\|\mathbf{x}^*-\mathbf{x}_0\right\|^2.$$

Hence we have that $\left\|\mathbf{x}_1-\mathbf{x}^*\right\| \le \beta\gamma\left\|\mathbf{x}^*-\mathbf{x}_0\right\|^2$. This is (c) for k=0. More over we have hat $\left\|\mathbf{x}_1-\mathbf{x}^*\right\| \le \beta\gamma\left\|\mathbf{x}_0-\mathbf{x}^*\right\|^2=\beta\gamma\left\|\mathbf{x}_0-\mathbf{x}^*\right\|\left\|\mathbf{x}_0-\mathbf{x}^*\right\| \le \frac{\varepsilon}{2}<\varepsilon$.

This is (b). And the proof is completed for k=0. Now assume that the theorem is also rue for k=0, 1, 2, 3…n, .i.e. $\{\mathbf{x}_k\}$ generated by the rule exists,

$$\left\|\mathbf{x}_{k+1}-\mathbf{x}^*\right\| \le \beta\gamma\left\|\mathbf{x}_k-\mathbf{x}^*\right\|^2,\ \forall k=0,1,2,...n-1 \text{ and also } \left\|\mathbf{x}_{k+1}-\mathbf{x}^*\right\| \le \frac{1}{2}\left\|\mathbf{x}_k-\mathbf{x}^*\right\|$$

$\forall k=0,1,2,...n-1$. And we are going to show that the theorem also holds for $k=n+1$.

We have that,

$$\left\|J(\mathbf{x}^*)^{-1}\left(J(\mathbf{x}_n)-J(\mathbf{x}^*)\right)\right\| \le \left\|J(\mathbf{x}_n)^{-1}\right\|\left\|J(\mathbf{x}_n)-J(\mathbf{x}^*)\right\| \le \beta\gamma\left\|\mathbf{x}_n-\mathbf{x}^*\right\| \le \frac{1}{2}<1, \text{ using the}$$

nduction assumption and the fact that J is Lipchitz continuous function at $\mathbf{x}^*$ with constant γ.This implies $J(\mathbf{x}_n)$ is nonsingular and its inverse exists. Also the following estimation holds.

$$\left\|J(\mathbf{x}_n)^{-1}\right\| \le \frac{\left\|J(\mathbf{x}^*)^{-1}\right\|}{1-\left\|J(\mathbf{x}^*)^{-1}[J(\mathbf{x}_n)-J(\mathbf{x}^*)]\right\|}$$ Now $\mathbf{x}_{n+1} = \mathbf{x}_n - J(\mathbf{x}_n)^{-1} F(\mathbf{x}_n)$.

I.e. $\mathbf{x}_{n+1}$ is exists and well defined i.e. (a) holds for k=n.

Moreover,

$$\begin{aligned}\mathbf{x}_{n+1} - \mathbf{x}^* &= \mathbf{x}_n - \mathbf{x}^* - J(\mathbf{x}_n)^{-1} F(\mathbf{x}_n) = \mathbf{x}_n - \mathbf{x}^* - J(\mathbf{x}_n)^{-1}\left(F(\mathbf{x}_n) - F(\mathbf{x}^*)\right) \\ &= J(\mathbf{x}_n)^{-1}\left(F(\mathbf{x}_n) - F(\mathbf{x}^*) - J(\mathbf{x}_n)(\mathbf{x}^* - \mathbf{x}_n)\right).\end{aligned}$$

From this we get,

$$\left\|\mathbf{x}_{n+1} - \mathbf{x}^*\right\| \le \left\|J(\mathbf{x}_n)^{-1}\right\|\left\|F(x_n) - F(\mathbf{x}^*) - J(\mathbf{x}_n)(\mathbf{x}_n - \mathbf{x}^*)\right\| \le \beta\gamma\left\|\mathbf{x}_n - \mathbf{x}^*\right\|^2,$$

by the lemma2.2.1.So (c) holds. Moreover from induction assumption it follows that,

$\left\|\mathbf{x}_{n+1} - \mathbf{x}^*\right\| \le \beta\gamma\left\|\mathbf{x}_n - \mathbf{x}^*\right\|^2 = \beta\gamma\left\|\mathbf{x}_n - \mathbf{x}^*\right\|\left\|\mathbf{x}_n - \mathbf{x}^*\right\| \le \beta\gamma\frac{1}{2}\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|\left\|\mathbf{x}_n - \mathbf{x}^*\right\|$. Since

we have $\left\|\mathbf{x}_n - \mathbf{x}^*\right\| \le \frac{1}{2^n}\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|$, we get

$\left\|\mathbf{x}_{n+1} - \mathbf{x}^*\right\| \le \frac{1}{2}\left\|\mathbf{x}_n - \mathbf{x}^*\right\| \le \frac{1}{2^{n+1}}\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|$. This implies $\mathbf{x}_n \to \mathbf{x}^*, n \to \infty$.

This completes the proof of the theorem by induction.//

Remark: From the theorem above we have seen that, for each $n \in \mathbf{N}$,

$\left\|\mathbf{x}_{n+1} - \mathbf{x}^*\right\| \le \beta\gamma\left\|\mathbf{x}_n - \mathbf{x}^*\right\|^2$, this shows that the local error in the affine model used to produce each iteration of the Newton method is almost a constant times .,This shows that the sequence converges quadratically.

In the previous section we have seen that the Jacobian is useful in formulating models of multivariable nonlinear functions for finding roots. In many applications, however this derivative is not analytically available. In this case we approximate the Jacobian by finite differences, i.e. it is reasonable to approximate the Jacobian by finite differences. To get details refer **[7].**

3. Newton's Method for Unconstrained Minimization of a Nonlinear Function of Several Variables

This chapter considers the local algorithm (Newton method) for unconstrained minimization. Section 3.1 considers some facts from Linear Algebra and multivariable calculus which are relevant to minimization. Section 3.2 considers Newton's method for unconstrained minimization of twice continuously differentiable functions.

3.1. Introduction

3.1.1. Eigen Values and Positive Definiteness

Definition 3.1.1.1: Let $\mathbf{A} \in \mathbf{R}^{n\times n}$ be a matrix , then the Eigen values and Eigen vectors of matrix $\mathbf{A}$ are real or complex scalar λ and an n-dimensional vector $\mathbf{v}$, (which is a nonzero), respectively such that $\mathbf{Av} = \lambda \mathbf{v}$.

Example 1:

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 6 & 1 \end{pmatrix}.$$

Then we can easily solve that the eigen values of this matrix are 4 and 1 and the corresponding eigen vectors are $t(1,-2)^T$ and $t(1,3)^T$ for a non zero constant, t respectively. In this example one of the eigen values is positive and the other is negative. But there are cases where all of the Eigen values of a matrix $\mathbf{A}$ are positive.

Example 2: Let

$$\mathbf{A} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}.$$

Which is real and symmetric matrix and its eigen values are 3 and 1. In this example all eigen values of the matrix $\mathbf{A}$ are positive.

Recall that the characteristic polynomial of a matrix, $\mathbf{A}$ is $p(\lambda) = \lambda^n + c_0\lambda^{n-1} + \ldots + |\mathbf{A}|$.

Where $|\mathbf{A}|$ is the determinant of (nxn)-matrix $\mathbf{A}$. If all eigen values of a matrix $\mathbf{A}$ are positive then $|\mathbf{A}| \neq 0$. Hence $\mathbf{A}$ is nonsingular.

Definition 3.1.1.2: Let $\mathbf{A} \in \mathbf{R}^{n\times n}$ a real symmetric matrix. Then $\mathbf{A}$ is said to be

a) Positive definite if and only if $\mathbf{v}^T\mathbf{Av} > 0$ for every nonzero vector $\mathbf{v} \in \mathbf{R}^n$.
b) Positive semi definite if and only if $\mathbf{v}^T\mathbf{Av} \geq 0, \forall \mathbf{v} \in \mathbf{R}^n$.
c) Negative definite if and only if $-\mathbf{A}$ Positive definite.
d) Negative semi definite if and only if $-\mathbf{A}$ Positive semi definite.

We have the fact that a real symmetrical matrix is positive definite if and only if all of its eigen values are positive. And it is positive semi definite if and only if its eigen values are nonnegative.With the help of these properties we have the followingtheorem which is given without proof. To see some properties and necessary and sufficientconditions of positive definiteness refer [2].

Theorem 3.1.1.1: If $\mathbf{A}$ is any positive definite matrix. Then its inverse exists and also positive definite.

Example: Let

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}.$$

Be a given matrix. We can easily see that the given matrix is real and symmetrical with eigen values 1and 4, which are all positive and hence the given matrix is positive definite, and consequently the matrix $\mathbf{A}$ has an inverse.

Definition 3.1.1.3: A set $\mathbf{U} \subset \mathbf{R}^n$ is said to be convex if for any $\mathbf{x}, \mathbf{y} \in \mathbf{U}$, and $\lambda \in (0,1)$, $\lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in \mathbf{U}$. Equivalently, a set $U \subset \mathbf{R}^n$ is said to be convex if for any two points in U, the line segment joining those two points must be entirely contained in U.

Definition 3.1.1.4: Let $U \subset \mathbf{R}^n$ is convex set and $f : U \to \mathbf{R}$ be a function. Then

a) $\mathbf{f}$ is said to be convex if for any $\mathbf{x}, \mathbf{y} \in \mathbf{U}$ and $\lambda \in (0,1)$, $f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \le \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$. And similarly
b) f is said to be strictly convex if for any $\mathbf{x}, \mathbf{y} \in U$ and $\lambda \in (0,1)$, $f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$.
c) f is said to be concave if $-f$ is convex.
d) f is said to be strictly concave if $-f$ is strictly convex.

3.1.2. Second Order Derivatives of $f : \mathbf{R}^n \to \mathbf{R}$: The Hessian of f

Definition 3.1.2.1:

1. Let $f : \mathbf{R}^n \to \mathbf{R}$ be continuously differentiable then

a) f is said to be twice continuously differentiable at $\mathbf{x} \in \mathbf{R}^n$, if $\dfrac{\partial^2 f(\mathbf{x})}{\partial x_i x_j}$ exists

and is continuous for $1 \le i, j \le n$.

b) The Hessian of f at $\mathbf{x}$ is then defined as the (nxn)-matrix whose $\mathbf{ij}^{th}$ element is $\nabla^2 f(\mathbf{x})_{ij} = \dfrac{\partial^2 f(\mathbf{x})}{\partial x_i x_j}$, $1 \le i, j \le n$.

c) The function f is said to be twice continuously differentiable in an open region $D \subset \mathbf{R}^n$, denoted by $f \in C^{(2)}(D)$, if it is twice continuously differentiable every point in D.

2. Let $f : \mathbf{R}^n \to \mathbf{R}$ be continuously differentiable in an open convex set $D \subset \mathbf{R}^n$. Then for any $\mathbf{x} \in D$ and any nonzero perturbation $\mathbf{p} \in \mathbf{R}^n$ the second partial derivative of f at $\mathbf{x}$ in the direction of $\mathbf{p}$ is defined by

$$\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{p}^2} = \lim_{\varepsilon \to 0} \frac{\dfrac{\partial f(\mathbf{x}+\varepsilon\mathbf{p})}{\partial \mathbf{p}} - \dfrac{\partial f(\mathbf{x})}{\partial \mathbf{p}}}{\varepsilon}.$$

Lemma 3.1.2.1: Let $f : \mathbf{R}^n \to \mathbf{R}$ be twice continuously differentiable in an open region $D \subset \mathbf{R}^n$ then,

a) For any $\mathbf{x} \in D$ and any nonzero perturbation $\mathbf{p} \in \mathbf{R}^n$ the second directional derivative of f at $\mathbf{x}$ in the direction of $\mathbf{p}$ exists and equals $\mathbf{p}^T \nabla^2 f(\mathbf{x}) \mathbf{p}$. (3.1.2.1)

b) For any $\mathbf{x}, \mathbf{x}+\mathbf{p} \in D$, there exists $\mathbf{z} \in (\mathbf{x}, \mathbf{x}+\mathbf{p})$ such that

$$f(\mathbf{x}+\mathbf{p}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T \mathbf{p} + \frac{1}{2}\mathbf{p}^T \nabla^2 f(\mathbf{z}) \mathbf{p}. \qquad (3.1.2.2)$$

Remark: lemma (3.1.2.1) suggests that we might model the functional f around a point $\mathbf{x}_c$, by the quadratic model,

$$m_c(\mathbf{x}_c + \mathbf{p}) := f(\mathbf{x}) + \nabla f(\mathbf{x})^T \mathbf{p} + \frac{1}{2}\mathbf{p}^T \nabla^2 f(\mathbf{x}) \mathbf{p}. \qquad (3.1.2.3)$$

Unconstrained Minimization

Definition 3.1.2.2: The unconstrained Optimization problem is a problem to minimize or maximize an objective function without restriction on the variable over thedomain of interest. That is, it is the problem (for minimization) of the form, $f(\mathbf{x}) \to \min$, $\mathbf{x} \in \mathbf{R}^n$, where $f : \mathbf{R}^n \to \mathbf{R}$ is twice continuously differentiable.

Definition 3.1.2.3:

a) A point $\mathbf{x}^*$ is said to be a local minimizer of f, if there is a neighborhood $S(\mathbf{x}^*, \varepsilon)$ of $\mathbf{x}^*$ such that $f(\mathbf{x}^*) \le f(\mathbf{x})$ for all $\mathbf{x} \in S(\mathbf{x}^*, \varepsilon)$.

b) A point $\mathbf{x}^*$ is a minimizer of (global minimizer) of f if $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x}$, where $\mathbf{x}$ ranges over all $\mathbf{x} \in \mathbf{R}^n$ (or at least over the domain of interest to the modeler).

c) A point $\mathbf{x}^*$ is said to be strict local (global) minimizer, if the inequality is strict inequality.

If a function is twice continuously differentiable, we may be able to tell that $\mathbf{x}^*$ is a local minimizer (and possibly strict local minimizer) by examining just the gradient $\nabla f(\mathbf{x}^*)$ and the Hessian $\nabla^2 f(\mathbf{x}^*)$. The mathematical tool used to study minimizer of smooth function is Taylor's theorem.

The Necessary and Sufficient Conditions

Theorem 3.1.2.2: Suppose that $f : \mathbf{R}^n \to \mathbf{R}$ is continuously differentiable and that $\mathbf{p} \in \mathbf{R}^n$. Then we have $f(\mathbf{x}+\mathbf{p}) = f(\mathbf{x}) + \nabla \mathbf{f}(\mathbf{x}+t\mathbf{p})^T \mathbf{p}$, for some $t \in (0,1)$. And if f is twice continuously differentiable, we have

$$\nabla f(\mathbf{x}+\mathbf{p}) = \nabla f(\mathbf{x}) + \int_0^1 \nabla^2 f(\mathbf{x}+t\mathbf{p})\mathbf{p}\,dt \text{ .That}$$

$$f(\mathbf{x}+\mathbf{p}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T \mathbf{p} + \frac{1}{2}\mathbf{p}^T \nabla^2 f(\mathbf{x}+t\mathbf{p})\mathbf{p} \text{ for some } t \in (0,1).$$

The necessary condition for optimality are derived by assuming that $\mathbf{x}^*$ is a local minim izer and then proving facts about $\nabla f(\mathbf{x}^*)$ and $\nabla^2 f(\mathbf{x}^*)$.

Theorem 3.1.2.3:(First order necessary conditions)
Suppose that $f : \mathbf{R}^n \to \mathbf{R}$ is continuously differentiable in an open convex set $D \subset \mathbf{R}^n$. If $\mathbf{x}^* \in D$ is a local minimizer of f, then $\nabla f(\mathbf{x}^*) = \mathbf{0}$.

Remark: The theorem above says that a continuously differentiable function attains its local minimum at the point where the gradient is zero, to see the proof refer [2].

Theorem 3.1.2.4: (Second order necessary conditions)
Let $f : \mathbf{R}^n \to \mathbf{R}$ be twice continuously differentiable in an open region $D \subset \mathbf{R}^n$. If $\mathbf{x}^* \in D$ and $\mathbf{x}^*$ is a local minimize of f. Then $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}^*)$ is positive semi definite.

Remark: The theorem above says that a twice continuously differentiable function f attains its local minimizer at the point where the gradient is zero and $\nabla^2 f(\mathbf{x}^*)$ is positive semi definite.

Theorem 3.1.2.5:(Second order sufficient conditions)Let $f:\mathbf{R}^n \to \mathbf{R}$ and $f \in C^{(2)}(D)$ in an open region $D \subset \mathbf{R}^n$. If $\mathbf{x}^* \in D, \nabla f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}^*)$ is positive definite then, $\mathbf{x}^*$ is a strict local minimizer.

The following corollary is directly comes from (theorem 3.1.3.3) and (theorem 3.1.3.4), which is most relevant to the conditions for Newton's method to find a local minimizer by finding the zero of ∇f.

Corollary 3.1.2.6: Let $f:\mathbf{R}^n \to \mathbf{R}$ be twice continuously differentiable in an open convex set $D \subset \mathbf{R}^n$. If $\mathbf{x}^* \in D$, $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f$ is a Lipchitz continuous at $\mathbf{x}^*$ with $\nabla^2 f(\mathbf{x}^*)$ nonsingular. Then $\mathbf{x}^*$ is a local minimize of f if and only if $\nabla^2 f(\mathbf{x}^*)$ is positive definite.

The sufficient condition for minimizations partially explains our interest in symmetric and positive semi definite matrices. They are very important, because our minimization algorithms commonly model f at perturbations $\mathbf{p}$ of $\mathbf{x}_c$ by a quadratic function

$$m_c(\mathbf{x}_c + \mathbf{p}) = f(\mathbf{x}_c) + \nabla f(\mathbf{x}_c)^T \mathbf{p} + \frac{1}{2}\mathbf{p}^T \mathbf{H}\mathbf{p}, \text{ where } \mathbf{H} \text{ is the Hessian of } f \text{ at } \mathbf{x}_c.$$

It is also important to understand the shapes of multivariable quadratic functions. They are said to be;

a) Strictly convex if $\mathbf{H}$ is positive definite.
b) Convex if $\mathbf{H}$ is positive semi definite.
c) Concave if $-\mathbf{H}$ is positive semi definite.
d) Strictly concave if $-\mathbf{H}$ is positive definite.
e) Saddle shaped if $\mathbf{H}$ is indefinite.

When a function f is convex, any local minimizer and global minimizer are simple to characterize. The following theorem gives this fact and the proof is available in [2].

Theorem 3.1.2.7: When the function f is convex, any local minimizer is also lobalminimizer.

3.2. Newton's Method

In this section we discuss the Newton's method fro the following problem.
Let $f:\mathbf{R}^n \to \mathbf{R}$ be twice continuously differentiable function. Then find the minimum of $f(\mathbf{x})$ for $\mathbf{x} \in \mathbf{R}^n$, i.e. solving the problem,

$$f(\mathbf{x}) \to \min \quad , \mathbf{x} \in \mathbf{R}^n. \tag{3.2.1}$$

The aim is now to formulate an iteration method which gives us the possibility to find a local minimum point of a given function f. Now let $\mathbf{x}_0 \in \mathbf{R}^n$ be any arbitrary initial point. Then we consider the quadratic function,

$m_0(\mathbf{x}_0 + \mathbf{p}) := f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^T \mathbf{p} + \frac{1}{2}\mathbf{p}^T \nabla^2 f(\mathbf{x}_0)\mathbf{p}$ and we consider it as an approximation of $f(\mathbf{x}_0 + \mathbf{p})$.i.e.

$$m_0(\mathbf{x}_0 + \mathbf{p}) \approx f(\mathbf{x}_0 + \mathbf{p}) .(3.2.2)$$

If the function f has a local minimum at $\mathbf{x}_0 + \mathbf{p}$. Then it has to be,

$$\nabla m_0(\mathbf{x}_0 + \mathbf{p}) = \nabla f(\mathbf{x}_0) + \nabla^2 f(\mathbf{x}_0)\mathbf{p} = \mathbf{0},$$
$$\nabla^2 f(\mathbf{x}_0)\mathbf{p} = -\nabla f(\mathbf{x}_0). \qquad (3.2.3)$$

The last is a system of linear equations and there for possible to solve with the aid of linear algebra. Now to solve (3.2.3) the new point $\mathbf{x}^*$ (it may be a local minimum of m_0) can be considered as an approximated vector for a local minimum point of f.

So we are ready to formulate an iteration to approximate a local minimum point of f. What we have to do is that in each iteration step, we use a quadratic model,

$$m_k(\mathbf{x}_k + \mathbf{p}) := f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T \mathbf{p} + \frac{1}{2}\mathbf{p}^T \nabla^2 f(\mathbf{x}_k)\mathbf{p} .(3.2.4)$$

We have the following Theorem which generalizes all things above.

Theorem 3.2.1: (Newton's Algorithm for Unconstrained Minimization)
Let $f : \mathbf{R}^n \to \mathbf{R}$ be twice continuously differentiable function and $\mathbf{x}_0 \in \mathbf{R}^n$ be a fixed vector. Then a sequence $\{\mathbf{x}_k\}$ is generated by
$\nabla^2 f(\mathbf{x}_k)\mathbf{s}_k = -\nabla f(\mathbf{x}_k)$, $\mathbf{x}_{k+1} := \mathbf{x}_k + \mathbf{s}_k$, where $k = 0,1,2,3,\ldots,$.

If the generated sequence is convergent, then we check by theorem 3.1.2.4 whether the limit is a local minimum point of f.

Note that the theorem (3.2.1) is simply the application of Newton's method theorem (2.2.1) to the system nonlinear equations, $\nabla f(\mathbf{x}) = \mathbf{0}$.

Example: Let $f : \mathbf{R}^2 \to \mathbf{R}$ be given by

$$f(x, y) = (x-2)^4 + (x-2)^2 y^2 + (y+1)^2$$

Which has its minimum at $\mathbf{x}^* = (2,-1)^T$. By the application of theorem (3.2.1) starting from the initial point $\mathbf{x}_0 = (1,1)^T$. The following sequence of points is calculated by (program 2).

k	x_k	$f(x_k)$
0	$x_0 = (1,1)^T$	6
1	$x_1 = (1,-0.5)^T$	1.5
2	$x_2 = (1.3913043,-0.6956217)^T$	$4.09x10^{-1}$
3	$x_3 = (1.7459441,-0.94879809)^T$	$6.49x10^{-2}$
4	$x_4 = (1.9862783,-1.0482081)^T$	$2.53x10^{-3}$
5	$x_5 = (1.9987342,-1.0001700)^T$	$1.63x10^{-6}$
6	$x_6 = (1.9999996,-1.0000016)^T$	$2.75x10^{-12}$

Table 3.2.1

From the above example we can se that Newton's algorithm gives the solution at step 6 which is very close to the real solution.

The Major advantages and disadvantages of Newton's method for unconstrained minimization are given below.

Advantages

1. If $\mathbf{x}_0$ is sufficiently close to the solution $\mathbf{x}^*$ of f, with $\nabla^2 f(\mathbf{x}^*)$ nonsingular then the sequence $\{\mathbf{x}_k\}$ generated by Theorem3.2.1 converges quadratically to $\mathbf{x}^*$.
2. If f is strictly convex and quadratic, then ∇f is affine and so $\mathbf{x}_1$ will be the unique minimizer of f.

Disadvantages

1. The algorithm is not globally convergent.
2. In each iteration step the algorithm requires to solve a system of linear equations, some systems may be ill conditioned in some iteration steps.
3. In each iteration step one has to calculate the derivatives $\nabla f(\mathbf{x}_k)$ and $\nabla^2 f(\mathbf{x}_k)$.
4. The method is not specifically geared to a minimization problem; theorem 3.2.1 may also proceed toward a maximizer or a saddle point of f, where ∇f is zero. That is each step simply goes to the critical point of the current local quadratic model. This is only consistent with trying to minimize $f(\mathbf{x})$ if $\nabla^2 f(\mathbf{x}_c)$ is positive definite.

Lemma 3.2.1: Let $f:\mathbf{R}^n \to \mathbf{R}$ be twice continuously differentiable function in an open convex set $D \subset \mathbf{R}^n$. Let $\nabla^2 f(\mathbf{x})$ be Lipchitz continuous at $\mathbf{x}$ in the neighborhood D, using the vector norm and the induced matrix norm and a constant γ. Then for any $\mathbf{x}+\mathbf{p} \in D$,

$$\left| f(\mathbf{x}+\mathbf{p}) - [f(\mathbf{x}) + \nabla f(\mathbf{x})^T \mathbf{p} + \frac{1}{2}\mathbf{p}^T \nabla^2 f(\mathbf{x})\mathbf{p}] \right| \le \frac{\gamma}{4}\|\mathbf{p}\|^3 . \qquad (3.2.5)$$

Proof: We have

$f(\mathbf{x}+\mathbf{p}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T \mathbf{p} + \frac{1}{2}\mathbf{p}^T \nabla^2 f(\mathbf{x}+t\mathbf{p})\mathbf{p}$, for some $0 \le t \le 1$, by extended Mean value theorem. And from this we have,

$f(\mathbf{x}+\mathbf{p}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^T \mathbf{p} = \frac{1}{2}\mathbf{p}^T \nabla^2 f(\mathbf{x}+t\mathbf{p})\mathbf{p}$. And integrating both sides of this equation with respect to t from 0 to 1, provided that there is an interval , I in $\mathbf{R}$ containing the points 0 and 1 such that $\forall t \in I, \mathbf{x} \in D, \mathbf{0} \ne \mathbf{p} \in \mathbf{R}^n, \mathbf{x}+t\mathbf{p} \in D$. We get

$$f(\mathbf{x}+\mathbf{p}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^T \mathbf{p} = \int_0^1 \frac{1}{2}\mathbf{p}^T \nabla^2 f(\mathbf{x}+t\mathbf{p})\mathbf{p}\,dt . \qquad (3.2.6)$$

Now we have,

$$f(\mathbf{x}+\mathbf{p}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^T \mathbf{p} - \frac{1}{2}\mathbf{p}^T \nabla^2 f(\mathbf{x})\mathbf{p} = \int_0^1 \frac{1}{2}[\mathbf{p}^T \nabla^2 f(\mathbf{x}+t\mathbf{p})\mathbf{p} - \mathbf{p}^T \nabla^2 f(\mathbf{x})\mathbf{p}]dt .$$

From this we get

$$\left\| f(\mathbf{x}+\mathbf{p}) - \mathbf{f}(\mathbf{x}) - \nabla f(\mathbf{x})^T \mathbf{p} - \frac{1}{2}\mathbf{p}^T \nabla^2 f(\mathbf{x})\mathbf{p} \right\| = \left\| \int_0^1 \frac{1}{2}[\mathbf{p}^T \nabla^2 f(\mathbf{x}+t\mathbf{p})\mathbf{p} - \mathbf{p}^T \nabla^2 f(\mathbf{x})\mathbf{p}]dt \right\|$$

$$\le \int_0^1 \frac{1}{2}\left\| \mathbf{p}^T \nabla^2 f(\mathbf{x}+t\mathbf{p})\mathbf{p} - \mathbf{p}^T \nabla^2 f(\mathbf{x})\mathbf{p} \right\|$$

$$\le \int_0^1 \frac{1}{2}\|\mathbf{p}\|\left\| \nabla^2 f(\mathbf{x}+t\mathbf{p}) - \nabla^2 f(\mathbf{x}) \right\|\|\mathbf{p}\|\,dt$$

$$\le \int_0^1 \frac{1}{2}\|\mathbf{p}\|^3 t\,dt = \frac{\gamma}{4}\|\mathbf{p}\|^3 , \text{ using then fact that}$$

$\nabla^2 f(\mathbf{x})$ is Lipchitz continuous at $\mathbf{x}$ in the neighborhood D .//

Remark: lemma (3.2.1) provides a useful bound on the error in the quadratic model (3.2.2), as an approximation to the $f(\mathbf{x}+\mathbf{p})$.

4. Newton's Method for Constrained Minimization

4.1. Introduction

Constrained Minimization

An optimization problem begins with a set of independent variables or parameters and often includes conditions or restrictions that define the acceptable values of the variables. Such restrictions are called the constraints of the problem.

Definition 4.1.1: And any optimization problems containing constraints are called the constrained optimization problems or equivalently a constrained optimization problem is a maximization or minimization problem subject to some constraints.

Example 1:

$$f(\mathbf{x}) \to \min \ , \ \mathbf{x} \in S.$$

Where S is a collection of points satisfying all the constraints.The problem is minimizing a function f subject to (s.t.) some constraints.

Any nonlinear constrained optimization problem can be one of the following; equality constrained, inequality constrained, or mixed constrained.

Example 2: $\min \ (x+y)$
s.t. $x\text{-}3y=5$

Example 3: $\min \ (x^2+y^2)$
s.t. $x+y\leq 4$

Example 4: $\min (x-1)^2+(y-2)^2$
s.t. $x+y=4, x-2y\leq 1$

Definitions of different types of local solutions are simple extensions of the corresponding definitions for the unconstrained case, except that we restrict consideration to the feasible points in the neighborhood of $\mathbf{x}^*$.

4.1.1. The Lagrange Method

The Lagrange method is an important instrument to transform a nonlinear constrained optimization problem to an equivalent unconstrained optimization problem (or a problem with easy constraints) so that the optimal solution of the constrained problem can be characterized using the transformed form. This is based on the following idea.

Let $\mathbf{R}^n$ where $n \in \mathbf{N}$, be considered as an inner product space and let $U \subseteq \mathbf{R}^n$ be an open set. And $f : U \to \mathbf{R}$ be differentiable and $S \subseteq U$ (S is not necessarily open). Now consider the following constrained nonlinear optimization problem,

$$(P) \qquad f(\mathbf{x}) \to \min , \quad \mathbf{x} \in S. \tag{4.1.1}$$

Note: Through the whole chapter the set U is considered to be an open.
To convert (P) into an equivalent unconstrained problem we try to find a functional Λ on U with the following properties. Let

$\Lambda : U \to \mathbf{R}$ be differentiable and $\Lambda(\mathbf{x}) = 0, \forall \mathbf{x} \in S$.

Then we consider the auxiliary functional L such that $L : U \to \mathbf{R}$ be given by

$$L(\mathbf{x}) = f(\mathbf{x}) + \Lambda(\mathbf{x}) \quad , \quad \mathbf{x} \in U.$$

Such L is called the Lagrange functional or Lagrangian for (P). And we consider the Lagrange problem,

$$(P_L) \qquad L(\mathbf{x}) \to \min , \quad \mathbf{x} \in U. \tag{4.1.2}$$

In this case it is clear that the functional L is differentiable, as both f and Λ are differentiable. Now the relation between (P) and (P_L) is given below in the following theorem.

Theorem 4.1.1: (Lagrange Lemma)

(a) If $\mathbf{x}^*$ is an optimal solution (global minimizer) of (P_L) and $\mathbf{x}^* \in S$, then $\mathbf{x}^*$ is also a global minimizer of (P).

(b) If $\mathbf{x}^*$ is an optimal solution (local minimizer) of (P_L) and $\mathbf{x}^* \in S$, then $\mathbf{x}^*$ is also a local minimizer of (P).

Proof :(a) given that $\mathbf{x}^*$ is a global minimizer of L and also we have

$$L(\mathbf{x}) = f(\mathbf{x}) + \Lambda(\mathbf{x}) \text{ and } \Lambda(\mathbf{x}) = 0, \ \forall \mathbf{x} \in S.$$

Now,

$L(\mathbf{x}) \geq L(\mathbf{x}^*), \ \forall \mathbf{x} \in U$. Also we have $\mathbf{x}^* \in S$. This implies

$f(\mathbf{x}) + \Lambda(\mathbf{x}) \geq f(\mathbf{x}^*) + \Lambda(\mathbf{x}^*), \forall \mathbf{x} \in U$. From this we get

$f(\mathbf{x}) + \Lambda(\mathbf{x}) \geq f(\mathbf{x}^*) + \Lambda(\mathbf{x}^*), \forall \mathbf{x} \in S$ in particular.

There for we have that,

$f(\mathbf{x}) \geq f(\mathbf{x}^*), \forall \mathbf{x} \in S$ as $\Lambda(\mathbf{x}) = 0, \forall \mathbf{x} \in S$.

There for $\mathbf{x}^*$ is also a global minimizer of (P).

(b) Given that $\mathbf{x}^*$ is an optimal solution (local minimizer) of (P_L). This implies there is a neighborhood $B_\varepsilon(\mathbf{x}^*) \subseteq U$, such that

$L(\mathbf{x}) \geq L(\mathbf{x}^*), \forall \mathbf{x} \in B_\varepsilon(\mathbf{x}^*) \cap U$. In particular also we have,

$L(\mathbf{x}) \geq L(\mathbf{x}^*), \forall \mathbf{x} \in B_\varepsilon(\mathbf{x}^*) \cap S$. From this we get

$f(\mathbf{x}) + \Lambda(\mathbf{x}) \geq f(\mathbf{x}^*) + \Lambda(\mathbf{x}^*), \forall \mathbf{x} \in B_\varepsilon(\mathbf{x}^*) \cap S$. I.e.

$f(\mathbf{x}) \geq f(\mathbf{x}^*), \forall \mathbf{x} \in B_\varepsilon(\mathbf{x}^*) \cap S$ as $\Lambda(\mathbf{x}) = 0, \ \forall \mathbf{x} \in S$.

There for $\mathbf{x}^*$ is also a local minimizer of (P). //

In general given any nonlinear constrained problem, to determine its Lagrangian depends on the types of constraints.

4.1.1.1. Lagrange Method for Equality Constraints

Consider the following optimization problem. Let

$f, g_i : U \to \mathbf{R}$, $i = 1, 2, 3, ...m$ are twice continuously differentiable. Then we have the following equality constrained problem.

$$(P_=) \qquad f(\mathbf{x}) \to \min, \quad \mathbf{x} \in S$$
$$S := \{\mathbf{x} \in U : g_i(\mathbf{x}) = 0, \ \forall i = 1, 2, ..., m\}$$

If we take for $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_m)^T$ the functional Λ is defined as follows.

$$\Lambda(\mathbf{x}) = \sum_{i=1}^{m} \mu_i \, g_i(\mathbf{x}), \ \boldsymbol{\mu} \in \mathbf{R}^m \text{ and } \Lambda(\mathbf{x}) = 0, \forall \mathbf{x} \in S.$$

Then we define the Lagrangian L as follows.

$$L(\mathbf{x}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{i=1}^{m} \mu_i \, g_i(\mathbf{x}) \ , \quad (\mathbf{x}, \boldsymbol{\mu}) \in U \mathrm{x} \mathbf{R}^m.$$

And the corresponding unconstrained problem is

$$(P_\mu) \qquad \mathbf{L}(\mathbf{x}, \boldsymbol{\mu}) \to \min \quad , \mathbf{x} \in U.$$

From the Lagrange Lemma we have the following result.

Theorem 4.1.2: If $\mathbf{x}^*$ is an optimal solution of (P_μ) for some $\boldsymbol{\mu}^* \in \mathbf{R}^m$ and $\mathbf{x}^* \in S$, then $\mathbf{x}^*$ is also an optimal solution of $(P_=)$. (The statement is also holds for local optimum solution).

As U is open and $\mathbf{x}^*$ is an optimal solution of (P_μ) for some $\boldsymbol{\mu}^* \in \mathbf{R}^m$. This implies $\nabla L_x(\mathbf{x}^*, \boldsymbol{\mu}^*) = 0$ and also $\mathbf{x}^* \in S$ implies $g_i(\mathbf{x}^*) = 0, \ \forall i = 1,2,3,...,m$.

Thus we have the following necessary condition for $\mathbf{x}^*$ to be the optimal solution of $(P_=)$.

The Karush-Kuhn-Tuker (KKT) Conditions

If $\mathbf{x}^*$ is an optimal solution of $(P_=)$, then for some $\boldsymbol{\mu}^* \in \mathbf{R}^m$, the following holds,

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \mu_i^* \nabla g_i(\mathbf{x}^*) = \mathbf{0}$$

$$g_i(\mathbf{x}^*) = \mathbf{0}, \forall i = 1,2,3,...m .$$

If $\mathbf{x}^*$ together with some $\boldsymbol{\mu}^* \in \mathbf{R}^m$ satisfy the KKT conditions then $\mathbf{x}^*$ is called the KKT point for $(P_=)$.

Note: The KKT conditions are only the necessary conditions. However, under the convexity assumption the KKT conditions are also sufficient.

Theorem 4.1.3: Let $\mathbf{x}^*$ be a KKT point for $(P_=)$ with the Lagrange multiplier $\boldsymbol{\mu}^*$, and $L(., \boldsymbol{\mu}^*)$ be a convex functional over a convex set U, then $\mathbf{x}^*$ is an optimal solution of $(P_=)$.

4.1.1.2. Lagrange Method for Inequality Constraints

Consider the following constrained problem. Let

$f, g_i : U \to \mathbf{R}$, $i = 1,2,3,...m$ are all twice continuously differentiable functions.

Then we have the following constrained problem.

$$(P_<) \qquad f(\mathbf{x}) \to \min \ , \quad \mathbf{x} \in S ,$$

$$S := \{\mathbf{x} \in U : g_i(\mathbf{x}) \le 0, \ \forall i = 1,2,...,m\} \ .$$

And we define the Lagrangian for $(P_<)$ as follows. For any $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_k)^T \in \mathbf{R}_+^k$

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) \ , \quad (\mathbf{x}, \boldsymbol{\lambda}) \in U x \mathbf{R}_+^k .$$

Thus the corresponding unconstrained problem is given as follows.

$$(p_\lambda) \qquad L(\mathbf{x}, \boldsymbol{\lambda}) \to \min , \ \mathbf{x} \in U \ .$$

The following theorem is the corresponding Lagrange lemma relating $(P_<)$ and (P_λ).

Theorem 4.1.4: Let $\lambda^* = (\lambda_1^*, \lambda_2^*, ..., \lambda_k^*)^T \in \mathbf{R}_+^k$, $\mathbf{x}^* \in S$ and $\sum_{i=1}^{m} \lambda_i^* g_i(\mathbf{x}^*) = \mathbf{0}$. If $\mathbf{x}^*$ is an optimal of (P_λ) and $\mathbf{x}^* \in S$, then $\mathbf{x}^*$ is an optimal solution of $(P_<)$.

If for some $\lambda^* = (\lambda_1^*, \lambda_2^*, ..., \lambda_k^*)^T \in \mathbf{R}_+^k$ and $\sum_{i=1}^{k} \lambda_i^* g_i(\mathbf{x}^*) = \mathbf{0}$. Then $\mathbf{x}^*$ is an optimal solution of (P_λ) over U implies $\nabla L_x(\mathbf{x}^*, \lambda^*) = \mathbf{0}$. And $\mathbf{x}^* \in S$ implies $g_i(\mathbf{x}^*) \le 0, \forall i = 1,2,3,...,k$. We have the following necessary condition $\mathbf{x}^*$ to be the optimal solution of $(P_<)$.

Karush -Kuhn -Tuker (KKT) conditions

If $\mathbf{x}^*$ is an optimal solution of $(P_<)$, then for some $\lambda_+^* \in \mathbf{R}^m$, the following holds,

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(\mathbf{x}^*) = \mathbf{0}$$
$$\lambda_i^* \ge 0, \forall i = 1,2,...,m$$
$$g_i(\mathbf{x}^*) \le \mathbf{0}, \forall i = 1,2,...,m$$
$$\lambda_i g_i(\mathbf{x}^*) = \mathbf{0}, \forall i = 1,2,...,m$$

Under the convexity assumption, the KKT points are also sufficient.

Theorem 4.1.5: Let $\mathbf{x}^*$ be a KKT point for $(P_<)$, with a corresponding Lagrange multiplier λ^*. If f and g_i are all convex functions over a convex set U, then $\mathbf{x}^*$ is an optimal solution of $(P_<)$.

4.1.1.3. Lagrange Method for Mixed Constraints

Consider the following constrained problem. Let

$f, g_i, h_j : U \to \mathbf{R}$, $i = 1,2,3,...m, j = 1,2,3,...,k$,

are all twice continuously differentiable functions. Then we consider the minimization problem,

$$(p_\le) \qquad f(\mathbf{x}) \to \min \quad , \quad \mathbf{x} \in S$$
$$S := \{\mathbf{x} \in U : g_i(\mathbf{x}) \le \mathbf{0}, \forall i = 1,2,3,...m \ , h_j(\mathbf{x}) = \mathbf{0}, \ \forall j = 1,2,3,...,k\} \ .$$

And we define the Lagrangian for $(P_{\leq})$ as follows. For any $\boldsymbol{\lambda} \in \mathbf{R}_+^{\,m}$ and $\boldsymbol{\mu} \in \mathbf{R}^k$,

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i \, g_i(\mathbf{x}) + \sum_{j=1}^{k} \mu_j \, h_j(\mathbf{x}) \quad , \quad (\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \in U \times \mathbf{R}_+^{\,m} \times \mathbf{R}^k .$$

Thus the corresponding auxiliary problem is,

$$(P_{\lambda,\mu}) \qquad L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \to \min \quad , \quad \mathbf{x} \in U.$$

Theorem 4.1.6: Let $\boldsymbol{\lambda}^* \in \mathbf{R}_+^{\,m}, \boldsymbol{\mu}^* \in \mathbf{R}^k, \mathbf{x}^* \in S$ and $\sum_{i=1}^{k} \lambda_i^* \, g_i(\mathbf{x}^*) = \mathbf{0}$. If $\mathbf{x}^*$ is an optimal solution of $(P_{\lambda,\mu})$, then $\mathbf{x}^*$ is also an optimal solution of $(P_{\leq})$.

The necessary condition for $\mathbf{x}^*$ to be an optimal solution of $(P_{\leq})$ is given below.

Karush -Kuhn -Tuker (KKT) conditions

If $\mathbf{x}^*$ is an optimal solution of $(P_{\leq})$, then for some $\boldsymbol{\lambda}^* \in \mathbf{R}_+^{\,m}$ and $\boldsymbol{\mu}^* \in \mathbf{R}^k$, the following holds.

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^{k} \mu_j^* \nabla h_j(\mathbf{x}^*) = \mathbf{0}$$

$$\lambda_i^* \geq 0, \forall i = 1, 2, ..., m$$

$$g_i(\mathbf{x}^*) \leq \mathbf{0}, \forall i = 1, 2, ..., m$$

$$h_j(\mathbf{x}^*) = \mathbf{0}, \forall j = 1, 2, ..., k$$

$$\lambda_i^* \, g_i(\mathbf{x}^*) = \mathbf{0}, \forall i = 1, 2, .., m$$

Under the convexity assumption the KKT points are optimal solutions.

4.2. Newton's Method for Constrained Minimization

Consider the general nonlinear constrained optimization problem. Let,

$f, g_i, h_j : U \to \mathbf{R}$ for $i = 1, 2, 3, ...m$ and $j = 1, 2, 3, ...k$

are all twice continuously differentiable. We have the optimization problem,

$$(p_{\leq}) \qquad f(\mathbf{x}) \to \min \quad , \quad \mathbf{x} \in S,$$

$$S := \{\mathbf{x} \in U : g_i(\mathbf{x}) \leq 0, \forall i - 1, 2, 3, ...m \, , h_j(\mathbf{x}) = 0, \; \forall j = 1, 2, 3, ..., k\} \; .$$

And the corresponding unconstrained problem is as follows. For $\boldsymbol{\lambda} \in \mathbf{R}_+^{\,m}, \boldsymbol{\mu} \in \mathbf{R}^k$,

$$(p_{\lambda,\mu}) \qquad L(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\mu}) \to \min \quad , \quad \mathbf{x} \in U.$$

Our aim is to solve $(P_{\lambda,\mu})$ using the Newton's method and to find the corresponding solution for $(P_{\leq})$. We have the Lagrangian L is given by,

$$L(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^{k} \mu_j h_j(\mathbf{x}), \qquad (\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\mu}) \in U \mathrm{x} \mathbf{R}_+^{\,m} \mathrm{x} \mathbf{R}^k.$$

The necessary condition for the point $(\mathbf{x}^*,\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*)$ to be an optimal solution of $(P_{\lambda,\mu})$ is $\nabla_{\mathbf{x}} \mathbf{f}(\mathbf{x}^*,\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*) = \mathbf{0}$. And the KKT condition is,

$$\begin{aligned} \nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^{k} \mu_j^* \nabla h_j(\mathbf{x}^*) &= \mathbf{0} \\ \lambda_i^* &\geq 0, \forall i = 1,2,...,m \\ g_i(\mathbf{x}^*) &\leq \mathbf{0}, \forall i = 1,2,...,m \\ h_j(\mathbf{x}^*) &= \mathbf{0}, \forall j = 1,2,...,k \\ \lambda_i^* g_i(\mathbf{x}^*) &= \mathbf{0}, \forall i = 1,2,..,m \end{aligned}$$

Now let us discuss the Newton's method for all cases.

4.2.1. Newton's Method for Equality Constrained Minimization

The given optimization problem with equality constraints is of the following form.

$$\begin{aligned} (P_{=}) \qquad & f(\mathbf{x}) \to \min, \quad \mathbf{x} \in S \\ & S := \{\mathbf{x} \in U : g_i(\mathbf{x}) = 0, \ \forall i = 1,2,...,m\} \ . \end{aligned}$$

And the corresponding transformed unconstrained optimization problem is

$$(p_{\mu}) \qquad \mathbf{L}(\mathbf{x},\boldsymbol{\mu}) \to \min \quad , \mathbf{x} \in U.$$

The necessary condition for $\mathbf{x}^*$ to be an optimal solution of $(P_{=})$ is the KKT condition which is,

$$\begin{aligned} \nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \mu_i^* \nabla g_i(\mathbf{x}^*) &= \mathbf{0} \\ g_i(\mathbf{x}^*) &= \mathbf{0}, \forall i = 1,2,3,...m. \end{aligned}$$

From the KKT conditions we get a system of nonlinear equations, which has $(n+m)$-equations and $(n+m)$-unnowns given as follows.

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \mu_i^* \nabla g_i(\mathbf{x}^*) = \mathbf{0}$$

$$g_i(\mathbf{x}^*) = \mathbf{0}, \forall i = 1,2,3,...m$$

Now set

$$F(\mathbf{y}) := \begin{pmatrix} \nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \mu_i^* \nabla g_i(\mathbf{x}^*) \\ g_i(\mathbf{x}^*), \forall i = 1,2,3,...m \end{pmatrix} = 0 \ . \qquad (4.2.1.1)$$

Clearly (4.2.1.1) is solvable by Newton's methods. And for a given initial point $\mathbf{y}_c$, we have by the (Newton's Algorithm 2.2.1), $J(\mathbf{y}_k)\mathbf{s}_k = -F(\mathbf{y}_k)$ and $\mathbf{y}_{k+1} := \mathbf{y}_k + \mathbf{s}_k$, where $k \in \mathbf{N}_0 = \{0\} \cup \mathbf{N}$ generates a sequence $\{\mathbf{y}_k\}$. And if our initial point $\mathbf{y}_c$ is good enough we get the solution $\mathbf{y}^* = (\mathbf{x}^*, \boldsymbol{\lambda}^*)$ of (4.2.1.1). Which is the KKT point of $(P_=)$. Moreover, if $\mathbf{x}^* \in S$ for some $\boldsymbol{\lambda}^* \in \mathbf{R}^m$. Then $\mathbf{x}^*$ is a candidate for the solution of $(P_=)$. Additionally if L is assumed to be convex, then $\mathbf{x}^*$ is also an optimal solution of $(P_=)$.

4.2.2. Newton's Method for Inequality Constrained Minimization

The given optimization problem with inequality constraints is given as

$$(P_<) \qquad f(\mathbf{x}) \to \min \ , \qquad \mathbf{x} \in S \ ,$$
$$S := \{\mathbf{x} \in U : g_i(\mathbf{x}) \le 0, \ \forall i = 1,2,...,m\} \ .$$

And the corresponding unconstrained problem is, for $\boldsymbol{\lambda} \in \mathbf{R}_+^m$

$$(p_\lambda) \qquad L(\mathbf{x}, \boldsymbol{\lambda}) \to \min, \ \mathbf{x} \in U \ .$$

The necessary condition for $\mathbf{x}^*$ to be an optimal solution of $(P_<)$ is the KKT condition. Which is,

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(\mathbf{x}^*) = \mathbf{0}$$

$$\lambda_i^* \ge 0, \forall i = 1,2,...,m$$

$$g_i(\mathbf{x}^*) \le \mathbf{0}, \forall i = 1,2,...,m$$

$$\lambda_i g_i(\mathbf{x}^*) = \mathbf{0}, \forall i = 1,2,...,m$$

Introducing slack variables in KKT condition, we get the following system of nonlinear equations,

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(\mathbf{x}^*) = \mathbf{0}$$
$$-\lambda_i^* + u_i^2 = 0, \forall i = 1,2,...,m$$
$$g_i(\mathbf{x}^*) + v_i^2 = \mathbf{0}, \forall i = 1,2,...,m$$
$$\lambda_i g_i(\mathbf{x}^*) = \mathbf{0}, \forall i = 1,2,...,m.$$

The above system of nonlinear equations, which has $(n+4m)$-equations and $(n+4m)$-unkowns. This is solvable by the Newton's method.
Now set

$$F(\mathbf{y}) = \begin{pmatrix} \nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(\mathbf{x}^*) \\ -\lambda_i^* + u_i^2, \forall i = 1,2,...,m \\ g_i(\mathbf{x}^*) + v_i^2, \forall i = 1,2,...,m \\ \lambda_i^* g_i(\mathbf{x}^*), \forall i = 1,2,..,m \end{pmatrix} = \mathbf{0} . \qquad (4.2.2.1)$$

Where $\mathbf{u} = (u_1, u_2, ..., u_m)^T$ and $\mathbf{v} = (v_1, v_2, ..., v_m)^T$. And for a given initial point $\mathbf{y}_c$, we have by Newton (Algorithm 2.2.1), $J(\mathbf{y}_k)\mathbf{s}_k = -F(\mathbf{y}_k)$ and $\mathbf{y}_{k+1} := \mathbf{y}_k + \mathbf{s}_k$, where $k \in \mathbf{N}_0 = \{0\} \cup \mathbf{N}$ generates a sequence $\{\mathbf{y}_k\}$. Finally if our initial point $\mathbf{y}_c$ is good enough we get the solution $\mathbf{y}^* = (\mathbf{x}^*, \boldsymbol{\lambda}^*, \mathbf{u}^*, \mathbf{v}^*)$ of (4.2.1.1). Then $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ is the KKT point of (P_c).

If $\mathbf{x}^* \in S$, for some $\boldsymbol{\lambda}^* \in \mathbf{R}_+^m$, then $\mathbf{x}^*$ is a candidate for the solution of (P_c). Moreover, if L is assumed to be convex, then $\mathbf{x}^*$ is also an optimal solution of (P_c).

4.2.3. Newton's Method for Mixed Constrained Minimization

The optimization problem is,

$$(p_c) \qquad f(\mathbf{x}) \to \min \quad , \quad \mathbf{x} \in S$$
$$S := \{\mathbf{x} \in U : g_i(\mathbf{x}) \le \mathbf{0}, \forall i = 1,2,3,...m \ , h_j(\mathbf{x}) = \mathbf{0}, \ \forall j = 1,2,3,...,k\} .$$

And the corresponding unconstrained problem is, for $\boldsymbol{\lambda} \in \mathbf{R}_+^m, \boldsymbol{\mu} \in \mathbf{R}^k$

$$(p_{\lambda,\mu}) \qquad L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \to \min \quad , \quad \mathbf{x} \in U.$$

The necessary condition for $\mathbf{x}^*$ to be an optimal solution of $(\mathbf{P}_{\leq})$ is the KKT condition, which is,

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^{k} \mu_j^* \nabla h_j(\mathbf{x}^*) = \mathbf{0}$$
$$\lambda_i^* \geq 0, \forall i = 1,2,...,m$$
$$g_i(\mathbf{x}^*) \leq \mathbf{0}, \forall i = 1,2,...,m$$
$$h_j(\mathbf{x}^*) = \mathbf{0}, \forall j = 1,2,...,k$$
$$\lambda_i^* \mathbf{g}_i(\mathbf{x}^*) = 0, \forall i = 1,2,..,m.$$

Introducing slack variables in KKT condition, the following system of nonlinear equations.

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^{k} \mu_j^* \nabla h_j(\mathbf{x}^*) = \mathbf{0}$$
$$-\lambda_i^* + u_i^2 = 0, \forall i = 1,2,...,m$$
$$g_i(\mathbf{x}^*) + v_i^2 = \mathbf{0}, \forall i = 1,2,...,m$$
$$h_j(\mathbf{x}^*) = \mathbf{0}, \forall j = 1,2,...,k$$
$$\lambda_i^* g_i(\mathbf{x}^*) = \mathbf{0}, \forall i = 1,2,..,m.$$

This is a system of nonlinear equations with $(n+3m+k)$-equations and $(n+3m+k)$-unknowns. This is solvable by Newton's numerical methods. Now set

$$F(\mathbf{y}) := \begin{pmatrix} \nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* g_i(\mathbf{x}^*) + \sum_{j=1}^{k} \mu_j^* \nabla h_j(\mathbf{x}^*) \\ -\lambda_i^* + u_i^2, \forall i = 1,2,...,m \\ g_i(\mathbf{x}^*) + v_i^2, \forall i = 1,2,...,m \\ h_j(\mathbf{x}^*), \forall j = 1,2,...,k \\ \lambda_i^* \mathbf{g}_i(\mathbf{x}^*), \forall i = 1,2,..,m \end{pmatrix} = 0 \qquad (4.2.1.1)$$

Where $\mathbf{u} = (u_1, u_2, ..., u_m)^T$ and $\mathbf{v} = (v_1, v_2, ..., v_m)^T$. And for a given initial point $\mathbf{y}_c$, we have by Newton's Algorithm (2.2.1), $J(\mathbf{y}_k)\mathbf{s}_k = -F(\mathbf{y}_k)$ and $\mathbf{y}_{k+1} := \mathbf{y}_k + \mathbf{s}_k$, where $k \in \mathbf{N}_0 = \{0\} \cup \mathbf{N}$ generates a sequence $\{\mathbf{y}_k\}$.

Finally if our initial point $\mathbf{y}_c$ is good enough, we get the solution $\mathbf{y}^* = (\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*, \mathbf{u}^*, \mathbf{v}^*)$ of (4.2.1.2). Then $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ is the KKT point for $(P_{\leq})$. If $\mathbf{x}^* \in S$, for some $\boldsymbol{\lambda}^* \in \mathbf{R}_+^m$,

$\mu^* \in \mathbf{R}^k$, then $\mathbf{x}^*$ is a candidate for the solution of $(P_{\leq})$. Moreover, if L is assumed to be convex, then $\mathbf{x}^*$ is also an optimal solution of $(P_{\leq})$.

Example 1:

$$(x_1^2 + x_2^2) \to \min \ , \ (x_1, x_2) \in S,$$
$$S := \{(x_1, x_2) \in \mathbf{R}^2 : x_1 + x_2 = 1\}$$

The real solution for this problem is $\mathbf{x}^* = (1/2, 1/2)^T$ and the Lagrange multiplier is $\lambda = -1$. Now let us test the method by this problem. Take the initial point $(1, 1.5, -1.5)^T$

The Lagrange functional is given by,

$L(x_1, x_2, \lambda) = f(x) + \lambda g(x) = x_1^2 + x_2^2 + \lambda(x_1 + x_2 - 1)$. And

$$\nabla_x L(x_1, x_2, \lambda) = \nabla f(\mathbf{x}) + \lambda \nabla g(\mathbf{x}) = \begin{pmatrix} 2x_1 + \lambda \\ 2x_2 + \lambda \end{pmatrix}, \text{where}$$

$f(\mathbf{x}) = x_1^2 + x_2^2, g(\mathbf{x}) = x_1 + x_2 - 1$

and $\lambda \in \mathbf{R}$. The KKT condition is

$$\begin{pmatrix} 2x_1 + \lambda \\ 2x_2 + \lambda \\ x_1 + x_2 - 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$ And our functional, F is given by

$$F(\mathbf{x}) := \begin{pmatrix} 2x_1 + \lambda \\ 2x_2 + \lambda \\ x_1 + x_2 - 1 \end{pmatrix} = \mathbf{0} .$$

This is a system of equations with three variables and three equations. Then the solution for $F(\mathbf{x}) = \mathbf{0}$ using the Newton method is given by the following table, which is calculated by (program1) for the first five iterations.

k	s_k	y_k
0	(-0.5, -1. ,0.5)	(1 ,1.5 , -1.5)
1	(0., 0., 0.)	(0.5 ,0.5 , -1)
2	(0., 0., 0.)	(0.5 ,0.5 , -1)
3	(0., 0., 0.)	(0.5 ,0.5 , -1)
4	(0., 0., 0.)	(0.5 ,0.5 , -1)
5	(0., 0., 0.)	(0.5 ,0.5 , -1)

Table (4.2.2a)

Therefore the KKT point is $\mathbf{x}^* = (1/2, 1/2, -1)^T$, with $x_1 = 0.5, x_2 = 0.5$ and $\lambda = -1$. Since we have $(x_1, x_2) = (0.5, 0.5) \in S$. Moreover the functional L is convex. There for the point $(0.5, 0.5)^T$ is optimal solution of the given constrained problem.

We are very lucky in this case, since the solution is obtained in the second iteration. This is due to the initial point is close enough to the solution and the Jacobean is nonsingular.

Example 2:

$$(x_1^2 - x_2) \rightarrow \min \quad , (x_1, x_2) \in S \ ,$$
$$S = \{(x_1, x_2) \in \mathbf{R}^2 : x_1 + x_2 \leq 1\}$$

This is a constrained optimization with only inequality constraint. And let us apply the Newton's method to solve this problem. The true solution for this problem is $\mathbf{x}^* = (-0.5, 1.5)^T$, for $\lambda = 1$.

The Lagrangian of the problem is,

$L(x_1, x_2, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x}) = x_1^2 - x_2 + \lambda(x_1 + x_2 - 1)$. And

$\nabla_x L(x_1, x_2, \lambda) = \nabla f(\mathbf{x}) + \lambda \nabla g(\mathbf{x}) = \begin{pmatrix} 2x_1 + \lambda \\ -1 + \lambda \end{pmatrix}$ where,

$f(\mathbf{x}) = x_1^2 - x_2$, $g(\mathbf{x}) = x_1 + x_2 - 1$ and $\lambda \in R_+$. The KKT condition is

$$\begin{pmatrix} 2x_1 + \lambda \\ -1 + \lambda \\ \lambda(x_1 + x_2 - 1) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$x_1 + x_2 - 1 \leq 0, \mu \geq 0$. The functional F is given by

$$F(\mathbf{y}) = \begin{pmatrix} 2x_1 + \lambda \\ -1 + \lambda \\ \lambda(x_1 + x_2 - 1) \\ x_1 + x_2 - 1 + u^2 \\ -\lambda + v^2 \end{pmatrix} = \mathbf{0}$$. This is a system of nonlinear equations containing five equations and five unknowns. Then by Newton's method, following table is calculated by (program 1) for the first five iterations. Where $\mathbf{y}_0 = (x_1, x_2, \lambda, u, v)^T = (2, 2, 2, 2, 2)^T$ is the initial point.

k	s_k	y_k
0	(-2.5, 1, -1, -1.375, -0.75)	(2, 2, 2, 2, 2)
1	(0., -1.5, 0. , -0.3125 , -0.225)	(-0.5, 3, 1, 0.625, 1.25)
2	(0., 0., 0., -0.15625, -0.0246951)	(-0.5, 1.5, 1, 0.3125 , 1.025)
3	(0., 0., 0., -0.078125, -0.000304832)	(-0.5 , 1.5, 1, 0.15625 , 1.0003)
4	(0., 0., 0., -0.0390625, -4.64611×10^{-8})	(-0.5, 1.5, 1, 0.078125, 1.)
5	(0., 0., 0., -0.0195313, -1.11022×10^{-15})	(-0.5, 1.5, 1, 0.0390625, 1.)

Table (4.2.2b)

From the fifth row we have $\mathbf{y}^* = (x_1, x_2, \lambda, u, v)^{\mathrm{T}} = (-0.5, 1.5, 1, 0.0390625, 1)^{\mathrm{T}}$ is the solution for $\mathrm{F}(\mathbf{x}) = \mathbf{0}$ .And $(x^*_1, x^*_2, \lambda^*) = (-0.5, 1.5, 1)$ is the KKT point for the given constrained optimization problem. Since L is convex the solution $(x^*_1, x^*_2) = (-0.5, 1.5)$ is also the optimal solution of the given constrained problem. The minimum value of the objective function is $-5/4$.

Example 3:

$$\min\ (x_1 + x_2)\quad ,\ (x_1, x_2) \in \mathrm{S}\ ,$$
$$\mathrm{S} := \{(x_1, x_2) \in \mathbf{R}^2 : x_1 - x_2 = 1,\ x_2 \geq 0\}$$

The true solution of this problem is $\mathbf{x}^* = (1, 0)^T$ and the Lagrange multipliers are $\lambda_1 = 2, \lambda_2 = -1$. Solve this problem using the Newton's method. Take the initial point (0.5, -1, 1, -0.5, 1, 0.5).

The Lagrange functional is given by,

$$\mathrm{L}(x_1, x_2, \lambda_1, \lambda_2) = \mathrm{f}(\mathbf{x}) + \lambda_1\, \mathrm{g}(\mathbf{x}) + \lambda_2\, \mathrm{h}(\mathbf{x}) = x_1 + x_2 + \lambda_1(-x_2) + \lambda_2(x_1 - x_2 - 1)\text{, where}$$

$$\mathrm{f}(\mathbf{x}) = x_1 + x_2\,,\ \mathrm{g}(x) = -x_2\,,\ \mathrm{h}(\mathbf{x}) = x_1 - x_2 - 1 \text{ and } \lambda_2 \in \mathbf{R}, \lambda_1 \in \mathbf{R}_+\,.$$

$$\nabla_x \mathrm{L}(x_1, x_2, \lambda, \mu) = \nabla \mathrm{f}(\mathbf{x}) + \lambda_1 \nabla \mathrm{g}(\mathbf{x}) + \lambda_2 \nabla \mathrm{h}(\mathbf{x}) = \begin{pmatrix} 1 + \lambda_2 \\ 1 - \lambda_1 - \lambda_2 \end{pmatrix}\text{.The KKT condition is}$$

$$\begin{pmatrix} 1 + \lambda_2 \\ 1 - \lambda_1 - \lambda_2 \\ x_1 - x_2 - 1 \\ \lambda_1(-x_2) \end{pmatrix} = \mathbf{0}\,,\ x_1 - x_2 \leq 1, \lambda_1 \geq 0\,.$$

Now adding slack variables in the KKT condition, our functional F becomes

$$F(\mathbf{y}) := \begin{pmatrix} 1+\lambda_2 \\ 1-\lambda_1-\lambda_2 \\ x_1-x_2-1 \\ \lambda_1(-x_2) \\ -x_2+u^2 \\ -\lambda_1+v^2 \end{pmatrix} = \mathbf{0}.$$

This is a system of nonlinear equations with six unknowns and six equations. Where u and v are real numbers. Taking the initial point $(0.5,-1,1,-0.5,1,0.5)^T$. Finally we have the following table calculated by (program1) for the first six iterations.

k	s_k	y_k
0	(1.5, 2, 1, -0.5, 0., 1.75)	(0.5, -1, 1, -0.5, 1, 0.5)
1	(-1, -1, 0., 0., -0.5, -0.680556)	(2, 1, 2, -1, 1, 2.25)
2	(0. ,0. ,0. ,0., -0.25, -0.147554)	(1, 0., 2, -1, 0.5, 1.56944)
3	(0., 0., 0., 0., -0.125, -0.007656608)	(1, 0., 2, -1, 0.25, 1.42189)
4	(0., 0., 0., 0., -0.0625, -0.0000207234)	(1, 0., 2, -1, 0.125, 1.41423)
5	(0., 0., 0. ,0., -0.03125, -1.51837×10^{-10})	(1 ,0., 2., -1, 0.0625, 1.41421)

Table (4.2.2c)

From the last row of he table above we get the solution of $F(\mathbf{x}) = \mathbf{0}$ is

$\mathbf{y}^* = (x_1, x_2, \lambda_1, \lambda_2, u, v) = (1, 0., 2, -1, 0.0625, 1.41421)$ and
$(x_1, x_2, \lambda_1, \lambda_2) = (1, 0., 2, -1)$ is the KKT point of the given mixed constrained problem. Moreover, since L is convex functional (linear) we have see that the point $(x_1, x_2)^T = (1, 0)^T$ is the optimal solution of the problem. And the minimal value of the objective function is 1.

Conclusion

As we have already seen, in this seminar report the Newton method is depends on the initial point (approximations), that is if our initial approximation is not good enough, t-hen the algorithm may not convergent all to the real solution. That is precisely to mean hat the Newton method is locally convergent. In order to converge for any initial point the method has to be modified. This is not contained in his paper. This is one limitation of this seminar report.

In this paper we need the analytic Jacobeans, Hessians and Gradients to implement our algorithm. But sometimes these derivatives are not available. In this case we may need finite difference approximations to these derivatives and which is not considered in this paper. This is another limitation of the paper.

Finally if we have reasonably good starting point, we can solve a system of nonlinear equations, nonlinear unconstrained minimization and nonlinear constrained minimizati-on under some suppositions.

```
Programming 1
Newton ' s Methos for solving the function F (x) = 0

Remove["Global *"]
(* Input *)
(*  1) Input of number of variables,i.e. dimenssion of the space *)
n = 1;
(*  2) Inputfor n≥2  *)
If[n ≥ 2, x0 = {-2.5, 2.5}; fF = {x2 * e^x1 - 1, x1^2 + x2^2 - 2}];
(*  3) Input for n=1   *)
If[n == 1, x0 = 2; f[x_] = x^3 + x^2 - 2 * x + 0.5];
(*  4) number of desired iterations *)
q = 10;
(*  5) Printing the input  *)
If[n ≥ 2,
  cx0[1] = Sequence["(", x0[[1]], ","];
  cx0[m_] := Sequence[cx0[m - 1], x0[[m]], ","];
  Print["Initial point x0 =", Sequence[cx0[n - 1], x0[[n]], ")"]];
  Print["The given vector function is F(x)=", MatrixForm[fF]] ];
(*  6) Assignments  *)
Table[u[i], {i, 0, n}];
Table[s[i], {i, 0, n}];
x0 = x0 // N;

If  n ≥ 2 ∧ n == Length[x0] ∧ n == Length[fF],
 δfF = Table[D[fF[[i]], xj], {i, 1, n}, {j, 1, n}];
 Print["The Jacobian is calculated as J(x)=", MatrixForm[δfF]];
 δf[z_] := δfF /. Table[xi → z[[i]], {i, 1, n}];
 f[z_] := fF /. Table[xi → z[[i]], {i, 1, n}];
 (*  7) Iteration for n≥2  *)
 v = x0;
 k = 0;
 While k ≤ q,
  w = LinearSolve[δf[v], -f[v]];
  u[k] = v;
  s[k] = w;
  u[q + 1] = u[q];
  lin[k] = Line[{u[k], u[k + 1]}];
  cu[1, k] = {u[k][[1]], ","};
  cu[m_, k] := Append[Append[cu[m - 1, k], u[k][[m]]], ","];
  cs[1, k] = {s[k][[1]], ","};
  cs[m_, k] := Append[Append[cs[m - 1, k], s[k][[m]]], ","];
  v = v + w;
  k = k + 1;
```

```
(*  9) Results in table form for n≥2    *)
  Print["Results of iteration steps in the table form"];
  Print[
   TableForm[
    Prepend[
     Prepend[
      Table[{k, MatrixForm[{Append[cs[n-1, k], s[k][[n]]]}],
        MatrixForm[{Append[cu[n-1, k], u[k][[n]]]}]}, {k, 0, q}],
      {"---", "------------", "---------------"}],
     {"k", "                    s[k]", "                u[k]"}]]] ];
If[n ≠ 1 ∧ (Length[x0] ∨ n ≠ Length[fF]), Print["Input is not correct"];
  Abort[]];

(*   All Calculations for n=1     *)
If[n == 1 ∧ n == Length[{x0}],
  Print["a)  Initial point x0 = ", x0];
  Print["b)   The given function is f(x)=  ", f[x]];
  Print["c)  The derivative is f'(x)=  ", f'[x]];
  (*  Iterations   *)
  v = x0;
  k = 0;
  While[k ≤ q,
   w = Solve[f'[v] * ξ == -f[v], ξ];
   ω = ξ /. w[[1]];
   u[k] = v;
   s[k] = ω;
   lin[k] = Line[{{u[k], 0}, {u[k], f[u[k]]}}];
   v = v + ω;
   k = k + 1;
   (*  Results in table form     *)
   Print["d)  The results of iterations steps in the table
      form:"];
   Print[];
   Print[
    TableForm[
     Prepend[
      Prepend[
       Table[{k, {{s[k]}}, {{u[k]}}}, {k, 0, q}],
       {"---", "------", "-----"}],
      {"k", "    s[k]", "              u[k]"}]]] ];
```

```
Program 2
Newton ' s Method for Unconstarined Minimization
Remove["Global *"]
(* Input *)
(*  1) Input of number of variables,
 i.e. dimenssion of the space *)
n = 2;
(*  2) Inputfor n≥2  *)
If[n ≥ 2, x0 = {2, 3}; f = x1^2 + x2^2 + 8];
(*  3) Input for n=1   *)
If[n == 1, x0 = 2; f[x_] = x^3 - x^2 - 3*x - 1];
(*  4) number of desired iterations *)
q = 10;
(*  ) Printing the input  *)
If[n ≥ 2,
  cx0[1] = Sequence["(", x0[[1]], ","];
  cx0[m_] := Sequence[cx0[m - 1], x0[[m]], ","];
  Print["Initial point x0 =", Sequence[cx0[n - 1], x0[[n]], ")"]];
  Print["The given function is f(x)=", f] ];
(*  5) Assignments  *)
Table[u[i], {i, 0, n}];
Table[s[i], {i, 0, n}];
x0 = x0 // N;
If n ≥ 2 ∧ n == Length[x0],
 δf = Table[D[f, xi], {i, 1, n} ];
 δ2f = Table[D[f, xi, xj], {i, 1, n}, {j, 1, n}];
 Print["Then the Gradient is calculated as gradf(x)= "
   MatrixForm[δf] ];
 Print["Then the Hessian is calculated as "];
 Print[];
 Print["                    H(x) = ", MatrixForm[δ2f]];
 Print[];
 δgf[z_] := δf /. Table[xi → z[[i]], {i, 1, n}];
 δhf[z_] := δ2f /. Table[xi → z[[i]], {i, 1, n}];
 ff[z_] := f /. Table[xi → z[[i]], {i, 1, n}];
 (*  6) Iteration for n≥2  *)
 v = x0;
 k = 0;
 While k ≤ q,
  w = LinearSolve[δhf[v], -δgf[v]];
```

```
u[k] = v;
   s[k] = w;
   u[q + 1] = u[q];
   lin[k] = Line[{u[k], u[k + 1]}];
   cu[1, k] = {u[k][[1]], ","};
   cu[m_, k] := Append[Append[cu[m - 1, k], u[k][[m]]], ","];
   cs[1, k] = {s[k][[1]], ","};
   cs[m_, k] := Append[Append[cs[m - 1, k], s[k][[m]]], ","];
   v = v + w;
   k = k + 1;
   (*  7) Results in the table form for n≥2      *)
   Print["Results of iteration steps in the table form"];
   Print[TableForm[
     Prepend[
      Prepend[
       Table[{k, MatrixForm[{Append[cs[n - 1, k], s[k][[n]]]}],
         MatrixForm[{Append[cu[n - 1, k], u[k][[n]]]}], ff[u[k]]},
        {k, 0, q}],
       {"---", "------------", "---------------", "--------"}],
      {"k", "              s[k]", "                 u[k]",
       "    f(u(k))"}]]];
If[n ≠ 1 ∧ n ≠ Length[x0], Print["Input is not correct"]; Abort[]];
(*    All Calculations for n=1      *)
If[n == 1 ∧ n == Length[{x0}],
 Print["a)  Initial point x0 = ", x0];
 Print["b)   The given function is f(x)=  ", f[x]];
 Print["c)  The derivative is f'(x)=  ", f'[x]];
 Print["d)  The second derivative is f''(x) =", f''[x]];
 (*  Iterations  for n=1 *)
 v = x0;
 k = 0;
 While[k ≤ q,
  w = Solve[f''[v] * ξ == -f'[v], ξ];
  ω = ξ /. w[[1]];
  u[k] = v;
  s[k] = ω;
  lin[k] = Line[{{u[k], 0}, {u[k], f[u[k]]}}];
  v = v + ω;
  k = k + 1;
```

```
(*  Results in the table form      *)
  Print["d)  The results of iterations steps in the
     table form:"];
  Print[];
  Print[
   TableForm[
    Prepend[
     Prepend[
      Table[{k, {{s[k]}}, {{u[k]}}, f[u[k]]}, {k, 0, q}],
      {"---", "--------", "--------", "-------"}],
     {"k" , "      s[k]", "        u[k]" , "f(u(k))"}]]] ];
```

References

[1] Dennis, Jr., J.E., Schenable, R.B: *Numerical Methods for Unconstrained optimization and Nonlinear Equations.* SIAM, Society for Industrial and Applied Mathematics, Philadelphia (1996).

[2] Deumlich, R.: *Optimization and Theory of Approximation.* Lecture Notes, Addis Ababa University (1997).

[3] Deumlich, R.: *Functional Analysis I.* Lecture Notes, Addis Ababa University (1997).

[4] Deumlich, R.: *Numerical Solutions of Optimization Problems By Means of Mathematica.* Unpublished Paper, Addis Ababa University (2003).

[5] M.K.Jain, S.R.K Iyengar, R.K.Jain: *Numerical Methods for Scientific and Engineering Computation.* 5th Edition, 2007.

[6] J.Stoer, R.Bulirsch: *Introduction to Numerical Analysis.* 3rd Edition, January 2002.

[7] Tadesse Bekeshe: *Newton's Method for Nonlinear Equations and Unconstrained Minimization.* Seminar Report, Addis Ababa University 2001

[8] Hissab: *Solution of Nonlinear Equations by Newton's Method and Application to Unconstrained optimization problem.* Ethiopian Journal of Mathematics, Volume 24, No. 2, 2004.