Hybrid Methods for Optimization Problems with Positive Semi–Definite Matrix Constraints

By

Suliman Saleh Al–Homidan

PhD Thesis

Department of Mathematics and Computer Science

University of Dundee

Dundee

December 2003

Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my sincere gratitude to my supervisor Professor Roger Fletcher for his kind supervision, suggestions and advice throughout the preparation of this thesis. His patience has been a constant source of inspiration throughout this work.

I would like to thank King Saud University for its financial support.

I would also like to thank the following:

My friends in the Department of Mathematics and Computer Science, University of Dundee, in particular George Mathai for his help in overcoming language difficulties.

My family and friends, with special regards to my beloved wife for her moral support.

Declaration

I declare that I am the author of this thesis; that I have consulted every reference cited except those for which I have indicated otherwise; that the work of which this thesis is a record has been done by myself, and that it has not been previously accepted for a higher degree.

Suliman Saleh Al–Homidan

Certification

This is to certify that Suliman Saleh Al–Homidan has complied with all the requirements for the submission of his PhD thesis to the University of Dundee.

Professor Roger Fletcher

Abstract

Three problems are handled in this thesis, all of which are involved with the positive semi–definite matrix as a convex constraint set. One problem is the Euclidean distance problem and the other two problems are different forms of the educational testing problem. Projection methods which solves least distance problems subject to the intersection of convex sets are used to solve these problems. It is found that the methods are globally convergent, but the rate of convergence is slow. However these methods do have the capability of determining the correct rank of the solution matrix, and this can be done in relatively few iterations. On the other hand there are conventional unconstrained and l_1 Sequential Quadratic Programming (SQP) methods which enable rapid convergence to be obtained. However, the correct rank is needed by these methods. Hence is is the purpose of this thesis to study hybrid methods. These hybrid methods have two different modes of operation. One is a projection method which provides global convergence and enables the correct rank to be determined. The other is either a quasi–Newton method or a nonlinear programming method, depending on the problem. An important feature concerns the interfacing of these modes of operation. Thus it has to be decided which method to use first, and when to switch between methods. Also it may not be straightforward, as we shall see here, to use the output of one method to start the other method. Difficulties such as these are addressed in the thesis. Many comparative numerical results are reported.

Chapter 0

Introduction

This thesis considers methods for solving certain optimization problems in which there are constraints on the variables. Many advances have taken place in this subject over the last forty years or so. There are now effective methods for situations in which the objective and constraint functions are smooth functions. Under reasonable assumptions, these methods can be shown to converge globally (that is from any starting point) to a point which satisfies optimality conditions for the problems. Also the rate of convergence can often be shown to be superlinear. Some progress has also been made for problems in which non–smooth functions occur. If these functions are a composition of a convex polyhedral function and a smooth function, then again globally and superlinear convergent methods have been suggested. This thesis addresses a rather more difficult situation in which some matrix, defined in terms of the problem variables, has to be positive semi–definite. One way to handle this problem is to impose a functional constraint in which the least eigenvalue of the matrix is non–negative. However, if there are multiple eigenvalues at the solution which is usually the case, such a constraint is usually non–smooth, and this non–smoothness cannot be modelled by a convex polyhedral composite function. An important factor is the determination of the multiplicity of the zero eigenvalues, or alternatively the *rank* of the matrix at the solution. If this rank is known it is usually possible to solve the problem by conventional techniques.

In this thesis the positive semi–definite matrix constraint is handled in a different way, by regarding it as a convex set. There are certain methods, known as *projection methods,* which can be used to solve least distance problems constrained by the intersection of convex sets. In this thesis the application of such methods to certain problems with positive semi–definite matrix constraint is considered. It is found that the methods are globally convergent, but the rate of convergence is linearly or slower. It is this latter feature that has probably contributed to the relatively little interest that has been shown in such methods. However it is demonstrated here that the methods do have the capability of determining the correct rank of the solution matrix, and this can be done in relatively few iterations.

Thus we are led to study hybrid methods in this thesis. The hybrid method has two different modes of operation. One is a projection method which provides global convergence and enables the correct rank to be determined. The other is either a quasi–Newton method or a conventional nonlinear programming method, depending on the problem, which enables rapid convergence to be obtained. An important feature concerns the interfacing of these modes of operation. Thus it has to be decided which method to use first, and when to switch between methods. Also it may not be straightforward, as we shall see here, to use the output of one method to start the other method. Difficulties such as these are addressed in the thesis. Hybrid methods have often been used successfully in optimization, for example Powell [1970], Hald and Madsen [1981] and Al–Baali and Fletcher [1985].

There are two main problems that are addressed in this thesis. Firstly, there is the *Euclidean distance matrix problem* which arises in many experimental sciences. The problem is to find the best Euclidean distance matrix which approximates a given non–Euclidean distance matrix. For solving this problem two methods are given. One is a projection method which is globally convergent. The other method for solving the Euclidean distance matrix problem is a quasi–Newton method, in particular the BFGS method. This method is superlinearly convergent but requires a knowledge of a certain characteristic rank. Hence new methods are established for solving this problem using the advantage of both methods by switching from one method to the other in a suitable way.

The second problem we going to study in this thesis is the *educational testing problem* which arises in statistics. In this problem there is given a symmetric positive definite matrix and it is required to determine how much can be subtracted from the diagonal of that matrix and still retain a positive semi–definite matrix. In the standard form the l_1–norm is used to measure the amount subtacted from the diagonal. Unfortunately this problem is not in the correct format for projection methods to be used directly. However there is an ingenious device due to Glunt [1991] which transforms this problem to a related one in which a least distance measure is used. We are therefore able to study the application of projection methods to the problem with the least distance measure. Then Glunt's transformation is used to enable the original educational testing problem to be solved.

These methods are again seen to be typified by being globally and slow convergent. When the correct rank for the matrix is known we are also able to use the l_1 *Sequential Quadratic Programming* (SQP) method to solve both problems, and this converges at second order. Subsequently hybrid methods are investigated to combine the advantageous features of both methods.

0.1 Outline of the thesis

Chapter 1 provides a general background to the optimization problem. This chapter includes a brief review of linear algebra and other various results. The concept of convex cones and normal cones with some important convex sets are also given. This chapter also introduces the concept of feasibility along with various expressions for feasible directions and describes optimality conditions relating to positive semi–definite matrix constraints. Finally, this chapter is concluded by a description of the Newton, quasi–Newton and Sequential Quadratic Programming (SQP) methods.

Chapter 2 provides a background about the projection methods for solving certain linear and least distance convex programming problems in which the feasible region is the intersection of a convex sets. Such optimization problems potentially arise in many practical situations, for example in linear programming problems, although projection methods are not the best for solving such problems. Here we are interested in the case where one of the convex cones is related to a positive semi–definite matrix cone. This chapter includes a description of the von Neumann [1950], Dykstra [1983] and Han [1988] projection methods for solving least distance convex programming problems and Glunt [1991] method for solving linear convex programming problems.

The aim of Chapter 3 is to find the best Euclidean distance matrix which approximates a given non–Euclidean distance matrix. Some applications of the above problem are given along with the definition of the Euclidean distance matrix and its characterization. Various methods for solving this problem are considered including a projection algorithm described by Glunt, Hayden, Hong and Wells [1990] and some new unconstrained methods based on using quasi–Newton methods. Other projection methods are also given and at the end of this chapter numerical comparisons of these methods are described.

In Chapter 4 some new methods for solving the Euclidean distance matrix problem are considered. These methods are developed from the methods of Chapter 3 using a hybrid method. A feature of some interest is how to move between the two methods. Numerical comparisons are also given in this chapter.

Chapter 5 considers a problem in which the objective function is a least distance function subject to a positive semi–definite matrix constraint where the diagonal of the matrix is allowed only to change. Two methods are developed for solving this problem. Firstly, a projection algorithm is given for solving this problem which converges globally. Secondly an implementation of the l_1 Sequential Quadratic Programming (SQP) method is used which converges quadratically. A transformation due to Fletcher [1985] is used to enable this method to be used. This chapter also includes a hybrid method between the projection method and the l_1 SQP method in a similar way to Chapter 4. Finally, numerical comparisons of these methods are carried out in the end of the chapter.

The problem to be considered in Chapter 6 is the educational testing problem. Previous attempts to solve the problem are described. The definition of the educational testing problem is given. This chapter also contains projection algorithm and l_1 SQP methods. At the end of this chapter numerical comparisons of these methods are given.

In Chapter 7 new methods for solving the educational testing problem are considered. The methods described here are similar to those in Chapter 4 and depend upon the two methods of Chapter 6 using a hybrid method. The projection method converges globally but often converges at very slow order. The l_1 SQP method converges quadratically but often requires the correct rank. Combining these two methods together produces a method with a better speed of convergence. Therefore this chapter describes two hybrid methods and also gives numerical comparisons.

The achievements of the thesis are summarized in Chapter 8 and suggestions for further research are discussed.

0.2 Notation

If $f(\mathbf{x})$ is continuously differentiable (C^1) then for any point $\mathbf{x}$ the vector of first partial derivatives, or gradient vector is referred to by $\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x})$ and ∇ denotes the gradient operator $(\partial/\partial x_1 \, , \, \ldots \, , \, \partial/\partial x_n)^T$. If $f(\mathbf{x})$ is twice continuously differentiable (C^2) then there exists a matrix of second partial derivatives, or Hessian matrix, written $\nabla^2 f(\mathbf{x})$ which is square and symmetric.

Superscript "k" generally denotes quantities related to the kth iterate. For instance $f^{(k)} = f(\mathbf{x}^{(k)})$, $\mathbf{g}^{(k)} = \mathbf{g}(\mathbf{x}^{(k)})$, *etc*, and $f^* = f(\mathbf{x}^*)$, $\mathbf{g}^* = \mathbf{g}(\mathbf{x}^*)$, *etc.*

Throughout this thesis the lower case boldface letters such as $\mathbf{x}$, $\mathbf{y}, \mathbf{v}$ are used to denote vectors. Matrices are denoted by capital letters such as A, B, C and sometimes A written as $A = [a_{ij}]$.

Chapter 1

Optimization review

1.1 Introduction

The purpose of this chapter is to provide a general background to the optimization problem. This chapter includes some important concepts of optimization theory along with a description of the quasi–Newton method and the Sequential Quadratic Programming (SQP) method.

Section 1.2 contains a brief review of linear algebra and other various results. The concept of convex cones is given in Section 1.3. Also in that section two important convex cones are given. These are the cone of all $n \times n$ symmetric positive semi–definite matrices and the convex cone which is a subset of the positive semi–definite matrix cone. Section 1.3 also includes expressions for the normal cones of these convex cones. In Section 1.4 the concept of feasibility is described, along with various expressions for feasible directions. Section 1.5 describes optimality conditions relating to positive semi–definite matrix constraints. In Section 1.6 some details of how Newton and quasi–Newton methods work are given together with a proof of second order convergence. The SQP method is an efficient method for solving nonlinear programming problems when first and second derivatives are available. This method is Newton's method applied to find the stationary point of a Lagrangian function. The SQP method converges locally at second order. The global properties of the SQP method are improved by associating it with an exact penalty function. This method is described in Section 1.7.

1.2 Various results

The analysis of the optimization methods in this thesis requires results from linear algebra along with some definitions of rates of convergence. These are reviewed below.

Definition 1.2.1 $(Inner\ \ product)$

If $A,\ B\ \in\ \Re^{n\times n}$ then their inner product is defined by

$$\langle\ A\ ,\ B\ \rangle\ =\ \sum_{i,j=1}^{n} a_{ij}b_{ij}\ =\ tr(A^TB).$$

where $tr(A^TB)$ means the trace of the matrix A^TB which is the sum of the elements on the diagonal of A^TB.

Here $\Re^{n\times n}$ denotes the space of all real $n\times n$ matrices. Also we distinguish between $Diag\ A$ which denotes the diagonal matrix whose entries are the diagonal elements of A, and $diag\ \mathbf{a}$ which is the diagonal matrix whose entries are the elements of vector $\mathbf{a}$. The *null space* of A is defined by $N(A)\ =\ \{\mathbf{x}\ \in\ \Re^n:\ A\mathbf{x}\ =\ \mathbf{0}\}$.

Definition 1.2.2 $(\ Frobenius\ \ norm)$

A useful matrix norm in $\Re^{n\times n}$ is the Frobenius norm defined by

$$\|A\|_F\ =\ \langle\ A\ ,\ A\ \rangle^{\frac{1}{2}}\ =\ \{\sum_{i,j=1}^{n}\ |a_{ij}|^2\ \}^{\frac{1}{2}}$$

Definition 1.2.3 $(Householder\ matrix)$

A matrix $Q\ \in\ \Re^{n\times n}$ is said to be orthogonal if $Q^TQ\ =\ I$. A particular Householder matrix may be defined by

$$Q\ =\ I\ -\ \frac{2}{\boldsymbol{\nu}^T\boldsymbol{\nu}}\ \boldsymbol{\nu}\boldsymbol{\nu}^T,\quad \boldsymbol{\nu}\ =\ [1,\ \ldots,\ 1,\ 1+\sqrt{n}]^T. \tag{1.2.1}$$

This Householder matrix is a special case for which if $\mathbf{e}\ =\ [1,1,\ldots,1]^T$ then

$$Q\mathbf{e}\ =\ \begin{bmatrix}\mathbf{0}\\ -\|\mathbf{e}\|_2\end{bmatrix}\ =\ \begin{bmatrix}\mathbf{0}\\ -\sqrt{n}\end{bmatrix}. \tag{1.2.2}$$

Definition 1.2.4 $(Irreducibly\ \ embeddable)$

If there exist n vectors $\mathbf{p}_1, \ldots, \mathbf{p}_n$ in $\Re^r$ $(r \leq n - 1)$ such that

$$a_{ij} = \| \mathbf{p}_i - \mathbf{p}_j \|_2^2 \quad (1 \leq i,j \leq n). \qquad (1.2.3)$$

for set of vectors in $\Re^r$ but not in $\Re^{r-1}$ then the points $\mathbf{p}_1, \ldots, \mathbf{p}_n$ and the matrix $A = [a_{ij}]$ are said to be irreducibly embeddable in $\Re^r$.

Definition 1.2.5 $(Positive\ \ definite\ \ matrices)$

An $n \times n$ symmetric matrix A is said to be positive definite if

$$\mathbf{x}^T A\mathbf{x} > 0 \quad \forall \mathbf{x} \in \Re^n \quad \mathbf{x} \neq \mathbf{0} \qquad (1.2.4)$$

and is denoted by $A > 0$. If the inequality in (1.2.4) replaced by $\mathbf{x}^T A\mathbf{x} \geq 0$ then A is said to be *positive semi–definite* and is denoted by $A \geq 0$.

Positive definite matrices are an important class of matrices and arise naturally in many applications. The above definition cannot be checked numerically. Equivalent definitions which can be checked are the following

i. All eigenvalues of $A > 0$.

ii. There exists a unique lower triangular $L \in \Re^{n\times n}$ such that $LL^T = A$ with $l_{ii} > 0$ *(Choleski factors).*

iii. LDL^T factors exist with $l_{ii} = 1$ and $d_{ii} > 0$.

If A is a positive definite matrix, then the largest entry in A is on the diagonal and the diagonal elements are all positive.

Definition 1.2.6 $(First\ and\ second\ order\ convergence)$

Let $\mathbf{x}^*$ be a local minimum point with error defined as

$$\mathbf{h}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}^*.$$

If $\mathbf{h}^{(k)} \to \mathbf{0}$ we have convergence. If the errors behave as

$$\frac{\|\mathbf{h}^{(k+1)}\|}{\|\mathbf{h}^{(k)}\|^p} \to a$$

where $a > 0$ then the order of convergence is defined to be p order. The most important cases are where $p = 1$ *(first order or linear convergence)* in which $a < 1$ must hold, and $p = 2$ *(second order or quadratic convergence)*. If $\frac{\|\mathbf{h}^{(k+1)}\|}{\|\mathbf{h}^{(k)}\|} \to 0$ then this is known as *superlinear convergence*. Often it is only possible to obtain bounds, for example

$$\frac{\|\mathbf{h}^{(k+1)}\|}{\|\mathbf{h}^{(k)}\|} \leq a$$

or

$$\mathbf{h}^{(k+1)} = O(\|\mathbf{h}^{(k)}\|)$$

for first order convergence and

$$\frac{\|\mathbf{h}^{(k+1)}\|}{\|\mathbf{h}^{(k)}\|^2} \leq a$$

or,

$$\mathbf{h}^{(k+1)} = O(\|\mathbf{h}^{(k)}\|^2).$$

for second order convergence.

1.3 Cones and normal cones

The concept of a convex cone and its properties are very useful when applying convex analysis, for instance the normal cone is important in the development of optimality conditions. In this section the notion of cones and normal cones is described.

Definition 1.3.1 (*Convex set and convex function*)

A subset C of $\Re^n$ is said to be a convex set if

$$\mathbf{x}_\lambda = (1-\lambda)\mathbf{x}_1 + \lambda\mathbf{x}_2 \in C$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in C$ and $0 \leq \lambda \leq 1$. A convex function $f(\mathbf{x})$ on the domain C is defined by the condition that for any $\mathbf{x}_1, \mathbf{x}_2 \in C$ it follows that

$$f(\mathbf{x}_\lambda) \leq (1-\lambda)f(\mathbf{x}_1) + \lambda f(\mathbf{x}_2) \qquad \forall\ \lambda \in [0,1]$$

where $\mathbf{x}_\lambda = (1-\lambda)\mathbf{x}_1 + \lambda\mathbf{x}_2$.

Definition 1.3.2 (*Convex cone*)

A subset K of $\Re^n$ is called a convex cone if and only if $\mathbf{x}_1, \mathbf{x}_2 \in K$, $\alpha, \beta \geq 0$ implies that $\alpha\mathbf{x}_1 + \beta\mathbf{x}_2 \in K$.

The set of all $n \times n$ symmetric positive semi–definite matrices

$$K_{\Re} = \{A : A \in \Re^{n\times n}, \;\; A^T = A \; and \; \mathbf{z}^T A \mathbf{z} \geq 0 \quad \forall\, \mathbf{z} \in \Re^n\} \tag{1.3.1}$$

is a convex cone of dimension $n(n+1)/2$. The dimension is the number of free parameters in a symmetric matrix A. Let $A, B \in K_{\Re}$ then $\mathbf{z}^T(\alpha A + \beta B)\mathbf{z} \geq 0 \quad \forall\, \mathbf{z} \in \Re^n$, and $\alpha, \beta \geq 0$. This is because $\alpha\mathbf{z}^T A\mathbf{z} \geq 0$, $\beta\mathbf{z}^T B\mathbf{z} \geq 0 \quad \forall\, \mathbf{z} \in \Re^n$, which implies that $\alpha A + \beta B \in K_{\Re}$. This proves that $K_{\Re}$ is a convex cone. (The subscript $\Re$ is used to distinguish this case from the restricted cone in the next paragraph).

Another convex cone which will be used for the projection method given in Chapter 3 is the set of all $n \times n$ symmetric positive semi–definite matrices with respect to M, where

$$M = \{\, \mathbf{x} \in \Re^n : \;\; \mathbf{e}^T\mathbf{x} = 0 \,\} \tag{1.3.2}$$

and $\mathbf{e} = [1, 1, \ldots, 1]^T \in \Re^n$. $K_{\Re}$ is subset of this set which may be denoted by

$$K_M = \{A : A \in \Re^{n\times n}, \;\; A^T = A \; and \; \mathbf{x}^T A \mathbf{x} \geq 0 \quad \forall\, \mathbf{x} \in M\} \tag{1.3.3}$$

which is a convex cone. Let $A, B \in K_M$ then

$$\mathbf{z}^T(\alpha A + \beta B)\mathbf{z} \geq 0 \quad \forall\, \mathbf{z} \in M, \;\; \alpha \geq 0 \; and \; \beta \geq 0. \tag{1.3.4}$$

This is because $\alpha\mathbf{z}^T A\mathbf{z} \geq 0$, $\beta\mathbf{z}^T B\mathbf{z} \geq 0 \quad \forall\, \mathbf{z} \in M$, which implies that $\alpha A + \beta B \in K_M$. Thus K_M is convex cone.

It is also convenient to define two other convex sets for the purposes of Chapters 5 and 6. If $F \in \Re^{n\times n}$ is any given symmetric positive definite matrix then define

$$K_{off} = \{A : A \in \Re^{n\times n}, \; A - Diag\, A = \bar{F}\}. \tag{1.3.5}$$

where $\bar{F} = F - Diag\, F$. This is the set of matrices whose off–diagonal elements are equal to those of F. Also, let $diag\, \mathbf{v} = Diag\, F$ then define

$$K_b = \{A : A \in \Re^{n\times n}, \;\; A = \bar{A} + diag\, \mathbf{x}, \;\; x_i \leq v_i \;\; i = 1, 2, ...n\} \tag{1.3.6}$$

where $\bar{A} = A - Diag\,A$. This is the set of matrices that is obtained by reducing the diagonal of A. K_{off} and K_b are convex subspaces.

Next, the concept of the normal cone, denoted by ∂K, is introduced which is of importance when deriving optimality conditions for problems which involve any convex set. If $\mathbf{a}$ is on the boundary of K, then a vector $\mathbf{x}$ is said to be normal to a convex set K at $\mathbf{a}$, if $\mathbf{x}$ does not make an acute angle with any line segment in K emanating from $\mathbf{a}$. Therefore any vector $\mathbf{x} \in \partial K(\mathbf{a})$ must satisfy $\langle \mathbf{y} - \mathbf{a}, \mathbf{x} \rangle \leq 0$ for every $\mathbf{y} \in K$, (see Figure 1.3.1). The set of all vectors $\mathbf{x}$ normal to K at $\mathbf{a}$ is called the normal cone to K at $\mathbf{a}$, and denoted by

$$\partial K(\mathbf{a}) = \{\mathbf{x} : \mathbf{x} \in \Re^n, \quad \langle \mathbf{y} - \mathbf{a}\,,\, \mathbf{x} \rangle \leq 0 \qquad \forall\, \mathbf{y} \in K\}. \tag{1.3.7}$$

Equivalently the normal cone can be defined by

$$\partial K(\mathbf{a}) = \{\mathbf{x} : \mathbf{x} \in \Re^n, \quad \langle\, \mathbf{x},\ \mathbf{a} \rangle = \sup_{\mathbf{y} \in K} \langle \mathbf{x},\ \mathbf{y} \rangle\}. \tag{1.3.8}$$

It is convenient to define $\partial K(\mathbf{a}) = \{0\}$ if $\mathbf{a}$ is interior to K, and $\partial K(\mathbf{a}) = \emptyset$ (the empty set) if $\mathbf{a}$ is exterior to K, this is consistent with (1.3.7) and (1.3.8).

Let K_1 and K_2 be convex sets in $\Re^n$ whose relative interiors have a point $\mathbf{a}$ in common. Then

$$\partial(K_1 \cap K_2)(\mathbf{a}) = \partial K_1(\mathbf{a}) + \partial K_2(\mathbf{a}) \tag{1.3.9}$$

(see [Rockafellar 1970]).

In this thesis we consider the case of the convex cone in which the elements are matrices instead of vectors and we use the matrix inner product in Definition 1.2.1. It follows from (1.3.7) that

$$\partial K(A) = \{B : B \in \Re^{n \times n} \ \ and \ \ \langle Z - A\,,\, B \rangle \leq 0 \qquad \forall\, Z \in K\} \tag{1.3.10}$$

where K is a matrix cone.

It follows from (1.3.8) that the normal cone for (1.3.1) is

$$\partial K_{\Re}(A) = \{B : B \in \Re^{n \times n}, \langle A, B \rangle = \sup_{V \in K_{\Re}} \langle V, B \rangle\}.$$

However since unsymmetric matrices in $\partial K_{\Re}$ are not of interest here it is more convenient to define $\partial K_{\Re}$ by restricting it to the symmetric normal cone

Figure 1.3.1: The normal cone ∂K for a convex cone K at point $\mathbf{a}$.

$$\partial K_{\Re}(A) = \{B : B \in \Re^{n\times n}, B = B^T, \langle A, B\rangle = \sup_{V\in K_{\Re}} \langle V, B\rangle\}. \tag{1.3.11}$$

The most interesting case concerns the elements of the boundary of $K_{\Re}$, since $\partial K_{\Re}(A) = \{0\}$ when A is interior to $K_{\Re}$ $(A > 0)$.

In the following a theorem due to Fletcher [1985] is given to show how to find the normal cone $\partial K_{\Re}(A)$ at A, such that A belongs to the boundary of $K_{\Re}$

Theorem 1.3.3

If the columns of Z are an orthonormal basis for the null space of A, and Λ is any symmetric positive semi–definite matrix, then an equivalent form to (1.3.11) where A lies on the boundary of $K_{\Re}$ is the following

$$\begin{aligned}\partial K_{\Re}(A) = \{B : B \in \Re^{n\times n}, B = B^T, B = &-Z\Lambda Z^T,\\ &\Lambda = \Lambda^T,\ \Lambda \geq 0\}.\end{aligned} \tag{1.3.12}$$

Proof

Consider $\sup_{V \in K_{\Re}} \langle V, B \rangle$ for fixed B, let $B = X \Omega X^T$ be the spectral decomposition of B with X being the orthogonal matrix of eigenvectors and $\Omega = diag\ [\omega_1, \omega_2, \ldots, \omega_n]$ the diagonal matrix of eigenvalues. Since A is positive semi–definite there exists $C = X^T V X$ which is positive semi–definite. Using Definition 1.2.1 of the inner product it follows that

$$\begin{aligned} \sup_{V \in K_{\Re}} \langle V, B \rangle &= \sup_{C \in K_{\Re}} \langle C, \Omega \rangle \\ &= \sup_{c_{ii} \geq 0} \sum c_{ii} \omega_i. \end{aligned}$$

This follows because

$$\begin{aligned} \langle V, B \rangle &= tr(VB) \\ &= tr(VXX^T BXX^T) \\ &= tr(X^T VXX^T BX) \\ &= tr(C\Omega). \end{aligned}$$

Hence

$$\sup_{V \in K_{\Re}} \langle V, B \rangle = 0 \qquad iff \ \ \omega_i \leq 0 \qquad \forall\ i \tag{1.3.13}$$

and this is equivalent to $B \leq 0$ since ω_i are the eigenvalues of B. Hence an equivalent form to (1.3.11) is

$$\partial K_{\Re}(A) = \{B : B \in \Re^{n \times n}, B = B^T,\ \langle A, B \rangle = 0,\ B \leq 0\}. \tag{1.3.14}$$

Let $A = Y \Lambda_r Y^T$, with Λ_r being the diagonal matrix whose elements are the nonzero eigenvalues of A and the columns of Y are the corresponding orthonormal set of eigenvectors, so that $[Y \quad Z]$ is an orthogonal matrix. Express B as

$$B = [Y \quad Z] \begin{bmatrix} R & S \\ S^T & T \end{bmatrix} [Y \quad Z]^T. \tag{1.3.15}$$

Since $\langle A,\ B\rangle = 0$ then $tr(\Lambda_r Y^T BY) = 0$. The diagonal elements of $Y^T BY$ are zero because Λ_r is positive definite and diagonal. Also from (1.3.15) $Y^T BY = R$ so it follows that R has zero diagonal elements. Hence from (1.3.14) $B \leq 0$ implies that $R = \mathbf{0}$, and thus $T \leq 0$. Therefore from (1.3.15) $B = ZTZ^T$, and (1.3.12) follows since $\Lambda = -T$ $\Box$.

Example 1.3.4

If $n = 2$ then the cone in (1.3.1) becomes

$$K_{\Re} = \{A : A = \begin{bmatrix} x & z \\ z & y \end{bmatrix} \ x \geq 0,\ y \geq 0,\ xy \geq z^2 \ \ and \ \ x, y, z \in \Re\}$$

and is illustrated in Figure 1.3.2. Clearly the matrices in the interior of the cone are positive definite, whereas those on the boundary are singular. For example the matrix

$$\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

on the boundary is positive semi–definite. Then $Z = [1 \quad 1]^T$ and $\Lambda = [\alpha] \geq 0$, so the normal cone (1.3.12) at this point is

$$\partial K_{\Re}(\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}) = \{B : B = -\alpha \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},\ \alpha \geq 0\}.$$

The normal cone for $K_{off} \cap K_b$ is given in the following.

Theorem 1.3.5

Figure 1.3.2: The positive semi–definite matrix cone $K_{\Re}$.

Let $F \in \Re^{n\times n}$ be a given symmetric positive definite matrix and define K_{off} and K_b as in (1.3.5) and (1.3.6) respectively. Let $A \in K_{off} \cap K_b$. Then

$$\partial(K_{off} \cap K_b)\ (A) = \{B : B \in \Re^{n\times n}, \\ \left\{\begin{array}{lllll} b_{ii} & \geq 0 & if & x_i & = v_i \\ b_{ii} & = 0 & if & x_i & < v_i \end{array}\right\} \ i = 1, ..., n\}. \tag{1.3.16}$$

where $A = \bar{A} + diag\ \mathbf{x}$.

Proof

From the normal cone definition (1.3.10) it is clear that

$$\partial K_{off}(\bar{A} + diag\ \mathbf{x}) \ = \ \{B : B \ = \begin{bmatrix} 0 & b_{21} & \dots & b_{n1} \\ b_{21} & 0 & \dots & b_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \dots & 0 \end{bmatrix} \} \tag{1.3.17}$$

because $\bar{Z} = \bar{A} = \bar{F}$ implies $Z - A = \mathbf{0} \quad \forall\, i \neq j,$ in (1.3.10) where $\bar{Z} = Z - Diag\ Z.$

Consider

$$\sup_{Z \in K_{off} \cap K_b} \langle B, Z \rangle$$

and assume for some i that $b_{ii} < 0$. Let $z_{ii} = a_{ii} - \beta$ then $Z \in K_{off} \cap K_b$. By making β sufficiently large we can make $\langle B, Z \rangle$ as large as we like. Thus, if $b_{ii} < 0$ for any i, we have that

$$\langle A, B \rangle \ = \ \sup_{Z \in K_{off} \cap K_b} \langle B, Z \rangle \ = \ \infty.$$

Now suppose $b_{ii} \geq 0 \quad \forall\, i$. Then from the normal cone definition (1.3.8)

$$\begin{aligned} \partial K_{off} \cap K_b(\bar{A} + diag\ \mathbf{x}) \ &= \ \{B : B \ \in \ \Re^{n \times n}, \\ &\langle B, \bar{A} \ + \ diag\ \mathbf{x} \rangle \ = \ \sup_{Z \in K_{off} \cap K_b} \langle B, Z \rangle \}. \end{aligned}$$

Now since $\bar{Z} = \bar{A}$, $b_{ii} \geq 0 \quad \forall\, i$ and

$$\begin{aligned} \langle B, diag\ \mathbf{z} \rangle \ &= \ \sum_{i=1}^{n} b_{ii} z_{ii} \\ &\leq \ \sum_{i=1}^{n} b_{ii} v_i \\ &= \ \langle B, diag\ \mathbf{v} \rangle \end{aligned}$$

where $diag\ \mathbf{z} = Z - \bar{Z}$. Then

$$\sup_{Z \in K_{off} \cap K_b} \langle B, Z \rangle \ \leq \ \langle B, \bar{A} \ + \ diag\ \mathbf{v} \rangle$$

but since $\bar{A} \ + \ diag\ \mathbf{v} \ \in \ K_{off} \cap K_b$ then

$$\sup_{Z \in K_{off} \cap K_b} \langle B, Z \rangle \ = \ \langle B, \bar{A} \ + \ diag \ \mathbf{v} \rangle.$$

Thus

$$\langle B, \bar{A} \ + \ diag \ \mathbf{x} \rangle \ = \ \langle B, \bar{A} \ + \ diag \ \mathbf{v} \rangle \ = \ \langle A, \ B \rangle$$

Now

$$\sup_{Z \in K_{off} \cap K_b} \langle B, Z \rangle \ = \ \begin{Bmatrix} \infty & if \quad b_{ii} \ < \ 0 \quad for \ any \ i \\ \langle A, \ B \rangle & otherwise \end{Bmatrix} \tag{1.3.18}$$

this implies from (1.3.8)

$$\begin{aligned} &\partial(K_{off} \cap K_b) \ (A) \ = \\ &\quad \{B : B \ \in \ \Re^{n \times n}, \ \langle B, \bar{A} \ + \ diag \ \mathbf{x} \rangle \ = \ \langle B, \bar{A} \ + \ diag \ \mathbf{v} \rangle\} \end{aligned} \tag{1.3.19}$$

which implies that

$$\sum_{i=1}^{n} b_{ii}(v_i \ - \ x_i) \ = \ 0. \tag{1.3.20}$$

Therefore if $x_i \ < \ v_i$ then $b_{ii} \ = \ 0$ since each term of (1.3.20) is nonnegative. □

In addition to the normal cone $\partial K_{\Re}$ another set of interest is the normal cone ∂K_M. This set is important when deriving the optimality conditions for the projection method given in Chapter 3. A theorem for the expression of the normal cone ∂K_M is stated and proved. Firstly though a theorem used in the proof is given. An example for the convex cone (1.3.3) when $n \ = \ 3$ is given later on.

The following theorem is based on Hayden and Wells [1988].

Theorem 1.3.6

Let Q be the Householder matrix in (1.2.1). If $A \ = \ A^T \in \Re^{n \times n}$ and M is given in (1.3.2), then

$$\mathbf{x}^T A \mathbf{x} \ \geq \ 0 \qquad \forall \ \mathbf{x} \ \in \ M \tag{1.3.21}$$

if and only if

$$QAQ \ = \ \begin{bmatrix} A_1 & \mathbf{a} \\ \mathbf{a}^T & \alpha \end{bmatrix}, \qquad A_1 \ \geq \ 0. \tag{1.3.22}$$

Proof

For all $\mathbf{x} \in M$ denote $\mathbf{y} = Q\mathbf{x}$, and it follows that $\mathbf{x} = Q\mathbf{y}$ since Q is orthogonal and symmetric. The condition $\mathbf{x} \in M$ is equivalent to $\mathbf{e}^T\mathbf{x} = 0$, or $\mathbf{e}^T Q\mathbf{y} = 0$, and hence to $\mathbf{e}_n^T\mathbf{y} = 0$ where $\mathbf{e}_n^T = [0,\ 0,\ \ldots,\ 0,\ 1]$. Thus (1.3.21) can be written as

$$(Q\mathbf{y})^T AQ\mathbf{y} \geq 0 \qquad \forall\, \mathbf{y} \in \Re^n \ \textit{such that} \ y_n = 0. \tag{1.3.23}$$

Thus (1.3.22) follows. $\square$

In what follows we denote the rank of A_1 by r, and hence the spectral decomposition of A_1 can be expressed as

$$A_1 = U\Lambda U^T = U \begin{bmatrix} \Lambda_r & \mathbf{0} \\ \mathbf{0}^T & 0 \end{bmatrix} U^T \tag{1.3.24}$$

where U is an orthogonal matrix and $\Lambda_r > 0$ is an $r \times r$ diagonal matrix.

A theorem due to Glunt et. al. [1990] will be given to show how to find the normal cone $\partial K_M(A)$ at $A \in K_M$.

Theorem 1.3.7

Given any A, then the normal cone $\partial K_M(A)$ is given by

$$\partial K_M\,(A) = \{B : \ B = Q \begin{bmatrix} UGU^T & \mathbf{0} \\ \mathbf{0}^T & 0 \end{bmatrix} Q, \ \ H \leq 0\} \tag{1.3.25}$$

where

$$G = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0}^T & H \end{bmatrix},$$

U is an orthogonal matrix given by (1.3.24) and H is a symmetric matrix in $\Re^{(n-r-1)\times(n-r-1)}$ (The partitioning of G reflects that of A_1).

Proof

Let $B \in \partial K_M(A)$ and define B_1, $\mathbf{b}$ and β by

$$B \;=\; Q \begin{bmatrix} B_1 & \mathbf{b} \\ \mathbf{b}^T & \beta \end{bmatrix} Q.$$

Now let $X_1 \in \Re^{n-1\times n-1}$ be any positive semi–definite matrix. Then for any $\mathbf{x}$, ξ by Theorem 1.3.6 the matrix

$$X \;=\; Q \begin{bmatrix} X_1 & \mathbf{x} \\ \mathbf{x}^T & \xi \end{bmatrix} Q$$

is in K_M. By (1.3.10)

$$\langle X - A, B \rangle \;\leq\; 0$$

and since Q is orthogonal we have

$$\langle QXQ, QBQ \rangle \;\leq\; \langle A, B \rangle$$

which implies that

$$\langle X_1, B_1 \rangle \;+\; 2\mathbf{x}^T\mathbf{b} \;+\; \xi\beta \;\leq\; \langle A, B \rangle. \tag{1.3.26}$$

Let either $\mathbf{b} \neq \mathbf{0}$ or $\beta \neq 0$. Choose $\mathbf{x} = \lambda\mathbf{b}$ and $\xi = \lambda\beta$ for sufficiently large $\lambda > 0$ then (1.3.26) is false (contradiction). This implies that $\mathbf{b} = \mathbf{0}$ and $\beta = 0$.

Following a similar strategy as in the previous proof (Theorem 1.3.3), let $V\Omega V^T$ be the spectral decomposition of B_1 with V being the orthogonal matrix of eigenvectors and $\Omega = diag\,[\omega_1,\ \omega_2,\ \ldots,\ \omega_{n-1}]$ the diagonal matix of eigenvalues. Since X_1 is positive semi–definite there exists a positive semi–definite matrix $C = V^T X_1 V$, and using (1.3.26)

$$\begin{aligned} \langle A, B \rangle \;\geq\; \langle X_1, B_1 \rangle \quad &= \quad \langle V^T X_1 V, \Omega \rangle \;=\; \langle C, \Omega \rangle \\ &= \quad \sum_{j=1}^{n-1} c_{jj}\omega_j. \end{aligned}$$

Hence

$$\sup_{A \in K_M} \langle A, B \rangle \;\geq\; 0 \qquad iff \;\; \omega_i \;\leq\; 0 \qquad i = 1, \ldots, n-1 \tag{1.3.27}$$

and this is equivalent to $B_1 \leq 0$ since ω_i are the eigenvalues of B_1 and $c_{11},\ c_{22},\ \ldots,\ c_{n-1\ n-1}$ are nonnegative scalars since C is positive semi–definite matrix. Therefore, if $B \in \partial K_M(A)$ it has the form

$$B \;=\; Q \begin{bmatrix} B_1 & 0 \\ 0 & 0 \end{bmatrix} Q, \quad B_1 \;\leq\; 0. \tag{1.3.28}$$

From (1.3.28) we have $\langle X, B\rangle \;\leq\; 0 \;\forall\; X \;\in\; K_M$ and since $\langle X - A, B\rangle \;\leq\; 0$, then

$$\langle A, B\rangle \;\leq\; \sup_{X \in K_M} \langle X, B\rangle \;\;\leq\;\; 0 \tag{1.3.29}$$

$$\leq\;\; \langle A, B\rangle. \tag{1.3.30}$$

from (1.3.27). Thus from (1.3.29) and (1.3.30)

$$\langle A, B\rangle \;=\; 0. \tag{1.3.31}$$

Then (1.3.28), (1.3.31) and (1.3.22) imply that

$$\langle A_1, B_1\rangle \;=\; 0.$$

Then from the spectral decomposition A_1 in (1.3.24) we have

$$\langle \Lambda_r, U^T B_1 U\rangle \;=\; \langle \Lambda_r, G\rangle \;=\; \sum_{j=1}^{r} \lambda_j g_{jj} \;=\; 0 \tag{1.3.32}$$

where $G \;=\; U^T B_1 U \;\leq\; 0.$ Now $G \;\in\; \Re^{(n-1 \times n-1)}$ has the following structure

$$G \;=\; \begin{bmatrix} N & S \\ S^T & H \end{bmatrix}$$

where $H \;\in\; \Re^{(n-r-1)\times(n-r-1)}$ but since $\lambda_j \;>\; 0$ and $g_{jj} \;\leq\; 0$ for $1 \;\leq\; j \;\leq\; r$ then from(1.3.32) $g_{jj} \;=\; 0$ for $1 \;\leq\; j \;\leq\; r$. Since G is negative semi–definite G has the form

$$G \;=\; \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & H \end{bmatrix}, \quad H \leq 0.$$

Therefore

$$B_1 = U \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & H \end{bmatrix} U^T, \quad H \leq 0.$$

and B has the form of (1.3.25).

Conversely, if B is written in the form (1.3.25) then since $B \leq 0$ and $X - A \geq 0 \quad \forall\, X \in K_M$, (since A is the nearest positive semi–definite matrix) then

$$\langle X - A, B \rangle \leq 0 \quad \forall\, X \in K_M$$

which implies that $B \in \partial K_M(A)$ □

In the rest of this section an example of the convex cone (1.3.3) where $n = 3$ is given.

Example 1.3.8

For the example let $n = 3$, and

$$A = \begin{bmatrix} 0 & x & y \\ x & 0 & z \\ y & z & 0 \end{bmatrix}$$

It is convenient for what follows later to express the Householder matrix Q as

$$Q = \begin{bmatrix} a - c & b - d & -c - d \\ b - d & a - c & -c - d \\ -c - d & -c - d & -c - d \end{bmatrix}$$

where $a = 0.911$, $b = 0.244$, $c = 0.122$ and $d = 0.455$ accurate to 3 decimal places. Then

$$QAQ = \begin{bmatrix} bz - \frac{1}{3}x - ay & \frac{1}{3}(2x - y - z) & dz - \frac{1}{3}x - cy \\ \frac{1}{3}(2x - y - z) & by - \frac{1}{3}x - az & dy + cz - \frac{1}{3}x \\ dz - \frac{1}{3}x - cy & dy + cz - \frac{1}{3}x & \frac{2}{3}(x + y + z) \end{bmatrix}$$

and the matrix A_1 is given by

$$A_1 = \begin{bmatrix} bz - ay - \frac{1}{3}x & \frac{1}{3}(2x - y - z) \\ \frac{1}{3}(2x - y - z) & by - \frac{1}{3}x - az \end{bmatrix}.$$

Figure 1.3.3: The positive semi–definite matrix cone K_M in M.

Using (1.3.23) the cone K_M in (1.3.3) is defined by the inequalities

$$bz - ay - \tfrac{1}{3}x \geq 0$$

$$by - \tfrac{1}{3}x - az \geq 0$$

$$(bz - ay - \tfrac{1}{3}x)(by - \tfrac{1}{3}x - az) \geq [\tfrac{1}{3}(2x - y - z)]^2 \tag{1.3.33}$$

where inequality (1.3.33) implies that

$$z^2 - 2z(x + y) + (x - y)^2 \leq 0.$$

The matrix

$$A' \;=\; \begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}$$

is on the boundary of the cone K_M. Then

$$U \;=\; \begin{bmatrix} -0.2588 & 0.966 \\ 0.966 & 0.2588 \end{bmatrix} \;\; and \;\; G \;=\; \begin{bmatrix} 0 & 0 \\ 0 & \lambda \end{bmatrix},$$

so the normal cone (1.3.25) at this point is

$$\partial K_M(A') \;=\; \{\lambda \begin{bmatrix} 0.5 & 0 & -0.5 \\ 0 & 0 & 0 \\ -0.5 & 0 & 0.5 \end{bmatrix}, \;\; \lambda \;\geq\; 0\}.$$

The cone for this example is illustrated in Figure 1.3.3.

1.4 The set of feasible directions

In this section results are given which are used subsequently to derive optimality conditions.

A *feasible point* $\mathbf{x}$ is a point which satisfies all the constraints in an optimization problem and the set of all such points is referred to as the *feasible region.* Here we consider problems in which the the feasible region is a convex set $K \subset \Re^n$. Let $\{\mathbf{x}^{(k)}\} \rightarrow \mathbf{x}$ where $\mathbf{x}^{(k)} \neq \mathbf{x} \;\; \forall \; k$ is an infinite sequence of feasible points. It is possible to express

$$\mathbf{x}^{(k)} \;-\; \mathbf{x} \;=\; \delta^{(k)}\mathbf{s}^{(k)} \quad \forall \;\; k \tag{1.4.1}$$

where $\delta^{(k)} \;>\; 0$ is a scalar. The sequence $\mathbf{x}^{(k)}$ is said to be a directional sequence if $\{\mathbf{s}^{(k)}\} \;\rightarrow\; \mathbf{s}$. The limiting vector $\mathbf{s}^{(k)}$ is referred to as a *feasible direction.* Then the set of feasible directions can be expressed as

$$\mathcal{F}(\mathbf{x}) \;=\; \{\mathbf{s} : \; \exists \; \{\mathbf{x}^{(k)}\} \;\; such \;\; that \;\; \{\mathbf{x}^{(k)}\} \;\rightarrow\; \mathbf{x}, \; \{\mathbf{s}^{(k)}\} \;\rightarrow\; \mathbf{s}, \; \delta^{(k)} \;\rightarrow\; 0\}. \tag{1.4.2}$$

A related set of feasible directions which is easier to manipulate is the set

$$F(\mathbf{x}) \;=\; \{\mathbf{s} : \; \mathbf{s} \;\in\; \Re^n, \;\; \mathbf{s}^T\mathbf{g} \;\le\; 0 \quad \forall \mathbf{g} \;\in\; \partial K(\mathbf{x})\}. \tag{1.4.3}$$

which is the set of feasible directions for the cone of all supporting hyperplanes at $\mathbf{x}$. For future reference it is important to prove that $\mathcal{F}(\mathbf{x}) \subseteq F(\mathbf{x})$.

Let $\mathbf{s} \in \mathcal{F}(\mathbf{x})$ then from (1.4.2) there exists a directional sequence $\mathbf{x}^{(k)} \to \mathbf{x}$ such that $\mathbf{s}^{(k)} \to \mathbf{s}$. Using (1.4.1) and dividing by $\delta^{(k)} > 0$ it follows that

$$\mathbf{s}^{(k)T}\mathbf{g} \;=\; \frac{(\mathbf{x}^{(k)} \;-\; \mathbf{x})^T\mathbf{g}}{\delta^{(k)}}. \qquad \forall\, \mathbf{g} \;\in\; \partial K(\mathbf{x}) \tag{1.4.4}$$

Now any vector $\mathbf{g} \in \partial K(\mathbf{x})$ satisfies $(\mathbf{z} - \mathbf{x})^T\mathbf{g} \le 0 \quad \forall\, \mathbf{z} \in K$. Then since $\mathbf{x}^{(k)}$ are feasible points

$$(\mathbf{x}^{(k)} \;-\; \mathbf{x})^T\mathbf{g} \;\le\; 0.$$

Hence taking limits in (1.4.4) as $k \to \infty$ and $\mathbf{s}^{(k)} \to \mathbf{s}$ implies that

$$\mathbf{s}^T\mathbf{g} \;\le\; 0$$

or $\mathbf{s} \in F(\mathbf{x})$. Therefore this proves that $\mathcal{F}(\mathbf{x}) \subseteq F(\mathbf{x})$ for the general case.

For the positive semi–definite matrix cone (1.3.1) similar definitions to (1.4.2) and (1.4.3) hold. If S is a symmetric matrix which is equivalent to a feasible direction in (1.4.3), Z is a basis matrix for the null space of A and Λ is any symmetric positive semi–definite matrix, then using Theorem 1.3.3, (1.4.3) and the inner product Definition 1.2.1, it follows that

$$\begin{aligned} F(A) \;&=\; \{S : \; S \;=\; S^T, \; \langle B, S\rangle \;\le\; 0 \quad \forall\, B \in \partial K_{\Re}(A)\} \\ &=\; \{S : \; S \;=\; S^T, \;\; \langle -Z\Lambda Z^T, S\rangle \;\le\; 0 \quad \forall\, \Lambda \;\ge\; 0\} \\ &=\; \{S : \; S \;=\; S^T, \;\; \langle \Lambda, Z^TSZ\rangle \;\ge\; 0 \quad \forall\, \Lambda \;\ge\; 0\} \end{aligned}$$

and hence

$$F(A) \;=\; \{S : \; S \;=\; S^T, \;\; Z^TSZ \;\ge 0\}. \tag{1.4.5}$$

The following theorem is due to Fletcher [1985].

Theorem 1.4.1

For $A \in K_{\Re}$

$$\mathcal{F}(A) \equiv F(A) \tag{1.4.6}$$

Proof

In general we proved that $\mathcal{F}(A) \subseteq F(A)$ above. Now the converse is considered.

Take a direction $S \in F$ and let $X = [Y \;\; Z]$ be the eigenvector matrix for A which is described in Theorem 1.3.3 and Λ_r the diagonal matrix of nonzero eigenvalues.

There are two cases, first when $Z^TSZ \geq 0$ and singular, consider the trajectory

$$A_\epsilon = A + \epsilon S + \beta\epsilon^2 I \tag{1.4.7}$$

which gives

$$X^T A_\epsilon X = [Y \;\; Z]^T A + \epsilon S + \beta\epsilon^2 I [Y \;\; Z]$$

$$= \begin{bmatrix} Y^TAY + \epsilon Y^TSY + \beta\epsilon^2 Y^TY & Y^TAZ + \epsilon Y^TSZ + \beta\epsilon^2 Y^TZ \\ Z^TAY + \epsilon Z^TSY + \beta\epsilon^2 Z^TY & Z^TAZ + \epsilon Z^TSZ + \beta\epsilon^2 Z^TZ \end{bmatrix}.$$

Then, since $A = Y\Lambda_r Y^T$ and Z is the basis matrix for the null space of A, it follows that

$$X^T A_\epsilon X = \begin{bmatrix} \Lambda_r + \epsilon Y^TSY + \beta\epsilon^2 & \epsilon Y^TSZ \\ \epsilon Z^TSY & \epsilon Z^TSZ + \beta\epsilon^2 \end{bmatrix}. \tag{1.4.8}$$

Now

$$\Lambda_r + \epsilon Y^TSY + \beta\epsilon^2 > 0 \tag{1.4.9}$$

and

$$\epsilon Z^TSZ + \beta\epsilon^2 - \epsilon^2 Y^TSZ(\Lambda_r + \epsilon Y^TSY + \beta\epsilon^2)^{-1} Z^TSY \geq 0 \tag{1.4.10}$$

are going to be proved. If $\beta > \|\Lambda_r^{-1}\|\|S\|^2$ is chosen and for ϵ sufficiently small, then clearly (1.4.9) and (1.4.10) are true by strength of $\Lambda_r > 0$ for (1.4.9) and $Z^TSZ > 0$ for (1.4.10). Hence there exist Choleski factors for (1.4.9) and (1.4.10) which enable us to construct a Choleski factor for (1.4.8). Therefore $X^TA_\epsilon X$ is positive semi–definite or equivalently A_ϵ is feasible.

For the second case when $Z^TSZ > 0$ consider the trajectory

$$A_\epsilon = A + \epsilon S. \tag{1.4.11}$$

Similar to the first case it gives

$$X^T A_\epsilon X \;=\; \begin{bmatrix} \Lambda_r \;+\; \epsilon Y^T SY & \epsilon Y^T SZ \\ \epsilon Z^T SY & \epsilon Z^T SZ \end{bmatrix}, \tag{1.4.12}$$

hence A_ϵ is feasible since

$$\Lambda_r \;+\; \epsilon Y^T SY \;>\; 0 \tag{1.4.13}$$

and

$$\epsilon Z^T SZ \;-\; \epsilon^2 Y^T SZ(\Lambda_r \;+\; \epsilon Y^T SY)^{-1} Z^T SY \;\geq\; 0. \tag{1.4.14}$$

Thus in both cases a direction $S \in \mathcal{F}(A)$ is constructed and if we take $\epsilon = \epsilon_k$ for any sequence $\epsilon_k \to 0$ then there exists a feasible directional sequence in $\mathcal{F}$. Therefore $F \subset \mathcal{F}$ proving that these sets are in fact equivalent. $\Box$

From this theorem we can deal with F which is easier to operate than $\mathcal{F}$. In this section expression (1.4.5) provides a characterization of a feasible direction of search. The benefit of this expression along with the normal cone expression (1.3.12) lies in their application to optimization problems. The expression for the normal cone plays the part of the subdifferential in the statement of optimality conditions. The expression for the feasible direction accommodates a characterization of a feasible direction of search which is easily verified.

1.5 First and second order conditions

The content of this section is useful in deriving the methods in Sections 5.3 and 6.4.

This section includes a useful theorem of first order conditions. Also at the end of this section second order conditions are stated. It is also shown how to compute a basis matrix Z for the null space of A in connection with the partial LDL^T factorization of A.

Consider the following problem

$$\begin{aligned} &minimize \quad f(A) \\ &subject\;\; to \;\; A \in K_{\Re}, \quad c_i(A) \;\leq\; 0. \quad i = 1,\ldots,m \end{aligned} \tag{1.5.1}$$

The problem of minimizing a convex function $f(A)$ on a general convex set K is said to be a convex programming problem. A special case of (1.5.1) occurs when $K = K_{\Re} \cup K_c$ where

$$K_c = \{A : A \in \Re^{n\times n} \; c_i(A) \le 0. \quad i = 1, \ldots, m\}.$$

is a convex set (this is assured if the functions $c_i(A)$ are convex).

A *local solution* is a point at which, in a neighbourhood about that point, has no feasible point that gives a smaller value of the objective function.

Theorem 1.5.1

Every local solution $\mathbf{x}^*$ to a convex programming problem is a global solution.

Proof

Let A^* be a local but not global solution. Then $\exists\, A \in K$ such that $f(A) < f(A^*)$. By convexity of K

$$A_\lambda = (1 - \lambda)A^* + \lambda A.$$

By convexity of f

$$\begin{aligned} f(A_\lambda) &\le (1 - \lambda)f(A^*) + \lambda f(A) \\ &= f(A^*) + \lambda(f(A) - f(A^*)) \\ &< f(A^*). \end{aligned} \tag{1.5.2}$$

Taking $\lambda \to 0$ in the limit there exists $f(A_\lambda)$ in the neighbourhood of $f(A^*)$ which contradictes the local solution property. Thus local solutions are global. $\Box$

In a convex programming problem every local solution is a global solution which has been proved above. If $f(A)$ and $c_i(A)$ $i = 1, \ldots, m$ are convex and nonsmooth then the first order necessary conditions can be given in the following theorem

Theorem 1.5.2 (*First order conditions*)

If A^* solves (1.5.1) and if the condition in Theorem 1.4.1 holds then A^* is feasible and there exist Lagrange multipliers $\Lambda^* \geq 0$ and $\boldsymbol{\pi}^* \geq 0$ satisfying the following

$$\sum_{i=1}^{m} \pi_i^* c_i^* = 0$$

and

$$\nabla_A \mathcal{L}(A^*, \Lambda^*, \boldsymbol{\pi}^*) = G^* + B^* + \sum_{i=1}^{m} \pi_i^* C_i^* = 0 \tag{1.5.3}$$

where $G^* \in \nabla f^*$, $B^* \in \partial K_{\Re}^*$ and $C_i^* = \nabla c_i^*$ $i = 1, \ldots, m$. (Note that the operator ∇ maps a scalar into a matrix).

Proof (see for example Rockafellar [1981] Chapter 5)

This theorem is related to the usual Kuhn–Tucker (KT) conditions (e.g. see Fletcher [1987]) and the π_i^* are KT multipliers. However an additional term derived from $\partial K_{\Re}^*$ also occurs.

The conditions in Theorem 1.5.2 are certainly sufficient when all feasible directions are strict ascent directions. However consider situation in which there exist feasible directions along which $f(A)$ has a zero directional derivative. Now higher order terms become significant. Second order information is required in order to provide algorithms that converge rapidly. Also, it is difficult to deal with the matrix cone constraint in (1.5.1), since it is not in the form of a functional constraint. An equivalent problem to (1.5.1) with equality and inequality constraints which are easier to manipulate is considered here. This formulation will enable us to derive algorithms with a second order rate of convergence.

Assume that r, the rank of A^*, $(1 < r < n)$ is known. Permuting rows and columns if necessary, then for A sufficiently close to A^* (which ensure that $D_1 > 0$) the partial factors

$$A = LDL^T \tag{1.5.4}$$

can be calculated, where

$$L = \begin{bmatrix} L_{11} & \\ L_{21} & I \end{bmatrix}, \quad D = \begin{bmatrix} D_1 & \\ & D_2 \end{bmatrix} \tag{1.5.5}$$

$L_{11} \in \Re^{r\times r}$ is unit lower triangular, $D_1 \in \Re^{r\times r}$ is diagonal and positive definite and $D_2 \in \Re^{n-r\times n-r}$. $D_2 = \mathbf{0}$ at the solution, in general we can calculate D_2 as follows, partitioning

$$A = \begin{bmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix} \tag{1.5.6}$$

where $A_{11} \in \Re^{r\times r}$, from (1.5.5)

$$LDL^T = \begin{bmatrix} L_{11}D_1L_{11}^T & L_{11}D_1L_{21}^T \\ L_{21}D_1L_{11}^T & L_{21}D_1L_{21}^T + D_2 \end{bmatrix} \tag{1.5.7}$$

then

$$A_{22} = L_{21}D_1L_{21}^T + D_2 \tag{1.5.8}$$

and since

$$\begin{aligned} L_{21}D_1L_{21}^T &= (L_{21}D_1L_{11}^T)(L_{11}^{-T}D_1^{-1}L_{11}^{-1})(L_{11}D_1L_{21}^T) \\ &= A_{21}A_{11}^{-1}A_{21}^T, \end{aligned}$$

therefore

$$D_2(A) = A_{22} - A_{21}A_{11}^{-1}A_{21}^T. \tag{1.5.9}$$

Thus A is positive semi–definite if and only if $D_2 = \mathbf{0}$, thus the constraint $A \in K_{\Re}$ can be expressed as

$$D_2(A) = \mathbf{0} \tag{1.5.10}$$

This gives a ready expression which can be used to compute both first and second derivatives of the constraints with respect to the elements of A.

The orthonormal basis matrix Z for the null space of A^* can be calculated using (1.5.5). Define

$$V = L^{-T} = \begin{bmatrix} L_{11}^{-T} & -L_{11}^{-T}L_{21}^T \\ \mathbf{0} & I \end{bmatrix}$$

then using (1.5.7)

$$V = \begin{bmatrix} L_{11}^{-T} & -A_{11}^{-1}A_{21}^T \\ \mathbf{0} & I \end{bmatrix}$$

$$= \begin{bmatrix} V_{11} & V_{21} \\ \mathbf{0} & I \end{bmatrix}. \tag{1.5.11}$$

Then

$$Z = \begin{bmatrix} V_{21} \\ I \end{bmatrix}.$$

From (1.5.9)

$$\begin{aligned} D_2(A) &= A_{22} - A_{21}A_{11}^{-1}A_{21}^T \\ &= [\,-A_{11}^{-1}A_{21}^T \quad I\,] \begin{bmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} -A_{11}^{-1}A_{21}^T \\ I \end{bmatrix} \\ &= Z^TAZ = \mathbf{0}. \end{aligned} \tag{1.5.12}$$

Then problem (1.5.1) can be expressed in the equivalent form

$$\begin{aligned} &\textit{minimize} \quad f(A) \\ &\textit{subject to} \;\; Z^TAZ = \mathbf{0} \quad c_i(A) \le 0. \quad i = 1, \ldots, m. \end{aligned} \tag{1.5.13}$$

It is convenient to introduce the Lagrangian function

$$\mathcal{L}(A, \Lambda, \boldsymbol{\pi}) = f(A) - \langle \Lambda, Z^TAZ \rangle + \sum_{i=1}^{m} \pi_i c_i(A) \tag{1.5.14}$$

in which Λ and $\boldsymbol{\pi}$ are Lagrange multipliers for the constraints (1.5.12) and $\mathbf{c}(A) \leq \mathbf{0}$ respectively. (Fletcher [1987] (Theorem 9.1.1)). Since $\langle \Lambda, Z^T A Z \rangle = \langle A, Z^T \Lambda Z \rangle$, then A^*, Λ^* and $\boldsymbol{\pi}^*$ satisfy

$$\nabla_A \mathcal{L}(A^*, \Lambda^*, \boldsymbol{\pi}^*) = \nabla_A f^* - Z^T \Lambda^* Z + \sum_{i=1}^{m} \pi_i^* \nabla_A c_i^* = \mathbf{0} \tag{1.5.15}$$

This equation corresponds to (1.5.3). (Note that Λ and $\boldsymbol{\pi}$ not necessarily the same as the Λ and $\boldsymbol{\pi}$ in Theorem 1.5.2).

The matrix Λ that appears in the normal cone expression (1.3.12) can be defined as the Lagrange multiplier matrix for the constraints $D_2(A) = \mathbf{0}$ relative to the basis Z.

This treatment of second order conditions was given by Fletcher [1985].

1.6 Quasi–Newton methods

In this section the problem of finding a local solution to the problem

$$\underset{\mathbf{x}}{\text{minimize}} \; f(\mathbf{x}), \qquad \mathbf{x} \in \Re^n \tag{1.6.1}$$

is considered. The function f is smooth and not necessary convex. In the previous section we dealt with optimization problems which have various types of constraint. However in this section the optimum value is sought of an objective function of many variables without any constraint. This type of problem will arise in Chapter 3.

The present section will be devoted to the study of quasi–Newton methods. First, the Newton method will be discussed in order to show how the quasi–Newton method is derived from it.

The idea behind Newton's method is to replace the function $f(\mathbf{x})$ in the equation to be minimized $(f(\mathbf{x}) = 0)$ by a quadratic model that approximates the function. The quadratic model is obtained from the first three terms of a Taylor series expansion of $f(\mathbf{x})$ about $\mathbf{x}^{(k)}$ as follows

$$f(\mathbf{x}^{(k)} + \boldsymbol{\delta}) \approx f^{(k)} + \mathbf{g}^{(k)T}\boldsymbol{\delta} + \frac{1}{2}\ \boldsymbol{\delta}^T G^{(k)}\boldsymbol{\delta} = q^{(k)}(\boldsymbol{\delta}) \tag{1.6.2}$$

where $\boldsymbol{\delta} = \mathbf{x} - \mathbf{x}^{(k)}$, and $q^{(k)}(\boldsymbol{\delta})$ is the resulting quadratic approximation for iteration k. With the requirement that the first and the second derivatives of $f(\mathbf{x})$ are known at any point, then the coefficients $f^{(k)}$, $\mathbf{g}^{(k)}$ and $G^{(k)}$ are also known. The kth iteration of Newton's method can be stated as follows:

Algorithm 1.6.1 (*Newton method*)

Let $G^{(k)}$ be an $n \times n$ positive definite matrix then the following algorithm computes the local minimum $\mathbf{x}^*$ for $f(\mathbf{x})$

i. Select initial point $\mathbf{x}^{(0)} \in \Re^n$.

ii. Solve $G^{(k)}\boldsymbol{\delta} = -\mathbf{g}^{(k)}$ for $\boldsymbol{\delta} = \boldsymbol{\delta}^{(k)}$

iii. Set $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \boldsymbol{\delta}^{(k)}$.

iv. If $\mathbf{g}^{(k)} \approx \mathbf{0}$

stop

else

go to (ii).

The following theorem by Fletcher [1987] gives the rate of convergence for Newton method.

Theorem 1.6.2

If f is twice continuously differentiable, $\mathbf{x}^{(k)}$ is sufficiently close to $\mathbf{x}^*$ for some k, G^* is nonsingular and $G(\mathbf{x})$ satisfies a *Lipschitz condition* $\|G(\mathbf{x})-G(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|$ in a neighbourhood of a local minimizer $\mathbf{x}^*$, then $\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}^*$ and Newton's algorithm converges at second order.

Proof

Since f is differentiable, a Taylor series for $\mathbf{g}\,(\mathbf{x}^{(k)} + \mathbf{h}\,)$ about $\mathbf{x}^{(k)}$ exists and can be written as

$$\mathbf{g}(\mathbf{x}^{(k)} + \mathbf{h}) = \mathbf{g}^{(k)} + G^{(k)}\mathbf{h} + O(\|\mathbf{h}\|^2). \tag{1.6.3}$$

Denote $\mathbf{h}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}^*$. Letting $\mathbf{h} = -\mathbf{h}^{(k)}$ gives

$$\mathbf{0} = \mathbf{g}^* = \mathbf{g}(\mathbf{x}^{(k)} - \mathbf{h}^{(k)}) = \mathbf{g}^{(k)} - G^{(k)}\mathbf{h}^{(k)} + O(\|\mathbf{h}^{(k)}\|^2). \tag{1.6.4}$$

Multiplying equation (1.6.4) by $G^{(k)-1}$ gives

$$\mathbf{0} = -\boldsymbol{\delta}^{(k)} - \mathbf{h}^{(k)} + O(\|\mathbf{h}^{(k)}\|^2) = -\mathbf{h}^{(k+1)} + O(\|\mathbf{h}^{(k)}\|^2) \tag{1.6.5}$$

Hence, by definition of $O(\mathbf{h})$ (see Definition 1.2.6), there exists a constant $c > 0$ such that

$$\|\mathbf{h}^{(k+1)}\| \leq c\,\|\mathbf{h}^{(k)}\|^2. \tag{1.6.6}$$

Let $\mathbf{x}^{(k)}$ be in a neighbourhood of $\mathbf{x}^*$ with $\|\mathbf{h}\| \leq \alpha/c$, where $0 < \alpha < 1$, then this implies that $\|\mathbf{h}^{(k+1)}\| \leq \alpha\|\mathbf{h}^{(k)}\|$. Thus $\mathbf{x}^{(k)} \to \mathbf{x}^*$ since $\|\mathbf{h}^{(k)}\| \to 0$ and relation (1.6.6) shows that Newton's algorithm converges at second order. □

The basic Newton method as it stands is not suitable for general purposes because $G^{(k)}$ may not be positive definite when $\mathbf{x}^{(k)}$ is far away from the solution $\mathbf{x}^*$. Sometimes even if $G^{(k)}$ is positive definite Newton's method may not converge, and $\{f^{(k)}\}$ may not even decrease.

Now, the concept of the *line search* is introduced. The line search algorithms have the following structure:

given an initial estimate $\mathbf{x}^{(0)}$ the basic structure of the kth iteration is

i. determine a direction of search $\mathbf{s}^{(k)}$

ii. find $\alpha^{(k)}$ to minimize $f(\mathbf{x}^{(k)} + \alpha^{(k)}\mathbf{s}^{(k)})$ with respect to $\alpha^{(k)}$

iii. set $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha^{(k)}\mathbf{s}^{(k)}$.

There are different methods which correspond to different ways of choosing $\mathbf{s}^{(k)}$. The line search subproblem in step (ii) is carried out by repeatedly sampling $f(\mathbf{x})$ and possibly its derivatives for different points $\mathbf{x} = \mathbf{x}^{(k)} + \alpha\mathbf{s}^{(k)}$ along the line. In practice step (ii) is solved approximately and the aim of the line search is to find a step $\alpha^{(k)}$ which gives a significant reduction in f on each iteration. (see Fletcher [1987] Sections 2.5 and 2.6 for more about the line search).

However the main difficulty in Newton's method arises from supplying the second derivative matrix G. Methods similar to Newton's method, and not requiring the second derivative, can be derived. Quasi–Newton methods are descent methods which approximate G^{-1} by a symmetric positive definite matrix $H^{(k)}$. The popularity of the most successful of these methods stems from the fact that they exhibit a fast rate of convergence while avoiding the second derivative calculations associated with Newton's method.

The quasi–Newton algorithm takes the following form.

Algorithm 1.6.3 (*quasi–Newton method*)

i. Select initial point $\mathbf{x}^{(0)} \in \Re^n$.

ii. Set $\mathbf{s}^{(k)} = -H^{(k)}\mathbf{g}^{(k)}$

iii. Line search along $\mathbf{s}^{(k)}$ giving $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha^{(k)}\mathbf{s}^{(k)}$

iv. Update $H^{(k)}$ giving $H^{(k+1)}$.

v. If $\mathbf{g}^{(k)} \approx \mathbf{0}$

stop

else

go to (ii).

The initial matrix $H^{(0)}$ is an arbitrary positive definite matrix, $H^{(0)} = I$ is the first choice if there is no better estimate. There are various possible formulas for updating the positive definite matrix H. An important formula was suggested independently by Broyden [1970], Fletcher [1970], Goldfarb [1970] and Shanno [1970], and is known as the BFGS formula

$$\begin{aligned} H^{(k+1)}_{BFGS} = H^{(k)} &+ \{1 + \frac{\boldsymbol{\gamma}^{(k)T}H^{(k)}\boldsymbol{\gamma}^{(k)}}{\boldsymbol{\delta}^{(k)T}\boldsymbol{\gamma}^{(k)}}\} \frac{\boldsymbol{\delta}^{(k)}\boldsymbol{\delta}^{(k)T}}{\boldsymbol{\delta}^{(k)T}\boldsymbol{\gamma}^{(k)}} \\ &- \frac{\boldsymbol{\delta}^{(k)}\boldsymbol{\gamma}^{(k)T}H^{(k)} + H^{(k)}\boldsymbol{\gamma}^{(k)}\boldsymbol{\delta}^{(k)T}}{\boldsymbol{\delta}^{(k)T}\boldsymbol{\gamma}^{(k)}}. \end{aligned} \tag{1.6.7}$$

where

$$\boldsymbol{\gamma}^{(k)} = \mathbf{g}^{(k+1)} - \mathbf{g}^{(k)}.$$

and

$$\boldsymbol{\delta}^{(k)} = \alpha^{(k)}\mathbf{s}^{(k)} = \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}.$$

There is growing evidence that the BFGS formula is the best general purpose quasi–Newton method currently available and it is an efficient technique for unconstrained optimization. Therefore, this formula will be used in this thesis. For more discussion about Newton and quasi–Newton methods with references see Fletcher [1987].

1.7 The l_1 SQP method

This section is devoted to constrained optimization in which additional constraints arise while in the previous section we had only objective functions. The methods in this section deal with constraints which are easier to handle than the constraints in previous sections. The constraints here are expressed in terms of equations and inequalities instead of sets and cones. Methods arising in this section are useful in deriving related methods in Sections 5.3 and 6.4.

This section will be devoted to the study of l_1 SQP method. First it will be shown how the l_1 SQP method is derived from the SQP method. The SQP method is also called the *Lagrange–Newton method*.

Consider the following equality constraint problem

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} \quad f(\mathbf{x}) \\ &\textit{subject to} \;\; \mathbf{c}(\mathbf{x}) = \mathbf{0}. \end{aligned} \tag{1.7.1}$$

The idea behind the SQP method is to iterate on the basis of certain approximations to the problem function $f(\mathbf{x})$ and $\mathbf{c}(\mathbf{x})$ using a linear approximation to the constraint function $\mathbf{c}(\mathbf{x})$. This method is Newton's method applied to find the stationary point of the Lagrangian function

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) \;=\; f(\mathbf{x}) \;-\; \sum_i \lambda_i c_i(\mathbf{x}) \tag{1.7.2}$$

The variables in the Lagrangian function are $\mathbf{x}$ and $\boldsymbol{\lambda}$. The method generates a sequence of approximations $\mathbf{x}^{(k)}$ and $\boldsymbol{\lambda}^{(k)}$ to the solution vector $\mathbf{x}^*$ and the Lagrange multipliers $\boldsymbol{\lambda}^*$.

A Taylor series for $\nabla_{\mathbf{x},\boldsymbol{\lambda}}\, \mathcal{L}$ about $\mathbf{x}^{(k)}$, $\boldsymbol{\lambda}^{(k)}$ gives

$$\begin{aligned}
\nabla_{\mathbf{x},\boldsymbol{\lambda}}\, \mathcal{L}(\mathbf{x}^{(k)} \;+\; \boldsymbol{\delta}\mathbf{x},\; \boldsymbol{\lambda}^{(k)} \;+\; \boldsymbol{\delta\lambda}) \;&=\; \nabla_{\mathbf{x},\boldsymbol{\lambda}}\, \mathcal{L}(\mathbf{x}^{(k)},\; \boldsymbol{\lambda}^{(k)}) \\
&+\; [\nabla^2_{\mathbf{x},\boldsymbol{\lambda}}\, \mathcal{L}(\mathbf{x}^{(k)},\; \boldsymbol{\lambda}^{(k)})] \begin{bmatrix} \boldsymbol{\delta}\mathbf{x} \\ \boldsymbol{\delta\lambda} \end{bmatrix} \;+\; \ldots
\end{aligned}$$

where $\boldsymbol{\delta\lambda} \;=\; \boldsymbol{\lambda}^* \;-\; \boldsymbol{\lambda}^{(k)}$, $\boldsymbol{\delta}\mathbf{x} \;=\; \mathbf{x}^* \;-\; x^{(k)}$. Neglecting higher order terms

$$\nabla^2_{\mathbf{x},\boldsymbol{\lambda}}\, \mathcal{L}(\mathbf{x}^{(k)},\; \boldsymbol{\lambda}^{(k)}) \begin{bmatrix} \boldsymbol{\delta}\mathbf{x} \\ \boldsymbol{\delta\lambda} \end{bmatrix} \;=\; -\,\nabla_{\mathbf{x},\boldsymbol{\lambda}}\, \mathcal{L}(\mathbf{x}^{(k)},\; \boldsymbol{\lambda}^{(k)}). \tag{1.7.3}$$

Since $\nabla_{\mathbf{x},\boldsymbol{\lambda}}\, \mathcal{L}(\mathbf{x}^*,\; \boldsymbol{\lambda}^*) \;=\; \nabla_{\mathbf{x},\boldsymbol{\lambda}}\, \mathcal{L}(\mathbf{x}^{(k)} \;+\; \boldsymbol{\delta}\mathbf{x},\; \boldsymbol{\lambda}^{(k)} \;+\; \boldsymbol{\delta\lambda}) \;=\; \mathbf{0}$. Using (1.7.2) to find $\nabla_{\mathbf{x},\boldsymbol{\lambda}}\, \mathcal{L}$ and $\nabla^2_{\mathbf{x},\boldsymbol{\lambda}}\, \mathcal{L}$, gives the system

$$\begin{bmatrix} W^{(k)} & -A^{(k)} \\ -A^{(k)T} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}\mathbf{x} \\ \boldsymbol{\delta\lambda} \end{bmatrix} \;=\; \begin{bmatrix} -\mathbf{g}^{(k)} \;+\; A^{(k)}\boldsymbol{\lambda}^{(k)} \\ \mathbf{c}^{(k)} \end{bmatrix} \tag{1.7.4}$$

where $\mathbf{g} \;=\; \nabla_{\mathbf{x}}\, f$, A is the *Jacobian matrix* of constraint $\mathbf{c}(\mathbf{x}^{(k)})$, and

$$W^{(k)} \;=\; \nabla^2_{\mathbf{x}} f(\mathbf{x}^{(k)}) \;-\; \sum_i \lambda_i^{(k)} \nabla^2_{\mathbf{x}} c_i(\mathbf{x}^{(k)}) \tag{1.7.5}$$

is the Hessian matrix $\nabla^2_{\mathbf{x}}\, \mathcal{L}(\mathbf{x}^{(k)}, \boldsymbol{\lambda}^{(k)})$. The system (1.7.4) is solved to give corrections $\boldsymbol{\delta}\mathbf{x}$ and $\boldsymbol{\delta\lambda}$.

An equivalent system to (1.7.4) is

$$\begin{bmatrix} W^{(k)} & -A^{(k)} \\ -A^{(k)T} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}^{(k)} \\ \boldsymbol{\lambda}^{(k+1)} \end{bmatrix} \;=\; \begin{bmatrix} -\mathbf{g}^{(k)} \\ \mathbf{c}^{(k)} \end{bmatrix} \tag{1.7.6}$$

where $\boldsymbol{\lambda}^{(k+1)} = \boldsymbol{\lambda}^{(k)} + \boldsymbol{\delta\lambda}$ and $\boldsymbol{\delta}^{(k)} = \boldsymbol{\delta}\mathbf{x}$. System (1.7.6)is used to determine $\boldsymbol{\delta}^{(k)}$ and $\boldsymbol{\lambda}^{(k+1)}$, then $\mathbf{x}^{(k+1)}$ is given by

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \boldsymbol{\delta}^{(k)}. \tag{1.7.7}$$

This method requires initial approximations $\mathbf{x}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$, and uses (1.7.6) and (1.7.7) to generate the iterative sequence $\{\mathbf{x}^{(k)}, \boldsymbol{\lambda}^{(k)}\}$.

Similar to Newton's method in the previous section it is possible to restate this method in terms of one in which the subproblem involves the minimization of a quadratic function. Consider the subproblem

$$\begin{aligned} &\underset{\boldsymbol{\delta}}{\text{minimize}} \quad q^{(k)}(\boldsymbol{\delta}) = \tfrac{1}{2}\, \boldsymbol{\delta}^T W^{(k)} \boldsymbol{\delta} + \mathbf{g}^{(k)T} \boldsymbol{\delta} + f^{(k)} \\ &\textit{subject to } \mathbf{l}^{(k)}(\boldsymbol{\delta}) = A^{(k)T} \boldsymbol{\delta} + \mathbf{c}^{(k)} = \mathbf{0} \end{aligned} \tag{1.7.8}$$

This problem is the *quadratic programming subproblem* (QPS). Equations (1.7.6) gives the first order conditions for problem (1.7.8). If the reduced matrix $Z^{(k)T} W^{(k)} Z^{(k)}$ is positive definite then $\boldsymbol{\delta}^{(k)}$ minimizes (1.7.8), where $Z^{(k)}$ is the null matrix for $A^{(k)}$. Hence the following algorithm is suggested.

Algorithm 1.7.1

Given initial estimate $\mathbf{x}^{(0)}, \boldsymbol{\lambda}^{(0)}$

$For \quad k = 1, 2, \ldots$

i. Solve (1.7.8) to determine $\boldsymbol{\delta}^{(k)}$ and $\boldsymbol{\lambda}^{(k+1)}$ the vector of Lagrange multipliers of the linear constraints.

ii. Set $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \boldsymbol{\delta}^{(k)}$.

This algorithm is known as the SQP algorithm.

Algorithm 1.7.1 suggests a generalization for solving the nonlinear inequality constraint problem

$$\underset{\mathbf{x}}{\text{minimize}} \quad f(\mathbf{x})$$

$$subject\ to\ \ \mathbf{c}(\mathbf{x}) \ \geq \ \mathbf{0}. \tag{1.7.9}$$

Replacing $\mathbf{c}(\mathbf{x})$ by $\mathbf{l}^{(k)}(\boldsymbol{\delta})$ and $f(\mathbf{x})$ by $q^{(k)}(\boldsymbol{\delta})$ leads to the subproblem

$$\begin{aligned} &\underset{\boldsymbol{\delta}}{\text{minimize}} \quad q^{(k)}(\boldsymbol{\delta}) \ = \ \tfrac{1}{2}\, \boldsymbol{\delta}^T W^{(k)} \boldsymbol{\delta} \ + \ \mathbf{g}^{(k)T} \boldsymbol{\delta} \ + \ f^{(k)} \\ &subject\ \ to\ \ \mathbf{l}^{(k)}(\boldsymbol{\delta}) \ = \ A^{(k)T} \boldsymbol{\delta} \ + \ \mathbf{c}^{(k)} \ \geq \ \mathbf{0}. \end{aligned} \tag{1.7.10}$$

This QPS can be used in an iterative scheme like Algorithm 1.7.1 in a similar way using (1.7.10) instead of (1.7.8) in step i.

The second order convergence of iteration (1.7.6) and (1.7.7) follows by using the technique of Theorem 1.6.2 applied to the system of $n + m$ equations $\nabla_{\mathbf{x},\boldsymbol{\lambda}}\, \mathcal{L}(\mathbf{x},\ \boldsymbol{\lambda}) \ = \ \mathbf{0}$. The convergence of this method at a rapid rate can be proved when $\mathbf{x}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$ are sufficiently close to $\mathbf{x}^*$ and $\boldsymbol{\lambda}^*$ for some k. In fact a stronger result is given by Fletcher [1987] in the following theorem.

Theorem 1.7.2

If $\mathbf{x}^{(0)}$ is sufficiently close to $\mathbf{x}^*$, the Lagrangian matrix

$$\begin{bmatrix} W^{(k)} & -A^{(k)} \\ -A^{(k)T} & \mathbf{0} \end{bmatrix}$$

is non–singular, and if second order sufficient conditions hold at $\mathbf{x}^*, \boldsymbol{\lambda}^*$ with A^* having full rank, then the QPS iteration (1.7.6) and (1.7.7) converges at second order. If $\boldsymbol{\lambda}^{(k)}$ is sufficiently close to $\boldsymbol{\lambda}^*$, $\boldsymbol{\lambda}^{(0)}$ is suitably chosen and if (1.7.8) is solved uniquely by $\boldsymbol{\delta}^{(0)}$ then the SQP method converges at second order.

Proof (see Fletcher [1987] Chapter 12)

Globally the SQP method may not converge, especially when $\mathbf{x}^{(0)}$ is remote from $\mathbf{x}^*$. However the SQP method is usually modified by the l_1 exact penalty function. The l_1 exact penalty function associated with (1.7.9) is

$$\phi(\mathbf{x}) \;=\; \mu f(\mathbf{x}) \;+\; \sum_{i\in E}|c_i(\mathbf{x})| \;+\; \sum_{i\in I}\max(-c_i(\mathbf{x}),0) \qquad (1.7.11)$$

where E is the set of equality constraints and I the set of inequality constraints. For sufficiently small μ local solutions of the nonlinear programming problem (1.7.9) are equivalent to local solutions of (1.7.11) under wide asssumptions.

Various algorithms based on the use of (1.7.11) have been tried. The function (1.7.11) is not differentiable so it cannot be minimized by conventional methods. The most simple is Han's [1977] method which uses the solution of SQP subproblem (1.7.10) as a search direction, and the next point is accepted only if it significantly reduces the value of $\phi(\mathbf{x})$. An algorithm with better convergence properties is suggested by Fletcher [1981a] in which a different subproblem to (1.7.10) is solved, which takes into account the structure of (1.7.11), but uses the same approximating functions as in (1.7.10). The l_1 SQP method is a direct and efficient approach to nonlinear programming. Fletcher [1981a] shows how to use a step restriction (or trust region) so that the difficulties mentioned above are removed. The method can be explained easily as follows: instead of substituting the Taylor series approximations (1.7.10) into the nonlinear programming problems they are substituted directly into the l_1 exact penalty function (1.7.11), giving a piecewise quadratic approximating function $\psi^{(k)}(\boldsymbol{\delta})$ and hence a QP subproblem

$$\begin{aligned} &\underset{\boldsymbol{\delta}}{\text{minimize}} \quad \psi^{(k)}(\boldsymbol{\delta}) \\ &\textit{subject to} \;\; \|\boldsymbol{\delta}\| \;\le\; \rho^{(k)} \end{aligned} \qquad (1.7.12)$$

where

$$\psi^{(k)}(\boldsymbol{\delta}) \;=\; q^{(k)}(\boldsymbol{\delta}) \;+\; \sum_{i\in E}|l_i^{(k)}(\boldsymbol{\delta})| \;+\; \sum_{i\in I}\max(-l_i^{(k)}(\boldsymbol{\delta}),0). \qquad (1.7.13)$$

The subproblem (1.7.12) is solved on each iteration which is of a similar to the QP subproblem (1.7.10). Subproblem (1.7.12) differs from (1.7.10) in that there are no explicit constraints derived from the linear approximations $\mathbf{l}^{(k)}(\boldsymbol{\delta})$. Thus there are no difficulties with an infeasible subproblem. The use of a trust region guarantees boundedness of the subproblem. The norm in (1.7.12) is arbitrary but either the $\|\cdot\|_\infty$ or the $\|\cdot\|_2$ is most likely choice since the subproblem (1.7.12) can then be solved by QP methods.

The radius $\rho^{(k)}$ is the step restriction which is adjusted adaptively in a customary way to be as large as possible subject to reasonable agreement between $\phi(\mathbf{x}^{(k)}+\boldsymbol{\delta})$ and $\psi^{(k)}(\boldsymbol{\delta})$, thus ensuring a significant decrease in the function $\phi(\mathbf{x})$. The *ratio* measures the extent to which ϕ and $\psi^{(k)}$ agree in neighbourhood of $\mathbf{x}^{(k)}$ is defined by

$$r^{(k)} = \frac{\nabla\phi(\mathbf{x}^{(k)})}{\nabla\psi^{(k)}(\boldsymbol{\delta}^{(k)})} \tag{1.7.14}$$

where

$$\nabla\phi(\mathbf{x}^{(k)}) = \phi(\mathbf{x}^{(k)}) - \phi(\mathbf{x}^{(k)} + \boldsymbol{\delta}^{(k)}) \tag{1.7.15}$$

is the *actual reduction* and

$$\nabla\psi^{(k)} = \phi(\mathbf{x}^{(k)}) - \psi^{(k)}(\boldsymbol{\delta}^{(k)}). \tag{1.7.16}$$

is the *predicted reduction.*

These features can be observed in the following algorithm from problem (1.7.9) given by Fletcher [1981a].

Algorithm 1.7.3

This algorithm solves problem (1.7.9)

i. Given $\mathbf{x}^{(k)}, \boldsymbol{\lambda}^{(k)}$ and $\rho^{(k)}$, calculate $f^{(k)}$, $\mathbf{g}^{(k)}$, $\mathbf{c}^{(k)}$, $A^{(k)}$ and $W^{(k)}$ which determine $\phi(\mathbf{x}^{(k)})$ and $\psi^{(k)}(\boldsymbol{\delta}^{(k)})$.

ii. Find a global solution $\boldsymbol{\delta}^{(k)}$ to (1.7.12).

iii. Evaluate $\phi(\mathbf{x}^{(k)} + \boldsymbol{\delta}^{(k)})$ and calculate $\nabla\phi(\mathbf{x}^{(k)}), \nabla\psi^{(k)}$ and $r^{(k)}$.

iv.

$$\begin{aligned} &If \quad r^{(k)} < 0.25 \quad set \quad \rho^{(k+1)} = \|\boldsymbol{\delta}^{(k)}\|/4 \\ &if \ \ r^{(k)} > 0.75 \ \ and \ \ \|\boldsymbol{\delta}^{(k)}\| = \rho^{(k)} \ \ set \ \ \rho^{(k+1)} = 2\rho^{(k)} \\ &otherwise \ \ set \quad \rho^{(k+1)} = \rho^{(k)}. \end{aligned}$$

v.

$$\begin{aligned} If \quad r^{(k)} \le 0 \qquad & set \quad \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)}, \ \ \boldsymbol{\lambda}^{(k+1)} = \boldsymbol{\lambda}^{(k)} \\ & else \quad \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \boldsymbol{\delta}^{(k)} \\ & \qquad \boldsymbol{\lambda}^{(k+1)} = multipliers \ \ from \ \ (1.7.12). \end{aligned}$$

The iteration based on (1.7.12) is guaranteed to converge to a Kuhn–Tucker point of (1.7.11) (Fletcher [1987]). Therefore, this algorithm will be used in this thesis (Chapters 5 and 6). For more about SQP and l_1 SQP methods see Fletcher [1987] Sections 12.4 and 14.5.