THESES, SIS/LIBRARY
R.G. MENZIES BUILDING NO.2
Australian National University
Canberra ACT 0200 Australia

Telephone: +61 2 6125 4631
Facsimile: +61 2 6125 4063
Email: library.theses@anu.edu.au

USE OF THESES

**This copy is supplied for purposes
of private study and research only.
Passages from the thesis may not be
copied or closely paraphrased without the
written consent of the author.**

E R R A T A

'Sequential Algorithms in Nonlinear Programming' by K. Jittorntrum

- page 18, line 3 : replace "Condition 2.2.1" by "Condition 2.2.0"
- page 23, line 12 : for " $J_3(\tilde{x}^*, \tilde{u}^*)$ " read " $J_2(\tilde{x}^*, \tilde{u}^*)$ "
- page 76, line 11 : for "sufficiently" read "sufficient"
- page 92, lines 4 & 5 : replace "e" by " $\tilde{e}$ " and " $J_2(\tilde{x}^*) \neq 0$ " by " $J_2(\tilde{x}^*) \neq \emptyset$ "
- page 113, line 12 : replace " $\tilde{u}^{(k+1)}$ " by " $\tilde{u}^{(k-1)}$ "
- page 151, line 4 : for "locally unique" read "locally unique in a neighbourhood of $\tilde{x}^*$ "
- page 155, line 9 : for "compelementarity" read "complementarity"
- page 167, line 7 : replace "Topia by "Tapia" (also on page 249, line 12)
- page 191, line -4 : for "Hestene" read "Hestenes"
- page 194, lines 4 & 8 : for "convex" read "strongly convex"
- page 236, line 8 : replace "0.3375" by "0.3335"
- line -3 : replace "11.8496" by "11.8396"
- page 237, line 12 : replace $a_{24} = "4"$ by $a_{24} = "0.4"$
-

SEQUENTIAL ALGORITHMS
IN
NONLINEAR PROGRAMMING

by

Krisorn Jittorntrum

A thesis submitted to the
Australian National University
for the degree of Doctor of Philosophy

July, 1978

ACKNOWLEDGEMENTS

This work was carried out at the Computing Research Group (previously known as Computer Centre), the Australian National University.

I wish to thank my supervisor and the head of the group, Mike Osborne, for the general support and encouragement that he has lent me and the constructive criticism that he has lent my work.

I wish to thank Bob Anderssen for his careful checking and weeding out many of my non-English expressions. I would also like to thank John Paine, Graeme Chandler, Mark Lukas and Andreas Griewank for checking the manuscripts and for their help with the many rules of English grammar.

I wish to thank all the members of the Computing Research Group for the welcome and friendliness that they have extended to me during my stay. I would also like to acknowledge my appreciation of the many courses, seminars and discussions organized by the staff and students of the Computing Research Group, which I feel have broadened my knowledge of mathematics/computing and experience in academic life.

I also wish to thank the typist, Mrs. Susan Love, whose patience and undoubted skill are clearly evident in the thesis.

PREFACE

Some of the results of this thesis have been established during or as the result of collaboration and discussion with Dr Mike Osborne. In particular, §3.2 has been published jointly as Jittorntrum and Osborne (1978).

The results of Appendix 2, subsection 4.2.5 and §4.3 have been published as Jittorntrum (1977, 1978a, 1978b). Several other papers based on the results of §2.2, §2.3, §3.3 and subsections 4.1.3 and 4.2.3 are being prepared/submitted for publication.

Unless otherwise stated in the text, the work described is my own.

K. Jittorntrum

K. Jittorntrum

ABSTRACT

In this thesis, a class of methods for solving the following non-linear programming problem:

$$\begin{aligned} & \text{minimize } f(\underline{x}) \\ & \text{subject to } \underline{x} \in S = \{h_i(\underline{x}) = 0, i = 1, \dots, l; g_j(\underline{x}) \geq 0, j = 1, \dots, m\} \end{aligned}$$

is investigated. The class of methods under investigation consists of the ones which reduce the computational process to a sequence of unconstrained minimization subproblems. These methods are generally noted for their simplicity and robustness. The aim is to analyse and extend existing results on the structure and convergence of these methods, and, in certain cases, to introduce modifications for improving the numerical performance of existing algorithms.

Because of the possible effects degeneracy and near-degeneracy may have on the numerical performance of algorithms, this thesis is initially concerned with the structural differences between degenerate and non-degenerate problems. This leads naturally to a new interpretation of degeneracy, some structural differences between degenerate and non-degenerate problems, necessary and sufficient conditions for 'isolated' solutions, the concept and properties of an 'extended' problem, and a sensitivity analysis for degenerate problems.

Barrier function methods, specifically first order (logarithmic function) methods, are then investigated. It is shown that, if certain regularity conditions are satisfied, the successive estimates $\underline{x}(r)$ of the solution $\underline{x}(0)$ lie on a smooth trajectory. If the non-degeneracy assumption is not satisfied, then the convergence to $\underline{x}(0)$ is order $r^{1/2}$ compared with order r for non-degeneracy. A measure of sensitivity is introduced which becomes large when the non-degeneracy assumption is close

to violation. It is shown that this sensitivity measure is related to the growth of $\frac{d^i x}{dr^i}$ with respect to i for fixed r . Using this information the extrapolation procedure can be analysed and the number of useful extrapolation steps can be predicted. A source for a number of shortcomings in the classical algorithm is identified, and a modification is proposed. Properties and convergence of the modified algorithm are analysed. The numerical performance of the classical and the modified algorithms is compared for a range of test problems.

Penalty function and multiplier methods, specifically the quadratic penalty function method and the corresponding augmented Lagrangian methods, are treated next. Order of convergence results for the quadratic loss penalty function and the augmented Lagrangian methods are extended to include problems with degenerate inequality constraints. For an equality constrained problem, it is known that an augmented Lagrangian method is just an unconstrained maximization of the augmented dual function, $D_\sigma(\underline{v}) = \min_x \{f(\underline{x}) - \underline{v}^T \underline{h} + \frac{\sigma}{2} \underline{h}^T \underline{h}\}$, with respect to $\underline{v}$ for σ large enough. It is now shown that the Newton iteration for $\underline{v}$ based on maximizing $D_\sigma(\underline{v})$ can be decomposed into a Powell/Hestenes iteration followed by a Newton-like correction to $\underline{v}$. Utilizing the decomposition result, two acceleration techniques are proposed for improving the convergence of the Powell/Hestenes multiplier method. The first of these two techniques is a simple device superimposed on the original Powell/Hestenes iteration to make use of information from the previous iteration. For problems with only one constraint, this acceleration technique is equivalent to replacing the second (Newton-like) part of the decomposition by a finite difference approximation. The other acceleration technique is just a quasi-Newton technique. Numerical results using various updating schemes for $\underline{v}$ are reported and discussed.

Finally, possible uses for the methods developed in this thesis

(v)

are discussed including constrained optimization without derivatives and optimal control problems with terminal constraints.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	(i)
PREFACE	(ii)
ABSTRACT	(iii)
 CHAPTER 1: INTRODUCTION AND SUMMARY	
1.1 INTRODUCTION	1
1.2 SUMMARY	4
 CHAPTER 2: MATHEMATICAL PROGRAMMING - THEORY	
2.1 CONDITIONS FOR A CONSTRAINED MINIMUM	10
2.1.1 Introduction	10
2.1.2 First Order Conditions	12
2.1.3 Second Order Conditions	14
2.2 STRICT COMPLEMENTARITY AND RELATED RESULTS	18
2.2.1 Introduction	18
2.2.2 First Order Analysis	25
2.2.3 Second Order Analysis	27
2.2.4 The Unique Solution of 1st Order and its Relation to Linear Programming	33
2.2.5 The Extended Problem for Inequality Constraints	38
2.3 SENSITIVITY ANALYSIS	46
2.3.1 Introduction and Classical Results	46
2.3.2 Sensitivity Analysis without Strict Complementarity	49
 CHAPTER 3: BARRIER FUNCTION METHODS	
3.1 CLASSICAL BARRIER FUNCTION METHODS	65
3.1.1 Introduction	65
3.1.2 Logarithmic Barrier Function Methods	67
3.1.3 Difficulties Associated with Barrier Function Methods	78
3.2 TRAJECTORY ANALYSIS AND EXTRAPOLATION	83
3.2.1 Introduction	83
3.2.2 Nature of the Trajectory in the Degenerate Case	86
3.2.3 A Measure of Degeneracy	101
3.2.4 The Numerical Performance of Extrapolation	107
3.3 MODIFIED BARRIER FUNCTION METHODS	112
3.3.1 Introduction	112

	Page
3.3.2 Properties of the Modified Barrier Function Algorithm	117
3.3.3 Implementation and Numerical Performance	132
CHAPTER 4: PENALTY FUNCTION/MULTIPLIER METHODS	
4.1 PENALTY FUNCTION METHODS	147
4.1.1 Introduction	147
4.1.2 The Quadratic Penalty Function Method	150
4.1.3 The Quadratic Loss Penalty Function Method	153
4.1.4 Difficulties with Penalty Function Methods	163
4.2 MULTIPLIER/AUGMENTED LAGRANGIAN METHODS	165
4.2.1 Introduction	165
4.2.2 Properties of Augmented Lagrangian Methods	168
4.2.3 Inequality Constraints and Degeneracy	174
4.2.4 Newton's Method and its Decomposition for the Augmented Dual Problem	189
4.3 ACCELERATED CONVERGENCE FOR THE POWELL/HESTENES MULTIPLIER METHOD	197
4.3.1 Introduction	197
4.3.2 Derivation of a Simple Accelerating Device	198
4.3.3 Quasi-Newton Update for the Multiplier	204
4.3.4 Numerical Performance and Discussion	208
CHAPTER 5: APPLICATIONS AND CONCLUSION	
5.1 CONSTRAINED OPTIMIZATION WITHOUT DERIVATIVES	214
5.2 OPTIMAL CONTROL PROBLEMS WITH TERMINAL CONSTRAINTS	223
5.3 CONCLUSION	226
APPENDIXES	
A1 - CONVEX FUNCTIONALS	229
A2 - AN IMPLICIT FUNCTION THEOREM	230
A3 - EXTRAPOLATION	233
A4 - TEST PROBLEMS	236
REFERENCES	242

CHAPTER 1

INTRODUCTION AND SUMMARY

1.1 INTRODUCTION

The computational character of numerous optimization applications is essentially that of a *mathematical programming problem* (MPP) which can be formulated as follows:

$$\begin{aligned} & \text{minimize } f(\underline{x}) \\ & \text{subject to } h_i(\underline{x}) = 0 \quad , \quad i = 1, 2, \dots, l, \\ & \text{and } g_i(\underline{x}) \geq 0 \quad , \quad i = 1, 2, \dots, m, \end{aligned} \tag{1.1-1}$$

where $f(\underline{x})$, the *objective function*, $\{h_i(\underline{x})\}$ and $\{g_i(\underline{x})\}$, the *constraint functions*, are real-valued functions of a vector $\underline{x}$ in the n -dimensional vector space E^n . It will be assumed throughout this thesis that the problem functions, $f(\underline{x})$, $\{h_i(\underline{x})\}$ and $\{g_i(\underline{x})\}$, are at least continuous and further restrictions, in the form of differentiability conditions, will be imposed when necessary. Much of the interest in the MPP (often referred to as a non-linear programming problem) has developed since the basic results of Dantzig (1951) and Kuhn and Tucker (1951) were published, and has been stimulated by the advent of modern computers and advances in many fields of application including operations research, economics, applied mathematics and industrial design and control. By placing restrictions on the choice of $f(\underline{x})$, $\{h_i(\underline{x})\}$ and $\{g_i(\underline{x})\}$, classes of problems (for example: linear programming, quadratic programming and convex programming problems), which are more tractable than the general problem (1.1-1), can be identified. In this thesis, however, we consider the solution of the MPP under the assumption that $f(\underline{x})$, $\{h_i(\underline{x})\}$ and $\{g_i(\underline{x})\}$ are non-linear.

Depending on the values of ℓ and m , non-linear programming problems are classified as follows:

- (i) $\ell = 0, m = 0$: unconstrained minimization with no constraints present;
- (ii) $\ell = 0, m \neq 0$: inequality constrained problem (ICP), when only inequality constraints are present;
- (iii) $\ell \neq 0, m = 0$: equality constrained problem (ECP), when only equality constraints are present;
- (iv) $\ell \neq 0, m \neq 0$: mixed constrained problem, when both equality and inequality constraints are present.

In this thesis, we restrict attention to constrained problems, and to methods for solving constrained problems which reduce the computational process to a *sequence* of unconstrained minimization subproblems, minimizing a *compound function* combining in a particular way the objective function, f , the constraint functions, $\{h_i\}$ and $\{g_i\}$, and possibly one or more *controlling parameters*.

A survey of the literature shows that three classes of methods can be identified:

- (i) *Barrier Function and Interior Point Techniques*: These methods are constrained to search within the feasible region, and therefore are only applicable to inequality constrained problems. In part, the development of these methods has been motivated by the desirability and necessity to maintain feasibility throughout a computation.

- (ii) *Penalty Function and Exterior Point Techniques:* In their original form, these methods have fallen from favour in recent years. However, because they are related to other methods, such as multiplier methods, there is still a strong theoretical interest in them.
- (iii) *Lagrangian and Multiplier Techniques:* These methods are based on the use of the Lagrangian function associated with the MPP.

Our interest in these sequential methods is motivated in part by the development of efficient unconstrained minimization techniques, and by the simplicity, robustness and wide applicability of the resulting algorithms. There is an extensive literature on these methods. However, when inequality constraints are involved, most of the existing theoretical results on the convergence of these methods is limited to non-degenerate problems. The extension of such results to degenerate problems is a major aim of the thesis.

With an analysis of degeneracy as the main goal, we analyse in Chapter 2 the structure of the MPP, and establish several new results about mathematical programming. Barrier function methods are treated in Chapter 3, penalty function and multiplier methods in Chapter 4. When and where appropriate, we shall briefly survey the development of each class of methods and list appropriate classical results before introducing the new results. Possible applications of methods developed are discussed in Chapter 5. Several results depend heavily on a stronger form of the implicit function theorem. This stronger form is stated and proved in Appendix 2.

1.2 SUMMARY

Because the discussion in all the following chapters is involved and detailed, we initially summarise the major results presented, so that the nature of the original motivation is not obscured.

Chapter 2

In this chapter, the various results on mathematical programming which are required subsequently are developed. Most of the new results in this chapter are consequences of the examination of possible effects of degeneracy. Some classical results on necessary conditions and sufficient conditions for constrained minima are presented in §2.1.

In §2.2, possible structural differences between degenerate and non-degenerate problems are investigated. As a result of this investigation, several new results are proved which give new insights into the nature of the problem, particularly those which reflect the nature of degeneracy. A new interpretation of degeneracy is given in subsection 2.2.1. The local solution, $\tilde{x}^*$, of the MPP is also a local solution of one or more equality constrained problems which are constructed from the original MPP by dropping some inequality constraints and treating the remaining inequality constraints as equality constraints. The original MPP is degenerate if more than one such equality constrained problem can be constructed, and non-degenerate otherwise. This new interpretation is important in the sense that it gives rise to a technique for analysing the convergence of a class of algorithms when applied to degenerate problems. In subsection 2.2.2, it is shown that if the linear independence condition (a 'well-known' first order condition) is satisfied, then degeneracy is equivalent to the existence of a trajectory with 'special' properties. This type of trajectory is of interest because, when it exists, it is utilized by classical barrier function methods even though it results in slower convergence. In subsection 2.2.3, we introduce

the concept of a unique (local) solution of 2nd order: $\tilde{x}^*$ is a unique local solution of 2nd order if there exist $\delta > 0$ and $a > 0$ such that

$$f(\tilde{x}) - f(\tilde{x}^*) \geq a \|\tilde{x} - \tilde{x}^*\|^2, \quad \forall \tilde{x} \in S, \|\tilde{x} - \tilde{x}^*\| \leq \delta. \quad (1.2-1)$$

We go on to give a necessary and sufficient condition for a local solution, $\tilde{x}^*$, to be a unique (local) solution of 2nd order. This condition is independent of degeneracy considerations. In subsection 2.2.4, we supplement this result by giving some necessary and sufficient conditions for $\tilde{x}^*$ to be a unique (local) solution of 1st order, and then use them to explain the role of degeneracy in linear programming. These results on necessary and sufficient conditions for a local solution, $\tilde{x}^*$, to be a unique (local) solution of 1st or 2nd order are of fundamental interest, and could have far-reaching consequences. Their full exploitation is however beyond the scope of this thesis. In subsection 2.2.5, it is shown that, if inequality constraints are present, then, by using slack variables, they can be transformed into equality constraints, resulting in an *extended* equality constrained problem. The structure of this extended problem is of interest because many methods for solving an equality constrained problem can be generalized in a natural way to handle inequality constraints, implicitly solving the extended problem. The structural dependence of the extended problem on the original problem is analysed, and it is shown that degeneracy of the original problem is a crucial factor determining the structure of the extended problem. This result gives insight into how degeneracy may effect the performance of certain algorithms which were originally developed for equality constrained problems.

Finally in §2.3, because of the present importance of degenerate problems, classical results on sensitivity analysis are extended to degenerate problems. It is shown that even when the problem is degenerate, most

classical sensitivity results still hold with the exception that the optimal solution, $\tilde{x}(\epsilon)$, is no longer guaranteed to be differentiable with respect to the perturbation parameter ϵ ; it may only be continuous in ϵ . The differentiability of $\tilde{x}(\epsilon)$ is in a sense recovered by showing that, even if $\tilde{x}(\epsilon)$ is not differentiable, the one-sided directional derivatives of $\tilde{x}(\epsilon)$ at $\tilde{x}(0)$ always exist and are well defined.

Chapter 3

This chapter is devoted to barrier function methods. In §3.1, some results on the classical barrier function method are presented. It is shown that if certain regularity conditions are satisfied, then the successive estimates $\tilde{x}(r)$ of the solution $\tilde{x}(0)$ lie on a smooth trajectory. These results are generalized to *separable* barrier function methods which have separate controlling parameter for each constraint. The attractiveness of barrier function methods and their computational difficulties are discussed.

In §3.2, a detailed trajectory analysis is presented for the classical barrier function method as the controlling parameter $r \rightarrow 0$. It is shown that if the problem is degenerate, then the classical barrier function method follows a trajectory with 'special' properties of the type discussed in subsection 2.2.2, resulting in lower order of convergence $O(r^{1/2})$ instead of the usual order of $O(r)$ for non-degenerate problems. A sensitivity measure γ is then introduced which is finite for non-degenerate problems and becomes large when the non-degeneracy assumption is close to violation:

$$\gamma = \max_{i,j} \frac{u_i^* \|\nabla g_i(\tilde{x}^*)\|}{u_j^* \|\nabla g_j(\tilde{x}^*)\|}, \quad (1.2-2)$$

where g_i, g_j are active at $\tilde{x}^*$, and u_i^*, u_j^* are the corresponding multipliers. We then go on to show that γ is simply related to the

growth, for fixed sufficiently small r , of $\frac{d^i \tilde{x}(r)}{dr^i}$ for increasing i :

$$\left\| \frac{d^i \tilde{x}}{dr^i} \right\| = O(\gamma^{2i-1}) . \quad (1.2-3)$$

This analysis shows how both degeneracy and near-degeneracy can effect the performance of classical barrier function algorithms, particularly the performance with respect to extrapolation. Finally, in this section, we analyse the extrapolation procedure and give a heuristic criterion for predicting the number of useful extrapolation steps. Numerical results are given.

In §3.3, several shortcomings of the classical barrier function method are identified in terms of its inability to discriminate between constraints: the classical barrier function method assigns equal weight to each constraint. A modified barrier function method is proposed. It assigns to each constraint at each iteration a relative weight based on the corresponding multiplier estimate obtained during the previous iteration. It is shown that, by weighting each constraint according to its multiplier estimate, the modified barrier function method possesses several desirable properties which are not possessed by the classical barrier function method. In particular:

- (i) It automatically identifies and discards inactive constraints.
- (ii) For non-degenerate problems, there is a basis for extrapolation which is not sensitive to the size of γ .
- (iii) For degenerate problems, the convergence is of order $O(r_k)$, the same as that for non-degenerate problems.

Details are given for an implementation of the modified barrier function algorithm. Numerical performance and other characteristics of

the modified barrier function method and the classical barrier function method are compared for a range of test problems.

Chapter 4

This chapter is devoted to the use of penalty function and Lagrangian function techniques for the construction of unconstrained minimization subproblems. Specifically, the quadratic penalty function method and related augmented Lagrangian methods are investigated. In §4.1, some classical results on the penalty function method are presented. Existing results are supplemented by establishing convergence for the penalty function method for degenerate problems.

In §4.2, basic results on multiplier/augmented Lagrangian methods are presented. Again existing results are supplemented by deriving corresponding results for degenerate problems. For an equality constrained problem, if certain regularity conditions are satisfied, augmented Lagrangian methods are equivalent to unconstrained optimization techniques for maximizing the augmented dual function

$$D_{\sigma}(\underline{v}) = \min_{\underline{x}} \left\{ f(\underline{x}) - \underline{v}^T \underline{h} + \frac{\sigma}{2} \underline{h}^T \underline{h} \right\}, \quad (1.2-4)$$

with respect to $\underline{v}$ for σ sufficiently large. We then show that the Newton update for the multiplier, $\underline{v}$, based on maximizing the augmented dual function, $D_{\sigma}(\underline{v})$, can be decomposed into taking the Powell/Hestenes update followed by another Newton-like update. The second part of the decomposition does not depend on the parameter σ , and in this sense the Newton update for maximizing the augmented dual function, $D_{\sigma}(\underline{v})$, is separated into a part depending only on σ , and another part depending only on the problem itself.

In §4.3, utilizing the decomposition result, two techniques are proposed for accelerating, without derivatives, the convergence of the Powell/Hestenes multiplier method. The first acceleration technique is a simple device superimposed on the original Powell/Hestenes iteration making use of information from the previous iteration. For problems with only one constraint, this device becomes a secant method, replacing the second (Newton-like) part of the decomposition by a finite difference approximation. The other acceleration technique is just a quasi-Newton technique. Numerical results comparing various updating schemes for y are presented. The relative merits of each scheme are discussed.

Chapter 5

In this chapter we discuss possible applications of methods developed in this thesis such as constrained optimization without derivatives and optimal control problems with terminal constraints. We conclude by discussing some of the implications for mathematical programming theory of the results presented.

CHAPTER 2

MATHEMATICAL PROGRAMMING - THEORY

2.1 CONDITIONS FOR A CONSTRAINED MINIMUM2.1.1 Introduction

A seemingly trivial but rather important feature underlying mathematical programming theory is that the MPP is solved by finding, instead of an optimal value f^* , an $\tilde{x}^*$ such that $f(\tilde{x}^*)$ is optimal. Therefore, for the MPP to be solvable, suitable regularity conditions need to be imposed to ensure that $\tilde{x}^*$ can be uniquely determined. The basic element of an optimization problem consists of a point set $S \subseteq E^n$ and a real-valued function $f(\tilde{x})$ defined on S . In the sequel, the *constraint set* S will be explicitly defined by a number of function inequalities and equations:

$$S = \left\{ \tilde{x} \mid h_i(\tilde{x}) = 0, i \in I ; g_i(\tilde{x}) \geq 0, i \in J \right\} \quad (2.1-1)$$

where $I = \{1, 2, \dots, \ell\}$ and $J = \{1, 2, \dots, m\}$ denote the sets of equality and inequality constraints, respectively. Points $\tilde{x}$ in S are called *feasible*; S is often referred to as the *feasible region*.

Definition

$\tilde{x}^*$ is a *global minimizer* of f on S if

$$f(\tilde{x}^*) \leq f(\tilde{x}), \quad \forall \tilde{x} \in S ; \quad (2.1-2)$$

correspondingly, $f(\tilde{x}^*)$ is the *global minimum* of f on S .

Remark 2.1.1

$\tilde{x}^*$ exists, for example, if S is finite or if S is compact and $f(\tilde{x})$ is continuous on S .

Definition

$\tilde{x}^*$ is a *local minimizer* of f on S , if there exists $\delta > 0$ such that

$$f(\tilde{x}^*) \leq f(\tilde{x}), \quad \forall \tilde{x} \in N(S, \tilde{x}^*, \delta), \quad (2.1-3)$$

where

$$N(S, \tilde{x}^*, \delta) = \{\tilde{x} | \tilde{x} \in S; \|\tilde{x} - \tilde{x}^*\| < \delta\}. \quad (2.1-4)$$

If strict inequality holds in either (2.1-2) or (2.1-3) whenever $\tilde{x} \neq \tilde{x}^*$, then the minimum is said to be *isolated*.

Definition

Associated with the MPP, we define the *Lagrangian function*, $L_0(\tilde{x}, \tilde{v}, \tilde{u})$, the *tangent cone*, $T(S, \tilde{x})$, the *active inequality constraint set*, $J_0(\tilde{x})$, and the *set of constrained directions*, $C(\tilde{x})$, as follows:

$$L_0(\tilde{x}, \tilde{v}, \tilde{u}) = f(\tilde{x}) - \sum_{i \in I} \tilde{v}_i h_i(\tilde{x}) - \sum_{j \in J} \tilde{u}_j g_j(\tilde{x}); \quad (2.1-5)$$

$$T(S, \tilde{x}) = \left\{ \tilde{t} | \exists \{\lambda_n\} \geq 0, \{\tilde{x}_n\} \rightarrow \tilde{x}, \{\tilde{x}_n\} \subseteq S \right. \\ \left. \text{such that } \lim_{n \rightarrow \infty} \|\lambda_n(\tilde{x}_n - \tilde{x}) - \tilde{t}\| = 0 \right\}; \quad (2.1-6)$$

$$J_0(\tilde{x}) = \{j | j \in J; g_j(\tilde{x}) = 0\}; \quad (2.1-7)$$

$$C(\tilde{x} | \tilde{x} \in S) = \left\{ \tilde{t} | \nabla h_i(\tilde{x})^T \tilde{t} = 0, \forall i \in I; \nabla g_j(\tilde{x})^T \tilde{t} \geq 0, \forall j \in J_0(\tilde{x}) \right\}. \quad (2.1-8)$$

Definition (Kuhn-Tucker Points)

Assume that $f, \{h_i\}$ and $\{g_j\}$ are once differentiable, and that f has a minimum on S at $\tilde{x}^0$, then f is said to satisfy the *Kuhn-Tucker condition* at $\tilde{x}^0$ if there exist vectors of multipliers $\tilde{v}$ and $\tilde{u}$ such that

$$(i) \quad \nabla f(\tilde{x}^0) = \sum_{i \in I} v_i \nabla h_i(\tilde{x}^0) + \sum_{i \in J_0(\tilde{x}^0)} u_i \nabla g_i(\tilde{x}^0), \quad (2.1-9)$$

$$(ii) \quad u_i \geq 0, \quad i \in J_0(\tilde{x}^0); \quad (2.1-10)$$

or, alternatively,

$$(i) \quad \nabla f(\tilde{x}^0) = \sum_{i \in I} v_i \nabla h_i(\tilde{x}^0) + \sum_{i \in J} u_i \nabla g_i(\tilde{x}^0), \quad (2.1-9a)$$

$$(ii) \quad \tilde{u}^T \tilde{g}(\tilde{x}^0) = 0, \quad (2.1-11)$$

and

$$(iii) \quad \tilde{u} \geq 0, \quad (2.1-10a)$$

where zero multipliers are introduced corresponding to the inactive constraints. If $u_i > 0$ whenever $g_i(\tilde{x}^0) = 0$, *strict complementarity* is said to hold. If strict complementarity does not hold, then the situation is *degenerate*. By convention, set $u_i = 0$ if $\nabla g_i(\tilde{x}^0) = 0$.

Remark 2.1.2

It is important to appreciate that the Kuhn-Tucker condition need not hold. A simple counter-example has been constructed by Kuhn and Tucker (1951).

2.1.2 First Order Conditions

Conditions which ensure that the constraint set is capable of supporting Lagrange multiplier type necessary conditions (2.1-9) and (2.1-10) are known as *constraint qualifications*. Extensive discussion and proofs of these conditions have been given by many authors including John (1948), Kuhn and Tucker (1951), Arrow, Hurwicz and Uzawa (1961), Hestenes (1966), Mangasarian and Fromowitz (1967), Fiacco and McCormick (1968), Lootsma (1970) and Gould and Tolle (1971). A condition which is both sufficient and necessary has been given by Gould and Tolle (1971).

This condition will be stated here because it is of fundamental importance in mathematical programming.

Definition

The *polar cone* to a point set $A \subseteq E^n$ is the set

$$A^* = \{ \tilde{x} | \tilde{x}^T \tilde{y} \leq 0, \forall \tilde{y} \in A \} . \quad (2.1-12)$$

Lemma 2.1.1

Assuming that all the problem functions are once differentiable, the Kuhn-Tucker condition is satisfied at $\tilde{x}^0$ for every function f which has a minimum on S at $\tilde{x}^0$ if and only if

$$T(S, \tilde{x}^0)^* = C(\tilde{x}^0)^* . \quad (2.1-13)$$

The term first order (or C1) refers to the once differentiability assumption imposed on the problem functions. A well-known first order condition introduced by Kuhn and Tucker (1951) is the following:

Definition (C1 Constraint Qualification)

Let $\tilde{x}^0 \in S$ and assume that $\{h_i\}$ and $\{g_j\}$ are *once* differentiable, then the C1 constraint qualification holds at $\tilde{x}^0$, if $\forall t \neq 0$, where $\tilde{t} \in C(\tilde{x}^0)$, $\tilde{t}$ is tangential to a *once* differentiable arc which starts from $\tilde{x}^0$ and is contained in S .

Remark 2.1.3

From their definitions in (2.1-6) and (2.1-8), it follows that $T(S, \tilde{x}^0) \subseteq C(\tilde{x}^0)$. The C1 constraint qualification implies that $C(\tilde{x}^0) \subseteq T(S, \tilde{x}^0)$. Hence $C(\tilde{x}^0) = T(S, \tilde{x}^0)$ and Lemma 2.1.1 holds trivially.

Corollary 2.1.1 (*Kuhn-Tucker Necessary Theorem*; Fiacco and McCormick (1968), p.20)

If the functions f , $\{h_i\}$ and $\{g_i\}$ are once differentiable at $\tilde{x}^* \in S$, and if the C1 constraint qualification holds at $\tilde{x}^*$, then a necessary condition that $\tilde{x}^*$ be a local minimum of the MPP is that there exist vectors $\tilde{v}^*$ and $\tilde{u}^*$ such that $(\tilde{x}^*, \tilde{v}^*, \tilde{u}^*)$ satisfies (2.1-9) .. (2.1-11). □

Condition 2.1.1 (*Linear Independence Condition*)

The active constraint gradients $\{\nabla h_i(\tilde{x}^*), i \in I\}$ and $\{\nabla g_i(\tilde{x}^*), i \in J_0(\tilde{x}^*)\}$ are linearly independent.

The C1 constraint qualification does not imply that the multipliers $\tilde{v}^*$ and $\tilde{u}^*$ are necessarily unique. For this reason and the reason that it is computationally useful and simple to verify, the linear independence condition is often imposed.

Lemma 2.1.2 (Fiacco and McCormick (1968), p.22)

A sufficient condition that the C1 constraint qualification holds at a point $\tilde{x}^0 \in S$ is that the constraint gradients $\{\nabla h_i(\tilde{x}^0), i \in I\}$ and $\{\nabla g_i(\tilde{x}^0), i \in J_0(\tilde{x}^0)\}$ be linearly independent. □

2.1.3 Second Order Conditions

In certain cases, it is possible to further characterise local minima of f on S by looking at second derivative information. In particular, it is possible to state sufficient conditions for $\tilde{x}^*$ to be an isolated (local) minimizer of f on S . When the problem functions are twice differentiable, taking into account the curvature of the problem functions, we introduce a stronger requirement than the C1 constraint qualification (cf., Fiacco and McCormick (1968), p.25).

Definition (C2 Constraint Qualification)

Let $\tilde{x}^0 \in S$, and assume that the problem functions are *twice* differentiable, then the C2 constraint qualification holds at $\tilde{x}^0$ if $\forall \tilde{t} \neq 0$, where $\tilde{t} \in C(\tilde{x}^0)$, there exists a *twice* differentiable curve $\tilde{x}(\theta)$, $(0 \leq \theta \leq \epsilon; \epsilon > 0)$, along which $h_i(\tilde{x}(\theta)) = 0, \forall i \in I; g_j(\tilde{x}(\theta)) = \theta \cdot \nabla g_j(\tilde{x}^0)^T \tilde{t}, \forall j \in J_0(\tilde{x}^0); \tilde{x}(0) = \tilde{x}^*$ and $d\tilde{x}(0)/d\theta = \tilde{t}$.

Lemma 2.1.3 (Hestenes (1966), p.35)

Let $\tilde{x}^0 \in S$, and assume that the problem functions are twice differentiable, then a sufficient condition for the C2 constraint qualification to hold at $\tilde{x}^0$ is that the constraint gradients $\{\nabla h_i(\tilde{x}^0), i \in I\}$ and $\{\nabla g_j(\tilde{x}^0), j \in J_0(\tilde{x}^0)\}$ be linearly independent. $\square$

Definition

Let $\tilde{x}^* \in S$ be a Kuhn-Tucker point of the MPP, with $(\tilde{v}^*, \tilde{u}^*)$ as the associated multipliers (not necessarily unique). Define the sets $J_1(\tilde{x}^*, \tilde{u}^*), J_2(\tilde{x}^*, \tilde{u}^*), J_3(\tilde{x}^*, \tilde{u}^*)$ and $G(\tilde{x}^*, \tilde{u}^*)$ as follows:

$$J_1(\tilde{x}^*, \tilde{u}^*) = \left\{ j \mid g_j(\tilde{x}^*) = 0, \tilde{u}_j^* > 0 \right\}; \quad (2.1-14)$$

$$J_2(\tilde{x}^*, \tilde{u}^*) = \left\{ j \mid g_j(\tilde{x}^*) = 0, \tilde{u}_j^* = 0 \right\}; \quad (2.1-15)$$

$$J_3(\tilde{x}^*, \tilde{u}^*) = \left\{ j \mid g_j(\tilde{x}^*) > 0, \tilde{u}_j^* = 0 \right\}; \quad (2.1-16)$$

$$G(\tilde{x}^*, \tilde{u}^*) = \left\{ \tilde{t} \mid \nabla h_i(\tilde{x}^*)^T \tilde{t} = 0, \forall i \in I; \nabla g_j(\tilde{x}^*)^T \tilde{t} = 0, \forall j \in J_1(\tilde{x}^*, \tilde{u}^*) \right\}. \quad (2.1-17)$$

Remark 2.1.4

Eqn. (2.1-11) implies that $J_3(\tilde{x}^*, \tilde{u}^*)$ is independent of $\tilde{u}^*$. When no ambiguities can arise, or when the sets are uniquely defined by $\tilde{x}^*$ alone (e.g., when $\tilde{u}^*$ is unique), the dependence of the sets $J_1(\tilde{x}^*, \tilde{u}^*), J_2(\tilde{x}^*, \tilde{u}^*)$ and $J_3(\tilde{x}^*, \tilde{u}^*)$ on $\tilde{u}^*$ will be suppressed.

A second order necessary condition for a constrained minimum can be stated as follows:

Lemma 2.1.4 (Fiacco and McCormick (1968), p.25)

Let $\tilde{x}^* \in S$, and assume that the C2 constraint qualification holds at $\tilde{x}^*$, then a necessary condition for $\tilde{x}^*$ to be a local minimum of the MPP is that there exist multipliers $\tilde{v}^*$ and $\tilde{u}^*$ such that the Kuhn-Tucker condition ((2.1-9)-(2.1-11)) is satisfied, and such that the inequality

$$\tilde{t}^T \nabla_{\tilde{x}}^2 L_0(\tilde{x}^*, \tilde{v}^*, \tilde{u}^*) \tilde{t} \geq 0 \quad (2.1-18)$$

holds for every $\tilde{t} \in \left\{ \tilde{t} \mid \nabla h_i(\tilde{x}^*)^T \tilde{t} = 0, \forall i \in I; \nabla g_j(\tilde{x}^*)^T \tilde{t} = 0, \forall j \in J_0(\tilde{x}^*) \right\}$. $\square$

In the theory of unconstrained minimization, a second order necessary condition is that $\nabla_{\tilde{x}}^2 f(\tilde{x}^*)$ be positive semi-definite. By converting positive semi-definite to positive definite, a second order sufficient condition is thereby obtained. By analogy, a second order sufficient condition for a constrained minimum can be obtained by replacing positive semi-definite with positive definite requirement in (2.1-18). This is indeed sufficient in most situations. However, if there are degenerate inequality constraints, we must require $\nabla_{\tilde{x}}^2 L_0$ to be positive definite on a larger subspace.

Condition 2.1.2 (*Second Order Sufficiency Condition*)

(at a Kuhn-Tucker point $(\tilde{x}^*, \tilde{v}^*, \tilde{u}^*)$). There exists $a > 0$ such that

$$\tilde{t}^T \nabla_{\tilde{x}}^2 L_0(\tilde{x}^*, \tilde{v}^*, \tilde{u}^*) \tilde{t} \geq a \|\tilde{t}\|^2, \forall \tilde{t} \in G(\tilde{x}^*, \tilde{u}^*) \quad (2.1-19)$$

Lemma 2.1.5 (Fiacco and McCormick (1968), p.30)

Assuming that the problem functions are twice differentiable, sufficient conditions for a point $\tilde{x}^* \in S$ to be an isolated local solution

of the MPP are that there exist vectors $\underline{v}^*$ and $\underline{u}^*$ such that $(\underline{x}^*, \underline{v}^*, \underline{u}^*)$ is a Kuhn-Tucker point, and that there exists $a > 0$ such that

$$\underline{t}^T \nabla_{\underline{x}}^2 L_0(\underline{x}^*, \underline{v}^*, \underline{u}^*) \underline{t} \geq a \|\underline{t}\|^2, \forall \underline{t} \in G(\underline{x}^*, \underline{u}^*) . \quad \square$$

Definition (*Regular Solution Condition*)

A point $\underline{x}^* \in S$ is said to be a *regular* local solution of the MPP, if the following conditions are satisfied:

- (i) the constraint gradients $\{\nabla h_i(\underline{x}^*), i \in I\}$ and $\{\nabla g_i(\underline{x}^*), i \in J_0(\underline{x}^*)\}$ are linearly independent;
- (ii) the multipliers $\underline{v}^*$ and $\underline{u}^*$ are unique;
- (iii) the second order sufficiency condition (2.1-19) holds at $\underline{x}^*$.

Remark 2.1.5

(a) Condition (ii) above is explicitly stated here to make the definition self-contained. By Lemma 2.1.2 and Eqn. (2.1-9), it follows that $\underline{v}^*$ and $\underline{u}^*$ exist and are unique as a consequence of (i). By (i) and (iii), a *regular* local solution implicitly implies that the problem functions are at least twice differentiable and that the C1 and C2 constraint qualifications are satisfied.

(b) Clearly, there exists an $A > 0$ such that

$$\underline{t}^T \nabla_{\underline{x}}^2 L_0(\underline{x}^*, \underline{v}^*, \underline{u}^*) \underline{t} \leq A \|\underline{t}\|^2 \quad (2.1-20)$$

for all $\underline{t}$ satisfying the requirements of the second order sufficiency condition.

2.2 STRICT COMPLEMENTARITY AND RELATED RESULTS

2.2.1 Introduction

Condition 2.2.1 (*Strict Complementarity/Non-degeneracy*)

(at a Kuhn-Tucker point $(\tilde{x}^*, \tilde{v}^*, \tilde{u}^*)$) $u_j^* > 0, \forall j \in J$ such that $g_j(\tilde{x}^*) = 0$.

Strict complementarity is often imposed on the MPP along with the linear independence and second order sufficiency conditions. However, unlike the other conditions, the role of strict complementarity has attracted very little attention in the past. In most texts on non-linear programming, it is merely stated without motivation. Two reasons that seem to justify its use are:

(i) In linear programming, it has an important role in determining that $\tilde{x}^*$ is an isolated solution (see subsection 2.2.4); it is also a necessary and sufficient condition for certain primal-dual results (see: Luenberger (1973), p.78). Because linear programming is just a special case of non-linear programming, it seems logical that strict complementarity should be imposed on non-linear programming problems.

(ii) Because the Kuhn-Tucker point $(\tilde{x}^*, \tilde{v}^*, \tilde{u}^*)$ satisfies the following system of equations

$$\nabla_x L_0(\tilde{x}, \tilde{v}, \tilde{u}) = 0, \quad (2.2-1)$$

$$h(\tilde{x}) = 0, \quad (2.2-2)$$

$$\tilde{u}^T g(\tilde{x}) = 0, \quad (2.2-3)$$

the isolation or uniqueness of the solution $(\tilde{x}^*, \tilde{v}^*, \tilde{u}^*)$ is closely related to the condition that requires the Jacobian of the system of equations (2.2-1)-(2.2-3) with respect to $(\tilde{x}, \tilde{v}, \tilde{u})$ to be non-singular at $(\tilde{x}^*, \tilde{v}^*, \tilde{u}^*)$. This condition is usually referred to as the *Jacobian uniqueness condition*.

$$J(\tilde{x}, \tilde{v}, \tilde{u}) = \begin{bmatrix} \nabla_{\tilde{x}}^2 L_0(\tilde{x}, \tilde{v}, \tilde{u}) & [-\nabla h_i(\tilde{x}), i \in I] & [-\nabla g_i(\tilde{x}), i \in J] \\ [\nabla h_i(\tilde{x})^T, i \in I] & 0 & 0 \\ [u_i \nabla g_i(\tilde{x})^T, i \in J] & 0 & [\text{diag } g_i(\tilde{x}), i \in J] \end{bmatrix} \quad (2.2-4)$$

$J(\tilde{x}^*, \tilde{v}^*, \tilde{u}^*)$ is called the *Jacobian* of the MPP. From the structure of $J(\tilde{x}^*, \tilde{v}^*, \tilde{u}^*)$, it follows that for $J(\tilde{x}^*, \tilde{v}^*, \tilde{u}^*)$ to be non-singular, strict complementarity must hold.

Lemma 2.2.1

Let $\tilde{x}^*$ be a local solution of the MPP and assume that the problem functions are twice differentiable, then the Jacobian uniqueness condition is satisfied if and only if

- (i) the linear independence condition,
- (ii) the second order sufficiency condition, and
- (iii) strict complementarity

are simultaneously satisfied.

Proof

Sufficiency: Consider the following matrix

$$K = \begin{bmatrix} \nabla_{\tilde{x}}^2 L_0(\tilde{x}^*, \tilde{v}^*, \tilde{u}^*) & [-\nabla h_i(\tilde{x}^*), i \in I] & [-\nabla g_i(\tilde{x}^*), i \in J_0(\tilde{x}^*)] & [-\nabla g_i(\tilde{x}^*), i \notin J_0(\tilde{x}^*)] \\ [\nabla h_i(\tilde{x}^*)^T, i \in I] & 0 & 0 & 0 \\ [\nabla g_i(\tilde{x}^*)^T, i \in J_0(\tilde{x}^*)] & 0 & 0 & 0 \\ 0 & 0 & 0 & [\text{diag } g_i(\tilde{x}^*), i \notin J_0(\tilde{x}^*)] \end{bmatrix} \quad (2.2-5)$$

Because strict complementarity holds, to show that $J(\underline{x}^*, \underline{v}^*, \underline{u}^*)$ is non-singular it is only necessary to show that K is non-singular.

Assume K is singular. Then there exists a non-trivial vector $\begin{bmatrix} \underline{a} \\ \underline{b} \\ 0 \end{bmatrix}$, ($\underline{a}$ has n components and $\underline{b}$ has $|I| + |J_0(\underline{x}^*)|$ components), such that

$$K \begin{bmatrix} \underline{a} \\ \underline{b} \\ 0 \end{bmatrix} = 0. \quad (2.2-6)$$

This implies that

$$\nabla h_i(\underline{x}^*)^T \underline{a} = 0, \quad i \in I; \quad (2.2-7a)$$

$$\nabla g_i(\underline{x}^*)^T \underline{a} = 0, \quad i \in J_0(\underline{x}^*). \quad (2.2-7b)$$

Premultiplying by $[\underline{a}^T, \underline{b}^T, 0]$ gives

$$\underline{a}^T \nabla_{\underline{x}}^2 L_0(\underline{x}^*, \underline{v}^*, \underline{u}^*) \underline{a} = 0, \quad (2.2-8)$$

so that the second order sufficiency condition implies that $\underline{a} = 0$.

But then the linear independence condition gives that $\underline{b} = 0$ which establishes a contradiction.

Necessity: Clearly, for $J(\underline{x}^*, \underline{v}^*, \underline{u}^*)$ to be non-singular the linear independence condition and strict complementarity must hold, and it is left to be shown that the second order sufficiency condition also holds.

If $J(\underline{x}^*, \underline{v}^*, \underline{u}^*)$ is singular, there is a non-trivial vector $\begin{bmatrix} \underline{y} \\ \underline{a} \\ \underline{b} \end{bmatrix}$ satisfying

$$J(\underline{x}^*, \underline{v}^*, \underline{u}^*) \begin{bmatrix} \underline{y} \\ \underline{a} \\ \underline{b} \end{bmatrix} = 0. \quad (2.2-9)$$

This relation gives

$$\nabla h_i(x^*)^T \underline{y} = 0, \quad i \in I, \quad (2.2-10)$$

$$u_i^* \nabla g_i(x^*)^T \underline{y} + b_i g_i(x^*) = 0, \quad i \in J, \quad (2.2-11)$$

and

$$\nabla_x^2 L_0(x^*, v^*, u^*) \underline{y} - \sum_{i \in I} a_i \nabla h_i(x^*) - \sum_{i \in J} b_i \nabla g_i(x^*) = 0. \quad (2.2-12)$$

From (2.2-11), using strict complementarity, it follows that $u_i^* > 0$ implies $\nabla g_i(x^*)^T \underline{y} = 0$ while $u_i^* = 0$ implies $b_i = 0$. Now consider the problem:

$$\begin{aligned} & \text{minimize } \underline{y}^T \nabla_x^2 L_0(x^*, v^*, u^*) \underline{y} \\ & \text{subject to } \nabla h_i(x^*)^T \underline{y} = 0, \quad i \in I, \end{aligned} \quad (2.2-13)$$

$$\nabla g_i(x^*)^T \underline{y} = 0, \quad i \in J_0(x^*),$$

and

$$\|\underline{y}\|^2 = 1.$$

Clearly, the constraint gradients are linearly independent as $2\underline{y} = \nabla_y(\|\underline{y}\|^2)$ is in the orthogonal complement of the set spanned by the other constraint gradients. Let $\underline{y}^*$ minimize the objective function (the minimum exists as the constraint set is compact), then linear independence of the constraint gradients ensures that there exist multipliers λ , a_i , $i \in I$, b_i , $i \in J_0(x^*)$ such that

$$2\nabla_x^2 L_0(x^*, v^*, u^*) \underline{y}^* - 2\lambda \underline{y}^* - \sum_{i \in I} a_i \nabla h_i(x^*) - \sum_{i \in J_0(x^*)} b_i \nabla g_i(x^*) = 0, \quad (2.2-14)$$

whence

$$\lambda = \underline{y}^{*T} \nabla_x^2 L_0(x^*, v^*, u^*) \underline{y}^* = \min \underline{y}^T \nabla_x^2 L_0(x^*, v^*, u^*) \underline{y}. \quad (2.2-15)$$

Now if $\lambda = 0$, (2.2-14) shows that (2.2-10)-(2.2-12) are satisfied and

hence $J(\underline{x}^*, \underline{v}^*, \underline{u}^*)$ is singular. Thus if $J(\underline{x}^*, \underline{u}^*, \underline{v}^*)$ is non-singular then $\lambda \neq 0$. In this situation, (2.2-15) and Lemma 2.1.4 imply that $\lambda > 0$ and thus the second order sufficiency condition is satisfied. $\square$

Remark 2.2.1

The proof that the matrix K in (2.2-5) is non-singular does not depend on strict complementarity. Hence, K will be non-singular even when strict complementarity does not hold.

Structural consequences associated with the use of the strict complementarity condition are now examined. They appear to imply that the Jacobian uniqueness condition (i.e., strict complementarity) should be imposed. However, in subsequent sections and chapters, it will be evident that the structure and properties of several algorithms and of the problem itself can be proved without the use of strict complementarity. This implies that strict complementarity need not be imposed on the problem. It is therefore concluded that the *Jacobian uniqueness condition* is too strong and should be replaced by the *regular solution condition*.

Let us recall that the inequality constraint set J is partitioned into the sets $J_0(\underline{x}^*)$ and $J_3(\underline{x}^*)$; or $J_1(\underline{x}^*, \underline{u}^*)$, $J_2(\underline{x}^*, \underline{u}^*)$ and $J_3(\underline{x}^*)$. $J_0(\underline{x}^*) = J_1(\underline{x}^*, \underline{u}^*) \cup J_2(\underline{x}^*, \underline{u}^*)$ is the *active* (inequality) constraint set for the primal problem. If the dual problem is well defined, the set $J_2(\underline{x}^*, \underline{u}^*) \cup J_3(\underline{x}^*, \underline{u}^*) = \{j \mid u_j^* = 0\}$ will be its *active* (inequality) constraint set. For the original (primal) problem, the set $J_2(\underline{x}^*, \underline{u}^*) \cup J_3(\underline{x}^*, \underline{u}^*)$ can be identified as the *inactive* (inequality) constraint set in the multiplier sense: $J_2(\underline{x}^*, \underline{u}^*) \cup J_3(\underline{x}^*, \underline{u}^*)$ is the index set of constraints with *zero* multiplier. The set $J_2(\underline{x}^*, \underline{u}^*)$ is the *degenerate* constraint set. With this partitioning of the set J , strict complementarity can be reinterpreted as follows:

Lemma 2.2.2

At a Kuhn-Tucker point $(\tilde{x}^*, \tilde{v}^*, \tilde{u}^*)$, strict complementarity is equivalent to

$$(J_1(\tilde{x}^*, \tilde{u}^*) \cup J_2(\tilde{x}^*, \tilde{u}^*)) \cap (J_2(\tilde{x}^*, \tilde{u}^*) \cup J_3(\tilde{x}^*, \tilde{u}^*)) \equiv J_2(\tilde{x}^*, \tilde{u}^*) = \phi, \quad (2.2-16)$$

or that there exists only *one* index set R such that

$$J_1(\tilde{x}^*, \tilde{u}^*) \subseteq R \subseteq J_1(\tilde{x}^*, \tilde{u}^*) \cup J_2(\tilde{x}^*, \tilde{u}^*) . \quad (2.2-17)$$

A trivial consequence of Lemma 2.2.2 is the following Corollary.

Corollary 2.2.1

At a Kuhn-Tucker point $(\tilde{x}^*, \tilde{v}^*, \tilde{u}^*)$, if strict complementarity does not hold, then there exist *more than one*, but only a *finite* number, of different index sets R such that

$$J_1(\tilde{x}^*, \tilde{u}^*) \subseteq R \subseteq J_1(\tilde{x}^*, \tilde{u}^*) \cup J_3(\tilde{x}^*, \tilde{u}^*) . \quad \square$$

When strict complementarity does not hold, $J_2(\tilde{x}^*, \tilde{u}^*) \neq \phi$.

This means that some constraints are in a sense both active and inactive.

Thus the sets $J_1(\tilde{x}^*, \tilde{u}^*) \cup J_2(\tilde{x}^*, \tilde{u}^*)$ and $J_2(\tilde{x}^*, \tilde{u}^*) \cup J_3(\tilde{x}^*, \tilde{u}^*)$ are unlikely to be stable with respect to perturbations.

One way of handling inequality constraints is to ignore inactive constraints and to treat active inequality constraints as equality constraints. Algorithms based on this approach have been studied by several authors, such as Osborne and Ryan (1972) and Fletcher (1973). An essential feature of this approach is the existence of some strategy for identifying the set of 'active' constraints.

Lemma 2.2.3

If, at $\tilde{x}^* \in S$, both the Kuhn-Tucker condition and the second order sufficiency condition are satisfied for the MPP, then, for any index set R such that $J_1(\tilde{x}^*, \tilde{u}^*) \subseteq R \subseteq J_1(\tilde{x}^*, \tilde{u}^*) \cup J_2(\tilde{x}^*, \tilde{u}^*)$, both the Kuhn-Tucker condition and the second order sufficiency condition are satisfied at $\tilde{x}^*$ for the following problem $P[R]$:

$$\begin{aligned} & \text{minimize } f(\tilde{x}) \\ & \text{subject to } h_i(\tilde{x}) = 0, \quad i \in I, \\ & \text{and } g_i(\tilde{x}) = 0, \quad i \in R. \end{aligned} \tag{2.2-18}$$

Proof

By using the same multipliers $\tilde{v}^*$ and $\tilde{u}^*$ (only those components $u_i^*, i \in R$, of $\tilde{u}^*$ are needed), the results follow directly from the definitions of the Kuhn-Tucker condition and the second order sufficiency condition. □

Theorem 2.2.1

If $\tilde{x}^*$ is a *regular* local solution of the MPP, then $\tilde{x}^*$ is a *regular* local solution of the problem $P[R]$, for any index set R such that

$$J_1(\tilde{x}^*) \subseteq R \subseteq J_1(\tilde{x}^*) \cup J_2(\tilde{x}^*).$$

Proof

Similar to Lemma 2.2.3, the result follows directly from the definition of a regular solution and Remark 2.1.4. □

Remark 2.2.2

If strict complementarity holds, there is only one such problem

$P[R]$, with $R = J_1(\underline{x}^*, \underline{u}^*)$. If strict complementarity does not hold, then there are several different problems $P[R]$. The key condition that allows the removal of some or all degenerate constraints is the second order sufficiency condition.

2.2.2 First Order Analysis

Lemma 2.2.4

If $\underline{x}^* \in S$ is a local solution of the MPP and the linear independence condition holds at $\underline{x}^*$, then $\forall i \in J_0(\underline{x}^*)$, $u_i^* = 0$ if and only if $\exists \underline{t} \in C(\underline{x}^*)$ such that $\nabla f(\underline{x}^*)^T \underline{t} = 0$ but $\nabla g_i(\underline{x}^*)^T \underline{t} > 0$.

Proof

Sufficiency: As $\underline{x}^*$ is a Kuhn-Tucker point, it follows from (2.1-9) that

$$\nabla f(\underline{x}^*) - \sum_{j \in I} v_j^* \nabla h_j(\underline{x}^*) - \sum_{j \in J_0(\underline{x}^*)} u_j^* \nabla g_j(\underline{x}^*) = 0. \quad (2.2-19)$$

Hence

$$\nabla f(\underline{x}^*)^T \underline{t} - \sum_{j \in I} v_j^* \nabla h_j(\underline{x}^*)^T \underline{t} - \sum_{\substack{j \in J_0(\underline{x}^*) \\ j \neq i}} u_j^* \nabla g_j(\underline{x}^*)^T \underline{t} = u_i^* \nabla g_i(\underline{x}^*)^T \underline{t}. \quad (2.2-20)$$

Since $\nabla h_j(\underline{x}^*)^T \underline{t} = 0$, $\forall j \in I$, and $\nabla g_j(\underline{x}^*)^T \underline{t} \geq 0$, $u_j^* \geq 0$, $\forall j \in J_0(\underline{x}^*)$, Eqn. (2.2-20) can only hold if $u_i^* = 0$.

Necessity: Let $u_i^* = 0$, then $\nabla g_i(\underline{x}^*)$ can be expressed as $\underline{t} + \underline{s}$, where $\underline{t}$ is orthogonal to and $\underline{s}$ is a linear combination of $\{\nabla h_j(\underline{x}^*), j \in I; \nabla g_j(\underline{x}^*), j \neq i, j \in J_0(\underline{x}^*)\}$. The linear independence condition implies that $\underline{t} \neq 0$. Thus

$$\nabla g_i(\underline{x}^*)^T \underline{t} = \|\underline{t}\|^2 + \underline{s}^T \underline{t} = \|\underline{t}\|^2 > 0. \quad (2.2-21)$$

From (2.2-19) we now have

$$\nabla f(\underline{x}^*) = \sum_{j \in I} v_j^* \nabla h_j(\underline{x}^*) + \sum_{\substack{j \in J_0(\underline{x}^*) \\ j \neq \lambda}} u_j^* \nabla g_j(\underline{x}^*) . \quad (2.2-22)$$

Therefore $\nabla f(\underline{x}^*)^T \underline{t} = 0$ and $\underline{t} \in C(\underline{x}^*)$ by construction. $\square$

Corollary 2.2.2

If $\underline{x}^* \in S$ is a local solution of the MPP, and the linear independence condition holds at $\underline{x}^*$, then $\forall i \in J_0(\underline{x}^*)$, $u_i^* > 0$ if and only if $\forall \underline{t} \in C(\underline{x}^*)$, $\nabla f(\underline{x}^*)^T \underline{t} = 0$ implies $\nabla g_i(\underline{x}^*)^T \underline{t} = 0$.

Proof

A direct consequence of Lemma 2.2.4. $\square$

If the linear independence condition holds at $\underline{x}^* \in S$, a local solution of the MPP, then the C1 constraint qualification, Lemma 2.2.4 and Corollary 2.2.2 imply that degeneracy corresponds to the existence of a feasible trajectory tangential to the surface $f(\underline{x}) = f(\underline{x}^*)$, and non-tangential to the constraint surfaces, $g_j(\underline{x}) = 0$, for some $j \in J_0(\underline{x}^*)$. More specifically, we state this result in the following Theorem.

Theorem 2.2.2

If $\underline{x}^* \in S$ is a local solution of the MPP, and the linear independence condition holds at $\underline{x}^*$, then the MPP is *degenerate* if and only if there exists a trajectory $\underline{x}(\theta) \in S$, $(0 \leq \theta \leq \epsilon; \epsilon > 0)$, $\underline{x}(0) = \underline{x}^*$, such that for $\underline{t} = d\underline{x}(0)/d\theta$, $\nabla f(\underline{x}^*)^T \underline{t} = 0$, $\nabla h_i(\underline{x}^*)^T \underline{t} = 0$, $\forall i \in I$, and $\nabla g_j(\underline{x}^*)^T \underline{t} > 0$ for some $j \in J_0(\underline{x}^*)$. Moreover, if such a trajectory exists, it can be chosen such that $\forall j \in J_0(\underline{x}^*)$, $\nabla g_j(\underline{x}^*)^T \underline{t} = 0$ if and only if $u_j^* > 0$.

Proof

The first part of the theorem follows from Lemmas 2.1.2 and 2.2.4 and Corollary 2.2.2. For the second part of the theorem, because strict complementarity does not hold, $\forall j \in J_2(\tilde{x}^*)$ there exists $\tilde{t}_j \in C(\tilde{x}^*)$ such that $\nabla f(\tilde{x}^*)^T \tilde{t}_j = 0$ but $\nabla g_j(\tilde{x}^*)^T \tilde{t}_j > 0$. Let $\tilde{t} = \sum_{j \in J_2(\tilde{x}^*)} \tilde{t}_j$. The result then follows from Lemma 2.1.2 and Corollary 2.2.2. $\square$

Theorem 2.2.2 shows that degeneracy corresponds to the existence of a trajectory with 'special' properties. One characteristic of such a trajectory is that the approach to $\tilde{x}^*$ is tangential to some constraint surfaces, $g_j(\tilde{x}^*) = 0$, $j \in J_0(\tilde{x}^*)$, and non-tangential to others. For an inequality constrained problem, this means that the approach to $\tilde{x}^*$, which is on the boundary of S , is tangential to the boundary of S . The result in Theorem 2.2.2 is therefore of fundamental importance to algorithms which generate feasible trajectories approaching $\tilde{x}^*$. If for some reason an algorithm has preference towards such a trajectory, then degeneracy can be expected to have adverse effects on the performance of the algorithm. An example of such an algorithm (see §3.2) is the classical logarithmic barrier function algorithm.

2.2.3 Second Order Analysis

As already mentioned, in linear programming (see subsection 2.2.4), strict complementarity is an important factor for determining whether or not the solution $\tilde{x}^*$ is isolated. Therefore, in non-linear programming, it is only appropriate that we should determine the effect of strict complementarity on the degree of isolation of $\tilde{x}^*$. The following definition is based on the terminology of Cromme (1978).

Definition (Unique Solution of 2nd Order)

A point $\tilde{x}^* \in S$ is a unique local solution of 2nd order of the MPP if there exist $a > 0$ and $\delta > 0$ such that

$$f(\tilde{x}) - f(\tilde{x}^*) \geq a \|\tilde{x} - \tilde{x}^*\|^2, \quad \forall \tilde{x} \in N(S, \tilde{x}^*, \delta) . \quad (2.2-23)$$

In the unconstrained case (i.e., $S = E^n$), it is possible to verify that the requirement for $\tilde{x}^*$ to be a unique local solution of 2nd order is equivalent to the requirement that $\nabla_x^2 f(\tilde{x}^*)$ be positive definite. Correspondingly, it would appear that the second order sufficiency condition (Condition 2.1.2) should be the required condition for constrained problems. However, this is not so. The corresponding condition is in fact slightly weaker than the second order sufficiency condition.

Definition

At a Kuhn-Tucker point $(\tilde{x}^*, \tilde{v}^*, \tilde{u}^*)$, the set $W(\tilde{x}^*, \tilde{u}^*)$ is defined as

$$W(\tilde{x}^*, \tilde{u}^*) = T(S, \tilde{x}^*) \cap G(\tilde{x}^*, \tilde{u}^*) . \quad (2.2-24)$$

Condition 2.2.1

(at a Kuhn-Tucker point $(\tilde{x}^*, \tilde{v}^*, \tilde{u}^*)$) There exists $a > 0$ such that

$$\tilde{t}^T \nabla_x^2 L_0(\tilde{x}^*, \tilde{v}^*, \tilde{u}^*) \tilde{t} \geq a \|\tilde{t}\|^2, \quad \forall \tilde{t} \in W(\tilde{x}^*, \tilde{u}^*) . \quad (2.2.25)$$

Theorem 2.2.3

Let $\tilde{x}^* \in S$ be a local solution of the MPP, and assume that the C2 constraint qualification holds at $\tilde{x}^*$, then a necessary and sufficient condition for $\tilde{x}^*$ to be a unique local solution of 2nd order is that Condition 2.2.1 holds at $\tilde{x}^*$.

Proof

Sufficiency: It is a well-known result that Condition 2.2.1 implies $\tilde{x}^*$ to be a unique local solution of 2nd order (Hestenes (1966), p.37).

Necessity: Assume that $\tilde{x}^*$ is a unique local solution of 2nd order, but Condition 2.2.1 does not hold at $\tilde{x}^*$. Then $\exists \tilde{t} \neq 0$, $\tilde{t} \in W(\tilde{x}^*, \tilde{u}^*)$ such that

$$\tilde{t}^T \nabla_x^2 L_0(\tilde{x}^*, \tilde{v}^*, \tilde{u}^*) \tilde{t} \leq 0. \quad (2.2.26)$$

The C2 constraint qualification then implies that there exists a twice differentiable trajectory $\tilde{x}(\theta) \in S$, $(0 \leq \theta \leq \varepsilon; \varepsilon > 0)$, such that $\tilde{x}(0) = \tilde{x}^*$, $d\tilde{x}(0)/d\theta = \tilde{t}$, and

$$h_{\tilde{i}}(\tilde{x}(\theta)) = 0, \quad \forall \tilde{i} \in I, \quad (2.2-27)$$

$$g_{\tilde{j}}(\tilde{x}(\theta)) = 0, \quad \forall \tilde{j} \in J_1(\tilde{x}^*, \tilde{u}^*). \quad (2.2.28)$$

From (2.2-27), using twice differentiability, it follows that

$$\frac{d\tilde{x}(0)^T}{d\theta} \nabla_x^2 h_{\tilde{i}}(\tilde{x}(0)) \frac{d\tilde{x}(0)}{d\theta} + \nabla h_{\tilde{i}}(\tilde{x}(0))^T \frac{d^2 \tilde{x}(0)}{d\theta^2} = 0, \quad \forall \tilde{i} \in I. \quad (2.2-29)$$

Similarly, from (2.2-28)

$$\frac{d\tilde{x}(0)^T}{d\theta} \nabla_x^2 g_{\tilde{j}}(\tilde{x}(0)) \frac{d\tilde{x}(0)}{d\theta} + \nabla g_{\tilde{j}}(\tilde{x}(0))^T \frac{d^2 \tilde{x}(0)}{d\theta^2} = 0, \quad \forall \tilde{j} \in J_1(\tilde{x}^*, \tilde{u}^*). \quad (2.2-30)$$

Because $\tilde{x}^*$ is a unique local solution of 2nd order, there exist $a > 0$ and $\delta > 0$ such that $\forall \theta \in [0, \varepsilon]$ with $\|\tilde{x}(\theta) - \tilde{x}^*\| < \delta$,

$$f(\tilde{x}(\theta)) - f(\tilde{x}^*) \geq a \|\tilde{x}(\theta) - \tilde{x}^*\|^2. \quad (2.2.31)$$

The Taylor's expansion about $\theta = 0$ of (2.2-31) gives

$$0. \nabla f(\tilde{x}(0))^T \frac{d\tilde{x}(0)}{d\theta} + \theta^2 \left[\frac{d\tilde{x}(0)}{d\theta}^T \nabla_x^2 f(\tilde{x}(0)) \frac{d\tilde{x}(0)}{d\theta} + \nabla f(\tilde{x}(0))^T \frac{d^2 \tilde{x}(0)}{d\theta^2} \right] \\ \geq a \cdot \theta^2 \cdot \frac{d\tilde{x}(0)}{d\theta}^T \cdot \frac{d\tilde{x}(0)}{d\theta} + o(\theta^2) \quad (2.2-32)$$

At the Kuhn-Tucker point $\tilde{x}(0)$, $\nabla f(\tilde{x}(0))^T \tilde{t} = 0$. Hence (2.2-32) becomes

$$\frac{d\tilde{x}(0)}{d\theta}^T \nabla_x^2 f(\tilde{x}(0)) \frac{d\tilde{x}(0)}{d\theta} + \nabla f(\tilde{x}(0))^T \frac{d^2 \tilde{x}(0)}{d\theta^2} \geq a \frac{d\tilde{x}(0)}{d\theta}^T \frac{d\tilde{x}(0)}{d\theta} \quad (2.2-33)$$

By combining (2.2-29), (2.2-30) and (2.2-33), setting $\frac{d\tilde{x}(0)}{d\theta} = \tilde{t}$, $\tilde{x}(0) = \tilde{x}^*$ and using the fact that $\nabla_x L_0(\tilde{x}^*, \tilde{v}^*, \tilde{u}^*) = 0$, the required contradiction

$$\tilde{t}^T \nabla_x^2 L_0(\tilde{x}^*, \tilde{v}^*, \tilde{u}^*) \tilde{t} \geq a \|\tilde{t}\|^2$$

is obtained. □

For illustrative purpose, a simple example is introduced:

Example 2.2.1

$$\begin{aligned} \text{minimize} \quad & f(\tilde{x}) = x_2 \\ \text{subject to} \quad & g_1(\tilde{x}) = x_2 \geq 0, \\ & g_2(\tilde{x}) = x_2 - x_1 \geq 0, \\ \text{and} \quad & g_3(\tilde{x}) = x_1 + x_2 \geq 0. \end{aligned}$$

The solution is $\tilde{x}^* = (0, 0)$ with a family of multipliers

$$\tilde{v}^* = \left(\alpha, \frac{1-\alpha}{2}, \frac{1-\alpha}{2} \right), \quad 0 \leq \alpha \leq 1.$$

Remark 2.2.3

(a) The C2 constraint qualification requirement in Theorem 2.2.3 is too strong. The condition that is actually needed in the proof is that a twice differentiable trajectory $\tilde{x}(\theta) \in S$, with $d\tilde{x}(0)/d\theta = \tilde{t}$, can be found for every $\tilde{t} \in T(S, \tilde{x}^*) \cap G(\tilde{x}^*, \tilde{u}^*)$. An example for which the C2 constraint qualification does not hold but the above condition holds will be presented and discussed in subsection 2.2.5 (Remark 2.2.5d).

(b) Condition 2.2.1 is trivially satisfied if $W(\tilde{x}^*, \tilde{u}^*) = \{0\}$, containing only the null vector. Some special properties of the case $W(\tilde{x}^*, \tilde{u}^*) = \{0\}$ will be covered later in subsection 2.2.4.

(c) The second order necessary and sufficient condition (Condition 2.2.1) is actually weaker than the second order sufficiency condition (Condition 2.1.2). In Ex. 2.2.1, Condition 2.2.1 is satisfied for $\alpha = 1$, but the second order sufficiency condition is not.

(d) The second order sufficiency condition depends on the particular choice of the multipliers, $\tilde{u}^*$. In Ex. 2.2.1, the second order sufficiency condition is satisfied for all choices of $\tilde{u}^*$ with $0 \leq \alpha < 1$, but not for the choice $\alpha = 1$.

(e) Condition 2.2.1 does not depend on any particular choice of the multipliers, $\tilde{u}^*$. In Ex. 2.2.1, Condition 2.2.1 is satisfied for all choices of $\tilde{u}^*$, $0 \leq \alpha \leq 1$. By Theorem 2.2.3, if Condition 2.2.1 is satisfied for a particular choice of $\tilde{u}^*$, then $\tilde{x}^*$ is a unique (local) solution of 2nd order. Hence, by Theorem 2.2.3, Condition 2.2.1 is satisfied for any other choice of the multipliers, $\tilde{u}^*$.

Lemma 2.2.5

If $\tilde{x}^* \in S$ is a local solution of the MPP, and the C1 constraint qualification holds at $\tilde{x}^*$, then for equality constrained problems, mixed

constrained problems and inequality constrained problems in non-degenerate situations, Condition 2.2.1 and the second order sufficiency condition are equivalent.

Proof

It is sufficient to consider only inequality constrained problems in non-degenerate situations. From Remark 2.1.3, Eqns. (2.1-8) and (2.1-17) and non-degeneracy, it follows that

$$\begin{aligned} W(\tilde{x}^*, \tilde{u}^*) &= T(S, \tilde{x}^*) \cap G(\tilde{x}^*, \tilde{u}^*) \\ &= C(\tilde{x}^*) \cap G(\tilde{x}^*, \tilde{u}^*) \\ &= G(\tilde{x}^*, \tilde{u}^*) . \end{aligned}$$

□

Remark 2.2.4

Lemma 2.2.5 shows that Condition 2.2.1 and the second order sufficiency condition are equivalent except for some degenerate situations. Computationally, the second order sufficiency condition is the one that is usually imposed on the problem (cf. regular solution condition). This is because it gives a useful and computationally verifiable structure that $\nabla_x^2 L_0$ is positive definite in a tangent subspace.

The original aim of this subsection was to examine the effect of strict complementarity on the isolation of the solution $\tilde{x}^*$. Theorem 2.2.3 shows that the (2nd order) isolation of the solution $\tilde{x}^*$ does not depend on strict complementarity.

Consider a linear programming problem with a solution $\tilde{x}^*$. It is known (for example, see subsection 2.2.4) that if the active constraint gradients are linear independent and span the whole space E^n , then the isolation of $\tilde{x}^*$ depends critically on whether or not strict complementarity

is satisfied. This and the above result seem to indicate a contradiction. Note, however, that if a non-linear term $\|\underline{x} - \underline{x}^*\|^2$ is added to the objective function of the linear problem in question, then the solution $\underline{x}^*$ will be isolated even when strict complementarity does not hold. Lemma 2.1.5 shows that in order that $\underline{x}^*$ be isolated, it is sufficient to replace (as opposed to superimpose) the first order conditions and strict complementarity by the second order sufficiency condition.

2.2.4 The Unique Solution of 1st Order and its Relation to Linear Programming.

When all the problem functions are linear, the isolation of the solution $\underline{x}^*$ is equivalent to the existence of $\alpha > 0$ and $\delta > 0$ such that

$$f(\underline{x}) - f(\underline{x}^*) \geq \alpha \|\underline{x} - \underline{x}^*\|, \forall \underline{x} \in N(S, \underline{x}^*, \delta). \quad (2.2.34)$$

If the linear independence condition holds at $\underline{x}^*$, the properties of linear functions and Theorem 2.2.2 imply that if the problem is degenerate, then there exists a feasible trajectory $\underline{x}(\theta) = \underline{x}^* + \theta \underline{t}$, $\underline{t} \neq 0$, such that $f(\underline{x}(\theta)) = f(\underline{x}^*)$, $\forall \theta \in [0, \epsilon]$, $\epsilon > 0$. This confirms that strict complementarity has an important role in determining the isolation of $\underline{x}^*$. Similar to other branches of mathematics (e.g., differential equation), non-linear problems are often examined using some form of linearization. This is probably one of the main reasons why strict complementarity is required in non-linear programming problems. In this subsection, linear programming will be treated as a special case of non-linear programming. This approach, as the results will show, yields more insight and understanding about both linear and non-linear programming problems.

Let us consider Eqn. (2.2-34). Following the terminology used by Cromme (1978), $\underline{x}^*$ is called a *unique (local) solution of 1st order*

of the MPP. From subsection 2.2.3, Theorem 2.2.3 gives a necessary and sufficient condition for $\tilde{x}^*$ to be a unique solution of 2nd order. The question of interest is whether or not similar necessary and sufficient conditions can be given to ensure that $\tilde{x}^*$ is a unique solution of 1st order. Such conditions will first be given for a general non-linear programming problem; and then the special case of linear programming problem will be discussed.

Theorem 2.2.4

If $\tilde{x}^* \in S$ is a local minimizer of f , and the C1 constraint qualification holds at $\tilde{x}^*$, then the following statements are equivalent.

- (i) $\exists a > 0, \delta > 0$ such that $f(\tilde{x}) - f(\tilde{x}^*) \geq a \|\tilde{x} - \tilde{x}^*\|, \forall \tilde{x} \in N(S, \tilde{x}^*, \delta)$;
- (ii) $\nabla f(\tilde{x}^*)^T \tilde{t} > 0, \forall \tilde{t} \neq 0, \tilde{t} \in T(S, \tilde{x}^*)$;
- (iii) $W(\tilde{x}^*, \tilde{u}^*) = \{0\}$, independent of the choice of $\tilde{u}^*$.

Proof

(i) $\Rightarrow$ (ii) $T(S, \tilde{x}^*) = C(\tilde{x}^*)$ by Remark 2.1.3. Because the C1 constraint qualification holds at $\tilde{x}^*$, it follows that $\forall \tilde{t} \neq 0, \|\tilde{t}\| = 1, \tilde{t} \in T(S, \tilde{x}^*)$, there exists a feasible trajectory $\tilde{x}(\theta)$ such that $\tilde{x}(0) = \tilde{x}^*$ and $d\tilde{x}(0)/d\theta = \tilde{t}$. From (i), for θ small enough

$$f(\tilde{x}(\theta)) - f(\tilde{x}^*) \geq a \|\tilde{x}(\theta) - \tilde{x}^*\|. \quad (2.2-35)$$

Hence

$$\lim_{\theta \rightarrow 0^+} \frac{f(\tilde{x}(\theta)) - f(\tilde{x}^*)}{\|\tilde{x}(\theta) - \tilde{x}^*\|} = \nabla f(\tilde{x}^*)^T \tilde{t} \geq a. \quad (2.2-36)$$

(ii) $\Rightarrow$ (i) Suppose that (i) does not hold, then for every integer q there is a point $\tilde{x}_q \in S$ such that

$$\|x_q - x^*\| < \frac{1}{q}, \text{ and } f(x_q) - f(x^*) < \frac{1}{q} \|x_q - x^*\|. \quad (2.2-37)$$

Clearly, $x_q \neq x^*$. The sequence $\{(x_q - x^*)/\|x_q - x^*\|\}$ is a bounded sequence, and hence has a convergent subsequence. Replace the sequence $\{x_q\}$ by a subsequence, again denoted by $\{x_q\}$, which converges to x^* in a direction $\underline{t}$, with $\|\underline{t}\| = 1$, $\underline{t} \in T(S, x^*)$. It then follows that

$$\lim_{q \rightarrow \infty} \frac{f(x_q) - f(x^*)}{\|x_q - x^*\|} = \nabla f(x^*)^T \underline{t} \leq 0 \quad (2.2-38)$$

which is the required contradiction.

(ii) $\Rightarrow$ (iii) Because x^* is a Kuhn-Tucker point, there exist $\underline{v}^*$ and $\underline{u}^*$ such that

$$\nabla f(x^*) = \sum_{i \in I} v_i^* \nabla h_i(x^*) + \sum_{i \in J} u_i^* \nabla g_i(x^*). \quad (2.2-39)$$

$\forall \underline{t} \neq 0$, $\underline{t} \in T(S, x^*) = C(x^*)$, multiplying (2.2-39) by $\underline{t}$ gives

$$\nabla f(x^*)^T \underline{t} = \sum_{i \in I} v_i^* \nabla h_i(x^*)^T \underline{t} + \sum_{i \in J_0(x^*)} u_i^* \nabla g_i(x^*)^T \underline{t}. \quad (2.2-40)$$

Thus, $\nabla g_i(x^*)^T \underline{t} > 0$ for some $i \in J_0(x^*)$ such that $u_i^* > 0$, and hence $\underline{t} \notin G(x^*, \underline{u}^*)$. But trivially, the null vector belongs to both $T(S, x^*)$ and $G(x^*, \underline{u}^*)$, therefore $W(x^*, \underline{u}^*) = \{0\}$. The argument does not depend on any particular choice of $\underline{u}^*$, hence $W(x^*, \underline{u}^*) = \{0\}$ independent of the choice of $\underline{u}^*$.

(iii) $\Rightarrow$ (ii) If $W(x^*, \underline{u}^*) = \{0\}$ for some $\underline{u}^*$, this implies that given a vector $\underline{t} \neq 0$, $\underline{t} \in T(S, x^*) = C(x^*)$, then $\underline{t} \notin G(x^*, \underline{u}^*)$. This means that $\nabla g_i(x^*)^T \underline{t} > 0$ for some $i \in J_0(x^*)$ such that $u_i^* > 0$. Hence, by (2.2-40), $\nabla f(x^*)^T \underline{t} > 0$. Then (ii) $\Rightarrow$ (iii) also implies that $W(x^*, \underline{u}^*) = \{0\}$

for every choice of $\underline{u}^*$. □

Theorem 2.2.4 is a fundamental result in mathematical programming (see: Cromme (1978)). However, the full extent of its possible implications will not be exploited here. Attention is restricted to the implications of this result for linear programming, and possible connections with strict complementarity.

Corollary 2.2.3

For a linear programming problem, let $\underline{x}^*$ be a local solution of the problem, then the following statements are equivalent:

- (i) $\underline{x}^*$ is an isolated solution of the problem;
- (ii) $\exists a_2 > 0, \delta_2 > 0$ such that $f(\underline{x}) - f(\underline{x}^*) \geq a_2 \|\underline{x} - \underline{x}^*\|^2, \forall \underline{x} \in N(S, \underline{x}^*, \delta_2)$;
- (iii) $\exists a_1 > 0, \delta_1 > 0$ such that $f(\underline{x}) - f(\underline{x}^*) \geq a_1 \|\underline{x} - \underline{x}^*\|, \forall \underline{x} \in N(S, \underline{x}^*, \delta_1)$;
- (iv) $\nabla f(\underline{x}^*)^T \underline{t} > 0, \forall \underline{t} \neq 0, \underline{t} \in T(S, \underline{x}^*)$;
- (v) $W(\underline{x}^*, \underline{u}^*) = \{0\}$, independent of the choice of $\underline{u}^*$.

Proof

(i), (ii) and (iii) are equivalent because all the problem functions are linear. Also because the problem functions are linear, it follows that the C1 and C2 constraint qualifications are satisfied, hence $T(S, \underline{x}^*) = C(\underline{x}^*)$ and the remainder follows from Theorem 2.2.4. □

Corollary 2.2.4

If $\underline{x}^*$ is a local solution of a linear programming problem, and the linear independence condition holds at $\underline{x}^*$, then the following statements are equivalent:

- (i) $\tilde{x}^*$ is an isolated solution of the problem;
- (ii) $G(\tilde{x}^*, \tilde{u}^*) = \{0\}$;
- (iii) strict complementarity holds and the set of active constraint gradients span the whole space E^n ;
- (iv) the second order sufficiency condition holds at $\tilde{x}^*$.

Proof

(ii) $\Rightarrow$ (i) Since $W(\tilde{x}^*, \tilde{u}^*) = T(S, \tilde{x}^*) \cap G(\tilde{x}^*, \tilde{u}^*) = \{0\}$, the result follows from Corollary 2.2.3.

(ii) $\Leftrightarrow$ (iii) The result follows from the linear independence condition and definition of $G(\tilde{x}^*, \tilde{u}^*)$.

(ii) $\Rightarrow$ (iv) Trivially true.

(iv) $\Rightarrow$ (i) By Lemma 2.1.5.

(i) $\Rightarrow$ (ii) By Corollary 2.2.3, (i) $\Leftrightarrow W(\tilde{x}^*, \tilde{u}^*) = \{0\}$.

If strict complementarity does not hold, then $u_j^* = 0$ for some $j \in J_0(\tilde{x}^*)$. The linear independence condition implies that there exists $\tilde{t}_j \neq 0$ such that

$$\forall h_i(\tilde{x}^*)^T \tilde{t}_j = 0, \quad \forall i \in I; \quad \nabla g_i(\tilde{x}^*)^T \tilde{t}_j = 0, \quad \forall i \neq j, \quad i \in J_0(\tilde{x}^*);$$

$$\text{and} \quad \nabla g_j(\tilde{x}^*)^T \tilde{t}_j > 0.$$

This shows that $\tilde{t}_j \in C(\tilde{x}^*)$ and $\tilde{t}_j \in G(\tilde{x}^*, \tilde{u}^*)$, and hence $W(\tilde{x}^*, \tilde{u}^*) \neq \{0\}$ which yields a contradiction.

Therefore strict complementarity holds, and $G(\tilde{x}^*, \tilde{u}^*) \subseteq C(\tilde{x}^*)$.

Thus

$$G(\tilde{x}^*, \tilde{u}^*) = C(\tilde{x}^*) \cap G(\tilde{x}^*, \tilde{u}^*) = W(\tilde{x}^*, \tilde{u}^*) = \{0\}.$$

□

As a direct consequence of Corollary 2.2.4, we have the following result.

Corollary 2.2.5

Let $\tilde{x}^*$ be a local solution of a linear programming problem, and assume that (i) the linear independence condition holds at $\tilde{x}^*$, (ii) the active constraint gradients span the whole space E^n (i.e., the full rank condition holds), then strict complementarity holds at $\tilde{x}^*$ if and only if the second order sufficiency condition holds at $\tilde{x}^*$. □

Corollary 2.2.5 shows that, for linear programming problems, strict complementarity is in fact just a special case of the second order sufficiency condition. Hence, for non-linear programming problems, in the presence of the second order sufficiency condition, it seems illogical to require that strict complementarity is also satisfied (as a carry over from linear programming).

2.2.5 The Extended Problem for Inequality Constraints

Many mixed constrained minimization algorithms were first developed for equality constrained problems and then generalized to include inequality constraints.

Example: Consider the following quadratic penalty function method for an equality constrained problem:

$$P(\tilde{x}, \sigma) = f(\tilde{x}) + \frac{\sigma}{2} \sum_{i \in I} h_i(\tilde{x})^2. \quad (2.2-41)$$

This method has been generalized to handle inequality constraints in the following way:

$$P(\tilde{x}, \sigma) = f(\tilde{x}) + \frac{\sigma}{2} \sum_{i \in I} h_i(\tilde{x})^2 + \frac{\sigma}{2} \sum_{i \in J} [g_i(\tilde{x})]_-^2, \quad (2.2-42)$$

where

$$[a]_- = \min(a, 0).$$

In this subsection, a framework will be developed for generalizing methods for solving equality constrained problems to include inequality constrained and mixed constrained problems.

By introducing (square) slack variables, the original MPP can be transformed into an equality constrained problem. This equality constrained problem will be referred to as *the extended problem* or *the extended MPP*:

$$\begin{aligned} & \underset{x, z}{\text{minimize}} \quad F(\tilde{x}, \tilde{z}) = f(\tilde{x}) \\ & \text{subject to} \quad H_i(\tilde{x}, \tilde{z}) = h_i(\tilde{x}) = 0, \quad i \in I, \quad (2.2-43) \\ & \quad \text{and} \quad G_i(\tilde{x}, \tilde{z}) = g_i(\tilde{x}) - z_i^2 = 0, \quad i \in J, \end{aligned}$$

where $z_1, \dots, z_m$ represent slack variables.

Notation For convenience, the following notation will be adopted:

$$(\tilde{a}, \tilde{b}) = \begin{bmatrix} a_1 \\ \vdots \\ a_n \\ b_1 \\ \vdots \\ b_m \end{bmatrix}, \quad (2.2-44)$$

where $\tilde{a}$ is an n -vector, $\tilde{b}$ an m -vector. Note also that in this notation

it follows that

$$\tilde{x} = [x_1, x_2, \dots, x_n]^T = (x_1, x_2, \dots, x_n) .$$

It follows from the above construction that the sets of solutions for the original MPP and the extended $\overline{\text{MPP}}$ have a one-to-one correspondence. Specifically, $\tilde{x}^*$ is a solution of the original MPP if and only if $(\tilde{x}^*, \tilde{z}^*) = (x^*, g_1(x^*)^{\frac{1}{2}}, \dots, g_m(x^*)^{\frac{1}{2}})$ is a solution of the extended $\overline{\text{MPP}}$. In this way, any methods for solving equality constrained problem can be generalized to handle inequality constraints by solving the extended $\overline{\text{MPP}}$. Because the dependence on slack variables is known explicitly, the unconstrained minimization subproblem for many methods, such as penalty function and multiplier methods, can be carried out analytically with respect to $\tilde{z}$. Hence, slack variables need not appear explicitly in the computation. For example, consider the quadratic penalty function

$$\begin{aligned} P(\tilde{x}, \sigma) &= \min_{\tilde{z}} \left\{ f(\tilde{x}) + \frac{\sigma}{2} \sum_{i \in I} h_i(\tilde{x})^2 + \frac{\sigma}{2} \sum_{i \in J} [g_i(\tilde{x}) - z_i^2]^2 \right\} \\ &= f(\tilde{x}) + \frac{\sigma}{2} \sum_{i \in I} h_i(\tilde{x})^2 + \frac{\sigma}{2} \sum_{i \in J} [g_i(\tilde{x})]^2 . \end{aligned} \quad (2.2-45)$$

Because the extended $\overline{\text{MPP}}$ is being solved implicitly, it is appropriate to study possible effects of strict complementarity on the structure of the extended $\overline{\text{MPP}}$.

Lemma 2.2.6

Let $\tilde{x}^* \in S$ be a local solution of the MPP, then the linear independence condition is satisfied at $\tilde{x}^*$ for the MPP if and only if it is also satisfied at $(\tilde{x}^*, g_1(\tilde{x}^*)^{\frac{1}{2}}, \dots, g_m(\tilde{x}^*)^{\frac{1}{2}})$ for the extended $\overline{\text{MPP}}$.

Proof

Let $\tilde{c}_j^T = [0, \dots, 0, 1, 0, \dots, 0]$ with '1' in the j th column and '0' elsewhere. Then

$$\nabla_{(x,z)} G_j(\tilde{x}^*, \tilde{z}^*) = (\nabla g_j(\tilde{x}^*), -2g_j(\tilde{x}^*)^{\frac{1}{2}} \tilde{c}_j), \quad \forall j \in J. \quad (2.2-46)$$

If $\left\{ \begin{bmatrix} \nabla g_j(\tilde{x}^*) \\ -2g_j(\tilde{x}^*)^{\frac{1}{2}} \tilde{c}_j \end{bmatrix} \right\}_{j \in J}$ is linearly independent, then it

follows that $\left\{ \nabla g_j(\tilde{x}^*) \right\}_{j \in J_0(\tilde{x}^*)}$ is linearly independent.

If $\left\{ \begin{bmatrix} \nabla g_j(\tilde{x}^*) \\ -2g_j(\tilde{x}^*)^{\frac{1}{2}} \tilde{c}_j \end{bmatrix} \right\}_{j \in J}$ is linearly dependent, then there

exists $\tilde{\lambda} \neq 0$ such that

$$\sum_{j \in J} \tilde{\lambda}_j \begin{bmatrix} \nabla g_j(\tilde{x}^*) \\ -2g_j(\tilde{x}^*)^{\frac{1}{2}} \tilde{c}_j \end{bmatrix} = 0. \quad (2.2-47)$$

But, since $g_j(\tilde{x}^*) = 0, \forall j \in J_0(\tilde{x}^*)$, and $g_j(\tilde{x}^*) > 0$ otherwise, it follows that $\tilde{\lambda}_j = 0, \forall j \notin J_0(\tilde{x}^*)$, and hence

$$\sum_{j \in J_0(\tilde{x}^*)} \tilde{\lambda}_j \nabla g_j(\tilde{x}^*) = 0. \quad \square$$

Theorem 2.2.5

Let $\tilde{x}^* \in S$ be a local solution of the MPP, and assume that the second order sufficiency condition holds at $\tilde{x}^*$, then the second order sufficiency condition (using the same multipliers) holds at the corresponding solution, $(\tilde{x}^*, g_1(\tilde{x}^*)^{\frac{1}{2}}, \dots, g_m(\tilde{x}^*)^{\frac{1}{2}})$, of the extended $\overline{\text{MPP}}$ if and

only if strict complementarity is satisfied for the original MPP.

Proof

Without loss of generality, we assume that only inequality constraints are present in the original MPP. The second derivative matrix of the Lagrangian for the extended $\overline{\text{MPP}}$ has the form:

$$\nabla_{(x,z)}^2 L_0((x^*, z^*), u^*) = \begin{bmatrix} \nabla_x^2 L_0(x^*, u^*) & 0 \\ \hline 0 & \begin{matrix} 2u_1^* & & \\ & \ddots & \\ & & 2u_m^* \end{matrix} \end{bmatrix}, \quad (2.2-48)$$

where $z^* = (g_1(x^*)^{\frac{1}{2}}, \dots, g_m(x^*)^{\frac{1}{2}})$, and $\nabla_x^2 L_0(x^*, u^*)$ is the corresponding matrix for the original MPP. The constraint gradients at (x^*, z^*) for the extended $\overline{\text{MPP}}$ have the form:

$$\nabla_{(x,z)} G_j(x^*, z^*)^T = [\nabla g_j(x^*)^T, 0, \dots, 0, -2z_j^*, 0, \dots, 0], \quad \forall j \in J. \quad (2.2-49)$$

Sufficiency: Assume that $u_j^* > 0, \forall j \in J_0(x^*)$, then for $\underline{s}$ as an n -vector and $\underline{w}$ an m -vector, consider $\underline{t} = (\underline{s}, \underline{w})$ such that

$$\nabla_{(x,z)} G_j(x^*, z^*)^T \underline{t} = \nabla g_j(x^*)^T \underline{s} - 2g_j(x^*)^{\frac{1}{2}} w_j = 0, \quad \forall j \in J. \quad (2.2-50)$$

Eqn. (2.2-50) implies that

$$\nabla g_j(x^*)^T \underline{s} = 0, \quad \forall j \in J_0(x^*), \quad (2.2-51)$$

with

$$w_j = \frac{\nabla g_j(x^*)^T \underline{s}}{2g_j(x^*)^{\frac{1}{2}}}, \quad \forall j \notin J_0(x^*). \quad (2.2-52)$$

Since the second order sufficiency condition is satisfied by the original MPP, there exists $a > 0$ such that

$$\begin{aligned} \tilde{t}^T \nabla_{(x,z)}^2 L_0((\tilde{x}^*, \tilde{z}^*), \tilde{u}^*) \tilde{t} &= \tilde{s}^T \nabla_x^2 L_0(\tilde{x}^*, \tilde{u}^*) \tilde{s} + 2 \sum_{j \in J} u_j^* (w_j)^2 \\ &\geq a \|\tilde{s}\|^2 + 2 \sum_{j \in J_0(\tilde{x}^*)} u_j^* (w_j)^2. \end{aligned} \quad (2.2-53)$$

Let $\gamma = 1 + \sum_{j \notin J_0(\tilde{x}^*)} \frac{\|\nabla g_j(\tilde{x}^*)\|^2}{|2g_j(\tilde{x}^*)|^{1/2}}^2$, then (2.2-52) implies that

$$\gamma \|\tilde{s}\|^2 + \sum_{j \in J_0(\tilde{x}^*)} w_j^2 \geq \|\tilde{t}\|^2, \quad (2.2-54)$$

and hence (2.2-53) becomes

$$\tilde{t}^T \nabla_{(x,z)}^2 L_0((\tilde{x}^*, \tilde{z}^*), \tilde{u}^*) \tilde{t} \geq \min \left(\frac{a}{\gamma}, 2u_j^* |j \in J_0(\tilde{x}^*)| \right) \|\tilde{t}\|^2. \quad (2.2-55)$$

Necessity: Assume that $\exists i \in J_0(\tilde{x}^*)$ such that $u_i^* = 0$, then for $\tilde{s}^0 = 0$ and $\tilde{w}^0 = (0, \dots, 0, 1, 0, \dots, 0)$ (with 1 in the i th position), it follows that for $\tilde{t}^0 = (\tilde{s}^0, \tilde{w}^0)$, $\|\tilde{t}^0\| \neq 0$,

$$\nabla_{(x,z)} G_j(\tilde{x}^*, \tilde{z}^*)^T \tilde{t}^0 = 0, \quad \forall j \in J,$$

and

$$\tilde{t}^{0T} \nabla_{(x,z)}^2 L_0((\tilde{x}^*, \tilde{z}^*), \tilde{u}^*) \tilde{t}^0 = 0. \quad \square$$

Remark 2.2.5

(a) Theorem 2.2.5 indicates that even if the second order sufficiency condition is satisfied by the original MPP for a family of multipliers,

the second order sufficiency condition may or may not be satisfied by the extended $\overline{\text{MPP}}$, depending on the particular choice of multipliers.

In Ex. 2.2.1, the second order sufficiency condition is satisfied by the original MPP for the following choices of multipliers $0 \leq \alpha < 1$. Since degeneracy occurs when $\alpha = 0$, the second order sufficiency condition is only satisfied by the extended $\overline{\text{MPP}}$ when $0 < \alpha < 1$.

(b) By construction, there is a one-to-one correspondence not only between the solutions but also between every point in the feasible sets of the original MPP and the extended $\overline{\text{MPP}}$: the *natural* correspondence is $\tilde{x} \leftrightarrow (\tilde{x}, g_1(\tilde{x})^{\frac{1}{2}}, \dots, g_m(\tilde{x})^{\frac{1}{2}})$. Since every sequence of points in one set has a corresponding sequence in the other set, it can be shown (proof omitted) that there is a one-to-one (natural) correspondence between elements of the tangent cone at $\tilde{x}^*$ of the original MPP and elements of the tangent cone at $(\tilde{x}^*, \tilde{z}^*)$ of the extended $\overline{\text{MPP}}$.

(c) The C1 and C2 constraint qualifications need not be satisfied by the extended $\overline{\text{MPP}}$ even if they are satisfied by the original MPP. In Ex. 2.2.1, it follows from the linearity of the problem functions that the C1 and C2 constraint qualifications are satisfied by the original MPP. For the extended $\overline{\text{MPP}}$, consider $t = (0, 0, 0, 1, 0)$, it is clear that $t \in C((\tilde{x}^*, \tilde{z}^*))$. However, a feasible trajectory $(\tilde{x}(\theta), \tilde{z}(\theta))$ such that $d(\tilde{x}(\theta), \tilde{z}(\theta))/d\theta = t$ cannot be found.

(d) One consequence of the above discussion is that the C2 constraint qualification does not have to hold for a solution of the MPP to be isolated (the point made in Remark 2.2.3a). From the structure of the extended problem of Ex. 2.2.1, it can be shown that $(\tilde{x}^*, \tilde{z}^*) = (0, 0, 0, 0, 0)$ is the unique solution of 2nd order. As mentioned in Remark 2.2.5a, the second order sufficiency condition is not satisfied by the (extended) problem for

all choices of α . However, even if the C2 constraint qualification is not satisfied (Remark 2.2.5c), Condition 2.2.1 is satisfied by the (extended) problem for all choices of multipliers. This follows from the fact that, for every vector $\tilde{t}$ which belongs to the tangent cone at $(\tilde{x}^*, \tilde{z}^*)$ of the extended problem, there is a natural corresponding vector in the tangent cone at $\tilde{x}^*$ of the original problem. For this vector, a corresponding, twice differentiable and feasible trajectory exists, because the C2 constraint qualification is satisfied by the original problem. This trajectory can be extended (naturally) to be a twice differentiable trajectory in the feasible set of the extended problem with $d(\tilde{x}(0), \tilde{z}(0))/d\theta = \tilde{t}$. The proof of Theorem 2.2.3 and Remark 2.2.3c therefore apply.

Theorem 2.2.6

If $\tilde{x}^* \in S$ is a regular local solution of the MPP, then $(\tilde{x}^*, \tilde{z}^*)$ is a regular local solution of the extended $\overline{\text{MPP}}$ if and only if strict complementarity is satisfied by the original MPP.

Proof

The result follows from Lemma 2.2.6 and Theorem 2.2.5. □

Theorems 2.2.5 and 2.2.6 give some insight into how degeneracy might effect the performance of some constrained minimization algorithms. For many algorithms which were originally developed for equality constrained problems, their properties depend strongly on the conditions which guarantee a regular local solution. When solving a problem with inequality constraints, because these algorithms implicitly solve the extended problem, Theorem 2.2.6 implies that strict complementarity *might* be a necessary condition.

2.3 SENSITIVITY ANALYSIS

2.3.1 Introduction and Classical Results

Sensitivity (or stability) analysis is one of the most important aspects of numerical analysis. For a computed solution to be meaningful, it has to be stable with respect to perturbations. Perturbations can arise either from the accumulation of truncation (rounding) errors or from perturbations of the problem data itself. By using backward error analysis, the effect of truncation errors can be reinterpreted as due to perturbations in the problem data. In this section, backward error analysis for specific algorithms is not discussed. Attention is restricted to perturbations in the problem data. For some non-linear programming problems, the computed solution is of less interest than the sensitivity of the optimal solution with respect to small perturbations.

Consider the following perturbation, $P[\underline{\epsilon}]$, of the original MPP:

$$\begin{aligned} & \underset{x}{\text{minimize}} && f(\underline{x}, \underline{\epsilon}) \\ & \text{subject to} && h_i(\underline{x}, \underline{\epsilon}) = 0, \quad i \in I, \\ & && \text{and} \quad g_j(\underline{x}, \underline{\epsilon}) \geq 0, \quad j \in J, \end{aligned} \tag{2.3-1}$$

where $\underline{\epsilon} \in E^q$ for some integer q , and $P[0]$ is the original MPP. A perturbation $\underline{\epsilon}$ can be interpreted in two ways: as a *random* error, and as a *specific* change in the parameters defining the problem functions.

In a classical sensitivity analysis, the solution $\underline{x}^*$ is usually assumed to satisfy the Jacobian uniqueness condition. The theorem and its proof can be found in most advanced texts on the subject (e.g., Fiacco and McCormick (1968) and Luenberger (1972)).

Theorem 2.3.1 (Fiacco (1976), Theorem 2.1)

At a local solution $\tilde{x}^*$ of the MPP, $P[0]$, assume that (i) the linear independence condition, (ii) the second order sufficiency condition and (iii) strict complementarity are satisfied, and that the functions defining $P[\tilde{\epsilon}]$ are twice continuously differentiable with respect to $(\tilde{x}, \tilde{\epsilon})$ in a neighbourhood of $(\tilde{x}^*, 0)$, then

- (a) $\tilde{x}^*$ is a local isolated minimizing point of $P[0]$, and the associated Lagrange multipliers $\tilde{v}^*$ and $\tilde{u}^*$ are unique;
- (b) for $\tilde{\epsilon}$ in a neighbourhood of 0, there exists a unique once continuously differentiable vector function $(\tilde{x}(\tilde{\epsilon}), \tilde{v}(\tilde{\epsilon}), \tilde{u}(\tilde{\epsilon}))$ satisfying the second order sufficiency condition for a local minimum of the problem $P[\tilde{\epsilon}]$ such that $(\tilde{x}(0), \tilde{v}(0), \tilde{u}(0)) = (\tilde{x}^*, \tilde{v}^*, \tilde{u}^*)$ and, hence, $\tilde{x}(\tilde{\epsilon})$ is a locally unique local minimum of $P[\tilde{\epsilon}]$ with associated unique Lagrange multipliers $\tilde{v}(\tilde{\epsilon})$ and $\tilde{u}(\tilde{\epsilon})$; and
- (c) strict complementarity (with respect to $\tilde{u}(\tilde{\epsilon})$ and the inequality constraints) and linear independence of the binding constraint gradients hold at $\tilde{x}(\tilde{\epsilon})$ for $\tilde{\epsilon}$ near 0. □

Notation: For differentiation with respect to $\tilde{x}$, the usual notation $\nabla_{\tilde{x}} = \left(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n} \right)$ is adopted. For differentiation with respect to $\tilde{\epsilon}$, in order to avoid ambiguities, $\frac{\partial}{\partial \tilde{\epsilon}}$ will be used for the usual partial differentiation with respect to $\tilde{\epsilon}$ when $\tilde{x}$ is treated as an independent variable, and $\frac{d}{d\tilde{\epsilon}}$ will be used to denote differentiation with respect to $\tilde{\epsilon}$ when any function of $\tilde{x}$ and $\tilde{\epsilon}$ is treated as a function only of $\tilde{\epsilon}$, implicitly defined by setting $\tilde{x} = \tilde{x}(\tilde{\epsilon})$.

Corollary 2.3.1 (Sensitivity of $\tilde{x}(\tilde{\epsilon}), \tilde{v}(\tilde{\epsilon}), \tilde{u}(\tilde{\epsilon})$ and $f(\tilde{x}(\tilde{\epsilon}), \tilde{\epsilon})$ at $\tilde{\epsilon}=0$)

Under the assumptions of Theorem 2.3.1, it follows that, at $\tilde{\epsilon}=0$,

$$\frac{d}{d\tilde{\epsilon}} \begin{bmatrix} \tilde{x}(0) \\ \tilde{v}(0) \\ \tilde{u}(0) \end{bmatrix} = J(0)^{-1} V(0) , \quad (2.3-2)$$

and

$$\frac{d}{d\tilde{\epsilon}} f(\tilde{x}(0), 0) = \frac{\partial f(\tilde{x}(0), 0)}{\partial \tilde{\epsilon}} - \sum_{i \in I} v_i^* \frac{\partial h_i(\tilde{x}(0), 0)}{\partial \tilde{\epsilon}} - \sum_{i \in J} u_i^* \frac{\partial g_i(\tilde{x}(0), 0)}{\partial \tilde{\epsilon}} , \quad (2.3-3)$$

where

$$J = \begin{bmatrix} \nabla_x^2 L_0 & [-\nabla h_i, i \in I] & [-\nabla g_i, i \in J] \\ [\nabla h_i^T, i \in I] & 0 & 0 \\ [u_i \nabla g_i^T, i \in J] & 0 & [\text{diag } g_i, i \in J] \end{bmatrix} , \quad (2.3-4)$$

and

$$V = \left[-\frac{\partial}{\partial \tilde{\epsilon}} (\nabla_x L_0^T), -\frac{\partial h_1}{\partial \tilde{\epsilon}}, \dots, -\frac{\partial h_\ell}{\partial \tilde{\epsilon}}, -u_1 \frac{\partial g_1}{\partial \tilde{\epsilon}}, \dots, -u_m \frac{\partial g_m}{\partial \tilde{\epsilon}} \right]^T . \quad (2.3-5)$$

Proof

The first part (2.3-2) is just Corollary 2.1 of Fiacco (1976).

The proof for the second part (2.3-3) is as follows: Using the chain rule,

$$\begin{aligned} \frac{df}{d\tilde{\epsilon}} &= \frac{\partial f}{\partial \tilde{\epsilon}} + \frac{dx}{d\tilde{\epsilon}}^T \cdot \nabla_x f \\ &= \frac{\partial f}{\partial \tilde{\epsilon}} + \frac{dx}{d\tilde{\epsilon}}^T \cdot \left(\sum_{i \in I} v_i^* \nabla_x h_i + \sum_{i \in J} u_i^* \nabla_x g_i \right) . \end{aligned} \quad (2.3-6)$$

Because the problem is non-degenerate, every active constraint remains active for $\tilde{\epsilon}$ sufficiently small. Hence, using the chain rule,

$$\frac{dh_i}{d\tilde{\epsilon}} = \frac{\partial h_i}{\partial \tilde{\epsilon}} + \frac{dx}{d\tilde{\epsilon}}^T \cdot \nabla_x h_i = 0 , \quad i \in I , \quad (2.3-7)$$

and

$$\frac{dg_i}{d\tilde{\epsilon}} = \frac{\partial g_i}{\partial \tilde{\epsilon}} + \frac{dx^T}{d\tilde{\epsilon}} \cdot \nabla_x g_i = 0, \quad i \in J_0(\tilde{x}^*) \quad (2.3-8)$$

Since $u_i^* = 0, \forall i \notin J_0(\tilde{x}^*)$, the result in (2.3-3) follows from substituting (2.3-7) and (2.3-8) into (2.3-6). $\square$

2.3.2 Sensitivity Analysis without Strict Complementarity[†]

In this subsection, various classical results in subsection 2.3.1 are strengthened by removing the strict complementarity requirement, replacing the Jacobian uniqueness condition by the regular solution condition. The stronger results (or weaker conditions) are achieved by utilizing properties of convex functionals (Appendix 1) and a stronger version of the implicit function theorem (Appendix 2). Results in this subsection are stated for a general MPP, but without loss of generality and in the interest of simplicity, proofs will only be presented for a degenerate ICP.

Consider an ICP. Assume that $\tilde{x}^*$ (together with $\tilde{u}^*$) is a regular local solution and that the situation is degenerate. For the perturbed problem $P[\tilde{\epsilon}]$, and a constant $\sigma > 0$, define

$$\bar{L}_\sigma(\tilde{x}, \tilde{u}, \tilde{\epsilon}) = f(\tilde{x}) + \sum_{i \in I} \begin{cases} -u_i g_i(\tilde{x}, \tilde{\epsilon}) + \frac{\sigma}{2} g_i(\tilde{x}, \tilde{\epsilon})^2 & \text{if } g_i(\tilde{x}, \tilde{\epsilon}) < u_i/\sigma \\ -\frac{1}{2} u_i^2/\sigma & \text{if } g_i(\tilde{x}, \tilde{\epsilon}) \geq u_i/\sigma, \end{cases} \quad (2.3-9)$$

[†] Proofs of results in this subsection refer to techniques which are developed in subsection 4.2.3. Since results in this subsection will not be required elsewhere in this thesis, and that proofs of these results rely on results and techniques of subsection 4.2.3, it is recommended that this subsection or the proofs of various results should be skipped until subsection 4.2.3 has been covered.

$$M(\underline{x}, \underline{u}, \underline{\epsilon}, \sigma) = \{j | j \in J, g_j(\underline{x}, \underline{\epsilon}) < u_j/\sigma\} , \quad (2.3-10)$$

and

$$Z(\underline{x}, \underline{u}, \underline{\epsilon}, \sigma) = \{j | j \in J, g_j(\underline{x}, \underline{\epsilon}) = u_j/\sigma\} . \quad (2.3-11)$$

Furthermore, if $\bar{L}_\sigma(\underline{x}, \underline{u}, \underline{\epsilon})$ is convex in $\underline{x}$, define

$$\bar{D}_\sigma(\underline{u}, \underline{\epsilon}) = \min_{\underline{x}} \bar{L}_\sigma(\underline{x}, \underline{u}, \underline{\epsilon}) . \quad (2.3-12)$$

Now, for any index set R such that

$$J_1(\underline{x}^*) \subseteq R \subseteq J_1(\underline{x}^*) \cup J_2(\underline{x}^*) , \quad (2.3-13)$$

let $P[R, \underline{\epsilon}]$ denote the following (equality constrained) problem:

$$\begin{aligned} & \underset{\underline{x}}{\text{minimize}} && f(\underline{x}, \underline{\epsilon}) \\ & \text{subject to} && g_i(\underline{x}, \underline{\epsilon}) = 0 , \quad i \in R . \end{aligned} \quad (2.3-14)$$

Similar to (2.3-9), for the problem $P[R, \underline{\epsilon}]$, define

$$L_\sigma^R(\underline{x}, \underline{u}, \underline{\epsilon}) = f(\underline{x}) - \sum_{i \in R} u_i g_i(\underline{x}, \underline{\epsilon}) + \frac{\sigma}{2} \sum_{i \in R} g_i(\underline{x}, \underline{\epsilon})^2 . \quad (2.3-15)$$

If $L_\sigma^R(\underline{x}, \underline{u}, \underline{\epsilon})$ is convex in $\underline{x}$, define

$$D_\sigma^R(\underline{u}, \underline{\epsilon}) = \min_{\underline{x}} L_\sigma^R(\underline{x}, \underline{u}, \underline{\epsilon}) , \quad (2.3-16)$$

and

$$U_\sigma^R(\underline{u}, \underline{\epsilon}) = D_\sigma^R(\underline{u}, \underline{\epsilon}) - \frac{1}{2\sigma} \sum_{i \notin R} u_i^2 . \quad (2.3-17)$$

The main Theorem of this subsection can now be stated and proved:

Theorem 2.3.2

Assume that $\underline{x}^*$ (together with $\underline{u}^*$) is a regular local solution of a degenerate ICP, $P[0]$, and that the functions defining $P[\underline{\epsilon}]$ and their first and second derivatives with respect to $\underline{x}$ are continuous with respect to $\underline{\epsilon}$, then there exists a unique *continuous* vector function $(\underline{x}(\underline{\epsilon}), \underline{u}(\underline{\epsilon}))$ such that $\underline{x}(\underline{\epsilon})$ (together with $\underline{u}(\underline{\epsilon})$) is a regular local solution of the problem $P[\underline{\epsilon}]$, with $(\underline{x}(0), \underline{u}(0)) = (\underline{x}^*, \underline{u}^*)$.

Proof

The proof of this Theorem will rely heavily on results and techniques developed in subsection 4.2.3. The parameter σ is assumed to be fixed and sufficiently large.

(i) *Similar to Lemma 4.2.1:* Because the functions defining $P[\underline{\epsilon}]$ are continuous in both $\underline{x}$ and $\underline{\epsilon}$, there exists $\delta > 0$ such that $\forall \underline{x}' \in N(\underline{x}^*, \delta)$, $\forall \underline{u}' \in N(\underline{u}^*, \delta)$ and $\forall \underline{\epsilon}' \in N(0, \delta)$,

$$J_1(\underline{x}^*) \subseteq M(\underline{x}', \underline{u}', \underline{\epsilon}', \sigma) \subseteq M(\underline{x}', \underline{u}', \underline{\epsilon}', \sigma) \cup Z(\underline{x}', \underline{u}', \underline{\epsilon}', \sigma) \subseteq J_1(\underline{x}^*) \cup J_2(\underline{x}^*) \quad (2.3-18)$$

(ii) *Similar to Lemma 4.2.2:* At a given point $(\underline{x}', \underline{u}', \underline{\epsilon}')$, because the functions defining $P[\underline{\epsilon}]$ are continuous in both $\underline{x}$ and $\underline{\epsilon}$, there exists $\delta' > 0$ such that $\forall \underline{x} \in N(\underline{x}', \delta')$, $\forall \underline{u} \in N(\underline{u}', \delta')$ and $\forall \underline{\epsilon} \in N(\underline{\epsilon}', \delta')$,

$$M(\underline{x}', \underline{u}', \underline{\epsilon}', \sigma) \subseteq M(\underline{x}, \underline{u}, \underline{\epsilon}, \sigma) \subseteq M(\underline{x}, \underline{u}, \underline{\epsilon}, \sigma) \cup Z(\underline{x}, \underline{u}, \underline{\epsilon}, \sigma) \subseteq M(\underline{x}', \underline{u}', \underline{\epsilon}', \sigma) \cup Z(\underline{x}', \underline{u}', \underline{\epsilon}', \sigma) \quad (2.3-19)$$

(iii) *Similar to Lemma 4.2.3:* Assume that $\underline{u}', \underline{\epsilon}'$ and the constant $\sigma > 0$ are given. If there exists $\delta > 0$ such that (2.3-18) is satisfied at $(\underline{x}, \underline{u}', \underline{\epsilon}')$ for all $\underline{x} \in N(\underline{x}^*, \delta)$ and that, for every index set R satisfying (2.3-13), $\nabla_{\underline{x}}^2 L_{\sigma}^R(\underline{x}, \underline{u}', \underline{\epsilon}')$ is positive definite for all

$\tilde{x} \in N(\tilde{x}^*, \delta)$, then $\bar{L}_\sigma(\tilde{x}, \tilde{u}', \tilde{\varepsilon}')$ is strictly convex on $N(\tilde{x}^*, \delta)$. This result follows because (i) and (ii) above imply that, for any point $\tilde{x}' \in N(\tilde{x}^*, \delta)$ and a direction $\tilde{t}$ ($\|\tilde{t}\| = 1$), if α is sufficiently small, then $\nabla_{\tilde{x}} \bar{L}_\sigma(\tilde{x}' + \alpha \tilde{t}, \tilde{u}', \tilde{\varepsilon}') - \nabla_{\tilde{x}} \bar{L}_\sigma(\tilde{x}', \tilde{u}', \tilde{\varepsilon}')$ can be identified with $\nabla_{\tilde{x}} L_\sigma^R(\tilde{x}' + \alpha \tilde{t}, \tilde{u}', \tilde{\varepsilon}') - \nabla_{\tilde{x}} L_\sigma^R(\tilde{x}', \tilde{u}', \tilde{\varepsilon}')$ for some index set R satisfying (2.3-13). Since, for every index set R satisfying (2.3-13), $\nabla_{\tilde{x}}^2 L_\sigma^R$ is positive definite, it follows that $\bar{L}_\sigma$ is strictly convex in $\tilde{x}$ (on $N(\tilde{x}^*, \delta)$).

(iv) *Similar to Theorem 4.2.4:* Since $\tilde{x}^*$ is a regular local solution and σ is sufficiently large, it follows from Theorems 2.2.1 and 4.2.1 that $\nabla_{\tilde{x}}^2 L_\sigma^R(\tilde{x}^*, \tilde{u}^*, 0)$ is positive definite for every index set R satisfying (2.3-13). Because $\nabla_{\tilde{x}}^2 L_\sigma^R(\tilde{x}, \tilde{u}, \tilde{\varepsilon})$ is continuous with respect to $\tilde{x}$, $\tilde{u}$ and $\tilde{\varepsilon}$, there exists $\delta > 0$ such that $\forall \tilde{x} \in N(\tilde{x}^*, \delta)$, $\forall \tilde{u} \in N(\tilde{u}^*, \delta)$ and $\forall \tilde{\varepsilon} \in N(0, \delta)$, $\nabla_{\tilde{x}}^2 L_\sigma^R(\tilde{x}, \tilde{u}, \tilde{\varepsilon})$ is positive definite. From (i), (ii) and (iii), using properties of convex functionals (Appendix 1) and a stronger version of the implicit function theorem (Appendix 2), it follows that there exists $\tilde{x}(\tilde{u}, \tilde{\varepsilon}, \sigma)$ which is a continuous function of $(\tilde{u}, \tilde{\varepsilon})$ such that, for every $\tilde{u}$ in a neighbourhood of $\tilde{u}^*$ and for every $\tilde{\varepsilon}$ in a neighbourhood of 0, $\tilde{x}(\tilde{u}, \tilde{\varepsilon}, \sigma)$ is the unique (local) minimizer of $\bar{L}_\sigma(\tilde{x}, \tilde{u}, \tilde{\varepsilon})$, with $\tilde{x}(\tilde{u}^*, 0, \sigma) = \tilde{x}^*$.

Locally, in a neighbourhood of $(\tilde{x}^*, \tilde{u}^*, 0)$, because $L_\sigma^R(\tilde{x}, \tilde{u}, \tilde{\varepsilon})$, for R satisfying (2.3-13), and $\bar{L}_\sigma(\tilde{x}, \tilde{u}, \tilde{\varepsilon})$ are strictly convex in $\tilde{x}$, it follows that $U_\sigma^R(\tilde{u}, \tilde{\varepsilon})$ and $\bar{D}_\sigma(\tilde{u}, \tilde{\varepsilon})$ are well defined (i.e., $\bar{D}_\sigma(\tilde{u}, \tilde{\varepsilon}) = \bar{L}_\sigma(\tilde{x}(\tilde{u}, \tilde{\varepsilon}, \sigma), \tilde{u}, \tilde{\varepsilon})$). By using $\bar{D}_\sigma(\tilde{u}, \tilde{\varepsilon})$ and $\nabla_{\tilde{u}}^2 U_\sigma^R(\tilde{u}, \tilde{\varepsilon})$ instead of $\bar{L}_\sigma(\tilde{x}, \tilde{u}, \tilde{\varepsilon})$ and $\nabla_{\tilde{x}}^2 L_\sigma^R(\tilde{x}, \tilde{u}, \tilde{\varepsilon})$, respectively, the arguments (iii) and (iv) can now be repeated. Since $\tilde{x}^*$ is a regular local solution, it follows from (2.3-17), Theorems 2.2.1 and 4.2.2 that, for every index set R satisfying (2.3-13), $\nabla_{\tilde{u}}^2 U_\sigma^R(\tilde{u}^*, 0)$ is negative definite. Because $\nabla_{\tilde{u}}^2 U_\sigma^R(\tilde{u}, \tilde{\varepsilon})$ is continuous in both $\tilde{u}$ and $\tilde{\varepsilon}$, $\nabla_{\tilde{u}}^2 U_\sigma^R(\tilde{u}, \tilde{\varepsilon})$ is negative definite for $\tilde{u}$ close to $\tilde{u}^*$ and $\tilde{\varepsilon}$ close to 0. Similar to (iii), if $\tilde{u}'$ close to $\tilde{u}^*$, $\tilde{\varepsilon}'$ close to 0

and a direction $\underline{t}$ ($\|\underline{t}\| = 1$) are given, then, for all α sufficiently small, $\nabla_{\underline{u}} \bar{D}_{\sigma}(\underline{u}' + \alpha \underline{t}, \underline{\varepsilon}') - \nabla_{\underline{u}} \bar{D}_{\sigma}(\underline{u}', \underline{\varepsilon}')$ can be identified with $\nabla_{\underline{u}} U_{\sigma}^R(\underline{u}' + \alpha \underline{t}, \underline{\varepsilon}') - \nabla_{\underline{u}} U_{\sigma}^R(\underline{u}', \underline{\varepsilon}')$ for some index set R satisfying (2.3-13). Hence, for every $\underline{\varepsilon}$ sufficiently close to 0, $\bar{D}_{\sigma}(\underline{u}, \underline{\varepsilon})$ is strictly concave on a neighbourhood of $\underline{u}^*$. It then follows from properties of concave functionals (Appendix 1) and the stronger version of the implicit function theorem that there exists $\underline{u}(\underline{\varepsilon}, \sigma)$ which is a continuous function of $\underline{\varepsilon}$ such that, for every $\underline{\varepsilon}$ in a small neighbourhood of 0, $\underline{u}(\underline{\varepsilon}, \sigma)$ is the unique local maximizer of $\bar{D}_{\sigma}(\underline{u}, \underline{\varepsilon})$, with $\underline{u}(0, \sigma) = \underline{u}^*$.

Therefore, for any $\underline{\varepsilon}$ sufficiently small, $\underline{x}(\underline{u}(\underline{\varepsilon}, \sigma), \underline{\varepsilon}, \sigma)$ and $\underline{u}(\underline{\varepsilon}, \sigma)$ are uniquely determined. Let

$$R(\underline{\varepsilon}) = \{j \mid j \in J, u_j(\underline{\varepsilon}, \sigma) > 0\}, \quad (2.3-20)$$

then, from construction, it follows that $(\underline{x}(\underline{u}(\underline{\varepsilon}, \sigma), \underline{\varepsilon}, \sigma), \underline{u}(\underline{\varepsilon}, \sigma))$ is the saddle point of $L_{\sigma}^{R(\underline{\varepsilon})}(\underline{x}, \underline{u}, \underline{\varepsilon})$ (only those components $i \in R(\underline{\varepsilon})$ of $\underline{u}(\underline{\varepsilon}, \sigma)$ are needed). This saddle point condition implies that $(\underline{x}(\underline{u}(\underline{\varepsilon}, \sigma), \underline{\varepsilon}, \sigma), \underline{u}(\underline{\varepsilon}, \sigma))$ solves the problem $P[R(\underline{\varepsilon}), \underline{\varepsilon}]$ (Arrow *et al.* (1973)). It is clear that the solution of $P[R(\underline{\varepsilon}), \underline{\varepsilon}]$ is independent of σ . Without any ambiguities, $(\underline{x}(\underline{u}(\underline{\varepsilon}, \sigma), \underline{\varepsilon}, \sigma), \underline{u}(\underline{\varepsilon}, \sigma))$ can now be written as $(\underline{x}(\underline{\varepsilon}), \underline{u}(\underline{\varepsilon}))$. From the above arguments, it follows that $(\underline{x}(\underline{\varepsilon}), \underline{u}(\underline{\varepsilon}))$ is a continuous function of $\underline{\varepsilon}$ for $\underline{\varepsilon}$ in a small neighbourhood of 0, with $(\underline{x}(0), \underline{u}(0)) = (\underline{x}^*, \underline{u}^*)$. Hence, it is only left to be shown that $\underline{x}(\underline{\varepsilon})$ (together with $\underline{u}(\underline{\varepsilon})$) is a regular local solution of the problem $P[\underline{\varepsilon}]$.

Second order sufficiency condition: This is a direct consequence of the existence of $(\underline{x}(\underline{\varepsilon}), \underline{u}(\underline{\varepsilon}))$ as a continuous function of $\underline{\varepsilon}$. Let $\underline{\varepsilon}$ be in a small neighbourhood of 0. Because $\nabla_{\underline{x}}^2 L_{\sigma}^{R(\underline{\varepsilon})}(\underline{x}^*, \underline{u}^*, 0)$ is positive definite and that $\nabla_{\underline{x}}^2 L_{\sigma}^{R(\underline{\varepsilon})}(\underline{x}, \underline{u}, \underline{\varepsilon})$ is continuous with respect to $\underline{x}$, $\underline{u}$ and $\underline{\varepsilon}$, it

follows that, if $\underline{\epsilon}$ is sufficiently small, $\nabla_{x_0}^2 L^R(\underline{\epsilon})(\underline{x}(\underline{\epsilon}), \underline{u}(\underline{\epsilon}), \underline{\epsilon})$ is positive definite. Consider

$$\begin{aligned} \nabla_{x_0}^2 L^R(\underline{\epsilon})(\underline{x}(\underline{\epsilon}), \underline{u}(\underline{\epsilon}), \underline{\epsilon}) &= \nabla_{x_0}^2 L^R(\underline{\epsilon})(\underline{x}(\underline{\epsilon}), \underline{u}(\underline{\epsilon}), \underline{\epsilon}) \\ &+ \frac{\sigma}{2} \sum_{i \in R(\underline{\epsilon})} \nabla_x g_i(\underline{x}(\underline{\epsilon}), \underline{\epsilon}) \cdot \nabla_x g_i(\underline{x}(\underline{\epsilon}), \underline{\epsilon})^T. \end{aligned} \quad (2.3-21)$$

It follows from (2.3-21) that, for the problem $P[\underline{\epsilon}]$, $\nabla_{x_0}^2 L(\underline{x}(\underline{\epsilon}), \underline{u}(\underline{\epsilon}), \underline{\epsilon}) = \nabla_{x_0}^2 L^R(\underline{\epsilon})(\underline{x}(\underline{\epsilon}), \underline{u}(\underline{\epsilon}), \underline{\epsilon})$ is positive definite in the orthogonal complement of the subspace spanned by $\{\nabla_x g_i(\underline{x}(\underline{\epsilon}), \underline{\epsilon}) \mid i \in J, u_i(\underline{\epsilon}) > 0\}$.

Linear independence condition: This is also a direct consequence of the existence of $(\underline{x}(\underline{\epsilon}), \underline{u}(\underline{\epsilon}))$ as a continuous function of $\underline{\epsilon}$. Consider a matrix

$$K(\underline{\epsilon}) = [\nabla_x g_i(\underline{x}(\underline{\epsilon}), \underline{\epsilon}), \nabla_x g_j(\underline{x}(\underline{\epsilon}), \underline{\epsilon}), \dots \mid i, j \in J_0(\underline{x}^*)] \quad (2.3-22)$$

$K(\underline{\epsilon})^T K(\underline{\epsilon})$ is positive definite at $\underline{\epsilon} = 0$. Because $K(\underline{\epsilon})^T K(\underline{\epsilon})$ is a continuous function of $\underline{\epsilon}$, its smallest eigenvalue remains positive in a sufficiently small neighbourhood of 0. This means that $\{\nabla_x g_i(\underline{x}(\underline{\epsilon}), \underline{\epsilon})\}_{i \in J_0(\underline{x}^*)}$ remains linearly independent for $\underline{\epsilon}$ sufficiently small. Since this set contains all binding constraints at $\underline{x}(\underline{\epsilon})$ (see (2.3-18)), the result follows. $\square$

Utilizing Theorem 2.3.2, the equivalence of Theorem 2.3.1 and Corollary 2.3.1 without strict complementarity can be stated as follows:

Theorem 2.3.3

At a local solution $\underline{x}^*$ of the MPP, $P[0]$, assume that (i) the linear independence condition and (ii) the second order sufficiency condition are satisfied, and that the functions defining $P[\underline{\epsilon}]$ are twice continuously

differentiable with respect to $(\underline{x}, \underline{\epsilon})$ in a neighbourhood of $(\underline{x}^*, 0)$, then

- (a) $\underline{x}^*$ is a local isolated minimizing point of $P[0]$, and the associated Lagrange multipliers $\underline{v}^*$ and $\underline{u}^*$ are unique;
- (b) for $\underline{\epsilon}$ in a neighbourhood of 0, there exists a unique continuous vector function $(\underline{x}(\underline{\epsilon}), \underline{v}(\underline{\epsilon}), \underline{u}(\underline{\epsilon}))$ satisfying the second order sufficiency condition for a local minimum of the problem $P[\underline{\epsilon}]$ such that $(\underline{x}(0), \underline{v}(0), \underline{u}(0)) = (\underline{x}^*, \underline{v}^*, \underline{u}^*)$ and, hence, $\underline{x}(\underline{\epsilon})$ is a locally unique minimizer of $P[\underline{\epsilon}]$ with associated unique Lagrange multipliers $\underline{v}(\underline{\epsilon})$ and $\underline{u}(\underline{\epsilon})$;
- (c) linear independence of the binding constraint gradients holds at $\underline{x}(\underline{\epsilon})$ for $\underline{\epsilon}$ near 0;
- (d) there exist $0 < \alpha, \beta, \gamma < \infty$ and $\epsilon_0 > 0$ such that, $\forall \underline{\epsilon}$ with $\|\underline{\epsilon}\| < \epsilon_0$,

$$\|\underline{x}(\underline{\epsilon}) - \underline{x}^*\| \leq \alpha \|\underline{\epsilon}\|, \quad (2.3-23)$$

$$\|\underline{v}(\underline{\epsilon}) - \underline{v}^*\| \leq \beta \|\underline{\epsilon}\|, \quad (2.3-24)$$

and

$$\|\underline{u}(\underline{\epsilon}) - \underline{u}^*\| \leq \gamma \|\underline{\epsilon}\|; \quad (2.3-25)$$

and

- (e) $f(\underline{\epsilon}) = f(\underline{x}(\underline{\epsilon}), \underline{\epsilon})$ is differentiable with respect to $\underline{\epsilon}$ at $\underline{\epsilon} = 0$:

$$\frac{d}{d\underline{\epsilon}} f(\underline{x}(0), 0) = \frac{\partial f(\underline{x}(0), 0)}{\partial \underline{\epsilon}} - \sum_{i \in I} v_i^* \frac{\partial h_i(\underline{x}(0), 0)}{\partial \underline{\epsilon}} - \sum_{i \in J} u_i^* \frac{\partial g_i(\underline{x}(0), 0)}{\partial \underline{\epsilon}}. \quad (2.3-26)$$

Proof

If strict complementarity is satisfied, the results follow directly from Theorem 2.3.1 and Corollary 2.3.1. Without loss of generality, it is sufficient to prove this theorem for a degenerate ICP.

Part (a): The result follows from Lemma 2.1.5 and Remark 2.1.5a.

Part (b) and Part (c): Direct consequences of Theorem 2.3.2.

Part (d): Let $\tilde{x}^R(\underline{\varepsilon})$ (together with $\tilde{u}^R(\underline{\varepsilon})$) denote the solution of the problem $P[R, \underline{\varepsilon}]$, for any index set R satisfying (2.3-13). By applying Theorem 2.3.1 and Corollary 2.3.1 to the problem $P[R, \underline{\varepsilon}]$, it follows that, for $\underline{\varepsilon}$ sufficiently small, $\tilde{x}^R(\underline{\varepsilon})$ is a continuously differentiable function of $\underline{\varepsilon}$, and hence there exist $\varepsilon_0^R > 0$ and $\alpha^R > 0$ ($\alpha^R = \|dx^R(0)/d\underline{\varepsilon}\| + o(1)$) such that

$$\|\tilde{x}^R(\underline{\varepsilon}) - x^*\| \leq \alpha^R \|\underline{\varepsilon}\|, \quad \forall \underline{\varepsilon} \in N(0, \varepsilon_0^R). \quad (2.3-27)$$

From (2.3-18), it follows that, for every $\underline{\varepsilon}$ sufficiently small, $\tilde{x}(\underline{\varepsilon})$ can be identified with $\tilde{x}^R(\underline{\varepsilon})$ for some index set R satisfying (2.3-13). From (2.3-27), because there are only a finite number of possible index sets R , it follows that

$$\|\tilde{x}(\underline{\varepsilon}) - x^*\| \leq \alpha \|\underline{\varepsilon}\|, \quad \forall \underline{\varepsilon} \in N(0, \varepsilon_0), \quad (2.3-28)$$

where $\varepsilon_0 = \min_R \varepsilon_0^R$, and $\alpha = \max_R \alpha^R$.

Similar arguments can be applied for the multiplier $\tilde{u}(\underline{\varepsilon})$.

Note that in this case there is a difference between the sizes of J and R . For a given $\underline{\varepsilon}$, $\tilde{u}(\underline{\varepsilon})$ can be identified with $\tilde{u}^{R(\underline{\varepsilon})}(\underline{\varepsilon})$ of the problem $P[R(\underline{\varepsilon}), \underline{\varepsilon}]$. Consider

$$\|\tilde{u}^{R(\underline{\varepsilon})}(\underline{\varepsilon}) - \tilde{u}^{R(\underline{\varepsilon})}(0)\| = \left[\sum_{i \in R(\underline{\varepsilon})} (u_i^{R(\underline{\varepsilon})}(\underline{\varepsilon}) - u_i^{R(\underline{\varepsilon})}(0))^2 \right]^{1/2}, \quad (2.3-29)$$

and

$$\|\tilde{u}(\underline{\varepsilon}) - \tilde{u}(0)\| = \left[\sum_{i \in R(\underline{\varepsilon})} (u_i(\underline{\varepsilon}) - u_i(0))^2 + \sum_{i \notin R(\underline{\varepsilon})} (u_i(\underline{\varepsilon}) - u_i(0))^2 \right]^{1/2}. \quad (2.3-30)$$

Because each term in the second summation on the RHS of (2.3-30) is identically zero, Eqns. (2.3-29) and (2.3-30) are equivalent. Thus, the difference between the sizes of J and R does not present any difficulty.

Part (e): From the differentiability assumption on the problem functions and (2.3-28), it follows that

$$f(\tilde{x}(\tilde{\epsilon}), \tilde{\epsilon}) - f(\tilde{x}^*, 0) = \frac{\partial f(\tilde{x}^*, 0)^T}{\partial \tilde{\epsilon}} \cdot \tilde{\epsilon} + \nabla_x f(\tilde{x}^*, 0)^T \cdot (\tilde{x}(\tilde{\epsilon}) - \tilde{x}^*) + o(\max\{\|\tilde{\epsilon}\|, \|\tilde{x}(\tilde{\epsilon}) - \tilde{x}^*\|\}) \quad (2.3-31)$$

$$= \frac{\partial f(\tilde{x}^*, 0)^T}{\partial \tilde{\epsilon}} \cdot \tilde{\epsilon} + \sum_{i \in J} u_i^* \nabla_x g_i(\tilde{x}^*, 0)^T \cdot (\tilde{x}(\tilde{\epsilon}) - \tilde{x}^*) + o(\|\tilde{\epsilon}\|). \quad (2.3-32)$$

Consider the summation on the RHS of (2.3-32). For terms with $u_i^* = 0$, it is trivially true that

$$u_i^* \nabla_x g_i(\tilde{x}^*, 0)^T \cdot (\tilde{x}(\tilde{\epsilon}) - \tilde{x}^*) = -u_i^* \frac{\partial g_i(\tilde{x}^*, 0)^T}{\partial \tilde{\epsilon}} \cdot \tilde{\epsilon} \quad (2.3-33)$$

For terms with $u_i^* > 0$, because $u_i^* > 0$, this means that the corresponding constraints will remain active for $\tilde{\epsilon}$ sufficiently small, hence

$$0 = g_i(\tilde{x}(\tilde{\epsilon}), \tilde{\epsilon}) - g_i(\tilde{x}^*, 0) = \frac{\partial g_i(\tilde{x}^*, 0)^T}{\partial \tilde{\epsilon}} \cdot \tilde{\epsilon} + \nabla_x g_i(\tilde{x}^*, 0)^T \cdot (\tilde{x}(\tilde{\epsilon}) - \tilde{x}^*) + o(\max\{\|\tilde{\epsilon}\|, \|\tilde{x}(\tilde{\epsilon}) - \tilde{x}^*\|\}) \quad (2.3-34)$$

Therefore, it follows from (2.3-28) that, $\forall i \in J$ such that $u_i^* > 0$,

$$u_i^* \nabla_x g_i(\tilde{x}^*, 0)^T \cdot (\tilde{x}(\tilde{\epsilon}) - \tilde{x}^*) = -u_i^* \frac{\partial g_i(\tilde{x}^*, 0)^T}{\partial \tilde{\epsilon}} \cdot \tilde{\epsilon} + o(\|\tilde{\epsilon}\|). \quad (2.3-35)$$

Using (2.3-33) and (2.3-35), Eqn. (2.3-32) becomes

$$f(\tilde{x}(\tilde{\epsilon}), \tilde{\epsilon}) - f(\tilde{x}^*, 0) = \frac{\partial f(\tilde{x}^*, 0)^T}{\partial \tilde{\epsilon}} \cdot \tilde{\epsilon} - \sum_{i \in J} u_i^* \frac{\partial g_i(\tilde{x}^*, 0)^T}{\partial \tilde{\epsilon}} \cdot \tilde{\epsilon} + o(\|\tilde{\epsilon}\|) . \quad (2.3-36)$$

Thus

$$\frac{df(\tilde{x}(0), 0)}{d\tilde{\epsilon}} = \frac{\partial f(\tilde{x}(0), 0)}{\partial \tilde{\epsilon}} - \sum_{i \in J} u_i^* \frac{\partial g_i(\tilde{x}(0), 0)}{\partial \tilde{\epsilon}} . \quad \square$$

Remark 2.3.1

Similar to results stated in subsections 4.1.3 and 4.2.3, without the strict complementarity condition, differentiability is no longer guaranteed. It is only possible to guarantee that $\tilde{x}(\tilde{\epsilon})$ is continuous with respect to $\tilde{\epsilon}$. However, Part (e) of Theorem 2.3.3 shows that $f(\tilde{x}(\tilde{\epsilon}), \tilde{\epsilon})$ is always differentiable with respect to $\tilde{\epsilon}$, even if $\tilde{x}(\tilde{\epsilon})$ is not. Hence, (2.3-26) can be used to estimate the first order change in the optimal (objective) function value when the problem is perturbed or some parameters are varied.

The strength and weakness of various stability and sensitivity results usually depend strongly on the underlying implicit function theorems. For standard implicit function theorems (see Appendix 2 for an example), certain differentiability and non-singularity conditions are required. When these standard implicit function theorems are applied to a non-linear programming problem, the differentiability requirement is the condition that requires $\bar{L}_\sigma$ to be twice differentiable with respect to $\tilde{x}$ at $\tilde{x}^*$; the non-singularity requirement corresponds to the requirement that $\nabla_{(x,z)}^2 L_\sigma((\tilde{x}^*, \tilde{z}^*), \tilde{u}^*)$ (of the extended problem) be positive definite for σ sufficiently large. Both of these mentioned conditions depend strongly on strict complementarity. In this study, it has been possible to remove the strict complementarity requirement by using a weaker form of the implicit function theorem (Appendix 2).

From studying the underlying implicit function theorems, we note that in both Theorems 2.3.1 and 2.3.3, the requirement that the functions defining $P[\underline{\epsilon}]$ be *twice* continuously differentiable in $(\underline{x}, \underline{\epsilon})$ is too strong. In fact, what is actually needed is that $\nabla_{\underline{x}} L_0$, $\{h_i\}$ and $\{g_i\}$ (or the function gradients with respect to $\underline{x}$) are *once* differentiable with respect to both $\underline{x}$ and $\underline{\epsilon}$ in a neighbourhood of $(\underline{x}^*, 0)$.

The imposed differentiability with respect to $\underline{\epsilon}$ is responsible for the differentiability property of $\underline{x}(\underline{\epsilon})$ in Theorem 2.3.1, the results in Corollary 2.3.1 and Parts (d) and (e) of Theorem 2.3.3. Consequently, if these results are not required, the differentiability condition with respect to $\underline{\epsilon}$ can be dropped. Robinson (1974) seems to be the first to notice this. By using an appropriate implicit function theorem, he has proved, without imposing the differentiability condition with respect to $\underline{\epsilon}$, a theorem analogous to Theorem 2.3.1. It is relevant to note that, in Theorem 2.3.2, Robinson's (1974) results have been strengthened by removing the strict complementarity requirement.

Theorem 2.3.3 gives a (first order) estimate of the change in the optimal (objective) function value for a small change in the parameter $\underline{\epsilon}$ (Part (e)), but it only gives a bound on the change in $\underline{x}(\underline{\epsilon})$ (Part (d)). Generally, without strict complementarity, sensitivity results of the form (2.3-2) cannot be obtained. However, in order to give a (first order) estimate of the change in $\underline{x}(\underline{\epsilon})$, only directional derivatives are required. Some work in this direction has been attempted by Bigelow and Shapiro (1974).

For a particular choice of direction $\dot{\underline{\epsilon}} \neq 0$, consider a family of perturbed problems $P[t\dot{\underline{\epsilon}}]$, $t \geq 0$, of an ICP (i.e., $P[\underline{\epsilon}]$ with $\underline{\epsilon} = t\dot{\underline{\epsilon}}$, $t \geq 0$). If $P[t\dot{\underline{\epsilon}}; t \geq 0]$ has a differentiable solution $(\underline{x}(t), \underline{u}(t))$, then, because the Kuhn-Tucker condition is always satisfied, it follows that the following system of inequalities and equations for $(\dot{\underline{x}}, \dot{\underline{u}})$

$$\left. \begin{aligned}
 \frac{\partial^2}{\partial \tilde{x}^2} L_0(\tilde{x}^*, \tilde{u}^*, 0) \dot{\tilde{x}} - \sum_{i \in J} \dot{u}_i \frac{\partial g_i}{\partial \tilde{x}} &= - \left[\frac{\partial^2 \Gamma}{\partial \tilde{\varepsilon} \partial \tilde{x}} - \sum_{i \in J} u_i^* \frac{\partial^2 g_i}{\partial \tilde{\varepsilon} \partial \tilde{x}} \right] \dot{\tilde{\varepsilon}} \\
 \frac{\partial g_i}{\partial \tilde{x}}^T \cdot \dot{\tilde{x}} &= - \frac{\partial g_i}{\partial \tilde{\varepsilon}}^T \cdot \dot{\tilde{\varepsilon}}, \quad i \in J_1(\tilde{x}^*) \\
 \frac{\partial g_i}{\partial \tilde{x}}^T \cdot \dot{\tilde{x}} &\geq - \frac{\partial g_i}{\partial \tilde{\varepsilon}}^T \cdot \dot{\tilde{\varepsilon}}, \quad i \in J_2(\tilde{x}^*) \\
 \dot{u}_i &= 0, \quad i \notin J_0(\tilde{x}^*) ; \quad \dot{u}_i \geq 0, \quad i \in J_2(\tilde{x}^*) \\
 \dot{u}_i \left[\frac{\partial g_i}{\partial \tilde{x}}^T \cdot \dot{\tilde{x}} + \frac{\partial g_i}{\partial \tilde{\varepsilon}}^T \cdot \dot{\tilde{\varepsilon}} \right] &= 0, \quad i \in J_2(\tilde{x}^*)
 \end{aligned} \right\} \quad (2.3-37)$$

admits $(dx(0)/dt, du(0)/dt)$ as a solution. Bigelow and Shapiro's (1974) results can then be stated as follows: If the system of inequalities and equations (2.3-37) admits a unique solution $(\dot{\tilde{x}}, \dot{\tilde{u}})$, then $\tilde{x}(t)$ and $\tilde{u}(t)$ are differentiable from the right in the direction $\dot{\tilde{\varepsilon}}$, and that the derivatives are $\dot{\tilde{x}}$ and $\dot{\tilde{u}}$. The mentioned results are incompleting in the sense that: (i) the existence of a unique continuous function $(\tilde{x}(t), \tilde{u}(t))$ has not been guaranteed; (ii) the results depend on the assumption that the system (2.3-37) has a unique solution. These results are strengthened in the following Theorem.

Theorem 2.3.4

For a particular choice of direction $\dot{\tilde{\varepsilon}} \neq 0$ consider the family of perturbed problems $P[t\dot{\tilde{\varepsilon}}; t \geq 0]$. Under the assumptions of Theorem 2.3.3, for sufficiently small t , there exists a continuous function $(\tilde{x}(t), \tilde{v}(t), \tilde{u}(t))$ as the (local) solution of $P[t\dot{\tilde{\varepsilon}}; t \geq 0]$. Furthermore, the right handed derivatives of $\tilde{x}(t)$, $\tilde{v}(t)$ and $\tilde{u}(t)$ exist; namely,

$$\lim_{t \rightarrow 0^+} \begin{bmatrix} \frac{\tilde{x}(t) - \tilde{x}(0)}{t} \\ \frac{\tilde{v}(t) - \tilde{v}(0)}{t} \\ \frac{\tilde{u}(t) - \tilde{u}(0)}{t} \end{bmatrix} = \begin{bmatrix} \frac{d\tilde{x}(0^+)}{dt} \\ \frac{d\tilde{v}(0^+)}{dt} \\ \frac{d\tilde{u}(0^+)}{dt} \end{bmatrix} = \begin{bmatrix} \dot{\tilde{x}} \\ \dot{\tilde{v}} \\ \dot{\tilde{u}} \end{bmatrix}. \quad (2.3-38)$$

Hence, the system of inequalities and equations corresponding to (2.3-37) has a solution $(\dot{\tilde{x}}, \dot{\tilde{v}}, \dot{\tilde{u}})$.

Proof

Without loss of generality, it is assumed that only inequality constraints are present, and that strict complementarity does not hold. From Theorem 2.3.3, the continuous solution $(\tilde{x}(t), \tilde{u}(t))$ exists. Thus, it is only necessary to show that $(\tilde{x}(t), \tilde{u}(t))$ is differentiable from the right at $t = 0$.

For every $j \in J$, g_j will be reclassified as 'active', 'inactive' or 'degenerate' depending on its behaviour with respect to the directional perturbation $t\dot{\tilde{e}}, t \geq 0$. The continuities of $g_j(t) = g_j(\tilde{x}(t), t)$ and $u_j(t)$ imply that, for sufficiently small t , $g_j(t)$ will remain active, (i.e., $u_j(t) > 0$ and $g_j(t) = 0$) if g_j is non-degenerate (active) at $t = 0$, and remain inactive (i.e., $u_j(t) = 0$ and $g_j(t) > 0$) if g_j is inactive at $t = 0$. Hence, it is only necessary to examine the behaviour of those constraints which are degenerate at $t = 0$.

For every $j \in J$, $g_j(t)$ is a continuous real-valued function of t . Therefore, there either exists or does not exist an interval $(0, \delta_1(j))$, $\delta_1(j) > 0$, such that $g_j(t) = 0, \forall t \in (0, \delta_1(j))$. Let $\delta_1(\dot{\tilde{e}}) = \min_{\delta_1(j) > 0} \delta_1(j)$, then define

$$R_1(\dot{\tilde{e}}) = \{j \mid j \in J, g_j(t) = 0, \forall t \in (0, \delta_1(\dot{\tilde{e}}))\}. \quad (2.3-39)$$

Similarly, $\forall j \in J$, there either exists or does not exist an interval $(0, \delta_3(j))$, $\delta_3(j) > 0$, such that $g_j(t) > 0, \forall t \in (0, \delta_3(j))$. Let $\delta_3(\dot{\tilde{e}}) = \min_{\delta_3(j) > 0} \delta_3(j)$, and define

$$R_3(\dot{\tilde{e}}) = \{j \mid j \in J, g_j(t) > 0, \forall t \in (0, \delta_3(\dot{\tilde{e}}))\}. \quad (2.3-40)$$

From their definitions, it is clear that $R_1(\dot{\underline{e}})$ and $R_3(\dot{\underline{e}})$ are disjoint, $J_1(\underline{x}^*) \subseteq R_1(\dot{\underline{e}})$ and $J_3(\underline{x}^*) \subseteq R_3(\dot{\underline{e}})$. For any $j \in J_2(\underline{x}^*)$, if $j \in R_1(\dot{\underline{e}})$, then this implies that g_j can be treated as an equality (or active) constraint for this directional perturbation. Similarly, if $j \in R_3(\dot{\underline{e}})$, then this implies that g_j can be ignored or treated as an inactive constraint for this directional perturbation.

If $R_1(\dot{\underline{e}}) \cup R_3(\dot{\underline{e}}) = J$, then all the constraints are included. By applying Theorem 2.3.1 to the problem $P[R_1(\dot{\underline{e}}), t\dot{\underline{e}}]$ (see (2.3-14)), and then identifying the solution $\underline{x}(t)$ of $P[t\dot{\underline{e}}; t \geq 0]$ with $\underline{x}^{R_1(\dot{\underline{e}})}(t)$, $\forall t \in [0, \delta_1(\dot{\underline{e}}))$, it follows that $\frac{d\underline{x}(0^+)}{dt} = \frac{d\underline{x}^{R_1(\dot{\underline{e}})}(0)}{dt}$.

If $R_1(\dot{\underline{e}}) \cup R_3(\dot{\underline{e}}) \subsetneq J$, this implies that $R_2(\dot{\underline{e}}) \neq \emptyset$, if

$$R_2(\dot{\underline{e}}) = J - R_1(\dot{\underline{e}}) - R_3(\dot{\underline{e}}). \quad (2.3-41)$$

The definitions of $R_1(\dot{\underline{e}})$ and $R_3(\dot{\underline{e}})$ imply that $R_2(\dot{\underline{e}}) \subseteq J_2(\underline{x}^*)$. Therefore, it follows that, $\forall j \in R_2(\dot{\underline{e}})$, $g_j(0) = 0$. For some $t_1 > 0$, if $g_j(t_1) > 0$, then, because $j \notin R_3(\dot{\underline{e}})$, there exists t_2 such that $0 < t_2 < t_1$ and $g_j(t_2) = 0$. In addition, because $j \notin R_1(\dot{\underline{e}})$, there exists t_3 such that $0 < t_3 < t_2$ and $g_j(t_3) > 0$. That is, if $j \in R_2(\dot{\underline{e}})$, then for any $t > 0$, as $t \rightarrow 0^+$, $g_j(t)$ alternates between $g_j(t) > 0$ and $g_j(t) = 0$ infinitely often.

Consider the special case where $R_2(\dot{\underline{e}})$ has only one element, (say) $R_2(\dot{\underline{e}}) = \{i\}$. By applying Theorem 2.3.1 to the problems $P[R_1(\dot{\underline{e}}), t\dot{\underline{e}}]$ and $P[R_1(\dot{\underline{e}}) \cup \{i\}, t\dot{\underline{e}}]$, it follows that their respective solutions $\underline{x}^{R_1(\dot{\underline{e}})}(t)$ and $\underline{x}^{R_1(\dot{\underline{e}}) \cup \{i\}}(t)$ are differentiable functions of t . For any sufficiently small $\bar{t} > 0$, if $g_i(\bar{t}) > 0$, then $\underline{x}(\bar{t})$ can be identified with $\underline{x}^{R_1(\dot{\underline{e}})}(\bar{t})$. If $g_i(\bar{t}) = 0$, $\underline{x}(\bar{t})$ can be identified with $\underline{x}^{R_1(\dot{\underline{e}}) \cup \{i\}}(\bar{t})$. Because $\underline{x}(t)$ is continuous in t , this means that, for any $t > 0$, as

$t \rightarrow 0^+$, $\tilde{x}^{R_1(\dot{\tilde{e}})}(t)$ and $\tilde{x}^{R_1(\dot{\tilde{e}}) \cup \{\dot{i}\}}(t)$ intersect infinitely often, with $\tilde{x}^{R_1(\dot{\tilde{e}})}(0) = \tilde{x}^{R_1(\dot{\tilde{e}}) \cup \{\dot{i}\}}(0) = \tilde{x}^*$. From this, it follows that

$$\frac{d\tilde{x}^{R_1(\dot{\tilde{e}})}(0)}{dt} = \frac{d\tilde{x}^{R_1(\dot{\tilde{e}}) \cup \{\dot{i}\}}(0)}{dt} . \quad (2.3-42)$$

Thus, because $\tilde{x}(t)$ can be identified with either $\tilde{x}^{R_1(\dot{\tilde{e}})}(t)$ or $\tilde{x}^{R_1(\dot{\tilde{e}}) \cup \{\dot{i}\}}(t)$, it follows that

$$\lim_{t \rightarrow 0^+} \frac{\tilde{x}(t) - \tilde{x}(0)}{t} = \frac{d\tilde{x}^{R_1(\dot{\tilde{e}})}(0)}{dt} = \frac{d\tilde{x}^{R_1(\dot{\tilde{e}}) \cup \{\dot{i}\}}(0)}{dt} . \quad (2.3-43)$$

When $R_2(\dot{\tilde{e}})$ has more than one element, then there will be several index sets R ($R = R_1(\dot{\tilde{e}}) \cup A$, $A \subseteq R_2(\dot{\tilde{e}})$), such that, for each R and for all $t > 0$, as $t \rightarrow 0^+$, $\tilde{x}(t)$ can be identified with $\tilde{x}^R(t)$ infinitely often. From the continuity of $\tilde{x}(t)$, it follows that $\frac{d\tilde{x}^R(0)}{dt}$ must be the same for each R , and this is the value of $\frac{d\tilde{x}(0^+)}{dt}$.

Similar arguments can then be applied to find $\frac{d\tilde{u}(0^+)}{dt}$. $\square$

Remark 2.3.2

(a) In the proof for Theorem 2.3.4, even for the case where $R_2(\dot{\tilde{e}}) \neq \emptyset$, the one-sided directional derivatives $\dot{\tilde{x}}$ and $\dot{\tilde{u}}$ can be obtained by considering only the 'active' set $R_1(\dot{\tilde{e}})$. For every $j \in R_2(\dot{\tilde{e}}) \cup R_3(\dot{\tilde{e}})$, because $j \notin R_1(\dot{\tilde{e}})$, it follows that $\dot{\tilde{u}}_j = 0$. Because of this, the system (2.3-37) of inequalities and equations can be reduced to a system of equations for $\dot{\tilde{x}}$ and $\dot{\tilde{u}}_j$, $j \in R_1(\dot{\tilde{e}})$. Clearly, this system of equations is satisfied by $\frac{d\tilde{x}^{R_1(\dot{\tilde{e}})}(0)}{dt}$ and $\frac{d\tilde{u}^{R_1(\dot{\tilde{e}})}(0)}{dt}$. The linear independence and second order sufficiency conditions also imply that this system of equations has a unique solution (see Remark 2.2.1). Thus $(\dot{\tilde{x}}, \dot{\tilde{u}})$ can be obtained by considering the problem $P[R_1(\dot{\tilde{e}}), t_{\tilde{e}}]$.

(b) If $R_1(\dot{\tilde{e}}) = R_1(-\dot{\tilde{e}})$, then both one-sided derivatives $\frac{dx(0^+)}{dt}$ and $\frac{dx(0^-)}{dt}$ are the same. The difference between degenerate and non-degenerate problems is that for a non-degenerate problem there is only one possible index set $R_1(\dot{\tilde{e}}) = J_1(x^*)$, but for a degenerate problem there are several possible index sets $R_1(\dot{\tilde{e}})$. Therefore, the trade-off against strict complementarity is that it may not be possible to assume that $\frac{dx(0^+)}{dt}$ and $\frac{dx(0^-)}{dt}$ are the same.

CHAPTER 3

BARRIER FUNCTION METHODS

3.1 CLASSICAL BARRIER FUNCTION METHODS

3.1.1 Introduction

Barrier function methods represent procedures for approximating constrained optimization problems by unconstrained problems. The approximation is accomplished by adding a term that favours points interior to the feasible region over those near the boundary. Such terms are designed so that a 'barrier' is constructed along the boundary of the feasible region S . The solution of the problem is approached from the interior of S , which is assumed to be non-empty. Barrier function methods are often referred to as *interior* point methods. It is clear that such methods are not applicable to problems with equality constraints. Associated with these methods there is a parameter r that determines the severity of the barrier and consequently the degree to which the unconstrained problem approximates the original constrained problem. As $r \rightarrow 0$, the approximation becomes increasingly accurate. Following Lootsma (1970, 1972), parametric barrier functions will be represented as:

$$B(\tilde{x}, r) = f(\tilde{x}) - r^\lambda \sum_{i \in J} \phi[g_i(\tilde{x})] \quad , \quad (3.1-1)$$

where r denotes a positive controlling parameter, ϕ is a function of a single variable η , defined and continuously differentiable for positive value of η , such that

$$\phi(0^+) = -\infty \quad . \quad (3.1-2)$$

The barrier function is said to be of *order* λ , if the function ϕ' is analytic and has a *pole of order* λ at $\eta=0$. Raising r to the power λ in (3.1-1) has certain advantages when working with the Taylor series expansion for the minimizing trajectory, $\underline{x}(r)$, associated with a given barrier function.

Simple examples of barrier functions can be obtained from (3.1-1) by putting

$$\phi'(\eta) = \eta^{-\lambda}, \quad (3.1-3)$$

with λ positive. The *logarithmic* barrier function corresponds to $\lambda=1$; while the *inverse* and *inverse quadratic* barrier functions correspond to $\lambda=2$ and $\lambda=3$, respectively.

Following Fiacco and McCormick (1968) and Lootsma (1970,1972), attention is restricted to the (first order) logarithmic barrier function

$$B(\underline{x}, r) = f(\underline{x}) - r \sum_{i \in J} \log[g_i(\underline{x})] . \quad (3.1-4)$$

The barrier term on the RHS of (3.1-4) can be generalized to depend on more than one controlling parameters. Specifically, separate controlling parameter s_i can be assigned to each constraint. The name *separable* logarithmic barrier function will be used to denote the barrier function

$$B(\underline{x}, \underline{s}) = f(\underline{x}) - \sum_{i \in J} s_i \log[g_i(\underline{x})] . \quad (3.1-5)$$

Parametric barrier function methods can be modified into methods which do not (explicitly) operate with controlling parameters. Such methods, often referred to as *parameter free*, have been investigated by Fiacco and McCormick (1967), Fiacco (1967) and Lootsma (1968,1970). However, the

study of Lootsma (1970) indicates that their *parametric* counterparts are computationally more desirable. The use of parameter-free methods seems to have fallen from favour in recent years.

Remark 3.1.1

The logarithmic family of barrier functions of Osborne and Ryan (1970) do not conform to the representation of (3.1-1). The first member of this family is

$$\phi^{(1)}(\eta) = \log(F - \log(\eta)) , \quad (3.1-6)$$

with F constant. Successive members are defined recursively by

$$\phi^{(i+1)}(\eta) = \log(\sigma + \phi^{(i)}(\eta)) , \quad (3.1-7)$$

with $\sigma > 1$. Let $\tilde{x}^{(i)}(\gamma)$ denote the minimizer of the barrier function

$$B^{(i)}(\tilde{x}, \gamma) = f(\tilde{x}) + \gamma \sum_{j \in J} \phi^{(i)}(g_j(\tilde{x})) . \quad (3.1-8)$$

Then an interesting result (Osborne and Ryan (1970)) is that not only does $\tilde{x}^{(i)}(\gamma) \rightarrow \tilde{x}^*$ as $\gamma \rightarrow 0$ for fixed i , but also $\tilde{x}^{(i)}(\gamma) \rightarrow \tilde{x}^*$ as $i \rightarrow \infty$ for fixed γ . As pointed out by Osborne and Ryan (1970), the minimizer $\tilde{x}^{(i)}(\gamma)$ of $B^{(i)}(\tilde{x}, \gamma)$ is also the minimizer of $B(\tilde{x}, \tilde{s})$ in (3.1-5) for some $\tilde{s}^{(i)}$, and that $i \rightarrow \infty$ corresponds to $\tilde{s}^{(i)} \rightarrow 0$. Their method may therefore be interpreted as a separable logarithmic barrier function method without the usual explicit control on the controlling parameters, $\tilde{s}$.

3.1.2 Logarithmic Barrier Function Methods

As indicated in subsection 3.1.1, attention will be restricted to an examination of logarithmic barrier functions. Initially, basic

properties of these barrier functions are developed and discussed. For an inequality constrained problem, the transformed objective function, obtained using logarithmic barrier functions, takes the form:

$$B(\underline{x}, r) = f(\underline{x}) - r \sum_{i \in J} \log[g_i(\underline{x})] \quad (3.1-9)$$

Convergence

Both the existence of a local minimizing point $\underline{x}^{(k)}$ of $B(\underline{x}, r_k)$ for sufficiently small r_k , and the convergence, as $r_k \rightarrow 0$, of $\underline{x}^{(k)}$ to a solution $\underline{x}^*$ of the ICP can be established under mild topological conditions. The conditions amount to requiring that the problem functions be continuous and that $\underline{x}^*$ be in the closure of the interior of S (the feasible region).

Theorem 3.1.1 (Fiacco and McCormick (1968), p.47)

- If (a) the functions f and $\{g_i\}$ are continuous,
- (b) there exists a non-empty isolated compact set, A^* , of local minima corresponding to a local minimum value f^* ,
- (c) at least one point in A^* is in the closure of the interior of S , and
- (d) $\{r_k\}$ is a strictly decreasing null sequence,

then, for $B(\underline{x}, r_k)$ defined by (3.1-9)

- (i) there exists a compact set W such that $A^* \subset \text{interior of } W$ and, for r_k sufficiently small, the *unconstrained minima* of $B(\underline{x}, r_k)$ in $\{\text{interior of } W\} \cap \{\text{interior of } S\}$ exist and every limit point of any subsequence $\{\underline{x}^{(k)}\}$ of the minimizing points is in A^* ,

$$(ii) \quad \lim_{k \rightarrow \infty} r_k \sum_{i \in J} \log [g_i(\tilde{x}^{(k)})] = 0 ,$$

$$(iii) \quad \lim_{k \rightarrow \infty} f(\tilde{x}^{(k)}) = f^* ,$$

$$(iv) \quad \lim_{k \rightarrow \infty} B(\tilde{x}^{(k)}, r_k) = f^* ,$$

$$(v) \quad \{f(\tilde{x}^{(k)})\} \text{ is a monotonically decreasing sequence, and}$$

$$(vi) \quad \left\{ - \sum_{i \in J} \log [g_i(\tilde{x}^{(k)})] \right\} \text{ is a monotonically increasing sequence.}$$

□

A more general result is available for the separable barrier function

$$B(\tilde{x}, s_k) = f(\tilde{x}) - \sum_{i \in J} s_{i,k} \log [g_i(\tilde{x})] , \quad (3.1-10)$$

where $\{s_k\}$ is a strictly decreasing null sequence with the ordering determined componentwise.

Theorem 3.1.2 (Osborne and Ryan (1970))

If conditions (a), (b) and (c) of Theorem 3.1.1 are satisfied and $\{s_k\}$ is a strictly decreasing null sequence, then conclusions (i), (ii), (iii) and (iv) of Theorem 3.1.1 remain valid.

Remark 3.1.2

(a) The barrier function method behaves as if it is solving the problem: $\inf f(\tilde{x})$ subject to $g_i(\tilde{x}) > 0, \forall i \in J$. This implies that the barrier function approach will not converge to local minima which are isolated points of the feasible region.

(b) The convergence result in Theorem 3.1.1 is independent of assumptions such as the Kuhn-Tucker condition. Indeed, the logarithmic barrier function guarantees convergence to local minima at which the Kuhn-Tucker condition does not hold. For illustration, consider the use of the logarithmic barrier function method to solve the following problem: minimize $f(\underline{x})$ subject to $g_1(\underline{x}) \geq 0$ and $g_2(\underline{x}) \geq 0$. Assume that, as $r_k \rightarrow 0$, the unconstrained minimizer, $\underline{x}^{(k)}$, of

$$B(\underline{x}, r_k) = f(\underline{x}) - r_k \log[g_1(\underline{x})] - r_k \log[g_2(\underline{x})]$$

converges to $\underline{x}^*$, and that both g_1 and g_2 are active at $\underline{x}^*$. The properties of logarithmic functions then implies that, for

$$G(\underline{x}) = g_1(\underline{x}) \cdot g_2(\underline{x}),$$

$$B(\underline{x}, r_k) = f(\underline{x}) - r_k \sum_{j=1}^2 \log[g_j(\underline{x})] = f(\underline{x}) - r_k \log[G(\underline{x})].$$

Hence, $\underline{x}^{(k)}$ converges to $\underline{x}^*$ which is also a solution of the problem: minimize $f(\underline{x})$ subject to $G(\underline{x}) \geq 0$. Because $\nabla G(\underline{x}^*) = 0$, the Kuhn-Tucker condition cannot be satisfied at $\underline{x}^*$ (except when $\nabla f(\underline{x}^*) = 0$). Thus, the logarithmic barrier function method behaves as if it is solving the problem with a single constraint (product of all constraints) and converges to a solution $\underline{x}^*$ which is generally not a Kuhn-Tucker point.

(c) Since we are only interested in a particular solution, $\underline{x}^*$, it follows that $A^* = \{\underline{x}^*\}$ and the set W is a closed bounded neighbourhood of $\underline{x}^*$. Because W is compact, and there is only one limit point, $\underline{x}^*$, for any subsequence of $\{\underline{x}^{(k)}\} \subseteq W$, it follows that, as $k \rightarrow \infty$,

$$\underline{x}^{(k)} \rightarrow \underline{x}^* \quad . \quad (3.1-11)$$

(d) At the local minimizer, $\underline{x}^{(k)}$, of $B(\underline{x}, r_k)$

$$\nabla f(\tilde{x}^{(k)}) - \sum_{i \in J} \frac{r_k}{g_i(\tilde{x}^{(k)})} \nabla g_i(\tilde{x}^{(k)}) = 0. \quad (3.1-12)$$

Therefore, if the linear independence condition holds at $\tilde{x}^*$, (3.1-11) and (3.1-12) imply that, as $k \rightarrow \infty$,

$$u_i^{(k)} = \frac{r_k}{g_i(\tilde{x}^{(k)})} \rightarrow u_i^*, \quad \forall i \in J, \quad (3.1-13)$$

where $u_i^{(k)}$ will be referred to as the *multiplier estimate* associated with r_k (or associated with s_k for the separable barrier function where $u_i^{(k)} = s_{i,k}/g_i(\tilde{x}^{(k)})$).

Rate of Convergence

By imposing suitable structure on the ICP, it is possible to show that the minimizing point, $\tilde{x}^{(k)}$, of $B(\tilde{x}, s_k)$ (discussed in Theorem 3.1.2) is locally unique and a continuous function of s , for $s > 0$.

Theorem 3.1.3

Assume that the problem functions are twice continuously differentiable, and that, as $s_k \rightarrow 0$, the sequence of minimizing points, $\{\tilde{x}^{(k)}\}$, of $B(\tilde{x}, s_k)$ converges to $\tilde{x}^*$ which is a regular local solution of the problem, then, for sufficiently small $\bar{s} > 0$, the minimizing point $\bar{\tilde{x}}$ is locally unique and $\nabla_x^2 B(\bar{\tilde{x}}, \bar{s})$ is positive definite.

Proof

It is only necessary to show that $\nabla_x^2 B(\bar{\tilde{x}}, \bar{s})$ is positive definite. As $s_k \rightarrow 0$, Remarks 3.1.2c and 3.1.2d imply that

$$\lim_{k \rightarrow \infty} \frac{s_{i,k}}{g_i(\tilde{x}^{(k)})^2} = \lim_{k \rightarrow \infty} \frac{u_i^{(k)}}{g_i(\tilde{x}^{(k)})} = \infty, \quad \forall i \in J_1(\tilde{x}^*). \quad (3.1-14)$$

Therefore, it follows that $\min_{i \in J_1(\bar{x})} \bar{s}_i / g_i(\bar{x})^2$ can be made as large as required by taking $\bar{s} > 0$ suitably small. At the minimizer, $\bar{x}$, of $B(\bar{x}, \bar{s})$, for a bounded constant $\sigma > 0$,

$$\nabla_{\bar{x}}^2 B(\bar{x}, \bar{s}) = \nabla^2 f(\bar{x}) - \sum_{i \in J} \frac{\bar{s}_i}{g_i(\bar{x})} \nabla^2 g_i(\bar{x}) + \sum_{i \in J} \frac{\bar{s}_i}{g_i(\bar{x})^2} \nabla g_i(\bar{x}) \cdot \nabla g_i(\bar{x})^T \quad (3.1-15)$$

$$= \nabla_{x^*}^2 L_0(x^*, u^*) + \sigma N(x^*) N(x^*)^T + E(\bar{x}, \sigma) + K(\bar{x}, \sigma) \quad (3.1-16)$$

where

$$\nabla_{x^*}^2 L_0(x^*, u^*) = \nabla^2 f(x^*) - \sum_{i \in J} u_i^* \nabla^2 g_i(x^*) ;$$

$$N(x^*) = [\nabla g_i(x^*), \nabla g_j(x^*), \dots] ; i, j \in J_1(x^*) ;$$

$$K(\bar{x}, \sigma) = \sum_{i \in J_1(x^*)} \left(\frac{\bar{s}_i}{g_i(\bar{x})^2} - \sigma \right) \nabla g_i(\bar{x}) \cdot \nabla g_i(\bar{x})^T + \sum_{i \notin J_1(x^*)} \frac{\bar{s}_i}{g_i(\bar{x})^2} \nabla g_i(\bar{x}) \cdot \nabla g_i(\bar{x})^T ;$$

and

$$E(\bar{x}, \sigma) = \nabla^2 f(\bar{x}) - \nabla^2 f(x^*) - \sum_{i \in J} \left(\frac{\bar{s}_i}{g_i(\bar{x})} \cdot \nabla^2 g_i(\bar{x}) - u_i^* \nabla^2 g_i(x^*) \right) + \sigma \sum_{i \in J_1(x^*)} \left(\nabla g_i(\bar{x}) \nabla g_i(\bar{x})^T - \nabla g_i(x^*) \nabla g_i(x^*)^T \right) .$$

Since σ is bounded, it follows that, for suitably small $\bar{s} > 0$, $K(\bar{x}, \sigma)$ is positive semi-definite, and $E(\bar{x}, \sigma)$ is small with components of order $o(1)$. Hence, it is sufficient to show that $\nabla_{x^*}^2 L_0(x^*, u^*) + \sigma N(x^*) N(x^*)^T$ is positive definite for some bounded σ .

Consider any vector t such that $\|t\| = 1$. Let t be decomposed into a component p orthogonal to the columns of $N(x^*)$ plus a component which is a linear combination of the columns of $N(x^*)$; i.e.,

$$\underline{t} = \underline{p} + N(\underline{x}^*)^+ \underline{q}, \quad (3.1-17)$$

where N^+ is the pseudo inverse of N : $N^+ = (N^T N)^{-1} N^T$.

Thus

$$\begin{aligned} \underline{t}^T (\nabla_{\underline{x}}^2 L_0(\underline{x}^*, \underline{u}^*) + \sigma N(\underline{x}^*) N(\underline{x}^*)^T) \underline{t} &= \underline{p}^T \nabla_{\underline{x}}^2 L_0(\underline{x}^*, \underline{u}^*) \underline{p} + 2 \underline{p}^T \nabla_{\underline{x}}^2 L_0(\underline{x}^*, \underline{u}^*) N(\underline{x}^*)^+ \underline{q} \\ &\quad + \underline{q}^T N(\underline{x}^*)^+ \nabla_{\underline{x}}^2 L_0(\underline{x}^*, \underline{u}^*) N(\underline{x}^*)^+ \underline{q} + \sigma \|\underline{q}\|^2. \end{aligned} \quad (3.1-18)$$

The second order sufficiency condition implies that there exists $a > 0$ such that

$$\underline{p}^T \nabla_{\underline{x}}^2 L_0(\underline{x}^*, \underline{u}^*) \underline{p} \geq a \|\underline{p}\|^2, \quad \forall \underline{p} \text{ such that } N(\underline{x}^*)^T \underline{p} = 0. \quad (3.1-19)$$

Writing $\|\nabla_{\underline{x}}^2 L_0(\underline{x}^*, \underline{u}^*) N(\underline{x}^*)^+\| = b$ and $\|N(\underline{x}^*)^+ \nabla_{\underline{x}}^2 L_0(\underline{x}^*, \underline{u}^*) N(\underline{x}^*)^+\| = c$,
Eqn. (3.1-18) then becomes

$$\underline{t}^T (\nabla_{\underline{x}}^2 L_0(\underline{x}^*, \underline{u}^*) + \sigma N(\underline{x}^*) N(\underline{x}^*)^T) \underline{t} \geq a \|\underline{p}\|^2 - 2b \|\underline{p}\| \|\underline{q}\| + (\sigma - c) \|\underline{q}\|^2 \quad (3.1-20)$$

$$\geq a \left[\|\underline{p}\| - \frac{b}{a} \|\underline{q}\| \right]^2 + \left(\sigma - c - \frac{b^2}{a} \right) \|\underline{q}\|^2. \quad (3.1-21)$$

Because $\underline{p} = \underline{q} = 0$ cannot hold for $\|\underline{t}\| = 1$, if σ is chosen such that $\sigma > c + \frac{b^2}{a}$, then the required result follows. $\square$

Remark 3.1.3

The result of Theorem 3.1.3 does not require strict complementarity.

Because, for all sufficiently small $\bar{s} > 0$, $\nabla_{\underline{x}}^2 B(\bar{\underline{x}}, \bar{s})$ is positive definite with $\nabla_{\underline{x}} B(\bar{\underline{x}}, \bar{s}) = 0$, the implicit function theorem (see Appendix 2) implies that there exists $\underline{x}(\underline{s})$ (and also $\underline{u}(\underline{s})$, $u_i(\underline{s}) = s_i / g_i(\underline{x}(\underline{s}))$)

which is a continuously differentiable function of $\underline{s}$ such that, for sufficiently small $\underline{s} > 0$, $\underline{x}(\underline{s})$ is the unique (local) minimizer of $B(\underline{x}, \underline{s})$.

For the classical logarithmic barrier function, $s_i = s_j = r$, $\forall i, j \in J$, $\underline{x}(r)$ describes a continuously differentiable trajectory for $r > 0$. If the derivatives of $\underline{x}(r)$ exist at $r = 0^+$, then, for small r , $\underline{x}(r)$ admits a Taylor series expansion about $r = 0$, providing a basis for extrapolation. In order to ensure that derivatives of $\underline{x}(r)$ exist at $r = 0^+$, additional structure must be imposed on the problem. At the minimizer $\underline{x}(r)$ of $B(\underline{x}, r)$, the following system of equations are satisfied

$$\nabla f(\underline{x}(r)) - \sum_{i \in J} u_i(r) \nabla g_i(\underline{x}(r)) = 0 \quad (3.1-22)$$

$$u_i(r) g_i(\underline{x}(r)) = r, \quad \forall i \in J.$$

By differentiating the system (3.1-22), it follows that $\underline{x}(r)$ and $\underline{u}(r)$ satisfy the system of differential equations

$$J(r) \begin{bmatrix} \frac{d\underline{x}(r)}{dr} \\ \frac{d\underline{u}(r)}{dr} \end{bmatrix} = \begin{bmatrix} \nabla_x^2 L_0(\underline{x}(r), \underline{u}(r)) \dots - \nabla g_i(\underline{x}(r)) \dots \\ \vdots \\ u_i(r) \nabla g_i(\underline{x}(r))^T \quad g_i(\underline{x}(r)) \\ \vdots \end{bmatrix} \begin{bmatrix} \frac{d\underline{x}(r)}{dr} \\ \frac{d\underline{u}(r)}{dr} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ e \end{bmatrix}, \quad (3.1-23)$$

where $J(r)$ is the Jacobian of the ICP (cf., (2.2-4) in subsection 2.2.1), and $\underline{e}^T = [1, 1, \dots, 1]$.

Theorem 3.1.4

Let $\underline{x}^*$ be a regular local solution of a non-degenerate ICP, then

- (i) $J(0)$ is non-singular,
- (ii) $\tilde{x}(r)$ and $\tilde{u}(r)$ lie on a smooth trajectory uniquely determined for sufficiently small r by the system of differential equations

$$J(r) \begin{bmatrix} \frac{d\tilde{x}(r)}{dr} \\ \frac{d\tilde{u}(r)}{dr} \end{bmatrix} = \begin{bmatrix} 0 \\ e \end{bmatrix}, \quad (3.1-24)$$

with initial conditions $\tilde{x}(0) = \tilde{x}^*$, $\tilde{u}(0) = \tilde{u}^*$,

and

- (iii) $\|\tilde{x}(r) - \tilde{x}^*\|, \|\tilde{u}(r) - \tilde{u}^*\| = o(r)$.

Proof

- (i) follows from Lemma 2.2.1
- (ii) follows from the above discussion and Remarks 3.1.2c and 3.1.2d.
- (iii) follows from (i) and (ii). □

Corollary 3.1.1

Assume that the problem functions are $k+1$ times continuously differentiable, and that the conditions of Theorem 3.1.4 hold, then $\tilde{x}(r)$ and $\tilde{u}(r)$ are k times continuously differentiable with respect to r , for sufficiently small $r \geq 0$.

Proof

By repeated differentiation of the system (3.1-24). □

Remark 3.1.4

(a) For a non-degenerate active constraint g_i ,

$$g_i(\tilde{x}(r)) = \frac{1}{u_i(r)} \cdot r \quad . \quad (3.1-25)$$

Eqn. (3.1-25) and (iii) of Theorem 3.1.4 imply

$$\|\tilde{x}(r) - \tilde{x}^*\| \geq \frac{1}{u_i^* \|\nabla g_i(\tilde{x}^*)\|} \cdot r + o(r) \quad . \quad (3.1-26)$$

Eqn. (3.1-26) and (iii) of Theorem 3.1.4 then imply that the convergence of $\tilde{x}(r)$ to $\tilde{x}^*$ is *strictly* order r . There exist constants $\mu_2 > \mu_1 > 0$ such that the asymptotic inequalities

$$\mu_1 r_k + o(r_k) \leq \|\tilde{x}(r_k) - \tilde{x}^*\| \leq \mu_2 r_k + o(r_k) \quad (3.1-27)$$

hold for sufficiently small r_k .

(b) The non-degeneracy assumption is necessary as well as sufficiently for the convergence of $\tilde{x}(r)$ to $\tilde{x}^*$ to be order r . Cases when non-degeneracy does not hold will be discussed and analysed in subsection 3.2.2.

Robustness

Barrier function methods have several desirable robustness properties. These properties can be identified as follows:

(i) The barrier function methods converge to a solution of the problem under mild topological conditions (see Theorems 3.1.1 and 3.1.2 and Remark 3.1.2b). The generality of the convergence results suggests that barrier function methods have wide applicability.

(ii) Barrier function methods are characterised by their property of preserving constraint feasibility at all times. This is a useful property when solving problems in which the objective function is only defined in the feasible region S . It also has the reassuring feature that, should the algorithm be terminated prematurely, a feasible solution is always returned.

(iii) For convex programming problems, Mifflin (1972,1975) has established, under strong convexity assumptions on the problem functions, that the classical logarithmic barrier function method will converge even when $B(\tilde{x}, r)$ is 'minimized' only to within a predetermined tolerance of $B(\tilde{x}(r), r)$. The importance of this result lies in the observation that inexact minimizations are an unavoidable consequence of any numerical implementation of the barrier function algorithm. In order to extend this result to non-convex cases, it is assumed that, as $r \rightarrow 0$, $\tilde{x}(r)$ converges to $\tilde{x}^*$, a *regular* local solution of the ICP. Theorem 3.1.3 then implies that, for suitably small $r_k > 0$, $\nabla_x^2 B(\tilde{x}(r_k), r_k)$ is positive definite with its smallest eigenvalue positive and bounded away from zero. Let $\tilde{x}^{(k)}$ be an estimate of the minimizer of $B(\tilde{x}, r_k)$. From the differentiability of the problem functions, it follows that

$$\nabla_x^2 B(\tilde{x}(r_k), r_k) (\tilde{x}^{(k)} - \tilde{x}(r_k)) = \nabla_x B(\tilde{x}^{(k)}, r_k) + o(\|\tilde{x}^{(k)} - \tilde{x}(r_k)\|), \quad (3.1-28)$$

and hence

$$\tilde{x}^{(k)} - \tilde{x}(r_k) = \left[\nabla_x^2 B(\tilde{x}(r_k), r_k) \right]^{-1} \left(\nabla_x B(\tilde{x}^{(k)}, r_k) + o(\|\tilde{x}^{(k)} - \tilde{x}(r_k)\|) \right). \quad (3.1-29)$$

Because $\nabla_x^2 B(\tilde{x}(r_k), r_k)$ is positive definite, it admits a decomposition of the form:

$$\nabla_x^2 B(\tilde{x}(r_k), r_k) = U D U^T, \quad (3.1-30)$$

where U is an orthogonal matrix, and D is a diagonal matrix with the eigenvalues of $\nabla_x^2 B(\tilde{x}(r_k), r_k)$ along its diagonal. By inverting the matrix in (3.1-30), it follows that

$$\| [\nabla_x^2 B(\tilde{x}(r_k), r_k)]^{-1} \nabla_x B(\tilde{x}^{(k)}, r_k) \| \leq \frac{1}{\lambda^*} \| \nabla_x B(\tilde{x}^{(k)}, r_k) \|, \quad (3.1-31)$$

where λ^* is the smallest eigenvalue of $\nabla_x^2 B(\tilde{x}(r_k), r_k)$.

Therefore, there exist $r^0, \epsilon^0 > 0$ and $0 < \gamma < \infty$ ($\gamma = 1/\lambda^* + o(1)$) such that, if $r_k \leq r^0$ and $\| \nabla_x B(\tilde{x}^{(k)}, r_k) \| \leq \epsilon^0$, then

$$\begin{aligned} \| \tilde{x}^{(k)} - \tilde{x}^* \| &= \| \tilde{x}^{(k)} - \tilde{x}(r_k) + \tilde{x}(r_k) - \tilde{x}^* \| \\ &\leq \| \tilde{x}(r_k) - \tilde{x}^* \| + \gamma \| \nabla_x B(\tilde{x}^{(k)}, r_k) \|. \end{aligned} \quad (3.1-32)$$

(3.1-32) implies that when the conditions of Theorem 3.1.4 are satisfied, even with inexact minimization, the $O(r_k)$ rate of convergence can be retained by ensuring that $\| \nabla_x B(\tilde{x}^{(k)}, r_k) \| \leq \alpha r_k$, for some $\alpha > 0$.

3.1.3 Difficulties Associated with Barrier Function Methods

In this subsection, difficulties encountered when using barrier function methods are identified and discussed. They are of a computational nature, and can be listed as follows:

(i) *Ill-conditioning*: As the controlling parameter r becomes small, this will allow the minimum of $B(\tilde{x}, r)$ to approach $\tilde{x}^*$, the solution of the problem. Generally, $\tilde{x}^*$ will be on the boundary of S , and is a singularity for those barrier terms corresponding to active constraints. Therefore, the unconstrained subproblem becomes more difficult to solve

because the required minimum lies in a steeply sided valley. This phenomenon has been discussed and studied by several authors (see, for example, Murray (1971), Lootsma (1969,1970) and Fletcher and McCann (1969)) in terms of the condition number or eigenvalues and eigenvectors of the matrix $\nabla_{\tilde{x}}^2 B(\tilde{x}(r), r)$. From (3.1-14) and (3.1-15), it is clear that, as $r \rightarrow 0$, the terms $r/g_{\lambda}(\tilde{x}(r))^2$, which correspond to non-degenerate active constraints, become unbounded, resulting in the ill-conditioning of the matrix $\nabla_{\tilde{x}}^2 B(\tilde{x}(r), r)$.

Lootsma (1970,1972) has used arguments concerning the condition number of $\nabla_{\tilde{x}}^2 B(\tilde{x}(r), r)$ to justify the choice of first order barrier functions for computational purpose.

Barrier function methods are parametric in nature. The definition of $\tilde{x}^{(k+1)}$ in no way depends on that of $\tilde{x}^{(k)}$. The sequential characteristic of these algorithms arises from attempts to overcome computational difficulties associated with ill-conditioning of the unconstrained problem. If the minimization of $B(\tilde{x}, r)$ is initiated from a poor starting point with a very small value of r , this could lead to all sorts of difficulties. It is therefore computationally preferable that $B(\tilde{x}, r)$ be first minimized for a moderate value of r , and then this minimizing point is used as a starting point for the minimization of $B(\tilde{x}, r)$ with a smaller value of r . Also, because $\tilde{x}(r)$ describes a 'smooth' trajectory converging to $\tilde{x}^*$ as r decreases, acceleration of the convergence can be obtained by extrapolation according to the Richardson-Romberg principle. Theorem 3.1.4 and Corollary 3.1.1 suggest that an extrapolation function, which is a polynomial in r , could be fitted to successive values of $\tilde{x}^{(k)}$. The extrapolation function could then be used to predict $\tilde{x}^*$ by evaluating it at $r=0$; or, alternatively, could be used to predict an initial estimate of $\tilde{x}^{(k+1)}$, the minimizing point of $B(\tilde{x}, r_{k+1})$. It should be clear that the extrapolation procedure does not remove any ill-conditioning of the unconstrained problem.

However, computational difficulties are partially overcome in the sense that: (i) the computational effort required to minimize $B(\underline{x}, r)$ is reduced by initiating the minimization from the better starting point given by the extrapolation; (ii) given a required accuracy for $\underline{x}^*$, the unconstrained problems which must be solved are not as ill-conditioned as they would be if extrapolation were not used.

(ii) *Unbounded minima*: Theorem 3.1.1 only guarantees the existence of bounded minima within a compact set W . Therefore, even if the original objective function $f(\underline{x})$ is bounded below in the feasible region S , $B(\underline{x}, r)$ may be unbounded below (see Powell (1972) for an example). Assuming that the objective function $f(\underline{x})$ is bounded below in S , then any unbounded behaviour comes from the barrier term $-r \sum_{i \in J} \log[g_i(\underline{x})]$. If it occurs, the problem can be rectified by any *ad hoc* means, or simply by reducing r to scale down the contribution of $-r \sum_{i \in J} \log[g_i(\underline{x})]$. However, because the feasible region S is often bounded and closed, the problem is not likely to occur.

(iii) *Linear search*: Barrier function methods work in the interior of the feasible region S . For most problems, the solution $\underline{x}^*$ usually lies on the boundary of S . As $r \rightarrow 0$, the minimizing point, $\underline{x}(r)$, of $B(\underline{x}, r)$ approaches $\underline{x}^*$ on the boundary of S . This implies that linear search procedures within the unconstrained minimization routine may need to be modified to accommodate attempts to evaluate $B(\underline{x}, r)$ at infeasible points. The simple expedient of setting the function value to 'infinity' may not be sufficient if this value is then used to determine an interpolating function to perform the linear search or if the function value is used to compute or update derivative information about $B(\underline{x}, r)$. The linear search procedures incorporated in many unconstrained algorithms are usually performed by fitting quadratic or cubic functions. It is not realistic to expect such

functions to model accurately the behaviour of the barrier function with its singularity at the boundary of the feasible region. This problem has been examined by several authors (e.g., Murray (1969a), Fletcher and McCann (1969), Lasdon *et al.* (1973) and Wright (1976)) who propose and numerically evaluate a number of alternative models incorporating singularities.

Remark 3.1.5

(a) It is questionable whether special linear search procedures for unconstrained minimization routines are computationally worth-while. In order to locate the singularity, additional function and gradient evaluations may be required. Some safeguards are also needed to take care of cases when there is no singularity along the search direction. This amounts to creating or tailoring a special version of an unconstrained minimization routine for the algorithm. In the author's experience, the algorithm performs very satisfactorily without any special linear search procedures. This is because, for moderate value of r , the unconstrained minimization of $B(\tilde{x}, r)$ is not so ill-conditioned, and can be handled efficiently by a standard minimization routine. For successively smaller values of r , the minimum, $\tilde{x}^{(k+1)}$, predicted by the extrapolation procedure is usually such a good starting point that any attempts to locate the singularity may actually be inefficient. Another point concerning linear search is that for most unconstrained minimization routines the linear search procedure has a stepsize control device which will automatically reduce the stepsize by some factor when a certain criterion is not satisfied (see Dixon (1972) and Sargent and Sebastian (1972)).

Some of these stepsize control devices can be exploited in the sense that, whenever an attempt is made to evaluate $B(\tilde{x}, r)$ outside the feasible region, the simple expedient of setting a 'large' value to $B(\tilde{x}, r)$

will force the stepsize to be reduced, and the algorithm thereby stays inside the feasible region. In numerical experimentation with the barrier function examples of subsections 3.2.4 and 3.3.3 (setting $B(\underline{x}, r) = \text{'large'}$ for $\underline{x} \notin S$ and using a quasi-Newton routine from the ANU's subroutine library), attempts to evaluate $B(\underline{x}, r)$ outside the feasible region were rare. Usually, they were made in early stages of the first unconstrained minimization where the starting point was far away from the minimum, and the minimization routine had little or no second derivative information. After the first minimization, with an estimate for the hessian and a better starting point, the unconstrained routine stayed inside the feasible region for the range of problems tested.

(b) Another difficulty associated with the minimization of $B(\underline{x}, r)$, using a standard minimization routine, arises in the specification of convergence criteria. Barrier functions provide very good examples for testing the robustness of unconstrained minimization routines. As r decreases, the barrier function becomes extremely steep-valleyed and in a very small neighbourhood of $\underline{x}(r)$, the magnitude of the gradient of $B(\underline{x}, r)$ can take very large values. In such circumstances, it is practically impossible to reduce the gradient magnitude numerically. Hence convergence criteria based solely on gradient magnitudes are unsuitable. Ill-conditioning is also likely to cause premature termination for criteria based on objective function improvement or step-lengths in successive iterations of the unconstrained algorithm. However, these are problems associated with the design of robust unconstrained algorithms, details of which will not be discussed here.

3.2 TRAJECTORY ANALYSIS AND EXTRAPOLATION

3.2.1 Introduction

When strict complementarity holds at $\underline{x}^*$, Theorem 3.1.4 and Corollary 3.1.1 provide the basic information necessary to use standard extrapolation procedures to obtain estimates of $\underline{x}^*$ and $\underline{u}^*$ from the results of minimizing $B(\underline{x}, r)$ for a sequence of r values: $r_1 < r_2 < \dots < r_k$. Standard extrapolation applied to $\underline{x}_i(r)$ at the points $r_1, r_2, \dots, r_k$ is equivalent to first fitting a Lagrange interpolation polynomial of degree $k-1$, $P_{k-1}(\{\underline{x}_i(r_j), j=1, \dots, k\}; r)$, to the data, and then evaluating it at $r=0$. A standard argument (Appendix 3) shows that the error associated with such an estimate is

$$P_{k-1}(\{\underline{x}_i(r_j), j=1, \dots, k\}; 0) - \underline{x}_i^* = \frac{(-1)^{k+1}}{k!} \prod_{j=1}^k r_j \frac{d^k \underline{x}_i}{dr^k}(\xi), \quad (3.2-1)$$

where ξ denotes a mean value. The extrapolation is known to be numerically stable provided r_{j+1}/r_j is sufficiently small for each j (Appendix 3).

The numerical performance of extrapolation when used to improve the classical barrier function algorithm has been considered by several authors (see, for example, Fiacco and McCormick (1968), Lootsma (1968, 1970) and Fletcher and McCann (1969)). Although the resulting algorithms tend to be very robust, the improvement in efficiency is not always satisfactory. Luenberger (1973) suggests that this may be due to the fact that the first k vectors $\underline{x}^{(1)}, \underline{x}^{(2)}, \dots, \underline{x}^{(k)}$ span only a k -dimensional subspace, whereas n dimensions are required to represent a general point. Even though no detailed investigation has yet been attempted, there is a common concern about the reason why the extrapolation procedure, which is generally a powerful tool for approximating the limit of an infinitesimal process, is not so successful in this situation.

In this section, an explanation of the numerical behaviour of extrapolation procedures, as applied to barrier function methods, is attempted. The strategy is to examine the implication of the non-degeneracy assumption $J_2(\tilde{x}^*, \tilde{u}^*) = \phi$. In subsection 3.2.2, it is shown that when strict complementarity does not hold, the trajectory analysis summarised in Theorem 3.1.4 is no longer valid. The resulting trajectory $\tilde{x}(r)$ for a degenerate problem turns out to be a trajectory tangential to the constraint surfaces $g_i(x) = 0$, $i \in J_1(\tilde{x}^*, \tilde{u}^*)$, and the rate of convergence is of order $O(r^{1/2})$ instead of the usual order $O(r)$ for a non-degenerate problem. It proves to be possible to introduce a 'measure of non-degeneracy' and this is done in subsection 3.2.3 where it is also shown that this measure is an important parameter in determining, at a point on the solution trajectory, the growth of $\frac{d^i x}{dr^i}$ and $\frac{d^i u}{dr^i}$ as i increases. In subsection 3.2.4, a possible strategy for estimating the worth of successive extrapolation steps is suggested and illustrated by numerical results.

The following simple example is used to illustrate the results of §3.2.

Example 3.2.1

For $\delta \geq 0$;

$$\text{minimize } f(\tilde{x}) = x_2$$

$$\text{subject to } g_1(\tilde{x}) = x_2 - x_1^2 \geq 0,$$

$$\text{and } g_2(\tilde{x}) = x_1 - \delta \geq 0.$$

The solution to this problem is

$$\tilde{x}^* = (\delta, \delta^2), \text{ and } \tilde{u}^* = (1, 2\delta).$$

Using the logarithmic barrier function method to solve Ex. 3.2.1

$$B(\underline{x}, r) = x_2 - r \log(x_2 - x_1^2) - r \log(x_1 - \delta) . \quad (3.2-2)$$

At the minimum of $B(\underline{x}, r)$, $\nabla_{\underline{x}} B(\underline{x}(r), r) = 0$

$$0 = (0, 1) - \frac{r}{x_2 - x_1^2} (-2x_1, 1) - \frac{r}{x_1 - \delta} (1, 0) . \quad (3.2-3)$$

By solving the two equations in (3.2-3) for the feasible solution, it follows that

$$x_1(r; \delta) = \frac{\delta + \sqrt{\delta^2 + 2r}}{2} , \quad (3.2-4a)$$

$$x_2(r; \delta) = \frac{3r}{2} + \frac{\delta^2}{2} + \frac{\delta \sqrt{\delta^2 + 2r}}{2} . \quad (3.2-4b)$$

For the limiting case, $\delta = 0$, the problem is degenerate with

$$x_1(r) = \left(\frac{r}{2}\right)^{1/2} , \quad x_2(r) = \frac{3r}{2} . \quad (3.2-5)$$

The convergence rate of $\underline{x}(r)$ to $\underline{x}^*$ is of order $O(r^{1/2})$, and the trajectory $\underline{x}(r)$ is tangential to the constraint surface $g_1(\underline{x}) \equiv x_2 - x_1^2 = 0$.

For other cases, where $\delta > 0$, the problem is non-degenerate and Theorem 3.1.4 applies. Using an expansion of $\sqrt{\delta^2 + 2r}$ for $r < \frac{\delta^2}{2}$, it follows that

$$x_1(r; \delta) = \frac{1}{2} \left(\delta + \delta + \frac{r}{\delta} - \frac{r^2}{2\delta^3} + \dots \right) , \quad (3.2-6a)$$

$$x_2(r; \delta) = \frac{1}{2} \left(3r + \delta^2 + \delta^2 + r - \frac{r^2}{2\delta^2} + \dots \right) . \quad (3.2-6b)$$

The expansion in (3.2-6) confirms that asymptotically the convergence rate of $\underline{x}(r; \delta)$ to $\underline{x}^*(\delta)$ is of order $O(r)$. However, the expansion is only

valid for $r < \frac{\delta^2}{2}$, and this in a sense reflects the fact that

$$\frac{d^i x_1(0; \delta)}{dr^i} = O(\delta^{-(2i-1)}) \quad (3.2-7)$$

For the above problem, the parameter δ (or δ^{-1}) reflects the degree of non-degeneracy of the problem. In addition, (3.2-1) and (3.2-7) show how the size of δ (or δ^{-1}) can effect the success of the extrapolation procedure.

3.2.2 Nature of the Trajectory in the Degenerate Case

The results of Theorem 3.1.4 and Corollary 3.1.1 rely heavily on the non-degeneracy assumption. However, the examination of situations where this does not hold has not attracted much attention to date. In subsection 3.2.1, an example (Ex. 3.2.1 with $\delta = 0$) is given in which the non-degeneracy condition is not satisfied and in which $\|x(r_k) - x^*\| = O(r_k^{1/2})$. In Mifflin (1972, 1975), it is shown that if the non-degeneracy condition is relaxed in Theorem 3.1.4, then, provided the problem is strictly convex, $\|x(r_k) - x^*\| \leq O(r_k^{1/2})$. Here, this result is improved by relaxing the convexity assumptions, and it is shown that $\|x(r_k) - x^*\| = O(r_k^{1/2})$ whenever x^* is a regular local solution of a degenerate problem (ICP). Further information is then derived about the nature of the convergence of $x(r)$ to x^* by deriving, in the degenerate case, the asymptotic form of the differential equations governing the trajectory as $r \rightarrow 0$.

Notation: Below, it will be convenient to denote a quantity of the form $A(r_k) - A(0)$ by $\Delta^{(k)}(A)$ for any function $A(r)$.

It is instructive to consider the differences between the degenerate and non-degenerate cases in more detail. Initially, recall that if x^* is a regular local solution and $J_2(x^*) = \phi$, then Theorem 3.1.4 holds;

namely,

$$J_0(\tilde{x}^*) = J_1(\tilde{x}^*) \quad , \quad \|\Delta^{(k)}(\tilde{x})\| = o(r) \quad . \quad (3.2-8)$$

In addition, from Remark 3.1.2d,

$$\frac{r_k}{\nabla g_i(\tilde{x}^*)^T \Delta^{(k)}(\tilde{x})} \rightarrow u_i^* > 0 \quad , \quad \forall i \in J_1(\tilde{x}^*) \quad . \quad (3.2-9)$$

Hence

$$\theta_i = \lim_{r_k \rightarrow 0} \cos^{-1} \left[\frac{|\nabla g_i(\tilde{x}^*)^T \Delta^{(k)}(\tilde{x})|}{\|\nabla g_i(\tilde{x}^*)\| \cdot \|\Delta^{(k)}(\tilde{x})\|} \right] < \frac{\pi}{2} \quad . \quad (3.2-10)$$

Now assume that $J_2(\tilde{x}^*) \neq \emptyset$. It follows that

$$\lim_{r_k \rightarrow 0} \frac{r_k}{g_i(\tilde{x}^{(k)})} = 0 \quad , \quad \forall i \in J_2(\tilde{x}^*) \quad , \quad (3.2-11)$$

where $\tilde{x}^{(k)}$ denotes $\tilde{x}(r_k)$. Since

$$g_i(\tilde{x}^{(k)}) = \nabla g_i(\tilde{x}^*)^T \Delta^{(k)}(\tilde{x}) + o(\|\Delta^{(k)}(\tilde{x})\|) \quad , \quad (3.2-12)$$

it follows that

$$r_k = o(\nabla g_i(\tilde{x}^*)^T \Delta^{(k)}(\tilde{x}) + o(\|\Delta^{(k)}(\tilde{x})\|)) \quad , \quad i \in J_2(\tilde{x}^*) \quad . \quad (3.2-13)$$

Thus, making use of the Cauchy-Schwartz inequality, we obtain

$$r_k = o(\|\Delta^{(k)}(\tilde{x})\|) \quad . \quad (3.2-14)$$

(3.2-9) then gives that

namely,

$$J_0(\tilde{x}^*) = J_1(\tilde{x}^*) \quad , \quad \|\Delta^{(k)}(\tilde{x})\| = o(r) \quad . \quad (3.2-8)$$

In addition, from Remark 3.1.2d,

$$\frac{r_k}{\nabla g_i(\tilde{x}^*)^T \Delta^{(k)}(\tilde{x})} \rightarrow u_i^* > 0 \quad , \quad \forall i \in J_1(\tilde{x}^*) \quad . \quad (3.2-9)$$

Hence

$$\theta_i = \lim_{r_k \rightarrow 0} \cos^{-1} \left[\frac{|\nabla g_i(\tilde{x}^*)^T \Delta^{(k)}(\tilde{x})|}{\|\nabla g_i(\tilde{x}^*)\| \cdot \|\Delta^{(k)}(\tilde{x})\|} \right] < \frac{\pi}{2} \quad . \quad (3.2-10)$$

Now assume that $J_2(\tilde{x}^*) \neq \emptyset$. It follows that

$$\lim_{r_k \rightarrow 0} \frac{r_k}{g_i(\tilde{x}^{(k)})} = 0 \quad , \quad \forall i \in J_2(\tilde{x}^*) \quad , \quad (3.2-11)$$

where $\tilde{x}^{(k)}$ denotes $\tilde{x}(r_k)$. Since

$$g_i(\tilde{x}^{(k)}) = \nabla g_i(\tilde{x}^*)^T \Delta^{(k)}(\tilde{x}) + o(\|\Delta^{(k)}(\tilde{x})\|) \quad , \quad (3.2-12)$$

it follows that

$$r_k = o(\nabla g_i(\tilde{x}^*)^T \Delta^{(k)}(\tilde{x}) + o(\|\Delta^{(k)}(\tilde{x})\|)) \quad , \quad i \in J_2(\tilde{x}^*) \quad . \quad (3.2-13)$$

Thus, making use of the Cauchy-Schwartz inequality, we obtain

$$r_k = o(\|\Delta^{(k)}(\tilde{x})\|) \quad . \quad (3.2-14)$$

(3.2-9) then gives that

$$\frac{|\nabla g_i(x^*)^T \Delta^{(k)}(x)|}{\|\nabla g_i(x^*)\| \cdot \|\Delta^{(k)}(x)\|} \rightarrow 0, \quad i \in J_1(x^*), \quad (3.2-15)$$

and hence that $\theta_i = \frac{\pi}{2}$. Thus the sequence $\{x^{(k)}\}$ approaches x^* along a trajectory that is tangential to the constraint surfaces $g_i(x) = 0$, $i \in J_1(x^*)$. If $\Delta^{(k)}(x)$ is decomposed into $\|\Delta^{(k)}(x)\|(s_k + t_k)$, where s_k is a linear combination of the $\nabla g_i(x^*)$, $i \in J_1(x^*)$, and t_k lies in the orthogonal complement of this set, then $\|t_k\| \rightarrow 1$ and $\|s_k\| \rightarrow 0$. It is convenient to formalise this result as the following Lemma.

Lemma 3.2.1 (cf., Theorem 2.2.2)

Let x^* be a regular local solution of an ICP, and $J_2(x^*) \neq \emptyset$, and

$$\Delta^{(k)}(x) = \|\Delta^{(k)}(x)\|(s_k + t_k),$$

where s_k is a linear combination of the $\nabla g_i(x^*)$, $i \in J_1(x^*)$, and t_k is in the orthogonal complement of this set, then

(i) $\exists \alpha > 0$ such that $1 \geq \|t_k\| \geq \alpha$ for k sufficiently large,

and

(ii) $\|s_k\| = o(1)$ as $r_k \rightarrow 0$. □

We are now in a position to prove the main result of this subsection.

Theorem 3.2.1

Let x^* be a regular local solution of an ICP, and $J_2(x^*) \neq \emptyset$, then we have the asymptotic inequalities

$$\sqrt{\frac{|J_2(x^*)|}{A}} \cdot r_k^{\frac{1}{2}} + o(r_k^{\frac{1}{2}}) \leq \|\Delta^{(k)}(x)\| \leq \sqrt{\frac{|J_2(x^*)|}{a\alpha}} r_k^{\frac{1}{2}} + o(r_k^{\frac{1}{2}}), \quad (3.2-16)$$

with a , Λ and α given by (2.1-19), (2.1-20) and Lemma 3.2.1.

Proof

The necessary condition for a minimum of $B(\tilde{x}, r_k)$ at $\tilde{x}^{(k)}$ gives

$$\nabla f(\tilde{x}^{(k)}) - \sum_{i \in J} u_i^{(k)} \cdot \nabla g_i(\tilde{x}^{(k)}) = 0 \quad (3.2-17)$$

Subtracting $\nabla_x L_0(\tilde{x}^*, \tilde{u}^*)$ from this system yields

$$\nabla_x^2 L_0(\tilde{x}^*, \tilde{u}^*) \cdot \Delta^{(k)}(\tilde{x}) - \sum_{i \in J} \Delta^{(k)}(u_i) \cdot \nabla g_i(\tilde{x}^{(k)}) = o(\|\Delta^{(k)}(\tilde{x})\|) \quad (3.2-18)$$

Multiplying this equation by $\Delta^{(k)}(\tilde{x})^T$ gives

$$\Delta^{(k)}(\tilde{x})^T \nabla_x^2 L_0(\tilde{x}^*, \tilde{u}^*) \cdot \Delta^{(k)}(\tilde{x}) - \sum_{i \in J} \Delta^{(k)}(u_i) \cdot \Delta^{(k)}(g_i) = o(\|\Delta^{(k)}(\tilde{x})\|^2), \quad (3.2-19)$$

so that

$$\begin{aligned} & \Delta^{(k)}(\tilde{x})^T \nabla_x^2 L_0(\tilde{x}^*, \tilde{u}^*) \Delta^{(k)}(\tilde{x}) - \sum_{i \in J_1(\tilde{x}^*)} \Delta^{(k)}(u_i) g_i(\tilde{x}^{(k)}) \\ & - \sum_{i \in J_2(\tilde{x}^*)} u_i^{(k)} g_i(\tilde{x}^{(k)}) \\ & - \sum_{i \in J_3(\tilde{x}^*)} \frac{r_k}{g_i(\tilde{x}^{(k)})} \cdot \Delta^{(k)}(g_i) = o(\|\Delta^{(k)}(\tilde{x})\|^2) \quad (3.2-20) \end{aligned}$$

From (3.2-20) and Lemma 3.2.1, it follows that

$$\|\Delta^{(k)}(\tilde{x})\|^2 (t_k^T \nabla_x^2 L_0(\tilde{x}^*, \tilde{u}^*) t_k + o(1)) = r_k (|J_2(\tilde{x}^*)| + o(1)), \quad (3.2-21)$$

where use has been made of the facts that $g_i(\tilde{x}^{(k)}) = o(r_k)$ and

$\Lambda^{(k)}(u_i) > 0$, $\forall i \in J_1(\tilde{x}^*)$; $r_k = u_i^{(k)} g_i(\tilde{x}^{(k)})$, $\forall i \in J_2(\tilde{x}^*)$; and that $g_i(\tilde{x}^{(k)}) > 0$ and $\Delta^{(k)}(g_i) = O(\|\Delta^{(k)}(\tilde{x})\|)$, $\forall i \in J_3(\tilde{x}^*)$. Thus, the required result in (3.2-16) follows by two easy estimates using the second order sufficiency condition, Remark 2.1.5b and Lemma 3.2.1.

In the remainder of this subsection, an analysis of the trajectory $\tilde{x}(r)$ as $r \rightarrow 0$ is presented in order to complement the rate of convergence result of Theorem 3.2.1.

Let $\tilde{x}^*$ be a regular local solution of an ICP, and $J_2(\tilde{x}^*) \neq \emptyset$. For r_k sufficiently small, it follows from Theorem 3.1.3 that $\nabla_x^2 B(\tilde{x}^{(k)}, r_k)$ is positive definite. Therefore there exists $\delta > 0$ such that, by the implicit function theorem, the system of equations $\nabla_x B(\tilde{x}, r) = 0$ defines a smooth trajectory $\tilde{x} = \tilde{x}(r)$ on the open interval $0 < r < \delta$. The trajectory $(\tilde{x}(r), \tilde{u}(r))$ satisfies the system of differential equations

$$J(r) \begin{bmatrix} \frac{d\tilde{x}(r)}{dr} \\ \frac{d\tilde{u}(r)}{dr} \end{bmatrix} = \begin{bmatrix} 0 \\ e \end{bmatrix}, \quad (3.2-22)$$

and, by assumption, $\lim_{r \rightarrow 0} (\tilde{x}(r), \tilde{u}(r)) = (\tilde{x}^*, \tilde{u}^*)$.

Lemma 3.2.2

Let the possible partitions of $J_2(\tilde{x}^*)$ into two disjoint sets be ordered with respect to an index i such that

$$E_i \cup F_i = J_2(\tilde{x}^*), \quad i = 1, 2, \dots, \sigma. \quad (3.2-23)$$

Then

$$\det(J(r)) = \sum_{i=1}^{\sigma} \Lambda_i(r) \prod_{j \in E_i} u_j(r) \prod_{j \in F_i} g_j(\tilde{x}(r)), \quad (3.2-24)$$

where (writing $\tilde{x}$ for $\tilde{x}(r)$ and $\tilde{u}$ for $\tilde{u}(r)$)

$$A_{\lambda}(r) = \theta_{\lambda} \det \begin{bmatrix} \nabla_{\tilde{x}}^2 L_0(\tilde{x}, \tilde{u}) & [-\nabla g_j(\tilde{x}), j \notin J_2(\tilde{x}^*)] & [-\nabla g_j(\tilde{x}), j \in E_{\lambda}] \\ [u_j \nabla g_j(\tilde{x})^T, j \notin J_2(\tilde{x}^*)] & [\text{diag } g_j(\tilde{x}), j \notin J_2(\tilde{x}^*)] & 0 \\ [\nabla g_j(\tilde{x})^T, j \in E_{\lambda}] & 0 & 0 \end{bmatrix}, \quad (3.2-25)$$

$\theta_{\lambda} = \pm 1$, and $A_{\lambda}(0)$ is non-zero, $\lambda = 1, 2, \dots, \sigma$.

Proof

Let $j \in J_2(\tilde{x}^*)$. Then, by suitable row and column interchanges, it follows that

$$\theta \det(J(r)) = \det \left(\begin{array}{cc|c} & & -\nabla g_j \\ & M & 0 \\ \hline u_j \nabla g_j^T & 0 & g_j \end{array} \right) \quad (3.2-26)$$

$$= u_j \det \left(\begin{array}{cc|c} & & -\nabla g_j \\ & M & 0 \\ \hline \nabla g_j^T & 0 & 0 \end{array} \right) + g_j \det(M), \quad (3.2-27)$$

where $\theta = \pm 1$. The result in (3.2-24) now follows by repeated applications of this device. That $A_{\lambda}(0)$ is non-zero follows by the argument used to establish Lemma 2.2.1. □

Corollary 3.2.1

$J(r)$ is non-singular for sufficiently small $r > 0$.

□

The trajectory $(\tilde{x}(r), \tilde{u}(r))$ must satisfy the system of differential equations (cf., (3.1.23))

$$J(r) \begin{bmatrix} \frac{d\tilde{x}(r)}{dr} \\ \frac{d\tilde{u}(r)}{dr} \end{bmatrix} = \begin{bmatrix} 0 \\ e \end{bmatrix}. \quad (3.2-28)$$

When $J_2(\tilde{x}^*) \neq 0$, it follows from (3.2-24) that this system becomes singular as $r \rightarrow 0$. The aim here is to show how this singular behaviour can be isolated. Eqn. (3.2-28) can be rewritten explicitly (writing $\tilde{x}$ for $\tilde{x}(r)$ and J for $J(r)$, etc.) as

$$\begin{bmatrix} \frac{d\tilde{x}}{dr} \\ \frac{d\tilde{u}}{dr} \end{bmatrix} = \frac{\text{adjoint}(J)}{\det(J)} \begin{bmatrix} 0 \\ e \end{bmatrix} = \frac{1}{\det(J)} \begin{bmatrix} \vdots \\ \sum_{\alpha=n+1}^{n+m} (-1)^{\alpha+\beta} |MJ_{\alpha\beta}| \\ \vdots \end{bmatrix}, \quad (3.2-29)$$

where $|MJ_{\alpha\beta}|$ denotes the determinant of the minor of J obtained by eliminating the row associated with constraint α and the β th column. By considering the form of expansion leading to (3.2-24), it will be seen that, for small r , the dominant terms on the RHS of (3.2-29) are those associated with $\alpha \in J_2(\tilde{x}^*)$. Attention is concentrated on the equations corresponding to $\beta \in J_2(\tilde{x}^*)$. To evaluate $|MJ_{\alpha\beta}|$, it is necessary to consider two cases:

- (i) $\alpha = \beta$: Let $J_\beta = J_2(x^*) - \{\beta\}$, and order all possible partitions of J_β into disjoint sets E_i^β and F_i^β , $E_i^\beta \cup F_i^\beta = J_\beta$, with $1 \leq i \leq \sigma_1$. Note that, for each i , $\exists j(i, \beta)$, which depends on i and β , such that

$$E_i^\beta = E_{j(i, \beta)} \quad (3.2-30)$$

This identification allows $|MJ_{\beta\beta}|$ to be expressed in the form

$$|MJ_{\beta\beta}| = \sum_{i=1}^{\sigma_1} \Lambda_{j(i, \beta)} \prod_{k \in E_i^\beta} u_k \prod_{k \in F_i^\beta} g_k \quad (3.2-31)$$

- (ii) $\alpha \neq \beta$: Let $J_{\alpha\beta} = J_2(x^*) - \{\alpha, \beta\}$, and order all possible partitions of $J_{\alpha\beta}$ into disjoint sets $E_i^{\alpha\beta}$ and $F_i^{\alpha\beta}$, $E_i^{\alpha\beta} \cup F_i^{\alpha\beta} = J_{\alpha\beta}$, with $1 \leq i \leq \sigma_2$. In this case the expansion procedure gives

$$|MJ_{\alpha\beta}| = u_\beta \sum_{i=1}^{\sigma_2} \Lambda_i^{\alpha\beta} \prod_{k \in E_i^{\alpha\beta}} u_k \prod_{k \in F_i^{\alpha\beta}} g_k, \quad (3.2-32)$$

where $\Lambda_i^{\alpha\beta}$ has the form

$$\Lambda_i^{\alpha\beta} = \theta_i \det \begin{bmatrix} \nabla_x^2 L_0 & [-\nabla g_j, j \notin J_2(x^*)] & -\nabla g_\alpha & [-\nabla g_j, j \in E_i^{\alpha\beta}] \\ [u_j \nabla g_j^T, j \notin J_2(x^*)] & [\text{diag } g_j, j \notin J_2(x^*)] & 0 & 0 \\ \nabla g_\beta^T & 0 & 0 & 0 \\ [\nabla g_j^T, j \in E_i^{\alpha\beta}] & 0 & 0 & 0 \end{bmatrix} \quad (3.2-33)$$

and $\theta_i = \pm 1$. In this case, it is not necessarily true that

$$\Lambda_i^{\alpha\beta}(0) \neq 0.$$

It can now be shown that asymptotically, as $r \rightarrow 0$, the system of differential equations (3.2-29) can be separated into a set of differential equations corresponding to degenerate constraints, and another set of differential equations which do not. The behaviour of the trajectory is governed asymptotically by the first set of differential equations. To see this, make the substitution $r/u_{\lambda}(r) = g_{\lambda}(\tilde{x}(r))$, $\lambda \in J$. Then, as $r \rightarrow 0$, the expansion for $\det(J(r))$ in (3.2-24) becomes

$$\det(J(r)) = \sum_{\lambda=1}^{\sigma} (\Lambda_{\lambda}(0) + o(1)) \prod_{j \in E_{\lambda}} u_j(r) \prod_{j \in F_{\lambda}} \frac{r}{u_j(r)}. \quad (3.2-34)$$

A similar asymptotic form can be obtained for each $|MJ_{\alpha\beta}|$. This means that, asymptotically, the right-hand side of (3.2-29) only depends on $\tilde{x}^*, u^*, r$ and $u_{\lambda}(r)$, $\lambda \in J_2(\tilde{x}^*)$. The differential equations corresponding to the degenerate constraints are the controlling equations in the sense that: (i) they are asymptotically independent of $\tilde{x}(r)$ and $u_{\lambda}(r)$, $\lambda \notin J_2(\tilde{x}^*)$; (ii) if they are solved to obtain the asymptotic behaviour of $u_{\lambda}(r)$, $\lambda \in J_2(\tilde{x}^*)$, then, by substituting back into the other set of differential equations, formulas are obtained for the asymptotic behaviour of $\frac{d\tilde{x}(r)}{dr}$ and $\frac{du_{\lambda}(r)}{dr}$, $\lambda \notin J_2(\tilde{x}^*)$, from which $\tilde{x}(r)$ and $u_{\lambda}(r)$, $\lambda \notin J_2(\tilde{x}^*)$ can be obtained using asymptotic integration (see Copson (1965)).

The controlling differential equations are non-linear and are generally difficult to solve. This system of differential equations is initially solved for the special case when only one constraint of the given problem is degenerate. A result for the convergence of $\tilde{x}(r)$ to $\tilde{x}^*$, which is sharper than (3.2-16), is obtained.

Theorem 3.2.2

Let $\tilde{x}^*$ be a regular local solution of a degenerate ICP of which only one constraint is degenerate, then there exist explicit constants b^* ,

$\mu^* > 0$ such that, as $r \rightarrow 0$,

$$u_j(r) \sim \left(\frac{r}{b^*}\right)^{\frac{1}{2}} \text{ for the degenerate constraint } g_j,$$

and

$$\|\tilde{x}(r) - \tilde{x}^*\| \sim \mu^* r^{\frac{1}{2}},$$

when using the classical barrier function method.

Proof

Without loss of generality, it is assumed that all the constraints are active and that the last constraint, g_m , is degenerate. The minimum of $B(\tilde{x}, r)$ with respect to $\tilde{x}$ defines a trajectory with r as the independent variable. The trajectory satisfies a system of singular differential equations (again writing $\tilde{x}$ for $\tilde{x}(r)$, J for $J(r)$, etc.)

$$\begin{bmatrix} \frac{d\tilde{x}}{dr} \\ \frac{d\tilde{u}}{dr} \end{bmatrix} = \frac{1}{\det(J)} \begin{bmatrix} \vdots \\ \sum_{\alpha=m+1}^{n+m} (-1)^{\alpha+\beta} |MJ_{\alpha\beta}| \\ \vdots \end{bmatrix}, \quad (3.2-35)$$

where the initial condition is $(\tilde{x}(0), \tilde{u}(0)) = (\tilde{x}^*, \tilde{u}^*)$. We recall that the Jacobian matrix, $J(r)$, has the form

$$J(r) = \begin{bmatrix} \nabla_{\tilde{x}}^2 L_0(\tilde{x}, \tilde{u}) & -\nabla g_1 & \dots & -\nabla g_m \\ u_1 \nabla g_1^T & g_1 & & \\ \vdots & & \ddots & \\ u_m \nabla g_m^T & & & g_m \end{bmatrix}. \quad (3.2-36)$$

From (3.2-24), (3.2-27) and (3.2-34), it follows that

$$\det(J) = u_m \det(B) + \frac{r}{u_m} |MJ_{m+n, m+n}|, \quad (3.2-37)$$

where

$$B = \begin{bmatrix} V_x^2 L_0(x, u) & -\nabla g_1 & \dots & -\nabla g_{m-1} & -\nabla g_m \\ u_1 \nabla g_1^T & g_1 & & & \\ \vdots & & \ddots & & \\ u_{m-1} \nabla g_{m-1}^T & & & g_{m-1} & \\ \nabla g_m^T & & & & 0 \end{bmatrix}. \quad (3.2-38)$$

Note that

$$\det(B) = \sum_{i=1}^n \frac{\partial g_m}{\partial x_i} (-1)^{i+m+n} |MJ_{m+n, i}|. \quad (3.2-39)$$

Let $|MB_{i,j}|$ denote the determinant of the minor of B obtained by eliminating the i th row and the j th column. Then, comparing (3.2-36) and (3.2-38), it follows that

$$|MJ_{i, m+n}| = u_m |MB_{i, m+n}|, \quad \forall i \neq m+n, \quad (3.2-40)$$

and

$$|MJ_{m+n, j}| = |MB_{m+n, j}|, \quad \forall j. \quad (3.2-41)$$

From the preceding discussion, the controlling differential equation is the last equation of (3.2-35) - the differential equation corresponding to the degenerate constraint g_m . Using (3.2-37) and (3.2-40) this differential equation becomes

$$\frac{du_m}{dr} = \frac{u_m(1+cu_m)}{r+bu_m^2}, \quad (3.2-42)$$

where

$$b = \det(B) / |MJ_{m+n, m+n}|, \quad (3.2-43)$$

and

$$c = \frac{1}{|MJ_{m+n, m+n}|} \sum_{i=n+1}^{m+n-1} (-1)^{m+n+i} |MB_{i, m+n}|. \quad (3.2-44)$$

As $r \rightarrow 0$, $b \rightarrow b^*$, $c \rightarrow c^*$ and $u_m \rightarrow 0$. Because $u_m \rightarrow 0$, it follows that $cu_m = o(1)$, and this confirms the earlier assertion that the dominant terms in $\sum_{\alpha=n+1}^{n+m} (-1)^{\alpha+\beta} |MJ_{\alpha\beta}|$ are the ones corresponding to the degenerate constraints.

To solve the differential equation (3.2-42), consider the equivalent system

$$\frac{du_m}{dr} = -u_m - u_m^2(c^* + o(1)) \quad (3.2-45a)$$

$$\frac{dr}{dr} = -r - u_m^2(b^* + o(1)) \quad (3.2-45b)$$

A result from the theory of ordinary differential equation (Hartman (1973), Theorems 3.5 and 3.6, p.218) implies that: If $(u_m(t), r(t)) \rightarrow (0, 0)$, then $t \rightarrow \infty$ and there exist uniquely determined constants a_1 and a_2 , at least one of which is not zero, such that

$$r(t) = e^{-t} \{a_1 + o(1)\} \quad (3.2-46a)$$

$$u_m(t) = e^{-t} \{a_2 + o(1)\}. \quad (3.2-46b)$$

Thus

$$\frac{u_m(t)}{r(t)} \rightarrow \frac{a_2}{a_1} \quad \text{if } a_1 \neq 0. \quad (3.2-47)$$

However, this is excluded, because $u_m/r = 1/g_m \rightarrow +\infty$.

Hence

$$a_1 = 0 \quad \text{and} \quad \frac{r(t)}{u_m(t)} \rightarrow 0. \quad (3.2-48)$$

Using the variation of constants formula, it follows that

$$r(t) = \int_t^\infty e^{-(t-s)} b(u_m(s)) u_m^2(s) ds. \quad (3.2-49)$$

Hence, using L'Hospital's rule,

$$\begin{aligned} \lim_{t \rightarrow \infty} \frac{e^t r(t)}{u_m(t)} &= \lim_{t \rightarrow \infty} \frac{e^t u_m^2(t) (b^{*+0}(1))}{u_m(t) + u_m^2(t) (c^{*+0}(1))} \\ &= a_2 b^*. \end{aligned} \quad (3.2-50)$$

Therefore

$$\begin{aligned} \lim_{t \rightarrow \infty} \frac{r(t)}{u_m^2(t)} &= \frac{1}{a_2} \lim_{t \rightarrow \infty} \frac{e^t r(t)}{u_m(t)} \\ &= b^*. \end{aligned} \quad (3.2-51)$$

Thus

$$u_m^2(r) \sim \frac{1}{b^*} \cdot r. \quad (3.2-52)$$

The fact that $u_m(r) \geq 0$ and $r \geq 0$ implies that $b^* \geq 0$. (3.2-43) and (3.2-25) imply that $b^* \neq 0$, and hence $b^* > 0$. For $i \leq j \leq n$, consider the j th equation of (3.2-35)

$$\frac{dx_j}{dr} = \frac{1}{\det(J)} \sum_{i=n+1}^{n+m} (-1)^{i+j} |MJ_{i,j}| \quad (3.2-53)$$

Again, using (3.2-37), (3.2-40) and (3.2-43), Eqn. (3.2-53) becomes

$$\frac{dx_j}{dr} = \frac{u_m}{r+bu_m^2} (a_j + d_j) \quad (3.2-54)$$

where

$$a_j = (-1)^{m+n+j} |MJ_{m+n,j}| / |MJ_{m+n,m+n}| \quad (3.2-55)$$

and

$$d_j = \frac{1}{|MJ_{m+n,m+n}|} \sum_{i=n+1}^{m+n-1} (-1)^{i+j} |MJ_{i,j}| \quad (3.2-56)$$

As $r \rightarrow 0$, $a_j \rightarrow a_j^*$, and $d_j \rightarrow 0$ (because $u_m \rightarrow 0$, $g_m \rightarrow 0$). By substituting $u_m \sim \left(\frac{r}{b^*}\right)^{1/2}$ into (3.2-54), this gives

$$\frac{dx_j}{dr} \sim \frac{r^{-1/2}}{2\sqrt{b^*}} (a_j^* + o(1)) \quad (3.2-57)$$

Using asymptotic integration (Copson (1965)), it follows that

$$x_j(r) \sim x_j^* + \frac{a_j^*}{\sqrt{b^*}} \cdot r^{1/2} \quad (3.2-58)$$

Therefore

$$\|x(r) - x^*\| \sim \frac{1}{\sqrt{b^*}} \left[\sum_{j=1}^n (a_j^*)^2 \right]^{1/2} \cdot r^{1/2} \quad (3.2-59)$$

From (3.2-39), (3.2-43) and (3.2-55), it follows that

$$\sum_{j=1}^n \frac{\partial g_m}{\partial x_j} a_j^* = b^* \neq 0 \quad (3.2-60)$$

Hence, there is at least one $a_j^* \neq 0$ such that

$$\mu^* = \frac{1}{\sqrt{b^*}} \left[\sum_{j=1}^n (a_j^*)^2 \right]^{1/2} > 0, \quad (3.2-61)$$

and

$$\|x(r) - x^*\| \sim \mu^* r^{1/2}. \quad \square$$

Theorem 3.2.3

Let $\tilde{x}^*$ be a regular local solution of a degenerate ICP. If, using the classical logarithmic barrier function method, as $r \rightarrow 0$,

$$u_\alpha(r) = c_\alpha r^q + o(r^q), \quad \forall \alpha \in J_2(\tilde{x}^*),$$

then $q = 1/2$.

Proof

Note that $u_\alpha(r) = r/g_\alpha(x(r)) \rightarrow 0$ as $r \rightarrow 0$ implies that $q < 1$. Substituting u_α , $\alpha \in J_2(\tilde{x}^*)$, into the equation for $\frac{du_\beta}{dr}$ and using (3.2-24), (3.2-29), (3.2-31) and (3.2-32) yields

$$\frac{du_\beta}{dr} = \frac{c}{r^q} + \text{small terms in } r, \quad \beta \in J_2(\tilde{x}^*),$$

where c is a non-zero constant. The result now follows. $\square$

Remark 3.2.1

If it is assumed that $\forall \alpha \in J_2(\tilde{x}^*)$, as $r \rightarrow 0$,

$$u_\alpha(r) = c_\alpha r^{q_\alpha} + o(r^{q_\alpha}), \quad (3.2-62)$$

with $c_\alpha \neq 0$ and $0 < q_\alpha < 1$. Using asymptotic differentiation (Copson (1965)), it follows that

$$\frac{du_{\alpha}(r)}{dr} = c_{\alpha} q_{\alpha} r^{q_{\alpha}-1} + o(r^{q_{\alpha}-1}), \quad (3.2-63)$$

and

$$g_{\alpha}(r) = \frac{1}{c_{\alpha}} r^{1-q_{\alpha}} + o(r^{1-q_{\alpha}}). \quad (3.2-64)$$

By substituting these expressions for $u_{\alpha}(r)$, $du_{\alpha}(r)/dr$ and $g_{\alpha}(r)$ into the set of differential equations corresponding to the degenerate constraints (a subsystem of (3.2-29)), and using (3.2-24), (3.2-31) and (3.2-32) a set of $|J_2(x^*)|$ asymptotic equations is obtained. Using asymptotic matching technique (see Bellman (1970), p.39) to match, as $r \rightarrow 0$ (or $1/r \rightarrow \infty$), both the coefficients and the order of terms in each equation, it is possible to show that (the proof is tedious and is omitted) $q_{\alpha} = \frac{1}{2}$, $\forall \alpha \in J_2(\tilde{x}^*)$. The coefficients, c_{α} , $\alpha \in J_2(\tilde{x}^*)$, can be determined. Then, similar to the proof of Theorem 3.2.2, a constant μ^* can be founded such that $\|\tilde{x}(r) - \tilde{x}^*\| \sim \mu^* r^{\frac{1}{2}}$, by substituting for $u_{\alpha}(r)$, $\alpha \in J_2(\tilde{x}^*)$, in the first n differential equations of the system (3.2-29) and using asymptotic integration. Note that, in (3.2-21), this means that, asymptotically, the quantity $t_k^T \nabla_x^2 L_0(\tilde{x}^*, \tilde{u}^*) t_k$ tends to the constant $|J_2(\tilde{x}^*)|/(\mu^*)^2$.

3.2.3 A Measure of Degeneracy

In this subsection, the aim is to introduce a measure of non-degeneracy of an ICP. Since, for a non-degenerate problem, $u_{\lambda}^* > 0$, $\forall \lambda \in J_0(x^*)$, a possible first guess at an appropriate measure is $\max_{i,j} u_{\lambda}^*/u_j^*$, $i, j \in J_0(\tilde{x}^*)$. However, this can be changed arbitrarily by multiplying the constraints by suitably chosen positive numbers. A more suitable choice which is independent of this kind of rescaling of the constraints is

$$\gamma = \max_{i,j} \frac{u_{\lambda}^* \|\nabla g_i(\tilde{x}^*)\|}{u_j^* \|\nabla g_j(\tilde{x}^*)\|}, \quad i, j \in J_0(\tilde{x}^*). \quad (3.2-65)$$

Clearly $\gamma \geq 1$, and γ is unbounded as the problem becomes degenerate. Also the results of the previous subsection show that, in the degenerate case, the derivatives $\frac{dx}{dr}$ and $\frac{du}{dr}$ are unbounded as $r \rightarrow 0$, because of the dependence on $r^{1/2}$. Thus it is natural to ask if there is a link between these two types of behaviour. The main result confirms that such a link exists and shows that the growth of $\frac{d^i x}{dr^i}$ and $\frac{d^i u}{dr^i}$ as i increases, for fixed sufficiently small r , is simply related to γ . To obtain this result, it is convenient first to rescale the problem as follows:

Let

$$\beta = \max_{i \in J_0(x^*)} (u_i^* \|\nabla g_i(x^*)\|) , \quad (3.2-66)$$

and consider

$$\hat{B}(\tilde{x}, p) = F(\tilde{x}) - p \sum_{i \in J} \log [G_i(\tilde{x})] , \quad (3.2-67)$$

where

$$p = r/\beta , \quad (3.2-68)$$

$$F(\tilde{x}) = f(\tilde{x})/\beta , \quad (3.2-69)$$

and

$$G_i(\tilde{x}) = g_i(\tilde{x})/\|\nabla g_i(x^*)\| \quad i \in J . \quad (3.2-70)$$

It follows from construction that

$$\nabla_{\tilde{x}} \hat{B}(\tilde{x}, p) = \frac{1}{\beta} \nabla_{\tilde{x}} B(\tilde{x}, r) , \quad (3.2-71)$$

so that $\hat{B}$ has the same stationary points as B and the rescaled problem has the same value of γ . The Lagrange multipliers for the rescaled problem are denoted by $\hat{u}_i$, $i \in J$, and the classical Lagrangian by $\hat{L}_0(\tilde{x}, \hat{u})$.

Thus

$$\gamma = \max_{\lambda \in J_0(\tilde{x}^*)} \frac{1}{\hat{u}_\lambda^*} . \quad (3.2-72)$$

Before constructing the trajectory for the rescaled problem, it is convenient to state a preliminary Lemma.

Lemma 3.2.3

Let $\tilde{x}^*$ be a regular local solution of the rescaled problem and let the matrix K be given by

$$K = \begin{bmatrix} \nabla_x^2 \hat{L}_0(\tilde{x}^*, \hat{u}^*) & [-\nabla G_\lambda(\tilde{x}^*), \lambda \in J_0(\tilde{x}^*)] & [-\nabla G_\lambda(\tilde{x}^*), \lambda \notin J_0(\tilde{x}^*)] \\ [\nabla G_\lambda(\tilde{x}^*)^T, \lambda \in J_0(\tilde{x}^*)] & 0 & 0 \\ 0 & 0 & [\text{diag } G_\lambda(\tilde{x}^*), \lambda \notin J_0(\tilde{x}^*)] \end{bmatrix} . \quad (3.2-73)$$

Then K is non-singular.

Proof

The result follows from the arguments used to establish Lemma 2.2.1. $\square$

Remark 3.2.2

The significance of Lemma 3.2.3 follows from the fact that, in the non-degenerate case, $\hat{J}(0)$ (where $\hat{J}(p)$ is the Jacobian of the rescaled problem) can be factored into the form:

$$\hat{J}(0) = \begin{bmatrix} I_n & & \\ & [\text{diag } \hat{u}_\lambda^*, \lambda \in J_0(\tilde{x}^*)] & \\ & & I_{J_3(\tilde{x}^*)} \end{bmatrix} K = QK , \quad (3.2-74)$$

where I_n and $I_{J_3(\tilde{x}^*)}$ are the $n \times n$ and $|J_3(\tilde{x}^*)| \times |J_3(\tilde{x}^*)|$ identity matrices, respectively. Consider a family of perturbations of the ICP which increases γ without bound but preserves the regular local solution property of the minimum. It is clear that the approach to degeneracy is captured by the diagonal matrix Q and that K remains non-singular provided the parameter a in the second order sufficiency condition is bounded away from zero (as it must if the minimum remains a regular local solution). Thus, an estimate of $\|\hat{J}(0)^{-1}\|$ of the form

$$\|\hat{J}(0)^{-1}\| = \kappa \frac{\gamma}{a}, \quad (3.2-75)$$

can be expected to hold, where κ is an (order 1) constant.

Remark 3.2.3

A convenient way to construct such a family of problems is to start with a degenerate problem having a regular local solution at $\tilde{x}^*$, and then to modify the objective function by adding a function $w(\tilde{x}, \tilde{\lambda})$ given by

$$w(\tilde{x}, \tilde{\lambda}) = \sum_{i \in J_2(\tilde{x}^*)} \lambda_i g_i(\tilde{x}), \quad \lambda_i \geq 0. \quad (3.2-76)$$

The effect of $w(\tilde{x}, \tilde{\lambda})$ is such that the multiplier u_i^* is unchanged for $i \notin J_2(\tilde{x}^*)$, and $u_i^* = \lambda_i$ for $i \in J_2(\tilde{x}^*)$. Clearly, the λ_i can be chosen such that γ takes any value ≥ 1 , while the contributions of $\nabla_{\tilde{x}}^2 w(\tilde{x}, \tilde{\lambda})$ and $\sum_{i \in J_2(\tilde{x}^*)} \lambda_i \nabla^2 g_i(\tilde{x})$ are such that $\nabla_{\tilde{x}}^2 L_0$ remains unchanged. Also, the second order sufficiency condition continues to hold with the same value of a as the set of allowable vectors $\tilde{t}$ in (2.1-19) is further constrained if any $\lambda_i > 0$ (i.e., the set $J_1(\tilde{x}^*)$ is enlarged). From this family, a sequence of problems can be chosen having the property that γ becomes unbounded while the regular local solution property holds uniformly.

Theorem 3.2.4

Assuming that the problem functions are suitably smooth, consider a family of problems indexed by a parameter α having the following properties:

- (i) for each α , $\tilde{x}^*$ is a regular local solution,
- (ii) there exists a constant $a > 0$ such that, for each α , the second order sufficiency condition (2.1-19) holds for this a , and
- (iii) $\lim_{\alpha \rightarrow \infty} \gamma(\alpha) = +\infty$,

then, for fixed sufficiently small p , the asymptotic estimates

$$\|D^{q+1}\tilde{x}_\alpha\| = O(\gamma(\alpha)^{2q+1}), \quad \|D^{q+1}\hat{\tilde{u}}_\alpha\| = O(\gamma(\alpha)^{2q+1}), \quad (3.2-77)$$

hold for $q = 0, 1, \dots$,

where $D \equiv \frac{d}{dp}$, and $\tilde{x}_\alpha(p)$, $\hat{\tilde{u}}_\alpha(p)$ are defined by minimizing $\hat{B}(\tilde{x}, p)$ in (3.2-67) for the α 'th member of the family.

Proof

It is convenient to drop the subscript α and to write $\tilde{z} = (\tilde{x}, \hat{\tilde{u}})$. From (3.1-23), it follows that

$$\hat{J} D\tilde{z} = Q K D\tilde{z} = \begin{bmatrix} 0 \\ \tilde{e} \end{bmatrix}, \quad (3.2-78)$$

so that, as K is non-singular by assumption,

$$\|D\tilde{z}\| = O(\|Q^{-1}\|) = O(\gamma), \quad (3.2-79)$$

and this establishes the result in the special case $q = 0$. In fact, the

estimate holds also for $DA(\underline{z})$ where Λ is any sufficiently smooth function of $\underline{z}$ for, by the chain rule,

$$|DA(\underline{z})| \leq \left\| \frac{\partial \Lambda(\underline{z})}{\partial \underline{z}} \right\| \|D\underline{z}\|, \quad (3.2-80)$$

where

$$\frac{\partial \Lambda(\underline{z})}{\partial \underline{z}} = \left(\frac{\partial \Lambda}{\partial z_1}, \frac{\partial \Lambda}{\partial z_2}, \dots \right). \quad (3.2-81)$$

The proof now follows by induction. We assume that

$$|D^j A(\underline{z})| = O(\gamma^{2j-1}), \quad j \leq q, \quad (3.2-82)$$

and show that this implies that $|D^{q+1} A(\underline{z})| = O(\gamma^{2q+1})$.

Differentiating (3.2-78) gives

$$\hat{J} D^{q+1} \underline{z} = - \sum_{j=1}^q \binom{q}{j} D^j \hat{J} D^{q+1-j} \underline{z}. \quad (3.2-83)$$

To estimate the order of each term on the right-hand side, we apply the induction hypothesis (3.2-82). This gives an order of

$$(2j-1) + (2q+2-2j-1) = 2q, \quad (3.2-84)$$

for each j . The estimate for $D^{q+1} \underline{z}$ follows as an extra order in γ must be counted for the inversion of $\hat{J}$ in (3.2-83). To complete the induction, it is necessary to show that (3.2-82) holds with $q := q+1$ as a consequence of the estimate established for $D^{q+1} \underline{z}$. We have

$$\begin{aligned} D^{q+1} A(\underline{z}) &= D^q \left[\frac{\partial \Lambda(\underline{z})^T}{\partial \underline{z}} \cdot D\underline{z} \right] \\ &= \sum_{i=0}^q \binom{q}{i} D^i \frac{\partial \Lambda(\underline{z})^T}{\partial \underline{z}} \cdot D^{q+1-i} \underline{z}. \end{aligned} \quad (3.2-85)$$

For $i = 0$, the corresponding term is the scalar product of $\frac{\partial A(\tilde{z})}{\partial \tilde{z}}$ and $D^{q+1}\tilde{z}$, so that it is necessarily of order $2q+1$ in γ . For $i > 0$, the order estimate is a consequence of (3.2-82) which gives

$$(2i-1) + 2(q+1-i) - 1 = 2q. \quad (3.2-86)$$

This is of smaller order than the term for $i = 0$, so that

$$|D^{q+1}A(\tilde{z})| = O(\gamma^{2q+1}). \quad (3.2-87)$$

Thus the induction hypothesis is verified and the theorem follows. $\square$

Remark 3.2.4

An appropriate formalism for developing expressions of the form (3.2-83) into their component parts has been developed by Butcher in a number of papers (for example, Butcher (1963)) for his work on Runge-Kutta methods. One deduction from this work is that the numerical coefficients hidden in the order estimates (3.2-77) are likely to grow like $(q+1)!$, while (3.2-75) and (3.2-78) suggest that a term $(1/a)^{q+1}$ is also likely to be present. It is convenient to assume a dependence of this form for the numerical coefficient. However, it is not critical to the development of our arguments.

3.2.4 The Numerical Performance of Extrapolation

In subsection 3.2.3, a measure of sensitivity is introduced which becomes large when the non-degeneracy assumption is close to violation. It is shown that this sensitivity measure is related to the growth, on the solution trajectory, of $\frac{d^i x}{dp^i}$ with respect to i for fixed sufficiently small p . With this information, it is possible to analyse the extrapolation procedure and to predict the number of stages of extrapolation which are useful.

If the results of subsection 3.2.3 are combined with (3.2-1) for the error in an extrapolation procedure, then it follows that an extrapolation based on the points $p = p_{k-\delta+1}, \dots, p_k$ is proportional to

$$e_{k,\delta} = \left[\prod_{i=k-\delta+1}^k p_i \right] \gamma^{2\delta-1}, \quad (3.2-88)$$

where, as a result of Remark 3.2.4, any explicit dependence of a coefficient on δ has been omitted. Any dependence on a has also been omitted as a 'second best' compromise because its estimation is relatively more difficult. The numerical situation considered is the one in which the successive minima $\tilde{x}(p_1), \dots, \tilde{x}(p_k)$ are available, and where the extrapolation has been applied to the points $p = p_{k-\delta+1}, \dots, p_k$. The question is examined whether it is worth-while to include data from the point $p = p_{k-\delta}$ in the extrapolation. As a basis for this decision, the following heuristic has been examined

$$\mu_{k,\delta} = \frac{e_{k,\delta+1}}{e_{k,\delta}} = p_{k-\delta} \gamma^2 < 0.1. \quad (3.2-89)$$

Hopefully, if the strategy is successful, and if the terms in $e_{k,\delta}$ are actually the most significant, then the new extrapolation would produce at least one more decimal place of accuracy. The performance in practise has been surprisingly successful, giving accurate predictions for a range of test problems including the 'easy' Rosen-Suzuki problem and a range of Colville test problems including the 'Shell dual' (Colville II) which is generally regarded as quite difficult and which proves to have a large value of γ .

The basic numerical data for the tabulated results is summarised in Table 3.2.1. The sequence of minimizations is designed to follow that of Fletcher and McCann (1969).

TEST PROBLEM	γ	β	p_i
Rosen-Suzuki	1.6	8	1.2×10^{-i}
Colville II	83	57	$.16 \times 10^{-i}$

Table 3.2.1: Basic data for the numerical results.

In Tables 3.2.2 and 3.2.3, values of $\max_{i \in J_0(\tilde{x}^*)} |g_i(\tilde{x})|$ are given for $\tilde{x}$ resulting from the extrapolation procedure. The minimizations have been carried out using a severe error tolerance to ensure that extraneous errors should not unduly perturb the results. It would seem that the results are accurate to at least 12 decimal places in both cases.

For the Rosen-Suzuki problem the heuristic is satisfied provided $k-s \geq 2$ suggesting that repeated extrapolation is justified on the set of points $p = p_2, \dots, p_k$, and this is borne out very well by Table 3.2.2. For Colville II the heuristic is satisfied provided the extrapolation is based on the points $p = p_4, \dots, p_k$, and the results in Table 3.2.3 again support this strategy.

Remark 3.2.5

Tables 3.2.2 and 3.2.3 both show that there is a tendency for the extrapolation to improve accuracy until a threshold is reached and from this point on this threshold is essentially preserved. This confirms that extrapolation with this choice of parameters is a very stable process. The threshold produced going down the table is a function of the accuracy of the successive minimizations and depends quite strongly on the accuracy of the final minimization (cf., Appendix 3).

$\begin{matrix} s \\ k \end{matrix}$	0	1	2	3	4	5	6	7	8	9
1	1.1									
2	1.1×10^{-1}	6.0×10^{-3}								
3	1.0×10^{-2}	1.4×10^{-3}	1.3×10^{-3}							
4	1.0×10^{-3}	1.9×10^{-5}	5.2×10^{-6}	3.8×10^{-6}						
5	1.0×10^{-4}	1.9×10^{-7}	5.9×10^{-9}	7.2×10^{-10}	3.4×10^{-10}					
6	1.0×10^{-5}	2.0×10^{-9}	7.2×10^{-12}	1.3×10^{-12}	1.2×10^{-12}	1.2×10^{-12}				
7	1.0×10^{-6}	2.1×10^{-11}	9.8×10^{-13}	9.8×10^{-13}	9.8×10^{-13}	9.8×10^{-13}	9.8×10^{-13}			
8	1.0×10^{-7}	3.0×10^{-13}	5.2×10^{-13}	5.2×10^{-13}	5.2×10^{-13}	5.2×10^{-13}	5.2×10^{-13}	5.2×10^{-13}		
9	1.0×10^{-8}	5.2×10^{-13}	5.2×10^{-13}	5.2×10^{-13}	5.2×10^{-13}	5.2×10^{-13}	5.2×10^{-13}	5.2×10^{-13}	5.2×10^{-13}	
10	1.0×10^{-9}	5.8×10^{-14}	6.2×10^{-14}	6.2×10^{-14}	6.2×10^{-14}	6.2×10^{-14}	6.2×10^{-14}	6.2×10^{-14}	6.2×10^{-14}	6.2×10^{-14}

Table 3.2.2: Numerical results for Rosen-Suzuki problem.

$\begin{smallmatrix} s \\ k \end{smallmatrix}$	0	1	2	3	4	5	6	7	8	9
1	4.2									
2	3.7×10^{-1}	1.6×10^{-1}								
3	4.3×10^{-2}	8.5×10^{-3}	7.0×10^{-3}							
4	4.4×10^{-3}	2.7×10^{-4}	1.9×10^{-4}	1.8×10^{-4}						
5	4.5×10^{-4}	4.4×10^{-6}	1.8×10^{-6}	1.6×10^{-6}	1.6×10^{-6}					
6	4.5×10^{-5}	4.8×10^{-8}	4.2×10^{-9}	2.4×10^{-9}	2.3×10^{-9}	2.3×10^{-9}				
7	4.5×10^{-6}	4.8×10^{-10}	1.0×10^{-11}	9.9×10^{-12}	9.8×10^{-12}	9.8×10^{-12}	9.8×10^{-12}			
8	4.5×10^{-7}	1.3×10^{-11}	8.7×10^{-12}	8.7×10^{-12}	8.7×10^{-12}	8.7×10^{-12}	8.7×10^{-12}	8.7×10^{-12}		
9	4.5×10^{-8}	7.4×10^{-12}	7.4×10^{-12}	7.4×10^{-12}	7.4×10^{-12}	7.4×10^{-12}	7.4×10^{-12}	7.4×10^{-12}	7.4×10^{-12}	
10	4.5×10^{-9}	4.1×10^{-12}	4.1×10^{-12}	4.1×10^{-12}	4.1×10^{-12}	4.1×10^{-12}	4.1×10^{-12}	4.1×10^{-12}	4.1×10^{-12}	4.1×10^{-12}

Table 3.2.3: Numerical results for Colville II.

Remark 3.2.6

The success of the heuristic for the Colville II problem suggests that near-degeneracy might well be the source of the numerical problems reported, and this raises the possibility that the difficulties could be the result of algorithmic shortcomings rather than a consequence of any particular difficulty inherent in the problem.

3.3 MODIFIED BARRIER FUNCTION METHODS3.3.1 Introduction

The results of subsection 3.2.2 show that degenerate constraints affect considerably the performance of the classical barrier function algorithm by forcing the solution trajectory against other active constraint surfaces $g_i(x) = 0$, $i \in J_1(x^*)$, with the resulting reduction in the rate of convergence (cf., Theorem 3.2.1). In an important sense, a degenerate constraint is redundant, as the first order necessary condition (the Kuhn-Tucker condition) is unchanged if the constraint is just ignored, and as degeneracy does not effect the second order sufficiency condition. However, it is an indication that the property of membership of the active constraint set is extremely sensitive to perturbations of the problem data. Thus, classical barrier function methods can be expected to encounter difficulty, because they are not capable of discriminating between degenerate and non-degenerate constraints.

A characteristic of classical barrier function methods is that there is no discrimination between members of the constraint set. The fact that classical barrier function methods put equal weights on each constraint results in several undesirable properties: (i) For near-degenerate problems, it limits the usefulness of extrapolation procedures for improving

the numerical performance, as reported in subsection 3.2.3; (ii) For degenerate problems, the rate of convergence is reduced to order $O(r^{1/3})$ as shown in subsection 3.2.2; (iii) For situations where the solution $\tilde{x}^*$ is in the interior of the feasible region, the inability to identify inactive constraints clearly slows down convergence.

Following the above arguments, it is therefore desirable to find ways which guarantee that appropriate weights are assigned to the constraints. In order to provide an automatic procedure for identifying and discarding inactive constraints, Osborne (1972) has suggested weighting each constraint by its multiplier estimate from the previous iteration. The transformed function now takes the form:

$$B(\tilde{x}, r_k \tilde{u}^{(k+1)}) = f(\tilde{x}) - r_k \sum_{i \in J} \tilde{u}_i^{(k-1)} \log[g_i(\tilde{x})] \quad , \quad (3.3-1)$$

where $\tilde{u}^{(0)}$, r_0 and $c > 1$ are given, and $r_k = r_{k-1}/c$ for $k = 1, 2, \dots$. The terms $\tilde{u}_i^{(k)} = r_k \tilde{u}_i^{(k-1)} / g_i(\tilde{x}^{(k)})$ denote the multiplier estimates after the k th iteration.

With the above modification, the barrier function algorithm is now truly sequential. The minimum $\tilde{x}^{(k)}$ depends not only on r_k but also on $\tilde{u}^{(k-1)}$, and hence the usual trajectory analysis is no longer applicable. However, the situation can be partially rectified by considering the separable barrier function $B(\tilde{x}, \tilde{s})$ with $\tilde{s} = r \tilde{\phi}(r)$, where $\tilde{\phi}(r)$ denotes a differentiable vector-valued function of r :

$$B(\tilde{x}, r \tilde{\phi}(r)) = f(\tilde{x}) - r \sum_{i \in J} \phi_i(r) \log[g_i(\tilde{x})] \quad . \quad (3.3-2)$$

This special form of the separable barrier function is of interest, because (i) trajectory analysis remains viable, and (ii) the modified barrier

function (3.3-1) generates points which follow the trajectory defined by a particular differentiable vector-valued function $\phi(r)$

$$\phi(r) = u(cr) . \quad (3.3-3)$$

In subsection 3.3.2, various properties and a basis for extrapolation of the modified barrier function algorithm are given. In subsection 3.3.3, a modified barrier function algorithm is proposed and implemented. The resulting algorithm is then tested and its numerical performance compared with the classical barrier function algorithm on a wide range of test problems.

The motivation for the development of this particular modified barrier function algorithm is best illustrated by considering a specific example:

Example 3.3.1

$$\begin{aligned} \text{minimize } f(x) &= x_2 \\ \text{subject to } g_1(x) &= x_2 - x_1^2 \geq 0 , \\ \text{and } g_2(x) &= x_1 \geq 0 . \end{aligned}$$

The solution is $x^* = (0,0)$, $u^* = (1,0)$.

Using the special separable barrier function (3.3-2), the transformed function for Ex. 3.3.1 takes the form:

$$B(x, r\phi(r)) = x_2 - r\phi_1(r)\log(x_2 - x_1^2) - r\phi_2(r)\log(x_1) . \quad (3.3-4)$$

At the minimum $x(r)$,

$$0 = (0,1) - \frac{r\phi_1(r)}{x_2 - x_1^2} (-2x_1, 1) - \frac{r\phi_2(r)}{x_1} (1,0) . \quad (3.3-5)$$

The second component of (3.3-5) gives

$$\frac{r\phi_1(r)}{x_2 - x_1^2} = u_1(r) = 1. \quad (3.3-6)$$

The first component gives

$$\frac{r\phi_2(r)}{x_1} = u_2(r) = 2x_1 \frac{r\phi_1(r)}{x_2 - x_1^2}. \quad (3.3-7)$$

For the special case of the classical barrier function (i.e., $\phi_1(r) = \phi_2(r) = 1$), solving (3.3-6) and (3.3-7) gives

$$x_1(r) = (r/2)^{1/2}; \quad x_2(r) = 3r/2 \quad (3.3-8a)$$

$$u_1(r) = 1; \quad u_2(r) = (2r)^{1/2}. \quad (3.3-8b)$$

This guarantees order $O(r^{1/2})$ convergence as predicted in subsection 3.2.2.

Now, consider the particular case when $\phi_1(r) = 1$ and $\phi_2(r) = 2c^2r$, for a constant $c \geq 1$. The result of solving (3.3-6) and (3.3-7) now gives

$$x_1(r) = cr; \quad x_2(r) = r + c^2r^2 \quad (3.3-9a)$$

$$u_1(r) = 1; \quad u_2(r) = 2cr. \quad (3.3-9b)$$

The convergence for this particular choice of $\phi(r)$ is then of order $O(r)$. Notice the coincidence that

$$\phi(r) = (1, 2c^2r) = \tilde{u}(cr). \quad (3.3-10)$$

Now consider using the modified barrier function algorithm (3.3-1) to solve

Ex. 3.3.1. At each iteration, we minimize

$$B(\tilde{x}, r_k u^{(k-1)}) = x_2 - r_k u_1^{(k-1)} \log(x_2 - x_1^2) - r_k u_2^{(k-1)} \log(x_1), \quad (3.3-11)$$

where $r_0, u^{(0)}$ and $c > 1$ are given, and $r_k = r_{k-1}/c$ for $k = 1, 2, \dots$.

From (3.3-6), it follows that

$$u_1^{(k)} = 1 \quad \text{for } k = 1, 2, \dots \quad (3.3-12)$$

From (3.3-6) and (3.3-7), the equation for $u_2^{(k)}$ becomes

$$u_2^{(k)} = 2^{\frac{1}{2}} r_k^{\frac{1}{2}} (u_2^{(k-1)})^{\frac{1}{2}}. \quad (3.3-13)$$

For $r_k = r_{k-1}/c$, the solution to (3.3-13) is

$$u_2^{(k)} = 2cr_k \left(\frac{u_2^{(0)}}{2r_0c} \right)^{\left(\frac{1}{2}\right)^k}. \quad (3.3-14)$$

By comparing $u^{(k)}$ from (3.3-9b), (3.3-12) and (3.3-14), it follows that, if the modified barrier function algorithm is started with $u_2^{(0)} = 2r_0c$, then the resulting multipliers are the same. The modified barrier function algorithm is in effect producing the continuous solution (3.3-9) for various values of r_k , and a second order extrapolation gives the exact solution $\tilde{x}(0) = (0, 0)$.

Usually, it cannot be expected that $u_2^{(0)} = 2r_0c$. However, the ratio of $u_2^{(k)}$ in (3.3-9b) and (3.3-14) has the form $\left(\frac{u_2^{(0)}}{2r_0c} \right)^{\left(\frac{1}{2}\right)^k}$, which tends to 1 as k becomes large, and hence the modified algorithm is stable. Furthermore, by simple manipulation, it follows that

$$u_2^{(k)} = 2cr_k \cdot p(r_k), \quad (3.3-15)$$

where

$$p(r) = \left(\frac{u_2^{(0)}}{2r_0c} \right) \left(\frac{r}{r_0} \right)^{\frac{\log 2}{\log c}}. \quad (3.3-16)$$

Therefore, the discrete points of (3.3-14) lie on the continuously differentiable trajectory

$$u_2(r) = 2cr.p(r) \quad . \quad (3.3-17)$$

Hence, there exists a basis for extrapolation (first order at least).

Remark 3.3.1

(a) This example illustrates the attractiveness of the modified barrier function algorithm which forces the rate of convergence to be of order $O(r)$ instead of the usual order $O(r^{1/2})$ for degenerate problems. The process also provides a basis for extrapolation.

(b) From (3.3-17), in order to provide a basis for higher order extrapolation, $p(r)$ must be (many times) differentiable as a function of r , $r \geq 0$. It also follows from (3.3-16) that the $\frac{d^k p(0)}{dr^k}$, $k = 1, 2, \dots$, exist and are finite if and only if $c = 2$. Hence $c = 2$ is the critical value.

3.3.2 Properties of the Modified Barrier Function Algorithm

For $u^{(0)}$, r_0 and $c > 1$ given, $r_k = r_{k-1}/c$, $k = 1, 2, \dots$, find $\tilde{x}^{(k)}$ which minimizes

$$B(\tilde{x}, r_k u^{(k-1)}) = f(\tilde{x}) - r_k \sum_{i \in I} u_i^{(k-1)} \log[g_i(\tilde{x})] \quad , \quad (3.3-18)$$

where $u_i^{(k)} = r_k u_i^{(k-1)} / g_i(\tilde{x}^{(k)})$ is the multiplier estimate after the k th iteration.

Throughout this subsection, it will be assumed that $\tilde{x}^*$ is a regular local solution of the problem (ICP), and that the sequences $\{\tilde{x}^{(k)}\}$

and $\{u^{(k)}\}$ generated by the modified algorithm converge to $\tilde{x}^*$ and $\tilde{u}^*$, respectively.

Inactive Constraints

The modified algorithm (3.3-18) was first proposed by Osborne (1972) for its ability to identify and discard inactive constraints.

Definition

Let $A(r)$ be a continuous function of r , $\{r_k\}$ a strictly decreasing null sequence, and $\{A^{(k)}\}$ a sequence of values, then the sequence $\{A^{(k)}\}$ is said to converge *super fast* to $A(r)$, if there exist bounded constants γ and ρ such that

$$\|A^{(k)} - A(r_k)\| \leq \gamma \rho^k \cdot \prod_{i=1}^k r_i, \quad \text{for } k = 1, 2, \dots \quad (3.3-19)$$

Below, the result of Osborne (1972) for inactive constraints is reproduced.

Theorem 3.3.1 (Osborne (1972))

Assume that $\tilde{x}^*$ is a regular local solution of the problem (ICP), and that for some i , g_i is inactive at $\tilde{x}^*$, then the sequence of multiplier estimates $\{u_i^{(k)}\}$ generated by the modified algorithm (3.3-18) converges super fast to zero.

Proof

$$\begin{aligned} u_i^{(k)} &= r_k u_i^{(k-1)} / g_i(\tilde{x}^{(k)}) \\ &\leq u_i^{(0)} \rho_i^k \prod_{j=1}^k r_j, \end{aligned} \quad (3.3-20)$$

where ρ_i is a bound for $1/g_i(\tilde{x}^{(k)})$, $k = 1, 2, \dots$. □

Remark 3.3.2

Eqn. (3.3-20) implies that, for any inactive constraint g_i , $u_i^{(k)}$ will be practically zero after a few iterations. Because of this property, for the remainder of this subsection, inactive constraints are discarded and attention is restricted to situations where all the constraints present are active.

Non-degenerate Cases

For non-degenerate problems, Theorem 3.1.4 shows that the trajectory, $\tilde{x}(r)$, for the classical barrier function algorithm is differentiable in r and supports polynomial extrapolation in r . In what follows, it is shown that the sequence of points, $\{\tilde{x}^{(k)}\}$, generated by the modified barrier function algorithm also supports polynomial extrapolation in r , and furthermore, the effect of the degeneracy measure (see subsection 3.2.3) on extrapolation is in a sense removed.

Theorem 3.3.2

Let $\tilde{x}^*$ be a regular local solution of a non-degenerate problem (ICP), and assume that all the problem functions are $(\ell+1)$ times continuously differentiable. Then, for a given constant $c \geq 1$, there exists a ℓ times continuously differentiable vector function $\hat{\phi}(r)$ (unique up to the first $(\ell+1)$ terms in its Taylor expansion about $r=0$) such that,

$$\hat{\phi}(r) = \hat{u}(cr) + o(r^\ell), \quad (3.3-21)$$

where $\hat{u}_i(r) = r\hat{\phi}_i(r)/g_i(\hat{x}(r))$, $i = 1, \dots, m$, are the multiplier estimates and $\hat{x}(r)$ is the minimizer of the compound function

$$B(\tilde{x}, r\hat{\phi}(r)) = f(\tilde{x}) - r \sum_{i \in J} \hat{\phi}_i(r) \log[g_i(\tilde{x})] \quad (3.3-22)$$

Proof

The proof is by construction. We first assume that $\hat{\phi}(r)$ exists and has the form:

$$\hat{\phi}(r) = \underline{u}^* + c \underline{a}_1 r + \dots + c^{\ell} \underline{a}_{\ell} \frac{r^{\ell}}{\ell!} + o(r^{\ell}) . \quad (3.3-23)$$

$\hat{\phi}(r)$ is then constructed by determining the coefficients $\underline{a}_1, \dots, \underline{a}_{\ell}$.

By assumption, $\hat{u}(r)$ has the form:

$$\hat{u}(r) = \underline{u}^* + \underline{a}_1 r + \dots + \underline{a}_{\ell} \frac{r^{\ell}}{\ell!} + o(r^{\ell}) . \quad (3.3-24)$$

At the minimum of $B(\underline{x}, r\hat{\phi}(r))$, the following system of equations is satisfied

$$\nabla f(\hat{\underline{x}}(r)) - \sum_{i \in J} \hat{u}_i(r) \nabla g_i(\hat{\underline{x}}(r)) = 0 \quad (3.3-25)$$

$$\hat{u}_i(r) g_i(\hat{\underline{x}}(r)) = r \hat{\phi}_i(r), i \in J .$$

By differentiating (3.3-25), the following system of differential equations is obtained

$$J(\hat{\underline{x}}(r), \hat{\underline{u}}(r)) \begin{bmatrix} \frac{d\hat{\underline{x}}(r)}{dr} \\ \frac{d\hat{\underline{u}}(r)}{dr} \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{d}{dr} (r\hat{\phi}(r)) \end{bmatrix} . \quad (3.3-26)$$

Because the problem is non-degenerate, $J(\underline{x}^*, \underline{u}^*)$ is non-singular.

Hence

$$\begin{bmatrix} \hat{\underline{x}}'(0) \\ \hat{\underline{u}}'(0) \end{bmatrix} = J(\underline{x}^*, \underline{u}^*)^{-1} \begin{bmatrix} 0 \\ \left. \frac{d}{dr} (r\hat{\phi}(r)) \right|_{r=0} \end{bmatrix} \quad (3.3-27)$$

From (3.3-23) and (3.3-24), it follows that

$$\begin{bmatrix} \hat{\tilde{x}}'(0) \\ \tilde{a}_1 \end{bmatrix} = J(\tilde{x}^*, \tilde{u}^*)^{-1} \begin{bmatrix} 0 \\ \tilde{u}^* \end{bmatrix}. \quad (3.3-28)$$

Therefore $\tilde{a}_1$ can be explicitly determined from (3.3-28). By differentiating the system of differential equations (3.3-26), it follows that (writing $\hat{\tilde{x}}$ for $\hat{\tilde{x}}(r)$, etc.)

$$J \begin{bmatrix} \hat{\tilde{x}}'' \\ \hat{\tilde{u}}'' \end{bmatrix} = \begin{bmatrix} 0 \\ (r\hat{\phi})'' \end{bmatrix} - \frac{dJ}{dr} \cdot \begin{bmatrix} \hat{\tilde{x}}' \\ \hat{\tilde{u}}' \end{bmatrix}. \quad (3.3-29)$$

Again, by evaluating at $r=0$, substituting for $\hat{\tilde{u}}$ and $\hat{\phi}$ from (3.3-23) and (3.3-24), it follows that

$$\begin{bmatrix} \hat{\tilde{x}}''(0) \\ \tilde{a}_2 \end{bmatrix} = J(\tilde{x}^*, \tilde{u}^*)^{-1} \left\{ \begin{bmatrix} 0 \\ 2c\tilde{a}_1 \end{bmatrix} - \frac{dJ(\tilde{x}^*, \tilde{u}^*)}{dr} \cdot \begin{bmatrix} \hat{\tilde{x}}'(0) \\ \tilde{a}_1 \end{bmatrix} \right\}. \quad (3.3-30)$$

Therefore $\tilde{a}_2$ and similarly $\tilde{a}_3, \dots, \tilde{a}_\ell$ can be successively determined. $\square$

Theorem 3.3.3

Let $\tilde{x}^*$ be a regular local solution of a non-degenerate problem (ICP), and assume that the conditions of Theorem 3.3.2 are satisfied. If r_0 sufficiently small, $\tilde{u}^{(0)} > 0$ and $c > 1$ are given, and $r_k = r_{k-1}/c$, $k=1, 2, \dots$, then the sequence of points $\{\tilde{x}^{(k)}\}$ generated by the modified barrier function algorithm converges super fast to $\hat{\tilde{x}}(r)$.

Proof

It is only necessary to show that the sequence of multiplier estimates $\{u^{(k)}\}$ converges super fast to $\hat{u}(r)$. Consider the separable barrier function

$$B(\underline{x}, \underline{s}) = f(\underline{x}) - \sum_{i=1}^m s_i \log[g_i(\underline{x})] \quad (3.3-31)$$

At the minimum of $B(\underline{x}, \underline{s})$, the system of equations are

$$\nabla f(\underline{x}(\underline{s})) - \sum_{i=1}^m u_i(\underline{s}) \nabla g_i(\underline{x}(\underline{s})) = 0 \quad (3.3-32)$$

$$u_i(\underline{s}) g_i(\underline{x}(\underline{s})) = s_i, \quad i = 1, \dots, m.$$

The Jacobian of the system of equations (3.3-32) with respect to $\underline{x}$ and $\underline{u}$ is non-singular. The behaviour of $\underline{x}(\underline{s})$ and $\underline{u}(\underline{s})$ is studied by considering the systems of differential equations (writing $\underline{x}$ for $\underline{x}(\underline{s})$, etc.)

$$J(\underline{x}, \underline{u}) \begin{bmatrix} \frac{\partial \underline{x}}{\partial s_i} \\ \frac{\partial \underline{u}}{\partial s_i} \end{bmatrix} = \begin{bmatrix} 0 \\ e_i \end{bmatrix}, \quad i = 1, \dots, m, \quad (3.3-33)$$

where $e_i = (0, 0, \dots, 1, \dots, 0)$ with 1 as the i th component. Note that for the situation when $s_i = r, \forall i$, the more familiar system of differential equations for the classical barrier function method is obtained

$$J(\underline{x}, \underline{u}) \begin{bmatrix} \sum_{i=1}^m \frac{\partial \underline{x}}{\partial s_i} \\ \sum_{i=1}^m \frac{\partial \underline{u}}{\partial s_i} \end{bmatrix} = J(\underline{x}, \underline{u}) \begin{bmatrix} \frac{d\underline{x}}{dr} \\ \frac{d\underline{u}}{dr} \end{bmatrix} = \begin{bmatrix} 0 \\ e \end{bmatrix}, \quad (3.3-34)$$

where $\underline{e} = (1, 1, \dots, 1)$.

From (3.3-33), it follows that

$$\frac{\partial u_i}{\partial s_j} = \frac{1}{\det(J)} (-1)^{n+i+n+j} |M_{n+i, n+j}|. \quad (3.3-35)$$

Hence, $\left| \frac{\partial u_i}{\partial s_j} \right|$ is bounded for $1 \leq i, j \leq m$.

Using differentiability, it follows that, for small $\Delta \underline{s}$,

$$\Delta u_i = \sum_{j=1}^m \frac{\partial u_i}{\partial s_j} \cdot \Delta s_j + \text{smaller terms}. \quad (3.3-36)$$

That is, for the deviation from $\hat{u}(r)$,

$$\begin{aligned} u_i^{(k)} - \hat{u}_i(r_k) &= \sum_{j=1}^m \frac{\partial u_i}{\partial s_j} \left(r_k u_j^{(k-1)} - r_k \hat{u}_j(r_{k-1}) \right) + \text{smaller terms} \\ &= r_k \cdot \sum_{j=1}^m \frac{\partial u_i}{\partial s_j} \left(u_j^{(k-1)} - \hat{u}_j(r_{k-1}) \right) + \text{smaller terms}. \end{aligned} \quad (3.3-37)$$

From (3.3-37), it follows that $\{u^{(k)}\}$ converges super fast to $\hat{u}(r)$.

Remark 3.3.3

(a) For the classical (logarithmic) barrier function algorithm, the convergence of $\underline{x}^{(k)}$ to $\underline{x}^*$ is such that, asymptotically,

$$g_j(\underline{x}^{(k)}) \sim \frac{r_k}{u_j^*}, \quad \text{for } g_j \text{ active}. \quad (3.3-38)$$

In some sense, the quantity $\max_{i \in J_0(\underline{x}^*)} |g_i(\underline{x}^{(k)})|$ represents a measure of convergence of $\underline{x}^{(k)}$ to $\underline{x}^*$, and the quantity $\min_{i \in J_0(\underline{x}^*)} |g_i(\underline{x}^{(k)})|$ represents the degree of difficulty of the unconstrained minimization.

The ratio of the two quantities is proportional to γ the measure of degeneracy (see subsection 3.2.3). Therefore, in this sense, the effect of the size of γ on the classical barrier function algorithm corresponds

to the fact that the (asymptotic) approach to $\tilde{x}^*$ is such that the distance of $\tilde{x}^{(k)}$ from each surface $g_i(\tilde{x}) = 0, i \in J_0(\tilde{x}^*)$, is inversely proportional to the size of the multiplier u_i^* .

For the modified algorithm,

$$g_i(\tilde{x}^{(k)}) = \frac{r_k u_i^{(k-1)}}{u_i^{(k)}} \sim r_k, \text{ for } g_i \text{ active.} \quad (3.3-39)$$

The modified algorithm automatically adjusts the weight for each constraint such that the approach of $\tilde{x}^{(k)}$ to $\tilde{x}^*$ is (asymptotically) equi-distance from each active constraint surface $g_i(\tilde{x}) = 0$. The effect of the size of γ is asymptotically removed.

(b) If the modified algorithm is initiated with an arbitrary $\tilde{u}^{(0)}$ (e.g., $\tilde{u}^{(0)} = \underline{c}$); for active constraints, $u_i^{(0)}$ is not necessarily close to $\hat{u}_i(r_0)$; for inactive constraints, $u_i^{(0)} \sim 0(1)$ also gives a considerable contribution. Thus, the first point, $\tilde{x}^{(1)}$, obtained is not very suitable for extrapolation. The very first minimization should be used to find an initial estimate for $\tilde{u}^{(0)}$ and a better starting point $\tilde{x}^{(0)}$.

Degenerate Cases

In subsection 3.3.1, it is shown that a degenerate problem (Ex. 3.3.1) can be handled effectively by the modified algorithm. In what follows, a general result is established that, for degenerate problems, the sequence of points, $\{\tilde{x}^{(k)}\}$, generated by the modified algorithm converges to $\tilde{x}^*$ with order $O(r_k)$ and not $O(r_k^{1/2})$, which was established for the classical barrier function algorithm in subsection 3.2.2.

Theorem 3.3.4

Let $\tilde{x}^*$ be a regular local solution of a degenerate problem (ICP),

and assume that the problem functions are at least twice continuously differentiable. If r_0 is sufficiently small with $\underline{u}^{(0)} > 0$ and $c > 1$ given, and $\{\tilde{x}^{(k)}\}$ is the sequence of points generated by the modified algorithm (3.3-18), then there exist constants μ_1 and $\mu_2 > 0$ such that the asymptotic inequalities

$$\mu_1 r_k + o(r_k) \leq \|\tilde{x}^{(k)} - x^*\| \leq \mu_2 r_k + o(r_k) \quad (3.3-40)$$

hold for sufficiently large k .

Proof

Without loss of generality and in the interest of simplicity, it is assumed that

- (i) all the constraints present are active (cf., Remark 3.3.2),

$$|J_1(\tilde{x}^*)| + |J_2(\tilde{x}^*)| = m; \quad (3.3-41)$$

- (ii) $\underline{u}^{(0)} = (1, 1, \dots, 1)$; and

- (iii) r_0 is sufficiently small such that $\{\tilde{x}^{(k)}\}$ belongs to a small neighbourhood of $\tilde{x}^*$: that is for every $k \geq 1$

$$\|\tilde{x}^{(k)} - \tilde{x}^*\| \leq \delta, \quad (3.3-42)$$

$$|u_i^{(k)} - u_i^*| \leq \delta, \quad i \in J_1(\tilde{x}^*), \quad (3.3-43a)$$

and

$$|u_i^{(k)}| \leq \delta, \quad i \in J_2(\tilde{x}^*), \quad (3.3-43b)$$

where $\delta \ll 1$.

Note that, for every $i \in J_1(x^*)$,

$$g_i(x^{(k)}) = \nabla g_i(x^*)^T \cdot (x^{(k)} - x^*) + o(\|x^{(k)} - x^*\|), \quad (3.3-44)$$

also that, from (3.3-43a), (cf., Remark 3.3.3a)

$$g_i(x^{(k)}) = \frac{r_k u^{(k-1)}}{u_i^{(k)}} \sim r_k (1 + o(\delta)) . \quad (3.3-45)$$

Hence

$$r_k \leq o(\|x^{(k)} - x^*\|) . \quad (3.3-46)$$

Therefore, to complete the proof for Theorem 3.3.4, it is only necessary to show that

$$\|x^{(k)} - x^*\| \leq o(r_k) . \quad (3.3-47)$$

Before proving (3.3-47), we first establish two preliminary results.

Result 1: There exists a bounded constant α such that, for every $k \geq 1$,

$$|u_i^{(k)} - u_i^*| \leq \alpha \|x^{(k)} - x^*\| , \quad i \in J_1(x^*) \quad (3.3-48a)$$

$$|u_i^{(k)}| \leq \alpha \|x^{(k)} - x^*\| , \quad i \in J_2(x^*) \quad (3.3-48b)$$

Proof of Result 1: Since $x^{(k)}$ is a stationary point of $B(x, r_k u^{(k-1)})$, it follows that

$$\nabla f(x^{(k)}) - \sum_{i \in J_1(x^*)} u_i^{(k)} \nabla g_i(x^{(k)}) = \sum_{i \in J_2(x^*)} u_i^{(k)} \nabla g_i(x^{(k)}) . \quad (3.3-49)$$

Because the problem functions are twice differentiable and that the

Kuhn-Tucker condition holds at $\tilde{x}^*$, (3.3-49) becomes

$$\begin{aligned} \nabla_{\tilde{x}}^2 L_0(\tilde{x}^*, \tilde{u}^*) \cdot (\tilde{x}^{(k)} - \tilde{x}^*) + o(\|\tilde{x}^{(k)} - \tilde{x}^*\|) \\ - \sum_{i \in J_1(\tilde{x}^*)} (u_i^{(k)} - u_i^*) \nabla_{\tilde{x}} g_i(\tilde{x}^{(k)}) = \sum_{i \in J_2(\tilde{x}^*)} u_i^{(k)} \nabla_{\tilde{x}} g_i(\tilde{x}^{(k)}) . \end{aligned} \quad (3.3-50)$$

Let w_i^T denote the i th row of the $m \times n$ matrix $N^+ = (N^T N)^{-1} N^T$, where $N = [\nabla_{\tilde{x}} g_1(\tilde{x}^{(k)}), \dots, \nabla_{\tilde{x}} g_m(\tilde{x}^{(k)})]$. Then, from the linear independence of $\{\nabla_{\tilde{x}} g_i(\tilde{x}^{(k)})\}_{i=1, \dots, m}$, it follows that

$$w_i^T \nabla_{\tilde{x}} g_j(\tilde{x}^{(k)}) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} . \end{cases} \quad (3.3-51)$$

Eqs. (3.3-50) and (3.3-51) give that

$$|u_i^{(k)} - u_i^*| \leq \|w_i^T \cdot \nabla_{\tilde{x}}^2 L_0(\tilde{x}^*, \tilde{u}^*)\| \cdot \|\tilde{x}^{(k)} - \tilde{x}^*\| + o(\|\tilde{x}^{(k)} - \tilde{x}^*\|), \quad i \in J_1(\tilde{x}^*) \quad (3.3-52a)$$

$$|u_i^{(k)}| \leq \|w_i^T \cdot \nabla_{\tilde{x}}^2 L_0(\tilde{x}^*, \tilde{u}^*)\| \cdot \|\tilde{x}^{(k)} - \tilde{x}^*\| + o(\|\tilde{x}^{(k)} - \tilde{x}^*\|), \quad i \in J_2(\tilde{x}^*) \quad (3.3-52b)$$

1. □

Result 2: For sufficiently large $\kappa_0 > 0$, there exists $a > 0$ such that, if

$$r_k \leq \frac{1}{\kappa_0} \|\tilde{x}^{(k)} - \tilde{x}^*\| , \quad (3.3-53)$$

then

$$(\tilde{x}^{(k)} - \tilde{x}^*)^T \cdot \nabla_{\tilde{x}}^2 L_0(\tilde{x}^*, \tilde{u}^*) \cdot (\tilde{x}^{(k)} - \tilde{x}^*) \geq a \|\tilde{x}^{(k)} - \tilde{x}^*\|^2 . \quad (3.3-54)$$

Proof of Result 2: Let $\tilde{x}^{(k)} - \tilde{x}^*$ be decomposed as

$$\tilde{x}^{(k)} - \tilde{x}^* = \|\tilde{x}^{(k)} - \tilde{x}^*\| (t^{(k)} + v^{(k)}) , \quad (3.3-55)$$

where $\tilde{v}^{(k)}$ is a linear combination of the $\nabla g_i(x^*)$, $i \in J_1(x^*)$, and $\tilde{t}^{(k)}$ is in the orthogonal complement of this set. Then, $\forall i \in J_1(x^*)$, (3.3-43a) implies that

$$\begin{aligned} g_i(x^{(k)}) &= \frac{r_k u_i^{(k-1)}}{u_i^{(k)}} = r_k(1+O(\delta)) \\ &\leq \frac{(1+O(\delta))}{\kappa_0} \cdot \|\tilde{x}^{(k)} - x^*\| \end{aligned} \quad (3.3-56)$$

Together, Eqns. (3.3-44), (3.3-55) and (3.3-56) imply that

$$\nabla g_i(x^*)^T \cdot \tilde{v}^{(k)} \leq 0 \left(\frac{1}{\kappa_0} \right) \quad (3.3-57)$$

Because $\tilde{v}^{(k)}$ is a linear combination of the $\nabla g_i(x^*)$, $i \in J_1(x^*)$, it follows that

$$\|\tilde{v}^{(k)}\| = \left(\tilde{v}^{(k)T} \cdot \tilde{v}^{(k)} \right)^{1/2} \leq 0 \left(\frac{1}{\kappa_0^{1/2}} \right) \quad (3.3-58)$$

Also, from (3.3-55), it follows that

$$1 - \|\tilde{v}^{(k)}\| \leq \|\tilde{t}^{(k)}\| \leq 1 + \|\tilde{v}^{(k)}\|, \quad (3.3-59)$$

and hence,

$$\|\tilde{t}^{(k)}\| = 1 + 0 \left(\frac{1}{\kappa_0^{1/2}} \right) \quad (3.3-60)$$

Thus, by using (3.3-55), (3.3-58) and (3.3-60) and the second order sufficiency condition, it follows that, if κ_0 is sufficiently large, a constant $a > 0$ can be found such that

$$\begin{aligned} (\tilde{x}^{(k)} - x^*)^T \cdot \nabla_x^2 L_0(x^*, u^*) \cdot (\tilde{x}^{(k)} - x^*) &= \|\tilde{x}^{(k)} - x^*\|^2 \left\{ \tilde{t}^{(k)T} \cdot \nabla_x^2 L_0(x^*, u^*) \cdot \tilde{t}^{(k)} + \right. \\ &\quad \left. O\|\tilde{v}^{(k)}\| \right\} \geq a \|\tilde{x}^{(k)} - x^*\|^2 \end{aligned} \quad (3.3-61)$$

Proof of (3.3-47): This is done by first assuming that (3.3-47) does not hold and then using *Results 1 and 2* to derive a contradiction to (3.3-42).

If (3.3-47) does not hold, then there exists a subsequence $\{k(l)\}$ such that, as l becomes large,

$$r_{k(l)} \sim o(\|x^{(k(l))} - x^*\|). \quad (3.3-62)$$

Choose an arbitrary large κ ($\kappa > \max(r_1^{-1}, \kappa_0)$), and $k = k(\kappa)$ such that

$$r_k \leq \frac{1}{\kappa} \cdot \|x^{(k)} - x^*\|, \quad (3.3-63)$$

and that the condition (3.3-53) of *Result 2* is satisfied. Multiplying

(3.3-50) by $(x^{(k)} - x^*)$ and making the substitution $g_i(x^{(k)} + o(\|x^{(k)} - x^*\|)) = \nabla g_i(x^{(k)})^T \cdot (x^{(k)} - x^*)$, $\forall i$, it follows that

$$\begin{aligned} & (x^{(k)} - x^*)^T \cdot \nabla_x^2 L_0(x^*, u^*) \cdot (x^{(k)} - x^*) + o(\|x^{(k)} - x^*\|^2) \\ & - \sum_{i \in J_1(x^*)} (u_i^{(k)} - u_i^*) (g_i(x^{(k)} + o(\|x^{(k)} - x^*\|))) \\ & = \sum_{i \in J_2(x^*)} u_i^{(k)} (g_i(x^{(k)} + o(\|x^{(k)} - x^*\|))) \end{aligned} \quad (3.3-64)$$

Using (3.3-63), *Results 1 and 2*, (3.3-64) becomes

$$\|x^{(k)} - x^*\|^2 - \sum_{i \in J_1(x^*)} |u_i^{(k)} - u_i^*| g_i(x^{(k)}) \leq \sum_{i \in J_2(x^*)} u_i^{(k)} g_i(x^{(k)}), \quad (3.3-65)$$

where terms of smaller order are absorbed into the inequality by further adjusting the coefficient a . Again, by using (3.3-63), (3.3-45) and

Result 1, it follows that the second term on the LHS of (3.3-65) is of smaller order than the first. Hence

$$\begin{aligned}
 \frac{a}{2} \| \tilde{x}^{(k)} - \tilde{x}^* \|^2 &\leq \sum_{i \in J_2(\tilde{x}^*)} u_i^{(k)} g_i(\tilde{x}^{(k)}) \\
 &= \sum_{i \in J_2(\tilde{x}^*)} r_k u_i^{(k-1)} \\
 &\leq r_k |J_2(\tilde{x}^*)| \cdot \alpha \| \tilde{x}^{(k-1)} - \tilde{x}^* \|, \quad (3.3-66)
 \end{aligned}$$

where the coefficient α is taken from *Result 1*. Together, (3.3-63) and (3.3-66) give that

$$\frac{a}{2} (\kappa r_k)^2 \leq r_k |J_2(\tilde{x}^*)| \cdot \alpha \| \tilde{x}^{(k-1)} - \tilde{x}^* \| \quad (3.3-67)$$

Thus

$$r_{k-1} \leq \frac{c}{\beta \kappa^2} \| \tilde{x}^{(k-1)} - \tilde{x}^* \|, \quad (3.3-68)$$

where $\beta = \frac{a}{2 \cdot \alpha \cdot |J_2(\tilde{x}^*)|}$.

This result implies that the condition (3.3-53) of *Result 2* is again satisfied for $k := k-1$. The above arguments can then be repeated using (3.3-68) instead of (3.3-63). Thus, by backtracking, it follows that

$$\| \tilde{x}^{(1)} - \tilde{x}^* \| \geq \left(\frac{\beta \kappa}{c} \right)^{\sum_{i=0}^{k-2} 2^i} \cdot \kappa r_1.$$

which provides the required contradiction. $\square$

Remark 3.3.4

(a) Note that, because of the nature of the second order sufficiency

condition, it is necessary to construct a proof by contradiction. On the other hand, if a stronger condition ($\nabla_{x_0}^2 L_0(\tilde{x}^*, \tilde{u}^*)$ is positive definite) is imposed on the problem (ICP), then it will be possible to construct a proof for (3.3-47) without first assuming that it does not hold.

(b) The modified algorithm automatically adjusts the weight for each constraint such that, asymptotically,

$$g_i(\tilde{x}^{(k)}) \sim r_k, \quad \forall i \in J_1(\tilde{x}^*) . \quad (3.3-69)$$

For degenerate constraints, the results of subsection 3.2.2 indicate that $u_i^{(k)} = O((r_k u_i^{(k-1)})^{1/2})$. Therefore, asymptotically, it is likely that $u_i^{(k)}$, $i \in J_2(\tilde{x}^*)$, are of order $O(r_k)$ (i.e., $u_i^{(k)} \sim b_i r_k$, b_i a constant). This implies, for degenerate constraints, a behaviour of the form

$$g_i(\tilde{x}^{(k)}) \sim c r_k, \quad i \in J_2(\tilde{x}^*) . \quad (3.3-70)$$

This behaviour (namely, (3.3-69) and (3.3-70)) is evident in the numerical results of subsection 3.3.3.

(c) Note that the proof of Theorem 3.3.4 utilizes every condition imposed on the problem. This suggests that any result stronger than Theorem 3.3.4 is very unlikely. In the light of this observation, it is concluded that, when the problem is degenerate, only first order improvement with extrapolation can in general be expected. This (intuitive) prediction is confirmed by the numerical results of subsection 3.3.3. On the other hand, for degenerate problems with some special structure, higher order improvement with extrapolation may be possible. Ex. 3.3.1 of subsection 3.3.1 is one such special case. Its two active constraint gradients happen to be orthogonal and the critical step size factor $c = 2$ is chosen to take care of $(r_k u_2^{(k+1)})^{1/2}$ convergence.

3.3.3 Implementation and Numerical Performance

In this subsection, some of the details associated with the implementation of the modified (logarithmic) barrier function algorithm are discussed. In addition, numerical results supporting the conclusions of subsection 3.3.2 are presented. Numerical justification for the development of the modified barrier function algorithm is obtained by comparing its numerical performance with that of the classical barrier function algorithm.

The Choice of c

The choice of c is a crucial factor in the numerical performance of the modified algorithm. The value of c must be large enough to ensure that the extrapolation process is numerically stable. In addition, it must not be too large, if the number of possible iterations, before extreme ill-condition is encountered, is not greatly limited. For the above reasons, the choice $c=2$ is taken as a suitable value. With the reduction factor $c=2$, the numerical process is numerically stable (see Appendix 3). It is relevant to note that a value of 2 for the reduction factor is widely used in other areas of numerical mathematics (for example, numerical integration and differentiation). Another possible argument supporting the choice $c=2$ is that, for the particular degenerate problem, Ex. 3.3.1, the choice $c=2$ is the critical value which takes care of order $(r_k u_2^{(k-1)})^{1/2}$ convergence and thereby makes higher order extrapolation possible.

The Choices of r_1 and $u^{(0)}$

The choice of r_1 is another critical factor in the numerical performance of the modified algorithm. The size of r_1 determines the degree of difficulty associated with the (first) unconstrained minimization subproblem. In addition, for a fixed reduction factor c , the size of r_1

determines the performance of extrapolation because of its presence in the error formula (see Eqn. (3.2-1) and Appendix 3). Without any other information, a natural choice for $\underline{u}^{(0)}$ is $\underline{u}^{(0)} = (1, 1, \dots, 1)$, which corresponds to starting as the classical barrier function algorithm with equal weights on the constraints. This is consistent with Remark 3.3.3b which suggests that the very first minimization should be used to obtain an initial estimate for $\underline{u}^{(0)}$. Thus, the algorithm proceeds as follows: Start as the classical barrier function method, with parameter $\bar{r}$,

$$B(\underline{x}, \bar{r}) = f(\underline{x}) - \bar{r} \sum_{i \in J} \log[g_i(\underline{x})] . \quad (3.3-71)$$

At the minimum $\bar{\underline{x}}$ of $B(\underline{x}, \bar{r})$, it is known that (for $\bar{u}_i = \bar{r}/g_i(\bar{\underline{x}})$)

$$\nabla f(\bar{\underline{x}}) = \sum_{i \in J_0(\underline{x}^*)} \bar{u}_i \nabla g_i(\bar{\underline{x}}) + \sum_{i \notin J_0(\underline{x}^*)} \frac{\bar{r}}{g_i(\bar{\underline{x}})} \cdot \nabla g_i(\bar{\underline{x}}) . \quad (3.3-72)$$

Multiplying (3.3-72) by $(\bar{\underline{x}} - \underline{x}^*)$, it follows that

$$\begin{aligned} f(\bar{\underline{x}}) - f(\underline{x}^*) &= \sum_{i \in J_0(\underline{x}^*)} \bar{u}_i g_i(\bar{\underline{x}}) + 0(\max\{\|\bar{\underline{x}} - \underline{x}^*\|^2, \bar{r} \|\bar{\underline{x}} - \underline{x}^*\|\}) \\ &= |J_0(\underline{x}^*)| \cdot \bar{r} + 0(\max\{\|\bar{\underline{x}} - \underline{x}^*\|^2, \bar{r} \|\bar{\underline{x}} - \underline{x}^*\|\}) . \end{aligned} \quad (3.3-73)$$

Assuming that $(\bar{\underline{x}} - \underline{x}^*)$ is not orthogonal to $\nabla f(\underline{x}^*)$, and using the fact that $|J_0(\underline{x}^*)| \leq n$, the following rough estimate is obtained

$$\|\nabla f\| \cdot \beta \cdot \|\bar{\underline{x}} - \underline{x}^*\| = n \cdot \bar{r} , \quad (3.3-74)$$

where β is the product of $n/|J_0(\underline{x}^*)|$ and the cosine of the angle between $(\bar{\underline{x}} - \underline{x}^*)$ and $\nabla f(\underline{x}^*)$. Therefore, if, with $\beta=1$, it is required that $\|\bar{\underline{x}} - \underline{x}^*\| \leq 10^{-2}$ holds for the very first estimate, then a suitable value

for $\bar{r}$ is given by

$$\bar{r} = \min \left(10^{-2}, \frac{\|\nabla f\| \cdot 10^{-2}}{n} \right) . \quad (3.3-75)$$

The modified algorithm is then restarted using $\bar{x}$ as the starting point, setting

$$\underline{u}^{(0)} = \underline{\bar{u}} \quad (3.3-76)$$

and

$$r_1 = \min \left(\frac{10^{-2}}{\|\underline{\bar{u}}\|_\infty}, \frac{\|\nabla f(\bar{x})\| \cdot 10^{-2}}{n \|\underline{\bar{u}}\|_\infty} \right) , \quad (3.3-77)$$

where the difference between (3.3-77) and (3.3-75) takes account of the difference between the modified and classical algorithms.

Convergence Criteria

If $\underline{x}_{\min}^{(k)}$ denotes the best estimate of the solution after the k th iteration, then a possible stopping criterion is

$$\|\underline{x}_{\min}^{(k)} - \underline{x}_{\min}^{(k-1)}\| \leq \varepsilon_0 . \quad (3.3-78)$$

For constrained optimization problems, the solution $\underline{x}^*$ usually occurs at the boundary of the feasible region (i.e., $g_i(\underline{x}^*) = 0$ for g_i active). Since $u_i^{(k)} \rightarrow 0$ for inactive constraints, another possible stopping criterion is

$$\max_{i: u_i^{(k)} \geq \varepsilon'} |g_i(\underline{x}_{\min}^{(k)})| \leq \varepsilon_1, \quad (3.3-79)$$

where ε_1 is proportional to ε_0 with the proportional constant (presumably) depending on $\max_{i \in J_0(\underline{x}^*)} \|\nabla g_i\|$, and ε' is used to discriminate between active and inactive constraints. However, because $u_i^{(k)} \rightarrow 0$ for

degenerate active constraints, the stopping criterion (3.3-79) needs to be modified. From the results of subsection 3.3.2, the convergence of $\tilde{x}^{(k)}$ to $\tilde{x}^*$ is such that

$$g_{\tilde{i}}(\tilde{x}^{(k)}) \sim r_k, \quad \tilde{i} \in J_1(\tilde{x}^*) \quad (3.3-80)$$

For degenerate constraints, $\tilde{i} \in J_2(\tilde{x}^*)$, Theorem 3.3.4 implies that

$$\begin{aligned} g_{\tilde{i}}(\tilde{x}^{(k)}) &\leq \|\nabla g_{\tilde{i}}(\tilde{x}^*)\| \cdot \|\tilde{x}^{(k)} - \tilde{x}^*\| + \text{smaller terms} \\ &\leq B r_k, \end{aligned} \quad (3.3-81)$$

where B is a bounded constant. Eqns. (3.3-80) and (3.3-81) and the improvement from extrapolation imply that, after a few iterations,

$$|g_{\tilde{i}}(\tilde{x}_{\min}^{(k)})| << r_k, \quad \tilde{i} \in J_0(\tilde{x}^*) \quad (3.3-82)$$

Thus, the following stopping criterion could be used

$$\max_{\tilde{i}: |g_{\tilde{i}}(\tilde{x}_{\min}^{(k)})| \leq 10 \cdot r_k} |g_{\tilde{i}}(\tilde{x}_{\min}^{(k)})| \leq \epsilon_1 \quad (3.3-83)$$

The actual stopping criterion adopted for implementation is a combination of (3.3-78) and (3.3-83). Condition (3.3-83) is used as the main stopping criterion with condition (3.3-78) as a safeguard, with $\epsilon_1 = 10 \epsilon_0$.

Another related issue is the tolerance (stopping criteria) for the unconstrained minimization subproblems. If a reduction factor $c=2$ is used, then the result in Appendix 3 suggests the use of the tolerance $\bar{\epsilon} = \epsilon_0/10$ (where ϵ_0 is the tolerance for stopping the modified algorithm).

Algorithm 3.3.1

Given the functions $f, g_1, \dots, g_m$ with a feasible starting point $\tilde{x}^{(0)}$, and the required tolerance ϵ_0 .

1) initialize:

1.1) set $\bar{\epsilon} := \epsilon_0/10$, $\epsilon_1 := 10\epsilon_0$, $u^{(0)} := (1, 1, \dots, 1)$ and
 $r_1 := \min \left(10^{-2}, \frac{\|\nabla f(\tilde{x}^{(0)})\| \cdot 10^{-2}}{n} \right)$;

1.2) find $\tilde{x}^{(1)}$ which minimizes $B(\tilde{x}, r_1 u^{(0)})$ from the starting point $\tilde{x}^{(0)}$ with a tolerance of $\bar{\epsilon}$;
 set $u_i^{(1)} := r_1 u_i^{(0)} / g_i(\tilde{x}^{(1)})$, $\forall i$;

1.3) set $u^{(0)} := u^{(1)}$, $\tilde{x}^{(0)} := \tilde{x}^{(1)}$,
 $r_1 := \min \left(\frac{10^{-2}}{\|u^{(0)}\|_\infty}, \frac{\|\nabla f(\tilde{x}^{(0)})\| \cdot 10^{-2}}{n \|u^{(0)}\|_\infty} \right)$, and $k := 1$;

2) find $\tilde{x}^{(k)}$ which minimizes $B(\tilde{x}, r_k u^{(k-1)})$ starting from $\tilde{x}^{(0)}$ with a tolerance of $\bar{\epsilon}$, and set $u_i^{(k)} := r_k u_i^{(k-1)} / g_i(\tilde{x}^{(k)})$, $\forall i$;

3) apply extrapolation of order $k-1$ to find the (best) solution estimate $\tilde{x}_{\min}^{(k)}$;

4) apply stopping criteria:

4.1) stop the algorithm if $\max_i |g_i(\tilde{x}_{\min}^{(k)})| \leq 10r_k$ and $|g_i(\tilde{x}_{\min}^{(k)})| \leq \epsilon_1$;

4.2) stop the algorithm if $\|\tilde{x}_{\min}^{(k)} - \tilde{x}_{\min}^{(k-1)}\| \leq \epsilon_0$, print a message indicating this and give the size of

$$\max_i \|\nabla g_i(\tilde{x}_{\min}^{(k)})\| ;$$

$$i: |g_i(\tilde{x}_{\min}^{(k)})| \leq 10r_k$$

4.3) stop the algorithm if $k > 11$, and print a message that the required accuracy is not achieved after the 11th iteration

(10th order extrapolation);

- 5) find an estimate for $\tilde{x}^{(k+1)}$ by extrapolation, set $\tilde{x}^{(0)} := \tilde{x}^{(k+1)}$ if it is an interior point, otherwise set $\tilde{x}^{(0)} := \tilde{x}^{(k)}$;
- 6) set $r_{k+1} := r_k/2$ and $k := k+1$, and return to (2).

Numerical Results

The two points verified by the numerical results are: (i) the behaviour of the modified algorithm as predicted in subsection 3.3.2; and (ii) the superior numerical performance of the modified algorithm over the classical algorithm.

In order to have a meaningful comparison between the modified and the classical algorithms, we make the first iterations compatible. To avoid confusion, the controlling parameter of the classical algorithm is denoted by p . The initial parameter p_1 is selected so that $p_1 = r_1 \|\tilde{u}^{(0)}\|_\infty$ (where r_1 and $\tilde{u}^{(0)}$ are defined by (3.3-77) and (3.3-76), respectively), and $p_{k+1} = p_k/2$ for subsequent iterations. The numerical performance and other characteristics of the two algorithms are examined and compared in terms of the quantities $LB(\tilde{x})$ and $UB(\tilde{x})$ which specify lower and upper bounds for the active constraints; namely,

$$LB(\tilde{x}) = \min_{i \in J_0(\tilde{x}^*)} |g_i(\tilde{x})| \quad (3.3-84)$$

$$UB(\tilde{x}) = \max_{i \in J_0(\tilde{x}^*)} |g_i(\tilde{x})| \quad (3.3-85)$$

A wide range of non-degenerate and degenerate problems are included in the test. For non-degenerate ($\gamma < \infty$) problems, there are the 'easy' Rosen-Suzuki problem, the Colville problem I and the 'difficult'

Colville problem 11. For degenerate ($\gamma = \infty$) problems, there are the 'simple' Ex. 3.3.1, a degenerate problem obtained by modifying the Rosen-Suzuki problem, and a 'highly degenerate' problem due to Rosen and Kreuser (1972).

Algorithm 3.3.1 is implemented using a quasi-Newton minimization routine from the ANU's subroutine library (this routine is an implementation of Fletcher's (1970a) algorithm, see also Remark 3.1.5a), the required tolerance in (3.3-78) is set at $\epsilon_0 = 10^{-10}$. The numerical results in terms of the quantities $LB(\tilde{x}^{(k)})$, $UB(\tilde{x}^{(k)})$ and $UB(\tilde{x}_{\min}^{(k)})$ are listed in Tables 3.3.1 - 3.3.6. Overall, the numerical performance of both the modified and the classical algorithms is as expected. Some observations are further discussed in the following Remark.

Remark 3.3.5

(a) The quantity $LB(\tilde{x}^{(k)})$ represents the degree of difficulty associated with the unconstrained minimization. The numerical results in Tables 3.3.1 - 3.3.6 show that the choice $p_1 = r_1 \|u^{(0)}\|_{\infty}$ is such that the quantities $LB(\tilde{x}^{(k)})$ for the modified and the classical algorithms are approximately the same for every iteration (and for every problem). This provides the basis for a comparison between the two algorithms.

(b) The quantity $UB(\tilde{x}^{(k)})$ represents a measure of the convergence of $\tilde{x}^{(k)}$ to $\tilde{x}^*$. For non-degenerate ($\gamma < \infty$) problems, Remark 3.3.3a indicates that, when using the classical algorithm, the ratio $UB(\tilde{x}^{(k)})/LB(\tilde{x}^{(k)})$ is asymptotically proportional to γ , whereas, when using the modified algorithm, the ratio $UB(\tilde{x}^{(k)})/LB(\tilde{x}^{(k)})$ is asymptotically equal to 1. The numerical results in Tables 3.3.1 - 3.3.3 agree with these two predictions. For degenerate ($\gamma = \infty$) problems, Remark 3.3.4b indicates that, when using the modified algorithm, the ratio $UB(\tilde{x}^{(k)})/LB(\tilde{x}^{(k)})$ converges asymptotically to 2 (when $c = 2$), whereas the results of subsection 3.2.2

indicate that, when using the classical algorithm, the ratio $UB(\tilde{x}^{(k)})/LB(\tilde{x}^{(k)}) \sim O(r_k^{1/2})$. Hence, asymptotically, the value of $UB(\tilde{x}^{(k)})/LB(\tilde{x}^{(k)})$ is expected to increase successively by a factor of $\sqrt{2}$ (when $c = 2$). Again, the numerical results in Tables 3.3.4-3.3.6 agree with these predictions.

(c) For non-degenerate problems with reasonably large values of γ , the modified algorithm should perform better than the classical algorithm. This is verified by the numerical results listed in Tables 3.3.2 and 3.3.3 for Colville problem I and II. The numerical results in Table 3.3.1 for the Rosen-Suzuki problem indicate that the classical algorithm performs better than the modified algorithm. However, this behaviour is expected. From the error estimate (3.2-1) and the results of subsection 3.2.3, because (for the Rosen-Suzuki problem) $\gamma = 1.6$, the classical algorithm (with extrapolation) is expected to perform well. Similar performance cannot be expected for the modified algorithm. This is due to the fact that, for the first few iterations, the 'trajectory' $\{\tilde{x}^{(k)}\}$ has yet to converge to $\{\tilde{x}(r_k)\}$ (see Theorems 3.3.2 and 3.3.3). As an additional supporting evidence for the theoretical results, a comparison of the numerical results for the Rosen-Suzuki problem with those for the other problems shows clearly that the performance of the modified algorithm is effectively independent of γ , whereas the performance of the classical algorithm is sensitive to the size of γ .

(d) For degenerate problems, the results in Tables 3.3.4 - 3.3.6 show clearly that the performance of the modified algorithm is superior to that of the classical algorithm. The classical algorithm performs poorly, specially on the Rosen-Kreuser problem, as shown in Table 3.3.6. Note also that, except for Ex. 3.3.1, for which higher order improvement is possible, the improvement from extrapolation is generally only first order (cf., Remark 3.3.4c).

(c) Additional numerical results, predicted by the theory, but not shown in Tables 3.3.1 - 3.3.6, are the associated multiplier estimates for inactive constraints. The classical algorithm gives values $\sim 10^{-4}$, while the modified algorithm gives values $\leq 10^{-24}$ for $u_i^{(k)}$, $i \in J_3(x^*)$ (for the last iteration, k th).

k	Classical Algorithm				Modified Algorithm		
	$LB(\tilde{x}^{(k)})$	$UB(\tilde{x}^{(k)})$	$UB(\tilde{x}_{\min}^{(k)})$		$LB(\tilde{x}^{(k)})$	$UB(\tilde{x}^{(k)})$	$UB(\tilde{x}_{\min}^{(k)})$
1	0.50×10^{-2}	0.10×10^{-1}	0.10×10^{-1}		0.50×10^{-2}	0.50×10^{-2}	0.50×10^{-2}
2	0.25×10^{-2}	0.50×10^{-2}	0.93×10^{-4}		0.25×10^{-2}	0.25×10^{-2}	0.67×10^{-4}
3	0.12×10^{-2}	0.25×10^{-2}	0.73×10^{-6}		0.12×10^{-2}	0.12×10^{-2}	0.20×10^{-4}
4	0.62×10^{-3}	0.13×10^{-2}	0.10×10^{-8}		0.62×10^{-3}	0.62×10^{-3}	0.27×10^{-5}
5	0.31×10^{-3}	0.63×10^{-3}	0.14×10^{-9}		0.31×10^{-3}	0.31×10^{-3}	0.16×10^{-6}
6					0.16×10^{-3}	0.16×10^{-3}	0.40×10^{-8}
7					0.77×10^{-4}	0.77×10^{-4}	0.41×10^{-10}

Table 3.3.1: Numerical results for the Rosen-Suzuki problem ($\gamma = 1.6$, $p_1 = 10^{-2}$).

k	Classical Algorithm			Modified Algorithm		
	$LB(\tilde{x}^{(k)})$	$UB(\tilde{x}^{(k)})$	$UB(\tilde{x}_{\min}^{(k)})$	$LB(\tilde{x}^{(k)})$	$UB(\tilde{x}^{(k)})$	$UB(\tilde{x}_{\min}^{(k)})$
1	0.84×10^{-3}	0.66×10^{-1}	0.66×10^{-1}	0.84×10^{-3}	0.12×10^{-2}	0.12×10^{-2}
2	0.42×10^{-3}	0.38×10^{-1}	0.10×10^{-1}	0.42×10^{-3}	0.42×10^{-3}	0.35×10^{-3}
3	0.21×10^{-3}	0.21×10^{-1}	0.19×10^{-2}	0.21×10^{-3}	0.21×10^{-3}	0.11×10^{-3}
4	0.11×10^{-3}	0.11×10^{-1}	0.29×10^{-3}	0.10×10^{-3}	0.10×10^{-3}	0.15×10^{-4}
5	0.53×10^{-4}	0.58×10^{-2}	0.23×10^{-4}	0.52×10^{-4}	0.52×10^{-4}	0.98×10^{-6}
6	0.26×10^{-4}	0.29×10^{-2}	0.16×10^{-5}	0.26×10^{-4}	0.26×10^{-4}	0.28×10^{-7}
7	0.13×10^{-4}	0.15×10^{-2}	0.52×10^{-7}	0.13×10^{-4}	0.13×10^{-4}	0.24×10^{-9}
8	0.66×10^{-4}	0.75×10^{-3}	0.14×10^{-8}			
	$\ x_{\min}^{(e)} - x_{\min}^{(7)}\ \leq 10^{-10}$					

Table 3.3.2: Numerical results for the Colville problem I ($\gamma = 91$, $p_1 = 10^{-2}$).

k	Classical Algorithm			Modified Algorithm		
	$LB(\tilde{x}^{(k)})$	$UB(\tilde{x}^{(k)})$	$UB(\tilde{x}_{\min}^{(k)})$	$LB(\tilde{x}^{(k)})$	$UB(\tilde{x}^{(k)})$	$UB(\tilde{x}_{\min}^{(k)})$
1	0.18×10^{-3}	0.43×10^{-1}	0.43×10^{-1}	0.18×10^{-3}	0.18×10^{-3}	0.18×10^{-3}
2	0.88×10^{-4}	0.21×10^{-1}	0.11×10^{-2}	0.88×10^{-4}	0.88×10^{-4}	0.80×10^{-5}
3	0.44×10^{-4}	0.11×10^{-1}	0.12×10^{-3}	0.44×10^{-4}	0.44×10^{-4}	0.27×10^{-5}
4	0.22×10^{-4}	0.55×10^{-2}	0.15×10^{-4}	0.22×10^{-4}	0.22×10^{-4}	0.38×10^{-6}
5	0.11×10^{-4}	0.28×10^{-2}	0.13×10^{-5}	0.11×10^{-5}	0.11×10^{-4}	0.25×10^{-7}
6	0.55×10^{-5}	0.14×10^{-2}	0.73×10^{-7}	0.55×10^{-5}	0.55×10^{-5}	0.86×10^{-9}
7	0.28×10^{-5}	0.70×10^{-3}	0.22×10^{-8}			
8	0.13×10^{-5}	0.35×10^{-3}	0.14×10^{-9}			

Table 3.3.3: Numerical results for the Colville problem II ($\gamma = 83$, $p_1 = 10^{-2}$).

k	Classical Algorithm			Modified Algorithm		
	$LB(\tilde{x}^{(k)})$	$UB(\tilde{x}^{(k)})$	$UB(\tilde{x}_{\min}^{(k)})$	$LB(\tilde{x}^{(k)})$	$UB(\tilde{x}^{(k)})$	$UB(\tilde{x}_{\min}^{(k)})$
1	0.25×10^{-2}	0.35×10^{-1}	0.35×10^{-1}	0.25×10^{-2}	0.94×10^{-2}	0.94×10^{-2}
2	0.13×10^{-2}	0.25×10^{-1}	0.15×10^{-1}	0.13×10^{-2}	0.34×10^{-2}	0.25×10^{-2}
3	0.63×10^{-3}	0.18×10^{-1}	0.89×10^{-2}	0.63×10^{-3}	0.15×10^{-2}	0.18×10^{-3}
4	0.31×10^{-3}	0.12×10^{-1}	0.59×10^{-2}	0.31×10^{-3}	0.68×10^{-3}	0.44×10^{-5}
5	0.16×10^{-3}	0.88×10^{-2}	0.41×10^{-2}	0.16×10^{-3}	0.33×10^{-3}	0.41×10^{-7}
6	0.78×10^{-4}	0.62×10^{-2}	0.28×10^{-2}	0.78×10^{-4}	0.16×10^{-3}	0.16×10^{-9}
7	0.39×10^{-4}	0.44×10^{-2}	0.20×10^{-2}			
8	0.20×10^{-4}	0.31×10^{-2}	0.14×10^{-2}			
9	0.98×10^{-5}	0.22×10^{-2}	0.99×10^{-3}			
10	0.49×10^{-5}	0.16×10^{-2}	0.70×10^{-3}			
11	0.24×10^{-5}	0.11×10^{-2}	0.50×10^{-3}			
	not converge after 11 iterations					

Table 3.3.4: Numerical results for Ex. 3.3.1 ($p_1 = 0.50 \times 10^{-2}$).

k	Classical Algorithm				Modified Algorithm		
	$LB(\tilde{x}^{(k)})$	$UB(\tilde{x}^{(k)})$	$UB(\tilde{x}_{\min}^{(k)})$		$LB(\tilde{x}^{(k)})$	$UB(\tilde{x}^{(k)})$	$UB(\tilde{x}_{\min}^{(k)})$
1	0.48×10^{-2}	0.13	0.13		0.41×10^{-2}	0.24×10^{-1}	0.24×10^{-1}
2	0.24×10^{-2}	0.88×10^{-1}	0.49×10^{-1}		0.23×10^{-2}	0.81×10^{-2}	0.80×10^{-2}
3	0.12×10^{-2}	0.62×10^{-1}	0.30×10^{-1}		0.12×10^{-2}	0.32×10^{-2}	0.49×10^{-3}
4	0.61×10^{-3}	0.43×10^{-1}	0.20×10^{-1}		0.59×10^{-3}	0.14×10^{-2}	0.39×10^{-4}
5	0.31×10^{-3}	0.30×10^{-1}	0.14×10^{-1}		0.29×10^{-3}	0.66×10^{-3}	0.86×10^{-5}
6	0.15×10^{-3}	0.21×10^{-1}	0.97×10^{-2}		0.14×10^{-3}	0.32×10^{-3}	0.28×10^{-5}
7	0.78×10^{-4}	0.15×10^{-1}	0.68×10^{-2}		0.74×10^{-4}	0.15×10^{-3}	0.87×10^{-6}
8	0.39×10^{-4}	0.11×10^{-1}	0.48×10^{-2}		0.37×10^{-4}	0.76×10^{-4}	0.25×10^{-6}
9	0.19×10^{-4}	0.75×10^{-2}	0.34×10^{-2}		0.19×10^{-4}	0.38×10^{-4}	0.80×10^{-7}
10	0.97×10^{-5}	0.53×10^{-2}	0.24×10^{-2}		0.93×10^{-5}	0.19×10^{-4}	0.27×10^{-7}
11	0.49×10^{-5}	0.38×10^{-2}	0.17×10^{-2}				
not converge after 11 iterations				$\ x_{\min}^{(10)} - x_{\min}^{(9)}\ \leq 10^{-11}$, the estimated solution, $x_{\min}^{(10)}$, is corrected to 9 decimal places.			

Table 3.3.5: Numerical results for the modified Rosen-Suzuki problem ($p_1 = 10^{-2}$).

k	Classical Algorithm				Modified Algorithm		
	$LB(\tilde{x}^{(k)})$	$UB(\tilde{x}^{(k)})$	$UB(\tilde{x}_{\min}^{(k)})$		$LB(\tilde{x}^{(k)})$	$UB(\tilde{x}^{(k)})$	$UB(\tilde{x}_{\min}^{(k)})$
1	0.50×10^{-2}	15.9	15.9		0.50×10^{-2}	0.34	0.34
2	0.25×10^{-2}	11.3	6.67		0.25×10^{-2}	0.36×10^{-1}	0.27
3	0.13×10^{-2}	8.01	4.08		0.17×10^{-2}	0.85×10^{-2}	0.65×10^{-1}
4	0.63×10^{-3}	5.69	2.71		0.63×10^{-3}	0.28×10^{-2}	0.60×10^{-2}
5	0.31×10^{-3}	4.03	1.87		0.31×10^{-3}	0.12×10^{-2}	0.18×10^{-3}
6	0.16×10^{-3}	2.85	1.30		0.15×10^{-3}	0.52×10^{-3}	0.13×10^{-4}
7	0.78×10^{-4}	2.02	0.92		0.78×10^{-4}	0.24×10^{-3}	0.37×10^{-5}
8	0.39×10^{-4}	1.43	0.65		0.39×10^{-4}	0.12×10^{-3}	0.10×10^{-5}
9	0.20×10^{-4}	1.01	0.46		0.20×10^{-4}	0.57×10^{-4}	0.37×10^{-6}
10	0.98×10^{-5}	0.72	0.32		$\ x_{\min}^{(9)} - x_{\min}^{(8)}\ \leq 10^{-10}$, the estimated solution, $\tilde{x}_{\min}^{(9)}$, is corrected to 9 decimal places.		
11	0.49×10^{-5}	0.51	0.23	not converge after 11 iterations			

Table 3.3.6: Numerical results for the Rosen-kreuser problem ($p_1 = 10^{-2}$).

CHAPTER 4

PENALTY FUNCTION / MULTIPLIER METHODS

4.1 PENALTY FUNCTION METHODS4.1.1 Introduction

Penalty function methods are procedures for approximating constrained optimization problems by unconstrained problems. The approximation is accomplished by adding to the objective function a term that prescribes a high cost for violation of constraints. Penalty function methods are attractive because they offer a simple and straightforward technique for handling constrained problems. Existing penalty function methods can be classified roughly into two categories:

(i) Exact penalty function methods: These are penalty function methods which yield an optimal solution of the constrained problem by means of a single unconstrained minimization (see, for example, Zangwill (1967) and Conn (1973)). These penalty functions generally have discontinuous first derivatives at the solution. For a class of exact penalty functions, Bertsekas (1975) has shown that the existence of discontinuities in first derivatives is a necessary feature. Because of these discontinuities, several efficient methods for unconstrained minimization cannot be utilized. For this reason, exact penalty function methods seem to have fallen from favour in recent years. Very recently, however, these exact penalty functions have been found to be useful for forcing convergence and for step size control in several algorithms (see, for example, Han (1975) and Powell (1977)).

There are other *exact* penalty functions which do not have discontinuities in their first derivatives (e.g., Fletcher (1970b, 1973) and

Martensson (1973)). However, because of their properties, it is more appropriate to refer to these methods as *continuous* augmented Lagrangian methods. These methods will be discussed later on in subsection 4.2.1 with other augmented Lagrangian methods.

(ii) Non-exact penalty function methods: These methods are often referred to as *parametric* methods. Associated with these methods is a parameter σ which determines the severity of the penalty and consequently the degree to which the unconstrained problem approximates the original constrained problem. As $\sigma \rightarrow \infty$, the approximation becomes increasingly accurate. These penalty functions are generally differentiable. However, as σ is increased to yield a good approximating problem, the corresponding structure of the resulting unconstrained problem becomes increasingly ill-conditioned thereby slowing the convergence rate of many unconstrained algorithms that might be applied.

These parametric penalty function methods can be modified into methods which do not explicitly operate with controlling parameters. Different approaches and their modifications are studied by many authors including: Rosenbrock (1960), Huard (1967), Fiacco and McCormick (1967), Lootsma (1968,1970), Schmit and Fox (1965), Morrison (1968) and Kowalik, Osborne and Ryan (1969). The study by Lootsma (1970) shows that these modifications do not make the resulting unconstrained problems any easier to minimize. Because their rate of convergence (the rate at which $\sigma \rightarrow \infty$) can be explicitly controlled, from computational considerations, parametric penalty function methods are more desirable than parametric-free methods.

Penalty function methods are now rarely used. Their theoretical properties are however still of much interest because of their relation and implication to other (more powerful) methods. Our interest in penalty function methods arises from their relation to augmented Lagrangian methods.

In what follows, attention is restricted to parametric penalty function methods. These parametric methods have been analysed in detail by Fiacco and McCormick (1968) and Lootsma (1970). Similar to the treatment of barrier functions, following Lootsma (1970,1972), parametric penalty functions are represented as : (for an ECP)

$$f(\tilde{x}) + \sigma^\mu \sum_{i \in I} \psi(h_i(\tilde{x})) \quad , \quad (4.1-1)$$

where σ is a positive controlling parameter, ψ is a continuously differentiable function of one variable η such that

$$\psi(0) = 0 \quad ,$$

and

$$\psi(-\eta) = \psi(\eta) > 0 \quad , \quad \text{for } \eta \neq 0 \quad .$$

(4.1-2)

The penalty function is then said to be of *order* μ if the derivative ψ' of ψ is analytic and if it has a *zero of order* μ at $\eta=0$.

Simple examples of penalty functions can be obtained from (4.1-1) by putting

$$\psi'(\eta) = (\eta)^\mu \quad , \quad (4.1-3)$$

with μ positive. For $\mu=1$ we have the quadratic penalty function. Again, similar to barrier functions, following the arguments used by Fiacco and McCormick (1968) and Lootsma (1970), the quadratic penalty function is chosen in preference to other penalty functions.

4.1.2 The Quadratic Penalty Function Method

For an equality constrained problem (ECP), the quadratic penalty function has the form:

$$P(\tilde{x}, \sigma) = f(\tilde{x}) + \frac{\sigma}{2} \sum_{i \in I} h_i(\tilde{x})^2, \quad (4.1-4)$$

where the multiple $\frac{1}{2}$ is included to avoid the appearance of a factor of 2 in the derivative.

Convergence

Both the existence of a local minimizing point $\tilde{x}^{(k)}$ of $P(\tilde{x}, \sigma_k)$ for sufficiently large σ_k , and the convergence, as $\sigma_k \rightarrow \infty$, of $\tilde{x}^{(k)}$ to a solution $\tilde{x}^*$ of the constrained problem can be established under mild topological conditions. The conditions amount to requiring that the problem functions be continuous and that the solution set be non-empty.

Theorem 4.1.1 (Fiacco and McCormick (1968), p.57)

- If (a) the problem functions are continuous,
- (b) there exists a non-empty isolated compact set, A^* , of local minima corresponding to a local minimum value f^* , and
- (c) $\{\sigma_k\}$ is a strictly increasing unbounded non-negative sequence,

then, for $P(\tilde{x}, \sigma_k)$ defined by (4.1-4),

- (i) there exists a compact set W such that $A^* \subset \text{interior of } W$ and, for σ_k sufficiently large, the *unconstrained minima* of $P(\tilde{x}, \sigma_k)$ in the interior of W exist and every limit point of any subsequence $\{\tilde{x}^{(k)}\}$ of the minimizing points is in A^*

$$(ii) \quad \lim_{k \rightarrow \infty} \frac{\sigma_k}{2} \sum_{i \in I} h_i(\tilde{x}^{(k)})^2 = 0 ,$$

$$(iii) \quad \lim_{k \rightarrow \infty} f(\tilde{x}^{(k)}) = f^* ,$$

$$(iv) \quad \lim_{k \rightarrow \infty} P(\tilde{x}^{(k)}, \sigma_k) = f^* , \quad P(\tilde{x}^{(k)}, \sigma_k) \leq f^* , \quad \forall k ,$$

$$(v) \quad \{f(\tilde{x}^{(k)})\} \text{ is a monotonically increasing sequence, and}$$

$$(vi) \quad \left\{ \sum_{i \in I} h_i(\tilde{x}^{(k)})^2 \right\} \text{ is a monotonically decreasing sequence.} \quad \square$$

Remark 4.1.1

(a) Since we are only interested in a particular solution, $\tilde{x}^*$, it follows that $A^* = \{\tilde{x}^*\}$ and the compact set W is just a closed bounded neighbourhood of $\tilde{x}^*$. Because W is compact, and there is only one limit point, $\tilde{x}^*$, for any subsequence of $\{\tilde{x}^{(k)}\} \subseteq W$, it follows that, as $k \rightarrow \infty$,

$$\tilde{x}^{(k)} \rightarrow \tilde{x}^* . \quad (4.1-5)$$

(b) Similar to Remark 3.1.2d, if the linear independence condition holds at $\tilde{x}^*$, then as $k \rightarrow \infty$,

$$\tilde{v}^{(k)} = -\sigma_k h(\tilde{x}^{(k)}) \rightarrow \tilde{v}^* . \quad (4.1-6)$$

(c) Similar to the treatment of barrier functions, by utilizing the technique for proving Theorem 3.1.3, it follows that if $\tilde{x}^*$ is a regular local solution of the ECP, then, for $\bar{\sigma}$ sufficiently large, the minimizing point $\bar{\tilde{x}}$ of $P(\tilde{x}, \bar{\sigma})$ is locally unique and $\nabla_{\tilde{x}}^2 P(\bar{\tilde{x}}, \bar{\sigma})$ is positive definite with the smallest eigenvalue bounded away from zero. The implicit function theorem also implies that, for $\sigma < \infty$ sufficiently large, there exists $\tilde{x}(\sigma)$ which is a continuously differentiable function

of σ such that $\tilde{x}(\sigma)$ is the unique (local) minimizer of $P(\tilde{x}, \sigma)$.

Rate of Convergence

Similar to the treatment of barrier functions, by imposing suitable conditions on the problem, the rate of convergence, as $\sigma \rightarrow \infty$, of $\tilde{x}(\sigma)$ to $\tilde{x}^*$ can be determined.

Theorem 4.1.2 (Fiacco and McCormick (1968), p.84)

If $\tilde{x}^*$ is a regular local solution of an ECP, then, for $s = 1/\sigma$ sufficiently small, there exists $\tilde{x}(s)$ which is a continuously differentiable function of s , $\tilde{x}(0) = \tilde{x}^*$, such that $\tilde{x}(s)$ is the unique (local) minimizer of $P(\tilde{x}, \sigma)$. Thus, there exist $\kappa > 0$ ($\kappa = \|d\tilde{x}(0)/ds\| + o(1)$) and $\bar{\sigma} > 0$ such that, $\forall \sigma \geq \bar{\sigma}$, (writing $\tilde{x}(\sigma)$ for $\tilde{x}(s)$)

$$\|\tilde{x}(\sigma) - \tilde{x}^*\| \leq \frac{\kappa}{\sigma} .$$

□

Robustness

Similar to barrier function methods (see §3.1), several robust characteristics of the quadratic penalty function method can be identified as follows:

(i) Convergence can be guaranteed under mild topological conditions (Theorem 4.1.1). Thus, the algorithm will give some sort of a 'solution' even when the Kuhn-Tucker condition is not satisfied.

(ii) If $\tilde{x}^*$ is a regular local solution, then, for σ sufficiently large, $\nabla_x^2 P(\tilde{x}(\sigma), \sigma)$ is positive definite with the smallest eigenvalue bounded away from zero (Remark 4.1.1c). This guarantees that, for $\tilde{x}^{(h)}$ as an approximated minimum of $P(\tilde{x}, \sigma_h)$, there exists a bounded constant $\gamma > 0$ (γ is inversely proportional to the smallest eigenvalue of $\nabla_x^2 P$,

see (3.1-31)) such that, for $\|\nabla_x P(\tilde{x}^{(k)}, \sigma_k)\|$ sufficiently small, (cf., (3.1-32))

$$\|\tilde{x}^{(k)} - \tilde{x}(\sigma_k)\| \leq \gamma \|\nabla_x P(\tilde{x}^{(k)}, \sigma_k)\| \quad (4.1-7)$$

Therefore, from Theorem 4.1.2, it follows that

$$\|\tilde{x}^{(k)} - \tilde{x}^*\| \leq \frac{\kappa}{\sigma_k} + \gamma \|\nabla_x P(\tilde{x}^{(k)}, \sigma_k)\| \quad (4.1-8)$$

This shows that the quadratic penalty function method is stable as regards non-exact minimizations.

4.1.3 The Quadratic Loss Penalty Function Method

When inequality constraints are present, a natural way to handle them is by solving the extended problem (see subsection 2.2.5). Without loss of generality and in the interest of simplicity, it is assumed that only inequality constraints are present. The quadratic penalty function for the extended problem has the form:

$$P((\tilde{x}, \tilde{z}), \sigma) = f(\tilde{x}) + \frac{\sigma}{2} \sum_{i \in J} (g_i(\tilde{x}) - z_i^2)^2 \quad (4.1-9)$$

The minimization with respect to $\tilde{z}$ can be performed explicitly, and this gives

$$z_i(\tilde{x}, \sigma)^2 = \max(0, g_i(\tilde{x})) \quad (4.1-10)$$

The quadratic (loss) penalty function for the original ICP therefore has the form:

$$\bar{P}(\tilde{x}, \sigma) = P((\tilde{x}, \tilde{z}(\tilde{x}, \sigma)), \sigma)$$

$$= f(\tilde{x}) + \frac{\sigma}{2} \sum_{i \in J} [g_i(\tilde{x})]_-^2, \quad (4.1-11)$$

where

$$[a]_- = \min(0, a).$$

Convergence

Because we are implicitly using the quadratic penalty function method to solve the extended problem, convergence follows from Theorem 4.1.1.

Rate of Convergence

The trajectory and rate of convergence results (for an ECP) in Theorem 4.1.2 depend strongly on the solution satisfying the regular solution condition. Because the extended problem is implicitly being solved, this suggests that the regular solution condition needs to be imposed on the extended problem. If so, Theorem 2.2.6 then implies that strict complementarity as well as the regular solution condition needs to be imposed on the original ICP. However, in what follows, it will be shown that strict complementarity need not be imposed on the original ICP.

With strict complementarity

If $\tilde{x}^*$ is a regular local solution of an ICP, and strict complementarity holds at $\tilde{x}^*$, then Theorem 2.2.6 implies that $(\tilde{x}^*, \tilde{z}^*)$ is a regular local solution of the extended problem. Theorem 4.1.2 therefore applies in this case. Note that since

$$\|(\tilde{x}(\sigma), \tilde{z}(\sigma)) - (\tilde{x}^*, \tilde{z}^*)\| = (\|\tilde{x}(\sigma) - \tilde{x}^*\|^2 + \|\tilde{z}(\sigma) - \tilde{z}^*\|^2)^{\frac{1}{2}} \leq \frac{\kappa}{\sigma}, \quad (4.1-12)$$

this gives

$$\|\tilde{x}(\sigma) - \tilde{x}^*\| \leq \frac{\kappa}{\sigma}. \quad (4.1-13)$$

And we have just given a proof of the following Theorem.

Theorem 4.1.3 (Piacco and McCormick (1968), p.82)

If $\tilde{x}^*$ is a regular local solution of a non-degenerate ICP, then, for $s=1/\sigma$ sufficiently small, there exists $\tilde{x}(s)$ which is a continuously differentiable function of s , $\tilde{x}(0)=\tilde{x}^*$, such that $\tilde{x}(s)$ is the unique (local) minimizer of $\bar{P}(\tilde{x},\sigma)$. Thus, there exist $\kappa > 0$ ($\kappa = \|d\tilde{x}(0)/ds\| + o(1)$) and $\bar{\sigma} > 0$ such that, $\forall \sigma \geq \bar{\sigma}$, (writing $\tilde{x}(\sigma)$ for $\tilde{x}(s)$)

$$\|\tilde{x}(\sigma) - \tilde{x}^*\| \leq \frac{\kappa}{\sigma} .$$

□

Without strict complementarity

As a consequence of Theorem 4.1.3, in order to show that strict complementarity need not be imposed, it is only necessary to establish the corresponding trajectory and convergence results for degenerate problems.

Consider solving a degenerate ICP, using the quadratic loss penalty function method. Implicitly, the extended problem is being solved, using the quadratic penalty function method. Because the original ICP is degenerate, Theorem 2.2.6 implies that the conditions of Theorem 4.1.2 cannot be satisfied, and hence a slower rate of convergence is expected. In what follows, it will be shown that the slower rate of convergence only occurs in the slack variables. Since the slack variables do not explicitly enter into the computation, the usual rate of convergence is retained.

Results in this subsection depend heavily on the properties of convex functionals (Appendix 1) and a stronger version of the implicit function theorem (Appendix 2). The trajectory result will be established by showing that $\bar{P}(\tilde{x},\sigma)$ satisfies certain convexity conditions. The new implicit function theorem then gives that the minimizing point $\tilde{x}(\sigma)$ of $\bar{P}(\tilde{x},\sigma)$ is locally unique and a continuous function of σ . The minimizing point $\tilde{x}(\sigma)$ can also be identified as the (local) minimizing point, for

some index set R , of $P^R(\underline{x}, \sigma) = f(\underline{x}) + \frac{\sigma}{2} \sum_{i \in R} [g_i(\underline{x})]^2$. By using this identification and properties of the problem $P[R]$ (see subsection 2.2.1), the rate of convergence, as $\sigma \rightarrow \infty$, of $\underline{x}(\sigma)$ to $\underline{x}^*$ can be derived.

Definition

$$M(\underline{x}) = \{i \mid i \in J, g_i(\underline{x}) < 0\}, \quad (4.1-14)$$

$$Z(\underline{x}) = \{i \mid i \in J, g_i(\underline{x}) = 0\}. \quad (4.1-15)$$

In what follows, it is assumed that:

- (i) the ICP under consideration is degenerate with $\underline{x}^*$ as a regular local solution;
- (ii) $\{\sigma_k\}$ is a strictly increasing unbounded sequence;
- (iii) for σ_k sufficiently large, a minimum $\underline{x}^{(k)}$ of $\bar{P}(\underline{x}, \sigma_k)$ exists, and, as $k \rightarrow \infty$, $\underline{x}^{(k)}$ converges to $\underline{x}^*$;
- (iv) as $k \rightarrow \infty$, $u_i^{(k)} = \sigma_k \max(-g_i(\underline{x}^{(k)}), 0)$ converges to u_i^* , $\forall i \in J$.

For convenience, the subscript or superscript k will be dropped. σ_k , $\underline{x}^{(k)}$ and $\underline{u}^{(k)}$ are written as $\bar{\sigma}$, $\bar{\underline{x}}$ and $\bar{\underline{u}}$, respectively. By using a sufficiently large value of $\bar{\sigma}$, $\bar{\underline{x}}$ (or $\bar{\underline{u}}$) can be made as close to $\underline{x}^*$ (or $\underline{u}^*$) as required.

Lemma 4.1.1

If $\bar{\sigma}$ is sufficiently large, then, for $\bar{\underline{x}}$ as a (local) minimizing point of $\bar{P}(\underline{x}, \bar{\sigma})$, it follows that

$$J_1(\underline{x}^*) \subseteq M(\bar{\underline{x}}) \subseteq M(\bar{\underline{x}}) \cup Z(\bar{\underline{x}}) \subseteq J_1(\underline{x}^*) \cup J_2(\underline{x}^*). \quad (4.1-16)$$

Proof

Since the constraint functions are continuous, it follows that, if $\bar{\sigma}$ is sufficiently large, then $g_i(\bar{x}) > 0$ for every constraint g_i which is inactive at $\tilde{x}^*$. Thus

$$M(\bar{x}) \cup Z(\bar{x}) \subseteq J_1(\tilde{x}^*) \cup J_2(\tilde{x}^*) . \quad (4.1-17)$$

Similarly, for every non-degenerate active constraint g_i , because $\sigma_k \max(-g_i(\tilde{x}^{(k)}), 0)$ converges to $u_i^* > 0$, it follows that $g_i(\bar{x}) < 0$ for $\bar{\sigma}$ sufficiently large. Thus

$$J_1(\tilde{x}^*) \subseteq M(\bar{x}) , \quad (4.1-18)$$

and (4.1-16) holds. □

Lemma 4.1.2

Let $\tilde{x}'$ be a given point, then there exists $\delta = \delta(\tilde{x}') > 0$ such that, $\forall \tilde{x} \in N(\tilde{x}', \delta)$,

$$M(\tilde{x}') \subseteq M(\tilde{x}) \subseteq M(\tilde{x}) \cup Z(\tilde{x}) \subseteq M(\tilde{x}') \cup Z(\tilde{x}') . \quad (4.1-19)$$

Proof

Because the constraint functions are continuous, there exists a small neighbourhood, $N(\tilde{x}', \delta)$, of $\tilde{x}'$ such that, $\forall \tilde{x} \in N(\tilde{x}', \delta)$,

$$g_i(\tilde{x}) < 0 \quad \text{if} \quad g_i(\tilde{x}') < 0 ,$$

and

$$g_i(\tilde{x}) > 0 \quad \text{if} \quad g_i(\tilde{x}') > 0 . \quad \square$$

Now consider any index set R such that

$$M(\bar{x}) \subseteq R \subseteq M(\bar{x}) \cup Z(\bar{x}). \quad (4.1-20)$$

Lemma 4.1.1 and Theorem 2.2.1 give that $\bar{x}^*$ is a regular (local) solution of the (equality constrained) problem $P[R]$. If the quadratic penalty function method is applied to $P[R]$, it follows that

$$P^R(\bar{x}, \sigma) = f(\bar{x}) - \frac{\sigma}{2} \sum_{i \in R} [g_i(\bar{x})]^2. \quad (4.1-21)$$

By putting $\sigma = \bar{\sigma}$, it follows that

$$\nabla_x P^R(\bar{x}, \bar{\sigma}) = \nabla_x \bar{P}(\bar{x}, \bar{\sigma}) = 0. \quad (4.1-22)$$

Therefore, $\bar{x}$ is also a stationary point and hence a minimizing point of $P^R(\bar{x}, \bar{\sigma})$. Remark 4.1.1c gives that $\nabla_x^2 P^R(\bar{x}, \bar{\sigma})$ is positive definite with the smallest eigenvalue bounded away from zero. Because $\nabla_x^2 P^R(\bar{x}, \bar{\sigma})$ is continuous in $\bar{x}$ and is positive definite at $\bar{x} = \bar{x}$, it follows that $\nabla_x^2 P^R(\bar{x}, \bar{\sigma})$ is positive definite in a small enough neighbourhood of $\bar{x}$. Thus, there exists $\delta = \delta(\bar{x}, \bar{\sigma}) > 0$ such that, for any index set R satisfying (4.1-20), $\nabla_x^2 P^R(\bar{x}, \bar{\sigma})$ is positive definite for all $\bar{x} \in N(\bar{x}, \delta)$. Note that the result is trivially true if the problem is non-degenerate.

Lemma 4.1.3

Assume that a constant $\sigma > 0$ and a point $\bar{x}$ are given. If there exists $\delta > 0$ such that the result of Lemma 4.1.2 holds at $\bar{x}$ and that, for every index set R satisfying (4.1-20), $\nabla_x^2 P^R(\bar{x}, \sigma)$ is positive definite for all $\bar{x} \in N(\bar{x}, \delta)$, then $\bar{P}(\bar{x}, \sigma)$ is strictly convex on $N(\bar{x}, \delta)$.

Proof

Given a point $\bar{x}' \in N(\bar{x}, \delta)$ and a direction $\bar{t}$ ($\|\bar{t}\| = 1$), there exists $a > 0$ such that, for all R satisfying (4.1-20),

$$\lim_{\alpha \rightarrow 0} \frac{\tilde{t}^T (\nabla_{\tilde{x}} P^R(\tilde{x}' + \alpha \tilde{t}, \sigma) - \nabla_{\tilde{x}} P^R(\tilde{x}', \sigma))}{\alpha} = \tilde{t}^T \cdot \nabla_{\tilde{x}}^2 P^R(\tilde{x}', \sigma) \cdot \tilde{t} \geq a \quad (4.1-23)$$

From the definition of $\bar{P}(\tilde{x}, \sigma)$, Lemma 4.1.2 implies that, for all α sufficiently small, $\nabla_{\tilde{x}} \bar{P}(\tilde{x}' + \alpha \tilde{t}, \sigma)$ can be identified with $\nabla_{\tilde{x}} P^R(\tilde{x}' + \alpha \tilde{t}, \sigma)$ for some R such that

$$M(\tilde{x}) \subseteq M(\tilde{x}') \subseteq R \subseteq M(\tilde{x}') \cup Z(\tilde{x}') \subseteq M(\tilde{x}) \cup Z(\tilde{x}) \quad (4.1-24)$$

Hence, by (4.1-23), there exists an $a' > 0$ such that

$$\frac{\tilde{t}^T (\nabla_{\tilde{x}} \bar{P}(\tilde{x}' + \alpha \tilde{t}, \sigma) - \nabla_{\tilde{x}} \bar{P}(\tilde{x}', \sigma))}{\alpha} \geq a' \quad (4.1-25)$$

for all α sufficiently small. It follows that the component of $\nabla_{\tilde{x}} \bar{P}$ along any line in $N(\tilde{x}, \delta)$ increases strictly monotonically, and it is then a straightforward application of the mean value theorem to show that $\bar{P}(\tilde{x}, \sigma)$ is strictly convex on $N(\tilde{x}, \delta)$. $\square$

Since $\nabla_{\tilde{x}} \bar{P}(\bar{\tilde{x}}, \bar{\sigma}) = 0$, Lemma 4.1.3 implies that $\bar{\tilde{x}}$ is the unique (local) minimizing point of $\bar{P}(\tilde{x}, \bar{\sigma})$. In fact, because $\nabla_{\tilde{x}}^2 P^{M(\bar{\tilde{x}})}(\bar{\tilde{x}}, \bar{\sigma})$ is positive definite, there exist $a > 0$ and $\delta > 0$ such that Lemma 4.1.2 holds and that, $\forall \tilde{x} \in N(\bar{\tilde{x}}, \delta)$,

$$\begin{aligned} \bar{P}(\tilde{x}, \bar{\sigma}) - \bar{P}(\bar{\tilde{x}}, \bar{\sigma}) &\geq P^{M(\bar{\tilde{x}})}(\tilde{x}, \bar{\sigma}) - P^{M(\bar{\tilde{x}})}(\bar{\tilde{x}}, \bar{\sigma}) \\ &\geq a \|\tilde{x} - \bar{\tilde{x}}\|^2 \end{aligned} \quad (4.1-26)$$

Thus, $\bar{\tilde{x}}$ is a unique minimum of 2nd order of $\bar{P}(\tilde{x}, \bar{\sigma})$.

Theorem 4.1.4

Let $\tilde{x}^*$ be a regular local solution of an ICP, then for σ sufficiently large, there exists $\tilde{x}(\sigma)$ as a *continuous* function of σ such that $\tilde{x}(\sigma)$ is the unique local minimizer of $\bar{P}(\tilde{x}, \sigma)$. Furthermore, there exist $\kappa > 0$ and $\sigma' > 0$ such that, $\forall \sigma \geq \sigma'$,

$$\|\tilde{x}(\sigma) - \tilde{x}^*\| \leq \frac{\kappa}{\sigma}.$$

Proof

In order to show that $\tilde{x}(\sigma)$ is a continuous function of σ , it is sufficient to show that $\tilde{x}(\sigma)$ is a continuous function of σ in a small neighbourhood of $\bar{\sigma}$, for every $\bar{\sigma}$ sufficiently large.

Given $\bar{\sigma}$ sufficiently large, let $\bar{x}$ be the corresponding minimizer of $\bar{P}(\tilde{x}, \bar{\sigma})$. From Lemma 4.1.1 and Remark 4.1.1c, for every index set R satisfying (4.1-20), $\nabla_x^2 P^R(\bar{x}, \bar{\sigma})$ is positive definite with the smallest eigenvalue bounded away from zero. Because $\nabla_x^2 P^R(\tilde{x}, \sigma)$ is continuous in both $\tilde{x}$ and σ , there exist $\delta_1 > 0$ and $\delta_2 > 0$ such that, for every index set R satisfying (4.1-20), $\nabla_x^2 P^R(\tilde{x}, \sigma)$ is positive definite $\forall \tilde{x} \in N(\bar{x}, \delta_1)$ and $\forall \sigma \in N(\bar{\sigma}, \delta_2)$. Lemmas 4.1.2 and 4.1.3 then give that there exists $\delta_3 > 0$ such that, $\forall \sigma \in N(\bar{\sigma}, \delta_2)$, $\bar{P}(\tilde{x}, \sigma)$ is strictly convex on $N(\bar{x}, \delta_3)$. Properties of convex functionals (Appendix 1) and the new implicit function theorem (Appendix 2) then give that there exists a continuous function $\tilde{x}(\sigma)$ such that, for every σ in a small enough neighbourhood of $\bar{\sigma}$, $\tilde{x}(\sigma)$ is the unique local minimizer of $\bar{P}(\tilde{x}, \sigma)$, with $\tilde{x}(\bar{\sigma}) = \bar{x}$. Thus, the trajectory result has been established.

The rate of convergence of $\tilde{x}(\sigma)$ to $\tilde{x}^*$ can now be derived using the technique of identification. Note that $\tilde{x}(\sigma)$ is also the (local) minimizing point of $P^{M(\tilde{x}(\sigma'))}(\tilde{x}, \sigma)$. For every index set R satisfying the condition of Theorem 2.2.1, Theorem 4.1.2 gives that there exist

$\kappa(R) > 0$ and $\sigma'(R) > 0$ such that, $\forall \sigma \geq \sigma'(R)$,

$$\| \tilde{x}^R(\sigma) - \tilde{x}^* \| \leq \frac{\kappa(R)}{\sigma}, \quad (4.1-27)$$

where $\tilde{x}^R(\sigma)$ denotes the minimizing point of $P^R_{(x, \sigma)}$. Now let

$$\kappa = \max_R \kappa(R), \quad (4.1-28)$$

and

$$\sigma' = \max_R \sigma'(R). \quad (4.1-29)$$

Then, by identifying $\tilde{x}(\sigma)$ with $\tilde{x}^R(\sigma)$ for some R , it follows that,
 $\forall \sigma \geq \sigma'$,

$$\| \tilde{x}(\sigma) - \tilde{x}^* \| \leq \frac{\kappa}{\sigma}. \quad \square$$

Note that, from Theorem 4.1.3, the results of Theorem 4.1.4 holds trivially if the problem is non-degenerate. Therefore, the trajectory and convergence results of the quadratic loss penalty function method are independent of degeneracy considerations.

From (4.1-10), it follows that, for a degenerate constraint $g_{\tilde{z}}$,

$$\begin{aligned} (z_{\tilde{z}}(\sigma) - z_{\tilde{z}}^*)^2 &= z_{\tilde{z}}(\sigma)^2 = 0(g_{\tilde{z}}(\tilde{x}(\sigma))) \\ &= 0(\| \tilde{x}(\sigma) - \tilde{x}^* \|) \end{aligned} \quad (4.1-30)$$

This implies that

$$\| (\tilde{x}(\sigma), z_{\tilde{z}}(\sigma)) - (\tilde{x}^*, z_{\tilde{z}}^*) \|^2 = \| \tilde{x}(\sigma) - \tilde{x}^* \|^2 + \| z_{\tilde{z}}(\sigma) - z_{\tilde{z}}^* \|^2 = 0(\| \tilde{x}(\sigma) - \tilde{x}^* \|) \quad (4.1-31)$$

The results of Theorem 4.1.4 then give that there exists $\kappa > 0$ such that

$$\|(\tilde{x}(\sigma), \tilde{z}(\sigma)) - (\tilde{x}^*, \tilde{z}^*)\| \leq \left(\frac{\kappa}{\sigma}\right)^{\frac{1}{2}}. \quad (4.1-32)$$

This result indicates that the rate of convergence of the quadratic penalty function method, applying to the extended problem, is of order $O(1/\sigma)$ or $O(1/\sigma^{\frac{1}{2}})$ depending on whether or not strict complementarity is satisfied by the original ICP. This phenomenon is in a sense similar to that of the classical barrier function method (see §3.1 and §3.2). However, for the quadratic penalty function method, the slower rate of convergence only occurs in the slack variables. Since slack variables are usually eliminated explicitly, the slower rate of convergence does not materialize in the computation.

Robustness

The robust characteristics of the quadratic loss penalty function method are the same as those of the quadratic penalty function method.

(i) A minimum $\tilde{x}^{(k)}$ of $\bar{P}(\tilde{x}, \sigma_k)$ exists and converges to $\tilde{x}^*$ under mild topological conditions. This is because, implicitly, the extended problem is being solved using the quadratic penalty function method. Hence, Theorem 4.1.1 applies, and $(\tilde{x}^{(k)}, \tilde{z}^{(k)}) \rightarrow (\tilde{x}^*, \tilde{z}^*)$.

(ii) The solution estimate $\tilde{x}^{(k)}$ is stable with respect to non-exact minimization. This is because, for σ_k sufficiently large and $\tilde{x}^{(k)}$ as the 'approximated' minimum of $\bar{P}(\tilde{x}, \sigma)$, it follows from Lemmas 4.1.1 and 4.1.2 that $J_1(\tilde{x}^*) \subseteq M(\tilde{x}^{(k)}) \subseteq J_1(\tilde{x}^*) \cup J_2(\tilde{x}^*)$. Hence, it is possible to identify $\nabla_{\tilde{x}} \bar{P}(\tilde{x}^{(k)}, \sigma_k)$ with $\nabla_{\tilde{x}} P^{M(\tilde{x}^{(k)})}(\tilde{x}^{(k)}, \sigma_k)$. If $\tilde{x}^*$ is a regular local solution, then (4.1-8) gives that

$$\|\tilde{x}^{(k)} - \tilde{x}^*\| \leq \frac{\kappa}{\sigma_k} + \gamma \cdot \|\nabla_{\tilde{x}} \bar{P}(\tilde{x}^{(k)}, \sigma)\|, \quad (4.1-33)$$

where $\kappa = \max_M \kappa(M)$ and $\gamma = \max_M \gamma(M)$. The quantities $\gamma(M) > 0$ and $\kappa(M) > 0$ (for the problem $P[M]$) are defined in (4.1-7) and Theorem 4.1.2, respectively.

The analysis in this subsection shows that the trajectory and convergence results of the quadratic loss penalty function method depend only on $\underline{x}^*$ being a regular local solution. When the problem is degenerate, jump discontinuities of $\nabla_x^2 \bar{P}$ may occur at the minimum $\underline{x}(\sigma)$, and hence, instead of differentiability, only continuity of $\underline{x}(\sigma)$ can be guaranteed.

4.1.4 Difficulties with Penalty Function Methods

Computational difficulties encountered by penalty function methods are similar to those encountered by barrier function methods.

(i) *Unbounded minima:* Because Theorem 4.1.1 only guarantees that $\underline{x}^{(k)}$ minimizes $P(\underline{x}, \sigma_k)$ within a compact set, it is possible for the penalty function to be unbounded below (see Powell (1972)) unless certain conditions are satisfied. For the case of the quadratic penalty function, the required condition is termed 'the quadratic growth condition' by Rockafellar (1974). This 'unbounded' difficulty can simply be overcome by either rescaling the objective function (e.g., replace $f(x)$ by $\exp(f(x))$), or increasing the penalty parameter σ so as to enlarge the region of attraction. Another technique is by using other 'faster growth' penalty functions.

(ii) *Ill-conditioning:* The numerical solution of the unconstrained subproblem becomes increasingly more difficult as the penalty parameter σ becomes large. The cause of this phenomenon has been analysed by Murray (1971) and Lootsma (1969) in terms of the condition number of the hessian matrix $\nabla_x^2 P$ at the minimum of the penalty function. Similar to barrier

function methods, extrapolation procedures can be employed to avoid the need for using excessively large values of σ . The ill-conditioned unconstrained subproblem can also be avoided by using the penalty trajectory method developed by Murray (1969), or similar methods developed by Biggs (1972, 1975 and 1978).

(iii) *Jump discontinuities in the hessian:* When inequality constraints are present, from (4.1-11), it follows that the hessian matrix $\nabla_x^2 \bar{p}(\underline{x}, \sigma)$ has jump discontinuities at points where $g_i(\underline{x}) = 0$ for some i . If strict complementarity holds at $\underline{x}^*$, then, for any $\sigma < \infty$, the discontinuities are bounded away from a small neighbourhood of $\underline{x}(\sigma)$, and no extra difficulties from these discontinuities are expected. If strict complementarity does not hold at $\underline{x}^*$, jump discontinuities may occur at $\underline{x}(\sigma)$. However, numerical difficulties may not necessarily follow, or can be simply overcome. More detailed arguments and analysis will be presented in subsection 4.2.3 for the more interesting case of augmented Lagrangian methods.

4.2 MULTIPLIER/AUGMENTED LAGRANGIAN METHODS

4.2.1 Introduction

Multiplier or dual methods were originally developed for convex or locally convex problems. For a convex ECP, where $f(\underline{x})$ is convex and $h_i(\underline{x}), i \in I$, are affine, the *classical Lagrangian function*, $L_0(\underline{x}, \underline{v})$, is a convex function of $\underline{x}$:

$$L_0(\underline{x}, \underline{v}) = f(\underline{x}) - \underline{v}^T \cdot \underline{h}(\underline{x}) \quad (4.2-1)$$

The *classical dual function*, $D_0(\underline{v})$, is a concave function of the multiplier $\underline{v}$:

$$D_0(\underline{v}) = \inf_{\underline{x}} L_0(\underline{x}, \underline{v}) \quad (4.2-2)$$

If $L_0(\underline{x}, \underline{v})$ is strictly convex in $\underline{x}$, then

$$D_0(\underline{v}) = L_0(\underline{x}(\underline{v}), \underline{v}) = \min_{\underline{x}} L_0(\underline{x}, \underline{v}) \quad (4.2-3)$$

And the dual problem is to determine a $\underline{v}^*$ such that $\underline{v}^*$ maximizes $D_0(\underline{v})$.

The merging of the multiplier and the penalty function methods resulted from two conceptually different approaches:

(i) Shifted Penalty Function: This idea was introduced by Powell (1969) into the quadratic penalty function method. For an ECP, the compound function takes the form:

$$P(\underline{x}, \underline{\theta}, \sigma) = f(\underline{x}) + \frac{\sigma}{2} \sum_{i \in I} (h_i(\underline{x}) - \theta_i)^2 \quad (4.2-4)$$

By adjusting the shift parameter $\underline{\theta}$, it is no longer necessary to increase

without bound the controlling parameter σ . The shift parameter, however, is essentially the multiplier parameter.

(ii) Multiplier Method for the Convexified Problem: For a non-convex problem (ECP) satisfying certain local conditions at $\tilde{x}^*$, the problem can be made locally convex by adding a term $\frac{\sigma}{2} \sum_{i \in I} h_i(\tilde{x})^2$ to the objective function. The modified problem, with the objective function $f(\tilde{x}) + \frac{\sigma}{2} \sum_{i \in I} h_i(\tilde{x})^2$, can then be solved using some multiplier algorithms. This idea of forcing (local) convexity by adding a penalty term to the objective function dates back to a paper by Arrow and Solow (1958). The first computational algorithm, however, was proposed by Hestenes (1969).

These methods have been called by several names: such as multiplier, dual, shifted penalty function, modified Lagrangian and augmented Lagrangian methods. In this study, these methods are referred to as *multiplier* or *augmented Lagrangian* methods.

Consider an ECP. Assume that, at a solution $\tilde{x}^*$, certain regularity conditions are satisfied, then, for σ sufficiently large, the modified problem is locally convex. For the modified problem, corresponding to (4.2-1) and (4.2-2), there are:

the *augmented Lagrangian function*,

$$L_{\sigma}(\tilde{x}, \tilde{v}) = f(\tilde{x}) - \tilde{v}^T \cdot \tilde{h}(\tilde{x}) + \frac{\sigma}{2} \tilde{h}(\tilde{x})^T \cdot \tilde{h}(\tilde{x}) ; \quad (4.2-5)$$

and the *augmented dual function*,

$$D_{\sigma}(\tilde{v}) = \min_{\tilde{x}} L_{\sigma}(\tilde{x}, \tilde{v}) . \quad (4.2-6)$$

Locally, $D_{\sigma}(\tilde{v})$ is well defined. And the augmented dual problem is to find a $\tilde{v}^*$ which maximizes $D_{\sigma}(\tilde{v})$.

Several different methods can be derived from various different techniques for maximizing $D_{\sigma}(\underline{v})$. These methods are sequential in the sense that they require the solutions of a sequence of unconstrained minimization subproblems, $\min_x L_{\sigma}(\underline{x}, \underline{v})$.

Modifications of these methods, so that 'incompleted' minimization of $L_{\sigma}(\underline{x}, \underline{v})$ can be used, have been studied by several authors including Miele, *et al.* (1972) and Bertsekas (1975). Topia (1977) has also proposed methods in which only one step in the minimization of $L_{\sigma}(\underline{x}, \underline{v})$ is taken before the multiplier $\underline{v}$ is updated. These modifications can be viewed to some degree as a shift from methods which attempt to find $(\underline{x}^*, \underline{v}^*)$ by maximizing $D_{\sigma}(\underline{v})$ to methods which attempt to find $(\underline{x}^*, \underline{v}^*)$ directly as the saddle point of $L_{\sigma}(\underline{x}, \underline{v})$. Han (1976, 1977) has also proposed methods for finding the saddle point $(\underline{x}^*, \underline{v}^*)$ of $L_{\sigma}(\underline{x}, \underline{v})$ using quadratic programming subproblems.

Fletcher (1970b, 1973) and Martensson (1973) have also developed related techniques where the multiplier $\underline{v}$ is adjusted continuously as a function of $\underline{x}$. The problem, therefore, can be solved by a single unconstrained minimization of the function $L_{\sigma}(\underline{x})$:

$$L_{\sigma}(\underline{x}) = L_{\sigma}(\underline{x}, \underline{v}(\underline{x})) \quad (4.2-7)$$

These continuous augmented Lagrangian methods are sometimes referred to as 'exact penalty function methods'. A notable disadvantage of this continuous approach is that the definition of the multiplier $\underline{v}(\underline{x})$ depends on first derivatives of the problem functions. Second derivatives of the problem functions (or estimates of) are consequently required in order to use gradient techniques for the unconstrained minimization of $L_{\sigma}(\underline{x})$.

In this study, attention is restricted to *sequential* augmented

Lagrangian methods: methods which solve the problem by maximizing the augmented dual function $D_{\sigma}(\underline{v})$, requiring a solution of an unconstrained subproblem for each evaluation of $D_{\sigma}(\underline{v})$. Our interest in these methods is motivated in part by their relation to penalty function methods and by the following arguments:

- (i) Modifications to accept 'incompleted' minimization of $L_{\sigma}(\underline{x}, \underline{v})$ usually require derivative information of the problem functions.
- (ii) Most algorithms for unconstrained minimization do have a super-linear rate of convergence, so obtaining high accuracy for 'completed' as opposed to 'incompleted' minimization need not be expensive.
- (iii) Most efficient unconstrained algorithms are of quasi-Newton type. These algorithms usually generate an estimate of $\nabla_{\underline{x}}^2 L_{\sigma}$. With this estimate and a good starting point, subsequent unconstrained minimizations are relatively easy.
- (iv) The resulting algorithms are usually very simple and robust using an established unconstrained minimization routine (see, for example, Powell (1969) and Hestenes (1969)).
- (v) An analysis of the augmented dual problem is of theoretical interest.

4.2.2 Properties of Augmented Lagrangian Methods

In this subsection, augmented Lagrangian methods are developed for equality constrained problems. Subsequently, they will be extended to include inequality constraints.

Definition (The Augmented Lagrangian Function)

For an ECP, the *augmented Lagrangian function* is defined as:

$$L_{\sigma}(\underline{x}, \underline{v}) = f(\underline{x}) - \underline{v}^T \cdot \underline{h}(\underline{x}) + \frac{\sigma}{2} \underline{h}(\underline{x})^T \cdot \underline{h}(\underline{x}) \quad (4.2-8)$$

Remark 4.2.1

(a) The parameter σ serves both as a penalty parameter in the penalty function sense, and as a convexifying parameter in the sense that it forces (local) convexity to hold.

(b) The parameter σ can be allowed to vary from constraint to constraint. But since this is only equivalent to a rescaling of the problem, without loss of generality, σ is taken to be the same for all constraints.

(c) Without loss of generality and in the interest of simplicity, the parameter σ is taken to be sufficiently large, and then held constant throughout the computation. The effect of increasing σ can be taken into account using related results developed for penalty function method.

Theorem 4.2.1 (Fletcher(1975), Theorem 1)

Let $\underline{x}^*$ be a regular local solution of an ECP, then there exists $\bar{\sigma} > 0$ such that, $\forall \sigma \geq \bar{\sigma}$, $\nabla_{\underline{x}}^2 L_{\sigma}(\underline{x}^*, \underline{v}^*)$ is positive definite, and $\underline{x}^*$ is a unique local minimizer of $L_{\sigma}(\underline{x}, \underline{v}^*)$. $\square$

Theorem 4.2.1 is proved using techniques which are used in the proof of Theorem 3.1.3. Because $\nabla_{\underline{x}}^2 L_{\sigma}(\underline{x}^*, \underline{v}^*)$ is positive definite, the implicit function theorem implies that there exist open neighbourhoods $N(\underline{x}^*, \delta) \subset E^n$, $N(\underline{v}^*, \delta) \subset E^k$, and $\underline{x}(\underline{v}, \sigma)$ as a continuously differentiable function of $\underline{v}$ such that, for every $\underline{v} \in N(\underline{v}^*, \delta)$, $\underline{x}(\underline{v}, \sigma) \in N(\underline{x}^*, \delta)$ is a unique local minimizer of $L_{\sigma}(\underline{x}, \underline{v})$, with $\underline{x}(\underline{v}^*, \sigma) = \underline{x}^*$.

Definition (The Augmented Dual Function)

For an ECP with a regular local solution $\tilde{x}^*$, locally, the augmented dual function is defined as:

$$D_{\sigma}(\tilde{v}) = L_{\sigma}(\tilde{x}(\tilde{v}, \sigma), \tilde{v}) = \min_{\tilde{x}} L_{\sigma}(\tilde{x}, \tilde{v}) . \quad (4.2-9)$$

Theorem 4.2.2 (Fletcher (1975), p.324)

Assume that $\tilde{x}^*$ is a regular local solution of an ECP, and that the conditions of Theorem 4.2.1 are satisfied, then $\nabla_{\tilde{v}}^2 D_{\sigma}(\tilde{v}^*)$ is negative definite, and $\tilde{v}^*$ is a unique local maximizer of $D_{\sigma}(\tilde{v})$. $\square$

The result of this Theorem was originally proved by Buys (1972). Using direct manipulation (see, for example, Buys (1972) or Fletcher (1975)), it follows that

$$D_{\sigma}(\tilde{v}) = L_{\sigma}(\tilde{x}(\tilde{v}, \sigma), \tilde{v}) , \quad (4.2-10)$$

$$\nabla_{\tilde{v}} D_{\sigma}(\tilde{v}) = -\tilde{h}(\tilde{x}(\tilde{v}, \sigma)) , \quad (4.2-11)$$

and

$$\nabla_{\tilde{v}}^2 D_{\sigma}(\tilde{v}) = -N(\tilde{x}(\tilde{v}, \sigma))^T \left(\nabla_{\tilde{x}}^2 L_{\sigma}(\tilde{x}(\tilde{v}, \sigma), \tilde{v}) \right)^{-1} N(\tilde{x}(\tilde{v}, \sigma)) , \quad (4.2-12)$$

where

$$N(\tilde{x}) = [\nabla h_1(\tilde{x}), \nabla h_2(\tilde{x}), \dots, \nabla h_{\ell}(\tilde{x})] \text{ is a matrix of rank } \ell .$$

Consider maximizing $D_{\sigma}(\tilde{v})$. Because $D_{\sigma}(\tilde{v})$ is twice differentiable, from (4.2-11) and (4.2-12), a Newton iteration for updating $\tilde{v}$ has the form:

$$\tilde{v}^{(k+1)} = \tilde{v}^{(k)} - \left[N(\tilde{x}^{(k)})^T \left(\nabla_{\tilde{x}}^2 L_{\sigma}(\tilde{x}^{(k)}, \tilde{v}^{(k)}) \right)^{-1} N(\tilde{x}^{(k)}) \right]^{-1} \tilde{h}^{(k)} , \quad (4.2-13)$$

where $\tilde{x}(\tilde{v}^{(k)}, \sigma)$ and $\tilde{h}(\tilde{x}(\tilde{v}^{(k)}, \sigma))$ are written as $\tilde{x}^{(k)}$ and $\tilde{h}^{(k)}$, respectively.

The Newton update (4.2-13) requires information about the second

derivatives of $f(\underline{x})$ and $h(\underline{x})$. This requirement is a serious set back. Due to the complexity or dimensions of the problem, it may not be economical nor practical to compute $\nabla^2 f$ and $\nabla^2 h$. However, as pointed out by Fletcher (1975), most good unconstrained minimization routines will give an estimate of $\nabla_{\underline{x}}^2 L_{\sigma}$ as a by-product. The Newton updating formula (4.2-13) may therefore be estimated without having to compute $\nabla^2 f$ and $\nabla^2 h$.

Because $\underline{x}^{(k)}$ is a minimizer of $L_{\sigma}(\underline{x}, \underline{v}^{(k)})$, it follows that

$$\nabla f(\underline{x}^{(k)}) - \sum_{i \in I} (v_i^{(k)} - \sigma h_i(\underline{x}^{(k)})) \nabla h_i(\underline{x}^{(k)}) = 0 \quad (4.2-14)$$

Eqn. (4.2-14) implies a natural formula, proposed independently by Powell (1969) and Hestenes (1969), for updating the multiplier:

$$\underline{v}^{(k+1)} = \underline{v}^{(k)} - \sigma \underline{h}^{(k)} \quad (4.2-15)$$

Substituting from (4.2-11), (4.2-15) becomes

$$\underline{v}^{(k+1)} - \underline{v}^{(k)} = \sigma \nabla_{\underline{v}} D_{\sigma}(\underline{v}^{(k)}) \quad (4.2-16)$$

Therefore, the Powell/Hestenes method is just a fixed step size gradient method, and only linear convergence can be expected.

Definition

$$\underline{\Lambda}(\underline{v}, \sigma) = \underline{v} - \sigma \underline{h}(\underline{x}(\underline{v}, \sigma)) \quad (4.2-17)$$

or, alternatively, in superscripted form

$$\underline{\Lambda}^{(k)} = \underline{v}^{(k)} - \sigma \underline{h}^{(k)} \quad (4.2-17a)$$

The Powell/Hestenes update now becomes

$$\tilde{v}^{(k+1)} = \tilde{\Lambda}^{(k)} \quad (4.2-18)$$

Theorem 4.2.3 (Bertsekas (1976))

Assume that $\tilde{x}^*$ is a regular local solution of an ECP, and that the conditions of Theorem 4.2.1 are satisfied, then, for an arbitrary bounded set B , there exist $\bar{\sigma} > 0$ and $\kappa > 0$ such that, for every $\sigma \geq \bar{\sigma}$ and $\tilde{v} \in B$, the augmented Lagrangian function $L_\sigma(\tilde{x}, \tilde{v})$ has a unique minimizing point $\tilde{x}(\tilde{v}, \sigma)$ within some open neighbourhood of $\tilde{x}^*$. Furthermore, $\forall \sigma \geq \bar{\sigma}$ and $\tilde{v} \in B$,

$$\|\tilde{x}(\tilde{v}, \sigma) - \tilde{x}^*\| \leq \frac{\kappa \|\tilde{v} - \tilde{v}^*\|}{\sigma},$$

and

$$\|\tilde{\Lambda}(\tilde{v}, \sigma) - \tilde{v}^*\| \leq \frac{\kappa \|\tilde{v} - \tilde{v}^*\|}{\sigma} \quad \square$$

Theorem 4.2.3 implies that, in using the Powell/Hestenes update, the rate of convergence can be made as fast as required by choosing σ sufficiently large. From (4.2-14), as a by-product of the unconstrained minimization, the Powell/Hestenes update reduces to a simple substitution. Because of its simplicity and arbitrary fast rate of convergence, the Powell/Hestenes update is often used instead of Newton-like updates, such as in (4.2-13).

Remark 4.2.2

(a) Theorem 4.2.3 is slightly stronger than the theorem given by Bertsekas (1976). In his theorem, Bertsekas requires an extra condition that the hessian matrices $\nabla^2 f, \{\nabla^2 h_i\}$ be Lipschitz continuous within an open neighbourhood of $\tilde{x}^*$. This extra requirement can be simply removed by replacing Lemma 2 of Bertsekas (1976) by a standard implicit function theorem (see Appendix 2 for an example).

In Bertsekas' (1976) notation, the equation in question is

$$A\tilde{s} = \tilde{t} + \tilde{r}(\tilde{s}) \quad , \quad (4.2-19)$$

where A is a non-singular matrix, $\tilde{r}(\tilde{s})$ is a differentiable function of $\tilde{s}$, and the equation is satisfied when $\tilde{s} = 0$, $\tilde{t} = 0$. The Lipschitz condition is required to ensure that

$$\|\nabla_{\tilde{s}} \tilde{r}(\tilde{s})\| \leq \alpha \|\tilde{s}\| \quad , \quad (4.2-20)$$

for $\|\tilde{s}\|$ sufficiently small, where α is a constant greater than zero. However, from his definition of $\tilde{r}(\tilde{s})$, it is clear that

$$\nabla_{\tilde{s}} \tilde{r}(0) = 0 \quad . \quad (4.2-21)$$

Now define

$$F(\tilde{s}, \tilde{t}) = A\tilde{s} - \tilde{r}(\tilde{s}) - \tilde{t} \quad . \quad (4.2-22)$$

Because $F(0,0) = 0$ and $\frac{\partial F(0,0)}{\partial \tilde{s}} = A$ is non-singular, the implicit function theorem implies that (locally) there exists $\tilde{s}(\tilde{t})$ as a continuously differentiable function of $\tilde{t}$ such that

$$\tilde{s}(0) = 0, \text{ and } F(\tilde{s}(\tilde{t}), \tilde{t}) = 0 \quad , \quad (4.2-23)$$

for $\tilde{t}$ in a neighbourhood of 0. Furthermore,

$$\nabla_{\tilde{t}} \tilde{s}(0) = A^{-1} \quad . \quad (4.2-24)$$

Using properties of continuously differentiable functions, it follows that there exist $\epsilon > 0$ and $\delta > 0$ such that, for $\kappa = \|A^{-1}\| + \epsilon$,

$$\|\tilde{s}(t)\| \leq \kappa \|\tilde{t}\|, \forall \tilde{t} \text{ such that } \|\tilde{t}\| \leq \delta.$$

Thus, the result of Lemma 2 of Bertsekas (1976) has been reproduced without the requirement that $\nabla^2 f, \{\nabla^2 h_i\}$ be Lipschitz continuous.

(b) Theorems 4.2.1 and 4.2.2 show only local properties of the augmented Lagrangian function. Note that, in any iteration, the unconstrained problem can be interpreted as minimizing a compound function resulting from the application of the quadratic penalty function method to the problem: minimize $f(\tilde{x}) - \tilde{v}^T \tilde{h}(\tilde{x})$ subject to $\tilde{h}(\tilde{x}) = 0$. Theorem 4.1.1 and Remark 4.1.1a then ensure that, for σ sufficiently large, the minimizing point $\tilde{x}(\tilde{v}, \sigma)$ will be sufficiently close to $\tilde{x}^*$ for the local properties to hold. Another interesting point to be noted here is that augmented Lagrangian methods can be interpreted as methods for solving an imbedded problem: that of finding, in a family of $\tilde{v}$ -problems, an optimal $\tilde{v}^*$ -problem such that the estimate $\tilde{x}(\tilde{v}, \sigma)$ given by the quadratic penalty function method coincides with $\tilde{x}^*$.

(c) The results in Theorem 4.2.3 are often interpreted as 'global' convergence results. This is because, for any *fixed* $\tilde{v}$, there exists a constant $\bar{\sigma} = \bar{\sigma}(\tilde{v})$ such that, $\forall \sigma \geq \bar{\sigma}$, the local minimizing point $\tilde{x}(\tilde{v}, \sigma)$ of $L_{\tilde{\sigma}}(\tilde{x}, \tilde{v})$ is in a neighbourhood of $\tilde{x}^*$, and the local properties hold. However, a constant $\bar{\sigma}$ cannot be found such that a required local minimizing point of $L_{\tilde{\sigma}}(\tilde{x}, \tilde{v})$ is guaranteed for *any* $\tilde{v}$. The results of Theorem 4.2.3 are therefore not global in this sense.

4.2.3 Inequality Constraints and Degeneracy

When inequality constraints are present, a natural way to handle them is by (implicitly) solving the extended problem (see subsection 2.2.5). Again, without loss of generality, it is assumed that only inequality

constraints are present. For the extended problem, the augmented Lagrangian function takes the form:

$$L_{\sigma}(\tilde{x}, \tilde{z}, u) = f(\tilde{x}) - \sum_{i \in J} u_i (g_i(\tilde{x}) - z_i^2) + \frac{\sigma}{2} \sum_{i \in J} (g_i(\tilde{x}) - z_i^2)^2 \quad (4.2-25)$$

The minimization with respect to the slack variable $\tilde{z}$ gives

$$z_i(\tilde{x}, u, \sigma)^2 = \max \left(0, g_i(\tilde{x}) - \frac{u_i}{\sigma} \right), \quad i \in J. \quad (4.2-26)$$

Therefore, as suggested by Rockafellar (1974), the augmented Lagrangian function for an inequality constrained problem has the form:

$$\bar{L}_{\sigma}(\tilde{x}, u) = f(\tilde{x}) + \sum_{i \in J} \begin{cases} -u_i g_i(\tilde{x}) + \frac{\sigma}{2} g_i(\tilde{x})^2 & \text{if } g_i(\tilde{x}) < u_i/\sigma \\ -\frac{1}{2} u_i^2/\sigma & \text{if } g_i(\tilde{x}) \geq u_i/\sigma \end{cases} \quad (4.2-27)$$

The hessian matrix $\nabla_{\tilde{x}}^2 \bar{L}_{\sigma}$ has jump discontinuities at points $\tilde{x}$ for which $g_i(\tilde{x}) = u_i/\sigma$ for some i . The possible effect of these discontinuities has been the cause of concern by several authors including Buys (1972), Kort and Bertsekas (1976) and Mangasarian (1975). Modifications to smooth out the discontinuities in $\nabla_{\tilde{x}}^2 \bar{L}_{\sigma}$ have been studied by Nakayama, *et al.* (1975), Kort and Bertsekas (1976) and Mangasarian (1975). These modifications, however, require that strict complementarity be satisfied at the solution $\tilde{x}^*$.

Corresponding to the shifted penalty function concept, an approach for handling inequality constraints is by using a shifted barrier function technique (Osborne (1972)), e.g.,

$$B(\tilde{x}, \theta, r) = f(\tilde{x}) - r \sum_{i \in J} \log[g_i(\tilde{x}) + \theta_i]. \quad (4.2-28)$$

Some nice properties of $B(\underline{x}, \underline{\theta}, r)$ are that: (i) strict complementarity implies that the solution $\underline{x}^*$ is bounded away from the barriers of $B(\underline{x}, \underline{\theta}, r)$; (ii) inside the barriers, $\nabla_{\underline{x}}^2 B$ has no jump discontinuities; and (iii) because of these barriers, the 'quadratic growth' condition of Rockafellar (1974) no longer needs to be satisfied outside the feasible region. But again, these shifted barrier function methods are only applicable to problems with strict complementarity.

When σ is bounded, the size of the discontinuities in $\nabla_{\underline{x}}^2 \bar{L}_{\sigma}$ are of order $\sigma \nabla g \nabla g^T$, and hence are bounded above. Furthermore, if strict complementarity is satisfied, these jump discontinuities are bounded away from a small neighbourhood of $\underline{x}^*$. Fletcher (1975) has investigated the effect of these discontinuities, and, on the basis of numerical experience, concluded that these jump discontinuities present no 'extra' difficulties for the unconstrained minimization subproblem. Thus, there is no 'real' evidence supporting any modification of $\bar{L}_{\sigma}(\underline{x}, \underline{u})$.

Non-degenerate Cases

For a non-degenerate problem (ICP) with $\underline{x}^*$ as a regular local solution, Theorem 2.2.6 implies that the extended problem has $(\underline{x}^*, \underline{z}^*)$ as a regular local solution. Thus, the results of subsection 4.2.2 apply.

Corresponding to (4.2-10)-(4.2-12), for an inequality constrained problem, the augmented dual function $\bar{D}_{\sigma}(\underline{u})$ takes the form:

$$\bar{D}_{\sigma}(\underline{u}) = \min_{\underline{x}} \bar{L}_{\sigma}(\underline{x}, \underline{u}) \quad , \quad (4.2-29)$$

with (in superscript form)

$$[\nabla_{\underline{u}} \bar{D}_{\sigma}(\underline{u}^{(k)})]_{\underline{i}} = -\min(g_{\underline{i}}(\underline{x}^{(k)}), u_{\underline{i}}^{(k)}/\sigma), \underline{i} \in J \quad , \quad (4.2-30)$$

$$\nabla_u^2 \bar{D}_\sigma(u^{(k)}) = \begin{cases} i \in M^{(k)} & \begin{cases} -N^T \left(\nabla_x^2 \bar{L}_\sigma \right)^{-1} N & 0 \\ 0 & -\frac{1}{\sigma} \end{cases} \\ i \notin M^{(k)} & \end{cases}, \quad (4.2-31)$$

where

$$M^{(k)} = \{i \mid g_i(x^{(k)}) < u_i^{(k)}/\sigma\}, \quad N = [\nabla g_i(x^{(k)}), \nabla g_j(x^{(k)}), \dots \mid i, j \in M^{(k)}].$$

Let $\tilde{x}(u, \sigma)$ be the local minimizer of $\bar{L}_\sigma(x, u)$, then, from (4.2-17) and (4.2-26), the Powell/Hestenes update for $\tilde{u}$ has the form:

$$\Lambda_i^{(k)} = \max(0, u_i^{(k)} - \sigma g_i^{(k)}), \quad i \in J. \quad (4.2-32)$$

Corresponding to the Newton update in (4.2-13), to make use of the fact that $\tilde{u}^* \geq 0$, Fletcher (1975) suggests choosing $\tilde{u}^{(k+1)}$ as the maximizer of the subproblem:

$$\text{maximize } U^{(k)}(\tilde{u}), \quad \text{subject to } \tilde{u} \geq 0,$$

where

$$U^{(k)}(\tilde{u}) = \frac{1}{2}(\tilde{u} - \tilde{u}^{(k)})^T \nabla_{\tilde{u}}^2 \bar{D}_\sigma(\tilde{u}^{(k)}) (\tilde{u} - \tilde{u}^{(k)}) + (\tilde{u} - \tilde{u}^{(k)})^T \nabla_{\tilde{u}} \bar{D}_\sigma(\tilde{u}^{(k)}). \quad (4.2-33)$$

For equality constrained problems, convergence is usually measured in terms of $\|h(\tilde{x}(\tilde{u}, \sigma))\|_\infty$. The corresponding quantity for inequality constrained problem has the form:

$$K = \max_{i \in J} \left| \min(g_i(\tilde{x}(\tilde{u}, \sigma)), u_i/\sigma) \right|. \quad (4.2-34)$$

Remark 4.2.3

Note that $\bar{D}_\sigma(\tilde{u})$ is also the augmented dual function of the extended problem. With strict complementarity, the extended problem has $(\tilde{x}^*, \tilde{z}^*)$ as a regular local solution, and the results of subsection 4.2.2

apply. (4.2-31) therefore indicates that, for an equality constrained problem, even if it has a regular local solution and all the problem functions are everywhere twice continuously differentiable, the augmented dual function D_σ need not have continuous second derivative. For the case of the extended problem, $\nabla_u^2 \bar{D}_\sigma(u)$ has jump discontinuities at points $\underline{u}$ for which $g_i(\underline{x}(\underline{u}, \sigma)) = u_i/\sigma$ for some i . When strict complementarity does not hold in the original ICP, $\nabla_u^2 \bar{D}_\sigma(u)$ has jump discontinuities at $\underline{u}^*$, the solution of the augmented dual problem.

Degenerate Cases

When the problem is degenerate, $\nabla_x^2 \bar{L}_\sigma(\underline{x}, \underline{u})$ has jump discontinuities at the solution $(\underline{x}^*, \underline{u}^*)$. In the extended problem's framework, this corresponds to the fact that $\nabla_{(x,z)}^2 L_\sigma((\underline{x}^*, \underline{z}^*), \underline{u}^*)$ cannot be made positive definite by increasing σ . Because $(\underline{x}^*, \underline{z}^*)$ is not a regular local solution, the results of subsection 4.2.2 are no longer applicable. However, similar to the quadratic loss penalty function method, the effect of degeneracy is not as bad as it may seem. Corresponding results without strict complementarity will now be developed.

Definition

The 'active' and 'degenerate' index sets M and Z are defined as:

$$M(\underline{x}, \underline{u}, \sigma) = \{i \mid i \in J, g_i(\underline{x}) < u_i/\sigma\}, \quad (4.2-35)$$

and

$$Z(\underline{x}, \underline{u}, \sigma) = \{i \mid i \in J, g_i(\underline{x}) = u_i/\sigma\}. \quad (4.2-36)$$

Note that both the sets $M(\underline{x}, \underline{u}, \sigma)$ and $Z(\underline{x}, \underline{u}, \sigma)$ have been introduced before in subsection 4.1.3, in the situation where the variable $\underline{u}$ is fixed as $\underline{u} = 0$. From (4.2-35) and (4.2-36), both $M(\underline{x}, 0, \sigma)$ and $Z(\underline{x}, 0, \sigma)$

are independent of σ , and hence they were introduced simply as $M(\underline{x})$ and $Z(\underline{x})$ in (4.1-14) and (4.1-15), respectively.

Lemma 4.2.1

Let $\underline{x}^*$ (together with $\underline{u}^*$) be a regular local solution of an ICP, then, for fixed $\sigma > 0$, there exists $\delta > 0$ such that, $\forall \underline{x} \in N(\underline{x}^*, \delta)$ and $\forall \underline{u} \in N(\underline{u}^*, \delta)$,

$$J_1(\underline{x}^*) \subseteq M(\underline{x}, \underline{u}, \sigma) \subseteq M(\underline{x}, \underline{u}, \sigma) \cup Z(\underline{x}, \underline{u}, \sigma) \subseteq J_1(\underline{x}^*) \cup J_2(\underline{x}^*) \quad (4.2-37)$$

Proof

For inactive constraints, $g_{\underline{i}}(\underline{x}^*) > \underline{u}^*/\sigma \equiv 0$. Therefore, for $\underline{x}$ sufficiently close to $\underline{x}^*$ and $\underline{u}$ sufficiently close to $\underline{u}^*$, $g_{\underline{i}}(\underline{x}) > \underline{u}/\sigma$, and

$$M(\underline{x}, \underline{u}, \sigma) \cup Z(\underline{x}, \underline{u}, \sigma) \subseteq J_1(\underline{x}^*) \cup J_2(\underline{x}^*) \quad (4.2-38)$$

Similarly, by considering non-degenerate active constraints, it follows that, for $\underline{x}$ sufficiently close to $\underline{x}^*$ and $\underline{u}$ sufficiently close to $\underline{u}^*$,

$$J_1(\underline{x}^*) \subseteq M(\underline{x}, \underline{u}, \sigma) \quad (4.2-39)$$

Hence, for sufficiently small $\delta > 0$, both (4.2-38) and (4.2-39) hold

$$\forall \underline{x} \in N(\underline{x}^*, \delta) \quad \text{and} \quad \forall \underline{u} \in N(\underline{u}^*, \delta): \quad \square$$

Lemma 4.2.2

Assume that $\underline{x}', \underline{u}'$ and a constant $\sigma > 0$ are given, then there exists $\delta = \delta(\underline{x}', \underline{u}', \sigma) > 0$ such that, $\forall \underline{x} \in N(\underline{x}', \delta)$ and $\forall \underline{u} \in N(\underline{u}', \delta)$,

$$M(\underline{x}', \underline{u}', \sigma) \subseteq M(\underline{x}, \underline{u}, \sigma) \subseteq M(\underline{x}, \underline{u}, \sigma) \cup Z(\underline{x}, \underline{u}, \sigma) \subseteq M(\underline{x}', \underline{u}', \sigma) \cup Z(\underline{x}', \underline{u}', \sigma).$$

Proof

Because the constraint functions are continuous, there exists $\delta = \delta(x', u', \sigma) > 0$ such that, $\forall \tilde{x} \in N(\tilde{x}', \delta)$ and $\forall \tilde{u} \in N(\tilde{u}', \delta)$,

$$g_{\tilde{\lambda}}(\tilde{x}) < u_{\tilde{\lambda}}/\sigma \quad \text{if} \quad g_{\tilde{\lambda}}(\tilde{x}') < u_{\tilde{\lambda}}'/\sigma ,$$

and

$$g_{\tilde{\lambda}}(\tilde{x}) > u_{\tilde{\lambda}}/\sigma \quad \text{if} \quad g_{\tilde{\lambda}}(\tilde{x}') > u_{\tilde{\lambda}}'/\sigma .$$

□

Now consider any index set R such that

$$J_1(\tilde{x}^*) \subseteq R \subseteq J_1(\tilde{x}^*) \cup J_2(\tilde{x}^*) . \quad (4.2-40)$$

Theorem 2.2.1 gives that if $\tilde{x}^*$ is a regular local solution of the original ICP, then $\tilde{x}^*$ is a regular local solution of the (equality constrained) problem $P[R]$. If an augmented Lagrangian method is applied to $P[R]$, the corresponding augmented Lagrangian function, denoted by $L_{\sigma}^R(\tilde{x}, \tilde{u})$, has the form:

$$L_{\sigma}^R(\tilde{x}, \tilde{u}) = f(\tilde{x}) - \sum_{\tilde{\lambda} \in R} u_{\tilde{\lambda}} g_{\tilde{\lambda}}(\tilde{x}) + \frac{\sigma}{2} \sum_{\tilde{\lambda} \in R} g_{\tilde{\lambda}}(\tilde{x})^2 . \quad (4.2-41)$$

From Theorem 4.2.1, there exists $\bar{\sigma}(R) > 0$ such that, $\forall \sigma \geq \bar{\sigma}(R)$, $\nabla_{\tilde{x}}^2 L_{\sigma}^R(\tilde{x}^*, \tilde{u}^*)$ is positive definite with the smallest eigenvalue bounded away from zero, with

$$\nabla_{\tilde{x}} L_{\sigma}^R(\tilde{x}^*, \tilde{u}^*) = 0 . \quad (4.2-42)$$

Let $\bar{\sigma} = \max \bar{\sigma}(R)$, then, from (4.2-27), (4.2-41), Lemma 4.2.1 and that $\nabla_{\tilde{x}}^2 L_{\sigma}^{J_1(\tilde{x}^*)}(\tilde{x}^*, \tilde{u}^*)$ is positive definite for $\sigma \geq \bar{\sigma}$, it follows that, for any fixed $\sigma \geq \bar{\sigma}$, there exists $\delta = \delta(\sigma) > 0$ and $a = a(\sigma) > 0$ such that, $\forall \tilde{x} \in N(\tilde{x}^*, \delta)$,

$$\begin{aligned}
\bar{L}_{\sigma}(\underline{x}, \underline{u}^*) - \bar{L}_{\sigma}(\underline{x}^*, \underline{u}^*) &\geq L_{\sigma}^{J_1(\underline{x}^*)}(\underline{x}, \underline{u}^*) - L_{\sigma}^{J_1(\underline{x}^*)}(\underline{x}^*, \underline{u}^*) \\
&\geq a \|\underline{x} - \underline{x}^*\|^2.
\end{aligned}
\tag{4.2-43}$$

Thus, $\underline{x}^*$ is a unique minimum of 2nd order of $\bar{L}_{\sigma}(\underline{x}, \underline{u}^*)$.

Because $\nabla_{\underline{x}}^2 L_{\sigma}^R(\underline{x}, \underline{u})$ is continuous in both $\underline{x}$ and $\underline{u}$ and positive definite at $(\underline{x}^*, \underline{u}^*)$, it follows that $\nabla_{\underline{x}}^2 L_{\sigma}^R(\underline{x}, \underline{u})$ is positive definite in a small enough neighbourhood of $(\underline{x}^*, \underline{u}^*)$. Therefore, there exists $\delta > 0$ such that, for any index set R satisfying (4.2-40), for any $\underline{x} \in N(\underline{x}^*, \delta)$ and for any $\underline{u} \in N(\underline{u}^*, \delta)$, $\nabla_{\underline{x}}^2 L_{\sigma}^R(\underline{x}, \underline{u})$ is positive definite. Similar to subsection 4.1.3, identification techniques will be used to show that $\bar{L}_{\sigma}(\underline{x}, \underline{u})$ satisfies certain convexity conditions. Properties of convex functionals (Appendix 1) and a stronger version of the implicit function theorem (Appendix 2) then imply that the minimizing point $\underline{x}(\underline{u}, \sigma)$ of $\bar{L}_{\sigma}(\underline{x}, \underline{u})$ is locally unique and (locally) a continuous function of $\underline{u}$. The minimizing point $\underline{x}(\underline{u}, \sigma)$ can be identified as the (local) minimizing point of $L_{\sigma}^R(\underline{x}, \underline{u})$ for some index set R . By using this identification and properties of the problem $P[R]$, the rate of convergence of various augmented Lagrangian methods, when solving degenerate problems, can then be derived.

Lemma 4.2.3

Assume that $\sigma > 0$ and $\underline{u}'$ are given. If there exists $\delta > 0$ such that (4.2-37) is satisfied at $(\underline{x}, \underline{u}', \sigma)$ for all $\underline{x} \in N(\underline{x}^*, \delta)$ and that, for all index sets R satisfying (4.2-40), $\nabla_{\underline{x}}^2 L_{\sigma}^R(\underline{x}, \underline{u}')$ is positive definite for all $\underline{x} \in N(\underline{x}^*, \delta)$, then $\bar{L}_{\sigma}(\underline{x}, \underline{u}')$ is strictly convex on $N(\underline{x}^*, \delta)$.

Proof

Given a point $\underline{x}' \in N(\underline{x}^*, \delta)$ and a direction $\underline{t}$ ($\|\underline{t}\| = 1$), there

exists $a > 0$ such that, for all R satisfying (4.2-40),

$$\lim_{\alpha \rightarrow 0} \frac{t^T (\nabla_{\underline{x}} L_{\sigma}^R(\underline{x}' + \alpha t, \underline{u}') - \nabla_{\underline{x}} L_{\sigma}^R(\underline{x}', \underline{u}'))}{\alpha} = t^T \nabla_{\underline{x}}^2 L_{\sigma}^R(\underline{x}', \underline{u}') t \geq a. \quad (4.2-44)$$

By the definition of $\bar{L}_{\sigma}(\underline{x}, \underline{u}')$, for α sufficiently small, Lemma 4.2.2 implies that $\nabla_{\underline{x}} \bar{L}_{\sigma}(\underline{x}' + \alpha t, \underline{u}')$ can be identified with $\nabla_{\underline{x}} L_{\sigma}^R(\underline{x}' + \alpha t, \underline{u}')$ for some R such that

$$M(\underline{x}', \underline{u}', \sigma) \subseteq R \subseteq M(\underline{x}', \underline{u}', \sigma) \cup Z(\underline{x}', \underline{u}', \sigma). \quad (4.2-45)$$

Hence by (4.2-44), there exists an $a' > 0$ such that

$$\frac{t^T (\nabla_{\underline{x}} \bar{L}_{\sigma}(\underline{x}' + \alpha t, \underline{u}') - \nabla_{\underline{x}} \bar{L}_{\sigma}(\underline{x}', \underline{u}'))}{\alpha} \geq a', \quad (4.2-46)$$

for all α sufficiently small. It follows that the component of $\nabla_{\underline{x}} \bar{L}_{\sigma}(\underline{x}, \underline{u}')$ along any line in $N(\underline{x}^*, \delta)$ increases strictly monotonically, and it is then a straightforward application of the mean value theorem to show that $\bar{L}_{\sigma}(\underline{x}, \underline{u}')$ is strictly convex on $N(\underline{x}^*, \delta)$. $\square$

Theorem 4.2.4

Let $\underline{x}^*$ (together with $\underline{u}^*$) be a regular local solution of a (degenerate) ICP, then, for σ sufficiently large, there exists $\underline{x}(\underline{u}, \sigma)$ which is a *continuous* function of $\underline{u}$ such that, for every $\underline{u} \in N(\underline{u}^*, \delta)$ for some $\delta > 0$, $\underline{x}(\underline{u}, \sigma)$ is the unique local minimizer of $\bar{L}_{\sigma}(\underline{x}, \underline{u})$, with $\underline{x}(\underline{u}^*, \sigma) = \underline{x}^*$.

Proof

Because, for every R satisfying (4.2-40), $\nabla_{\underline{x}}^2 L_{\sigma}^R(\underline{x}, \underline{u})$ is positive definite in a small neighbourhood of $(\underline{x}^*, \underline{u}^*)$, there exists $\delta > 0$ such that

the conditions of Lemmas 4.2.1 and 4.2.3 are satisfied, and hence $\bar{L}_\sigma(\underline{x}, \underline{u})$ is strictly convex on $N(\underline{x}^*, \delta)$ for all $\underline{u} \in N(\underline{u}^*, \delta)$. By using properties of convex functionals (Appendix 1) and a stronger version of the implicit function theorem (Appendix 2), the required result follows in similar fashion to the proof of Theorem 4.1.4. $\square$

Note here that the result of Theorem 4.2.4 was first discovered by Fletcher (1975). The technique of equating $\nabla_{\underline{x}} \bar{L}_\sigma$ with $\nabla_{\underline{x}} L_\sigma^R$ for some R was also due to Fletcher (1975). Lemmas 4.2.1-4.2.3 and Theorem 4.2.4 are presented here, partly to provide results analogous to Lemmas 4.1.1-4.1.3 and Theorem 4.1.4, and partly because these results and their proofs are needed in establishing the sensitivity results of subsection 2.3.2.

Switching of the 'active' set $M(\underline{x}, \underline{u}, \sigma)$

Recall that

$$M(\underline{x}, \underline{u}, \sigma) = \{i \mid g_i(\underline{x}) < u_i/\sigma\}.$$

For $\underline{x}'$ close to $\underline{x}^*$ and $\underline{u}'$ close to $\underline{u}^*$, Lemma 4.2.1 and Theorem 2.2.1 imply that the (equality constrained) problem $P[M(\underline{x}', \underline{u}', \sigma)]$ has $\underline{x}^*$ as a regular local solution. Consider applying an augmented Lagrangian method to $P[M(\underline{x}', \underline{u}', \sigma)]$, then (using the value of σ in $M(\underline{x}', \underline{u}', \sigma)$ as the penalty parameter)

$$L_\sigma^{M(\underline{x}', \underline{u}', \sigma)}(\underline{x}, \underline{u}) = f(\underline{x}) - \sum_{i \in M(\underline{x}', \underline{u}', \sigma)} u_i g_i(\underline{x}) + \frac{\sigma}{2} \sum_{i \in M(\underline{x}', \underline{u}', \sigma)} g_i(\underline{x})^2. \quad (4.2.47)$$

Clearly, at $(\underline{x}', \underline{u}')$, $\nabla_{\underline{x}} \bar{L}_\sigma(\underline{x}, \underline{u})$ can be identified with $\nabla_{\underline{x}} L_\sigma^{M(\underline{x}', \underline{u}', \sigma)}(\underline{x}, \underline{u})$. Hence at $(\underline{x}', \underline{u}')$, as far as the augmented Lagrangian algorithm is concerned, the original ICP is the same as the problem $P[M(\underline{x}', \underline{u}', \sigma)]$.

Remark 4.2.4

When the m -vector $\underline{u}$ is mentioned in any problem $P[M]$, it is clear that only those components $u_i, i \in M$, are needed. When a multiplier from $P[M]$ is referred to in the original ICP, the convention is to patch in the extra components by setting $u_i = 0$ for $i \notin M$.

As the algorithm proceeds, the 'active' set $M(\underline{x}, \underline{u}, \sigma)$ identified by the algorithm will be switching between those index sets $R, J_1(\underline{x}^*) \subseteq R \subseteq J_1(\underline{x}^*) \cup J_2(\underline{x}^*)$. The algorithm can therefore be identified as switching between different problems $P[R]$, each of which has the same solution $(\underline{x}^*, \underline{u}^*)$. In fact, this 'switching' by the algorithm can be separated into two stages:

- (i) For $\underline{u}'$ close to $\underline{u}^*$, Lemmas 4.2.1 and 4.2.2 imply that, for $\underline{x}$ close to $\underline{x}(\underline{u}', \sigma)$,

$$J_1(\underline{x}^*) \subseteq M(\underline{x}(\underline{u}', \sigma), \underline{u}', \sigma) \subseteq M(\underline{x}, \underline{u}', \sigma) \subseteq$$

$$M(\underline{x}(\underline{u}', \sigma), \underline{u}', \sigma) \cup Z(\underline{x}(\underline{u}', \sigma), \underline{u}', \sigma) \subseteq J_1(\underline{x}^*) \cup J_2(\underline{x}^*). \quad (4.2-48)$$

Therefore, sufficiently close to $\underline{x}(\underline{u}', \sigma)$, the minimization of $\bar{L}_\sigma(\underline{x}, \underline{u}')$ can be identified as switching between the minimization of different $L_\sigma^{M(\underline{x}, \underline{u}', \sigma)}(\underline{x}, \underline{u}')$, each of which has $\underline{x}(\underline{u}', \sigma)$ as the unique local minimizing point.

- (ii) The maximization of $\bar{D}_\sigma(\underline{u})$ can be identified as switching between the maximization of different $D_\sigma^{M(\underline{x}(\underline{u}, \sigma), \underline{u}, \sigma)}(\underline{u}) = \min_x L_\sigma^{M(\underline{x}(\underline{u}, \sigma), \underline{u}, \sigma)}(\underline{x}, \underline{u})$, each of which has $\underline{u}^*$ as the unique maximizing point.

By this process of identification, as the algorithm proceeds, a natural choice for choosing the next $\underline{x}$ or $\underline{u}$ is the one chosen by the

problem $P[M(\underline{x}, \underline{u}, \sigma)]$ or $P[M(\underline{x}(\underline{u}, \sigma), \underline{u}, \sigma)]$.

Updating of the multiplier and the rate of convergence

At the minimum $\underline{x}^{(k)}$ of $\bar{L}_\sigma(\underline{x}, \underline{u}^{(k)})$, the set $M^{(k)} = M(\underline{x}^{(k)}, \underline{u}^{(k)}, \sigma)$ is uniquely defined. If a Newton method is to be used, the Newton update (4.2-13) for the problem $P[M^{(k)}]$ is essentially the same as (4.2-33) or (4.2-30) and (4.2-31). Thus, (4.2-33) can be used as a general Newton update for the multiplier of the ICP. By a similar argument to that used in subsection 4.1.3, it can be shown that the convergence of $\underline{u}^{(k)}$ to $\underline{u}^*$ (or $\underline{x}^{(k)}$ to $\underline{x}^*$) is quadratic (super-linear) or linear depending on whether the Newton update or the Powell/Hestenes update is used. Note that, if the Powell/Hestenes update is used,

$$u_i^{(k+1)} = \lambda_i^{(k)} = \max(0, u_i^{(k)} - \sigma g_i^{(k)}) , i \in J . \quad (4.2-49)$$

And thus the index set $M(\underline{x}^{(k)}, \underline{u}^{(k)}, \sigma)$ need not be explicitly identified.

Unconstrained minimization subproblem

Most good unconstrained methods are of quasi-Newton type. An essential feature of quasi-Newton methods is that an estimate of the hessian is being kept and updated. Therefore, as a by-product, this approximated hessian can be used instead of the exact hessian for the updating of the multiplier in (4.2-33).

For a degenerate problem, since the 'active' set $M(\underline{x}, \underline{u}, \sigma)$ will be switching between several different index sets, it is appropriate that the estimate of $\nabla_{\underline{x}}^2 \bar{L}_\sigma$ should also be switching between the corresponding $\nabla_{\underline{x}}^2 L_\sigma^M$. In what follows, it will be shown how this switching can be achieved:

Locally, as the minimization of $\bar{L}_\sigma(\underline{x}, \underline{u})$ proceeds, suppose that after some computation the unconstrained algorithm is at a point $\underline{x}$, having

an estimate of $\nabla_x^2 L_{\sigma}^M$, $M = M(\underline{x}, \underline{u}, \sigma)$, as the current estimate of $\nabla_x^2 \bar{L}_{\sigma}$. Now assume that the next step takes the algorithm to a point $\underline{x}'$ with $M' = M(\underline{x}', \underline{u}, \sigma)$, $M' \neq M$. A natural way to proceed is to perform an appropriate update for the estimate of $\nabla_x^2 L_{\sigma}^M$ and then switch over to an estimate of $\nabla_x^2 L_{\sigma}^{M'}$. Since

$$\begin{aligned} \nabla_x^2 L_{\sigma}^M(\underline{x}', \underline{u}) &= \nabla_x^2 \bar{L}_{\sigma}(\underline{x}', \underline{u}) + \sum_{i \in M', i \notin M} \left(\frac{u_i}{\sigma} - g_i(\underline{x}') \right) \nabla g_i(\underline{x}') \\ &\quad - \sum_{i \notin M', i \in M} \left(\frac{u_i}{\sigma} - g_i(\underline{x}') \right) \nabla g_i(\underline{x}') \quad , \end{aligned} \quad (4.2-50)$$

therefore $\underline{x}' - \underline{x}$ and $\nabla_x^2 L_{\sigma}^M(\underline{x}', \underline{u}) - \nabla_x^2 L_{\sigma}^M(\underline{x}, \underline{u})$ are available for an appropriate updating of the estimate of $\nabla_x^2 L_{\sigma}^M$. The switch to an estimate of $\nabla_x^2 L_{\sigma}^{M'}$ can then be achieved by:

$$\nabla_x^2 L_{\sigma}^{M'} = \nabla_x^2 L_{\sigma}^M + \sigma \left(\sum_{i \in M', i \notin M} \nabla g_i(\underline{x}') \cdot \nabla g_i(\underline{x}')^T - \sum_{i \notin M', i \in M} \nabla g_i(\underline{x}') \cdot \nabla g_i(\underline{x}')^T \right) . \quad (4.2-51)$$

In this way, when the unconstrained algorithm stops at $\underline{x}^0$, the current estimate of $\nabla_x^2 \bar{L}_{\sigma}$ given by the algorithm will be an estimate of $\nabla_x^2 L_{\sigma}^M(\underline{x}^0, \underline{u}, \sigma)$, for the current 'active' set $M(\underline{x}^0, \underline{u}, \sigma)$. This estimate of $\nabla_x^2 L_{\sigma}^M$ can then be used in the Newton formula (4.2-33) for updating the multiplier $\underline{u}$.

However, as pointed out before, if the Powell/Hestenes update is used, the index set M need not be explicitly identified. Thus, there is a question of whether or not the overhead in keeping track of the index set M and switching between the estimates of different $\nabla_x^2 L_{\sigma}^M$ is computationally justified. Based on numerical experience, Fletcher (1975) has concluded that the discontinuities in $\nabla_x^2 \bar{L}_{\sigma}$ present no extra difficulties for the unconstrained minimization routine. Intuitively, since $\underline{x}(\underline{u}, \sigma)$ is a

unique minimum of 2nd order of $\bar{L}_\sigma(\underline{x}, \underline{u})$, it seems reasonable to expect any 'good' routine to perform well on minimizing $\bar{L}_\sigma(\underline{x}, \underline{u})$. An argument supporting Fletcher's conclusion will now be presented:

Consider minimizing $\bar{L}_\sigma(\underline{x}, \underline{u})$ with $\underline{x}(\underline{u}, \sigma)$ as its minimizer. For illustration, only the case where $\underline{u} = \underline{u}^*$ and where there is only one degenerate constraint g_i will be considered. From the definition of $\bar{L}_\sigma(\underline{x}, \underline{u})$, $\nabla_{\underline{x}}^2 \bar{L}_\sigma(\underline{x}, \underline{u}^*)$ has jumps $\pm \sigma \nabla g_i \nabla g_i^T$ across the surface $g_i(\underline{x}) = 0$ (i.e., $g_i(\underline{x}) - g_i(\underline{x}^*) = \underline{u}^*/\sigma$). For a point $\underline{x}$ close to $\underline{x}^*$ with $M = M(\underline{x}, \underline{u}^*, \sigma)$, it follows that

$$\nabla_{\underline{x}} \bar{L}_\sigma(\underline{x}, \underline{u}^*) = \nabla_{\underline{x}} L_\sigma^M(\underline{x}, \underline{u}^*) \quad (4.2-52)$$

Now suppose that the current estimate of $\nabla_{\underline{x}}^2 \bar{L}_\sigma$ has the form:

$$\nabla_{\underline{x}}^2 L_\sigma^M + \alpha \nabla g_i \nabla g_i^T \quad (4.2-53)$$

Then, by using a formula for matrix inversion (Householder (1974), p.123), the new point $\bar{\underline{x}}$ generated from taking a (quasi) Newton step satisfies the equation:

$$\begin{aligned} \bar{\underline{x}} &= \underline{x} - [\nabla_{\underline{x}}^2 L_\sigma^M + \alpha \nabla g_i \nabla g_i^T]^{-1} \nabla_{\underline{x}} L_\sigma^M \\ &= \underline{x} - [\nabla_{\underline{x}}^2 L_\sigma^M]^{-1} \nabla_{\underline{x}} L_\sigma^M + \tau [\nabla_{\underline{x}}^2 L_\sigma^M]^{-1} \nabla g_i \nabla g_i^T [\nabla_{\underline{x}}^2 L_\sigma^M]^{-1} \nabla_{\underline{x}} L_\sigma^M, \end{aligned} \quad (4.2-54)$$

where $\tau^{-1} = \nabla g_i^T [\nabla_{\underline{x}}^2 L_\sigma^M]^{-1} \nabla g_i + \frac{1}{\alpha}$.

Now note that $[\nabla_{\underline{x}}^2 L_\sigma^M]^{-1} \nabla_{\underline{x}} L_\sigma^M$ is just the (quasi) Newton correction for minimizing $L_\sigma^M(\underline{x}, \underline{u}^*)$. Let

$$\bar{\underline{x}}^M = \underline{x} - [\nabla_{\underline{x}}^2 L_\sigma^M]^{-1} \nabla_{\underline{x}} L_\sigma^M \quad (4.2-55)$$

Since $\tilde{x}^*$ is the minimizing point of $L_\sigma^M(\tilde{x}, u^*)$, and $\tilde{x}^M$ is produced from a (quasi) Newton step, it follows that

$$\|\tilde{x}^M - \tilde{x}^*\| = o(\|\tilde{x} - \tilde{x}^*\|) . \quad (4.2-56)$$

This implies that

$$\begin{aligned} \nabla g_{\tilde{\lambda}}(\tilde{x})^T \cdot [\nabla_x^2 L_\sigma^M(\tilde{x}, u^*)]^{-1} \nabla_x L_\sigma^M(\tilde{x}, u^*) &= \nabla g_{\tilde{\lambda}}(\tilde{x})^T \cdot (\tilde{x} - \tilde{x}^M) \\ &= g_{\tilde{\lambda}}(\tilde{x}) + o(\|\tilde{x} - \tilde{x}^*\|) . \end{aligned} \quad (4.2-57)$$

Therefore, after ignoring higher order terms, (4.2-54) gives

$$\tilde{x} = \tilde{x}^M + \tau g_{\tilde{\lambda}}(\tilde{x}) [\nabla_x^2 L_\sigma^M(\tilde{x}, u^*)]^{-1} \nabla g_{\tilde{\lambda}}(\tilde{x}) . \quad (4.2-58)$$

Eqn. (4.2-58) indicates that an error of the form $\alpha \nabla g_{\tilde{\lambda}} \nabla g_{\tilde{\lambda}}^T$ in the estimate of $\nabla_x^2 L_\sigma^M$ produces a term of order $O(g_{\tilde{\lambda}}(\tilde{x}))$ to the Newton correction.

The argument supporting computational experience now goes as follows:

As the minimization of $\bar{L}_\sigma(\tilde{x}, u^*)$ proceeds, if the approach to $\tilde{x}^*$ is on either side of the surface $g_{\tilde{\lambda}}(\tilde{x}) = 0$, then, clearly, the jump discontinuities in $\nabla_x^2 \bar{L}_\sigma$ have no effect on the rate of convergence. On the other hand, if the approach to $\tilde{x}^*$ is along or zig-zagging close to the surface $g_{\tilde{\lambda}}(\tilde{x}) = 0$ (this gives $g_{\tilde{\lambda}}(\tilde{x}) \sim o(\|\tilde{x} - \tilde{x}^*\|)$), any error in the estimated hessian of the form $\alpha \nabla g_{\tilde{\lambda}} \nabla g_{\tilde{\lambda}}^T$ will add a term of order $O(g_{\tilde{\lambda}}(\tilde{x})) \sim o(\|\tilde{x} - \tilde{x}^*\|)$ to the Newton correction. Overall, $\|\tilde{x} - \tilde{x}^*\| \sim o(\|\tilde{x} - \tilde{x}^*\|)$ and the super-linear rate of convergence is retained.

Remark 4.2.5

(a) The above argument is purely a heuristic one and should not be mistaken for a rigorous proof.

(b) The analysis in this subsection has in effect suggested that the basic Powell/Hestenes multiplier method does not require any modification for handling inequality constraints (with or without strict complementarity). The Powell/Hestenes update (4.2-49) can be computed without explicitly identifying the 'active' set M . The jump discontinuities have little or no effect on the unconstrained subproblem. And above all, the usual rate of convergence of $\tilde{u}^{(k)}$ to $\tilde{u}^*$ (or $\tilde{x}^{(k)}$ to $\tilde{x}^*$) is retained.

4.2.4 Newton's Method and its Decomposition for the Augmented Dual Problem

In subsection 4.2.2, the quoted results rely heavily on $\tilde{x}^*$ being a regular local solution of the ECP. It is therefore instructive to rephrase this condition in another form.

Lemma 4.2.4

If $\tilde{x}^*$ is a regular local solution of an ECP, then there exists $\delta > 0$ such that, $\forall \tilde{x} \in N(\tilde{x}^*, \delta)$ and $\forall \tilde{v} \in N(\tilde{v}^*, \delta)$, the matrix $J(\tilde{x}, \tilde{v})$ is non-singular,

$$J(\tilde{x}, \tilde{v}) = \begin{bmatrix} \nabla_{\tilde{x}}^2 L_0(\tilde{x}, \tilde{v}) & -N(\tilde{x}) \\ N(\tilde{x})^T & 0 \end{bmatrix}, \quad (4.2-59)$$

where $N(\tilde{x}) = [\nabla h_1(\tilde{x}), \nabla h_2(\tilde{x}), \dots, \nabla h_\ell(\tilde{x})]$.

Proof

As a direct consequence of Lemma 2.2.1, $J(\tilde{x}^*, \tilde{v}^*)$ is non-singular, and hence $J(\tilde{x}^*, \tilde{v}^*)J(\tilde{x}^*, \tilde{v}^*)^T$ is positive definite. Because $J(\tilde{x}, \tilde{v})J(\tilde{x}, \tilde{v})^T$ is continuous in both $\tilde{x}$ and $\tilde{v}$ and positive definite at $(\tilde{x}^*, \tilde{v}^*)$, it follows that $J(\tilde{x}, \tilde{v})J(\tilde{x}, \tilde{v})^T$ is positive definite in a sufficiently small neighbourhood of $(\tilde{x}^*, \tilde{v}^*)$. The required results then follow. $\square$

An immediate consequence of Lemma 4.2.4 is that, $\forall \underline{x} \in N(\underline{x}^*, \delta)$ and $\forall \underline{v} \in N(\underline{v}^*, \delta)$, $J(\underline{x}, \underline{v})$ has an inverse. From the structure of $J(\underline{x}, \underline{v})$, it follows that the inverse is of the form:

$$J(\underline{x}, \underline{v})^{-1} = \begin{bmatrix} \Lambda(\underline{x}, \underline{v}) & B(\underline{x}, \underline{v}) \\ -B(\underline{x}, \underline{v})^T & Q(\underline{x}, \underline{v}) \end{bmatrix}. \quad (4.2-60)$$

Note that the diagonal submatrices $\Lambda(\underline{x}, \underline{v})$ and $Q(\underline{x}, \underline{v})$ are symmetric.

The matrices $\Lambda(\underline{x}, \underline{v})$, $B(\underline{x}, \underline{v})$ and $Q(\underline{x}, \underline{v})$ are defined by the equations:

$$\nabla_{\underline{x}}^2 L_0(\underline{x}, \underline{v}) \cdot \Lambda(\underline{x}, \underline{v}) + N(\underline{x}) \cdot B(\underline{x}, \underline{v})^T = I, \quad (4.2-61)$$

$$N(\underline{x})^T \cdot \Lambda(\underline{x}, \underline{v}) = 0, \quad (4.2-62)$$

$$N(\underline{x})^T \cdot B(\underline{x}, \underline{v}) = I, \quad (4.2-63)$$

and

$$Q(\underline{x}, \underline{v}) = B(\underline{x}, \underline{v})^T \cdot \nabla_{\underline{x}}^2 L_0(\underline{x}, \underline{v}) \cdot B(\underline{x}, \underline{v}). \quad (4.2-64)$$

We are now in a position to prove the main Lemma of this subsection.

Lemma 4.2.5

Assume that $\underline{x}^*$ is a regular local solution of an ECP, and that the conditions of Theorem 4.2.2 are satisfied, then there exists $\delta > 0$ such that, $\forall \underline{v} \in N(\underline{v}^*, \delta)$,

$$[\nabla_{\underline{v}}^2 D(\underline{v})]^{-1} = -\sigma I - Q(\underline{x}(\underline{v}, \sigma), \Lambda(\underline{v}, \sigma)).$$

Proof

From the definitions of $L_{\sigma}(\underline{x}, \underline{v})$ and $L_0(\underline{x}, \underline{v})$, it follows that

$$\nabla_{\underline{x}}^2 L_{\sigma}(\underline{x}(\underline{v}, \sigma), \underline{v}) = \nabla_{\underline{x}}^2 L_0(\underline{x}(\underline{v}, \sigma), \Lambda(\underline{v}, \sigma)) + \sigma N(\underline{x}(\underline{v}, \sigma)) N(\underline{x}(\underline{v}, \sigma))^T. \quad (4.2-65)$$

With $\underline{v}$ fixed, $\underline{x}(\underline{v}, \sigma)$ and $\underline{\Lambda}(\underline{v}, \sigma)$ are uniquely determined. Since no ambiguity can arise, it is convenient to denote any matrix written as $E(\underline{x}(\underline{v}, \sigma))$, $E(\underline{x}(\underline{v}, \sigma), \underline{v})$, $E(\underline{v})$ or $E(\underline{x}(\underline{v}, \sigma), \underline{\Lambda}(\underline{v}, \sigma))$ simply as E . Using (4.2-12) and (4.2-61)-(4.2-65), it follows that

$$\begin{aligned}
 [\nabla_{\underline{v}}^2 D_{\sigma}] [-\sigma I - Q] &= N^T [\nabla_{\underline{x}}^2 L_{\sigma}]^{-1} N [\sigma I + Q] \\
 &= N^T [\nabla_{\underline{x}}^2 L_{\sigma}]^{-1} N [B^T [\sigma N N^T + \nabla_{\underline{x}}^2 L_0] B] \\
 &= N^T [\nabla_{\underline{x}}^2 L_0 + \sigma N N^T]^{-1} [I - \nabla_{\underline{x}}^2 L_0 \Lambda] [\sigma N N^T + \nabla_{\underline{x}}^2 L_0] B \\
 &= N^T [I - [\nabla_{\underline{x}}^2 L_0 + \sigma N N^T]^{-1} [\nabla_{\underline{x}}^2 L_0 + \sigma N N^T] A \nabla_{\underline{x}}^2 L_0] B \\
 &= N^T [I - A \nabla_{\underline{x}}^2 L_0] B \\
 &= I \quad .
 \end{aligned}$$

□

As a direct consequence of Lemma 4.2.5, we have the following Corollary.

Corollary 4.2.1

Assume that the conditions of Lemma 4.2.5 are satisfied, then, as $\sigma \uparrow \infty$,

$$[\nabla_{\underline{v}}^2 D_{\sigma}(\underline{v})]^{-1} \sim -\sigma I \quad .$$

□

This result was originally proved by Powell (1969) and also by Fletcher (1975). It gives an intuitive explanation for the fast rate of convergence, and supports the use of the Powell/Hestene update.

Note that the second part of the decomposition of $[\nabla_{\underline{v}}^2 D_{\sigma}]^{-1}$ does not involve the parameter σ . In what follows, from examining the decomposition of $[\nabla_{\underline{v}}^2 D_{\sigma}]^{-1}$, the Newton iteration for $\underline{v}$ based on maximizing

$D_0(\underline{v})$ is decomposed into a Powell/Hestenes iteration followed by a Newton-like correction. It is then shown that this Newton-like correction can be derived from several approaches.

Consider a perturbed problem $P[\underline{\theta}]$:

$$\begin{aligned} &\text{minimize} && f(\underline{x}) \\ &\text{subject to} && \underline{h}(\underline{x}) = \underline{\theta} . \end{aligned} \tag{4.2-66}$$

Lemma 4.2.6

Assume that $\underline{x}^*$ (together with $\underline{v}^*$) is a regular local solution of an ECP, and that the conditions of Lemma 4.2.4 are satisfied, then there exists $\delta > 0$ such that, $\forall \underline{\theta} \in N(0, \delta)$, $P[\underline{\theta}]$ has a unique solution pair $(\underline{x}(\underline{\theta}), \underline{v}(\underline{\theta}))$ as differentiable functions of $\underline{\theta}$, $(\underline{x}(0), \underline{v}(0)) = (\underline{x}^*, \underline{v}^*)$. Furthermore,

$$\nabla_{\underline{\theta}} \underline{x}(\underline{\theta}) = B(\underline{x}(\underline{\theta}), \underline{v}(\underline{\theta})) ,$$

and

$$\nabla_{\underline{\theta}} \underline{v}(\underline{\theta}) = Q(\underline{x}(\underline{\theta}), \underline{v}(\underline{\theta})) .$$

Proof

The system of non-linear equations to be satisfied is

$$\begin{aligned} \nabla f(\underline{x}) - \underline{v}^T \underline{h}(\underline{x}) &= 0 \\ \underline{h}(\underline{x}) &= 0 . \end{aligned} \tag{4.2-67}$$

Since (4.2-67) is satisfied when $(\underline{x}, \underline{v}) = (\underline{x}^*, \underline{v}^*)$ and $\underline{\theta} = 0$, and the Jacobian of (4.2-67) with respect to $(\underline{x}, \underline{v})$ is $J(\underline{x}, \underline{v})$, the results follow directly from (4.2-60) and the implicit function theorem. $\square$

Remark 4.2.6

Given a $\bar{v} \in N(v^*, \delta)$, let $\bar{\theta} = h(x(\bar{v}, \sigma))$, then $(x(\bar{v}, \sigma), \Lambda(\bar{v}, \sigma))$ is the unique solution pair of the perturbed problem $P[\bar{\theta}]$: minimize $f(x)$ subject to $h(x) = \bar{\theta}$. Therefore, for a given $v^{(k)}$, there is the correspondence $h^{(k)} = \theta^{(k)}$ and $\Lambda^{(k)} = v(\theta^{(k)})$.

Theorem 4.2.5

Assume that x^* is a regular local solution of an ECP, and that the conditions of Lemmas 4.2.5 and 4.2.6 are satisfied. Consider iterations for updating the multiplier v , then

$$\begin{array}{ccccc} \text{Newton step} & & \text{Powell/Hestenes} & & \text{Newton-like step from} \\ & = & & + & \\ \text{to maximize } D_{\sigma} & & \text{step} & & \text{truncated expansion of } v. \end{array}$$

Proof

By truncating the Taylor's expansion for $v(\theta)$, the update for $v^{(k+1)}$ at $v = v(\theta^{(k)})$ has the form:

$$v^{(k+1)} = v(\theta^{(k)}) - v_{\theta} v(\theta^{(k)}) \cdot \theta^{(k)}. \quad (4.2-68)$$

The result then follows immediately from Remark 4.2.6 and Lemmas 4.2.5 and 4.2.6. □

Convex Cases

If the constrained problem is (locally) convex, the convexifying parameter σ can be set equal to zero. In this case, the augmented dual problem becomes the classical dual problem. The decomposition result of Lemma 4.2.5 then implies that the difference between the Newton iteration for v based on maximizing D_{σ} and the Powell/Hestenes iteration is just

the Newton iteration for maximizing D_0 .

The following Lemma is a trivial consequence of Lemma 4.2.5.

Lemma 4.2.7

For a convex programming problem, assume that the conditions of Lemma 4.2.5 are satisfied, then

$$[\nabla_{\tilde{v}}^2 D_{\sigma}(\tilde{v})]^{-1} = -\sigma I + [\nabla_{\tilde{v}}^2 D_0(\Lambda(\tilde{v}, \sigma))]^{-1}.$$

□

Theorem 4.2.6

For a convex programming problem, assume that the conditions of Lemma 4.2.5 are satisfied. Consider iterations for updating the multiplier $\tilde{v}$, then

$$\begin{array}{ccccc} \text{Newton step} & & \text{Powell/Hestenes} & & \text{Newton step} \\ & = & & + & \\ \text{to maximize } D_{\sigma} & & \text{step} & & \text{to maximize } D_0. \end{array}$$

Proof

Given a $\tilde{v}^{(k)}$, since $\tilde{x}^{(k)}$ is the minimizer of $L_{\sigma}(\tilde{x}, \tilde{v}^{(k)})$, it follows from (4.2-11) that

$$\nabla_{\tilde{v}} D_{\sigma}(\tilde{v}^{(k)}) = -\tilde{h}^{(k)}. \quad (4.2-69)$$

Since $\nabla_{\tilde{x}} L_0(\tilde{x}^{(k)}, \tilde{\Lambda}^{(k)}) = 0$, $\tilde{x}^{(k)}$ is also the minimizer of $L_0(\tilde{x}, \tilde{\Lambda}^{(k)})$, and hence

$$\nabla_{\tilde{v}} D_0(\tilde{\Lambda}^{(k)}) = -\tilde{h}^{(k)} = \nabla_{\tilde{v}} D_{\sigma}(\tilde{v}^{(k)}). \quad (4.2-70)$$

The result then follows from Lemma 4.2.7.

□

Non-convex Cases

For a convex problem, the classical dual problem is always well defined. Hence, without involving the penalty parameter σ , the constrained problem can be treated as an unconstrained maximization problem. For a general non-convex problem, the classical dual problem may not be defined. But, as in the convex case, it is instructive to treat the constrained problem in some sense as an unconstrained problem without involving the penalty parameter σ .

Consider an *associated non-linear problem* (ANP) :

$$\nabla_{\tilde{\theta}} K(\tilde{\theta}) = 0$$

where

(4.2-71)

$$K(\tilde{\theta}) = f(\tilde{x}(\tilde{\theta})) - \tilde{v}^{*T} \cdot \tilde{\theta} .$$

By construction, it follows that

$$\nabla_{\tilde{\theta}} K(\tilde{\theta}) = \tilde{v}(\tilde{\theta}) - \tilde{v}^* . \quad (4.2-72)$$

Hence, if $\bar{\tilde{\theta}}$ is a solution of the ANP, then $\tilde{v}(\bar{\tilde{\theta}}) = \tilde{v}^*$.

Remark 4.2.7

From (4.2-72) and Lemma 4.2.6, it follows that

$$\nabla_{\tilde{\theta}}^2 K(\tilde{\theta}) = Q(\tilde{x}(\tilde{\theta}), \tilde{v}(\tilde{\theta})) . \quad (4.2-73)$$

Note that, for a general ECP, the matrix Q is not required to be non-singular. Thus, besides the obvious solution $\tilde{\theta}^* = 0$, a family of solutions may exist for the ANP.

Theorem 4.2.7

Assume that $\tilde{x}^*$ is a regular local solution of an ECP, and that the conditions of Lemmas 4.2.5 and 4.2.6 are satisfied. Consider iterations for updating the multiplier $\tilde{v}$, then

$$\begin{array}{lcl} \text{Newton step} & = & \text{Powell/Hestenes} \quad \text{Newton-like correction} \\ \text{to maximize } D_{\sigma} & & \text{step} \quad + \quad \text{based on solving } \nabla_{\theta} K(\tilde{\theta}) = 0. \end{array}$$

Proof

Since $K(\tilde{\theta})$ is twice differentiable, a linear estimate of its gradient is

$$\nabla_{\theta} K(\tilde{\theta}) = \nabla_{\theta}^2 K(\tilde{\theta}) \cdot \tilde{\theta} \quad . \quad (4.2-74)$$

The result then follows from (4.2-72), (4.2-73) and Remark 4.2.6. $\square$

4.3 ACCELERATED CONVERGENCE FOR THE POWELL/HESTENES MULTIPLIER METHOD

4.3.1 Introduction

In what follows, an augmented Lagrangian method will be referred to as a *Newton method* if it uses the update formula (4.2-13) or (4.2-33), a *Powell/Hestenes method* if it uses the update formula (4.2-18).

By viewing it as a Newton technique for maximizing the augmented dual function $D_{\sigma}(\underline{v})$, the Newton method has a super-linear rate of convergence. However, because the Newton method requires derivative information of the problem functions, it is not suitable for problems in which the derivatives are not available; or in which the first derivatives are available, but it is not practical nor economical to keep an estimate of $\nabla_{\underline{x}}^2 L_{\sigma}$.

On the other hand, the Powell/Hestenes method is very simple and attractive. It does not require any derivative information in updating the multiplier, and the rate of convergence can be controlled by adjusting the penalty parameter. However, the method can hardly be described as optimal: (i) the method is just a fixed step size gradient method for maximizing $D_{\sigma}(\underline{v})$; (ii) if the method is applied to two different problems, having the same constraints $\underline{h}(\underline{x}) = 0$, but different objective functions $f(\underline{x})$ and $f(\underline{x}) + \alpha \underline{h}(\underline{x})^T \underline{h}(\underline{x})$, the Powell/Hestenes method will use the same fixed step size σ for both problems.

By viewing the Powell/Hestenes method as a fixed step size gradient method for maximizing $D_{\sigma}(\underline{v})$, in the case of convex programming problems, Bertsekas (1975b) has devised a rule to modify the step size σ every second iteration. But, because it is only applied every second iteration and because it requires the problem to be convex, this device is not suitable for implementation in a general algorithm.

Because of its practical importance, there is a strong interest in any modification which will improve, *without requiring derivatives*, the convergence of the Powell/Hestenes method. In subsection 4.2.4, it is shown that the Newton update for $\underline{v}$, based on maximizing $D_{\sigma}(\underline{v})$, can be decomposed into taking the Powell/Hestenes update followed by another Newton-like update. This result implies that possible techniques for accelerating the convergence of the Powell/Hestenes method are the ones which take into account the second (Newton-like) part of the decomposition. In this section, utilizing the decomposition result, two techniques are proposed for accelerating the convergence of the Powell/Hestenes method. The first is a simple device superimposed on the original Powell/Hestenes method to make use of information from the previous iteration. The second is a quasi-Newton technique for maximizing $D_{\sigma}(\underline{v})$. Relative merits of the two techniques are discussed. Numerical results, comparing various updating schemes for the multiplier, are presented.

4.3.2 Derivation of a Simple Accelerating Device

In this subsection, utilizing the results of subsection 4.2.4, a simple device is proposed for improving, without requiring derivatives, the convergence of the Powell/Hestenes method. For problems with only one constraint, this accelerating technique is equivalent to replacing the second (Newton-like) part of the decomposition result of Lemma 4.2.5 by a finite difference approximation. By interpreting it as a relaxation technique, this device is then generalized for the general problem (MPP). The resulting method obtained from superimposing the accelerating device on the basic Powell/Hestenes method will be referred to as *the accelerated Powell/Hestenes method*.

From Lemma 4.2.6, a linear approximation to $\underline{v}^*$ is given by

$$\tilde{v}^* = \tilde{v}(\tilde{\theta}) - \nabla_{\tilde{\theta}} \tilde{v}(\tilde{\theta}) \cdot \tilde{\theta} \quad (4.3-1)$$

For the special case of a problem with only one constraint, by taking a finite difference approximation to the unknown $dv(\theta)/d\theta$, an appropriate update for v has the form:

$$v^{(k+1)} = v(\theta^{(k)}) - \left[\frac{v(\theta^{(k)}) - v(\theta^{(k-1)})}{\theta^{(k)} - \theta^{(k-1)}} \right] \cdot \theta^{(k)} \quad (4.3-2)$$

Following Remark 4.2.6, (4.3-2) can be viewed as a relaxation method:

$$v^{(k+1)} = \Lambda^{(k)} + w^{(k)} (\Lambda^{(k)} - \Lambda^{(k-1)}) \quad (4.3-3)$$

with $w^{(k)} = -\frac{h^{(k)}}{h^{(k)} - h^{(k-1)}}$ as the optimal parameter. In this relaxation framework, (4.3-3) can then be generalized for the general problem with an arbitrary number of constraints $l \leq n$. Consider

$$\tilde{v}(w) = \tilde{\Lambda}^{(k)} + w(\tilde{\Lambda}^{(k)} - \tilde{\Lambda}^{(k-1)}) \quad (4.3-4)$$

The relaxation problem is that of computing an 'optimal' $w^{(k)}$ which will improve the convergence of the Powell/Hestenes method, and which will reduce to the *secant* step (4.3-2) if the problem has only one constraint.

Before proceeding to find the 'optimal' $w^{(k)}$, it will be shown in the following Theorem that, as a consequence of the convergence of the Powell/Hestenes method, convergence is also assured for a more general class of updates (4.3-4).

Theorem 4.3.1

Assume that $\tilde{x}^*$ (together with $\tilde{v}^*$) is a regular local solution of an ECP, the sequence $\{w^{(k)}\} \subseteq [a, b]$ a finite interval, and that the

conditions of Theorem 4.2.3 are satisfied, then there exists $\bar{\sigma} > 0$ such that, $\forall \sigma \geq \bar{\sigma}$, the sequence $\{\tilde{v}^{(k)}\}$ generated by the update

$$\tilde{v}^{(k+1)} = \tilde{\Lambda}^{(k)} + w^{(k)}(\tilde{\Lambda}^{(k)} - \tilde{\Lambda}^{(k-1)})$$

converges to $\tilde{v}^*$.

Proof

Consider

$$\|\tilde{v}^{(k+1)} - \tilde{v}^{(k)}\| = \|(1+w^{(k)})\tilde{\Lambda}^{(k)} - w^{(k)}(\tilde{\Lambda}^{(k-1)} - \tilde{v}^*)\| \quad (4.3-5)$$

From Theorem 4.2.3, for σ sufficiently large, there exists $\kappa > 0$ such that

$$\|\tilde{v}^{(k+1)} - \tilde{v}^*\| \leq |1+w^{(k)}| \frac{\kappa}{\sigma} \|\tilde{v}^{(k)} - \tilde{v}^*\| + |w^{(k)}| \frac{\kappa}{\sigma} \|\tilde{v}^{(k-1)} - \tilde{v}^*\| \quad (4.3-6)$$

Since $\{w^{(k)}\} \subseteq [a, b]$ which is a fixed interval, there exists $\bar{\sigma} > 0$ such that, $\forall \sigma \geq \bar{\sigma}$ and $\forall k \geq 1$,

$$|1+w^{(k)}| \frac{\kappa}{\sigma} \leq \alpha, \quad |w^{(k)}| \frac{\kappa}{\sigma} \leq \beta, \quad \text{with } \alpha + \beta < 1 \quad (4.3-7)$$

Let $s_k = \|\tilde{v}^{(k)} - \tilde{v}^*\|$, then it follows that

$$s_{k+1} \leq \alpha s_k + \beta s_{k-1} \quad (4.3-8)$$

The linear difference equation $\phi_{k+1} = \alpha \phi_k + \beta \phi_{k-1}$ has the general solution

$\phi_k = \gamma_1 t_1^k + \gamma_2 t_2^k$, where t_1 and t_2 are roots of the equation

$t^2 - \alpha t - \beta = 0$, with $-1 < t_2 < 0 < t_1 < 1$; $\phi_0 = s_0 = \gamma_1 + \gamma_2$, $\phi_1 = s_1 = \gamma_1 t_1 +$

$\gamma_2 t_2$. Therefore

$$s_{k+1} \leq \alpha s_k + \beta s_{k-1} \leq \alpha \phi_k + \beta \phi_{k-1} = \phi_{k+1} \quad (4.3-9)$$

and hence

$$\|\tilde{v}^{(k)} - \tilde{v}^*\| \leq \gamma_1 t_1^k + \gamma_2 t_2^k \quad \square$$

Remark 4.3.1

(a) The convergence result of Theorem 4.3.1 does not guarantee that

$$\|\tilde{v}^{(k+1)} - \tilde{v}^*\| \leq \|\tilde{v}^{(k)} - \tilde{v}^*\|, \quad \forall k.$$

(b) From subsection 4.2.3, because Theorem 4.2.3 is also applicable to problems with inequality constraints (with or without strict complementarity), this implies that Theorem 4.3.1 is also applicable to problems with inequality constraints (with or without strict complementarity).

Convex Cases

For a general convex programming problem, it is known that the constrained problem can be treated as an unconstrained maximization of the classical dual function $D_0(\tilde{v})$. In this case, a natural 'optimal' choice for w in (4.3-4) is the value w which solves the subproblem:

$$\underset{w}{\text{maximize}} \quad D_0(\tilde{v}(w)) \quad (4.3-10)$$

Assuming that $D_0(\tilde{v}(w))$ is quadratic in w , then the optimal value for w is

$$w^{(k)} = - \frac{\tilde{h}^{(k)T} \cdot (\tilde{\Lambda}^{(k)} - \tilde{\Lambda}^{(k-1)})}{(\tilde{h}^{(k)} - \tilde{h}^{(k-1)})^T \cdot (\tilde{\Lambda}^{(k)} - \tilde{\Lambda}^{(k-1)})} \quad (4.3-11)$$

Non-convex Cases

For a general non-convex problem, the classical dual problem of maximizing $D_0(\tilde{v})$ may not be defined. However, as pointed out in subsection 4.2.4, an unconstrained interpretation of the constrained problem

exists: the ANP, solving $\nabla_{\theta} K(\theta) = 0$. Consider an update of the form:

$$\tilde{\theta}(w) = \tilde{\theta}^{(k)} + w(\tilde{\theta}^{(k)} - \tilde{\theta}^{(k-1)}) \quad (4.3-12)$$

For the ANP, the problem is to find an $\bar{\theta}$ such that $\nabla_{\theta} K(\bar{\theta}) = 0$. Hence, the appropriate subproblem for determining the optimal w is

$$\frac{\partial}{\partial w} K(\theta(w)) = 0 \quad (4.3-13)$$

Again, by assuming that $K(\theta(w))$ is quadratic in w , the optimal value for w is

$$w^{(k)} = - \frac{(\tilde{\theta}^{(k)} - \tilde{\theta}^{(k-1)})^T \cdot (\tilde{v}(\tilde{\theta}^{(k)}) - \tilde{v}^*)}{(\tilde{\theta}^{(k)} - \tilde{\theta}^{(k-1)})^T \cdot (\tilde{v}(\tilde{\theta}^{(k)}) - \tilde{v}(\tilde{\theta}^{(k-1)}))} \quad (4.3-14)$$

From considering solving the ANP, the 'optimal' update has the form (4.3-12) and (4.3-14). In the present form, the update (4.3-12) and (4.3-14) is not suitable. This is because it gives an update for θ instead of v , and also that the calculation for $w^{(k)}$ in (4.3-14) involves the unknown quantity v^* .

From Lemma 4.2.6, ignoring higher order terms, it follows that

$$\tilde{v}(\tilde{\theta}^{(k)}) = \tilde{v}^* + Q \cdot \tilde{\theta}^{(k)} \quad (4.3-15)$$

Using the above relation and Remark 4.2.6, (4.3-12) and (4.3-14) then become

$$\tilde{v}^{(k+1)} = \tilde{\Lambda}^{(k)} + w^{(k)}(\tilde{\Lambda}^{(k)} - \tilde{\Lambda}^{(k-1)}) \quad (4.3-16)$$

with

$$w^{(k)} = - \frac{\tilde{h}^{(k)T} \cdot (\tilde{\Lambda}^{(k)} - \tilde{\Lambda}^{(k-1)})}{(\tilde{h}^{(k)} - \tilde{h}^{(k-1)})^T \cdot (\tilde{\Lambda}^{(k)} - \tilde{\Lambda}^{(k-1)})} \quad (4.3-17)$$

As can be expected, in the case of convex problems, the 'optimal'

value for $w^{(k)}$ in (4.3-17) is the same as in (4.3-11). Therefore, the proposed accelerating device has the desired properties that it is the same for convex and non-convex problems and that it reduces to the secant correction (4.3-2) when the problem has only one constraint. Above all, this accelerating device is very simple and does not require derivative information. Its numerical performance (reported in subsection 4.3.4) has been *surprisingly* successful.

Remark 4.3.2

(a) As mentioned in Remark 4.2.1, the analysis in this subsection does not require that the parameter σ be held constant throughout. In fact, since the parameter σ is a crucial controlling parameter for the rate of convergence, it is usually allowed to vary not only from iteration to iteration but also differently for each constraint. Thus, the direction $\tilde{\Lambda}^{(k)} - \tilde{\Lambda}^{(k-1)}$ is arbitrary, depending on the strategy for changing σ . The proposed accelerating device is just an 'approximated' linear optimization along the direction $\tilde{\Lambda}^{(k)} - \tilde{\Lambda}^{(k-1)}$, with no guarantee for improvement.

(b) In the case of problems which have only one constraint, or which have special structure, (4.3-2) shows that the accelerating device is just a secant method, and hence a super-linear rate of convergence can be expected. Generally, however, only linear convergence can be expected.

(c) Note that the relaxation parameter $w^{(k)}$ is computed using only the function values at $\tilde{x}^{(k)}$ and $\tilde{x}^{(k-1)}$. Because the problem functions are non-linear, it is possible that 'optimal' $w^{(k)}$ is far from optimal. The convergence result of Theorem 4.3.1 requires that, for every iteration, the relaxation parameter $w^{(k)}$ belongs to a finite interval. Therefore, in actual implementations, it is recommended that $w^{(k)}$ be restricted to the interval $[-0.4, 3.0]$. It should be obvious that the choice $[-0.4, 3.0]$

is heuristic and arbitrary.

(d) The proposed accelerating device should be viewed as a modification superimposed on the original Powell/Hestenes method. The correction step is based on utilizing information from the previous iteration, and hence, by its very nature, this accelerating device cannot be used before the second iteration.

4.3.3 Quasi-Newton Update for the Multiplier

In using an augmented Lagrangian method, the original constrained problem is transformed into an unconstrained maximization of the augmented dual function $D_{\sigma}(\underline{v})$. Because the Powell/Hestenes method is just a fixed step size gradient method, it seems rather surprising that more efficient unconstrained methods such as quasi-Newton methods, which have super-linear rate of convergence, have not been considered. In this subsection, a possible strategy for implementing the quasi-Newton technique is proposed. Initially, the resulting method, referred to as the *quasi-Newton method*, is identical to the Powell/Hestenes method, and hence it can be viewed as a technique for accelerating the convergence of the Powell/Hestenes method. Some desirable and undesirable features of the quasi-Newton method are discussed. Its numerical performance is reported later on in subsection 4.3.4.

Updating the estimate hessian

One of the major obstacles for implementing the quasi-Newton technique seems to be that the parameter σ is usually allowed to increase from iteration to iteration. The size of the increase may also be different for each constraint. The objective function $D_{\sigma}(\underline{v})$ need not be the same function at each iteration, and hence the updating of the estimate for

$\nabla_{\mathbf{v}}^2 D_{\sigma}$ (or $[\nabla_{\mathbf{v}}^2 D_{\sigma}]^{-1}$) may not be meaningful. Utilizing the results of subsection 4.2.4, this difficulty can now be overcome.

Suppose that, after evaluating $D_{\sigma}(\mathbf{v}^{(k)})$, the algorithm updates $\mathbf{v}^{(k)}$ to $\mathbf{v}^{(k+1)}$ and also increases σ to $\sigma + \Delta\sigma_i$ for the i th constraint. Let $[\Delta\sigma]$ denote the diagonal matrix with $\Delta\sigma_i$ along its diagonal. Lemma 4.2.5 then gives that

$$[\nabla_{\mathbf{v}}^2 D_{\sigma+\Delta\sigma}(\mathbf{v}^{(k)} + [\Delta\sigma]\mathbf{h}^{(k)})]^{-1} = -[\Delta\sigma] + [\nabla_{\mathbf{v}}^2 D_{\sigma}(\mathbf{v}^{(k)})]^{-1}, \quad (4.3-18)$$

with

$$\nabla_{\mathbf{v}} D_{\sigma+\Delta\sigma}(\mathbf{v}^{(k)} + [\Delta\sigma]\mathbf{h}^{(k)}) = \nabla_{\mathbf{v}} D_{\sigma}(\mathbf{v}^{(k)}) = -\mathbf{h}^{(k)}. \quad (4.3-19)$$

If an estimate for $[\nabla_{\mathbf{v}}^2 D_{\sigma}]^{-1}$ is kept, then (4.3-18) gives an estimate for $[\nabla_{\mathbf{v}}^2 D_{\sigma+\Delta\sigma}]^{-1}$. The gradient of $D_{\sigma+\Delta\sigma}$ at $\mathbf{v}^{(k)} + [\Delta\sigma]\mathbf{h}^{(k)}$ is given by (4.3-19). Assuming that $\mathbf{v}^{(k+1)} \neq \mathbf{v}^{(k)} + [\Delta\sigma]\mathbf{h}^{(k)}$, then (4.3-19) together with the gradient of $D_{\sigma+\Delta\sigma}$ at $\mathbf{v}^{(k+1)}$ can be used to update $[\nabla_{\mathbf{v}}^2 D_{\sigma+\Delta\sigma}]^{-1}$, using some update formula (see, for example, Dennis and Moré (1977)).

Choice of the starting matrix

In general, without any second derivative information, quasi-Newton methods start with I (or $-I$ for maximization) as the initial estimate for the hessian. In this case, the decomposition result of Lemma 4.2.5 suggests that $-\sigma I$ is more appropriate as the initial estimate for $[\nabla_{\mathbf{v}}^2 D_{\sigma}]^{-1}$. This corresponds to starting off as the Powell/Hestenes method, ignoring the second part of the decomposition of $[\nabla_{\mathbf{v}}^2 D_{\sigma}]^{-1}$. The estimate of $[\nabla_{\mathbf{v}}^2 D_{\sigma}]^{-1}$ can then be updated as more gradient information becomes available.

Linear search

Another problem associated with the quasi-Newton technique is the linear search and step size control. Because every evaluation of $D_{\sigma}(\underline{v})$ gives an update for $\underline{v}$ using the Powell/Hestenes formula, any evaluation of $D_{\sigma}(\underline{v})$ for the linear search can be considered wasteful. However, these wasteful evaluations of $D_{\sigma}(\underline{v})$ may not occur. It seems reasonable to expect that if σ is sufficiently large for the Powell/Hestenes method (i.e., use $-\sigma I$ as the estimate for $[\nabla_{\underline{v}}^2 D_{\sigma}]^{-1}$) to converge, then, the quasi-Newton step (with step size = 1) should satisfy some weak line search conditions, such as the Goldstein condition (see, for example, Dennis and Moré (1977)). Computational results (reported in subsection 4.3.4) support this argument. When the Goldstein condition is not satisfied, this is in a sense an indication that the penalty parameter is not sufficiently large. Hence σ should be increased instead of reducing the step size as in normal unconstrained optimization.

Following the above arguments, the quasi-Newton method for maximizing $D_{\sigma}(\underline{v})$ is as follows:

Algorithm 4.3.1 (for an ECP)

- (i) set $H^{(1)} := -\sigma I$, $\underline{v}^{(1)} := 0$ and $k := 1$;
- (ii) find $\underline{x}^{(k)}$ which minimizes $L_{\sigma}(\underline{x}, \underline{v}^{(k)})$ (i.e., evaluating $D_{\sigma}(\underline{v}^{(k)})$ and $\nabla_{\underline{v}} D_{\sigma}(\underline{v}^{(k)})$);
- (iii) if $\|\underline{h}^{(k)}\|_{\infty} \leq \epsilon$, stop;
- (iv) if $k = 1$, go to (viii);
- (v) Goldstein step size control, for (say) $\alpha = 10^{-4}$,
 if $\left[\phi = - \frac{D_{\sigma}(\underline{v}^{(k)}) - D_{\sigma}(\underline{v}^{(k-1)})}{(\underline{v}^{(k)} - \underline{v}^{(k-1)})^T \cdot \underline{h}^{(k-1)}} \right] \geq \alpha$, go to (vii)

- (vi) increase in the penalty parameter, for (say) $\Delta\sigma_i = 10, \forall i$,
 set $\sigma_i := \sigma_i + \Delta\sigma_i$, $H^{(k)} := -[\Delta\sigma] + H^{(k-1)}$, $\tilde{v}^{(k+1)} := \tilde{v}^{(k)}$,
 $\tilde{v}^{(k)} := \tilde{v}^{(k)} + [\Delta\sigma]h^{(k)}$ and $k := k+1$, and return to (ii);
- (vii) use an update formula, such as the BFGS formula, to calculate
 $H^{(k)}$ from $H^{(k-1)}$, $\tilde{v}^{(k)} - \tilde{v}^{(k-1)}$ and $\tilde{h}^{(k)} - \tilde{h}^{(k-1)}$;
- (viii) set $\tilde{v}^{(k+1)} := \tilde{v}^{(k)} + H^{(k)}\tilde{h}^{(k)}$ and $k := k+1$, and return to (ii).

Practical consideration

The quasi-Newton method has the desired property that it accelerates the convergence of the Powell/Hestenes method without requiring derivative information of the problem functions. However, the only advantage of using a quasi-Newton technique is that, asymptotically, the rate of convergence is super-linear (Dennis and Moré (1977)). Even if, the problem of maximizing different functions $D_\sigma(v)$ at each iteration can be overcome, and the linear search subproblem can be avoided, the quasi-Newton method remains here purely as an academic exercise. It is not recommended for a general implementation. Some arguments are listed as follows:

- (i) The super-linear rate of convergence is asymptotic and will only occur after a good estimate of $[\nabla_v^2 D_\sigma]^{-1}$ has been achieved. Normally, because the rate of convergence can be improved by increasing σ , the required accuracy is achieved in a few iterations, well before the super-linear rate of convergence sets in.
- (ii) When inequality constraints are present, the number of constraints $(\ell+m)$ can be arbitrarily large. The task of storing and updating an estimate of $[\nabla_v^2 D_\sigma]^{-1}$ may not be worth-while.
- (iii) When inequality constraints are involved, if strict complementarity

does not hold, or is close to violation, then $[\nabla_v^2 D_\sigma]^{-1}$ has jump discontinuities at or near $\underline{v}^*$. These jump discontinuities may or may not slow down convergence. However, the very existence of such discontinuities requires special attention, and hence further complication. Simpler methods such as the Powell/Hestenes method or the accelerated Powell/Hestenes method are preferred.

(iv) The task of storing and updating $[\nabla_v^2 D_\sigma]^{-1}$ destroys the simplicity of the underlying Powell/Hestenes update. Because of the fast rate of convergence of the Powell/Hestenes method for σ large enough (see Theorem 4.2.3), numerical performance of the quasi-Newton method need not be better.

4.3.4 Numerical Performance and Discussion

In this subsection, numerical results comparing the effectiveness of various update formulas for the multiplier are reported. A general program has been written and modified for each update formula. For equality constrained problems, the algorithm is as follows:

Algorithm 4.3.2

Initially set $k := 1$, $\underline{v}^{(1)} := 0$ and $\sigma := \sigma^{(1)}$;

- (i) evaluate $\underline{x}^{(k)} = \underline{x}(\underline{v}^{(k)}, \sigma)$;
- (ii) if $\|\underline{h}^{(k)}\|_\infty \leq \varepsilon$, stop;
- (iii) update $\underline{v}^{(k)}$ by some scheme, set $k := k+1$, and return to (i).

The initial choice for $\underline{v}$ is $\underline{v}^{(1)} = 0$, which corresponds to starting as the penalty function method. If the convexifying parameter σ is allowed to vary, then an automatic scheme should increase σ when convergence is slow, and keep σ constant when the convergence rate is satisfactory. Since the purpose here is to compare the effectiveness of

various updating schemes for $\underline{v}$, the parameter σ is not allowed to vary.

The choice of σ used here has the form 10^{2i} , for an integer i . For each test problem, in the first run $\sigma = 10^{2i}$ is chosen such that it is just large enough to force convergence using the Powell/Hestenes update. The effect of the parameter σ on the rate of convergence is then observed by allowing σ to vary between subsequent runs.

A quasi-Newton subroutine FUNMIN (this routine is an implementation of Fletcher's (1970a) algorithm) from the ANU's subroutine library is used to minimize $L_\sigma(\underline{x}, \underline{v})$ with respect to $\underline{x}$; the tolerance in FUNMIN is set at 10^{-8} . The algorithm is terminated when an accuracy of $\epsilon = 10^{-7}$ is achieved.

Remark 4.3.3

Several authors (Kort and Bertsekas (1976), etc.) have suggested for less accurate unconstrained minimization in the first few iterations. In this situation, the appropriate Powell/Hestenes update, as pointed out by Haarhoff and Buys (1970), is given by $\underline{\Lambda}^{(k)} = (\underline{N}^T \underline{N})^{-1} \underline{N}^T \nabla f(\underline{x}^{(k)})$. However, because most efficient unconstrained routines have a super-linear rate of convergence, a more accurate result (above truncation error) is usually achieved by taking one or two more steps using only one or two more function/gradient evaluations. Because of this, the relatively simpler form $\underline{\Lambda}^{(k)} = \underline{v}^{(k)} - \sigma \underline{h}^{(k)}$ is preferred. Another point favouring the use of the update $\underline{\Lambda}^{(k)} = \underline{v}^{(k)} - \sigma \underline{h}^{(k)}$ over the update $\underline{\Lambda}^{(k)} = (\underline{N}^T \underline{N})^{-1} \underline{N}^T \nabla f(\underline{x}^{(k)})$ is that the latter requires derivative information.

As mentioned in subsection 4.2.2, after the minimization, the estimate of $\nabla_{\underline{x}}^2 L_\sigma$ is then extracted from FUNMIN to be used for the Newton update (4.2-13) or (4.2-33). The subproblem (4.2-13) or (4.2-33) is solved by the appropriate subroutine from the ANU's subroutine library, with

tolerance set at 10^{-8} .

For the quasi-Newton method (see subsection 4.3.3), the BFGS formula (see, for example, Dennis and Moré (1977)) is used for updating the estimate of $[\nabla_{\sigma}^2 D]^{-1}$, and the estimate is kept in Choleski decomposition. The constant α in the Goldstein condition is set at 10^{-4} .

A wide selection of test problems has been used. These are: the problem RS due to Rosen and Suzuki (1965), the problem P due to Powell (1969), the problems M1 and M4 of Miele, *et al.* (1972) (examples 8.1 and 8.4), and three test problems CV1, 2 and 7 collected by Colville (1968).

When the parameter σ is large enough for the Powell/Hestenes method to converge, all other methods (i.e., the accelerated Powell/Hestenes, the quasi-Newton and the Newton methods) should have no trouble converging. And this is borne out by numerical evidence. For the case of the quasi-Newton method this also means that the Goldstein condition (with $\alpha = 10^{-4}$) is always satisfied. For the range of test problems considered, when the parameter σ is large enough for convergence by the Powell/Hestenes method, the computed optimal value of $w^{(k)}$ for the accelerated Powell/Hestenes method always lies in the range $[-0.4, 3.0]$.

Recall that the accelerating device of the accelerated Powell/Hestenes method makes use of information from the previous iteration and that the quasi-Newton method has $-\sigma I$ as an initial estimate of $[\nabla_{\sigma}^2 D]^{-1}$. Hence, in interpreting the numerical results of Table 4.3.1, it should be kept in mind that both the accelerated Powell/Hestenes and the quasi-Newton methods are identical to the underlying Powell/Hestenes method for the first two iterations.

From Table 4.3.1, when σ is relatively small, the convergence by the Powell/Hestenes method is slow, and marked improvement using the

accelerated Powell/Hestenes method or the quasi-Newton method can be observed. The numerical results also confirm the faster rate of convergence of the Newton method. When the parameter σ is increased, as predicted in subsections 4.2.2 and 4.2.4, the rate of convergence is dominated by the size of σ . The performance is more or less the same for each updating scheme, again with the accelerated Powell/Hestenes method standing out as an improvement on the original Powell/Hestenes method.

For the problem CV2, in examining the numerical results more closely, the Powell/Hestenes method seems to be more efficient than the Newton method. This seemingly anomalous behaviour has been noticed before by Fletcher (1975). The structure of CV2 is such that the 15×15 hessian matrix $\nabla_x^2 L_0$ is zero everywhere except for the top-left 5×5 submatrix. The structure of the (constraint) gradients and the result of subsection 4.2.4 then imply that the matrix $Q = B^T \nabla_x^2 L_0 B$ is a fairly sparse matrix. Hence, $[\nabla_v^2 D_\sigma]^{-1} = -\sigma I - Q$ is likely to be better approximated by $-\sigma I$ then by the estimate from the minimization routine.

The Powell/Hestenes method V.S. the accelerated Powell/Hestenes method

The improvement in numerical performance and the simplicity of the accelerating device suggest strongly in favour of using the accelerating technique superimposed on the original Powell/Hestenes method.

The accelerated Powell/Hestenes method V.S. the quasi-Newton method

The arguments discussed in subsection 4.3.3 and the simplicity of the accelerating device both favour the use of the accelerated Powell/Hestenes method. Comparing numerical performance, because of its super-linear rate of convergence, the quasi-Newton method requires fewer iterations in situations when the parameter σ is kept relatively small. However, in practical implementations when the parameter σ is allowed to vary, the

simplicity of the accelerating device together with the fast rate of convergence of the underlying Powell/Hestenes method (for large σ) is the deciding factor.

The accelerated Powell/Hestenes method V.S. the Newton method

For moderate values of σ , the numerical performance of the Newton method is only slightly better than that of the accelerated Powell/Hestenes method. This slight improvement in numerical performance does not seem to be worth-while when comparing the simplicity of the accelerating device with the Newton update for $\underline{v}$. Some other arguments in favour of using the accelerated Powell/Hestenes method are as follows: (i) when inequality constraints are involved, the subproblem (4.2-33) for the Newton update can be much larger than the original problem; (ii) after the first minimization, with an estimate of $\nabla_{\underline{x}}^2 L_{\sigma}$ and a good starting point, subsequent unconstrained minimizations are relatively easy, and (iii) the rate of convergence of the accelerated Powell/Hestenes method can always be improved by increasing σ .

In the above arguments, it is assumed that the derivatives of the problem functions, specially $\nabla_{\underline{x}}^2 L_{\sigma}$ or an estimate of, are available. If an estimate of $\nabla_{\underline{x}}^2 L_{\sigma}$ is not available, then, clearly, there is no need for comparison between the accelerated Powell/Hestenes method and the Newton method. It is also relevant to point out that if a general quadratic programming subroutine is available for solving the subproblem (4.2-33), then several 'faster' methods for solving the MPP (see, for example, Powell (1977) and Han (1977)) are preferable to the Newton method.

Problem	Type of problem				Parameter σ	Performance			
						Powell/Hestenes	Accelerated P/H	Quasi-Newton	Newton
	n	m_e	m_i	m_a					
P	5	3	-	3	10^{-2}	88 [†]	20	12	9
					1	6	4	6	5
					10^2	3	3	3	3
M1	5	3	-	3	1	98	13	7	4
					10^2	6	4	4	2
M4	5	3	-	3	1	9	5	6	5
					10^2	3	3	3	3
RS	4	-	3	2	1	55	11	10	5
					10^2	5	4	4	4
CV1	5	-	15	4	1	164	51	15	7
					10^2	7	5	6	4
CV2	15	-	20	11	10^2	4	4	4	7
CV7	16	8	32	13	10^2	77	18	13	5
					10^4	5	4	5	4

† Total number of iterations

m_e , m_i and m_a indicate the numbers of equality, inequality and active constraints, respectively.

Table 4.3.1: Number of iterations for various updating schemes.

CHAPTER 5

APPLICATIONS AND CONCLUSION

5.1 CONSTRAINED OPTIMIZATION WITHOUT DERIVATIVES

The application of the methods developed in this thesis to constrained optimization problems without requiring derivative information is now examined. Methods for solving constrained optimization problems which do not require derivative information are of practical interest. The chief reasons are the following:

- (i) For some problems, derivatives are simply not available.
- (ii) For many problems, even though derivatives can be computed, it is not economical nor practical to do so.
- (iii) A method which does not require derivative information can be used to check independently the solution produced by a method which utilizes derivative information. Certain 'errors' in gradient calculations are very difficult to detect. In theory the gradient routines are corrected. Therefore, if they are not what they should be, any algorithm which utilizes derivative information is effectively solving a different problem. There is a likelihood that an algorithm which utilizes derivative information will terminate properly even when the function and gradient routines are inconsistent.
- (iv) When solving a constrained problem, if the derivative information is not required, only $(1+l+m)$ real-valued functions on E^n have to be coded and debugged. On the other hand, if the first derivative information is required, $(1+l+m) \times (n+1)$ real-valued functions on E^n have to be coded and debugged. Consequently, it is likely that, if constrained optimization methods which do not require derivative information are available, they will

be as popular as corresponding methods for unconstrained optimization.

(v) There are obvious savings in (human) resources, if it is not necessary to calculate (analytically), code and debug derivatives. In addition, even if such methods are generally less efficient than methods which utilize derivative information, the trade-off against cost effectiveness cannot be ignored, especially for the once only situations.

In this thesis, attention has been restricted to methods for solving constrained problems which reduce the computational process to a sequence of unconstrained minimization subproblems. When derivatives are not utilized, this implies that corresponding unconstrained minimization routines must be used. Several such unconstrained minimization methods, such as the methods proposed by Powell (1964), Stewart (1967), Gill and Murray (1971) and Brent (1973), are well known and readily available in standard subroutine libraries. Utilizing results developed in this thesis, two methods are proposed for solving constrained optimization problems. The first method has the property that, with respect to the inequality constraints, feasibility is maintained throughout the computation.

A Penalty Function - Modified Barrier Function Method

Consider solving an ICP. In order to maintain feasibility with respect to the inequality constraints, an appropriate technique is the barrier function technique. Clearly, as the work in Chapter 3 indicates, the best barrier function method to use is the modified barrier function method:

$$B(\tilde{x}, r_k u^{(k-1)}) = f(\tilde{x}) - r_k \sum_{i \in J} u_i^{(k-1)} \log[g_i(\tilde{x})] \quad (5.1-1)$$

Since the determination of the $u_i^{(k)} = r_k u_i^{(k-1)} / g_i(\tilde{x}^{(k)})$, $i \in J$, does not involve derivatives, provided a corresponding unconstrained routine is

available, the modified barrier function method can be used as a constrained optimization method without requiring derivatives.

When equality constraints are present, they can be handled through the use of penalty functions. The transformed function for the MPP takes the form:

$$PB(\tilde{x}, r_k, \tilde{u}^{(k-1)}) = f(\tilde{x}) + r_k^{-1} \sum_{i \in I} h_i(\tilde{x})^2 - r_k \sum_{i \in J} u_i^{(k-1)} \log[g_i(\tilde{x})]. \quad (5.1-2)$$

Properties of $PB(\tilde{x}, r_k, \tilde{u}^{(k-1)})$ can be readily established by combining the results of §3.3 and §4.1. It is re-emphasized that this method is intended for maintaining feasibility with respect to inequality constraints.

Algorithm 5.1.1 (for an ICP)

Assume that an interior starting point $\tilde{x}^{(0)}$ is given, and that an unconstrained minimization routine which does not require derivative information is available. Then, with respect to a given accuracy tolerance ϵ , the algorithm proceeds as follows:

1) initialize:

1.1) set $\bar{\epsilon} := \epsilon/10$, $\tilde{u}^{(0)} := (1, 1, \dots, 1)$ and $r_1 := 10^{-2}$;

1.2) find $\tilde{x}^{(1)}$ which minimizes $B(\tilde{x}, r_1 \tilde{u}^{(0)})$ from the starting point $\tilde{x}^{(0)}$ with a tolerance of $\bar{\epsilon}$;
set $u_i^{(1)} := r_1 u_i^{(0)} / g_i(\tilde{x}^{(1)})$, $\forall i$;

1.3) set $\tilde{u}^{(0)} := \tilde{u}^{(1)}$, $\tilde{x}^{(0)} := \tilde{x}^{(1)}$, $r_1 := 10^{-2} / \|\tilde{u}^{(0)}\|_\infty$ and $k := 1$;

2) find $\tilde{x}^{(k)}$ which minimizes $B(\tilde{x}, r_k \tilde{u}^{(k-1)})$ from the starting point $\tilde{x}^{(0)}$ with a tolerance of $\bar{\epsilon}$, and set $u_i^{(k)} := r_k u_i^{(k-1)} / g_i(\tilde{x}^{(k)})$, $\forall i$;

- 3) apply extrapolation of order $k-1$ to find the (best) solution estimate $\tilde{x}_{\min}^{(k)}$;
- 4) apply stopping criteria:
 - 4.1) if $\max_{i: |g_i(\tilde{x}_{\min}^{(k)})| \leq 10r_k} |g_i(\tilde{x}_{\min}^{(k)})| \leq \epsilon$, print the following message 'the required accuracy is achieved based on the difference from zero of the active constraints', and stop the algorithm;
 - 4.2) if $\|\tilde{x}_{\min}^{(k)} - \tilde{x}_{\min}^{(k-1)}\| \leq \epsilon$, print the following message 'the required accuracy is achieved based on the difference between successive solution estimates', and stop the algorithm;
 - 4.3) if $k > 11$, print the following message 'the required accuracy is not achieved after the 11th iteration', and stop the algorithm;
- 5) find an estimate for $\tilde{x}^{(k+1)}$ by extrapolation, set $\tilde{x}^{(0)} := \tilde{x}^{(k+1)}$ if it is an interior point, otherwise set $\tilde{x}^{(0)} := \tilde{x}^{(k)}$;
- 6) set $r_{k+1} := r_k/2$ and $k := k+1$, and return to (2).

Below, some characteristic and optional features of Algorithm 5.1.1 are further discussed.

- (a) It is assumed that an interior starting point $\tilde{x}^{(0)}$ is available. The problem of finding an interior point may be more difficult than the optimization problem itself.
- (b) The choice $r_1 = 10^{-2}$ is arbitrary, and it is possible that, starting from $\tilde{x}^{(0)}$, the unconstrained routine may fail to converge to a bounded minimum. In this situation, a smaller value of r_1 (e.g., $r_1 = 10^{-3}$ or

10^{-4}) should be used. Alternatively, some form of approximation can be used to obtain estimates for $\nabla f(\tilde{x}^{(0)})$ and $\{\nabla g_i(\tilde{x}^{(0)})\}$, and the parameter r_1 is chosen such that

$$\|\nabla f(\tilde{x}^{(0)})\| > r_1 \left\| \sum_{i \in J} \frac{1}{g_i(\tilde{x}^{(0)})} \nabla g_i(\tilde{x}^{(0)}) \right\|. \quad (5.1-3)$$

(c) The stopping criterion (4.1), which is based on the difference from zero of the active constraints, is the stopping criterion which is usually found in many constrained optimization algorithms. However, the stopping criterion (4.2), which is based on the difference between successive solution estimates, seems to be more appropriate in situations where the problem is poorly scaled (e.g., $\|\nabla g_i\|$, for some i , is relatively large), and where the optimal solution is in the interior of the feasible region. The stopping criterion (4.3) is also included, because, for most problems, the required accuracy should have been achieved before the 11th iteration. Additional iterations beyond the 11th ($r < 10^{-5}$) will tend to encounter difficulties associated with the ill-conditioning of the unconstrained subproblem.

(d) The algorithm could fail in situations where $\|u^{(k)}\|_\infty$ becomes unbounded. Such situations could occur when solving problems which do not satisfy the Kuhn-Tucker condition. When this happens the algorithm should be switched from the modified form of the barrier function method to the classical form.

(e) By using the transformed function $PB(\tilde{x}, r_k, u^{(k-1)})$ instead of $B(\tilde{x}, r_k u^{(k-1)})$, Algorithm 5.1.1 can be used to solve mixed-constrained problems for which it is necessary to maintain feasibility with respect to inequality constraints throughout the computation. In such situations, it is assumed that a starting point $\tilde{x}^{(0)}$, which is interior with respect

to inequality constraints, is available. It is also necessary that the term $\max_{i \in I} |h_i(x_{\min}^{(k)})|$ be included in the stopping criterion (4.1).

An Accelerated Powell/Hestenes Method

It follows from the results of Chapter 4 that, when an ECP is solved without using derivative information, the best augmented Lagrangian-type method is the accelerated Powell/Hestenes method. The transformed (or augmented Lagrangian) function takes the form:

$$L_{\sigma}(\underline{x}, \underline{v}) = f(\underline{x}) - \underline{v}^T \underline{h}(\underline{x}) + \frac{1}{2} \underline{h}(\underline{x})^T [\sigma] \underline{h}(\underline{x}), \quad (5.1-4)$$

where $[\sigma] = [\text{diag } \sigma_i, i \in I]$. By increasing σ while keeping $\underline{v}$ fixed, the method can be interpreted as a penalty function method. Consequently, desirable properties, such as the robust convergence property of penalty function methods, will hold for the proposed accelerated Powell/Hestenes method. Inequality constraints can be handled by (implicitly) solving the extended problem, using the results and techniques of subsection 4.2.3.

Algorithm 5.1.2 (for an ECP)

Assume that a suitable unconstrained minimization routine which does not require derivative information is available, and that a starting point $\underline{x}^{(0)}$ is given (otherwise set $\underline{x}^{(0)} := 0$). Then, with respect to a given accuracy tolerance ϵ , the algorithm proceeds as follows:

1) initialize:

1.1) set $\bar{\epsilon} := \epsilon/10$, $k := 1$, $\underline{v}^{(1)} := 0$ and $\sigma_i^{(1)} := 10^2, \forall i$;

1.2) find $\underline{x}^{(1)}$ which minimizes $L_{\sigma^{(1)}}(\underline{x}, \underline{v}^{(1)})$ from the starting point $\underline{x}^{(0)}$ with a tolerance of $\bar{\epsilon}$;

set $\underline{\Lambda}^{(1)} := \underline{v}^{(1)} - [\sigma^{(1)}] \underline{h}^{(1)}$, $B := \|\underline{h}^{(1)}\|_{\infty}$, $\underline{v}^{(2)} := \underline{\Lambda}^{(1)}$,

$$[\sigma^{(2)}] := [\sigma^{(1)}] \text{ and } k := k+1 ;$$

- 2) find $\tilde{x}^{(k)}$ which minimizes $L_{\sigma^{(k)}}(\tilde{x}, \tilde{v}^{(k)})$ from the starting point $\tilde{x}^{(k-1)}$ with a tolerance of $\frac{\epsilon}{2}$;
 set $\tilde{\Lambda}^{(k)} := \tilde{v}^{(k)} - [\sigma^{(k)}]_{\tilde{h}}^{(k)}$ and $K := \|\tilde{h}^{(k)}\|_{\infty}$;
 if $K \geq B$, go to (5);

- 3) apply stopping criteria:

- 3.1) if $K \leq \epsilon$, print the following message 'the required accuracy is achieved based on the difference from zero of the active constraints', and stop the algorithm;
- 3.2) if $\|\tilde{x}^{(k)} - \tilde{x}^{(k-1)}\| \leq \epsilon$, print the following message 'the required accuracy is achieved based on the difference between successive solution estimates', and stop the algorithm;
- 3.3) if $k > 15$, print the following message 'the required accuracy is not achieved after the 15th iteration', and stop the algorithm;

4) set

$$w^{(k)} := \begin{cases} 0 & \text{if } (\tilde{h}^{(k)} - \tilde{h}^{(k-1)})^T \cdot (\tilde{\Lambda}^{(k)} - \tilde{\Lambda}^{(k-1)}) = 0, \\ \max \left\{ -0.4, \min \left\{ -\frac{\tilde{h}^{(k)T} \cdot (\tilde{\Lambda}^{(k)} - \tilde{\Lambda}^{(k-1)})}{(\tilde{h}^{(k)} - \tilde{h}^{(k-1)})^T \cdot (\tilde{\Lambda}^{(k)} - \tilde{\Lambda}^{(k-1)})}, 0.3 \right\} \right\} & \text{otherwise,} \end{cases}$$

$$\tilde{v}^{(k+1)} := \tilde{\Lambda}^{(k)} + w^{(k)} (\tilde{\Lambda}^{(k)} - \tilde{\Lambda}^{(k-1)}),$$

$$\sigma_{\tilde{\lambda}}^{(k+1)} := \begin{cases} \sigma_{\tilde{\lambda}}^{(k)} & \text{if } |h_{\tilde{\lambda}}^{(k)}| < B/4 \\ 10\sigma_{\tilde{\lambda}}^{(k)} & \text{otherwise} \end{cases}, \quad \forall \tilde{\lambda},$$

$$B := K \text{ and } k := k+1, \text{ and return to (2).}$$

- 5) set $\underline{v}^{(k+1)} := \underline{v}^{(k)}$, $[\sigma^{(k+1)}] := 10[\sigma^{(k)}]$ and $k := k+1$, and return to (2).

Below, some characteristics and optional features of Algorithm 5.1.2 are further discussed.

(a) Similar to Algorithm 5.1.1, the success of Algorithm 5.1.2 depends strongly on the robustness and reliability of the unconstrained routine. Appropriate check should be carried out to ensure that the solution obtained for the unconstrained subproblem is meaningful. The choice $\sigma_i^{(1)} = 10^2$, $\forall i$, is arbitrary. If the unconstrained routine fails to converge to a bounded minimum, a larger value of $[\sigma^{(1)}]$ (e.g., $\sigma_i^{(1)} = 10^3$ or 10^4 , $\forall i$) should be used. Alternatively, some simple transformation of the objective function (e.g., $f(\underline{x}) := \exp(f(\underline{x}))$), which guarantees that the unconstrained subproblem has no unbounded minima, can be applied.

(b) The parameters $\underline{v}$ and $[\sigma]$ are allowed to vary independently, and both can vary in the same iteration. Clearly, this strategy is more efficient than a strategy which only allows either one of the parameters to vary at a time. However, if no improvement is obtained, this is an indication that $[\sigma]$ is too small, and hence only $[\sigma]$ should be updated (increased) for the next iteration (as in step (5)).

(c) Inequality constraints can be handled by using the techniques of subsection 4.2.3. In this situation, feasibility with respect to inequality constraints is no longer maintained as in Algorithm 5.1.1.

(d) Similar to Algorithm 5.1.1, the stopping criterion (3.1) is the one usually found in such algorithms. However, the stopping criterion (3.2) is more appropriate for problems which are poorly scaled (e.g., $\|\nabla h_i\|$ is large for some i). As previously, the stopping criterion (3.3) is included because, for most problems, the required accuracy should have been achieved

before the 15th iteration.

(e) The algorithm could fail in situations where $\|\Lambda^{(k)}\|_{\infty}$ becomes unbounded. If this happens, the algorithm should be switched over to the quadratic penalty function method. Provided that the unconstrained routine is sufficiently robust, a reasonable estimate of the solution can be expected before the algorithm breaks down due to ill-conditioning.

Both the penalty function-modified barrier function method and the accelerated Powell/Hestenes method are simple, efficient and robust: (i) they are simple to program because the underlying barrier function and Powell/Hestenes methods are simple and straightforward; (ii) if certain conditions, such as the regular local solution condition, are satisfied by the MPP, both compete favourably with other possible methods of a similar structure (nature); (iii) both retain the barrier function and penalty function approaches, and hence, a reasonable estimate of the solution can be expected even when the Kuhn-Tucker condition is not satisfied.

Because of their simplicity, efficiency and robustness, these two methods can be competitive even when derivative information is readily available. In these situations, the derivative information can be utilized by using the more efficient gradient techniques to solve the unconstrained subproblem.

5.2 OPTIMAL CONTROL PROBLEMS WITH TERMINAL CONSTRAINTS

The following class of optimal control problems is considered:

$$\text{minimize } \Phi = \phi(x(1)) + \int_0^1 L(\underline{x}, \underline{y}, t) dt, \quad (5.2-1)$$

$$\text{subject to } \dot{\underline{x}}(t) = F(\underline{x}, \underline{y}, t), \quad \underline{x}(0) = \underline{x}_0, \quad 0 \leq t \leq 1, \quad (5.2-2)$$

$$\text{and } \underline{\psi}(\underline{x}(1)) = 0, \quad (5.2-3)$$

where $\underline{x}_0$ is specified, Φ is a scalar objective function, $\underline{\psi}$ is an l -vector of terminal constraint relations, $\underline{x}$ is an n -vector of state variables, $\underline{y}$ is a p -vector of control variables, and t is the independent variable. Relative minima are sought, and all conditions for a well-posed problem are assumed. In particular, it is assumed that, given a control $\underline{y}(t)$, the state equation (5.2-2) can be solved for a unique $\underline{x} = \underline{x}(\underline{y})$, and thus $\Phi = \Phi(\underline{y})$ is a function of $\underline{y}$ alone. Hence the problem becomes

$$\begin{aligned} &\text{minimize}_{\underline{y}(t)} \quad \Phi(\underline{y}) \\ &\text{subject to} \quad \underline{\psi}(\underline{y}) = 0. \end{aligned} \quad (5.2-4)$$

Note that $\underline{y}$ is a p -vector of real-valued functions on $[0,1]$, not a p -vector of real variables. From (5.2-4), it is clear that the augmented Lagrangian function technique can be used to solve the original constrained problem via a sequence of unconstrained subproblems. A numerical study of this approach, using the Powell/Hestenes update for the multipliers, has been reported by Doherty and Pierson (1974). The reported numerical performance is very encouraging. Consider the following unconstrained problem:

$$\underset{y(t)}{\text{minimize}} \quad K(y, v, \sigma) = \phi(y) - \sum_{i=1}^{\ell} v_i \psi_i(y) + \frac{1}{2} \sum_{i=1}^{\ell} \sigma_i \psi_i(y)^2. \quad (5.2-5)$$

The gradient function of $K(y, v, \sigma)$ with respect to y has the form:

$$\frac{\partial K(t)}{\partial y} = \frac{\partial L(t)}{\partial y} + \frac{\partial F(t)^T}{\partial y} \cdot \lambda(t), \quad (5.2-6)$$

where $\lambda(t)$ is an adjoint n -vector of real-valued functions satisfying

$$\left. \begin{aligned} \dot{\lambda}(t) &= - \frac{\partial F(t)^T}{\partial x} \cdot \lambda(t) - \frac{\partial L(t)}{\partial x}, \\ \lambda(1) &= \frac{\partial}{\partial x} \left[\phi(x(t)) - \sum_{i=1}^{\ell} v_i \psi_i(x(t)) + \frac{1}{2} \sum_{i=1}^{\ell} \sigma_i \psi_i(x(t))^2 \right]_{t=1} \end{aligned} \right\} \quad (5.2-7)$$

with

The gradient $\frac{\partial K}{\partial y}$ can therefore be computed by first solving (5.2-2) and then (5.2-7). With the gradient $\frac{\partial K}{\partial y}$ available, many methods can be used to solve the unconstrained subproblem (5.2-5); see, for example, Oi, *et.al.* (1973), Klessig and Polak (1973), Lasdon (1970) and Tripathi and Narendra (1970).

Note that, in any numerical method for solving the unconstrained subproblem (5.2-5), some form of discretization of the time interval $[0,1]$ is necessary. That is, for $0 = t_1 < t_2 < \dots < t_N = 1$, any function $A(t)$ is represented by the N -tuples $(A(t_1), A(t_2), \dots, A(t_N))$. Therefore, the unconstrained control problem can be viewed as an unconstrained minimization problem in $p \times N$ dimensions. Because of the size of the problem, it is not practical to use the more efficient methods (e.g., quasi-Newton methods) which require a $(p \times N) \times (p \times N)$ hessian matrix to be stored and updated. Simpler methods, such as conjugate gradient methods, are usually preferred. Thus, due to the size of the problem, the Newton update (4.2-13) for the multipliers is not suitable. Because of its

simplicity and performance, it is clear that the accelerated Powell/Hestenes method is most suitable for such problems. Hence, the Algorithm 5.1.2 can be used to solve optimal control problems with terminal constraints by using either a suitable routine for solving unconstrained control problems, or a routine for solving large-scale unconstrained minimization problems.

5.3 CONCLUSION

Some of the major implications of the results presented in this thesis are:

Strict Complementarity / Non-degeneracy

The strict complementarity requirement has been effectively removed from many results, including sensitivity analysis. It has been shown that degeneracy does not effect the rate of convergence of certain methods including the quadratic loss penalty function and various augmented Lagrangian methods. In addition, it has been shown that a (modified) barrier function method can be constructed so that its rate of convergence is the same for both degenerate and non-degenerate problems. The resulting implication is that strict complementarity should not and need not be imposed on the MPP. A 'good' method could be and should be more or less independent of degeneracy considerations. Any numerical difficulties, which arise from degeneracy, should be regarded as shortcomings of the method under consideration.

On the other hand, degeneracy and near-degeneracy have to be taken into consideration when designing an algorithm. It has been shown how degeneracy and near-degeneracy can affect the convergence of barrier function methods. It has also been shown that, when solving a degenerate problem, some methods are effectively switching between solving several equality constrained problems each of which has the same solution x^* as the original problem. Near the solution of a degenerate (or near-degenerate) problem, the 'active' constraint set determined by using some criterion can be extremely sensitive to perturbations, and this could result in a 'failure' of the method under consideration. Difficulties of this nature have been reported by Lill (1972). This type of 'failure' can also occur when using methods which do not explicitly determine the correct 'active' set (sometimes referred to as the 'basis'), but which employ a quadratic

programming subroutine for solving intermediate subproblems. When solving a degenerate (or near-degenerate) problem, 'failure' may occur because the 'active' set identified by the quadratic programming subroutine is extremely sensitive to perturbations. Techniques and results presented in this thesis may not be exhaustive, but they should at least give some insight into these difficulties associated with degeneracy and their possible remedies.

It is also appropriate to point out that the modified barrier function method of §3.3 is just one possible technique for rectifying the shortcomings of classical barrier function methods. Other types of modifications are certainly worthy of further investigation.

Necessary and Sufficient Conditions for a Unique Solution of 1st or 2nd Order.

Such results have not been fully exploited. Their significance is, however, best described in terms of the title of Cromme's (1978) paper: 'Strong uniqueness. A far-reaching criterion for the convergence analysis of iterative procedures'.

The 'Extended' Problem

Any method for solving equality constrained problems can be generalized to include inequality constraints by (implicitly) solving the extended problem. This technique obviates the need for *ad hoc* modifications associated with each new method for solving equality constrained problems. Results on the structure of the extended problem also give insight into how degeneracy may affect the performance of the resulting algorithm.

The Accelerated Powell/Hestenes Method

Except in the convex case, the proposed accelerating device is based partly on heuristics. Its numerical performance has proved to be excellent. The mechanism which is responsible for such numerical performance

is not fully understood. Because of the practical importance of these techniques for updating multipliers without derivative information, further investigation will certainly be worth-while.

A Stronger Version of the Implicit Function Theorem

The version of the implicit function theorem given in Appendix 2 yields a necessary and sufficient condition. Its usefulness has been clearly demonstrated in this thesis. However, its true impact on mathematical analysis will lie not in strengthening existing results which depend on standard implicit function theorems, but in producing new results which were previously inaccessible.

APPENDIXES

APPENDIX 1 - CONVEX FUNCTIONALS

Definition

A functional $F: \mathcal{D} \subset E^n \rightarrow E^1$ is *convex* on a convex set $\mathcal{D}_0 \subseteq \mathcal{D}$ if, for all $x_1, x_2 \in \mathcal{D}_0$ and $0 < \alpha < 1$,

$$F(\alpha x_1 + (1-\alpha)x_2) \leq \alpha F(x_1) + (1-\alpha)F(x_2) \quad . \quad (A1-1)$$

The functional F is *strictly convex* on $\mathcal{D}_0$ if strict inequality holds in (A1-1) whenever $x_1 \neq x_2$.

Definition

A functional $F: \mathcal{D} \subset E^n \rightarrow E^1$ is (strictly) *concave* on a convex set $\mathcal{D}_0 \subseteq \mathcal{D}$ if and only if $-F$ is (strictly) convex on $\mathcal{D}_0$.

Suppose that the functional $F: \mathcal{D} \subset E^n \rightarrow E^1$ is (Gateaux) differentiable on a convex set $\mathcal{D}_0 \subseteq \mathcal{D}$. Then, F is convex on $\mathcal{D}_0$ if and only if

$$[F'(x_1) - F'(x_2)] \cdot (x_1 - x_2) \geq 0, \quad \forall x_1, x_2 \in \mathcal{D}_0, \quad (A1-2)$$

and F is strictly convex on $\mathcal{D}_0$ if and only if strict inequality holds in (A1-2) (for a proof of this result, see 3.4.5 of Ortega and Rheinboldt (1970)). Note that this result implies that, for $F': \mathcal{D}_0 \subset E^n \rightarrow E^n$, F' is a *one-to-one* mapping on $\mathcal{D}_0$ if and only if F is strictly convex on $\mathcal{D}_0$.

APPENDIX 2 - AN IMPLICIT FUNCTION THEOREM

In nonlinear programming, several conditions that are imposed on the problem can be identified as the appropriate conditions for some implicit function theorem. The classical theorem for continuously differentiable functions and its proof can be found in most advanced books.

For example:-

THEOREM (Ortega and Rheinboldt (1970), p.128)

Suppose that $F: \mathcal{D} \subset E^n \times E^m \rightarrow E^n$ is continuous on an open neighbourhood $\mathcal{D}_0 \subset \mathcal{D}$ of a point (x^0, y^0) for which $F(x^0, y^0) = 0$. Assume that $\partial_1 F$ exists and is strong at (x^0, y^0) , and that $\partial_1 F(x^0, y^0)$ is non-singular. Then there exist open neighbourhoods $A_0 \subset E^n$ and $B_0 \subset E^m$ of x^0 and y^0 , respectively, such that for any $y \in B_0$, the equation $F(x, y) = 0$ has a unique solution $x = Hy \in A_0$ and the mapping $H: A_0 \rightarrow E^n$ is continuous. Moreover, if $\partial_2 F$ exists at (x^0, y^0) , then H is differentiable at y^0 and

$$H'(y^0) = -[\partial_1 F(x^0, y^0)]^{-1} \partial_2 F(x^0, y^0) \quad \square$$

However, if we are only interested in $x(y)$ as a *continuous* function of y such that $F(x(y), y) = 0$. Both the non-singularity and differentiability requirements are obviously too strong. It is therefore of interest to find a theorem with the weakest requirement. The theorem that we will attempt to prove here is such a theorem.

IMPLICIT FUNCTION THEOREM

Suppose that $F: \mathcal{D} \subset E^n \times E^m \rightarrow E^n$ is a continuous mapping with $F(x^0, y^0) = 0$. Assume that there exist open neighbourhoods $A \subset E^n$ and $B \subset E^m$ of x^0 and y^0 , respectively, such that $\forall y \in B$,

$F(.,y) : A \subset E^n \rightarrow E^n$ is locally one-to-one. Then there exist open neighbourhoods $A_0 \subset A$ and $B_0 \subset B$ of x^0 and y^0 , respectively such that $\forall y \in B_0$, the equation $F(x,y) = 0$ has a unique solution $x = Hy \in A_0$ and the mapping $H : A_0 \rightarrow E^n$ is continuous.

To prove the above theorem, we first prove the following Lemma.

LEMMA (*Invariance of Domain*)

Let $U \subset E^n$ be open connected, if $F : U \rightarrow E^n$ is a one-to-one continuous map then $F(U)$ is open connected and F is a homeomorphism onto $F(U)$.

Proof

For $n=1$, F is one-to-one and continuous on U means that F is strictly monotone on U and hence the result is obvious.

For $n \geq 2$, this is a direct consequence of the Jordan-Brouwer separation theorem (see Corollary (18.9) of Greenberg (1971) for proof). $\square$

We are now ready to prove the main theorem.

Proof

Let's define

$$\Psi : A \times B \subset E^n \times E^m \rightarrow E^n \times E^m$$

$$\Psi : (x,y) \mapsto (F(x,y), y) \quad .$$

It is obvious from the definition that Ψ is one-to-one and continuous.

Thus, by the above Lemma, $\Psi(A \times B)$ is open and there is an inverse

$$\Phi : \Psi(A \times B) \rightarrow A \times B$$

$$\Phi : (\xi, \eta) \mapsto (\phi_1(\xi, \eta), \phi_2(\xi, \eta))$$

where $\phi_1(\xi, \eta) \in E^n$, $\phi_2(\xi, \eta) \in E^m$.

By local homeomorphism we have

$$I = \Psi\phi$$

hence

$$\begin{aligned} (0, y) &= \Psi\phi(0, y) \\ &= (F(\phi_1(0, y), \phi_2(0, y)), \phi_2(0, y)) \end{aligned}$$

i.e. $\phi_2(0, y) = y$ and $F(\phi_1(0, y), y) = 0$

or $\phi_1(0, y)$ is the required function H . □

REMARKS

1. The above proof is also valid for the more general theorem which we want to satisfy $F(x, y) = z$, $z \in C_0$ an open neighbourhood of 0 . In this case $x = \phi_1(z, y)$ is the required solution.
2. This version of the theorem is strongest in the sense that the one-to-one condition is obviously necessary as well as sufficient.

The one-to-one property of a function is a necessary condition for other properties such as strict monotonicity (5.4.3 of Ortega and Rheinboldt (1970)) and hence should not be too difficult to prove. In fact the original motivation for the new theorem is that: In solving some non-linear programming problems, the function F turns out to be the gradient $\partial_1 L$ of a generalized Lagrangian $L: E^n \times E^m \rightarrow E^1$. For some problems the classical theorem fails to hold because $\partial_{11}^2 L$ may become singular, or in some cases $\partial_{11}^2 L$ may not exist at all. But in using the new theorem, from (3.4-5) of Ortega and Rheinboldt (1970), we need only to show that $L(x, y)$ is strictly convex with respect to x .

APPENDIX 3 - EXTRAPOLATION

Assume that $x(r)$ is a suitably smooth function of r , $r \in [0, r_0]$, and that $\{r_k\}$ is a strictly decreasing null sequence. Then an estimate for $x(0)$ can be obtained by first fitting a polynomial of degree $k-1$ to the data $x(r_1), x(r_2), \dots, x(r_k)$ and then evaluating it at $r = 0$:

$$L_k(x) = \sum_{j=1}^k A_j^k x(r_j) . \quad (A3-1)$$

The coefficients A_j^k are such that the extrapolation is exact if x is a polynomial of degree $k-1$. This gives

$$A_j^k = \prod_{\substack{i=1 \\ i \neq j}}^k \frac{1}{1-r_j/r_i} , \quad j \leq k , \quad (A3-2)$$

and the relation

$$A_i^{k+1} = \frac{1}{1-r_i/r_{k+1}} \cdot A_i^k . \quad (A3-3)$$

Error estimate

A standard argument (see, for example, Conte and de Boor (1972)) gives that

$$L_k(x) - x(0) = \frac{(-1)^{k+1}}{k!} \prod_{j=1}^k r_j \frac{d^k x}{dr^k}(\xi) , \quad (A3-4)$$

where ξ is a mean value.

Stability analysis

Let ϵ_i denote the error associated with the value for $x(r_i)$, and ϵ denote the error bound, $|\epsilon_i| \leq \epsilon$, $\forall i$. Then, from (A3-1), it follows that a bound for the error associated with the computed $L_k(x)$ is

given by

$$\varepsilon \left(\sum_{j=1}^k |A_j^k| \right) . \quad (\text{A3-5})$$

In what follows, it is further assumed that

$$r_{k+1} = r_k/c , \quad (\text{A3-6})$$

where $c > 1$ is a given constant.

Note 1: The study by Laurent (1963) gives that, for $c = 2$,

$$\lim_{k \rightarrow \infty} \sum_{j=1}^k |A_j^k| \leq 8.2 .$$

Note 2: From (A3-2), it follows that

$$|A_j^k| = \left| \frac{c^{k-j+1}-1}{1-c^{-(j-1)}} \right| \cdot |A_{j-1}^k| , \quad 1 < j \leq k . \quad (\text{A3-7})$$

Hence $|A_1^k| < |A_2^k| < \dots < |A_k^k|$, and it is then followed from (A3-5) that the size of the error in the computed $L_k(x)$ depends strongly on $|\varepsilon_k|$.

Computation

(i) The estimate $L_k(x)$: Note that, from (A3-1),

$$L_{k-1}(x) = A_1^k x(r_1) + A_2^k x(r_2) + \dots + A_{k-1}^{k-1} x(r_{k-1}) . \quad (\text{A3-8})$$

Therefore, in order to compute $L_k(x)$, it is sufficient to multiply each term on the RHS of (A3-8), respectively, by

$$\frac{1}{1-c^{k-1}} , \quad \frac{1}{1-c^{k-2}} , \quad \dots , \quad \frac{1}{1-c} , \quad (\text{A3-9})$$

and then add the term $A_k^k x(r_k)$.

(ii) An estimate for $x(r_{k+1})$: Let $\bar{x}(r_{k+1})$ denote the estimate of $x(r_{k+1})$ obtained from evaluating the extrapolating polynomial (of degree $k-1$) at $r = r_{k+1}$. Using the uniqueness property of the extrapolating polynomial, it follows that

$$\bar{x}(r_{k+1}) = \frac{1}{A_{k+1}^{k+1}} \left\{ L_k(x) - \sum_{j=1}^k \frac{1}{1-c^{k+1-j}} \cdot A_j^k x(r_j) \right\}, \quad (\text{A3-10})$$

where $L_k(x)$ and $A_j^k x(r_j)$, $j=1, \dots, k$ are known quantities.

Similar (recursive) techniques for calculating $L_k(x)$ and $\bar{x}(r_{k+1})$ can also be found in Fiacco and McCormick (1968) and Joyce (1971).

APPENDIX 4 - TEST PROBLEMS

[1] Colville: Problem 1 (Colville (1968)) $n = 5, \ell = 0, m = 15$

$$\text{minimize } f(\underline{x}) = \sum_{j=1}^5 e_j x_j + \sum_{j=1}^5 \sum_{i=1}^5 c_{ij} x_i x_j + \sum_{j=1}^5 d_j x_j^3$$

$$\text{subject to } g_i(\underline{x}) = \sum_{j=1}^5 a_{ij} x_j + b_i \geq 0, \quad i = \overline{1, 10},$$

$$\text{and } g_{i+10}(\underline{x}) = x_i \geq 0, \quad i = \overline{1, 15},$$

where a_{ij}, c_{ij}, b_i, d_j and e_j are given in Table A4.1. The third, fifth, sixth and ninth constraints are active. The constrained minimum is at (to 4 decimal places) $\underline{x}^* = (0.3, 0.3375, 0.4, 0.4283, 0.2240)$. An initial feasible point is $\underline{x}^{(0)} = (0.125, 0.0625, 0.125, 0.125, 1.0)$.

[2] Colville: Problem 2 (Colville (1968)) $n = 15, \ell = 0, m = 20$

$$\text{minimize } f(\underline{x}) = \sum_{i=1}^{10} b_i x_{i+5} + \sum_{i=1}^5 \sum_{j=1}^5 c_{ij} x_i x_j + 2 \sum_{i=1}^5 d_i x_i^3$$

$$\text{subject to } g_j(\underline{x}) = e_j + 2 \sum_{i=1}^5 c_{ij} x_i + 3d_j x_j^2 - \sum_{i=1}^{10} a_{ij} x_{i+5} \geq 0, \quad j = \overline{1, 5},$$

$$\text{and } g_{i+5}(\underline{x}) = x_i \geq 0, \quad i = \overline{1, 15},$$

where a_{ij}, c_{ij}, b_i, d_j and e_j are given in Table A4.1. Constraints 1, 2, 3, 4, 5, 11, 12, 14, 17, 18 and 20 are active. The constrained minimum is at (to 4 decimal places) $\underline{x}^* = (0.3, 0.3335, 0.4, 0.4283, 0.2240, 0, 0, 5.1740, 0, 3.0611, 11.8496, 0, 0, 0.1039, 0)$. An initial feasible point is $x_i = 20, i = \overline{1, 5}, x_i = 0.001, i = \overline{6, 15}$. This problem is the dual formulation of Colville: Problem 1.

$i \backslash j$	1	2	3	4	5	
e_j	-15	-27	-36	-18	-12	
c_{ij}	1 2 3 4 5	30 -20 -10 32 -10	-20 39 -6 10 -6	-10 -6 10 -6 39	32 -31 -6 39 -20	-10 32 -10 -20 30
d_j	4	8	10	6	2	b_i
a_{ij}	1 2 3 4 5 6 7 8 9 10	-16 0 -3.5 0 0 2 -1 -1 1 1	2 -2 0 -2 -9 0 -1 -2 2 1	0 0 2 0 -2 -4 -1 -3 3 1	1 4 0 -4 1 0 -1 -2 4 1	0 2 0 -1 -2.8 0 -1 -1 5 1

Table A4.1: Coefficients for Colville: Problems 1 and 2

[3] Colville: Problem 7 (Colville (1968)) $n = 16$, $\ell = 8$, $m = 32$

$$\text{minimize } f(\underline{x}) = \sum_{i=1}^{16} \sum_{j=i}^{16} A_{ij} (x_i^2 + x_i + 1.0) (x_j^2 + x_j + 1.0)$$

$$\text{subject to } \left. \begin{aligned} g_i(\underline{x}) &= x_i \geq 0 \\ g_{i+16}(\underline{x}) &= 5 - x_i \geq 0 \end{aligned} \right\} i = \overline{1, 16},$$

$$h_1(\underline{x}) = 0.22x_1 + 0.2x_2 + 0.19x_3 + 0.25x_4 + 0.15x_5 + 0.11x_6 + 0.12x_7 + 0.13x_8 + x_9 - 2.5 = 0,$$

$$h_2(\underline{x}) = -1.46x_1 - 1.3x_3 + 1.82x_4 - 1.15x_5 + 0.8x_7 + x_{10} - 1.1 = 0,$$

$$h_3(x) = 1.29x_1 - 0.89x_2 - 1.16x_5 - 0.96x_6 - 0.49x_8 + x_{11} + 3.1 = 0,$$

$$h_4(x) = -1.1x_1 - 1.06x_2 + 0.95x_3 - 0.54x_4 - 1.78x_6 - 0.41x_7 + x_{12} + 3.5 = 0,$$

$$h_5(x) = -1.43x_4 + 1.51x_5 + 0.59x_6 - 0.33x_7 - 0.43x_8 + x_{13} - 1.3 = 0,$$

$$h_6(x) = -1.72x_2 - 0.33x_3 + 1.62x_5 + 1.24x_6 + 0.21x_7 - 0.26x_8 + x_{14} - 2.1 = 0,$$

$$h_7(x) = 1.12x_1 + 0.31x_4 + 1.12x_7 - 0.36x_9 + x_{15} - 2.3 = 0,$$

$$h_8(x) = 0.45x_2 + 0.26x_3 - 1.10x_4 + 0.58x_5 - 1.03x_7 + 0.1x_8 + x_{16} + 1.5 = 0,$$

where A_{ij} , $i = \overline{1,16}$, $j = \overline{1,16}$, is given in Table A4.2. For inequality constraints, g_{10} , g_{11} , g_{12} , g_{14} and g_{16} are active. The constraint minimum is at (to 4 decimal places) $\tilde{x}^* = (0.0398, 0.7920, 0.2039, 0.8444, 1.2699, 0.9347, 1.6820, 0.1553, 1.5679, 0, 0, 0, 0.6602, 0, 0.6743, 0)$. A starting point is $\tilde{x}^{(0)} = e$.

$i \backslash j$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1			1			1	1								1
2		1	1				1			1						
3			1				1		1	1				1		
4				1			1				1				1	
5					1	1				1		1				1
6						1		1							1	
7							1				1		1			
8								1		1					1	
9									1			1				1
10										1				1		
11											1		1			
12												1		1		
13													1	1		
14														1		
15															1	
16																1

Table A4.2: Coefficients for Colville: Problem 7

- [4] Miele, et al.: Example 8.1 (Miele, et al. (1972)) $n=5, \ell=3, m=0$

$$\text{minimize } f(\underline{x}) = (x_1 - x_2)^2 + (x_2 + x_3 - 2)^2 + (x_4 - 1)^2 + (x_5 - 1)^2$$

$$\text{subject to } h_1(\underline{x}) = x_1 - 3x_2 = 0,$$

$$h_2(\underline{x}) = x_3 + x_4 - 2x_5 = 0,$$

$$h_3(\underline{x}) = x_2 - x_5 = 0.$$

The constraint minimum is at $\underline{x}^* = (-0.7674, 0.2558, 0.6279, -0.1162, 0.2558)$.

- [5] Miele, et al.: Example 8.4 (Miele, et al. (1972)) $n=5, \ell=3, m=0$

$$\text{minimize } f(\underline{x}) = (x_1 - 1)^2 + (x_1 - x_2)^2 + (x_2 - x_3)^2 + (x_3 - x_4)^4 + (x_4 - x_5)^4$$

$$\text{subject to } h_1(\underline{x}) = x_1 + x_2^2 + x_3^3 - 2 - 3\sqrt{2} = 0,$$

$$h_2(\underline{x}) = x_2 - x_3^2 + x_4 + 2 - 2\sqrt{2} = 0,$$

$$h_3(\underline{x}) = x_1 x_5 - 2 = 0.$$

The constrained minimum is at $\underline{x}^* = (1.1911, 1.3626, 1.4728, 1.6350, 1.6790)$.

- [6] Powell's (1969) problem $n = 5, \ell = 3, m = 0$

$$\text{minimize } f(\underline{x}) = \exp(x_1 x_2 x_3 x_4 x_5)$$

$$\text{subject to } h_1(\underline{x}) = x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 - 10 = 0,$$

$$h_2(\underline{x}) = x_2 x_3 - 5 x_4 x_5 = 0,$$

$$h_3(\underline{x}) = x_1^3 + x_2^3 + 1 = 0.$$

The constrained minimum is at $\underline{x}^* = (-1.7172, 1.5957, 1.8272, -0.7636, -0.7636)$.

- [7] The Rosen-Suzuki problem (Rosen and Suzuki (1965)) $n=4, \ell=0, m=3$

$$\begin{aligned} \text{minimize} \quad f(\underline{x}) &= x_1^2 + x_2^2 + 2x_3^2 + x_4^2 - 5x_1 - 5x_2 - 21x_3 + 7x_4 \\ \text{subject to} \quad g_1(\underline{x}) &= -x_1^2 - x_2^2 - x_3^2 - x_4^2 - x_1 + x_2 - x_3 + x_4 + 8 \geq 0, \\ g_2(\underline{x}) &= -x_1^2 - 2x_2^2 - x_3^2 - 2x_4^2 + x_1 + x_4 + 10 \geq 0, \\ g_3(\underline{x}) &= -2x_1^2 - x_2^2 - x_3^2 - 2x_4^2 + x_1 + x_2 + x_4 + 5 \geq 0. \end{aligned}$$

Constraints 1 and 3 are active. The constrained minimum is at $\underline{x}^* = (0, 1, 2, -1)$. A feasible starting point is $\underline{x}^{(0)} = 0$.

- [8] The modified Rosen-Suzuki problem $n=4, \ell=0, m=3$

This problem is obtained by subtracting 1 from the 2nd constraint of the Rosen-Suzuki problem. The (modified) 2nd constraint

$$g_2(\underline{x}) = -x_1^2 - 2x_2^2 - x_3^2 - 2x_4^2 + x_1 + x_4 + 9 \geq 0$$

is degenerate at the solution.

- [9] The Rosen-Kreuser problem $n=15, \ell=0, m=10$

$$\begin{aligned} \text{minimize} \quad f(\underline{x}) &= - \sum_{i=1}^{15} c_i x_i \\ \text{subject to} \quad g_i(\underline{x}) &= b_i - \sum_{j=1}^{15} a_{ij} x_j^2, \quad i = \overline{1, 10}, \end{aligned}$$

where the coefficients a_{ij} , b_i and c_i are given in Table A4.3. All constraints are active and the first nine constraints are degenerate. The constrained minimum is at $\underline{x}^* = (1, 1, \dots, 1)$.

$i \backslash j$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
c_j	20	40	400	20	80	20	40	140	380	280	80	40	140	40	120
a_{ij}	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	100	100	10	5	10	0	0	25	0	10	55	5	45	20	0
2	90	100	10	35	20	5	0	35	55	25	20	0	40	25	10
3	70	50	0	55	25	100	40	50	0	30	60	10	30	0	40
4	50	0	0	65	35	100	35	60	0	15	0	75	35	30	65
5	50	10	70	60	45	45	0	35	65	5	75	100	75	10	0
6	40	0	50	95	50	35	10	60	0	45	15	20	0	5	5
7	30	60	30	90	0	30	5	25	0	70	20	25	70	15	15
8	20	30	40	25	40	25	15	10	80	20	30	30	5	65	20
9	10	70	10	35	25	65	0	30	0	0	25	0	15	50	55
10	5	10	100	5	20	5	10	35	95	70	20	10	35	10	30
b_i	385	470	560	565	645	430	485	455	390	460					

Table A4.3: Coefficients for the Rosen-Kreuser problem

REFERENCES

- Arrow, K.J. and Solow, R.M. (1958) Gradient methods for constrained maxima with weakened assumptions; in 'Studies in linear and non-linear programming', eds., K.J. Arrow, L. Hurwicz and H. Uzawa. Stanford Uni. Press, Stanford Calif.
- Arrow, K.J., Gould, F.J. and Howe, S.M. (1973). A general saddle point result for constrained optimization. Math. Programming, 5, 225-234.
- Arrow, K.J., Hurwicz, L. and Uzawa, H. (1961). Constraint qualifications in maximization problems. Nav. Res. Log. Q, 8, 175-191.
- Bellman, R.E. (1970). 'Methods in non-linear analysis'. VI, New York, Academic Press.
- Bertsekas, D.P. (1975). Necessary and sufficient conditions for a penalty method to be exact. Math. Programming, 9, 81-99.
- Bertsekas, D.P. (1975b). Combined primal-dual and penalty methods for constrained minimization. SIAM J. Control, 13, 521-544.
- Bertsekas, D.P. (1976). On penalty and multiplier methods for constrained minimization. SIAM J. Control, 14, 217-235.
- Bigelow, J.H. and Shapiro, N.E. (1974). Implicit function theorems for mathematical programming and for systems of inequalities. Math. Programming, 6, 141-156.
- Biggs, M.C. (1972). Constrained minimization using recursive equality quadratic programming; in 'Numerical methods for non-linear optimization', ed., F.A. Lootsma, Academic Press, London.
- Biggs, M.C. (1975). Constrained minimization using recursive quadratic programming: some alternative subproblem formulations; in 'Towards global optimization', eds., L.C.W. Dixon and G.P. Szegö, North-Holland Publishing Co. (Amsterdam).
- Biggs, M.C. (1978). On the convergence of some constrained minimization algorithms based on recursive quadratic programming. J. Inst.

- Maths. Applics., 21, 67-81.
- Brent, R.P. (1973). 'Algorithms for minimization without derivatives'.
Prentice Hall.
- Buys, J.D. (1972). Dual algorithms for constrained optimization problems.
Ph.D. Thesis, University of Leiden.
- Colville, A.R. (1968). A comparative study on non-linear programming
codes. IBM, New York Scientific Centre Tech. Report No. 320-2949.
- Conn, A.R. (1973). Constrained optimization using a non-differentiable
penalty function. SIAM J. Numer. Anal., 10, 760-784.
- Conte, S.D. and de Boor, C. (1972). 'Elementary numerical analysis',
McGraw-Hill.
- Copson, E.T. (1965). 'Asymptotic expansions'. Cambridge tracts in
mathematics and mathematical physics No. 55.
- Cromme, L. (1978). Strong uniqueness. A far reaching criterion for the
convergence analysis of iterative procedures. Numer. Math.,
29, 179-194.
- Dantzig, G.B. (1951). Maximization of a linear function of variables
subject to linear inequalities; in 'Activity analysis of produc-
tion and allocation', ed., T.C. Koopmans. Cowles Commission
Monograph No. 13, Wiley.
- Dennis, J.E., Jr. and Moré, J.J. (1977). Quasi-Newton methods, motivation
and theory. SIAM Review, 19, 46-89.
- Dixon, L.C.W. (1972). The choice of step length, a crucial factor in the
performance of variable metric algorithms; in 'Numerical methods
in non-linear optimization', ed., F.A. Lootsma, Academic Press,
London and New York.
- Fletcher, R. (1970a). A new approach to variable metric algorithm.
Comp. J., 13, 317-322.
- Fletcher, R. (1970b). A class of methods for non-linear programming with
termination and convergence properties; in 'Integer and non-

- linear programming', ed., J. Abadie, North-Holland, Amsterdam.
- Fletcher, R. (1973). An exact penalty function for non-linear programming with inequalities. *Math. Programming*, 5, 129-150.
- Fletcher, R. (1975). An ideal penalty function for constrained optimization. *J. Inst. Maths. Applics.*, 15, 319-342.
- Fletcher, R. and McCann, A.P. (1969). Acceleration techniques for non-linear programming; in 'Optimization', ed., R. Fletcher, Academic Press, London and New York.
- Fiacco, A.V. (1967). Sequential unconstrained minimization methods for non-linear programming. Ph.D. Thesis, North Western University, Evanston, Illinois.
- Fiacco, A.V. (1976). Sensitivity analysis for non-linear programming using penalty methods. *Math. Programming*, 10, 287-311.
- Fiacco, A.V. and McCormick, G.P. (1967). The sequential unconstrained minimization technique without parameters. *Opn. Res.*, 15, 820-827.
- Fiacco, A.V. and McCormick, G.P. (1968). 'Non-linear programming: Sequential unconstrained minimization techniques'. Wiley.
- Gill, P.E. and Murray, W. (1971). Quasi-Newton methods for unconstrained optimization. *Natn. Phys. Lab. Math.* 97.
- Gould, F. and Tolle, J.W. (1971). A necessary and sufficient constraint qualification for constrained optimization. *SIAM J. Appl. Math.*, 20, 164-172.
- Greenberg, M. (1971). 'Lectures on algebraic topology'. W.A. Benjamin, Menlo Park, California.
- Haarhoff, P.C. and Buys, J.D. (1970). A new method for the optimization of a non-linear function subject to non-linear constraints. *Comp. J.*, 13, 178-184.
- Han, S.P. (1975). A globally convergent method for non-linear programming. Dept. of Comp. Sci., Cornell Univ., Report No. 75-257.
- Han, S.P. (1976). Superlinearly convergent variable metric algorithms

- for general non-linear programming problems. *Math. Programming*, 11, 263-282.
- Han, S.P. (1977). Dual variable-metric algorithms for constrained optimization. *SIAM J. Control*, 15, 546-565.
- Hartman, P. (1973). 'Ordinary differential equation'. Baltimore, Wiley.
- Hestenes, M.R. (1966). 'Calculus of variations and optimal control theory'. Wiley.
- Hestenes, M.R. (1969). Multiplier and gradient methods. *J. Opt. Theory Appl.*, 4, 303-320.
- Householder, A.S. (1974). 'The theory of matrices in numerical analysis'. New York, Blaisdell.
- Huard, P. (1967). Resolution of mathematical programming with non-linear constraints by the method of centres; in 'Non-linear programming', ed., J. Abadie, North-Holland, Amsterdam.
- Jittorntrum, K. (1977). An implicit function theorem. *Notices Amer. Math. Soc.* 24, No. 7, A-615.
- Jittorntrum, K. (1978a). An implicit function theorem. *J. Opt. Theory Appl.*, 25, No. 4.
- Jittorntrum, K. (1978b). Accelerated convergence for the Powell/Hestenes multiplier method. *Math. Programming* - to be published.
- Jittorntrum, K. and Osborne, M.R. (1978). Trajectory analysis and extrapolation in barrier function methods. *J. Austral. Math. Soc. (series B)*, 21, 1-18.
- John, F. (1948). Extremum problems with inequalities as subsidiary conditions; in 'Studies and essays'. Courant anniversary volume, Interscience.
- Joyce, D.C. (1971). Survey of extrapolation processes in numerical analysis. *SIAM Review*, 13, 435-490.
- Klessig, R. and Polak, E. (1973). An adaptive precision gradient method for optimal control. *SIAM J. Control*, 11, 80-93.
- Kort, B.W. and Bertsekas, D.P. (1976). Combined primal-dual and penalty

- methods for convex programming. SIAM J. Control, 14, 268-294.
- Kowalik, J., Osborne, M.R. and Ryan, D.M. (1969). A new method for constrained optimization problems. Opn. Res., 17, 973-983.
- Kuhn, H.W. and Tucker, A.W. (1951). Non-linear programming; in 'Proceedings of the second Berkeley symposium on mathematical statistics and probability', ed., J. Neyman, Uni. of California Press, pp 481-493.
- Lasdon, L.S. (1970). Conjugate direction methods for optimal control. IEEE Tran. Automatic Control, April 1970, 267-268.
- Lasdon, L.S., Fox, R.L. and Rather, M.W. (1973). An efficient one dimensional search procedure for barrier functions. Math. Programming, 4, 279-295.
- Laurent, P.J. (1963). Convergence du procédé d'extrapolation de Richardson. Troisième Congrès de l'AFCALTI, Toulouse, pp 81-98.
- Lill, S.A. (1972). Generalization of an exact method for solving equality constrained problems to deal with inequality constraints; in 'Numerical methods for non-linear optimization', ed., F.A. Lootsma, Academic Press, London.
- Lootsma, F.A. (1968). Control optimization via penalty-free penalty functions. Philips Res. Repts. 23, 424-437.
- Lootsma, F.A. (1969). Hessian matrices of penalty functions for solving constrained optimization problems. Philips Res. Repts. 24, 322-331.
- Lootsma, F.A. (1970). Boundary properties of penalty functions for constrained minimization. Philips Res. Repts. Suppl. 1970, No. 3.
- Lootsma, F.A. (1972). A survey of methods for solving constrained minimization problems via unconstrained minimization; in 'Numerical methods in non-linear optimization', ed., F.A. Lootsma, Academic Press, London.
- Luenberger, D.G. (1973). 'Introduction to linear and non-linear programming'.

Addison-Wesley.

- Mangasarian, O.L. (1975). Unconstrained methods in non-linear programming. Comp. Sci. Tech. Report #253, Uni. of Wisconsin-Madison.
- Mangasarian, O.L. and Fromowitz, S. (1967). The Fritz-John necessary optimality conditions in the presence of equality and inequality constraints. J. Math. Anal. Appl., 17, 37-47.
- Martensson, K. (1973). A new approach to constrained function optimization. J. Opt. Theory Appl., 12, 531-554.
- Miele, A., Moseley, P.E., Levy, A.V. and Coggins, G.M. (1972). On the method of multiplier for mathematical programming problem. J. Opt. Theory Appl., 10, 1-32.
- Mifflin, R. (1972). On the convergence of the logarithmic barrier function method; in 'Numerical methods for non-linear optimization', ed., F.A. Lootsma, Academic Press, London.
- Mifflin, R. (1975). Convergence bounded for non-linear programming algorithms. Math. Programming, 8, 251-271.
- Morrison, D.D. (1968). Optimization by least squares. SIAM J. Numer. Anal., 5, 83-88.
- Murray, W. (1969a). Constrained optimization. Natn. Phys. Lab. Report MA79.
- Murray, W. (1969b). An algorithm for constrained minimization; in 'Optimization', ed., R. Fletcher, Academic Press, London & New York.
- Murray, W. (1971). Analytic expressions for the eigenvalues and eigenvectors of the hessian matrices of barrier and penalty functions. J. Opt. Theory Appl., 7, 189-196.
- Nakayama, H., Sayama, H. and Sawaragi, Y. (1975). A generalized Lagrangian function and multiplier method. J. Opt. Theory Appl., 17, 211-227.
- O'Doherty, R.S. and Pierson, B.L. (1974). A numerical study of augmented penalty function algorithm for terminally constrained optimal control problems. J. Opt. Theory Appl., 14, 393-403.

- Oi, K., Sayama, H. and Iakamatsu, T. (1973). Computational schemes of the Davidon-Fletcher-Powell method in infinite-dimensional space. *J. Opt. Theory Appl.*, 12, 447-458.
- Osborne, M.R. (1972). Topics in optimization. *Comp. Sci. Dept., Stanford Uni.*, STAN-CS-72-279.
- Osborne, M.R. and Ryan, D.M. (1970). On penalty function methods for non-linear programming problems. *J. Math. Anal. Appl.*, 31, 559-578.
- Osborne, M.R. and Ryan, D.M. (1972). A hybrid algorithm for non-linear programming; in 'Numerical methods for non-linear optimization', ed., F.A. Lootsma, Academic Press, London.
- Powell, M.J.D. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Comp. J.*, 7, 155-162.
- Powell, M.J.D. (1969). A method for non-linear constraints in minimization problems; in 'Optimization', ed., R. Fletcher, Academic Press, London.
- Powell, M.J.D. (1972). Problems related to unconstrained optimization; in 'Numerical methods for unconstrained optimization', ed., W. Murray, Academic Press, London & New York.
- Powell, M.J.D. (1977). A fast algorithm for non-linearly constrained optimization calculations. *Proceedings of the 1977 Dundee Conference on Numerical Analysis*, ed., G.A. Watson, Springer-Verlag.
- Robinson, S.M. (1974). Perturbed Kuhn-Tucker points and rates of convergence for a class of non-linear programming algorithms. *Math. Programming*, 7, 1-16.
- Rockafellar, R.T. (1974). Augmented Lagrange multiplier functions and duality in non-convex programming. *SIAM J. Control*, 12, 269-285.
- Rosen, J.B. and Kreuser, J. (1972). Private communications.
- Rosen, J.B. and Suzuki, S. (1965). Construction of non-linear programming test problems. *Comm. A.C.M.*, 8, 113.

- Rosenbrock, H.H. (1960). An automatic method for finding the greatest or least values of a function. *Comp. J.*, 3, 175-184.
- Sargent, R.W.H. and Sebastian, D.J. (1972). Numerical experience with algorithms for unconstrained minimization; in 'Numerical methods for non-linear optimization', ed., F.A. Lootsma, Academic Press, London and New York.
- Schmit, L.A. and Fox, R.L. (1965). Integrated approach to structural synthesis. *AIAA J.*, 3, 1104-1112.
- Stewart, G.W. (1967). A modification of Davidon's minimization method to accept difference approximations of derivatives. *J. ACM.*, 14, 72-83.
- Topia, R.A. (1977). Diagonalized multiplier methods and quasi-Newton methods for constrained optimization. *J. Opt. Theory Appl.*, 22, 135-194.
- Tripathi, S.S. and Narendra, K.S. (1970). Optimization using conjugate gradient methods. *IEEE Tran. Automatic Control*, April 1970, 268-270.
- Wright, M. (1976). Numerical methods for nonlinearly constrained optimization. Ph.D. Thesis, Stanford University.
- Zangwill, W.I. (1967). Non-linear programming via penalty functions. *Management Sci.*, 13, 344-358.