NOTES ON STATIC AND DYNAMIC OPTIMIZATION

René Victor Valqui Vidal

LYNGBY 1984

FORELÆSNINGSNOTE

imsor

Trykt af [illegible] - DTH

PREFACE

This is not an introductory textbook on optimization, there are so many of them especially in the American literature having some common features of: poor quality, misleading content and sometimes decidedly erroneous statements. This NOTES is not a topological approach to optimization, the mathematical background of most potential users will not be sufficient to cope with such an elegant but highly theoretical discussion.

This book pretends to be a unified presentation of the main theoretical and numerical results on optimization, and at the same time it provides an outlook to the many areas of application. It contains what I believe is the minimum knowledge required for a serious use of normative mathematical models. Most of the results presented here are available in the current literature although they are not well-known to most users of optimization methods, - what is different is the way they are presented: stepwise from general to particular results placing emphasis on the geometrical rather than the mathematical abstract approach.

This work has evolved from one of my teaching and research activities at The Institute of Mathematical Statistics and Operations Research, during the past five years, and it is intented to be essentially self-contained and should be suitable for classroom work or self-study. With this in mind, Hans F. Ravn and I have written (in Danish) a companion monograph with examples and case studies on optimization that provides a great deal of illustration on theory, method and applications.

I am indebted to my students at The Technical University for their critical comments and sincere suggestions while working with this material as the book evolved. Specially, I am grateful to Hans F. Ravn, MSc., who has read the whole manuscript, provided some examples and suggested several valuable improvements.

My thanks to Jørgen, Annelise, Kirsten, Bente, Birgit and Svend Åge for their work with the preparation of this manuscript.

Ø. VICTOR

Albertslund, August 1976.

This is a second edition of 'Notes on Static and Dynamic Optimization', only minor changes have been done and a few things have been added.

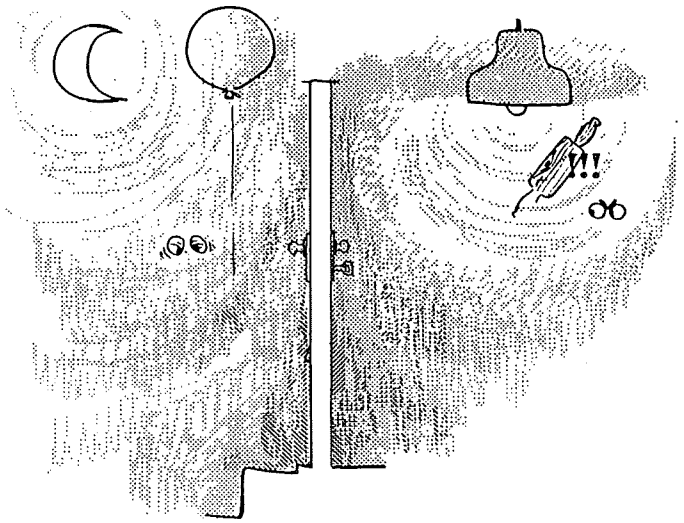
Once again

Albertslund, August 1977.

This is a third edition of 'Notes on Static and Dynamic Optimization'. Some few corrections have been added. Chapter 10: two application has been removed, because I am planning to publish a book on advanced applications where this chapter will be included. Moreover, the EPILOGUE has been extended.

Nyhavn, June 1981.

"OPTIMIZATION IS A TOOL"



A TOOL IS AN OBJECT THAT CAN SERVE MANY PURPOSES

CONTENTS

	Page
PREFACE	i
INTRODUCTION	9
PART ONE: STATIC OPTIMIZATION	13
Chapter 1: Fundamentals of Mathematical Programming	15
1. INTRODUCTION	17
2. SOME EXAMPLES	18
3. THE MATHEMATICAL PROGRAMMING PROBLEM	21
4. SOME GENERAL RESULTS	22
4.1 Local-global theorem	24
4.2 Existence problem	25
5. LOCATION OF OPTIMAL SOLUTIONS INSIDE X	27
5.1 Necessary conditions for local maximum	27
5.2 Sufficient conditions for local maximum	30
5.3 Concave and pseudoconcave problems	31
6. P-D (payoff-decision) SPACE - the geometry of the feasible set	32
7. P-R (payoff-resource) SPACE - family of constrained problems	34
8. REFERENCES	41

Page

Chapter 2: Theory 43

1. INTRODUCTION 45
2. SUFFICIENT CONDITIONS FOR OPTIMAL SOLUTION - Everett's theory 46

2.1 Case of only equality constraints (m=0) 53
The homogeneous strong Lagrange multiplier principle ($\mu \in L_{\bar{C}}$) 54
Lagrange multipliers 57

2.2 Case of only inequality constraints (p=0) 59
Duality theory 64

2.3 Final remarks 72

3. NECESSARY CONDITIONS FOR OPTIMAL SOLUTIONS - Kuhn-Tucker theory 73
A geometrical interpretation 73

3.1 Generalized Kuhn-Tucker theorem 77
Constraint qualifications 79
Kuhn-Tucker conditions 79

3.2 Second-order necessary conditions 84
3.3 Sufficiency of the Kuhn-Tucker conditions 84
3.4 The homogeneous weak Lagrangian principle 86

4. EVERETT'S THEORY VS. KUHN-TUCKER THEORY 89

5. REFERENCES 91

Chapter 3: Computational methods 93

1. INTRODUCTION 95

2. ONE-DIMENSIONAL MINIMIZATION 96

2.1 Methods which specify an interval 96
Fibonacci search 97
Search by golden section 99

2.2 Methods which specify the position 100
The algorithm of Davies - Swann - Campey 100
Powell's method 101
Coggin's method 102

2.3 Final remarks 102

Page

3. UNCONSTRAINED MINIMIZATION 103

3.1 Unconstrained minimizations procedures using derivatives 103
General procedure 104

A) The optimum gradient or steepest descent method 107
B) The Newton - Raphson method 109
C) Davidon - Fletcher - Powell method 110
D) Fletcher - Reeves conjugate gradient method 112

3.2 Unconstrained minimization procedures without using derivatives 113

A) Pattern Search 113
B) Powell's method 116

3.3 Comparison of the methods 119

4. CONSTRAINED MINIMIZATION 120

4.1 Linear Approximation Methods 122

A) Method of approximate programming 122
B) Projection methods 125
C) The generalized reduced gradient method 129

4.2 Penalty functions 132

A) Generalized Lagrange multiplier technique 132
B) Sequential unconstrained minimization 134

4.3 Comparison of the methods 138

5. REFERENCES 139

Chapter 4: Case studies 141

1. INTRODUCTION 143
2. CASE 1: A production planning problem 144
3. CASE 2: Economic dispatching of interconnected power systems 152
4. CASE 3: Least-cost allocation of reliability investment 156

Page

5. CASE 4: Optimal Control of processes 158
6. CASE 5: Portfolio Selection 163
7. CASE 6: Control of water quality in a stream 166
8. CASE 7: Linear Programming 171
9. REFERENCES 181

PART TWO: DYNAMIC OPTIMIZATION 183

CHAPTER 5: Fundamentals of optimal control theory 185

1. INTRODUCTION 187
2. SOME EXAMPLES 188
3. PROBLEM FORMULATION 193
4. IDENTIFICATION AND CONTROL 193
5. THE MATHEMATICAL MODEL OF A CONTINUOUS-TIME OPTIMAL CONTROL PROBLEM 195
   - 5.1 Types of optimal control 201
   - 5.2 Typical control problems 203
   - 5.3 Suboptimal control (heuristic methods) 206
   - 5.4 Controllability and observability 207
6. THE MATHEMATICAL MODEL OF A DISCRETE-TIME OPTIMAL CONTROL PROBLEM 209
7. REFERENCES 211

	Page
CHAPTER 6: Theory	213
1. INTRODUCTION	215
2. DISCRETE-TIME OPTIMAL CONTROL PROBLEMS	215
2.1 Sufficient conditions for optimality	216
A decision tree and the principle of optimality	216
Functional equations	219
Complementary remarks	224
Mitten's conditions	224
2.2 Necessary conditions for optimality derived from the static theory	226
A discrete maximum principle derived from a control theoretic approach	234
2.3 Interpretation of the discrete-time control problem in terms of Everett's theory	239
Complementary remarks	244
3. CONTINUOUS-TIME OPTIMAL CONTROL PROBLEMS	245
3.1 Sufficient conditions for optimality	246
Hamilton-Jacobi equation	246
Complementary remarks	251
3.2 Necessary conditions for optimality	252
The Pontryagin's maximum principle	252
Boundary conditions	258
Further comments	263
3.3 Variants of the problem	267
Higher-order equations of motion	267
Choice of extra parameters	268
Simultaneous control and state constraints	269
3.4 Some special continuous-time problems	269
Minimum-time control of linear systems	269
Calculus of variations	272
3.5 Relationship of the maximum principle to dynamic programming	275
4. REFERENCES	278

	Page
CHAPTER 7: Computational methods	281
1. INTRODUCTION	283
2. DISCRETE-TIME OPTIMAL CONTROL	283
The standard computational algorithm of dynamic programming	284
Remarks	286
Computational algorithms based on the discrete maximum principle	289
Remarks	290
3. CONTINUOUS-TIME OPTIMAL CONTROL	291
Two-point boundary-value problems	293
Remarks	295
4. REFERENCES	296
CHAPTER 8: Case studies	299
1. INTRODUCTION	301
2. CASE 1: A simple inventory problem	302
3. CASE 2: A production planning problem	303
4. CASE 3: Control of a chemical plant	307
5. CASE 4: A road building problem	309
6. CASE 5: A one-sector economy	311
7. REFERENCES	314
PART THREE	317
CHAPTER 9: Stochastic optimization	319
1. INTRODUCTION	321
2. STRATEGIES TO DEAL WITH STOCHASTIC OPTIMIZATION PROBLEMS	322

		Page
3.	STOCHASTIC PROGRAMMING	323
	3.1 The objective function	325
	3.2 The feasible set	326
	3.3 Chance constrained programming	327
	3.4 Two-stage programming	328
	3.5 Fixed-charge penalty	330
4.	STOCHASTIC CONTROL	331
	4.1 Functional equations	333
	4.2 Certainty equivalence or separation principle	335
	4.3 Stochastic maximum principle	336
5.	REFERENCES	337
EPILOGUE		339
1.	Use of geometric and algebraic inequalities: Duality	340
2.	Need for calculus: Lagrange's method of undetermined multiples	343
3.	Development of the calculus of variations	345
4.	Development after World War II	346
APPENDIX		349

INTRODUCTION

Optimization problems within engineering, economics and operations research are usually formulated as normative mathematical models, that is models where we want to find a vector that maximizes a given function or functional and that at the same time satisfies some constraints. A great deal of numerical methods are available to solve such practical problems. These numerical procedures are based on several theories that have been developed by studying the mathematical properties of the normative models. While preparing this book on optimization I have tried to find a suitable balance between theory, numerical methods and applications.

In what concerns theory, I present in a compact form the most important results, placing emphasis on understanding the geometrical properties of the problem while most proofs have been avoided. Here I give a well grounded discussion of Lagrange multiplier theories. Lagrange multipliers are very often employed but their users are seldom aware of the mathematical properties of these multipliers.

There exist many numerical methods, here I present the best known. The theoretical background of each method is outlined as well as its properties, advantages and disadvantages.

Finally, this NOTES contains many case studies within different application areas, some of them are classical examples while others are the product of my own research work.

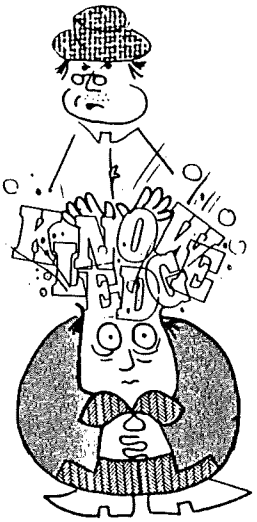
The contents of this book is divided in 3 parts. PART ONE deals with STATIC OPTIMIZATION, that is optimizations problems at a given instant of time, while PART TWO is concerned

with DYNAMIC OPTIMIZATION or optimization problems at a given interval of time. These two aspects of optimization problems are usually discussed in the available literature under the name of MATHEMATICAL PROGRAMMING and MODERN CONTROL THEORY, respectively. In this book I have attempted to show that these two areas have a common basic background. Finally, PART THREE contains an introduction to STOCHASTIC OPTIMIZATION, an EPILOGUE and a mathematical APPENDIX.

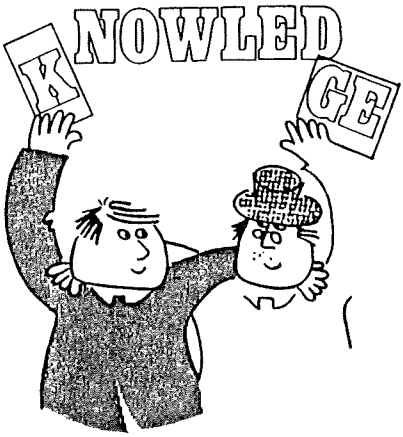
Each chapter contains a complete, in a qualitative sense, list of references. Most of these books and journals are available at IMSOR's library. Moreover at IMSOR some of my students have developed an OPTIMIZATION COMPUTER SYSTEM that can be used to find numerical solutions to small problems.

The essential mathematical prerequisite is the one corresponding to engineers from The Technical University and a knowledge on Linear Programming theory would be very desirable. Although this NOTES has been employed in a course of Operations Research, I believe that other researchers could have equal benefit in reading this book. This will also be the case for many of my operations research colleaques who need to refresh their knowledge on optimization.

At the end a few words on the way I have employed this book. First, I usually presented the material of this NOTES in a very informal lecture to motivate the more careful self-study that is necessary before the students were able to solve case studies during the practical exercises. Thus it is of extreme importance that the students participate in the lectures and the practical exercises, and at the same time by self-study get familiar with the contents of this NOTES.



MAN PASSIVELY LETS THE TEACHER FILL UP HIS MIND



RATHER MAN SHOULD CREATE AND RECREATE KNOWLEDGE IN COOPERATION WITH THE TEACHER

PART ONE

STATIC OPTIMIZATION

CHAPTER 1

STATIC OPTIMIZATION:

Fundamentals of Mathematical Programming

1. INTRODUCTION

The purpose of this chapter is to introduce the so called Mathematical Programming problem and to evaluate its difficulty by relating it to the well-known Linear Programming problem. Moreover, some definitions and general results will also be discussed. Special emphasis will be placed in understanding the geometrical properties of the problem. In this respect the understanding of the notion of the so called perturbation function will facilitate the comprehension of most of the fundamental results of static optimization.

Let us first begin by discussing some examples with the purpose to illustrate the type of problems we will be dealing with.

2. SOME EXAMPLES

a) Thomas manufacturers a smoking blended called Nyborg Gold. The blend is made up of tobacco and mary-john leaves. For legal reasons the fraction x_1 of mary-john in the mixture must satisfy $0 < x_1 < \frac{1}{2}$. From extensive market research Thomas has determined his expected volume of sales as a function of x_1 and the selling price x_2. Furthermore, tobacco can be purchased at a fixed price, whereas the cost of mary-john is a function of the amount purchased. If Thomas wants to maximize his profits, how much mary-john and tobacco should be purchase, and what price x_2 should he set ?

Our decision vector is $x = (x_1,x_2)'$, thus if x_1 is known we can calculate the total amount of mary-john purchased: $x_1 f(x_1,x_2)$ and the total amount of tobacco purchased $(1-x_1)f(x_1,x_2)$, where $f(x_1,x_2)$ is the expected sales volume of mixture as a function of $(x_1,x_2)'$, this is equal to the total amount to produce since it is not profitable for Thomas to produced more that can be sold.

The net profit is then: $N(x_1,x_2) = x_2 f(x_1,x_2) - C(x_1,x_2)$ where the total cost $C(x_1,x_2)$ is the addition of $p_1(x_1 f(x_1,x_2))$, the purchase price of $x_1 f(x_1,x_2)$ pounds of mary-john, and $p_2(x_1,x_2)$ where p_2 is the purchase price per pound of tobacco.

The set of admissible decisions is X, where

$$X = \{(x_1,x_2) \mid 0 < x_1 < \tfrac{1}{2},\ 0 < x_2 < \infty\} \subset R^2$$

Formally, we have the following decision problem: find $x \in R^2$, so that

$$\max z = N(x_1,x_2)$$

subject to: $x \in X \subset R^2$

Note that (x_1,x_2) take continuous values while all the functions defining $N(x_1,x_2)$ are well-defined real-valued functions.

b) A linear programming model, $\max\{z = cx | Ax \leq b, x \geq 0\}$, is a decision problem of a very special nature that will be contained in the more general model to be discussed in this chapter.

c) A manufacturer makes in lots a certain item which he in turn supplies to another producer. The manufacturer has a contract with the other producer to supply the following quantities of the part in each of the next 5 months: d_1,d_2,d_3,d_4,d_5. The manufacturer has to deliver the part in the month specified. However, he need not manufacture them during that month. He can, if desired, produce several months supply in advance and inventory the parts until needed. There is an inventory-carrying cost of h per unit per month. Imagine, for simplicity, that this cost is based on the inventory on hand of the end of the month. The cost of the setup for one production run is S. Setups are considered only at the beginning of a month. The time required to carry out a production run will be ignored, so that if a production run is made at the beginning of month j, the units produced are available to meet the demand in month j. We wish to determine when setups should be made and how many units should be produced in each of the production runs, given that the combined setup and inventory-carrying charges are to be minimized. Assume that no units are on hand at the beginning of the first month, and it is desired not to have any on hand at the end of the last month.

The <u>decision vector</u> is $x = (y_1,y_2,y_3,y_4,y_5; x_1,x_2,x_3,x_4,x_5)'$ where y_j is either 0 if a production run is not setup in month j, or 1 otherwise, and x_j denotes the amount produced in the j'th month. The fact that when $y_j=0 \Rightarrow x_j=0$ and $x_j > 0 \Rightarrow y_j=1$ can be expressed mathematically as

$$x_j(1-y_j) = 0$$

$$x_j \geq 0,\ y_j = \{0,1\} \quad \text{for } j=1,2,\cdots,5$$

As the inventory level at each month has to be ≥ 0, then

$$\sum_{i=1}^{j} x_i - \sum_{i=1}^{j} d_i \geq 0\ , \qquad j=1,\cdots,4$$

and the final inventory

$$\sum_{i=1}^{5} x_i - \sum_{i=1}^{5} d_i = 0$$

Formally our decision problem is:

Find $x \in R^{10}$, so that

$$\min z = S \sum_{j=1}^{5} y_j + h \sum_{j=1}^{4} \left(\sum_{i=1}^{j} x_i - \sum_{i=1}^{j} d_i \right)$$

subject to:

$$\sum_{i=1}^{j} x_i \geq \sum_{i=1}^{j} d_i, \qquad j=1,\cdots,4$$

$$\sum_{i=1}^{5} x_i = \sum_{i=1}^{5} d_i$$

$$x_j(1-y_j) = 0, \qquad j=1,2,\cdots,5$$

$$x_j \geq 0, \qquad j=1,2,\cdots,5$$

$$y_j \text{ either 0 or 1}, \qquad j=1,2,\cdots,5$$

Note that while x_j takes on continuous values, y_j is a so called binary variable.

3. THE MATHEMATICAL PROGRAMMING PROBLEM

The general mathematical or nonlinear programming problem can be formally formulated as follows:

Find $x \in R^n$, so that

$$\max f(x)$$

subject to: $x \in X$, $X \subseteq R^n$

where:

$x = (x_1, x_2, \cdots, x_n)'$ is the decision vector, a vector in Euclidean n-space, R^n;

X represents the set of feasible decision vectors, it is a subset of R^n, usually denominated as the feasible or constraint set; and

$f(x)$, $f: R^n \to R$, is a given real-valued function to be maximized, usually denominated as objective, criteria or utility function.

The solution of our problem, $x^* \in X$, will be denominated as the optimal decision or the global maximum, i.e.

$$f(x^*) \geq f(x), \quad \forall x \in X$$

Moreover, if x^* is unique, it is a strict global maximum.

Often we are interested in minimizing $f(x)$, therefore certain elementary relations between maxima and minima are of interest. If $\beta > 0$, and α arbitrary, then minimizing $f(x)$ is equivalent to minimizing $\alpha + \beta f(x)$, or to maximizing $\alpha - \beta f(x)$, that is

multiplying f(x) by -1 can be used to convert minimization problems to maximization problems.

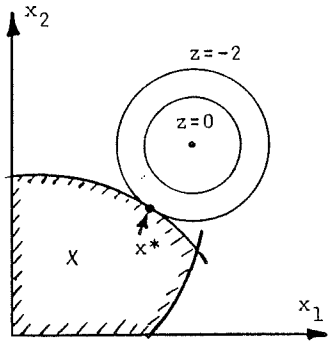
The feasible set, X, might be open, that is if and only if every point of X is an interior point. Sometimes X will be closed, that is if and only if X contains all its boundary points. If and only if given any two points in X, the distance between these two points is finite, then X is bounded, otherwise is unbounded. Sometimes X will be assumed to be compact, as $X \subset R^n$, X is compact if and only if it is closed and bounded. Finally, X is said to be convex, if and only if any linear convex combination of two points in X is also in X.

Example 1

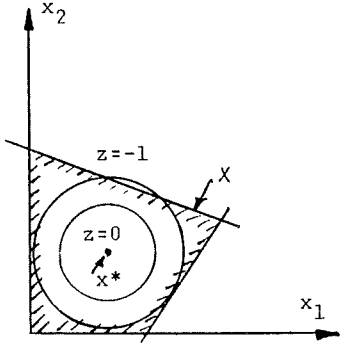
In linear programming, a special case of our problem, f(x) is linear and the feasible set X is a polyhedral convex set or, if bounded, a convex polyhedron, that is defined by a set of linear inequalities.

4. SOME GENERAL RESULTS

We know from one of the fundamental theorems of linear programming that the objective function achieves maximum, if one exists, at an extreme point of the feasible set. This is not necessarily true for the nonlinear programming problem as illustrated below:



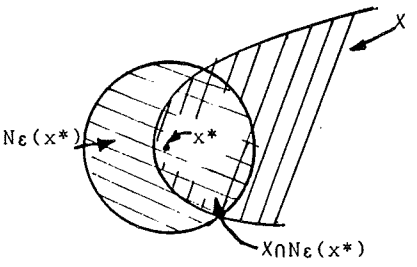
x* ∈ boundary [X]



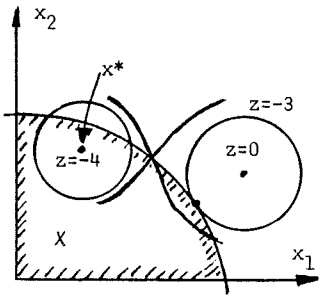
x* ∈ int [X]

Another further complication is the possibility of having several local maxima, giving different values of the objective function. $x^* \in X$ is a local maximum if

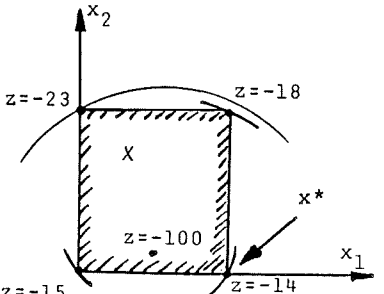
$$f(x^*) \geq f(x), \quad \forall x \in X \cap N_\varepsilon(x^*)$$



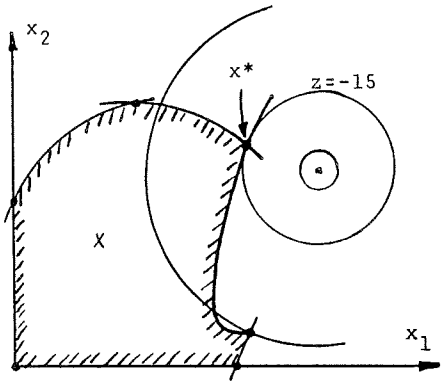
where $N_\varepsilon(x^*)$ is an ε-neighbourhood of x^* for some positive ε, however small, in this case the set of all x such that $|x - x^*| < \varepsilon$. Obviously a global maximum is also a local maximum but the viceversa is not necessarily true. Examples of local maxima are shown below:



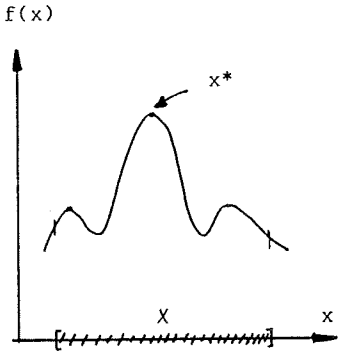
2 local maxima



4 local maxima



2 local maxima



3 local maxima

4.1 Local - global theorem

Existence of several local maxima makes the nonlinear problem difficult to solve. Sometimes it is practically impossible to guarantee that a local maximum is also global. Linear programming problems are easy to solve because of the fact that any local maxima is also a global one. This property can be generalized by the following theorem that gives sufficient conditions for local maxima to be global maxima.

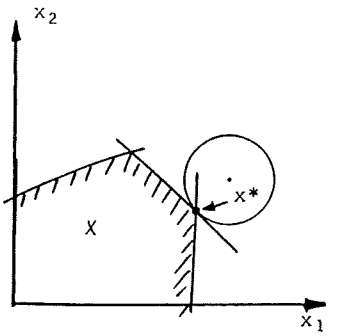
Theorem 1 [1]

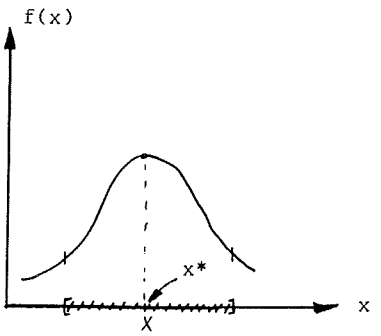
Let $f(x)$ be a concave function over the closed convex set X, then any local maximum of $f(x)$ is also the global maximum of $f(x)$ over X. The set of points at which $f(x)$ takes on its global maximum is a convex set. Moreover, if $f(x)$ is strictly concave, then the global maximum of $f(x)$ is taken on at a unique point.

Corollary

The last theorem is also true if $f(x)$ is a pseudoconcave function.

The next two figures show examples where this theorem is applicable:





4.2 Existence problem (Weierstrass' theorem)

We know that linear programming problems may have no solution because either X is empty or X is unbounded and the objective function can increase without bound in this set. This can also happen in nonlinear problems. Furthermore, solutions might not exist in the nonlinear case because either $f(x)$ is not continuous or X is an open set. In cases where a maximum or minimum does not exist, one introduces the more general idea of supremum or infimum. The supremum of $f(x)$ in X, denoted sup $f(x)$, is the least value of λ for which: $f(x) \leq \lambda$, $\forall x \in X$.

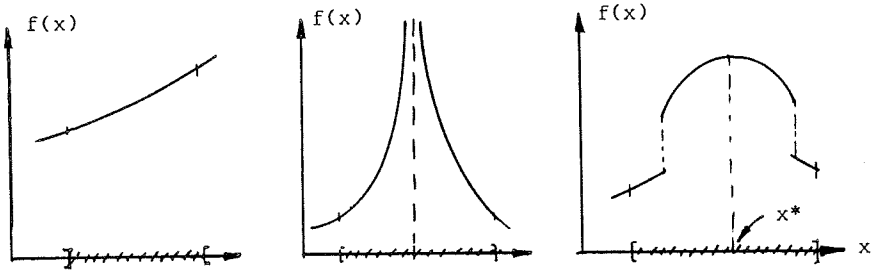
Correspondingly, the infimum, denoted inf $f(x)$, is the greatest value of λ for which $f(x) \geq \lambda$, $\forall x \in X$. These values represent extremes that can be approached arbitrarily closely, but not necessarily attained. The statement "a maximum exists" is equivalent to the statement "the supremum is attained".

Sufficient conditions for an objective function to attain its extreme values are given by the following theorem.

Theorem 2 [2]

If the feasible set X is nonempty and compact, and the objective function f(x) is continuous on X then f(x) has a global maximum either in the interior or on the boundary of X.

The examples below illustrate some cases where this theorem does not apply:



Actually, theorem 2 can be extended to be also applicable to the last figure.

Corollary

The last theorem is also true if f(x) is an upper semicontinuous function.

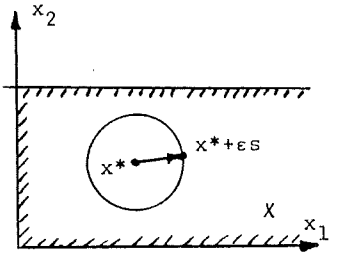
The last two theorems tell us that if X is a nonempty compact and convex set and f(x) is continuous and concave over X then a local maximum is a global maximum, and the set of points at which the maximum is obtained is convex.

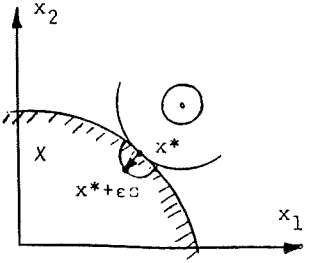
5. LOCATION OF OPTIMAL SOLUTIONS INSIDE X

Conceivably a maximizing value x* could be located by testing all the points of X, this is the so-called enumerative method of maximization. This blind method of search is applicable only for finite X, and even then it may take long time. However, one may be able to exploit special properties of the decision problem to show right from the beginning that only certain points of X could possibly yield a maximum. Such special properties may be analytic ones, such as continuity or differentiability, or they may be structural ones specific to the case in study.

5.1 Necessary conditions for local maxima

A partial test of whether a particular value x* truly maximizes f(x) in X can be made by considering local variations for x in the neighbourhood of the postulated x*. Thus, suppose that the point x* + εs, where ε is a nonnegative scalar and s a vector, belongs to X for all nonnegative ε less than some positive value ε(s). Then s will be described as directed into the interior of X from x*, or as a feasible direction from x*. The figures below illustrate this concept:





Now, since x* is optimal

$$f(x^*) \geq f(x^* + \varepsilon s)$$

whenever $0 \leq \varepsilon \leq \varepsilon(s)$.

Let us assume that f is differentiable, then by Taylor's theorem

$$f(x^* + \varepsilon s) = f(x^*) + \varepsilon \nabla f(x^*)s + o(\varepsilon)$$

where $\frac{o(\varepsilon)}{\varepsilon} \to 0$ as $\varepsilon \to 0$

Thus we obtain: $\nabla f(x^*)s + \frac{o(\varepsilon)}{\varepsilon} \leq 0$

Letting $\varepsilon \to 0$, we obtain that

$$\nabla f(x^*)s \leq 0$$

Now if x^* is an interior point of X or X is an open set the last inequality must hold for $\forall s \in R^n$, for instance s^0 and $-s^0$, then we must have:

$$\nabla f(x^*) = 0'$$

Thus we have found the well-known stationarity condition, the condition that x^* be a stationary (or critical) point of $f(x)$

Theorem 3

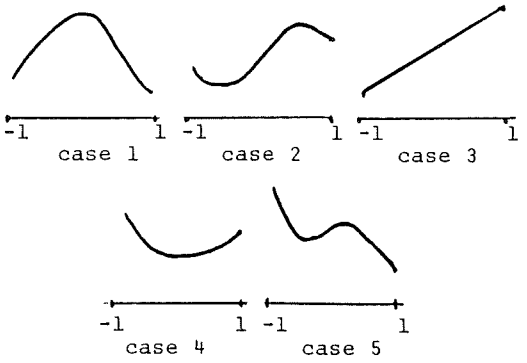
Let X be an open subset of R^n. Let $f: R^n \to R$ be a differentiable function. Then x^*, the optimal solution, must satisfy:

$$\nabla f(x^*) = 0'$$

The examples below illustrate this theorem.

Example 2

Assume X open and $f(x)$ has to be maximized where:
$X = \{x \mid -1 < x < 1\}$ and $f(x)$ is illustrated below:

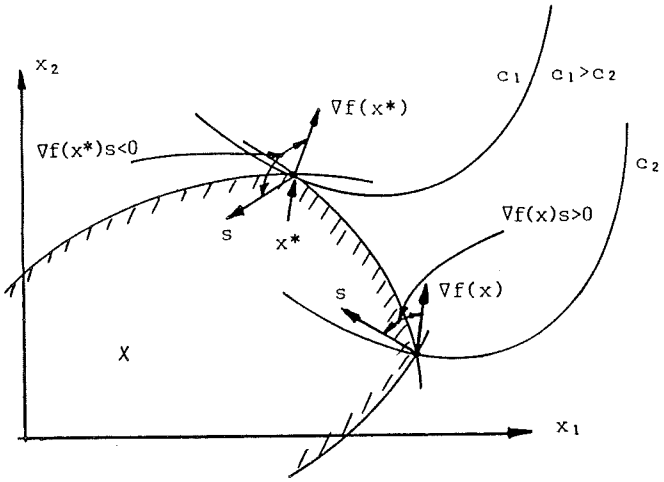


Case	Does there exist x^* ?	At how many points $\nabla f(x^*) = 0'$?	Further consequences
1	Yes	Exactly one point	x^* is the unique optimal
2	Yes	More than one point	
3	No	None	
4	No	Exactly one point	
5	No	More than one point	

Theorem 4

Suppose that X is compact. If the gradient vector, $\nabla f(x^*)$, exists, then: $\nabla f(x^*)s \leq 0$, for feasible directions s from x^*, and if x^* is an interior point of X then: $\nabla f(x^*) = 0'$.

This theorem is illustrated below:



In the next chapter, this theorem will give origin to the well-known Kuhn-Tucker conditions.

That is, a maximizing (or minimizing!) point of f(x) is either:

a) a stationary point of f(x),

b) a point where f(x) is nondifferentiable, or

c) a boundary point of X.

That f(x) has a stationary point at x* does not imply that it is maximal at x*, or even that it has a local maximum there. Stationarity is a necessary condition for a maximum (or minimum) if x* is a value in the interior of X at which f(x) is differentiable.

5.2 Sufficient conditions for local maxima

Stronger conditions can be deduced assuming f(x) twice differentiable. Then by extension of the Taylor expansion,

$$f(x^* + \varepsilon s) = f(x^*) + \varepsilon \nabla f(x^*)s + \tfrac{1}{2}\varepsilon^2 s'H(x^*)s + o(\varepsilon^2)$$

where $H(x^*)$ is the Hessian matrix of $f(x)$ evaluated at x^*, and by repetition of the previous argument we can prove the following theorem:

Theorem 5 [3]

Let x^* be a stationary point of $f(x)$. Let $f(x)$ have all its second partial derivatives continuous in a neighbourhood of x^*. Then at x^*:

(i) if $H(x^*)$ is positive definite, then $f(x^*)$ is a local minimum,

(ii) if $H(x^*)$ is negative definite, then $f(x^*)$ is a local maximum,

(iii) if $H(x^*)$ is indefinite, then $f(x^*)$ is neither a maximum nor a minimum. This point is a saddle point, and

(iv) if $H(x^*)$ is merely semidefinite without further calculations we cannot say what $f(x^*)$ is. $H(x^*)$ positive (negative) semidefinite is still a necessary condition for a local minimum (maximum).

5.3 Concave and pseudoconcave problems

The maximization of concave functions and the minimization of convex functions are usually desirable problems due to the following results.

Theorem 6 [4]

If x^* is a stationary point of the concave (convex) differentiable function $f(x)$ in the interior of X, then it is also a maximizing (minimizing) point of $f(x)$.

Corollary

The last theorem is also true if $f(x)$ is a pseudoconcave (pseudoconvex) function.

6. P-D (PAYOFF-DECISION) SPACE - THE GEOMETRY OF THE FEASIBLE SET

Let us now turn back to our problem

$$\max\{f(x) \mid x \in X\}$$

In the last section, we have investigated some of the properties of the optimal solution, if one exists, for the case of x^* being an interior point of X. We concluded that the general case is not an easy one, thus the system of equations to be solved, $\nabla f(x^*) = 0$, is in general nonlinear and might be difficult to solve.

The problem becomes more complex if x^* is not an interior point of X. Then x^*, if one exists, has to be searched at the boundary of X. Therefore a more precise definition of X is required. For instance in linear programming

$$X = \{x \in R^n \mid Ax \leq b\ ,\ x \geq 0\}$$

For the nonlinear case X might take different forms. For problems with only nonnegativity constraints we have: $X = \{x \in R^n \mid x \geq 0\}$. And for the so-called classical programming problems: $X = \{x \in R^n \mid h(x) = \bar{c}\}$ where $h: R^n \to R^p$, $n \geq p$, is a vector of functions whose $k^{\underline{th}}$ component is written $h_k(x)$, $k = 1,2,\cdots,p$, and $h_k: R^n \to R$.

The nonlinear programming problem will usually have as feasible set:

$$X = \{x \in R^n \mid x \geq 0 \ , \ g(x) \leq \bar{b}\}$$

where $g: R^n \to R^m$ and $g_j: R^n \to R$.

For the general mathematical programming problem:

$$X = \mathcal{D} \cap F \cap N$$

where: $\mathcal{D} \subseteq R^n$

$$F = \{x \in R^n \mid h(x) = \bar{c}\}$$

$$N = \{x \in R^n \mid g(x) \leq \bar{b}\}$$

our problem becomes

$$\max f(x)$$

$$x \in \mathcal{D}$$

subject to: $g(x) \leq \bar{b}$

$$h(x) = \bar{c}$$

Now, if for some j, $g_j(x^*) = \bar{b}_j$ then we say that this restriction is active or binding.

We have seen that a very important subclass of feasible sets are those that are convex. Then how can we define convex sets?

Theorem 7

If the functions $g_j(x)$ are quasi-convex, then the set

$$X = \{x \in R^n \mid g(x) \leq \gamma\}$$

is convex for any γ.

The proof is obvious by definition of quasi-convex functions. Any pseudoconvex and convex function is quasi-convex, therefore the theorem will also be valid for these cases.

A concave programming problem is one of maximizing a concave function over a convex constraint set, this family of problems are important because of Theorem 1 and other important results to be seen in the following chapters. Some authors call this as a convex programming problem due to the fact that it is analogue to the problem of minimizing a convex function over a convex constraint set. In what follows we will keep these two definitions, thus in both cases the feasibility set is convex, but if we maximize a concave function we have a concave problem while if we minimize a convex function we have a convex problem.

7. P-R (PAYOFF-RESOURCE) SPACE - FAMILY OF CONSTRAINED PROBLEMS

Let us consider the following problem

$$\max f(x)$$

$$x \in \mathcal{D}$$

$$g(x) \leq \bar{b}$$

$$h(x) = \bar{c}$$

where f, g_j and h_k are real-valued (not necessarily differentiable or continuous) functions defined on $\mathcal{D}$. We will define the so-called perturbation set as

$$\mathcal{B} = \{(b,c) \mid \exists x \in \mathcal{D} \text{ such that } g(x) \leq b, h(x) = c\}$$

That is, $\mathcal{B}$ is the set of right-hand sides or perturbation vectors such that our problem is feasible. Note that for given c if $(b^o,c) \in \mathcal{B}$ and if $b \geq b^o$, then also $(b,c) \in \mathcal{B}$. Now if

$\mathcal{D} \neq \phi$, then $\mathcal{B}$ is never empty since for any x in $\mathcal{D}$ the vector $(g(x),h(x)) \in \mathcal{B}$. Now define the so-called perturbation functions

$$f_{sup} : \mathcal{B} \to R_+ = R \cup \{+\infty\}$$

as:

$$f_{sup}(b,c) = \sup\{f(x) \mid x \in \mathcal{D}, g(x) \leq b, h(x) = c\}$$

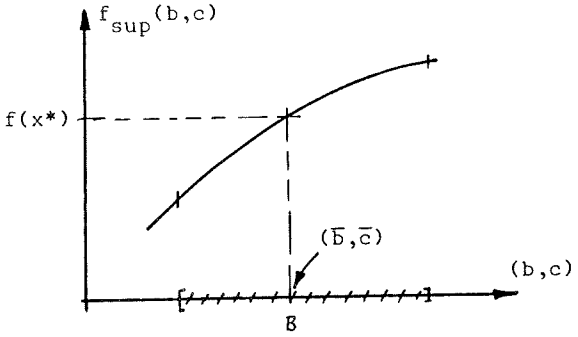
By definition of the set $\mathcal{B}$, and the admission of the value $+\infty$, the function f_{sup} is well defined. Thus, if our original problem is feasible, that is $(\bar{b},\bar{c}) \in \mathcal{B}$, then $f_{sup}(\bar{b},\bar{c})$ is defined even though $f(x)$ may not attain a maximum. On the other hand, if our problem has a solution say x^*, then $f_{sup}(\bar{b},\bar{c}) = f(x^*)$, and the supremum is actually attained by f. Conversely, if the supremum is attained by f at some feasible x^*, then x^* clearly is a solution to our problem. If $f_{sup}(\bar{b},\bar{c}) = +\infty$, the objective function of our problem is unbounded and the problem has no solution for the right-hand side $(\bar{b},\bar{c})$. For many problems of interest the maximum of f will be attained for all $(b,c) \in \mathcal{B}$. For example, this will be the case if $\mathcal{D}$ is compact and the functions f, g and h are continuous on $\mathcal{D}$. In this last case f_{sup} can be replaced with f_{max}*).

Added insight in the geometrical properties of our problem can be obtained by interpreting the point set $\{(b,c), f_{sup}(b,c) \mid (b,c) \in \mathcal{B}\}$ as a surface in R^{m+p+1}, where the first $(m+p)$ coordinates represent resource levels (right-hand sides), and the last coordinate refers to the supremum of f (the maximum

---

*) In parametric programming we solve $\max\{z = f(x) \mid g(x) \leq \bar{b} + \beta, h(x) = \bar{c} + \gamma\}$ for various values of β and γ, and observe how $f(x^*)_{\beta,\gamma}$ changes. It is seen that by suitable transformation $f(x^*)_{\beta,\gamma} = f_{sup}(b,c)$.

in well behaved cases). This is the so-called PR (payoff. vs. resource) space. This is illustrated below:



Let us further introduce the following function:

$$z_{\lambda,\mu}(b,c) = \sum_{j=1}^{m} \left[\lambda_j(b_j) - \lambda_j(\bar{b}_j)\right]$$

$$+ \sum_{k=1}^{p} \left[\mu_k(c_k) - \mu_k(\bar{c}_k)\right]$$

$$+ f_{sup}(\bar{b},\bar{c})$$

where the functions

$$\lambda(\alpha) = (\lambda_1(\alpha_1),\cdots,\lambda_m(\alpha_m)) \in S, \text{ and}$$

$$\mu(\gamma) = (\mu_1(\gamma_1),\cdots,\mu_p(\gamma_p)) \in F_{\bar{c}}$$

Now

$$F_{\bar{c}} = \{\mu \mid R^1 \to R_+^p \text{ such that } \mu_k(\bar{c}_k) < \infty, \forall k\}$$

and $L_{\bar{b}} \subseteq S \subseteq M$

where $M = \{\lambda \in F_{\bar{b}} \mid \lambda_j(\alpha_2) \geq \lambda_j(\alpha_1)\ ,\ \text{when}\ \alpha_2 > \alpha_1\ ,\ \forall_j\}$

and $F_{\bar{b}} = \{\lambda \mid R^1 \to R_+^m \ \text{ such that } \ \lambda_j(\bar{b}_j) < \infty\ ,\ \forall_j\}$

In other words a smaller class of functions $F_{\bar{b}}$ will be required namely the class of coordinatewise increasing (i.e. nondecreasing) functions in $F_{\bar{b}}$. Thus S denotes a specific subset of M. In particular the class of nondecreasing linear functions in M will be of interest, that is

$$L_{\bar{b}} = \{\lambda \in F_{\bar{b}} \mid \lambda_j(\alpha) = y_j\ \alpha\ ,\ y_j \geq 0\ ,\ \forall_j\}$$

analogously

$$L_{\bar{c}} = \{\mu \in F_{\bar{c}} \mid \mu_k(\alpha) = u_k\ \alpha\ ,\ \forall k\}$$

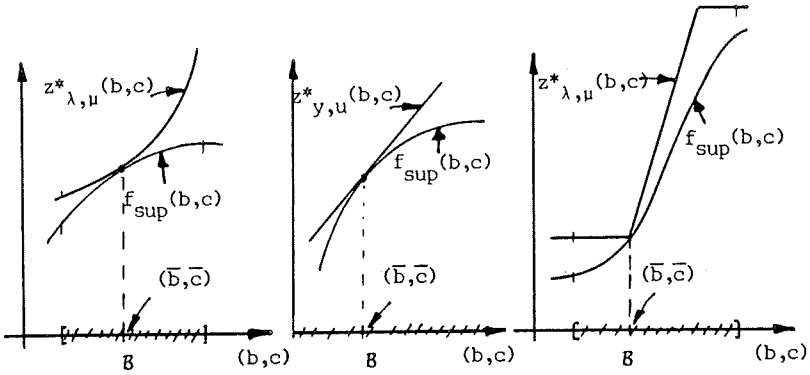
The y_j's and u_k's are usually called Lagrange multipliers, dual variables, shadow prices or implicit prices, and as we will see later on play a central role in mathematical programming. In the general case we are introducing the concept of Lagrange multiplier functions λ and μ. We should note that the m functions $\lambda_j(\ )$ are nondecreasing and are related to the m inequality constraints, whilst the p functions $\mu_k(\ )$ have not this restriction and are related to the p equality constraints.

That is the set of points $(b,c,z_{\lambda,\mu}(b,c))$ can be interpreted as a (nonvertical) surface in PR space. If we find a $\lambda^* \in S$ and a $\mu^* \in F_{\bar{c}}$ such that $z^*_{\lambda,\mu}(b,c)$ supports $f_{sup}(b,c)$ at the point $((\bar{b},\bar{c})\ ,\ f_{sup}(\bar{b},\bar{c}))$, i.e.

$$z^*_{\lambda,\mu}(\bar{b},\bar{c}) = f_{sup}(\bar{b},\bar{c})$$

and $\quad z^*_{\lambda,\mu}(b,c) \geq f_{sup}(b,c)\ ,\ \forall (b,c) \in B$

Then we denominate $z^*_{\lambda,\mu}(b,c)$ as a supporting surface, and $z^*_{y,u}(b,c)$ as a supporting hyperplane. These concepts are depicted below:



For a given (λ,μ) it is not sure that a supporting surface always exists at $\left((\bar{b},\bar{c}),f_{sup}(\bar{b},\bar{c})\right)$. This will be illustrated in the following example.

Example 3

Let us consider the following problem

$$\max\ x$$

$$\text{subject to:}\quad x \in X$$

where: $\quad X = \{x \in R \mid \frac{x^2}{2} + \delta(x)2 \leq b\ ,\ x \geq 0\}$

and $\quad b \geq 0$

$$\delta(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \end{cases}$$

Obviously if we do not take account of the feasible set, then at $x = +\infty$, we obtain $\sup f(x) = +\infty$, that is only feasible for $b = +\infty$.

Following our definition: $\mathcal{D} = \{x \mid x \geq 0\}$, $m = 1$, $p = 0$, $g_1(x) = \frac{x^2}{2} + \delta(x)2$ and $b_1 = b$.

The perturbation set is then

$$\mathcal{B} = \{b \mid \exists x \geq 0 \quad \text{such that} \quad \frac{x^2}{2} + \delta(x)2 \leq b\}$$

$$\mathcal{B} = \{b \mid b \geq 0\}$$

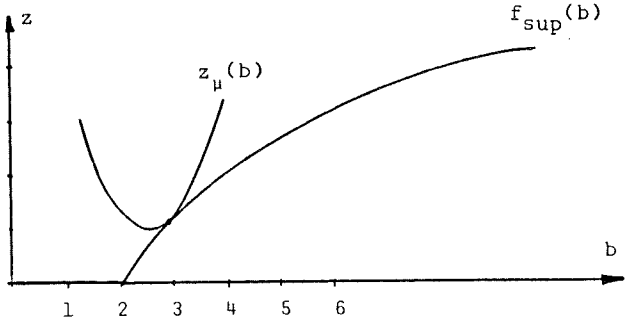
Now the optimal solution is easily obtained

$$x^* = \begin{cases} 0 & \text{if} \quad 0 \leq b \leq 2 \\ \sqrt{2(b-2)} & \text{if} \quad b > 2 \end{cases}$$

and the perturbation function is

$$f_{\sup}(b) = \begin{cases} 0 & \text{if} \quad 0 \leq b \leq 2 \\ \sqrt{2(b-2)} & \text{if} \quad b > 2 \end{cases}$$

and the PR space is illustrated below:



Although the problem is formulated with inequality constraint, the constraint is always active for $b \geq 2$, that is the optimal solution of our problem is the same as for the problem

$$\max \quad \{x \mid x \geq 0 \quad \text{and} \quad \frac{x^2}{2} + \delta(x)2 = b\}$$

This facilitates the construction of supporting surfaces because the monotonicity of multiplier functions will not be required.

Let us suppose $\bar{b} = 3$ and define

$$\mu(\varepsilon) = \alpha\varepsilon^2 + \beta\varepsilon$$

Then:

$$z_\mu(b) = \mu(b) - \mu(3) + f_{sup}(\bar{b} = 3)$$

Now if $\alpha = \sqrt{2}$ and $\beta = -5.5\sqrt{2}$, then

$$z_\mu(b) = \sqrt{2}\, b^2 - 5.5\sqrt{2}\, b + 8.5\sqrt{2}$$

and it can easily be seen from the diagram that $z_\mu(b)$ is a supporting surface at the point $\bar{b} = 3$.

Now if we assume: $\mu(\alpha) \in L_{\bar{b}}$ that is

$$\mu(\varepsilon) = y\varepsilon$$

then $$z_y(b) = (b-3)y + \sqrt{2}$$

In the interval, $-\infty \leq y \leq +\infty$, there is not a supporting hyperplane for $\bar{b} = 3$, a supporting hyperplane can be found when $\bar{b} = 0$ or $\bar{b} \geq 4$. Thus we have a gap interval (0,4) where a supporting surface using the scalar multipliers cannot be found.

A supporting surface with the monotonicity property can be obtained, by patching together the function $z_\mu(b)$ and a horizontal line.

Example 4

It is easy to show that for the LP-problem

$$\begin{aligned} \max\ & c'x \\ & Ax \leq b \\ & x \geq 0 \end{aligned}$$

the perturbation function $f_{sup}(b)$ is concave over the convex perturbation set $\mathcal{B}$. The variations of b are usually expressed as

$$\bar{b}_i + \alpha_i\, \theta$$

for $i = 1,\cdots,m$, where $\bar{b}_i$ and α_i are given constants and θ is a parameter, then the perturbation function $f_{sup}(\theta)$ is a concave function of θ (LP-problem will be further discussed in Chapter 4, case 7).

This concept of perturbation function and supporting surface will provide the geometrical background to understand most of the theoretical results on static optimization, this will be seen in the next chapter.

8. REFERENCES

[1] Walsh, G. R.: Methods of Optimization, 1975

[2] Intriligator, M. D.: Mathematical optimization and economic theory, 1971

[3] Hancock, H.: Theory of Maxima and Minima, 1960

[4] Zangwill, W.: Nonlinear Programming, 1969

OPTIMIZING

CHAPTER 2

STATIC OPTIMIZATION

Theory

1. INTRODUCTION

In this chapter we will be concerned with the main theoretical results on static optimization theory. This theory will provide necessary and sufficient conditions for optimality. Moreover it will also give us an understanding of the geometrical properties of the mathematical programming problem.

A clear knowledge of these theoretical results will facilitate the understanding of the many numerical methods that will be presented in the next chapter and the theoretical results for dynamic models to be seen in chapter 6.

Here some of the most important results in optimization theory are presented. This chapter provides a compact and informal, but not elementary, discussion of this subject. We will begin introducing the most general results and make clear the assumptions needed to prove more specific outcomes. For the sake of compactness most proofs will not be given, they are usually quite elementary and can be found in some of the references given in sec. 5.

2. SUFFICIENT CONDITIONS FOR OPTIMAL SOLUTIONS - EVERETT'S THEORY

In this section we will be dealing with the following problems:

(P1) The General Mathematical Programming problem

Find $\max f(x)$

$x \in \mathcal{D}$

so that $g(x) \leq \bar{b}$

$h(x) = \bar{c}$

where $\mathcal{D} \subseteq R^n$ and f, g and h are real-valued functions defined on $\mathcal{D}$.

(P2) Unconstrained Lagrangian problem

Find $\max\left\{f(x) - \sum_{j=1}^{m} \lambda_j(g_j(x)) - \sum_{k=1}^{p} \mu_k(h_k(x))\right\}$

$x \in \mathcal{D}$

where $\lambda \in S$ and $\mu \in F_{\bar{c}}$ are given functions.

(P3) Constrained Lagrangian problem

Find $x^* \in \mathcal{D}$, $\lambda^* \in S$ and $\mu^* \in F_{\bar{c}}$ such that

(i) x^* maximizes $\{f(x) - \sum_j \lambda_j^*(g_j(x)) - \sum_k \mu_k^*(h_k(x))\}$

(ii) $g(x^*) \leq \bar{b}$, $h(x^*) = \bar{c}$ (feasibility condition)

(iii) $\lambda_j^*(g_j(x^*)) = \lambda_j^*(\bar{b}_j)$ (complementary slackness)

(P4) Support problem

Suppose that $(\bar{b},\bar{c}) \in \mathcal{B}$. Find $\lambda^* \in S$ and $\mu^* \in F_{\bar{c}}$ such that $z^*_{\lambda,\mu}(b,c)$ supports $f_{sup}(b,c)$ at the point $((\bar{b},\bar{c}), f_{sup}(\bar{b},\bar{c}))$.

The relations among these problems will provide a great deal of information about the geometrical properties of the problem we want to solve, but let us first see an example to clarify what is meant with each problem.

Example 1

Consider (P1): $\max -(x-2)^2$

$g(x) = 3x \leq \bar{b} = 10$

$h(x) = 2x = \bar{c} = 6$

$\mathcal{D}$: $x \geq 0$

The solution to (P1) is easily seen to be $x^* = 3$, yielding $f(x^*) = -1$. From (P1) we construct (P2), where λ and μ are assumed given. Let us assume that $\lambda(\varepsilon) = \varepsilon^2$ and $\mu(\varepsilon) = +\varepsilon$. We then have

$$(P2): \quad \max \; -(x-2)^2 - (3x)^2 - 2x$$
$$\mathcal{D}: \quad x \geq 0 .$$

The solution to this is $x^* = \frac{1}{10}$. We note that this is not the solution to (P1). If, however, we let $\lambda(\varepsilon) = 0$ and $\mu(\varepsilon) = -\varepsilon$, we have

$$(P2): \quad \max \; -(x-2)^2 + 2x$$
$$\mathcal{D}: \quad x \geq 0$$

and here we get $x^* = 3$ as in (P1).

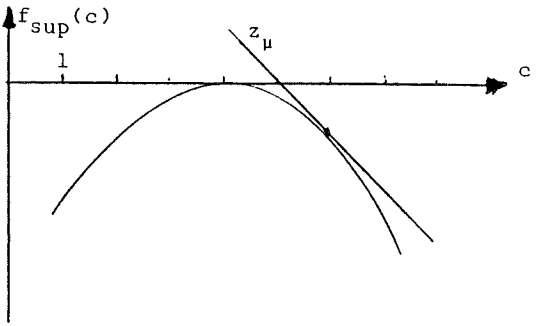
Can we always find λ and μ so that the solution to (P2) provides a solution to (P1) ?

In (P3) the problem is to find x^*, λ^* and μ^* to satisfy the 3 conditions. Note that in (i), x^* should be optimal for λ^* and μ^*, and that in this optimization x^* is restricted only to $\mathcal{D}$. The x^* thus found should then satisfy (ii) and (iii). If we let $\lambda(\varepsilon) = \varepsilon^2$ and $\mu(\varepsilon) = +\varepsilon$ se find from (i) $x^* = \frac{1}{10}$ which does not satisfy (ii), or (iii).

If however, we let $\lambda(\varepsilon) = 0$ and $\mu(\varepsilon) = -\varepsilon$ we find from (i) $x^* = 3$, which satisfies (ii) and (iii). Thus $\lambda^*(\varepsilon) = 0$, and $\mu^*(\varepsilon) = -\varepsilon$.

Will a solution to (P1) always satisfy the conditions stipulated in (P3) ?

To illustrate (P4) let us first take away the restriction $g(x) \leq \bar{b}$. We then have $f_{sup}(c)$:

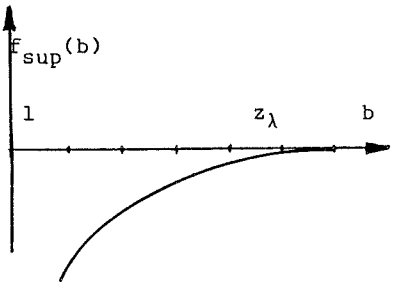


If we let $\mu(\varepsilon) = -\varepsilon$ we have

$$z_\mu(c) = -c + 6 - 1 = 5 - c \ ,$$

which supports $f_{sup}(c)$ at $(\bar{c}, f_{sup}(\bar{c}))$.

If we then take away the restriction $h(x) = \bar{c}$ we have $f_{sup}(b)$:



If we let $\lambda(\varepsilon) = 0$ we have

$$z_\lambda(b) = 0 - 0 + 0 = 0 \ ,$$

which supports $f_{sup}(b)$ at $(\bar{b}, f_{sup}(\bar{b}))$.

Considering these two solutions we conclude that the solution to (P4) is $\lambda^*(\varepsilon) = 0$ and $\mu^*(\varepsilon) = -\varepsilon$.

Will a solution to (P4) (does it always exist?) and a solution to (P1) give a solution to (P3)?

Answers to these questions will be provided in the following theoretical discussion.

Theorem 8 [1]

If x^* is a solution to (P2), then x^* is also a solution to the following relative to (P1):

(P1)'. Find $\max f(x)$

$$x \in \mathcal{D}$$

$$g_j(x) \leq \beta_j \,, \qquad j = 1,\cdots,m$$

$$h_k(x) = h_k(x^*) \,, \qquad k = 1,\cdots,p$$

where β_j is any number not less than $g_j(x^*)$ so that:

$$\lambda_j(\beta_j) = \lambda_j(g_j(x^*)) \,, \quad j = 1,\cdots,m$$

Theorem 9 [1]

x^*, λ^* and μ^* is a solution to (P3) if and only if

(i) x^* is a solution to (P1).

(ii) λ^*, μ^* is a solution to (P4).

What is worthy to note is the fact that to prove these theorems we need not restrict the feasibility set $\mathcal{D}$. Moreover the functions f, g and h are only restricted to real-valuedness.

$\mathcal{D}$ may therefore be a discrete finite set, or an infinite set of any cardinality.

Example 2

Consider the problem in example 1. With $\lambda(\varepsilon) = 0 \cdot \varepsilon$ and $\mu(\varepsilon) = -\varepsilon$ we get $x^* = 3$. Now $g(x^*) = 9$ and $\lambda(g(x^*)) = 0 \cdot 9 = 0$. Therefore β must satisfy (i) $\beta \geq 9$ and (ii) $0 \cdot \beta = 0$, which yields that β can be any number for which $\beta \geq 9$. Therefore $x^* = 3$ is also the solution to (P1) with e.g. $\bar{b} = 9$, $\bar{b} = 50$ and $\bar{b} = 60$. (Theorem 8).

Since $f_{sup}(b,c)$ is differentiable we see that the only possible solutions to (P4) are $\lambda^*(\varepsilon) = 0$ and $\mu^*(\varepsilon) = -\varepsilon$. Since $x^* = 3$ is a solution to (P1), λ^*, μ^* and x^* is a solution to (P3) and conversely (Theorem 9).

Theorem 8 tells us that if we can find λ and μ so that $\bar{b}_j = \beta_j$, $\forall j$, and $\bar{c}_k = h_k(x^*)$, $\forall k$, then the solution of (P2) is also solution to (P1). In other words this theorem permits us to transform a nonlinear programming problem to one unconstrained problem. Obviously, it is not easy to find λ and μ. Usually one has to postulate functions with unknown parameters that will be found using iterative procedures, but even then we are not sure that the desired right-hand side vector could be generated. Why is that so? The answer can be found in Theorem 9. That is if a right-hand side cannot be generated is because we cannot find a supporting surface at $(\bar{b}, \bar{c}, f_{sup}(\bar{b},\bar{c}))$, we are then dealing with an inaccessible region or gap. Later on we will see that the basic cause of a gap is nonconcavity of $f_{sup}(\bar{b},\bar{c})$. For the concave case usually functions in $L_{\bar{b}}$ and $L_{\bar{c}}$ are sufficient to solve (P1). Nevertheless Theorem 8, the generalized Everett's theorem [2], is "fail-safe" in the sense that any solution that it does yield are guaranteed to be optimal.

Example 3

Let us consider the problem of example 3 in the last chapter, for $\bar{b} = 3$. The Lagrange multiplier function was

$$\mu(\varepsilon) = \sqrt{2}\,\varepsilon^2 - 5.5\,\sqrt{2}\,\varepsilon$$

Now the Lagrangian function is

$$x - \sqrt{2}\left(\frac{x^2}{2} + \delta(x)2\right)^2 + 5.5\,\sqrt{2}\left(\frac{x^2}{2} + \delta(x)2\right)$$

that is maximized by $x^* = \sqrt{2}$, this is also a solution to (P2) because $x^* > 0$. Applying Theorem 8, $x^* = \sqrt{2}$ will also be a solution to

$$\max x$$

$$x \geq 0$$

$$\frac{x^2}{2} + \delta(x)2 = \frac{(\sqrt{2})^2}{2} + 2 = 3$$

a problem that is equivalent to our original one, that is $x^* = \sqrt{2}$ is the optimal solution to (P1) when $\bar{b} = 3$.

For the case that

$$\mu(\varepsilon) = u\,\varepsilon$$

the (P2) is

$$\max\left(x - u\left(\frac{x^2}{2} + \delta(x)2\right)\right)$$

$$x \geq 0$$

Now for $u > \frac{1}{2}$, we have $x^* = 0$ and $b = 0$. For $u \leq \frac{1}{2}$, we have $x^* \geq 2$ and $b \geq 4$. Thus the right-hand sides within the interval $(0,4)$, a gap, cannot be generated by varying the

Lagrange multiplier u. Therefore a solution to (P1) cannot be found by using $\mu \in L_{\bar{c}}$ and solving (P2). That is so because a solution to (P4) does not exist as shown in example 3 in the last chapter.

2.1 Case of only equality constraints (m = 0)

In this case stronger conditions can be found.

Theorem 10 [1]

x^* is a solution to (P1) if and only if there is a $\mu^* \in F_{\bar{c}}$ such that x^*, μ^* is a solution to (P3).

Thus while in Theorem 9 it was possible to have a solution to (P1) without having a solution to (P3), this is not longer possible in Theorem 10. See for instance example 1.

The last theorem expresses sufficient conditions under which a constrained problem can be solved as an unconstrained one. This sufficiency condition can be roughly expressed: "if the technique works, it is valid".

Example 4

Consider the maximization of

$$f(x) = \sum_{k=1}^{n} \left(x_k - x_k \log x_k\right)$$

in $x \geq 0$, subject to

$$\sum_{k=1}^{n} x_k = N$$

$$\sum_{k=1}^{n} \epsilon_k x_k = \xi$$

where N, ξ and the ε_k are strictly positive. x_k can be interpreted as the number of molecules in a certain volume possessing energy ε_k, the total number N and the total energy being prescribed. The probability of a partition x over energy levels is proportional to $\prod_{k=1}^{n} (x_k!)^{-1}$ in the unconstrained case, if the molecules do not interact; $f(x)$ is an approximation for large x_k to the logarithm of this quantity. Thus x_k^*, the Gibbs distribution, describe the most probable distribution of molecule numbers over energy levels in the limit of large N.

Suppose $\mu \in L_c^-$, then the Lagrangian function is

$$\sum_{k=1}^{n} (x_k - x_k \log x_k - u_1 x_k - u_2 \varepsilon_k x_k)$$

that has its unique maximum in $x \geq 0$ at the stationary point:

$$x_k^* = e^{-u_1 - u_2 \varepsilon_k}$$

If $\varepsilon_1 < \varepsilon_2 < \varepsilon_3 \cdots$ and $\varepsilon_1 \leq \frac{\xi}{N} < \lim \varepsilon_k$, then the multipliers u_1 and u_2 can certainly be chosen so that the two constraints are satisfied, that is x^* and u^* solve (P3). Therefore x^* solves (P1).

Example 5

The Theorem 10 cannot be applied for the problem of example 3, when $\mu \in L_c^-$ and $\bar{b} = 3$.

The homogeneous strong Lagrangian principle $(\mu \in L_c^-)$

Stronger results can be obtained by assuming $\mu \in L_c^-$, that is due to the fact that it is easier to work with scalars than with functionals.

In what follows the multiplier vector will be defined by: $u = (u_1, u_2, \cdots, u_p)$, a row vector.

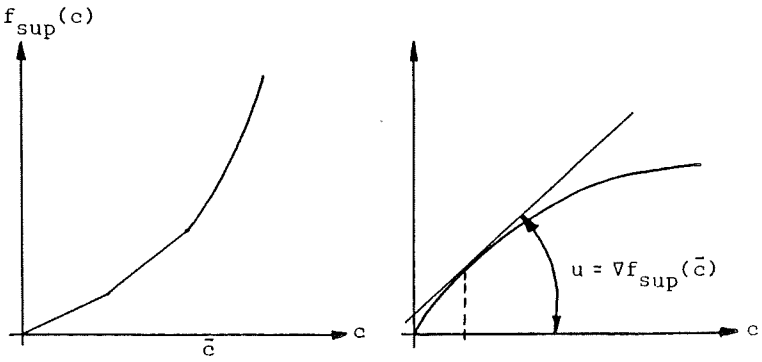
We can now stipulate the homogeneous strong Lagrangian principle:

> There exists a vector u and a non-negative scalar w such that $wf(x) - uh(x)$ is maximal for $x \in \mathcal{D}$, at x^*, the solution of $\max\{f(x) \mid x \in \mathcal{D} \wedge h(x) = \bar{c}\}$.

Now, there is a slight difference of this principle to our (P2), this is the introduction of the scalar w. This will permit us to drop the earlier assumption $u_k < \infty$, and to take account of some patological cases that might occur at the boundary of $\mathcal{B}$. The following theorem, that is proved in [3] expresses the conditions under which this principle is valid.

Theorem 11

The homogeneous strong Lagrangian principle is valid for (P1) if and only if a supporting hyperplane to the surface $(c, f_{sup}(c))$ exists at the point $(\bar{c}, f_{sup}(\bar{c}))$. The scalar may be normalized to unity if $\bar{c}$ is interior to $\mathcal{B}$, and $f_{sup}(c)$ is stable at $\bar{c}$ (e.g. it does not increase infinitely steeply in a neighbourhood of $\bar{c}$). If $\bar{c}$ is interior to $\mathcal{B}$ and the vector of derivatives $\nabla f_{sup}(c)$ exists at $\bar{c}$, then with $w = 1$, it is true that: $u = \nabla f_{sup}(\bar{c})$.



The Theorem is not applicable The Theorem is applicable

This theorem gives us more pricise characterizations of the geometrical properties of our problem at the optimal solution. The case $w = 0$ is that in which the supporting hyperplane is vertical and this can occur only if c is a boundary point of B. That is, the form of the surface $(c, f_{sup}(c))$ plays a central role for the validity of the homogeneous strong Lagrangian principle (see further our earlier discussion of Theorem 9).

Example 6

The last theorem is applicable for the problem of example 3, when $\bar{b}_1 \geq 4$, that is when a supporting hyperplane $(z_\mu = (b_1-3)u + \sqrt{2})$ exists. Moreover the Lagrange multiplier becomes

$$u_1 = \nabla f_{sup}(\bar{b}_1) = \frac{1}{\sqrt{2(b_1-2)}}, \quad \bar{b}_1 \geq 4$$

because $w = 1$. Note that the supporting hyperplane is also a tangent hyperplane in this case.

Example 7

Consider (P1): $\max f(x)$

$$h(x) = x_1 + x_2 = \bar{c} = 0$$

$$\mathcal{D}: \; x \geq 0$$

If we let $f(x) = -x_1^2$ we get $f_{sup}(c) = 0$, which has a supporting surface.

Since $f_{sup}(c)$ is stable we can solve (P1) by solving $\max(-x_1^2 - u(x_1+x_2))$. $\bar{c}$ is not interior, but we can nevertheless let $w = 1$ and $u = 0$ to yield $x^* = (0,0)'$. We observe that $\nabla f_{sup}(0) = 0$.

If we let $f(x) = x_1^2$ we cannot solve (P1) by solving $\max(x_1^2 - u(x_1+x_2))$ since no supporting surface exists. If we

let $f(x) = \sqrt{x_1}$ we can solve (P1) by solving max $w \cdot f(x) - u(x_1+x_2)$ if $w = 0$ and $u > 0$. However, u is not equal to $\nabla f_{sup}(0)$.

The characterization of the Lagrangian principle in terms of the geometric properties of the nonlinear problem in the PR-space is undoubtedly illuminating and fundamental. However, as it stands the characterization yields less information than it does insight, in that it appeals to properties of the derived quantity $f_{sup}(c)$ in the set $\mathcal{B}$, rather than of the given quantities $f(x)$, $h(x)$ and $\mathcal{D}$. Now, if $\mathcal{B}$ is convex, $f_{sup}(c)$ concave in $\mathcal{B}$ and finite at the finite value $c = \bar{c}$, a supporting plane will always exist and the conclusions of Theorem 11 are also applicable. Furthermore, if $\mathcal{D}$ is convex, $f(x)$ concave and $h(x)$ linear in x, then $\mathcal{B}$ is convex and $f_{sup}(c)$ is concave in $\mathcal{B}$, that is we have finally arrived to sufficient conditions expressed on our initial functions and the set $\mathcal{D}$ so that the homogeneous strong Lagrangian principle is applicable. It is remarkable that in order to do so one is forced to the very restrictive assumptions that $h(x)$ is linear. Later on, we will see that for the case of inequality constraints less restrictive assumptions are needed.

Note that under these sufficient conditions Theorem 1 is also applicable, but in general this is not always true.

Lagrange multipliers

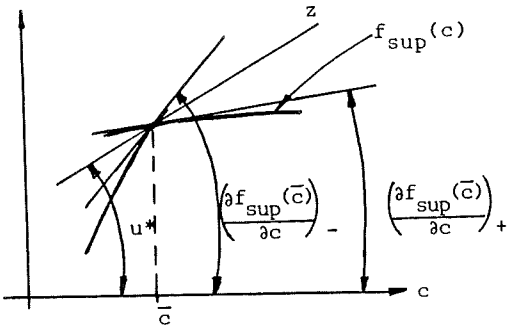
Together with the optimal solutions x^*, under some conditions, we will obtain a vector of Lagrange multipliers u^*, so that

$$u_k^* = \frac{\partial f_{sup}(\bar{c}_k)}{\partial c_k} \quad , \quad \forall k$$

Then the Lagrange multipliers at the optimal solution measure the sensivity of the optimal value of the objective function to variations in the constraint constants.

The Lagrange multipliers have an especially important interpretation in problems of economic allocation in which the objective function has the dimensions of value and the constraints specify a given value for a certain resource, then the Lagrange multiplier measures the sensitivity of a value to changes in a quantity and hence represents a price, often called in linear programming a shadow price of the resource.

In some situations the $\nabla f_{sup}(\bar{c})$ does not necessarily exist, this means that the supporting hyperplane is not tangent as shown in the figure below. At this point a new concept has to be introduced, this is the definition of a supergradient. Then we can show that $f_{sup}(c)$ is stable at $\bar{c}$ if and only if $f_{sup}(c)$ has a supergradient af $\bar{c}$.



Let us assume that the "left" and "right" derivatives exist.

Theorem 12 [4]

Suppose $\bar{c} \in \mathcal{B}$, $f_{sup}(\bar{c}) < \infty$ and $f_{sup}(c)$ stable at $\bar{c}$. Then if (x^*,u^*) gives and optimal solution to (P3), then

$$\left(\frac{\partial f_{sup}(\bar{c})}{\partial c_k}\right)_+ \leq u_k^* \leq \left(\frac{\partial f_{sup}(\bar{c})}{\partial c_k}\right)_-, \quad \forall k$$

In other words the "left" and "right" - derivatives bound the Lagrange multipliers. The two coincide if and only if $\nabla f_{sup}(\bar{c})$

exists. If both directional derivatives are finite, e.g. $f_{sup}(c)$ stable at $\bar{c}$, then the slope of the supporting hyperplane af $\bar{c}$ may adopt any intermediate value.

Example 8

The LP-problem

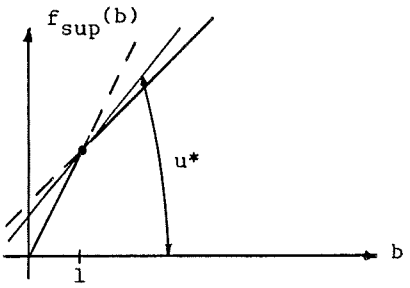
$$\max(2x_1+x_2)$$

$$x_1 + x_2 = b$$

$$0 \leq x_1 \leq 1$$

$$x_2 \geq 0$$

has the following perturbation function



for $b = 1$, the Lagrange multiplier u^* satisfies

$$1 \leq u^* \leq 2$$

2.2 Case of only inequality constraints (p = 0)

Obviously, constraints of the type $g_j(x) \leq \bar{b}_j$, can be always transformed to equality constraints by adding a slack variable. Nevertheless the case with pure inequality constraints deserves

some study because it appears so often and because it has some special properties. Let us now define the following problem:

(P5) Saddle point problem

Define the functional

$$\Phi(x,\lambda): \quad \mathcal{D} \times S \to R- = R \cup \{-\infty\}$$

as

$$\Phi(x,\lambda) = f(x) + \sum_{j=1}^{m} \left[\lambda_j(b_j) - \lambda_j(g_j(x))\right]$$

Recall that the functions $\lambda \in S$ have the property that $\lambda_j(b_j) < \infty$, but for $\epsilon \neq b_j$, $\lambda_j(\epsilon)$ can take value $+\infty$. Thus the range of $\Phi(\ )$ is the space $R-$. A saddle point of $\Phi(\ )$ is defined to be a pair (x^*,λ^*), $x^* \in \mathcal{D}$, $\lambda^* \in S$ such that $\Phi(x^*,\lambda^*)$ is finite and

$$\Phi(x,\lambda^*) \leq \Phi(x^*,\lambda^*) \leq \Phi(x^*,\lambda)$$

for $\forall x \in \mathcal{D}$, $\forall \lambda \in S$.

The saddle point problem is to find a saddle point of the functional $\Phi(x^*,\lambda^*)$.

The following results are easily proved.

Theorem 13 [1]

x^*,λ^* is a solution to (P3) if and only if x^*,λ^* is a solution to (P5).

Thus saddle points are equivalent to solutions of the constrained Lagrangian problem.

Corollary

A saddle point provides a solution to (P1).

Corollary

Suppose x^* is a solution to (P1) and that $f_{sup}(b)$ has linear support at $(\bar{b}, f_{sup}(\bar{b}))$. Then there exists a $\lambda^* \in L_{\bar{b}}$ such that (x^*, λ^*) is a saddle point.

It can easily be shown that the last two corollaries taken together, state the Kuhn-Tucker [5] equivalence theorem, but under weaker assumptions. One possible statement of this theorem is the following: If f and g are differentiable on the non-negative orthant, f concave, g_j convex, $\forall j$, and if the feasible set has an interior point then x^* is a solution to (P1) if and only if there exists $\lambda^* \in L_{\bar{b}}$, such that (x^*, λ^*) is a saddle point. Thus the sufficiency is implied by the first corollary. Concerning the necessity, it is easy to show that the above assumptions imply the existence of a linear sup-or, the assumption of the second corollary.

If λ is not restricted to $L_{\bar{b}}$, it can be shown that there is always a $\lambda^* \in M$ which solves (P4), provided $\bar{b} \in B$. Then

Theorem 14

x^* solves (P1), if and only if there exists a $\lambda^* \in M$ such that (x^*, λ^*) is a saddle point of $\Phi(\ )$.

Example 9

The problem in example 3, for $\bar{b} = 3$, has a saddle point at:

$$x^* = \sqrt{2}$$

$$\lambda^* = \sqrt{2\varepsilon^2} - 5.5\sqrt{2\varepsilon}$$

When $\lambda \in L_{\bar{b}}$, then $\Phi(\ )$ becomes

$$L(x,y) = f(x) + \sum_{j=1}^{m} y_j(\bar{b}_j - g_j(x))$$

and $y_j \geq 0$, $\forall j$. This is the so-called Lagrangian form. Note that the expression $f(x) - y\,g(x)$ that we have considered above, is different from the Lagrangian form by a quantity $(y\bar{b})$. Now, a saddle point is defined as (x^*,y^*).

Theorem 13 gives us a clue to identify saddle points to $L(x,y)$, thus

Theorem 15 [6]

Let $y^* \geq 0$ and $x^* \in \mathcal{D}$. Then (x^*,y^*) is a saddle point if and only if

(i) x^* maximizes $L(x,y^*)$ over $\mathcal{D}$

(ii) $g_j(x^*) \leq \bar{b}_j$, $\forall j$

(iii) $y_j^*(\bar{b}_j - g_j(x^*)) = 0$, $\forall j$

Then we can conclude that the existence of a saddle point to the Lagrangian form guarantees the existence of a supporting hyperplane. Then existence of saddle points, guarantee also the validity of the homogeneous strong Lagrangian principle. This can be shown by transforming our problem

$$\max\{f(x) \mid g(x) \leq \bar{b}\}$$

$$x \in \mathcal{D}$$

to one with equality constraints

$$\max\{f(x) \mid g(x) + q = \bar{b}\}$$

$$x \in \mathcal{D}$$

$$q \geq 0$$

and since a saddle point guarantees a supporting plane of $f_{sup}(b)$ at $\bar{b}$, then the homogeneous strong Lagrangian principle is also valid. Then the results and discussion around Theorem 11 are also valid. A very important property of the problems with inequality constraints is the fact that

$$y_j^* \geq 0 \quad \text{and} \quad y_j^* \, q_j^* = 0, \ \forall j \quad \text{(complementary slackness)}$$

Thus if $q_j^* > 0$, that is the $j^{\underline{th}}$ restriction is <u>not active</u> then $y_j^* = \dfrac{\partial f_{sup}(\bar{b})}{\partial b_j} = 0$; and if $y_j^* > 0$, then $q_j^* = 0$, that is the $j^{\underline{th}}$ restriction is <u>active</u>.

If $\mathcal{D}$ is convex, $f(x)$ concave and $g(x)$ convex, then $\mathcal{B}$ is convex and $f_{sup}(b)$ is concave in $\mathcal{B}$, then these are <u>sufficient conditions</u> for the existence of Lagrange multipliers thus concavity of $f_{sup}(b)$ is the principal sufficient condition required for the existence of Lagrange multipliers and essentially it is also necessary if the strong Lagrange principle is to hold for all b in $\mathcal{B}$.

<u>Example 10</u>

Consider (P1): $\max - (x-4)^2$

$$x \leq 3$$

$$\mathcal{D}: \ x \geq 0$$

If we let $\lambda(\varepsilon) = y \cdot \varepsilon$, that is, λ is linear, we get $\Phi(x,\lambda) = \Phi(x,y) = L(x,y) = -(x-4)^2 + y(3-x)$. Now, if we let $y = 2$ we find $L = -(x-4)^2 - 2x + 6$, which has a global maximum at $x = 3$. Therefore $L(x,2) \leq L(3,2)$. If we let $x = 3$ we

find $L = -1$. Therefore $L(3,2) \leq L(3,y)$. and we conclude that $(x^*,y^*) = (3,2)$ is a saddle point.

We see that

(i) $x^* = 3$ maximizes $-(x-4)^2 - 2x$ and $-(x-4)^2 - 2x + 6$

(ii) $g(3) = 3 \leq 3$

(iii) $2 \circ g(3) = 2 \circ 3 = 2 \circ 3$

That is (3,2) is a solution to (P3) (Theorem 13 and essentially also Theorem 15).

It is also seen that $x^* = 3$ is a solution to (P1) (Corollary 1). Further, since $f(x)$ is concave and $g(x)$ is linear, a linear support exists at $(\bar{b}, f_{sup}(\bar{b}))$. Therefore a y^* exists, such that $(3,y^*)$ is a saddle point (Corollary 2).

We note that the restriction is active and $y^* > 0$, so that $y^*(3-x^*) = 0$.

y^* is the shadow price. If therefore $\bar{b}$ is increased by ε, $f(x^*)$ will be increased by $y^* \cdot \varepsilon = 2\varepsilon$. (Or rather, this is the limit for $\varepsilon \to 0$). If for instance we let $\varepsilon = 0.1$ we expect that $f(x^*) \approx -1 + 0.2$. Note that we cannot find x^* directly from our knowledge of y^*, but have to solve (P1) again with $\bar{b} = 3.1$.

Duality theory $(\lambda \in L_{\bar{b}})$.

Let us now see some results that are useful to find saddle points to Lagrangian forms. By definition if $(x^* \in \mathcal{D} \wedge y^* \geq 0)$ is a saddle point, then

$$L(x,y^*) \leq L(x^*,y^*) \leq L(x^*,y)$$

The last inequalities can be interpreted as

$$L(x^*,y^*) = \max_{x\in\mathcal{D}}[\min_{y\geq 0} L(x,y)] = \max_{x\in\mathcal{D}} L_*(x)$$

$$= \min_{y\geq 0}[\max_{x\in\mathcal{D}} L(x,y)] = \min_{y\geq 0} L^*(y)$$

We have then two problems. The primal problem (P):

(P) Find x such that: $L_*(x^*) = \max_{x\in\mathcal{D}} L_*(x)$

correspondingly, the dual problem (D):

(D) Find y such that: $L^*(y^*) = \min_{y\geq 0} L^*(y)$

It is not difficult to prove that (P) is equivalent to (P1) [10]. Problems (P) and (D) always have optimal values (possibly $\mp\infty$) whether or not they have optimal solutions provided we invoke the customary convention that a supremum (infimum) taken over an empty set is $-\infty(+\infty)$. The function $L^*(y)$ is usually denominated as dual or Legendre function and its domain of definition is

$$\mathcal{Y} = \{y \mid y \geq 0\,,\ \max_{x\in\mathcal{D}} L(x,y) \text{ exists}\}$$

If for all $y \geq 0$, $L(x,y)$ is continuous in x for all $x \in \mathcal{D}$ and if $\mathcal{D}$ is compact, then

$$\mathcal{Y} = (R^m)_+ = \{y \mid y \geq 0\}$$

However, in general, $\mathcal{Y}$ need not be convex and may be empty.

Let us now see some useful results that are easily proved, see for the proofs [7].

Theorem 16 - (Weak duality)

If x is feasible in (P) and y is feasible in (D), then

$$L_*(x) \leq L^*(y)$$

This theorem holds even if the Lagrangian form does not have a saddle point. One consequence of this theorem is that any feasible solution of (D) provides an upper bound on the optimal value of (P); and any feasible solution of (P) provides a lower bound on the optimal value of (D). If the feasible regions of both primal and dual are non-empty, then the last theorem implies that $\sup L_*(x)$ and $\inf L^*(y)$, both taken over their respective constraint sets, are finite. Of course, neither the supremum nor the infimum need be attained at any feasible point. If they are and if both primal and dual are feasible, both have optimal solutions. The following results can also be easily proved:

(i) If $\inf\{L^*(y) \mid y \in Y\} = -\infty$, then the primal is infeasible.

(ii) If $\sup\{f(x) \mid x \text{ feasible}\} = +\infty$, then the dual is infeasible.

(iii) If there exist primal feasible x^* and dual feasible y^* such that $L_*(x^*) = L^*(y^*)$, then x^* solves the primal and y^* solves the dual.

Theorem 17 [7]

The dual function $L^*(y)$ is convex over any convex subset of its domain.

Then if Y is convex, the dual program requires the minimization of a convex function over a convex set, and thus has no local minima distinct from the global one. In particular, if f and g are continuous on $\mathcal{D}$ and if $\mathcal{D}$ is compact, then $\mathcal{D} = (R^m)_+$, which is convex Thus the dual problem may, in general, be made well behaved. In that case, the only question which remains is whether or not

$$\min_{y \geq 0} L^*(y) = \max_{x \in \mathcal{D}} L_*(x) .$$ The answer

is given by the following theorem.

Theorem 18 [6] (Strong duality)

(x^*,y^*) is a saddle point for $L(x,y)$ if and only if x^* is primal feasible, y^* is dual feasible, and $L_*(x^*) = L^*(y^*)$.

Example 11

Consider (P1): $\max\ x_1^2 + 3x_2$

$$x_1^2 - x_2 \leq 0$$

$$x_2 \leq 4$$

$$\mathcal{D}: \quad x \geq 0$$

It is easy to show that $x^* = (2,4)'$, $f(x^*) = 16$ and, since (P1) = P, $L_*(x^*) = f(x^*) = 16$.

Sufficient conditions for the existence of a saddle point are that $\mathcal{D}$ is convex, $f(x)$ is concave and $g(x)$ is convex. These are not satisfied here. So to find out whether a saddle point actually exists we could either see if $f_{sup}(b)$ has a linear support at $\bar{b} = (0,4)'$, or solve the dual and see if $L_*(x^*) = L^*(y^*)$ (Theorem 18). We will here solve the dual. We have

$$L(x,y) = x_1^2 + 3x_2 + y_1(-x_1^2 + x_2) + y_2(4 - x_2)$$

First we find $\max_x L(x,y)$. Since

$$\frac{\partial L}{\partial x} = (2x_1(1-y_1),\ 3 + y_1 - y_2)$$

we conclude that

if $1-y_1 < 0$ then $x_1^* = 0$ and $L^*(y) = 4y_2 + (3+y_1-y_2)x_2$

if $1-y_1 = 0$ then x_1^* is arbitrary and $L^*(y) = 4y_2+(3+y_1-y_2)x_2$

if $1-y_1 > 0$ then no maximum exists

if $3+y_1-y_2 < 0$ then $x_2^* = 0$ and $L^*(y) = 4y_2 + x_1^2(1-y_1)$

if $3+y_1-y_2 = 0$ then x_2^* is arbitrary and $L^*(y) = 4y_2 + x_1^2(1-y_1)$

if $3+y_1-y_2 > 0$ then no maximum exists.

From this we conclude that (study all combinations):

$$Y = \{(y_1,y_2) \mid y_1 \geq 0 \wedge y_2 \geq 0 \wedge y_1 \geq 1 \wedge -y_1+y_2 \geq 3\}$$

and $L^*(y) = 4y_2$

Second we find $\min L^*(y)$, $y \in Y$. This is a LP-problem with $y^* = (1,4)$ which yields $L^*(y^*) = 16$. Since $L^*(y^*) = 16 = L_*(x^*)$, a saddle point exists.

(Actually (P1) is a linear programming problem, therefore the dual is also a linear programming problem).

Example 12

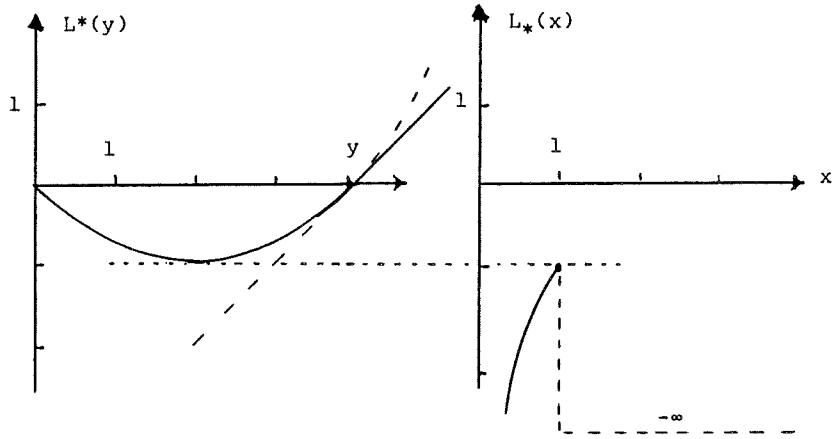
Consider (P1): $\max f(x) = -(x-2)^2$

$x \leq 1$

$\mathcal{D}$: $x \geq 0$

Form this we get $L(x,y) = -(x-2)^2 + y(1-x)$. Since $(P) = (P1)$, $f(x) = L_*(x)$ if x is feasible, and moreover $L_*(x) = -\infty$ when x is infeasible. It is easily seen that $x^* = 1$. The dual problem is $\min \max L(x,y)$. We find $\frac{\partial L}{\partial x} = -2(x-2) - y = 0 \Leftrightarrow x = 2 - \frac{y}{2}$. Since we find a maximum for all y, $Y = \{y \mid y \geq 0\}$. Therefore $x^* = 2 - \frac{y}{2}$ if $y \leq 4$, $x^* = 0$ otherwise. Consequently $L^*(y) = -\left(\frac{y}{2}\right)^2 + y\left(\frac{y}{2} - 1\right) = \frac{y^2}{4} - y$ if $y \leq 4$, $L^*(y) = -4 + y$ otherwise. Minimum of this is attained for $y^* = 2$ yielding $L^*(y^*) = -1$.

If we draw $L^*(y)$ and $L_*(x)$ it looks like this:



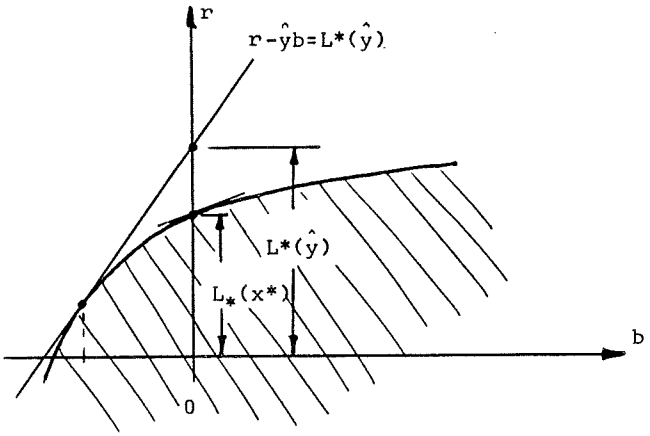
We see that $L_*(x) \leq L^*(y)$ (Theorem 16).

We see that $L_*(x^*) = L^*(y^*)$, thus we have a saddle point (Theorem 18).

We see that $L^*(y)$ is convex (Theorem 17).

The natural immediate questions are: Can the dual problem be given some intuitive meaning? What interpretation can be attached to $L(x,y)$ when $L_*(x^*) < L^*(y^*)$? When Lagrangian methods fail?

Let us again consider the PR-space and for simplicity suppose $\bar{b} = 0$.



Theorem 19 [6]

Let $x^* \in \mathcal{V} \wedge y^* \geq 0$. Then x^* maximizes $L_*(x)$ over $\mathcal{V}$ if and only if

$$H = \{(r,b) \mid r - y^*b = L(x^*,y^*)\}$$

is a supporting plane of $f_{sup}(b)$ at $\bar{b} = 0$ and $f_{sup}(0)$.

According to the above results, the process of max $L(x,y^*)$ over $\mathcal{V}$ is equivalent to finding a real number

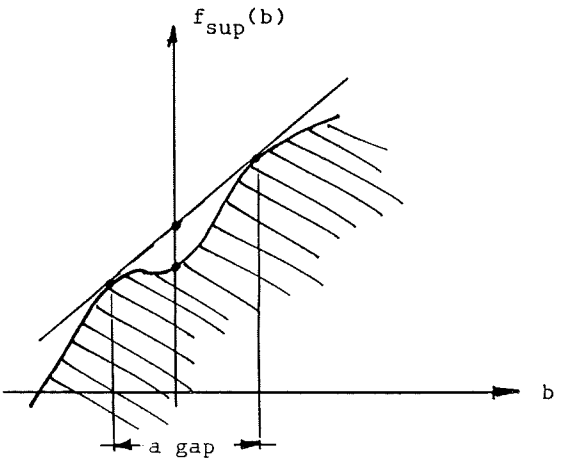
$$L^*(y^*) = L(x^*,y^*)$$

such that the hyperplane H is a supporting plane to $f_{sup}(b)$ at $(0,f_{sup}(0))$. The value of the dual function is the intercept of this support plane with the r axis, and the slopes of the plane are the numbers y^*. The dual problem is that of finding the support plane with positive slopes having minimal r intercept, while the primal is that of finding a point $(r,0)$ on $(0,f_{sup}(0))$ with maximal intercept, the number $f_{sup}(0)$. Since $f_{sup}(b)$ is nondecreasing, all support planes have slopes $y \geq 0$. It is evident from the figure that such plane has r intercept bigger than or equal to $f_{sup}(0)$, or Theorem 16 is verified. If $f_{sup}(b)$ is concave, supporting planes exists always, in particular to $(0,f_{sup}(0))$, yielding Theorem 19.

Linear supporting planes will not exist if $\bar{b}$ is in an interval gap, where obviously

$$L^*(y^*) > L_*(x^*) \qquad \text{(Theorem 16)}$$

This is illustrated below.



A non-linear duality theory has been elaborated in [8].

2.3 Final remarks

In this section we have gone through Everett's theory, and for the general case, that is for the case of not necessarily linear Lagrange functions, we have seen that:

a) In this section all functions are not necessarily differentiable or continuous, the point is that it is sufficient to find a saddle point to obtain a solution to (P1).

b) Does a saddle point always exist? In principle yes, the problem is just to find a suitable Lagrange function. Numerical methods based on this idea will be discussed in the next chapter.

c) The results obtained for ($m = 0$) can easily be modified so to be applicable for ($p = 0$) and viceversa, that is due to the fact that a problem with equality constraints can be transformed to a problem with inequality constraints and viceversa.

d) We have the following relationship among the solution of the above formulated problems:

$$(P5) \Leftrightarrow (P3) \Leftrightarrow ((P1) \wedge (P4))$$

In practice, linear Lagrange functions are usually employed under such conditions we have seen that:

a) It is not sure that a saddle point will exist, then the strong duality theorem is not applicable and a linear support to the perturbation function cannot be found, we have a gap.

b) If we are dealing with a concave programming problem where a solution exists then a saddle point will also exist, as well as a linear support to the perturbation function. If this linear support is also a tangent to the perturbation function, then under mild restrictions the Lagrange multiplier will be uniquely defined.

c) The results obtained can also be applied to the classic linear programming problem to obtain a clear understanding of its geometrical properties, but the results cannot be applied to find numerical solutions through (P2) because for some regions the Lagrange multipliers are constant (see further chapter 4, case 7).

d) It is extremely important to have in mind the way the concept of Lagrange multipliers were obtained in this section because in the following section another way to obtain Lagrange multipliers will be presented. Most introductory books on Mathematical Programing make no explicit difference between these two, in principle, different approaches.

3. NECESSARY CONDITIONS FOR OPTIMAL SOLUTIONS - KUHN AND TUCKER THEORY

In chapter 1 we developed necessary conditions for an optimal solution when x^* was an interior point of X.

We proved that when x^* was on the boundary of X, then necessarily: $\nabla f(x^*)\, s \leq 0$, for all feasible directions s from x^*. To obtain further results we have to specify the region of feasible directions and to do so we need a more precise definition of the feasible set X. We will assume throughout this section that the functions f, g and h are once differentiable functions.

A geometrical interpretation

Let us first consider the following example:

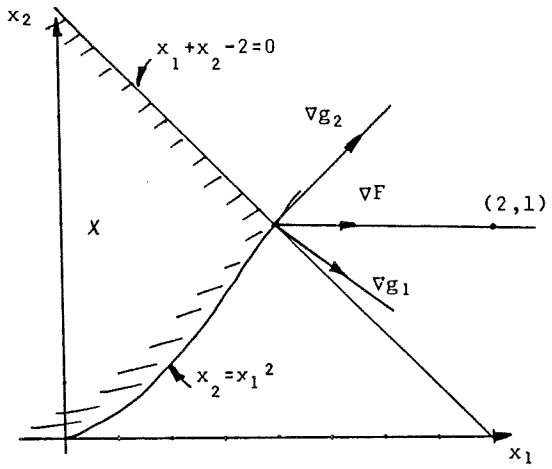
$$\max F = -(x_1-2)^2 - (x_2-1)^2$$

$$x \in X$$

where: $X = \{(x_1,x_2) \mid g_1 = +x_1^2 - x_2 \leq 0$, and

$g_2 = x_1 + x_2 - 2 \leq 0\}$

This problem is graphically represented below:



It is evident that the optimum is at the intersection of the 2 constraints at (1,1). We have defined a feasible direction as a vector s , such that a small move along that vector violates no constraints. At (1,1) the set of feasible directions lies between the line $x_1 + x_2 - 2 = 0$ and the tangent line to $x_2 = x_1^2$ at (1,1), e.g. $x_2 = 2x_1 - 1$. In other words this set is the cone generated by these lines. The vector ∇F points in the direction of maximum rate of increase of F and a small move along any direction making an angle of less than 90^0 with ∇F will increase F . Thus, at the optimum, necessarily, no feasible direction can have an angle of less than 90^0 between it and ∇F . In the last figure, the gradient vectors ∇g_1 and ∇g_2 are drawn. Note that the set of feasible directions can

also be defined as: $\{s \mid \nabla g_1\, s \leq 0$ and $\nabla g_2\, s \leq 0\}$. Moreover note also that ∇F is contained in the cone generated by ∇g_1 and ∇g_2. This is necessarily true at an optimal solution. If ∇F where slightly above ∇g_2, it would make an angle of less than 90^o with a feasible direction just below the line $x_1 + x_2 - 2 = 0$. If ∇F were sliggtly below ∇g_1 it would ake an angle of less than 90^o with a feasible direction just above the line $x_2 = 2x_1 - 1$. Neither case can occur at an optimal point, and both cases are excluded if and only if ∇F lies within the cone generated by ∇g_1 and ∇g_2. This is the usual statement of the so-called Kuhn-Tucker necessary conditions, i.e. at x^*, ∇F lies within the cone generated by the gradients of the active constraints. Now for ∇F to lie within the cone above described, it must be a nonnegative linear combination of the gradients of the active constraints, or symbolically there must exist numbers λ_j^*, so that:

$$\nabla F = \sum_{j \in A} \lambda_j^* \nabla g_j(x^*)$$

where: $\lambda_j^* > 0$, $\forall j \in A$

and A is the set of indices of the active constraints. This result may be restated to include all constraints by defining:

$$\lambda_j^* = 0 \quad \text{if} \quad g_j(x^*) < 0$$

Then the Kuhn-Tucker conditions for a problem $\max\{f(x) \mid g(x) \leq 0\}$ take the form

$$\nabla f(x^*) = \sum_{j=1}^{m} \lambda_j^* \nabla g_j(x^*)$$

$$\lambda_j^* \geq 0, \quad \lambda_j^* g_j(x^*) = 0, \quad g_j(x^*) \leq 0, \quad \forall j$$

Do the λ_j's satisfying the last conditions always exist? Let us see an example:

Example 13 [5]

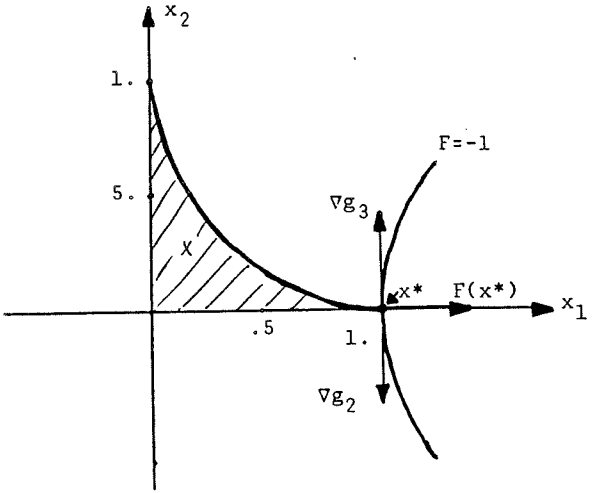
Consider the problem

$$\max F = -(x_1 - 2)^2 - x_2^2$$

$$x \in X$$

where: $X = \{x \mid -x_1 \leq 0 , -x_2 \leq 0 , x_2 - (1 - x_1)^3 \leq 0\}$

The figure below illustrates this problem.



From the figure $x^* = (1,0)'$, $\nabla F(x^*) = (2,0)$, $\nabla g_2(x^*) = (0,-1)$, $\nabla g_3(x^*) = (0,1)$, and $A = \{2,3\}$. We can see that it is impossible to define ∇F as a nonnegative linear combination of ∇g_2 and ∇g_3 .

The Fritz-John theorem [9] is a more general version of the necessary conditions and permits to take account of such pathological examples. The necessary conditions are modified to

$$\omega^* \nabla f(x^*) = \sum_{j=1}^{m} \lambda_j \nabla g_j(x^*)$$

$$\omega^* , \ \lambda_j^* \geq 0 , \ \lambda_j^* g_j(x^*) = 0 , \ g_j(x^*) \leq 0 , \ \forall j$$

and not all the Lagrange multipliers (ω^* , λ_j^*) equal to zero.

Thus for example 13 , $\omega^* = 0$ and $\lambda_2^* = \lambda_3^*$.

Then Kuhn-Tucker conditions will be necessary at an optimal solution only when $\omega^* \neq 0$, if that is so we can always put $\omega^* = 1$. The additional assumptions that will assure $\omega^* \neq 0$, are usually denominated as constraint qualifications or regularity conditions. Now if s is admissible, then $\nabla g_j(x^*)\, s \leq 0$, $\forall j \in A$, however the converse is not true without some additional conditions. Thus for our example 13: $\nabla g_2 \; s = -s_2 \leq 0$

$$g_3 \; s = s_2 \leq 0 .$$

The vector $(s_1,s_2)' = (1,0)'$ satisfies these inequalities, yet points out of the feasible set. This means we have to strengthen our definition of feasible directions, now as at x^* necessarily: $\nabla f(x^*)\, s \leq 0$, define

$$K_2^* = \{ s \mid \nabla g_j(x^*)\, s \leq 0 , \ \forall j \in A , \ \nabla f(x^*)\, s > 0 \}$$

Then at the optimal solution, necessarily, $K_2^* = \emptyset$. This can be shown to be a necessary and sufficient condition to set $\omega^* = 1$, that is for the multipliers λ_j^* to exist.

3.1 Generalized Kuhn-Tucker theorem

Let us consider (P1) with $\mathcal{D} = R^n$. The rather intuitive results we have obtained are special cases of the following theorem.

Theorem 20 [10]

(i) If x^* is feasible,

(ii) f, g, and h are once differentiable, and

(iii) at x^* the set $K_2^* = \emptyset$, where

$$K_2^* = \{s \mid \nabla g_j(x^*)\, s \leq 0,\ \forall j \in A;\ \nabla h_k\, s = 0,\ \forall k;\ \nabla f(x^*)\, s > 0\}$$

Then there exist row vectors λ^* and μ^* such that (x^*), (λ^*, μ^*) satisfies

$$\nabla f(x^*) = \sum_{j=1}^{m} \lambda_j^* \nabla g_j(x^*) + \sum_{k=1}^{p} \mu_k^* \nabla h_k(x^*)$$

$$g(x^*) \leq \bar{b}$$

$$h(x^*) = \bar{c}$$

$$\left.\begin{aligned} &\lambda_j^*(\bar{b}_j - g_j(x^*)) = 0 \\ &\lambda_j^* \geq 0' \end{aligned}\right\} \forall j$$

This theorem can be easily proved by applying directly Farkas lemma (see appendix). From the proof we can conclude that $K_2^* = \emptyset$ is a necessary and sufficient condition for the existence of the Lagrange multipliers satisfying Kuhn-Tucker necessary conditions for an optimal solution. Thus the Kuhn-Tucker conditions are the precise mathematical formulation of the following rough but intuitive concept: at any constrained optimum, no (small) feasible change in the decision variables can improve the objective function.

Compare Theorem 20 with Theorem 15 and see that the first one provides necessary conditions for a saddle point under the differentiability assumption.

Constraint qualifications

The condition $K_2^* = \emptyset$ has the serious defect, with regard to numerical applications, that it is not generally possible to verify computationally.

Some of the best known sufficient conditions guaranteeing $K_2^* = \emptyset$ are [6, 10]:

(i) All constraint functions are linear.

(ii) The g functions are convex, the h functions are linear, and the feasible set has a nonempty interior.

(iii) The gradients $\nabla g_j(x^*)$, $\forall j \in A$, and $\nabla h_k(x^*)$, $\forall k$, are linearly independent.

(iv) The Kuhn-Tucker constraint qualification, i.e. every feasible direction is tangent to a once differentiable arc, the arc emanating from x^* and contained in the constrained set [5].

A more complete, but mathematically more advanced, discussion around constraint qualifications can be found in [11].

Finally a fact worth noting is that the condition $K_2^* = \emptyset$ is not necessary at an optimal solution. Example 13 is a good illustration of this fact.

Kuhn-Tucker conditions

Care should be taken when stipulating the Kuhn-Tucker conditions because it is too easy to make mistakes. For our general problem the Lagrangian form is defined as:

$$L(x,\lambda,\mu) = f(x) + \lambda(\bar{b} - g(x)) + \mu(\bar{c} - h(x))$$

Then the Kuhn-Tucker conditions can be stipulated as:

$$\max f(x) \qquad g(x) \leq \bar{b} \qquad h(x) = \bar{c}$$

$$x \in R^n$$

$$\downarrow \qquad\qquad \downarrow \qquad\qquad \downarrow$$

$$\frac{\partial L(x^*,\lambda^*,\mu^*)}{\partial x} = 0' \qquad \frac{\partial L(x^*,\lambda^*,\mu^*)}{\partial \lambda} \geq 0 \qquad \frac{\partial L(x^*,\lambda^*,\mu^*)}{\partial \mu} = 0$$

$$\lambda^* \frac{\partial L(x^*,\lambda^*,\mu^*)}{\partial \lambda} = 0$$

$$\lambda^* \geq 0'$$

Very often we will have problems with restrictions: $x \geq 0$, then some extra relations have to be introduced. For the simple case:

$$\max f(x)$$

$$x \geq 0 \ ,$$

$$L(x,\lambda) = f(x) + \lambda x$$

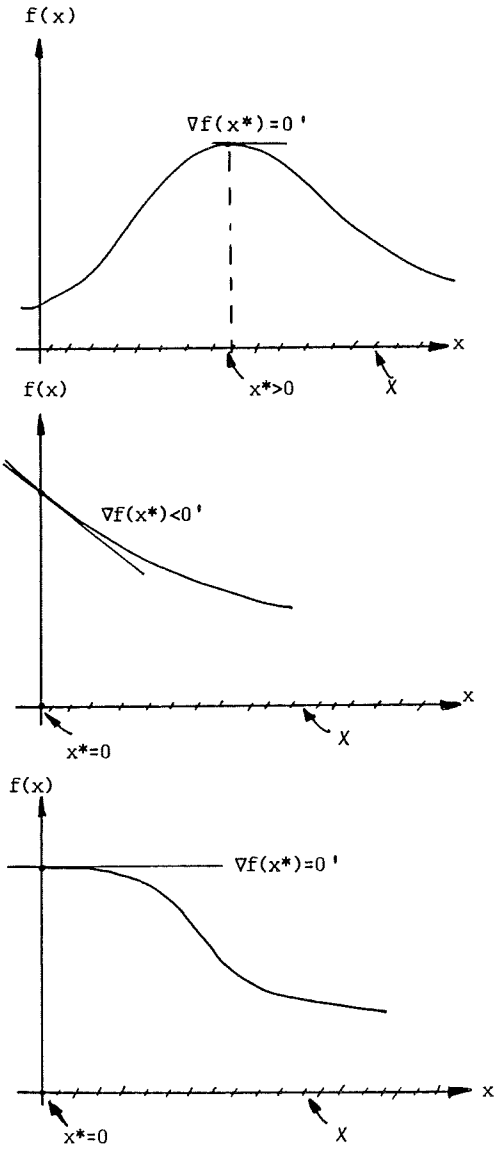
The Kuhn-Tucker conditions become

$$\nabla f(x^*) \leq 0'$$

$$\nabla f(x^*)x^* = 0$$

$$x^* \geq 0$$

The alternative possible solutions to the problem in the one dimensional case are illustrated below: an interior solution at which the slope is zero, a boundary solution at which the slope is negative, or a boundary solution at which the slope is zero.


f(x)
∇f(x*)=0'
x
x*>0
X
f(x)
∇f(x*)<0'
x
x*=0
X
f(x)
∇f(x*)=0'
x
x*=0
X

Example 14

It is easily proved that if instead of having $x \in R^n$, we have $x \geq 0$, then the condition $\frac{\partial L(x^*,\lambda^*,\mu^*)}{\partial x} = 0'$ becomes

$$\frac{\partial L(x^*,\lambda^*,\mu^*)}{\partial x} \leq 0'$$

$$\frac{\partial L(x^*,\lambda^*,\mu^*)}{\partial x} \cdot x^* = 0$$

$$x^* \geq 0$$

Now, we are in conditions to suggest a systematic way to stipulate the Kuhn-Tucker conditions for any problem.

The following table describes the conditions that have to be satisfied for an optimization problem under constraints

		equalities	inequalities
opt.f(x) x	$x \in R^n$	$\frac{\partial L}{\partial x} = 0'$ $\frac{\partial L}{\partial \mu} = 0$	$\lambda \frac{\partial L}{\partial \lambda} = 0$ $\frac{\partial L}{\partial x} = 0'$ $\lambda \geq 0'$
	$x \geq 0$	$\frac{\partial L}{\partial x}_{x=0}$ $\frac{\partial L}{\partial \mu} = 0'$ $x \geq 0$	$\frac{\partial L}{\partial x}_{x=0}$ $\lambda \frac{\partial L}{\partial \lambda} = 0$ $x \geq 0$ $\lambda \geq 0'$

To these conditions we have to add some conditions that depend on whether we want to max or min $f(x)$, and the type of inequalities on hand.

	$g(x) \leq \bar{b}$	$g(x) \geq \bar{b}$	$h(x) = \bar{c}$
max f(x)	$L = f(x) + \lambda(\bar{b}-g(x))$ $\frac{\partial L}{\partial x} \leq 0'$ $\frac{\partial L}{\partial \lambda} \geq 0$	$L = f(x) - \lambda(\bar{b}-g(x))$ $\frac{\partial L}{\partial x} \leq 0'$ $\frac{\partial L}{\partial \lambda} \geq 0$	$\frac{\partial L}{\partial x} \leq 0'$
min f(x)	$L = f(x) - \lambda(\bar{b}-g(x))$ $\frac{\partial L}{\partial x} \geq 0'$ $\frac{\partial L}{\partial \lambda} \leq 0$	$L = f(x) + \lambda(\bar{b}-g(x))$ $\frac{\partial L}{\partial x} \geq 0'$ $\frac{\partial L}{\partial \lambda} \leq 0$	$\frac{\partial L}{\partial x} \geq 0'$

In the last table for the case of equality constraints the Lagrangian form can be defined as

$$L(x,\mu) = f(x) - \mu(\bar{c} - h(x)) \quad \text{or}$$

$$L(x,\mu) = f(x) + \mu(\bar{c} - h(x))$$

That is so because the μ are not restricted in sign.

Example 15

Kuhn-Tucker conditions need not be satisfied at an optimal solution, example 3 and example 8 show this. Moreover the Kuhn-Tucker conditions are, by themselves, not sufficient for a point to be optimal. Let us for instance consider the problem:

$$\max_{x \in R} x^2$$
$$-1 \leq x \leq 2$$

here it is easily verify that the Kuhn-Tucker conditions are satisfied for x=-1, while the optimal solution is $x^*=2$. x=-1 is a local maximum. Another point that satisfy Kuhn-Tucker conditions is x=0, this is a global minimum.

3.2 Second-order necessary conditions

It must be recalled that Kuhn-Tucker conditions can identify points that are not optimal, even locally, as they are necessary conditions. That is Kuhn-Tucker conditions are just a generalization of the stationary condition for the unrestricted case. For the unrestricted case second-order conditions were developed in theorem 5, similar conditions can be developed for the restricted problem.

In [10] several necessary and sufficient conditions for a local maximum are stipulated. Essentially all the analyses are based on the study of the Hessian matrix of the Lagrangian form.

3.3 Sufficiency of the Kuhn-Tucker conditions

Sufficient conditions can be found by essaying to satisfy the assumptions of Theorem 1. Thus for our general problem if

(i) f is concave, g are convex and h linear functions, and

(ii) the constraint qualification is satisfied,

the x^* solves our problem if and only if there exist Lagrange multipliers such that the Kuhn-Tucker conditions are satisfied.

For concave differentiable problems if the Kuhn-Tucker conditions are satisfied at a point, this is an optimal solution. Linear programming problems are examples where this result works.

A more general sufficiency property is given by the following theorem.

Theorem 21 [12]

(i) f, g, h are once differentiable, and

(ii) f is pseudoconcave, g_j are quasi-convex and the h_k are linear

(iii) x^* satisfies the Kuhn-Tucker conditions.

Then x^* is optimal for the nonlinear programming problem (P1) with $\mathcal{D} = R^n$ or $\mathcal{D} = \{x \mid x \geq 0\}$.

Example 16

Consider $\max -(x_1-2)^2 + x_2$

$$x_1 + x_2 \leq 2$$

$$x_2 \leq 4$$

$$\mathcal{D}: \quad x \geq 0$$

Here $L(x,\lambda) = -(x_1-2)^2 + x_2 + \lambda_1(2-x_1-x_2) + \lambda_2(4-x_2)$. And since all constraint functions are linear, if a x^* exists, then there exists a λ^* that satisfies.

$$\frac{\partial L}{\partial x} = (-2(x_1-2) - \lambda_1 \, , 1 - \lambda_1 - \lambda_2) \leq (0,0)$$

$$\frac{\partial L}{\partial x} x = (-2(x_1-2) - \lambda_1)x_1 + (1 - \lambda_1 - \lambda_2)x_2 = 0$$

$$x \geq 0$$

$$\frac{\partial L}{\partial \lambda} = \begin{pmatrix} 2 - x_1 - x_2 \\ 4 - x_2 \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\lambda \frac{\partial L}{\partial \lambda} = \lambda_1(2 - x_1 - x_2) + \lambda_2(4 - x_2) = 0$$

$$\lambda \geq 0'$$

Moreover, since $f(x)$ and $g(x)$ are differentiable, $f(x)$ is pseudoconcave and $g(x)$ is quasi-convex, then if there is a x^*, it is optimal.

It is easily verified that $(x^{*\prime},\lambda^*) = (1\frac{1}{2}, \frac{1}{2}, 1, 0)$ satisfies the conditions.

3.4 The homogeneous weak Lagrangian principle

Here we are interested in discussing under which conditions the optimal solution to (P1), that satisfies the Kuhn-Tucker conditions, provides stationary solutions to (P2). For the sake of simplicity let us consider the following problem

$$(P6): \quad \max_{x \in \mathcal{D}} f(x)$$

$$h(x) = \bar{c}$$

The homogeneous weak Lagrangian principle says:

> There exists a row vector μ and a non-negative scalar ω such that
>
> $$(\omega \nabla f - \mu \nabla h(x))s \leq 0$$
>
> holds at x^*, the solution to (P6), for all feasible directions s from x^*. In particular, if x^* is interior to $\mathcal{D}$, then $\omega f(x) - \mu h(x)$ is stationary af x^*.

The validity of this principle can be stipulated by turning back to our PR space that will provide a geometrical interpretation of this principle.

Theorem 22 [3]

(i) Suppose a tangent hyperplane to $f_{sup}(c)$ exists at $(\bar{c}, f_{sup}(\bar{c}))$, and that the derivatives of f and h exist at x^*. Then the homogeneous weak Lagrangian principle is applicable.

(ii) If $\nabla f_{sup}(\bar{c})$ exists then one can normalize ω to unity, and necessarily

$$\mu = \nabla f_{sup}(\bar{c})$$

Note that the theorem makes no converse statement, to the effect that validity of the weak Lagrangian principle would imply the existence of a tangent hyperplane at $(\bar{c}, f_{sup}(\bar{c}))$. Such an assertion can be made if it is assumed that $x^*(c)$ is differentiable in c.

Sufficient conditions are given by the following result.

Theorem 23 [3]

The homogeneous weak Lagrangian principle is valid if the maximizing point x^* is interior to $\mathcal{D}$, and ∇f and ∇h exist at this point. The coefficient ω may be normalized to unity if either of the following conditions is fulfilled in addition:

(i) h is linear, or

(ii) ∇h is of full rank p.

Example 17

Consider (P1): $\max\ (x_1 - 2)^2$

$$x_1 + x_2 = 4$$

$$\mathcal{D}: \quad x \geq 0$$

Here we have $x^* = (0,4)'$ and $(4,0)'$.

At $(0,4)'$ (where Kuhn-Tucker conditions are satisfied with $\mu^* = 0$).

$$(\omega\nabla f - \mu\nabla h)s = (\omega(2(x_1 - 2), 0) - \mu(1,1))s =$$

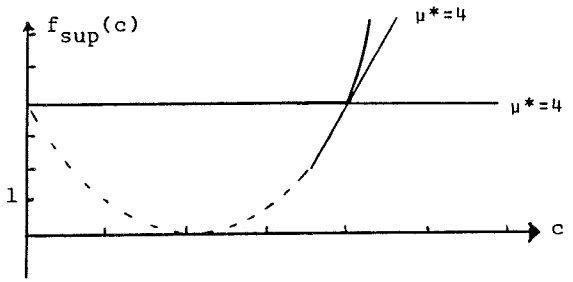
$$s_1(-4\omega - \mu) - \mu s_2 = -4\omega\alpha$$

A feasible s is $(\alpha,-\alpha)'$, $\alpha \geq 0$. This is non-positive if we let $\omega = 1$.

At $(4,0)'$ (where Kuhn-Tucker conditions are satisfied with $\mu^* = 4$) the only feasible s is $(-\alpha,\alpha)'$, $\alpha \geq 0$.

Therefore $(\omega\nabla f - \mu\nabla h)s \leq 0$ if $\omega = 1$.

We see that $f_{sup}(c)$ is not differentiable at $\bar{c} = 4$:



From this we conclude that the homogeneous weak Lagrangian principle is applicable for (P1), even though the conditions in Theorems 22 and 23 are not satisfied.

That is we have not tangent hyperplane but a sub-gradient to the perturbation function.

Kuhn-Tucker conditions are also satisfied at $x = (2,2)'$, this is a point at which the objective function is minimized. At

this point $(\omega\nabla f - \mu\nabla h)s = 0$, that is the homogeneous weak Lagrangian principle is applicable for

$$\max - (x_1 - 2)^2$$

$$x_1 + x_2 = 4$$

$$x \geq 0$$

Moreover a tangent (and also supporting) hyperplane exists at $\bar{c} = 4$ where $\mu^* = 0$. Thus the conditions in Theorem 22 are satisfied. Remark that a stationary solution to the Lagrangian function for $\mu = 0$, provides a minimum to (P1) and it is impossible to generate a maximum to (P1).

---

4. EVERETT'S THEORY VS. KUHN-TUCKER THEORY

a) We have seen two different theories of Lagrange multipliers. Everett's theory is more general in the sense that it permits nonlinear Lagrange multipliers function. Moreover Everett's theory does not assume differentiability. Finally Kuhn-Tucker theory, under some assumptions, provide necessary conditions while Everett's theory stipulates sufficient conditions for optimality. Most books on Mathematical Programming make not a clear distinction between these two approaches, see for instance [13]. In the next chapter we will see how these two approaches will give origin to two different ways to elaborate numerical methods to solve constrained nonlinear problems.

b) Assume we are dealing with linear Lagrange multiplier functions, the table on the following page gives examples of problems where all or some of the above discussed conditions are or are not satisfied.

$\exists x^*$ for (P1)	Kuhn-Tucker necessary conditions	Sufficient conditions	
		Saddle point	Theorem 21
max x $\frac{x^2}{2} + \delta(x)2 \leq 3$	./.	./.	./.
max x^4 $-\frac{1}{2} \leq x \leq \frac{1}{2}$	×	./.	./.
max $x^2 + x^3$ $x^4 \leq 1$	×	×	./.
max e^{-x} $x > 0$	×	./.	×
max f(x) $0 \leq x \leq 2$ where $f(x) = \begin{cases} 3x, 0 \leq x \leq 1 \\ 4-x^2, 1 \leq x \leq 2 \end{cases}$	./.	×	./.
max $-(x-2)^2$ $2x = 6$ $x \geq 0$	×	×	×

Note: ./. means that the conditions are not satisfied
× means that the conditions are satisfied.

c) A very important family of problems, where the two theories are applicable is the so-called concave (differentiable) programming problems that satisfies the constraint qualification assumption. Here Everett's theory, duality theory and Kuhn-Tucker theory are closely interconnected. Linear programming problems is the best known member of this family.

d) It has been consistently assumed that x and $f(x)$, $g(x)$, $h(x)$, take values in a finite dimensional space. However, applications often demand much more generality for example, it may be necessary for x to take values in the space of functions (as in the case of control problems). The methods employed hitherto do indeed generalize naturally to more complex vector spaces. Nevertheless, treatment of more general cases does require additional discussion, concepts and precautions, for an elucidating approach see [14].

5. REFERENCES

[1] Gould, F.J.: Extensions of Lagrange multipliers in nonlinear programming, SIAM, J. Appl. Math., vol. 17, 1969.

[2] Everett, H.: Generalized Lagrange multiplier method for solving problems of optimum allocation of resources Operations Research 11, 1963.

[3] Whittle, P.: Optimization under constraints, 1971.

[4] Varaiya, P.P.: Notes on optimization, 1972.

[5] Kuhn, H.W., and Tucker, A.W.: Nonlinear programming, Proc. Second Berkeley Symposium on Mathematical Statistics and Probability, 1951.

[6] Lasdon, L.S.: Optimization theory for large systems, 1970.

[7] Geoffrion, A.M.: Duality in nonlinear programming: A simplified applications-oriented development, SIAM Review, vol. 13, 1971.

[8] Evans, J.P., and Gould, F.J.: A nonlinear duality theorem without convexity, Econometrica, vol. 40, 1972.

[9] Aoki, M.: Introduction to optimization techniques, 1971.

[10] Fiacco, A.V., and Mc Cormick, G.P.: Nonlinear programming, 1968.

[11] Bazaraa, M.S., and Goode, J.J., and Shetty, C.M.: Constraint qualifications revisited, Management Science, vol. 18, No. 9, 1972.

[12] Zangwill, W.: Nonlinear programming, 1969.

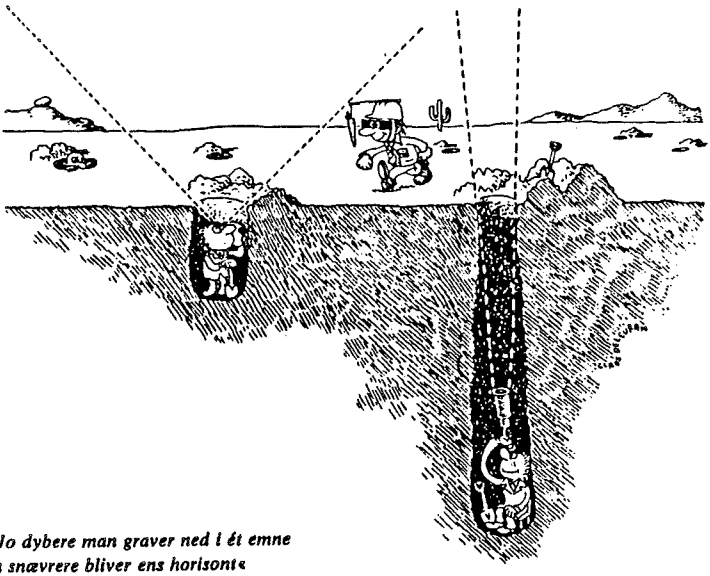
[13] Intriligator, M.D.: Mathematical optimization and economic theory, 1971.

[14] Luenberger, D.: Optimization by vector space methods, 1969.

CHAPTER 3

STATIC OPTIMIZATION:

Computational methods



»Jo dybere man graver ned i ét emne
ja snævrere bliver ens horisont«

1. INTRODUCTION

In the last chapter most of the theoretical results and geometrical properties of static optimization problems have been presented. Here we will concentrate on the best known computational methods that have been suggested or applied to solve optimization models. This is not an easy task due to the fact that this area of Mathematical Programming is still under development.

Our purpose is to describe, as simply as possible, and evaluate several of the most effective methods of nonlinear programming. Emphasis will be placed on the main theoretical principles on which a given numerical method is grounded and on its advantages and disadvantages when solving real life problems. A more complete and profund discussion belongs to a course on Numerical Analysis, the interested reader should consult [1], [7] and [9].

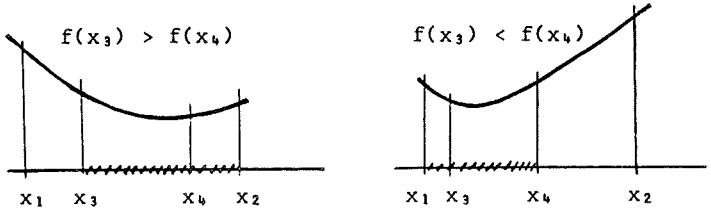
The methods we will discuss will hopefully converge to local optima, convergence to global optima can be guaranteed only under very restrictive assumptions. We start by discussing the simplest problem, that is one-dimensional optimization. Then in section 3 computational methods to solve min $f(x)$, $x \in R^n$, will be presented. A good complementary lecture to this section is [13], where most results are shown and cleary illustrated. Finally in the last section the most complex problem min $f(x)$, $x \in X$ will be solved.

2. ONE-DIMENSIONAL MINIMIZATION

It seems reasonable that we should begin our discussion of numerical methods with the simplest problem, namely the one in which the minimum of a function of just one variable is sought. It will be seen later on that many of the techniques for minimizing functions of several variables perform searches along each of a sequence of directions, and each of these linear searches is equivalent to a univariate search. Thus the efficiency of any such procedure depends critically on the efficiency of the method used to solve the single-dimensional search. The methods of univariate searching fall into two classes: methods which specify an interval in which the minimum lies, and methods which specify the position of the minimum by a point approximating to it. Let us now describe some of these methods.

2.1 Methods which specify an interval

Let us first assume that an initial interval known to contain the minimum is given and that the function is unimodal within this interval. These methods will reduce this interval until the minimum is located to the required accuracy.



Suppose that the minimum lies within $[x_1, x_2]$. Choose now x_3 and x_4 so that: $x_1 < x_3 < x_4 < x_2$. Now if

$f(x_3) \geq f(x_4) \rightarrow$ minimum lies $[x_3, x_2]$, whilst

if $f(x_3) \leq f(x_4)$ → minimum lies $[x_1, x_4]$.

This procedure can be continued until the interval containing the minimum has been reduced to a specified size. The central problem is then: how to locate x_3 and x_4? Let us see two methods.

Fibonacci Search

This method uses the sequence of positive integers known as Fibonacci numbers. These are defined by the relations:

$$F_0 = F_1 = 1$$

$$F_n = F_{n-1} + F_{n-2}, \quad n \geq 2$$

and therefore the sequence begins $1,1,2,3,5,8,13,\cdots$. If N is the total number of function evaluations to be performed, the test points for the i^{th} iteration are

$$x_3^i = \frac{F_{N-1-i}}{F_{N+1-i}} (x_2^i - x_1^i) + x_1^i$$

and

$$x_4^i = \frac{F_{N-i}}{F_{N+1-i}} (x_2^i - x_1^i) + x_1^i$$

for $i = 1,2,\cdots,N-1$, where $[x_1^i, x_2^i]$ is the initial interval. However, the use of these rules makes the last two test points coincident at the midpoint of the interval $[x_1^{N-1}, x_2^{N-1}]$. Therefore in order to determine in which half of the range $[x_1^{N-1}, x_2^{N-1}]$ the minimum actually lies, we displace one of these final test points by a small value ε .

It is easy to show that the i^{th} iteration reduces the interval containing the minimum by a factor $\frac{F_{N-i}}{F_{N+1-i}}$ [13], and after N

function evaluations (that is, N-1 iterations), the final interval is of length

$$L_N^* = \frac{1}{F_N}\left(x_2^1 - x_1^1\right) + \varepsilon$$

at most. Then if δ is the required accuracy, N must be chosen so that

$$F_N \geq \frac{x_2^1 - x_1^1}{\delta - \varepsilon} > F_{N-1}$$

The Fibonacci technique is an "optimal" search procedure in a particular sense. In the N-experiment case the final interval of uncertainty, ℓ_N, depends both on K, the index of the smallest experiment, and the position of the experiments $x_1, x_2, \cdots, x_N$, that is: $\ell_N = \ell_N\left(K, x_1, x_2, \cdots, x_N\right)$. An obvious requirement of a good search plan is that it makes the final interval of uncertainty as small as possible, and that it does this no matter what (unimodal) function it operates on. Since ℓ_N depends on K, thus on the function being minimized, minimizing it would yield a plan good only for a particular function. We can remedy this by defining

$$L_N\left(x_1, \cdots, x_N\right) = \max \ell_N\left(K, x_1, \cdots, x_N\right)$$

$$1 \leq K \leq N$$

where L_N is the maximum final interval of uncertainty obtained over all best outcomes, K, this is independent of K, hence independent of the function being minimized. It is easy to prove [12] that the Fibonacci search <u>minimizes the maximum final interval of uncertainty</u>, or

$$L_N^* = \min_{x_1, \cdots, x_N} \max_{1 \leq K \leq N} \ell_N\left(K, x_1, \cdots, x_N\right)$$

Thie criterion assures us that the final interval cannot be greater than L_N^*, assuming a unimodal function. It is a rather

conservative criterion, yet leads to very effective search results. Moreover, it is easily seen that each iteration, except the first, requires just one function evaluation.

Search by golden section

A disadvantage of Fibonacci search is that in many practical situations it is not possible to specify in advance the exact number of evaluations to be performed. Golden section is one method which does not have this disadvantage. This method is a simplification of the last method for which successive interval reduction factors are equal to some value τ. Then

$$x_3^i = \frac{\tau-1}{\tau}\left(x_2^i - x_1^i\right) + x_1^i$$

and

$$x_4^i = \frac{1}{\tau}\left(x_2^i - x_1^i\right) + x_1^i$$

Now $\tau = \frac{1}{2}(1+\sqrt{5})$, this value comes out from the fact that for large N, $\frac{F_{N+1}}{F_N} \approx \frac{1}{2}(1+\sqrt{5})$.

Since the length of the interval after the $(i-1)^{th}$ iteration is equal to the sum of the lengths of the intervals after the i^{th} and $(i+1)^{th}$ iterations, the interval after the $(i-1)^{th}$ iteration is cut into "golden section", since the ratio of the whole interval to the larger segment is equal to the ratio of the larger segment to the smaller. Golden section reduces the interval by a factor $1/\tau^{N-1}$. As

$$\frac{F_N}{\tau^{N-1}} = \frac{\tau^2}{\sqrt{5}} = 1.17$$

then Fibonacci search asymptotically achieves an interval reduction 17% greater that can be achieved with search by golden section.

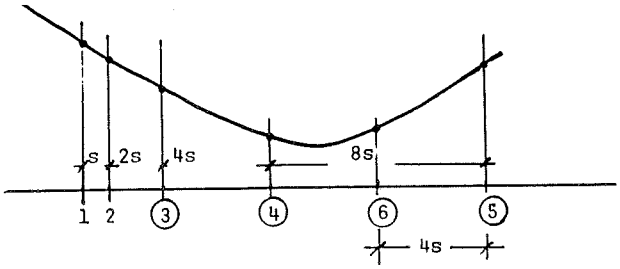
2.2 Methods which specify the position

To use these methods, an initial point approximating to the minimum must be provided. The methods proceed by fitting a low order polynomial through a number of points, and then finding the minimum of the fitted function, the procedure being repeated until the minimum has been found to the required accuracy. These methods vary according to the selection of points through which the fitted polynomial is to pass, and the order of this polynomial.

It has been found that the two methods to be described below are usually more efficient (as measured by the number of function evaluations necessary for the minimum to be located to a specified precision) [8] than methods such as Fibonacci search.

The algorithm of Davies-Swan-Campey

In this method, the function is first evaluated at the given initial point. A step is then taken along the line of search, and the function recalculated. If this function value is less than or equal to the initial function value, the step-length is doubled, and a furthermore made in the direction in which the function is decreasing. This process is repeated until a step resulting in an increase in the function value is performed, indicating that the minimum has been overshot. The step-length is then halved, and a step again taken from the last successful point. This gives four points equally spaced along the line of search, at each of which the function has been evaluated (points 3, 4, 6 and 5 in the figure below).



The end point farthest from the point corresponding to the smallest function value is rejected, and the remaining 3 points used for quadratic interpolation, that is points 3, 4 and 6 in our example. If the first step fails, the direction of search is reversed by changing the sign of the step. If this also fails then the minimum has been boxed in, and the interpolation may be performed. If the 3 points to be used for the interpolation are $x_1 = x_2 - s$, x_2 and $x_3 = x_2 + s$, and if the corresponding function values are f_1, f_2, and f_3, then it can be shown [8] that the minimum of the fitted quadratic is at $x_2 + s_m$, where

$$s_m = \frac{s(f_1-f_3)}{2(f_1-2f_2+f_3)}$$

A new stage, with reduced s, is then begun using as initial point whichever of x_2 and $x_2 + s_m$ corresponds to the smaller function value. One disadvantage of this method resides in the fact that its efficiency is highly dependent on the initial point and the initial step-length.

Powell's method [13]

In this procedure, the function is evaluated at a base point x_1 and at $x_2 = x_1 + s$. The point x_3 is chosen to be

$$x_1 + 2s \quad \text{if} \quad f_1 \geq f_2$$

but

$$x_1 - s \quad \text{if} \quad f_1 < f_2$$

The optimum x_m of the quadratic passing through these 3 points is given by

$$x_m = \frac{1}{2} \frac{(x_3^2-x_2^2)f_1 + (x_1^2-x_3^2)f_2 + (x_2^2-x_1^2)f_3}{(x_3-x_2)f_1 + (x_1-x_3)f_2 + (x_2-x_1)f_3}$$

If x_m and whichever of x_1, x_2 and x_3 corresponds to the smallest function value differ by less than the required accuracy, the minimum is assumed to have been located. Otherwise one of the 3 points x_1, x_2 and x_3 is discarded. The process of quadratic interpolation is then continued.

Coggin's method

Powell's algorithm is one of the most efficient procedures, but it has two undesirable properties. First, if s is too large one can introduce a point too distant from the minimum, which can be detrimental to the search, as for example when this wayward point returns to the vicinity of the minimum only slowly. Secondly, x_m could be a maximum instead of a minimum.

Coggin's method avoids these problems by using first Davies-Swann-Campey algorithm to bracket the minimum and all subsequent calculations follow Powell's algorithm. This procedure has been shown to the better [8] than either of the individual algorithms, therefore it is often recommended to solve univariate search problems.

2.3 Final remarks

The above described procedures will work successfully for the case of unimodal functions and they are nearly equally efficient [1]. Otherwise Coggin's method is recommended, but sometimes more complicated methods are needed, especially for the case with many local minima or for the case where the objective function is perturbed by some noise. Statistical (adaptive) procedures are usually recommended in such situations [14], [16].

3. UNCONSTRAINED MINIMIZATION

This type of problems can be classified in two groups:

a) Those problems with a given, once- or twice-differentiable, objective function. These are usually solved by procedures that use derivatives, some of which will be discussed in the next section. Moreover many theoretical results are also available, this will permit us a deeper insight on the problems under study.

b) Those where the objective function is not continous or it is not known a priori, then values has to be generated by experimentation. These problems are usually solved by procedures that do not use derivatives also known as search methods. These methods can obviously be applied to solve the first kind of problems.

As a general rule the derivative-type methods converge faster than search methods. That is due to the fact that these methods utilize more information on the objective function. However, in practice, the derivative-type methods have two main barriers to their implementation. First, in problems with a modestly large number of variables, it is laborious or impossible to provide analytical functions for the derivatives, and numerical methods to evaluate derivatives might introduce too large errors. Second, the derivative-type methods require a relatively large amount of problem preparation.

3.1 Unconstrained minimization procedures using derivatives

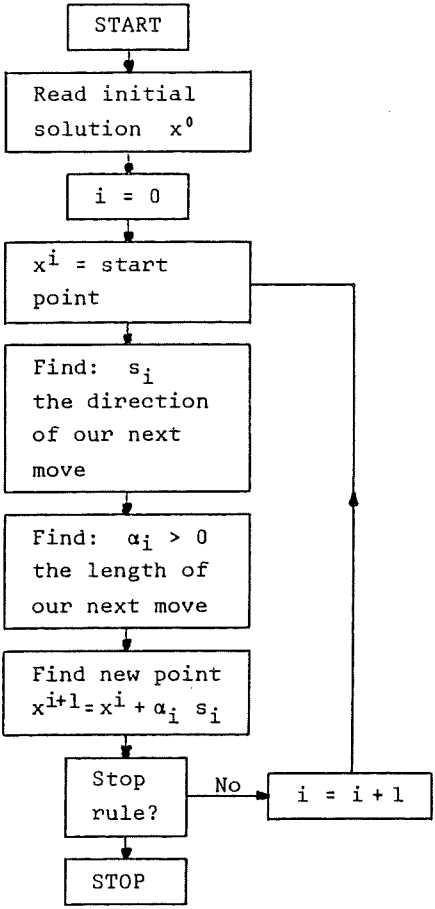
The problem we want to solve is:

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

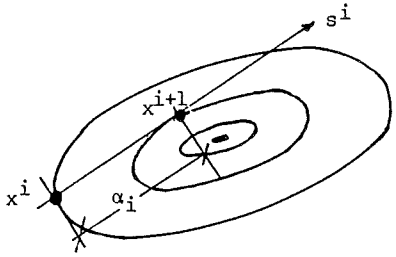
where $f(x)$ is a differentiable function. From our theoretical discussion we know that if a minimum exists, necessarily $\nabla f(x^*) = 0$' (see chapter 1). The methods we will discuss in this section will usually converge to stationary points, that can be optimal solutions. Convergence to global minima can only be guaranteed under very restrictive assumptions.

General procedure

The following procedure is usually applied to find a local minimum:



This procedure is illustrated in the figure below.



where given a point x^i we first have to find a direction s^i and thereafter a step-length α_i that will define a new point x^{i+1}.

This general procedure depends on two factors that are here the same for all these methods. These two factors are:

a) What is a good starting solution?

The answer will many times depend on the actual problem in study. The point to which a numerical procedure will converge, if it converges, is usually depending on the start point. For problems with many local maxima and minima, different stationary points could be generated by changing the initial solution.

b) Which stop rules will be used?

These rules will specify criteria for the termination of an iterative procedure, and the selection of one or more rules will have influence upon the degree of precision obtained. The following rules are often used:

- $x^{i+1} - x^i \simeq 0$
- $f(x^{i+1}) - f(x^i) \simeq 0$
- $\dfrac{f(x^{i+1}) - f(x^i)}{f(x^i)} \simeq 0$
- $\nabla f(x^{i+1}) \simeq 0$'
- $|\nabla f(x^{i+1})| \simeq 0$
- Number of times that $f(x)$ has been evaluated $\simeq N$
- Computation time $\simeq T$

Many times more than one stop rule are utilized [13].

The other three factors that will define a method are:

c) How to find s_i ?

That is the direction of the next move.

d) How to find α_i ?

Two general methods of selecting the step-size α_i are usually employed. In one method, the one we will employ, the objective function is minimized with respect to α_i in order to move from x^i to x^{i+1}, while in the other method a fixed, or variable, value is selected for the α_i's.

e) Under which assumptions can convergence be shown?

This is a theoretical question. It has become a tradition to prefer methods that can be proved to converge to optimal solutions for the case of "nice" functions, usually quadratic convex functions [9].

In what follows we will briefly describe some of the best known procedures for unconstrained minimization of differentiable functions. These procedures will be described by giving answer to the last three questions.

A) The Optimum Gradient or Steepest Descent method

c) As it is known that $-\nabla f(x)$ points in the direction of steepest descent from x, here we set:

$$s_i = -\nabla f(x^i)\,.$$

e) If $f(x)$ is assumed to have a continuous derivative and the α_i are chosen so that: $f(x_{i+1}) < f(x_i)$, $\forall i$, then the method will converge to a stationary point. Moreover it can be shown that if $f(x)$ is convex, having third-order differentiability, the method converges to a global minimum in the limit as $i \to \infty$, [1].

Characteristics of the method

- The function $f(x) = \sum_{j=1}^{n} x_j^2$, is minimized by this method in only one step.

- We observe that

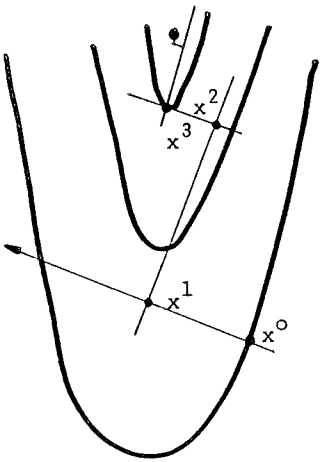
$$\frac{d}{d\alpha} f(x^i + \alpha s_i) = \nabla f(x^i + \alpha s_i) s_i = 0$$

that is

$$\nabla f(x^i + \alpha s_i) s_i = \nabla f(x^{i+1}) s_i = -s_{i+1}^{!} s_i = 0$$

so that successive directions are orthogonal to each other.

- The last fact lead to very slow convergence near the optimum for functions whose contours are in a way eccentric. Thus an inefficient zigzag behaviour results as shown below. This occurs because the gradient direction is generally quite different from the direction to the minimum. Most functions occurring in practice are of these kind, thus more efficient schemes are desirable.



- One advantage of this method is the fact that if our initial solution is very far from the optimum, the first initial steps are very large.

B) The Newton-Raphson method

Recently, a number of minimization techniques have been developed which substantially overcome the above difficulties. These methods have their origin in the so-called Newton-Raphson method that we are going to describe now. Suppose that the second-order derivatives exist, then we can approximate f in the vicinity of a given start point, say x_0, by the quadratic function Q,

$$Q(x) = f(x_0) + \nabla f(x_0)(x-x_0) + \tfrac{1}{2}(x-x_0)'H(x_0)(x-x_0)$$

As a next point of iteration we will select x_1 so that

$$\nabla Q(x_1) = 0$$

or

$$\nabla f(x_0)' + H(x_0)(x_1-x_0) = 0$$

Supposing our minimum is locally unique and H is positive definite in the vicinity of x^*, then

$$x_1 = x_0 - H(x_0)^{-1} \nabla f(x_0)'$$

This is the iteration scheme for Newton's method, then Newton-Raphson's method goes as follows:

c) $$s_i = -H(x^i)^{-1} \nabla f(x^i)'$$

e) The criterion to guarantee convergence when $i \to \infty$, assuming $f(x)$ twice-differentiable is that the inverse of the Hessian matrix, that is $H(x^i)^{-1}$, should be positive definite, [1].

Characteristics of the method

- A draw-back with this method is that we must evaluate H and H^{-1}. Many standard digital computer programs for matrix inversion are unsatisfactory.

- Experience has shown that due to round-off errors H has a tendence to become singular.

- The convex problem $\min[x' Ax]$ can theoretically be solved in 1 iteration. In practice this may not be true due to round-off errors.

- Convergence is very fast near the minimum, but usually convergence fails when the initial solution is too far from the optimum. One can conclude that a suitable combination of steepest descent for the initial steps and Newton's methods for the steps near the optimum should be desirable, the next procedure has somehow this property.

C) Davidon-Fletcher-Powell method (variable metric, quasi-Newton or large-step gradient method).

This method approximates H or its inverse but use only information from ∇f, thus:

c) $$s_i = -H_i \nabla f(x^i)'$$

Now the initial H_0 is a given positive definite matrix (usually $H = I$) and then

$$H_{i+1} = H_i + A_i + B_i$$

where

$$A_i = \frac{\sigma_i \sigma_i'}{\sigma_i' y_i} \quad , \quad B_i = -\frac{H_i y_i y_i' H_i}{y_i' H_i y_i}$$

and

$$\sigma_i = \alpha_i s_i \quad , \quad y_i = \nabla f(x^{i+1})' - \nabla f(x^i)'$$

see further [13].

e) If conjugate directions are employed (i.e., $s_i' Q s_j = 0$, $i \neq j$ for $i,j = 0,1,\cdots,n-1$ where $s_i \neq 0$, $\forall i$, and Q is the matrix of the function $(a+b'x + \frac{1}{2} x'Qx)$ any quadratic function of n variables that has a minimum can be minimized in n steps. It is easy to prove that this method generates directions s_i that are conjugate for a quadratic function, and therefore minimizes it theoretically in $\underline{n}$ steps. In other words this method has the property of quadratic termination, [1].

Characteristics of the method

- It is probably the most powerful general procedure for finding a local minimum which is known at the present time.
- H_i are all positive definite. As a consequense of this the method will usually converge.
- No matrix inversion is needed.
- When applied to strictly convex quadratic functions, $x' Ax$, it arrives at the minimum in n steps, and $H_n = A^{-1}$.
- $\lim_{i \to \infty} H_i = H(x^*)^{-1}$.
- It has been found that this method can take negative steps occasionally or terminate at a non-stationary point. This happens when H_i has become singular due to round-off errors.

- For large n, it may be cumbersome to work and store matrices, this is a draw-back with this method.

- Several other methods have been proposed that modify in some way Davidon-Fletcher-Powell method, although keeping some of its properties, that is for instance iterative approximation to the inverse Hessian matrix [1].

D) Fletcher-Reeves Conjugate Gradient method

c) $$s_{i+1} = -\nabla f(x^{i+1})' + \beta_i s_i$$

and we start with

$$s_0 = -\nabla f(x^0)'$$

and

$$\beta_i = \frac{\nabla f(x^{i+1})\ \nabla f(x^{i+1})'}{\nabla f(x^i)\ \nabla f(x^i)'}$$

d) Quadratic termination.

Characteristics of the method

- Generate conjugate directions.

- When applied to strictly convex quadratic functions, this method is analogue to the one described in C).

- It requires less storage than the last method.

- In general it is not so efficient as the last method.

- No matrix inversion is required.

- Several other methods have been proposed that modify in some way this method, although keeping some of its properties, that is for instance conjugate directions that are linear combinations of $-\nabla f(x^{i+1})'$, [1].

3.2 Unconstrained minimization procedures without using derivatives.

Optimization techniques that proceed by evaluating the function only, and do not estimate its derivatives, are often termed direct search techniques. We will discuss two of the best known methods. The first one, pattern search, that probes along predetermined directions, and Powell's method, that undertakes unidirectional searches for minima.

A) Pattern search

This method consists of two major phases, an exploratory search around the base point and a pattern search in a direction selected for minimization. The algorithm operates as follows. Initial values for all elements of x must be provided, as well as an initial incremental change Δx. To initiate an exploratory search, $f(x)$ is evaluated at a base point (the base point is the vector of initial guesses of the independent variables for the first cycle). Then each variable is changed in rotation, one at a time, by incremental amounts, until all the parameters have been so changed. To be concrete, x^0 is changed by an amount $+\Delta x_1^0$, so that $x_1^1 = x_1^0 + \Delta x_1^0$. If $f(x)$ is reduced, $x_1^0 + \Delta x_1^0$ is adopted as the new element in x, otherwise, x_1^0 is changed by $-\Delta x^0$, and the value of $f(x)$ again checked as before. If the value of $f(x)$ is not improved x_1^0 is left unchanged. Then x_2^0 is changed by an amount Δx_2^0, and so on, until all the independent variables have been changed to complete one exploratory search. For each step in the independent variable, the value of the objective function is compared with the value at the previous point. If the objective

function is improved for the given step, the new value of the objective function replaces the old one in the testing. However, if a perturbation is a failure, then the old value of $f(x)$ is retained. After making one (or more) exploratory searches in this manner, a pattern search is made. The successfully changed variables define a vector in R^n that represents a successful direction for minimization. Then a series of accelerating steps is made along this vector as long as $f(x)$ is decreased. The magnitude of the step for the pattern search in each coordinate direction is roughly proportional to the number of successful steps previously encountered in each coordinate direction during the exploratory searches for several previous cycles. If $f(x)$ is not decreased, the pattern search is said to fail, and a new type of exploratory search is made in order to define a new successful direction. If this also fails, Δx is reduced gradually, until either a new successful direction can be defined or each Δx_i becomes smaller than some given tolerance. Failure to decrease $f(x)$ for very small Δx indicates that a local minimum has probably been reached.

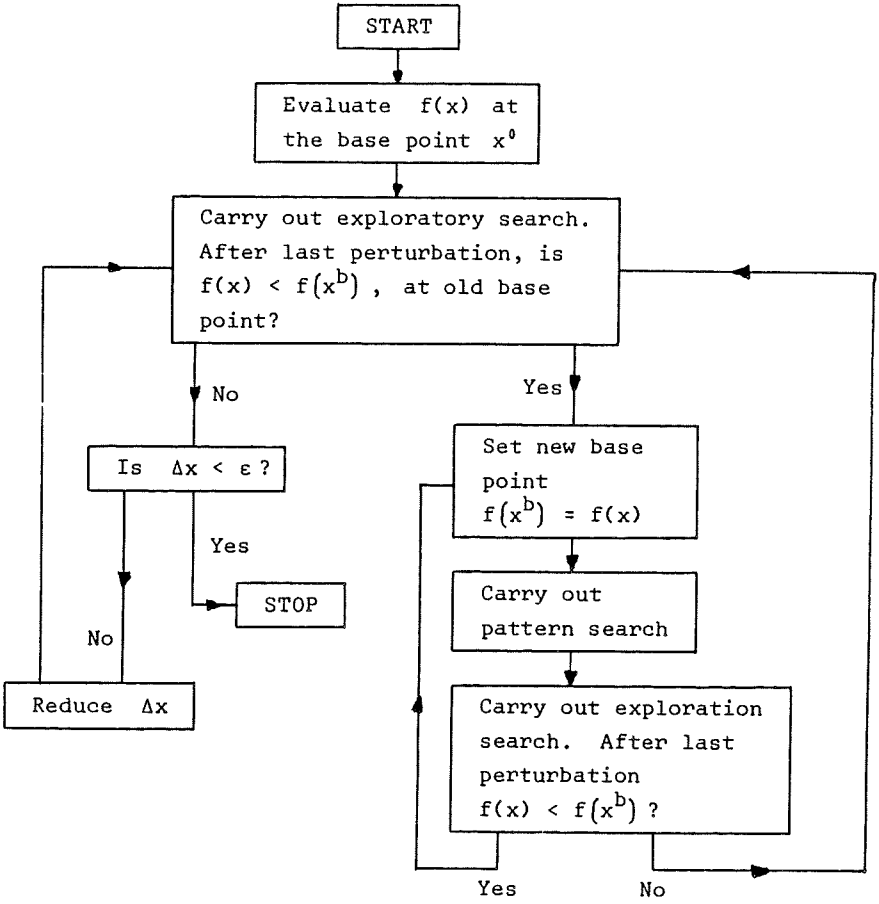
The flow chart next page illustrates this method.

Characteristics of the method

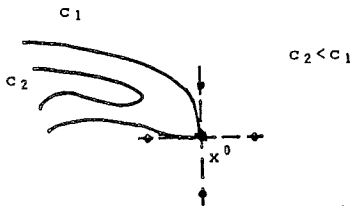
- In the literature there exists many versions of the pattern search method, where different exploration rules are used, [1], [15].

- Although the method lacks mathematical elegance, it is a highly efficient optimization procedure. The simplicity in computer programming furthers its appeal.

- This technique is particularly well suited to functions exhibiting a straight, sharp ridge or valley.

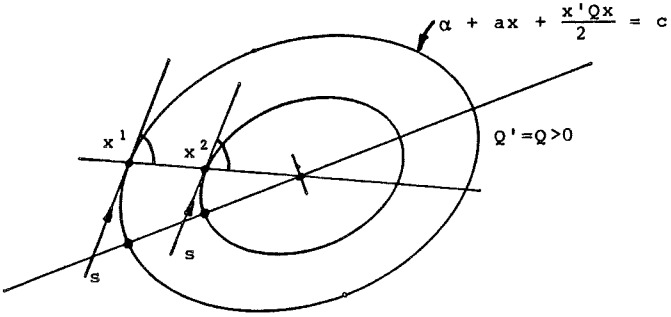


- The disadvantage of the pattern search method is that it may "get stuck", that is it may stop short of a local minimum unable to make further improvement, unless the strategy for exploratory moves is made sufficiently complicated. This is especially true if the objective function contours have sharp corners or very curved ridges, as shown below.



B) Powell's method

This method is based on the simply observation that any line that passes through the minimum of a quadratic function intersects the contours of the quadratic form at equal angles. This is illustrated below.



The corollary of this observation is that if two minima are found along two parallel directions s, the minimum of the quadratic form is located along $x^1 - x^2$. Moreover, this direction is Q-conjugate to s.

The directions of search are generated by the following four steps:

(i) Choose a starting point x^0 and $s_1,\cdots,s_n$ linearly independent directions (e.g. the co-ordinate directions).

(ii) For $r = 1,2,\cdots,n$ find α_r that minimizes

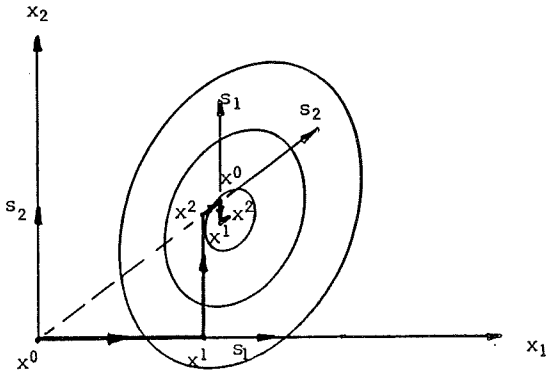
$$f(x^{r-1} + \alpha_r s_r)$$

and set $x^r = x^{r-1} + \alpha_r s_r$.

(iii) Replace s_n by $x^n - x^0$, and s_r by s_{r+1}, $r = 1,2,\cdots,n-1$.

(iv) Find α that minimizes $f(x^n + \alpha(x^n - x^0))$ and replace x^0 by $x^n + \alpha(x^n - x^0)$. Go to (ii).

This procedure is illustrated below.



This algorithm produces, for quadratic forms, conjugate directions and will therefore converge in n steps. However, it may fail if one of the α_r's is zero. The directions can also, in a practical example tend to be nearly dependent and so Powell proposed a revised version that has worked satisfactorily in practice, but unfortunately the property of quadratic conver-

gence in n steps is no longer true. The procedure to generate new directions is as follows (i) and (ii) are as before. But

(iii) Find m, $1 \leq m \leq n$, that solves $\max\left(f(x^{m-1}) - f(x^m)\right)$ (which we denote Δ).

(iv) Set $f_1 = f(x^0)$, $f_2 = f(x^n)$, and $f_3 = f(2x^n - x^0)$.

(v) If $f_1 > f_3$ and $(f_1 - 2f_2 + f_3)(f_1 - f_2 - \Delta)^2 < \frac{1}{2}\Delta(f_1 - f_3)^2$ set $s = x^n - x^0$, and find α that minimizes

$$f(x^n + \alpha s)$$

For the next iteration use the directions $s_1, \cdots, s_{m-1}$, $s_{m+1}, \cdots, s_n$, s and $x^n + \alpha s$ as our new x^0. Go back to (ii).

(vi) Otherwise, use the old directions $s_1, \cdots, s_n$ for the next iteration and x^n as our new x^0. Go back to (ii).

Powell recommends that convergence be established by continuing the search until each variable is altered during an iteration by less than 1/10 of its required accuracy [8]. This first solution will be denoted by a. A second solution b is then found by repeating the whole procedure from a starting point obtained by perturbing each variable from the first solution point a by an amount equal to ten times its required accuracy. The function is then minimized along $a - b$ to give point c. If all the components of $a - c$ and $b - c$ are less than 1/10 of the required accuracy, convergence is assumed. Otherwise $a - c$ is incorporated into the set of search directions in place of s_1, and the whole process restarted from c. This is a very safe criterion in that it has been found to ensure convergence to a minimum for most problems. However it has also been found that the first solution point a is usually sufficiently close to the minimum for all practical purposes, and a considerable reduction in the necessary number of func-

tion evaluations is achieved if the searches for b and c are omitted.

Characteristics of the method

- It is probably the most effective search technique, especially in the region near the optimum.

- It demands many one-dimensional searches, and these are usually time consuming.

- The method behaves poorly for bad scaled problems, because it has failed to choose any new directions.

3.3 Comparison of the methods

There is no one algorithm that is most suitable for all optimization problems. The number of variables (>5 ?), the possibility of having analytic expressions for the first derivatives, the existence of good auxiliary subroutines, etc., are factors that will influence on the decision concerning the selection of a method. And, even then, when a method has been selected a great deal of experimentation is necessary before being able to implement a model and for its solution.

Himmelblau [1] has performed an extensive experimentation with the purpose to rank the different methods, but this is not an easy task because it is difficult to find a single comparison criterion.

For test problems containing few variables he arrived to the following conclusion:

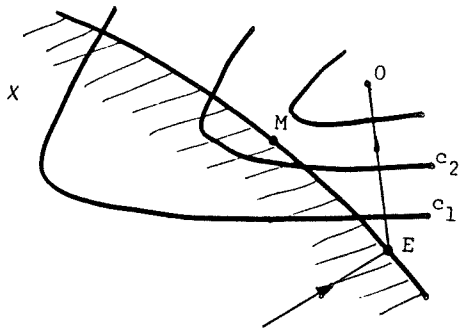
Classification	Method
Superior	Davidon-Fletcher-Powell
	Powell
Fair	Fletcher-Reeves
	Pattern search

When dimensionality increases Davidon-Fletcher-Powell and Powell's method are roughly equivalent and distinctly superior to other methods.

These results are practically equivalent to tests performed by other researches, [2], [6].

4. CONSTRAINED MINIMIZATION

We have seen that unconstrained optimization methods iterate by selecting a search direction and a step length, these are chosen from knowledge about the objective function, but when a constraint is present this too plays an important part in constrained minimization procedures. Let us assume that in the figure below the point E has been reached by an unconstrained optimizer.



Continued use of the optimizer may give a new direction pointing towards the unconstrained optimum, O , whereas the search should be directed towards the constrained minimum, M . Thus the problem of constrained optimization is to incorporate into the search direction the appropriate information about the constraint, and on nonlinear constraints, this can be very difficult and time consuming.

There exist many procedures for solving constrained problems and they mostly fall into two categories as follows.

a) Direction modification without altering the function (linear approximation methods).

   Some of these methods attempt to follow a constraint while others try to rebound from them and so continue the search in the feasible region. In principle they are based in Kuhn-Tucker theory.

b) Transformation of objective function followed by unconstrained minimization (penalty function methods).

   These methods seek to define a new function that has an unconstrained optimum at the same point as the optimum of the given problem. Optimization of this new function will then define the required change in the search direction. In principle they are based in Everett's theory.

In what follows we will describe some of the methods that have been most utilized in practice.

4.1 Linear approximation methods

Because linear programming methods have been successfully applied to large-dimensional problems containing both equality and inequality constraints, the linearization of nonlinear programming problems is one of the most obvious approaches, one of these procedures will be discussed below. Two somewhat more effective methods that we will discuss afterwards involve linearization of only the constraints.

There exist many other methods that are either modifications of the methods we will see in this section or of very special form applicable only for specific problems, [1], [6].

The conditions under which convergence to the solution of our problem is guaranteed are usually the following, [3], [9]:

- The problem is a convex programming problem.
- The objective and constraint functions are differentiable.
- The feasible region is non-empty, closed and convex.
- The constraint functions are bounded.

In practice any of the algorithms can reach a local minimum, that is directly or indirectly they try to satisfy Kuhn-Tucker conditions, even if these convergence conditions are not satisfied, depending highly on the nature of the problem.

A) Method of approximation programming (small-step gradient method).

Let us formally state our problem as

$$\min z = f(x)$$

subject to:

$$g(x) \left\{ \begin{matrix} \leq \\ = \end{matrix} \right\} b$$

$$k_1 \leq x \leq k_2$$

Notice that the method applies equally well to equality or inequality constraints. We then approximate our problem in the region about x^k, by

$$\min_x z = f(x^k) + \nabla f(x^k)(x - x^k)$$

$$g_i(x^k) + \nabla g_i(x^k)(x - x^k) \left\{ \begin{matrix} \leq \\ = \end{matrix} \right\} b_i \quad , \; i = 1, \cdots, m$$

$$k_1 \leq x \leq k_2$$

Since the partial derivatives are taken as constants, c_j and w_{ij}, the above may be written as:

$$\min z = z^k + \sum_{j=1}^{n} c_j \, \Delta x_j$$

subject to:

$$\sum_{j=1}^{n} w_{ij} \, \Delta x_j \left\{ \begin{matrix} \leq \\ = \end{matrix} \right\} b_i - g_i(x^k) \quad , \; i = 1, \cdots, m$$

$$k_{1j} - x_j^k \leq \Delta x_j \leq k_{2j} - x_j^k$$

We will obtain a standard linear programming problem by defining, $\Delta^+ x_j = \Delta x_j$, when $\Delta x_j \geq 0$ and $\Delta^- x_j = -\Delta x_j$ when $\Delta x_j \leq 0$. We finally have

$$\min(c\Delta^{+}x - c\Delta^{-}x)$$

subject to

$$w_i(\Delta^{+}x - \Delta^{-}x) \left\{ \begin{matrix} \leq \\ = \end{matrix} \right\} b_i - g_i(x^k) \quad , \; i = 1,\cdots,m$$

$$p_j \, \Delta^{+}x_j + q_j \, \Delta^{-}x_j \leq k_{3j} \quad , \; j = 1,\cdots,n$$

$$\Delta^{+}x_j \; , \; \Delta^{-}x_j \geq 0 \; , \; \forall j$$

$$p_j = \max\left[1 \; , \; \frac{k_{3j}}{k_{2j} - x_j^k}\right]$$

$$q_j = \max\left[1 \; , \; \frac{k_{3j}}{x_j^k - k_{1j}}\right]$$

The last inequalities keep the maximum distance any x_j can move no greater than k_{3j}. Furthermore, if a variable is near an upper or lower limit, the large value of p_j and q_j generated keeps the variable from exceeding the limit while allowing movement away from the limit.

After obtaining a solution to this problem, we obtain a new linear approximation and repeat the process. Because a complete relinearization of the problem is taken at each stage, all old information is discarded hence this method can be used to solve non-convex problems, but this causes also the problem that when non-feasible solutions are found the method works slowly.

Characteristics of the method

- In many problems, such as petroleum-refinery optimization, experience has shown the functions encountered to be well behaved so that convergence is usually obtained. When this is so, use of approximation programming is highly advantageous.

- This method may give rise to a poor search direction if the problem is highly nonlinear.

- It is difficult to devise acceleration procedures to make use of a priori information.

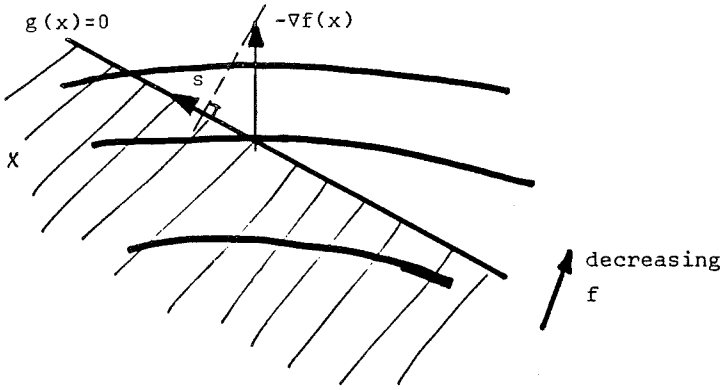
- Partial solutions are not necessarily feasible.

- Many times one has to add to the algorithm some rules to expand or contract the maximum step size to avoid oscillations near the optimal solution.

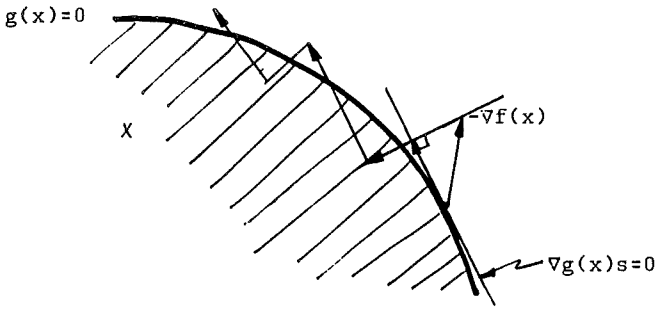
B) Projection methods (feasible directions or large-step gradient methods).

As we have seen earlier, if we have a feasible point x, then we must make two decisions:

We must pick a feasible direction s and we must decide how big a step to take in the direction s. There are generally many feasible directions. In deciding on a "best" direction s, suppose we simply minimize $(\nabla f(x)s)$ in the hope of decreasing f value as much as possible. Assume that at x at least one constraint is active. Suppose further that the direction of the negative gradient, $-\nabla f(x)$, is not feasible. If the constraint equation is linear, then such a "best" s is obtained by projecting $-\nabla f(x)$ on the manifold defined by the linear constraint equation, as shown below:



If the active constraint equation is nonlinear, the preceding procedure for picking s by projecting $-\nabla f(x)$ on the tangent plane of the constraint equation may not work, since any finite move along s such that $\nabla g(x)s = 0$ may violate the constraint $g(x) \leq 0$ *), as illustrated below:

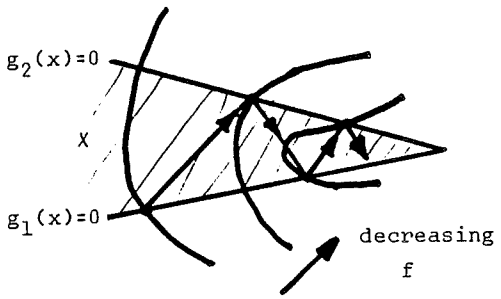


---

*) Note that the problem to be solved here is $\min_x\{f(x) \mid g(x) \leq 0\}$.

In such cases, a move must be followed by a "recovery" move to bring the next point back into X. We may then expect a "zigzagging" behaviour resulting in an inefficient algorithm.

Another type of difficulty may arise even for linear active constraints if we do not pay attention to constraint equations that are "almost active". This is because conceivably two or more constraints could become active and almost active alternately and the "zigzagging" phenomenon may again result, as illustrated below:



To avoid this zigzagging (also called jamming, since all the generated points get jammed into a corner), we must therefore include constraints that are almost active in choosing s. These considerations have to be incorporated in the direction-finding procedure. We have thus to modify our problem

$$\min_s \nabla f(x)s$$

$$\nabla g(x)s \leq 0$$

to the following problem that takes account of all these difficulties:

$$\min \xi$$

$$\nabla g_i(x)s \leq \theta_i \xi \quad , \quad \text{for all } i: \quad -\varepsilon \leq g_i(x)$$

$$\nabla f(x)s \leq \xi$$

$$s's = 1$$

where $0 \leq \theta_i \leq 1$, and $\varepsilon > 0$ are adjustable parameters and where ξ and s are the variables.

This is an almost linear problem and it can be shown that the problem can be handled by a modified version of linear programming [1]. Or, sometimes restrictions $s's = 1$ can be replaced by

$$L \leq s \leq U$$

a lower and an upper bound vector, and the use of linear programming is obvious.

Of course, once a direction has been selected, the problem of selecting the step size still remains. This problem may be dealt with almost as in the unconstrained case. It is still desirable to minimize the objective function along the vector s, but now no constraint may be violated. Thus we have to solve $\min_{\alpha} f(x^k + \alpha s_k)$ subject to $x^k + \alpha s_k \in X$. A new point is thus determined, and the direction-finding problem is repeated. If at some point $\xi \geq 0$, then the procedure terminates. The final point will generally be a local minimum of the problem.

Characteristics of the method

- There is not much information on the behaviour of this procedure in practice.

- Theoretically this method essays to satisfy the Kuhn-Tucker conditions discussed in chapter 2.

- A main disadvantage of this method is that at each iteration the direction-finding algorithm can be time consuming. An alternative is provided by the gradient projection method of Rosen [1]. Here a usable direction is found without solving an optimization problem, the direction, however, may not be locally "best" in any sense. The method is said to be efficient when all constraints are linear.

C) The generalized reduced gradient method [3]

For the sake of simplicity we will assume that all constraints are linear. We have to solve:

$$\min_x f(x)$$

subject to

$$Ax \leq b$$

where A is $(m \times n)$ and $m > n$. By introducing slack variables and suitable partitioning of A and b, our problem becomes:

$$\min_x f(x)$$

$$A_1 x + w^1 = b^1$$

$$A_2 x + w^2 = b^2$$

$$w^1, w^2 \geq 0$$

where A_1 is an $(n \times n)$ submatrix of A.

The last partition is chosen so that the vanishing components of the slack vector are all included in w^1. If $w^1 = 0$, then x is a vertex of the constraint polyhedron. Assuming that A has rank n, no more than n components of the slack

vector vanish. Then given x on a facet, a direction of search for a local minimum of $f(x)$ while maintaining the feasibility of x is in the direction $\Delta w_j^1 \geq 0$ if $w_j^1 = 0$, and if this does not cause w^2 vector to become negative. Since w^1 gives a convenient means of representing feasible moves, we regard w^1 as the independent variable and compute the gradient of $f(x)$ with respect to w^1, or

$$\nabla_{w^1} f(x) = -\nabla f(x) A_1^{-1}$$

This is called the reduced gradient. Therefore, a feasible move may be defined by

$$\Delta w_j^1 = -\left[\nabla_{w^1} f(x)'\right]_j \quad \text{if} \quad \left[\nabla_{w^1} f(x)'\right]_j < 0 \quad \text{or} \quad w_j^1 > 0,$$

$$\text{otherwise} \quad \Delta w_j^1 = 0$$

Now given a move $\Delta w^1 \geq 0$, then

$$\Delta x = -A_1^{-1} \Delta w^1$$

$$\Delta w^2 = A_2 A_1^{-1} \Delta w^1$$

To achieve a maximum length of move while maintaining the feasibility of x, w^1 and w^2 can be moved as much as

$$w^1 + \theta^* \Delta w^1 \quad \text{and} \quad w^2 + \theta^* \Delta w^2$$

where

$$\theta^* = \min(\theta_1, \theta_2)$$

$$\theta_1 = \max_{\theta \geq 0}(\theta \,;\, w^1 + \theta \Delta w^1 \geq 0)$$

$$\theta_2 = \max_{\theta \geq 0}(\theta \,;\, w^2 + \theta \Delta w^2 \geq 0)$$

without losing the feasibility of x. Within the move represented by $w^1 + \theta_1 \Delta w^1$ and $w^2 + \theta_2 \Delta w^2$, where $0 \leq \theta_1, \theta_2 \leq \theta^*$, we look for a minimum of $f(x)$,

$$\min_{\alpha} f(x + \alpha \Delta x)$$

If the value of α achieving this minimization is less than θ^*, then this move is completed and a new direction of move is computed. Otherwise choose the components of w^2 that vanishes as a component of w^1 and exchange it for the largest component of w^1, and a new direction of move is computed with respect to this new set of independent variables. Note that this is precisely the pivot operation of the simplex method. When $\Delta w^1 = 0$, then the search is completed.

This procedure is nothing else than the steepest descent method adapted to handle linear constraints by a procedure similar to the simplex method.

This algorithm can be extended so to be able to handle nonlinear equality constraints, with upper and lower bounds on the decision vector. In essence the method employs linear, or linearized constraints, defines new variables that are normal to some of the constraints, and transforms the gradient to this new basis. The best treatment can be found in [11].

Characteristics of the method

The procedure facilitates the introduction of many numerical methods to reduce computation time when elaborating a computer code.

Many experiences have been done with this algorithm and several standard programs have been elaborated. These programs contain many heuristic rules that are used to accelerate the procedure or to cope with complicated problems. Due to this fact this is probably the best general constrained optimization technique known today.

4.2 Penalty function methods

These methods are based on Everett's theory, let us see two of the best known procedures.

A) Generalized Lagrange multiplier technique

This procedure is grounded on our discussion of sufficient conditions to solve constrained problems. Assume we want to

$$\min f(x)$$

$$x \in \mathcal{D}$$

$$g_i(x) \leq b_i \quad , \quad i = 1,\cdots,m$$

This method is based on Everett's linear theory and proceeds as follows:

1. Select values y_i^0 , $i = 1,\cdots,m$

2. Find $x(y_i^0)$ that gives

$$\inf_{x\in\mathcal{D}} \left[f(x) + \sum_{i=1}^{m} y_i^0 \, g_i(x) \right]$$

3. The remainer of the algorithm consists of changing y_i until the b_i's are "fully utilized" at the current activity level.

Let x^k be the activity level that solves

$$\inf_{x\in\mathcal{D}} \left[f(x) + \sum_{i=1}^{m} y_i^k \, g_i(x) \right]$$

If for all i , where $y_i^k = 0$, $g_i(x^k) \leq b_i$

and for all i , where $y_i^k > 0$, $g_i(x^k) = b_i$

the algorithm is terminated.

The problem we have left is to elaborate an iterative method to update the Lagrange multipliers, this can be done by considering the y_i's as "shadow" prices. Everett has suggested an algorithm for adjusting the y_i's that has been quite successful, see further [2]. As we have seen, this procedure searches for saddle points and it will obviously not work if one does not exist, but when the procedure works it gives optimal solutions.

Several other Lagrange methods have been suggested some seek to satisfy the Kuhn-Tucker conditions, while others are based on duality theory, but practical experiences with such methods have been rather scarce.

Characteristics of the method

- This method can be very effective for problems with few constraints and large n.

- A lack of a precise statement of conditions on the programming problems that guarantee its convergence, the lack of published experience with the method, and its theoretical difficulties with certain problems keep it in the realm of conjectural rather than proved algorithms.

- This method has been used with success to solve the so-called cell problems [10], or

$$\min_{x_j \in \mathcal{D}_j} \sum_{j=1}^{n} f_j(x_j)$$

subject to

$$\sum_{j=1}^{n} g_{ij}(x_j) \leq b_i \quad , \quad i = 1, \cdots, m$$

For this problems, the total minimization of the Lagrangian can be achieved by separately minimizing

$$\left[f_j(x_j) + \sum_{i=1}^{m} y_i \, g_{ij}(x_j) \right]$$

so that $x_j \in \mathcal{D}_j$.

B) Sequential unconstrained minimization

Consider the problem of min $f(x)$ under the constraint that x must be chosen in a feasible set F, smaller than the basic set X. That is, F is a proper subset of X. One natural approach is to reformulate this as the problem

$$\min_{x \in X} [f(x) + R(x)]$$

where R is a penalty function having a behaviour at least approximating the prescription

$$R(x) = \begin{cases} 0 & \text{if } x \in F \\ +\infty & \text{if } x \in \overline{F} \end{cases}$$

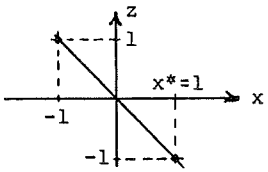
$\overline{F}$ being the complement of F in X.

In fact, it is computationally impossible to handle functions of this nature. An alternative approach is rather to consider a family of penalty functions $R(x,\lambda)$ which are well behaved for finite λ, but which tend to $R(x)$ in the limit as $\lambda \to +\infty$. One chooses a sequence $\{\lambda^k\}_1^\infty$, $\lambda^k \nearrow \infty$, $k \to \infty$, and successively minimizes the forms $\Phi(x,\lambda^k) = f(x) + R(x,\lambda^k)$ freely in X. The minimizing value x^k is used as an initial trial solution for the minimization of the next form $\Phi(x,\lambda^{k+1})$. The hope is that, as k increases, x^k will approximate arbitrarily closely to a solution of the original problem.

For the sake of illustration, consider the following example, where $x^* = 1$,

$$\min - x$$

$$x^2 \leq 1$$



Let us now define

$$\Phi(x,\lambda) = -x - \frac{1}{\lambda(x^2-1)}$$

so that x^2 is always < 1 and $\lambda \geq 0$.

Now if

$\lambda = 0.1 \rightarrow \Phi(x^*,\lambda)$ for $x^* = 0.050$

$\lambda = 1 \rightarrow \Phi(x^*,\lambda)$ for $x^* = 0.372$

$\lambda = 10 \rightarrow \Phi(x^*,\lambda)$ for $x^* = 0.778$

$\lambda = 100 \rightarrow \Phi(x^*,\lambda)$ for $x^* = 0.929$

$\lambda \rightarrow \infty \rightarrow \Phi(x^*,\lambda)$ for $x \rightarrow x^* = 1$

Let us suppose that the restrictions take the form

$$g(x) \leq \bar{b}$$

Then the function $R(x,\lambda^k)$ is usually additive, that is

$$R(x,\lambda^k) = \sum_{j=1}^{n} \lambda_j^k (g_j(x) - b_j)$$

and these functions should converge to

$$\lambda_j(\xi) = \begin{cases} 0 \ , & \xi \leq \bar{b}_j \\ \infty & \xi > \bar{b}_j \end{cases}$$

as $\lambda^k \not\to \infty$.

These $\lambda_j(\ )$'s are nothing else than the Lagrange multiplier functions of our discussion of Everett's theory, and hence the limit function [4]:

$$z_\lambda(b) = \sum_{j=1}^{m} \left[\lambda_j(b_j) - \lambda_j(\bar{b}_j)\right] + f_{sup}(\bar{b}) = \begin{cases} f_{sup}(\bar{b}) \ , & b \leq \bar{b} \\ \infty \ , & \text{otherwise} \end{cases}$$

will always solve the support problem. Thus this method as the generalized Lagrange multiplier technique seeks to find also a saddle point. The problem that is left is to discuss under which conditions x^k will converge to the optimal solution of the problem.

For the sake of concreteness, let us discuss a procedure called SUMT (Sequential underconstrained minimization technique). Consider the convex programming problem:

$$\min f(x)$$

$$\text{subject to:} \qquad g_i(x) \leq 0 \ , \quad i = 1,\cdots,m$$

The problem is solved using the auxiliary function

$$P(x,r) = f(x) - r \sum_{i=1}^{m} \frac{1}{g_i(x)}$$

defined for $x \in R^0 = \{x | g_i(x) < 0 \ , \ i = 1,\cdots,m\}$ and real $r > 0$. Choosing a real sequence $\{r_k\}_0^\infty$, $r_k \searrow 0$, $k \to \infty$, one then solves the sequence of unconstrained problems:

$$\min_{x \in R^0} P(x, r_k)$$

Let us now assume that:

(a) $R^0 \neq \emptyset$.

(b) $f(x)$, $g(x)$ are convex and twice continuously differentiable.

(c) For every finite α, $\{x | f(x) \leq \alpha, x \in R\}$ is a bounded set, where $R = \{x | g_i(x) \leq 0, i = 1, \cdots, m\}$.

(d) For each $r > 0$, $P(x,r)$ is a strictly convex function of x.

Under these assumptions it is possible to show that [5]:

- $P(x,r)$ has a unique minimum $x = x(r) \in R^0$.
- $P(x(r),r) \to f(x^*)$, $r \searrow 0$.
- $f(x(r)) \searrow f(x^*)$, $r \searrow 0$.
- If x^* is unique, $\lim_{k \to \infty} x_k = x^*$.

Characteristics of the method

- We have always to start with an initial interior point for the case of SUMT.
- The value $f(x_k) - P(x_k, r_k)$ provides an estimate of how far we are from the optimum value, this provides an excellent stop criterion, as this expression tends to zero as $k \to \infty$.

- If we have equality constraints SUMT becomes:

$$P(x,r) = f(x) - r \sum_{i=1}^{m} \frac{1}{g_i(x)} + r^{-\frac{1}{2}} \sum_{k=1}^{p} h_k(x)^2$$

and similar results as above can be found.

- The sequence $\{r_k\}$ is usually choosen as

$$r = 1, \quad r_k = \frac{r_{k-1}}{c}, \quad c > 1 \quad \text{(usually } c = 4\text{)}$$

in most standard codes.

- Despite the rather special character of its penalty function, SUMT seems to work well, judging from the considerable computing experience which has now been accumulated with it, [2].

4.3 Comparison of methods

Although most of the procedures we have discussed in this section have been used in practice, it is very difficult to compare them because their performance depends highly on the actual problem to be solved. Then given a concrete case study usually a great deal of experimentation is recommended before selecting a procedure, especially if the problem is going to be solved many times as it is the case of on-line process control.

The "best" known general procedure is now-a-days the Generalized Reduced Gradient, this is probably due to the fact that most of the accumulated experiences has been incorporated in the available standard codes. The other general procedure that has been often used in practice with many positive results is SUMT, [1].

5. REFERENCES

[1] Himmelblau, D.: Applied Nonlinear Programming, 1972.

[2] Fiacco, A., and Mc Cormick, G.: Nonlinear Programming, 1968.

[3] Aoki, M.: Introduction to optimization techniques, 1971.

[4] Gould, F.: Extensions of Lagrange multipliers in Nonlinear Programming, SIAM J. APPL. MATH., vol. 17, No. 6.

[5] Sandblom, C.: On the convergence of SUMT. Math. Prog., vol. 6, 1974.

[6] Sandblom, C.: Nonlinear Programming. University of Birmingham, 1972.

[7] Polak, E.: Computational methods in Optimization, 1971.

[8] Box, M. et al.: Nonlinear optimization techniques, 1969.

[9] Zangwill, W.: Nonlinear programming, 1969.

[10] Everett, H.: Generalized Lagrange multiplier method for solving problems of optimum allocation of resources, Operations Research, vol. 11, 1963.

[11] Fletcher, R.: Optimization, 1969.

[12] Nemhauser, G.: Introduction to dynamic programming, 1966.

[13] Jacobsen, H.K., and Pedersen, H.C.: Optimeringssystem til undervisningsbrug, IMSOR, 1976.

[14] Jenson, P.: Strongins algoritme til søgning af globalt ekstremum, IMSOR 1970.

[15] Wilde, D.: Optimum seeking methods, 1964.

[16] Christiansen, H.D.: Optimal sequential and nonsequential Monte Carlo maximization, IMSOR, 1974.

C H A P T E R 4

STATIC OPTIMIZATION:

Case studies

1. INTRODUCTION

Mathematical programming has been applied to analyse and solve many real-life decision problems of Operations Research, Economics and Engineering.

In the next sections seven typical case studies are discussed and formulated as static mathematical optimization problems. Mathematical optimization is a tool, and, like any tool, the more one understands its structure and its functional utility, the more valuable becomes. Thus, as a user of optimization procedures, one must cultivate an ability to recognize in an assigned engineering - or scientific, or economic, or managerial - situation the existence of an optimization problem, to develop the knowledge to characterize it, and to realize what practical techniques exist to solve it.

In many situations a qualitative information about the optimal solution is as important as the quantitative optimal value, therefore it is always recommended to be aware of the relations between theory and numerical methods.

The problems we will discuss are presented in their most elementary version further study can be followed by consulting the references' list.

2. CASE 1: A production planning problem

In many plants, production decisions are made periodically; let the discrete periods be denoted by the subscripts $t = 1,2,\cdots,n$. It is assumed that the demand for the product is known exactly for the next n periods. The plant manager can meet these requirements by varying the level of production with a constant-size work force, by working undertime or overtime, or by changing the size of the work force, or by changing the amount in inventory, or by some combination of these. Long range decisions, such as changing the capacity of the plants, will not be considered in this model.

The decision variables are:

w_t = the number of productive workers employed at the beginning of period t. This value will be kept constant during the whole period, and

p_t = the number of units produced during period t.

At each period the following balance equation has to be satisfied:

$$I_t = I_{t-1} + p_t - R_t$$

where

R_t = the requirement for the product, in units, during period t, and

I_t = "net" inventory in units at the end of period t (inventory minus unfilled orders).

Finally the initial conditions are given by the known values w_0 and I_0. It will be assumed that decisions are implemented at the end of a period and are effective immediately.

The objective of the plant manager is to minimize his cost in satisfying the requirements.

The costs at each period are the following:

1. Payroll costs:

$$C_{1t} = c_1 w_t \quad , \quad c_1 > 0$$

2. Hiring and firing costs:

$$C_{2t} = c_2 (w_t - w_{t-1})^2 \quad , \quad c_2 > 0$$

3. Over - or underproduction costs:

$$C_{3t} = c_3 (p_t - c_4 w_t)^2 + c_5 p_t - c_6 w_t$$
$$c_3 , c_4 , c_5 , c_6 > 0$$

4. Inventory costs:

$$C_{4t} = c_7 (I_t - c_8)^2 \quad , \quad c_7 , c_8 > 0$$

A mathematical model for our problem is then:

$$\min_{w,p \in R^{2n}} C_n(w_1,\cdots,w_n ; p_1,\cdots,p_n) = \sum_{t=1}^{n} C_t(w_t,p_t)$$

subject to:

$$I_t = I_{t-1} + p_t - R_t \quad , \quad t = 1,\cdots,n$$

given:

$$w_0 , I_0 \text{ and } R_1,\ldots,R_n$$

where:

$$C_t(w_t,p_t) = C_{1t} + C_{2t} + C_{3t} + C_{4t}$$

Note first that we assume that the components of w are continuous variables, this is realistic as far as they take relatively large values, otherwise they have to be considered as integer variables. And secondly, we do not impose in our model the restriction $w,p \geq 0$, that is we do assume that at the optimum these restrictions will be satisfied what is realistic enough for most practical cases.

We can reformulate our problem utilizing a matrix notation, thus:

$$C_t(w_t,p_t) = (c_2 + c_3\, c_4^2)w_t^2 - 2c_3\, c_4\, p_t\, w_t + c_3\, p_t^2 + c_2\, w_{t-1}^2$$
$$- 2c_2\, w_t\, w_{t-1} + (c_1 - c_6)w_t + c_5\, p_t + c_7\, I_t^2$$
$$- 2c_7\, c_8\, I_t + c_7\, c_8^2$$

Let us define our decision vector by

$$X' = [w_1,p_1,w_2,p_2,\cdots,w_n,p_n] \quad \text{and}$$

$$Z' = [I_1,I_2,\cdots,I_n]$$

Moreover:

$$E' = [c_1 - c_6 - 2c_2\, w_0\,,\, c_5\,,\, c_1 - c_6\,,\, c_5,\cdots,c_1 - c_6\,,\, c_5] \quad (1\times 2n)$$

$$G' = -2c_7\, c_8[1,1,\cdots,1] \quad (1\times n)$$

$$S' = -\left[R_1 - I_0\,,\, R_1 + R_2 - I_0,\cdots,\sum_{t=1}^{n} R_t - I_0\right]$$

and the matrices

$$B = c_7[I] \quad (n\times n)$$

$$A = \begin{bmatrix} 2c_2+c_3c_4^2 & -c_3c_4 & -c_2 & 0 & \cdots & 0 & 0 \\ -c_3c_4 & c_3 & 0 & 0 & \cdots & 0 & 0 \\ -c_2 & 0 & 2c_2+c_3c_4^2 & -c_3c_4 & \cdots & 0 & 0 \\ 0 & 0 & -c_3c_4 & c_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & c_2+c_3c_4^2 & -c_3c_4 \\ 0 & 0 & 0 & 0 & \cdots & -c_3c_4 & c_3 \end{bmatrix} \quad (2n\times 2n)$$

$$R = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 1 & \cdots & 0 \\ 0 & 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 1 & 0 & 1 & \cdots & 1 \end{bmatrix} \quad (n\times 2n)$$

Then our problem becomes:

$$\min_{X\in R^{2n}} C_n = X'AX + E'X + Z'BZ + G'Z + f$$

subject to:

$$Z = RX + S$$

where f is a constant.

FIRST SOLUTION: Optimization without constraints.

Due to the assumptions on the different constants and functions, C_n is a quadratic convex function in w and p. Therefore

$\nabla C_n = 0$, is a necessary and sufficient condition to have a global minimum.

First we eliminate the variables Z , then after rearranging terms we obtain:

$$C_n = X'(A+R'BR)X + (E' + 2S'BR + G'R)X + S'BS + G'S + f$$

since

$$(RX + S)' = X'R' + S' \quad \text{and} \quad X'R'BS = S'B'RX$$

because B is symmetric.

Then

$$\nabla C_n = 2(A + R'BR)X + E + 2R'BS + R'G$$

$$= KX + M$$

and the optimal solution is: $\underline{X^* = -K^{-1} M}$

SECOND SOLUTION: Optimization using Lagrange multipliers.

Our decision variables are now (X,Z). Then since we have to minimize a convex function subject to linear restrictions, unrestricted Lagrange multipliers will exist at the optimal solution. Moreover, Kuhn-Tucker conditions are both necessary and sufficient to have a minimum. The Lagrange function is

$$L(X,Z,\mu) = C_n(X,Z) + \mu(S - Z + RX)$$

and the Kuhn-Tucker conditions are:

$$E + 2AX + R'\mu' = 0$$

$$G + 2BZ - \mu' = 0$$

$$S + RX - Z = 0$$

The optimal solution is given by

$$\begin{bmatrix} X^* \\ Z^* \\ \mu^{*\prime} \end{bmatrix} = - \begin{bmatrix} 2A & 0 & R' \\ 0 & 2B & -I \\ R & -I & 0 \end{bmatrix}^{-1} \begin{bmatrix} E \\ G \\ S \end{bmatrix}$$

Moreover, it can be verified that at the optimal solution

$$\frac{\partial C_n}{\partial S} = \mu^*$$

the usual economical interpretation of μ^*.

THIRD SOLUTION: Iterative optimization without restrictions.

The objective function of real-life production planning problems is not quadratic and usually the approximation obtained is not satisfactory. For more complicated forms of objective functions analytical solutions are difficult or impossible to obtain. Therefore, we have to utilize computational procedures that find local minima. Which computational method to use? This is a very important decision to be taken due to the fact that the model is going to be used repetitively.

Siddall [2] and Himmelblau [3] are maybe the best references on more or less tested computer codes. We select the following methods:

a. Pattern Search [2]

b. Powell's method [3]

c. Fletcher-Reeves conjugate gradient method [3]

d. Davidon-Fletcher-Powell method [2]

e. Davidon-Fletcher-Powell method [3].

As a test example the well-known paint factory case study was used [4], where the objective function is quadratic, with

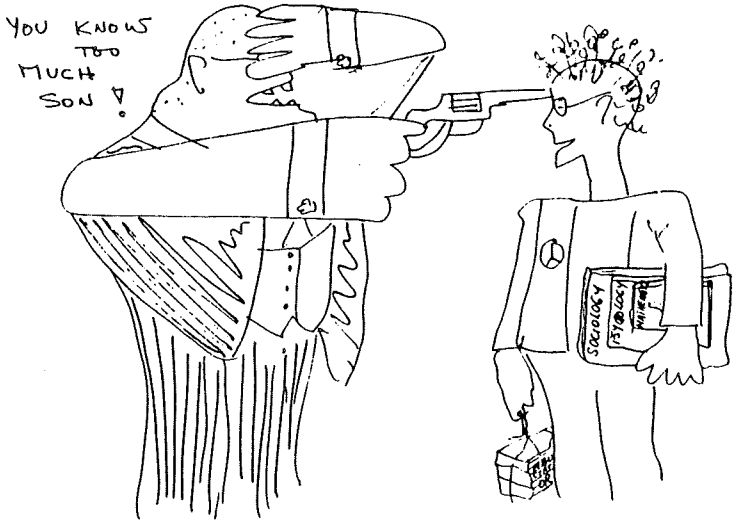
$c_1 = 340.0$, $c_2 = 64.3$, $c_3 = 0.2$, $c_4 = 5.67$ $c_5 = 51.2$,

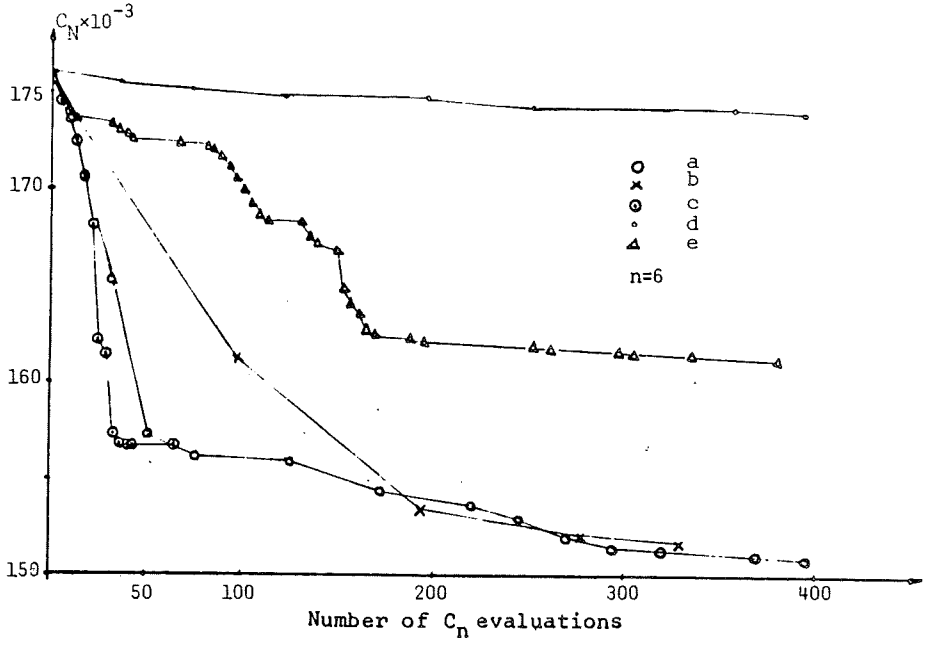
$c_6 = 281.0$, $c_7 = 0.0825$, $c_8 = 320.0$, $w_0 = 81.0$, $I_0 = 263.0$.

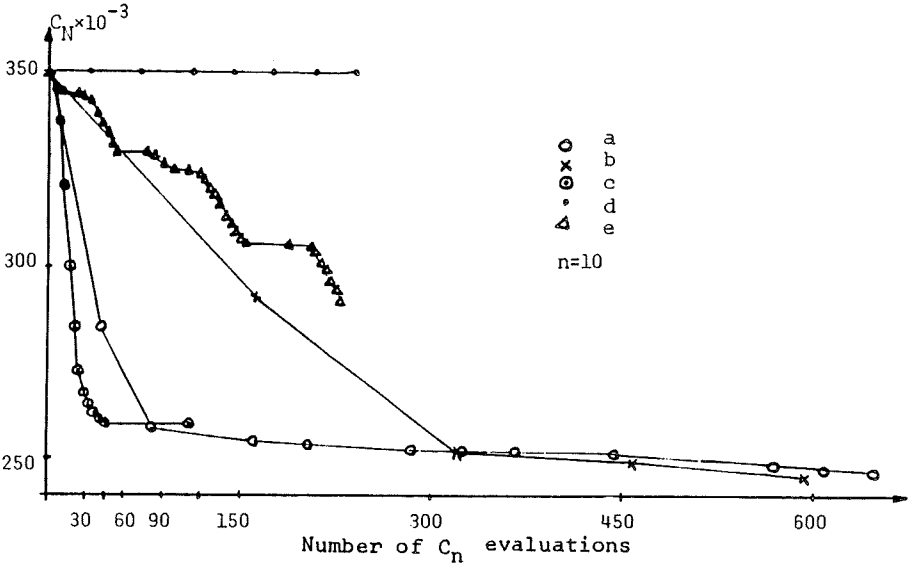
The figures below show the results obtained for $n = 6$ and $n = 10$, and the sales forecast

$R_1 = 430$, $R_2 = 447$, $R_3 = 440$, $R_4 = 316$, $R_5 = 397$, $R_6 = 375$,

$R_7 = 292$, $R_8 = 458$, $R_9 = 400$, $R_{10} = 350$.




$C_N \times 10^{-3}$
175
170
160
150
50
100
200
300
400
a
b
c
d
e
n=6
Number of C_n evaluations


$C_N \times 10^{-3}$
350
300
250
30
60
90
150
300
450
600
a
b
c
d
e
n=10
Number of C_n evaluations

The results obtained show that searching methods are better than gradient methods for our problem. This is a surprising result because theory tells us quite the opposite, thus the gradient methods should find the global optimum in n iterations. The bad behaviour of the gradient method for our problem resides in the fact that due to round-off errors the procedures get stuck in non-stationary points. The importance of round-off errors is further amplified due to the dynamical properties of our case study. Method e better performance than method d resides in the fact that the first one works with the real gradient vector while the second one uses numerical approximation to calculate the values of the gradient vector.

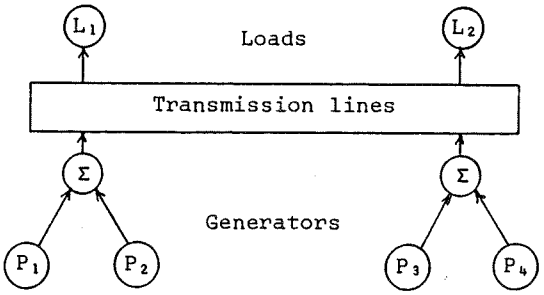
Pattern search and Powell's method are very fast and reliable. The table below shows some of the typical results obtained:

n = 10	Methods a	b	c	d	e	Analytical solution
$C_n \cdot 10^{-3}$	242	244	258.7	351.1	291.3	241.5
Computing time (sec.)	2.07	1.80	0.42	5.00	5.00	
Number of times C_n was calculated	1034	680	110	241	230	
Stop due to time trap				*	*	

3. CASE 2: Economic dispatching of interconnected power systems

This is a typical industrial application of the method of Lagrange multipliers where a set of so-called "coordination equations" is obtained.

As indicated below, interconnected power systems consist mainly of 3 parts: the generators which produce the electrical energy, the transmsision lines which transmit it, and the loads which use it.



Such a configuration applies for all interconnected networks (regional, national and international), where the number of elements vary. Since the sources of energy are so diverse (coal or gas, riverwater, marine tide, radioactive matter, sun-power), the choice of one or the other is made on economic, technical or geographic bases. The unit cost of the energy production depends on the category of the plant, and varies even from one plant to another of the same category. The transmission of the energy causes losses in the lines. For a given load, these losses depend on which plant or plants the energy is coming from. In other words, the losses depend on the complexity of the network between the generators and the load. The economic dispatching problem is then: for a pre-scribed schedule of loads, define the production level of each plant and the paths of transmission to the loads so that the total cost of production and transmission is minimum.

The decision variables are P_i, the power produced by the plant i. The mathematical model is then:

$$\min F = \sum_{i=1}^{n} F_i(P_i)$$

subject to:

$$\sum_{i=1}^{n} P_i - \sum_{i=1}^{n} P_{L_i} = P_R$$

where

$F_i(P_i)$ are the cost of producing P_i,

P_{L_i} are the losses for transmitting P_i to the loads, they are dependent on $P_1, \cdots, P_n$, and

P_R is the given total load.

Note that in this model we do not have the natural restrictions $P_i \geq 0$, that is we assume that these constraints will be automatically satisfied in real-life situations. Supposing that at the optimal solution μ, the Lagrange multipliers, exists the Lagrangian form becomes

$$\sum_{i=1}^{n} F_i(P_i) + \mu \sum_{i=1}^{n} P_i - \mu \sum_{i=1}^{n} P_{L_i} - \mu P_R$$

and defining: $P_L(P_1, \cdots, P_n) = \sum_{i=1}^{n} P_{L_i}$, then the Kuhn-Tucker necessary conditions become:

$$\frac{d\, F_i(P_i)}{d\, P_i} + \mu - \mu \frac{\partial P_L}{\partial P_i} = 0 \quad , \quad i = 1, \cdots, n$$

and

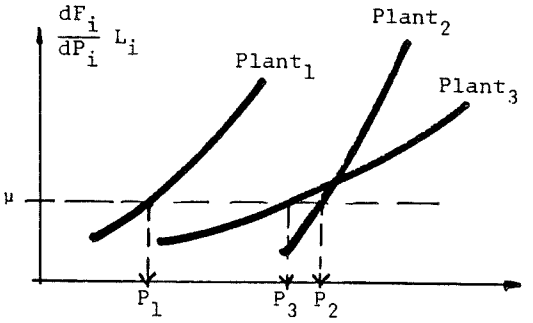
$$\sum_{i=1}^{n} P_i - P_L = P_R$$

The first set of equations can be written as

$$\frac{d\,F_i(P_i)}{d\,P_i}\,L_i = \mu \quad , \quad i = 1,\cdots,n$$

where

$$L_i = \frac{1}{\frac{\partial P_L}{\partial P_i} - 1} \quad ,$$ these are the so-called <u>coordination equations</u> which state that the levels P_i must be chosen so that all the weighted slopes $\left(\frac{d\,F_i(P_i)}{d\,P_i}\right) L_i$ of the various plants must be equal to the same quantity μ, as illustrated below:



For each value of μ, there is a combination of values of P_i. Obviously the value of μ depends on the total demand. The coordination equations permit the realization of an on-line economic dispatching control. Practical difficulties arise in obtaining the values $\frac{d\,F_i}{d\,P_i}$ and L_i considered as known quantities. The costs F_i and the cost rates $\frac{d\,F_i}{d\,P_i}$ are generally nonlinear functions of P_i, and can only be obtained by laborious experimentations. The values of L_i depend on the con-

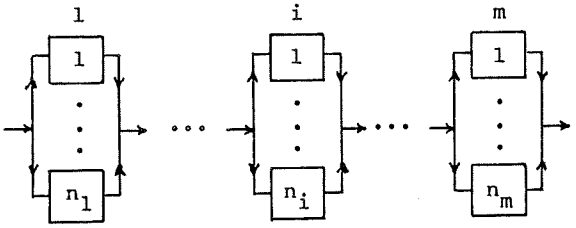
figuration of the overall network and on the knowledge of the instantaneous values of P_i. The general method of this computation is to assume

$$P_L = \sum_m \sum_n P_m B_{mn} P_n$$

where P_m and P_n are production levels, and B_{mn} coefficients to be determined according to the network configurations. The computation of L_i is also very tedious; in practice it can be aided by some network simulators.

4. CASE 3: Least-cost allocation of reliability investment

The problem is to optimize the redundancy of an m-stage system, each stage of which consists of a number n_i of parallel (redundant) components of costs c_i and reliability (availability) α_i. The separate stages are taken to be in series, so that the system is operable if, and only if, every stage contains at least one operable component. This system is illustrated below:



The allocation problem is then to choose the stage redundancies (n_i's) in such a manner as to minimize the cost of achieving some stated system reliability (or alternately, to maximize the system reliability subject to constrained total cost).

The system reliability is given by

$$A = \prod_{i=1}^{m} \left[1 - (1 - \alpha_i)^{n_i}\right]$$

since maximizing the logarithm of a positive function maximizes the function, our mathematical model is:

$$\max H = \ln A = \sum_{i=1}^{m} \ln\left[1 - (1 - \alpha_i)^{n_i}\right]$$

subject to:

$$\sum_{i=1}^{m} c_i n_i \leq C$$

$$n_i \geq 1, \quad n_i \text{ integer}, \quad i = 1, \cdots, n$$

This is a cell (or separable) problem, and in accord with the generalized Lagrange multiplier technique for a given y we have to maximize, independently for each stage (cell), the quantity

$$L_i(n_i) = \ln\left[1 - (1 - \alpha_i)^{n_i}\right] - y\, c_i\, n_i$$

over the integers $n_i \geq 1$. The allocation so produced is then guaranteed by Everett's theorem to be optimum for its cost. Since the $L_i(n_i)$'s are concave, they can be maximized by determining first analytically which real value of n_i maximizes $L_i(n_i)$, then testing the integer on each side to find which integer maximizes.

Thus

$$\frac{d\, L_i(n_i)}{d\, n_i} = -(1 - \alpha_i)^{n_i} \frac{\ln(1 - \alpha_i)}{1 - (1 - \alpha_i)^{n_i}} - y\, c_i = 0$$

leads to:

$$n_i = \frac{\ln\left\{\dfrac{1}{1 - \dfrac{\ln(1-\alpha_i)}{y\, c_i}}\right\}}{\ln(1-\alpha_i)}$$

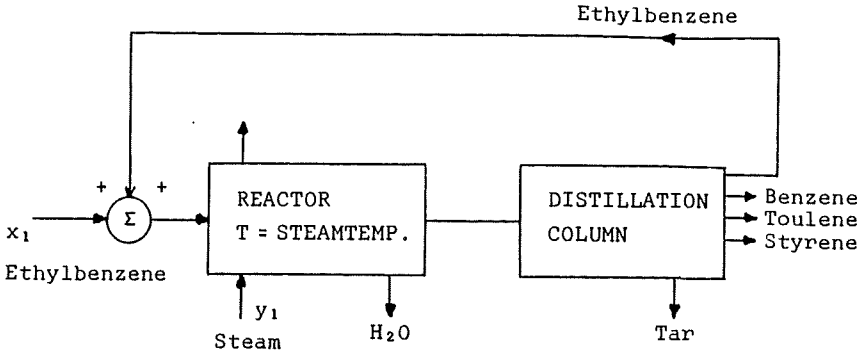
This formula is then applied to each stage, the nearest integer (not less than one), $[n_i]$ and $[n_i + 1]$, are tested to determine which maximizes $L_i(n_i)$, and the payoffs and costs summed to produce an optimal solution.

The entire procedure is repeated for a series of values of y to produce a series of optimal solutions. The amount of computation involved in producing a solution for a given y is linear with the number of stages - a considerable advantage over other methods for large-scale problems. Moreover for large-scale problems the importance of gaps is smallest. Consequently any small-scale numerical examples chosen cannot properly convey the value of the method, since for very small problems other methods are more competitive and the gaps between solutions produced by this method can have much greater significance.

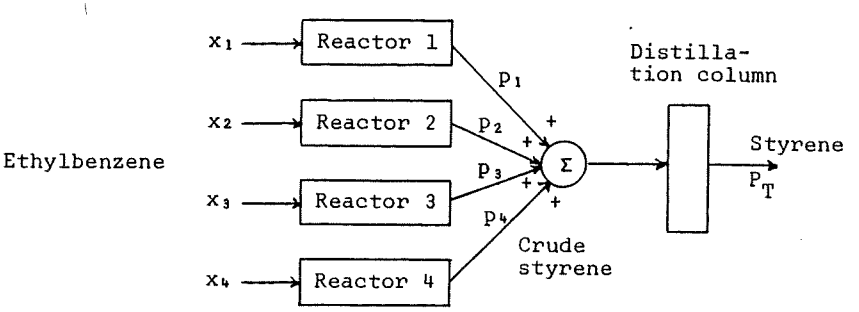
Occasions may rise where gaps occur in regions of critical interest. Under such circumstances there are several useful techniques and methods that can be attempted [7].

5. CASE 4: Optimal control of processes

A typical application of the Lagrange multiplier technique is the optimal control of the styrene manufacturing process. The figure below shows the functional model of the styrene manufacturing process:



To manufacture styrene ethylbenzene is heated in a reactor in the presence of a catalyst to form crude styrene. The crude styrene output stream from the reactor contains benzene, ethylbenzene, toulene, and styrene. The output stream from the reactor is fed to a distillation column to separate the components listed above. The ethylbenzene output of the distillation column is recycled to the reactor, the benzene is recycled to an ethylbenzene unit, and the styrene is used directly. In an actual process, there may be several reactors feeding into a common distillation column as shown below.



The basic reactions occurring in the reactors are quite simply in concept. The reaction rate depends on: temperature, catalyst age and history, and concentration of basic compounds.

The physical process model can be developed experimentally through the use of regression analysis techniques.

Let us define

x_i, $i = 1,\cdots,4$, ethylbenzene feed rate

y_i, $i = 1,\cdots,4$, steam feed rate

p_i, $i = 1,\cdots,4$, crude styrene production rate

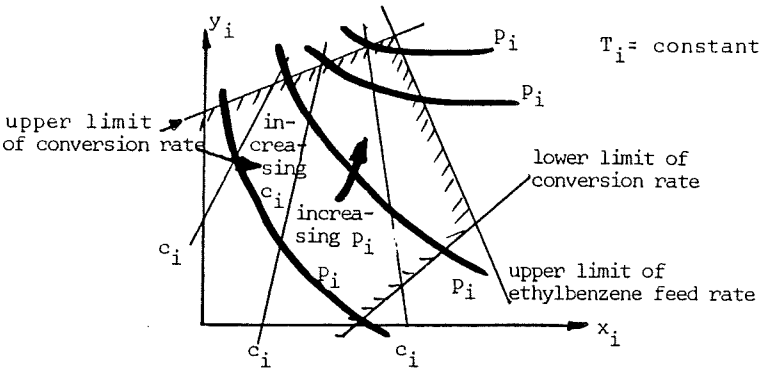
P_T, total styrene produced

C_i, $i = 1,\cdots,4$, relative production cost for reactors

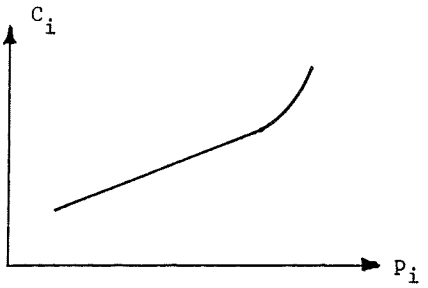
T_i, $i = 1,\cdots,4$, steam temperature of reactors

The objective of operation is to operate the four reactors at minimum cost to meet a specified total production requirement of styrene P_T. The production cost of the $i^{\underline{th}}$ reactor depends on x_i, y_i, p_i and T_i. For constant steam temperature T_i, the relative production cost C_i increases as the ethylbenzene feed rate x_i is increased. For constant steam temperature T_i and constant ethylbenzene feed rate x_i, it is possible to increase crude styrene output rate p_i by increasing catalyst deterioration and steam flow y_i. This, however, is at the expense of increasing the production cost C_i. The limits on the ethylbenzene flow rate and the upper and lower "conversion" limits are also shown below.

The "conversion" limits attempt to force the operation into a reasonable region and to conserve catalyst usage.



The most economic cost for any given crude styrene production rate is along the conversion limit. The figure below shows the relationship of relative production cost C_i plotted against crude styrene production p_i <u>for operation along the upper conversion limit</u>, thus $C_i = C_i(p_i)$



Each of the four reactors will have different characteristics of "cost versus production rate" curves. Our problem is then

$$\min \sum_{i=1}^{4} C_i(p_i)$$

subject to:

$$\sum_{i=1}^{4} p_i = P_T$$

The necessary and sufficient conditions for a global minimum are

$$\frac{\partial C_i(p_i)}{\partial p_i} = \mu \quad , \quad i = 1,\cdots,4$$

$$\sum_{i=1}^{4} p_i = P_T$$

that is at the optimum, all the reactors should operate at equal incremental costs. The minimum cost for any given specified total production can be solved iteratively by varying the values of the "incremental cost" μ, so that the sum of the production equals the total required. Constraints on the production rate for each reactor can be readily built into the iterative solution by introducing a few tests to ensure that the values p_i are within limits.

The above discussion has been centered on optimal control of well-defined processes, that is processes whose objective function, physical process model, and constraints are: steady-state, continuous-value, deterministic and well-behaved. Real-life process are often not well-defined and many times it is advisable to use on-line process control, in such situations analytical solutions are difficult to derive therefore the pattern search method has been implemented in some industrial equipment [9].

6. CASE 5: Portfolio selection

An investor has an amount which he wishes to invest. There are available n activities in which he may invest any amount. The returns (interest, dividend, etc.) from the activities differ; some consistently pay a reasonable return, other fluctuate widely. Data are available for the past m periods. Let a_{ij} denote the return per dollar invested in the $j^{\underline{th}}$ activity in the $i^{\underline{th}}$ period ($i = 1,\cdots,m$; $j = 1,\cdots,n$). The investor wishes to invest his money in the various activities in such amounts as to achieve at least a certain rate of return, r, and he wishes to minimize the deviation of his actual return from r.

This problem can be formulated as an optimization problem, if there are no other constraints other than that all money must be invested, if the total amount invested in certain sets of activities must be equal to predetermined constants, or if these conditions are stated as inequalities.

Assuem that conditions do not change; that is, on the average the pattern of returns that existed in the past can be expected to continue. Let z_j denote the proportion to be invested in the $j^{\underline{th}}$ activity, $j = 1,\cdots,n$. Then

$$\sum_{j=1}^{n} z_j = 1$$

The average yield of the $j^{\underline{th}}$ activity is

$$\bar{a}_j = \frac{1}{m} \sum_{i=1}^{m} a_{ij}$$

and the yield for any values $z_1, z_2, \cdots, z_n$ for year i is given by

$$x_i = \sum_{j=1}^{n} a_{ij} z_j$$

and the average or expected yield is

$$x = \frac{1}{m} \sum_{i=1}^{m} x_i = \sum_{j=1}^{n} \bar{a}_j z_j$$

The expected yield must equal or exceed the desired rate of return. This may be expressed as:

$$x \geq r \quad \text{or} \quad \bar{A}'Z \geq r$$

where

$$\bar{A}' = \left[\bar{a}_1, \bar{a}_2, \cdots, \bar{a}_n\right] \quad \text{and}$$

$$Z' = \left[z_1, z_2, \cdots, z_n\right]$$

The deviation (variance) is defined as

$$v = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - x)^2$$

that is

$$v = \frac{1}{m-1} \sum_{i=1}^{m} \left(\sum_{j=1}^{n} (a_{ij} - \bar{a}_j) z_j \right)^2$$

expanding

$$v = \frac{1}{m-1} \sum_{i=1}^{m} \left(b_{i1} \, b_{i1} \, z_1^2 + 2 b_{i1} \, b_{i2} \, z_1 \, z_2 + \cdots \right.$$

$$\left. + 2 b_{i1} \, b_{in} \, z_1 \, z_n + \cdots + b_{in} \, b_{in} \, z_n^2 \right)$$

where

$$b_{ij} = a_{ij} - \bar{a}_j$$

but since the terms z_j do not depend on i.

$$v = c_{11} z_1^2 + 2c_{12} z_1 z_2 + \cdots + c_{nn} z_n^2 = Z'CZ$$

where

$$c_{ij} = \frac{1}{m-1} \sum_{k=1}^{m} b_{ki} b_{kj}$$

Let $1' = [1,1,\cdots,1]$. If all the money must be invested and the desired rate of return is temporarily disregarded, the problem may be stated as:

$$\min Z'CZ$$

subject to:

$$1'Z = 1$$

Since C is a covariance matrix, $Z'CZ$ is convex, and the optimal solution is found easily by solving:

$$Z = -\tfrac{1}{2} \mu C^{-1} 1$$

$$1'Z = 1$$

If $Z^* \geq 0$ and $\bar{A}'Z^* \geq r$, the global solution to the portfolio selection problem has been found. Otherwise, the problem has to be solved as a nonlinear programming problem in which the objective function is (convex) quadratic and the constraints are linear. The statement of the general problem is

$$\min Z'CZ$$

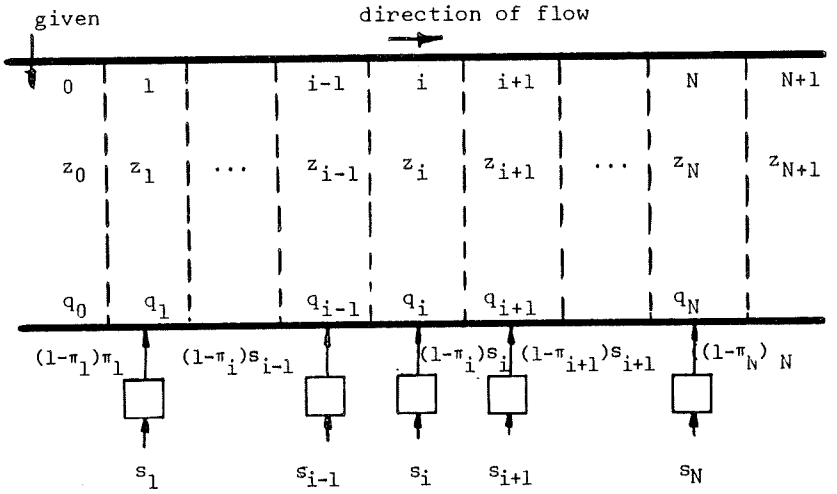
subject to:

$$1'Z \leq 1$$
$$\bar{A}'Z \geq r$$
$$Z \geq 0$$

This is a so-called quadratic programming problem. Several general procedures have been suggested to solve this type of problems, see for instance Boot [10]. Moreover this is a "nice" problem with only 2 restrictions and most of the well-known computational methods will give a solution. Since, it is desirable to generate solutions for different values of r, Everett's method seems to be the most suitable.

7. CASE 6: Control of water quality in a stream

The figure below is a schematic diagram of a part of a stream into which n sources (industries and municipalities) discharge polluting effluents.



The pollutants consist of various materials, but for simplicity we assume that their impact on the quality og the stream is measured in terms of a single quantity, namely the biochemical oxygen demand (BOD) which they place on the dissolved oxygen (DO) in the stream. Since the DO in the stream is used to break-down chemically the pollutans into harmless substances,

the quality of the stream improves with the amount of DO and decreases with in creasing BOD. It is a well-advertized fact that if the DO drops below a certain concentration, then life in the stream is seriously threatened; indeed, the stream can "die". Therefore, it is important to treat the effluents before they enter the stream in order to reduce the BOD to concentration levels which can be safely observed by the DO in the stream. The problem is then to find the optimal balance between costs of waste treatment and costs of high BOD in the stream.

Let us first derive the equations which govern the evolution in time of BOD and DO in the n areas of the streams. The fluctuations of BOD and DO will be cyclical with a period of 24 hours. Hence, it is enough to study the problem over a 24-hour period. We divide this period into T intervals, $t = 1,\cdots,T$.

During interval t and in area i let

$z_i(t)$ = concentration of BOD (mg/litre)

$q_i(t)$ = concentration of DO (mg/litre)

$s_i(t)$ = concentration of BOD of effluent discharge in mg/litre, and

$m_i(t)$ = amount of effluent discharge in litres.

The principle of conservation of mass gives us

$$z_i(t+1) - z_i(t) = -\alpha_i z_i(t) + \frac{\Psi_{i-1} z_{i-1}(t)}{v_i} - \frac{\Psi_i z_i(t)}{v_i} + \frac{s_i(t)m_i(t)}{v_i} \qquad (1)$$

$$q_i(t+1) - q_i(t) = \beta_i \left(q_i^s - q_i(t)\right) + \frac{\psi_{i-1}\, q_{i-1}(t)}{v_i} - \frac{\psi_i\, q_i(t)}{v_i} + \alpha_i\, z_i(t) - \eta_i\, v_i \qquad (2)$$

$t = 1,\cdots,T$

$i = 1,\cdots,N$

where

v_i = volume of water in area i (litres)

ψ_i = volume of water which flows from area i to area i+1 in each period (litres)

α_i = rate of decay of BOD per interval (this decay occurs by combination of BOD and DO)

β_i = rate of generation of DO (the increase in DO is due to various natural oxygen-producing biochemical reactions in the stream)

q^s = saturation level of DO in the stream, and

η_i = DO requirement in the bottom sludge.

All these are parameters of the stream and are assumed known. They may vary with the time interval t. Also $z_0(t)$ and $q_0(t)$ are assumed known. Finally, the initial concentrations $z_i(1)$, $q_i(1)$, $i = 1,\cdots,N$ are also given values.

Now suppose that the waste treatment facility in area i removes in interval t a fraction $\pi_i(t)$ of the concentration $s_i(t)$ of BOD. Then (1) becomes

$$z_i(t+1) - z_i(t) = -\alpha_i\, z_i(t) + \frac{\psi_{i-1}\, z_{i-1}(t)}{v_i} - \frac{\psi_i\, z_i(t)}{v_i} + \frac{\left(1-\pi_i(t)\right)s_i(t)m_i(t)}{v_i} \qquad (3)$$

We now turn to the cost associated with waste treatment and pollution. The cost of waste treatment can be readily identified. In period t the $i^{\underline{th}}$ facility treats $m_i(t)$ litres of effluent with a BOD concentration $s_i(t)$ mg/litre of which the facility removes a fraction $\pi_i(t)$. Hence, the cost in period t will be $f(\pi_i(t),s_i(t),m_i(t))$ where the function must be monotonically increasing in all of its arguments. We further assume that f is convex.

The cost associated with increased amounts of BOD and reduced amounts of DO are much more difficult to quantify since the stream is used by many institutions for a variety of purposes (e.g., agricultural, industrial, municipal, recreational), and the disutility caused by a decrease in the water quality varies with the user. Therefore, instead of attempting to quantify these costs let us suppose that some minimum water quality standards are set. Let $\underline{q}$ be the minimum acceptable DO concentration and let $\bar{z}$ be the maximum permissible BOD concentration. Then we face the following problem:

$$\max \; - \sum_{i=1}^{N} \sum_{t=1}^{T} f_i(\pi_i(t),s_i(t),m_i(t))$$

subject to: (2), (3) and

$$q_i(t) \geq \underline{q} \quad , \quad \forall(i,t)$$

$$z_i(t) \leq \bar{z} \quad , \quad \forall(i,t)$$

$$0 \leq \pi_i(t) \leq 1 \quad , \quad \forall(i,t)$$

Suppose that all the treatment facilities are in the control of a single public agency. Then assuming that the agency is required to maintain the standards $(\underline{q},\bar{z})$ and it does this at minimum cost it will solve the last mathematical model and arrive at an optimal solution.

It may be acceptable to tax individual polluters in proportion to the amount of pollutants discharged by the individual.

The question we now pose is whether there exist tax rates such that if each individual polluter minimizes its own total cost, then the resulting water quality will be acceptable and, furthermore, the resulting amount of waste treatment is carried out at the minimum expenditure of resources.

It should be clear from the duality theory that the answer is in the affirmative. To see this let $w_i(t) = \big(z_i(t), -q_i(t)\big)'$, let $w(t) = \big(w_1(t), \cdots, w_N(t)\big)$, and let $w = (w(1), \cdots, w(T))$. Then we can solve (2) and (3) for w and obtain

$$w = b + Ar \tag{5}$$

where the matrix A and the vector b depend upon the known parameters and initial conditions, and r is the NT-dimensional vector with components $\big(1 - \pi_i(t)\big)s_i(t)m_i(t)$. Note that the coefficients of the matrix must be non-negative because an increase in any component of r cannot decrease the BOD levels and cannot increase the DO levels. Using (5) we can rewrite (4) as follows:

$$\text{Maximize} \quad -\sum_i \sum_t f_i\big(\pi_i(t), s_i(t), m_i(t)\big)$$

$$\text{subject to:} \quad b + Ar \leq \bar{w}$$

$$0 \leq \pi_i(t) \leq 1 \quad , \quad i = 1, \cdots, N \; ; \quad t = 1, \cdots, T$$

where the 2NT-dimensional vector $\bar{w}$ has its components equal to $-\underline{q}$ or $\bar{z}$ in the obvious manner. By the duality theorem there exists a 2NT-dimensional vector $\lambda^* \geq 0$, and an optimal solution $\pi_i^*(t)$, $i = 1, \cdots, N$, $t = 1, \cdots, T$, of the problem:

$$\text{Maximize } -\sum_i \sum_t f_i\left(\pi_i(t), s_i(t), m_i(t)\right) - \lambda^{*\prime}(b + Ar - w) \tag{7}$$

$$\text{subject to: } 0 \leq \pi_i(t) \leq 1, \quad i = 1,\cdots,N; \quad t = 1,\cdots,T$$

such that $\{\pi_i^*(t)\}$ is also an optimal solution of (6) and, furthermore, the optimal values of (t) and (7) are equal. If we let $p^* = A'\lambda^* \geq 0$, and we write the components of p^* as $p_i^*(t)$ to match with the components $(1 - \pi_i(t))s_i(t)m_i(t)$ of r we can see that (7) is equivalent to the set of NT problems:

$$\text{Maximize } -f_i\left(\pi_i(t), s_i(t), m_i(t)\right) - p_i^*(t)\left(1 - \pi_i(t)\right)s_i(t)m_i(t)$$

$$0 \leq \pi_i(t) \leq 1 \tag{8}$$

$$i = 1,\cdots,N; \quad t = 1,\cdots,T$$

Thus $p_i^*(t)$ is the optimum tax per mg of BOD in area i during period t.

Note that the optimum dual variable or shadow price λ^* plays an important role in a larger framework. We noted earlier that the quality standard $(\underline{q}, \bar{z})$ was somewhat arbitrary. Now suppose it is proposed to change the standard in the $i^{\underline{th}}$ area during period t to $\underline{q} + \Delta q_i(t)$ and $\bar{z} + \Delta z_j(t)$. If the corresponding components of λ^* are $\lambda_i^{q*}(t)$ and $\lambda_i^{z*}(t)$, then the change in the minimum cost necessary to achieve the new standard will be approximately $\lambda_i^{q*}(t)\Delta q_i(t) + \lambda_i^{z*}(t)\Delta z_i(t)$. This estimate can now serve as a basis in making a benefit/cost analysis of the proposed new standard.

8. CASE 7: Linear Programming

We will apply our theoretical results of chapter 1 and 2 to the well-known family of mathematical programming problems known as LP-problems.

General results

Let us consider the following LP-problem

$$(P1): \quad \max f(x) = k' x$$
$$g(x) = Ax \leq \bar{b}$$
$$\mathcal{D}: \quad x \geq 0$$

where k and x are n-dimensional column vectors, and A is a $(m \times n)$ matrix.

Since linear functions are both convex and concave, and X, as will be seen below, is convex, then the first theorem becomes:

Theorem 1

Any local maximum is also a global maximum. The set of points at which $f(x)$ takes on its global maximum is a convex set.

Further:

Theorem 2

If X is non-empty and bounded, then $f(x)$ has a maximum.

Since: $\nabla f(x) = k'$, and assuming that $k \neq 0$, we can stipulate

Theorem 3

If a LP-problem has a solution, then it is a boundary point of X.

Theorem 4

If X is non-empty and bounded, then $\nabla f(x^*)s \leq 0$, for all feasible directions from x^*.

If the gradient of an active constraint, $\nabla g_j(x)$, is parallel to k', and if $g_j(x) = \bar{b}$ contains feasible x, then all the x that satisfies all the other constraints $g_i(x)$, $i \neq j$, and lay on the hyperplane $g_j(x) = \bar{b}$ are optimal, since for all these x^* we obtain $\nabla f(x^*)s \leq 0$. That is we get the well-known result.

Theorem 4[a]

The solution to a LP-problem is a convex set in a n-dimensional hyperplane. This convex set can be a part of a plane, a segment of a line or a point.

Note that Theorem 5 and 6 cannot be applied to LP.

Theorem 7

The feasible set X of a LP-problem is a convex set.

Everett's theory

In principle we could work with nonlinear Lagrange multiplier functions, but we will restrict us to linear functions, e.g. $\lambda(\varepsilon) = y\varepsilon$, where y is a row vector, and $y \geq 0'$, while $\mu(\varepsilon) = u\varepsilon$.

Obviously Theorem 8 and Theorem 9 are applicable, but are they always applicable? Another important question to be discussed below is whether (P1) can be solved with the help of (P2).

Let us first consider a LP-problem with only equality constraints:

$$\begin{aligned} \max\ & k'x \\ & h(x) = Ax = \bar{c} \\ \mathcal{D}: \quad & x \geq 0 \end{aligned}$$

Then the homogeneous strong Lagrangian principle says:

> There exists a vector u and a non-negative scalar w such that $(wk' - uA)x$ is maximal for $x \geq 0$, at x^*, the solution of
>
> $$\max\{h'x \mid x \geq 0 \wedge Ax = \bar{c}\}$$

The modified (P2) that has to be solved in the homogeneous strong Lagrange principle is again a LP-problem whose solution is either $x_i^* = 0$ or $x_i^* =$ undefined, $\forall i$. Therefore the Lagrange method cannot be applied to solve LP-problems since variation in u will not cause smooth variations in x and will not necessarily give a unique defined x. Note that for $w = 1$, $(k' - uA)$ are the so-called reduced costs of the simplex method.

Theorem 10

x^* is a solution to (P1) if and only if there is a vector u^* that maximizes $(k' - u^*A)x$ over $x \geq 0$, and $Ax^* = \bar{c}$.

For the case of inequality constraints we have to add the condition:

$$y^*(\bar{b} - Ax^*) = 0$$

the well-known complementary slackness condition.

Since $f_{sup}(c)$ is a concave function (to be shown) we can now stipulate:

Theorem 11

The homogeneous strong Lagrangian principle is valid for LP-problems that have a solution. If $f_{sup}(c)$ is differentiabel

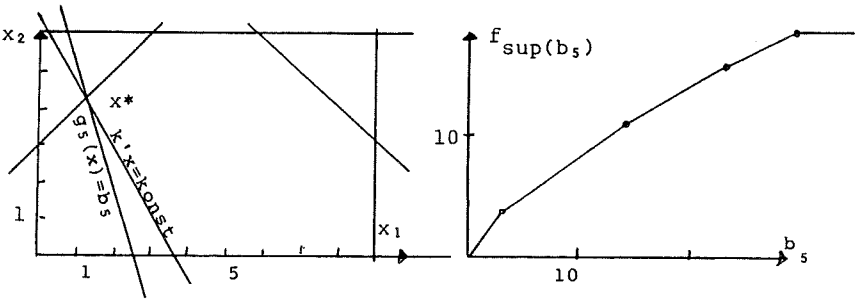
then with $w = 1$,

$$u = \nabla f_{sup}(\bar{c})$$

The optimal u and y are shadow prices for their respective constraints.

Let us consider the following example with only equality constraints, where

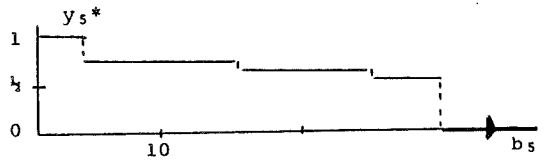
$$k' = (2,1) \quad , \qquad A = \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & 3 \end{pmatrix} \qquad \begin{pmatrix} 3 \\ 6 \\ 12 \\ 9 \\ b_5 \end{pmatrix}$$



The figures above show that the perturbation function is piecewise linear as a function of b_5. Note that the corners appear when a basis is changed. Moreover we verify that $f_{sup}(b)$ is concave, that is so because LP-problems satisfy the sufficient conditions, e.g. $\mathcal{D}$ convex, $f(x)$ concave and $g(x), h(x)$ linear.

At the corners, the Lagrange multipliers are not uniquely defined but are limited by the "left" and "right" derivative (Theorem 12).

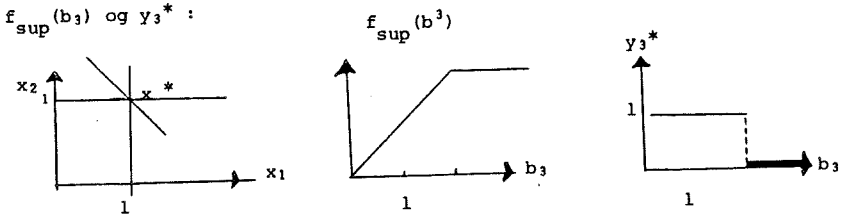
Note that in the last example the y* changes by jumps, this is illustrated below:



Let us now consider the problem

$$k = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad \bar{b} = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}$$

Then $f_{sup}(b_3)$ and y_3^* are:



Since $0 \leq y_3^* \leq 1$, and drawing $f_{sup}(b_1)$ and $f_{sup}(b_2)$, we will find out the y_1^* and y_2^* in the same interval, but it is only when we draw $f_{sup}(b)$ in R^3 that we can be able to find the relations between the y_j^*'s:

$$0 \leq y_3^* \leq 1 \quad , \quad y_1^* = 1 - y_3^* \quad , \quad y_2^* = 1 - y_3^*$$

These y_j^*'s are just the solution to the following dual LP-problem

$$\begin{aligned} \min \quad & y_1 + y_2 + 2y_3 \\ & y_1 \qquad + \quad y_3 \geq 1 \\ & \qquad y_2 + \quad y_3 \geq 1 \\ & y \geq 0' \end{aligned}$$

We will see below that this is a general result for LP-problems.

We will concentrate Theorems 13, 14 and 15 in the following one

Theorem

(x^*,y^*) is a saddle point if and only if x^* is a solution to (P1) and

$$y^*(\bar{b} - Ax^*) = 0$$

That is if a constraint is not active then $y_j^* = 0$, and if $y_j^* > 0$, then a constraint is not active.

Let us now study the connection between the primal and dual problems. For (P1) we obtain $L = k'x + y(\bar{b} - Ax)$, that has to be minimized with respect to y. This minimum exists only when $\bar{b} - Ax \geq 0$, then y_j^* is either zero or undefined, and $L_*(x) = k'x$. Thereafter we have to maximize $L_*(x)$ with respect to x, subject to $\bar{b} - Ax \geq 0$ and $x \geq 0$, that is we obtain (P1).

To solve the dual problem, $L = y\bar{b} + (k' - yA)x$, that has a maximum when $(k' - yA) \leq 0$, x_i^* has to be either zero or undefined, and $L^*(y) = y\bar{b}$. Then the dual problem becomes

$$\begin{aligned} \min \quad & y\bar{b} \\ & yA \geq k' \\ & y \geq 0' \end{aligned}$$

That is, if x^* and y^* are solutions to the primal and dual problems respectively, then (x^*,y^*) is a saddle point and viceversa (Theorem 18). Now if one of the problems primal/dual has an unbounded solution then the other has no solution. Moreover it is possible to have a primal and its respective dual so that both have no solution.

Theorem 16

$$L_*(x) \leq L^*(y)$$

Theorem 17

The dual problem of a LP-problem is also a LP-problem.

Note that for LP the dual problem can be formulated independent of x. We can now formulate the well-known Complementary Slackness Theorem of LP.

Theorem

x^* and y^* are solutions to the primal and dual problems, if and only if x^* and y^* are primal and dual feasible respectively and

$$y^*(\bar{b} - Ax^*) = 0$$

$$(y^*A - k')x^* = 0$$

Kuhn-Tucker Theory

Since in the LP-problem all the constraints are linear, the Kuhn-Tucker constraint qualification is satisfied, and since f and g are differentiabel we obtain

Theorem 21

x^*, the solution to the LP-problem satisfies Kuhn-Tucker conditions.

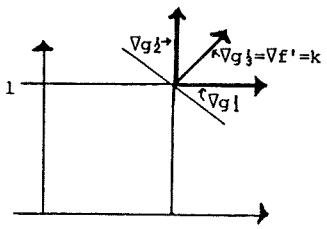
The stipulation of Kuhn-Tucker conditions will carry us to the same results obtained with duality theory. Thus if λ is denoted by y, we obtain the Kuhn-Tucker conditions:

$$\frac{\partial L}{\partial x} = k' - yA \leq 0' \qquad \frac{\partial L}{\partial y} = \vec{b} - Ax \geq 0$$

$$\frac{\partial L}{\partial x} x = (k' - yA)x = 0 \qquad y \frac{\partial L}{\partial y} = y(\vec{b} - Ax) = 0$$

$$y \geq 0' \qquad x \geq 0$$

This is nothing else but the complementary slackness theorem, e.g. the Kuhn-Tucker conditions are also sufficient.



For the problem

$$k' = (1,1) \quad , \quad A' = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}' \quad , \quad \bar{b}' = (1,1,2)$$

we note that Kuhn-Tucker conditions expresses ∇f as a linear combination of the ∇g_j. In this example ∇f can be expressed in infinite many ways as a linear combination, as far as the Lagrange multipliers satisfy

$$0 \leq y_3 \leq 1 \quad , \quad y_1 = 1 - y_3 \quad , \quad y_2 = 1 - y_3$$

Note also that the y_j's are shadow prices, but these are in nature obtained differently than those obtained with Everett's theory.

We can conclude that there is a beautiful correspondence between the results we have obtained in relation to the y's, with different approaches:

- as a (super) gradient to $f_{sup}(b)$
- as optimal dual variables, and
- as Kuhn-Tucker coefficients.

LP-problems are a family of problems where both Everett's theory and Kuhn-Tucker theory can be applied, but we cannot find analytically the optimal solution, we can only verify if a given solution is optimal. This is highly related to the piecewise linearity of the perturbation function, and the piecewise constancy of the Lagrange multipliers.

9. REFERENCES

CASE 1

[1] Valqui Vidal, R.V.: Operations Research in Production Planning, IMSOR 1970.

[2] Siddall, J.: OPTISEP, designers optimization, Mc Master University, 1970.

[3] Himmelblau, D.: Applied Nonlinear Programming, 1972.

[4] Holt et al.: Planning production, inventories and work force, 1960.

CASE 2

[5] Kirchmayer, L. : Economic Control of Interconnected systems, 1959.

[6] Pun, L.: Introduction to Optimization Practice, 1969.

CASE 3

[7] Everett, H.: Generalized Lagrange multiplier method for solving problems of optimum allocation of resources, Operations Research, 1963.

[8] Kettelle, J.: Least-Allocation of Reliability Investment, Operations Research, 1962.

CASE 4

[9] Lee, T., et al.: Computer Process Control: Modeling and Optimization, 1968.

CASE 5

[10] Boot, J.: Quadratic Programming, 1964.

[11] Markowitz, H.: Portfolio selection, 1959.

CASE 6

[12] Kendrich, H., et al.: Water Quality Regulation with Multiple Polluters. Proc., 1971, JT. Autom. Control Conference.

[13] Solow, R.: The economist's approach to Pollution and its control, Science, 1973.

PART TWO

DYNAMIC OPTIMIZATION

MATAJURA wanted to become a great swordsman, but his father said he wasn't quick enough and could never learn. So Matajura went to the famous dueller Banzo, and asked to become his pupil. "How long will it take me to become a master?" he asked. "Suppose I become your servant, to be with you every minute; how long?"

"Ten years," said Banzo.

"My father is getting old. Before ten years have passed I will have to return home to take care of him. Suppose I work twice as hard; how long will it take me?"

"Thirty years," said Banzo.

"How is that?" asked Matajura. "First you say ten years. Then when I offer to work twice as hard, you say it will take three times as long. Let me make myself clear: I will work unceasingly: no hardship will be too much. How long will it take?"

"Seventy years," said Banzo. "A pupil in such a hurry learns slowly."

CHAPTER 5

DYNAMIC OPTIMIZATION:

Fundamentals of optimal control theory

1. INTRODUCTION

During the past fifteen years, an intense amount of research has been carried out in the area of dynamic optimization. Many fine books have been published which deal, in whole or in part, with dynamic optimization techniques. In addition, many educational institutions are offering formal courses in the area of optimal control and its applications; such courses are well attended by students in all branches of engineering, mathematics, economics, operations research, management, etc. Moreover, many real life problems of significance have been attacked using the concepts, techniques, and tools provided by "modern control theory". Thus, at the present time, the field of dynamic optimization has reached a certain state of maturity, and it is regarded as one of the areas of most fervent research.

Optimal control theory has as its objective the maximization of the return from, or the minimization of the cost of, the operation of physical, social, and economic processes. Thus the field of dynamic optimization focus its attention to systems that are evoluating over time. The need to solve optimization problems has appeared within different areas, as for instance: mathematics (calculus of variations), the design of aerospace systems, process control, economics and operations research. From a historical viewpoint, the need to additional theory in the area of dynamic optimization arose, to a large degree, from the stringent requirements of aerospace systems. Before the development of what is today commonly called 'modern control theory', the only other body of theory available for control was that dealing with the analysis and design of servomechanisms. However, the theory of servomechanisms was not adequate for the analysis and design of aerospace system, due to the fact that most available tools could be applied only to the analysis and design of unconstrained, single-output, linear and stationary systems. The fact that it may be economically feas-

ible to use a digital computer as part of the control system for a complicated process, the need for more accurate designs, and the availability of powerful computational facilities naturally created the need for additional theoretical knowledge for understanding, analysing, and designing complex, non-linear, and time-varying systems subject to constraints.

The purpose of this chapter is to introduce some fundamentals, concepts and definitions of optimal control theory, as well as to specify the type of problems we will center around in the following chapters, but let us first discuss some elementary case studies.

2. SOME EXAMPLES

a) Mr. Thorkild is the manager of an economy which produces one output, wine. If $K(t)$ and $L(t)$ respectively are the capital stock used and the labour employed at time t, then the rate of output of wine $W(t)$ at time t is given by the production function: $W(t) = F(K(t),L(t))$.

As manager, Mr. Thorkild allocates some of the output rate $W(t)$ to the consumption rate $C(t)$, and the remainer $I(t)$ to investment in capital goods. Obviously W, C, I and K are being measured in a common currency. Thus,

$$W(t) = C(t) + I(t) = (1-s(t))W(t) + s(t)\cdot W(t)$$

where $s(t) = I(t)/W(t) \in [0,1]$ is the fraction of output which is saved and invested. Suppose that the capital stock decays exponentially with time at rate $\delta > 0$, so that

$$\dot{K}(t) = \frac{dK(t)}{dt} = -\delta K(t) + s(t)W(t)$$

$$-\delta K(t) + s(t)F(K(t),L(t))$$

The labour force is growing at a constant birth rate $\beta > 0$. Hence, $\dot{L}(t) = \beta L(t)$.

Suppose that the production function F exhibits constant returns to scale, i.e., $F(\lambda K,\lambda L) = \lambda F(K,L)$, $\forall \lambda > 0$. If we define the relevant variables in terms of per capita labour, $w = W/L$, $c = C/L$, $k = K/L$, and if we let $f(k) = F(K,1)$, then we see that $F(K,L) = LF\left(\frac{K}{L}, 1\right) = Lf(k)$, whence the consumption per capita of labour becomes $c(t) = (1-s(t))f(k(t))$. Using these definitions and the last equations we see that the economy will be described by the differential equation $\dot{k}(t) = s(t)f(k(t)) - \mu k(t)$ where $\mu = \delta + \beta$.

Suppose there is a planning horizon time T, and at time 0 Mr. Thorkild starts with a capital-to-labour ratio k_0. If 'welfare' over the planning period $[0,T]$ is identified with total consumption $\int_0^T c(t)dt$, what should Mr. Thorkild savings policy $s(t)$, $0 \leq t \leq T$, be so as to maximize welfare?

This control problem can be written as

$$\max_{\{s(t)\}} J = \int_0^T (1-s(t))f(k(t))dt$$

subject to: $\dot{k}(t) = s(t)f(k(t)) - \mu k(t)$

$0 \leq s(t) \leq 1$

given $k(0) = k_0$ and T

where

- J is the criterion or performance function,
- $k(t)$ is the <u>state</u> variable, and
- $s(t)$ is the <u>control</u> variable (or decision variable).

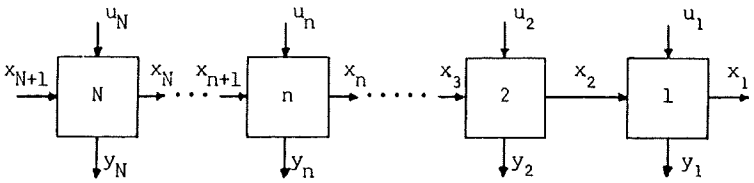
This decision problem differs from those considered under static optimization in that the decision variable, which is

function $s : [0,T] \to R$, cannot be represented as vectors in a finite dimensional space. This is a continuous-time, continuous-state control problem [1].

b) Another classic example of the control problem that can be formulated as the last example is that of determining optimal missile trajectories. In this problem the control variables are the timing, magnitude, and direction of various thrusts that can be exerted on the missile. These thrusts are chosen subject to certain constraints; for example, the total amount of propellant available. The state variables, which describe the missile trajectory, are the mass of the missile and the position and the velocity of the missile relative to a given coordinate system. The influence of the thrusts on the state variables is summarized by a set of differential equations obtained from the laws of physics. The mission to be accomplished is then represented as the maximization of a performance measure [2]. This is also a continuous-time, continuous-state control problem.

c) Extraction is a chemical engineering operation whereby a compound or compounds in liquid solution with one or more other compounds is removed from the solution by contacting it with a second liquid known as an extracting solvent. The compound in question dissolves in the extracting solvent, which is then removed by physical means from contact with the original liquid.

There are a number of ways in which this process is carried out in practice. The figure below depicts one such way, whereby N fresh streams of solvent are added at N individual locations in a cross-flow manner to the stream of feed solution. Such an arrangement is known as cross-current extraction. Each of the N contact points is considered to be an ideal stage in which the mixing between feed and solvent is accomplished and then separation of stream is effected.



The solute is assumed to equilibrate between the two streams. A solute material balance around the n^{th} stage is:

$$q(x_{n+1}-x_n) = w_n\, y_n \quad , \quad n = 1,2,\cdots,N$$

where q is the flow rate of feed solution, x_n is the concentration of solute in raffinate, y_n is the concentration of solute in extract and w_n is the flow rate of solvent. The equilibrium expression relating y to x may be expressed generally by stating $y_n = y_n(x_n)$.

Let $u_n = w_n/q$, so that the transformation equation becomes: $x_{n+1} - x_n = u_n\, y_n$, $n = 1,2,\cdots,N$. The last equations together with the equilibrium relationship constitute a set of equations which describe the behaviour of the cross-current extraction unit. It is possible to suggest several performance criteria for the extraction system; one often used is operating profit. For a typical stage, the operating profit is expressed as:

$$P_n = P_e\, w_n\, y_n - c_w\, w_n$$

where P_e is the price obtained per unit of extracted material and c_w is the cost per unit of solvent. This equation may be normalized, thus define $r_n = P_n/P_e\, q$, a normalized return, and a relative cost $\lambda = c_w/P_e$; then $r_n = u_n(y_n-\lambda)$ and the total profit becomes

$$J = \sum_{n=1}^{N} r_n$$

and it is J that we wish to maximize. J is a function of the 2 N variables u_n and x_n, $1,\cdots,N$. A logical distinction can be made between the two sets of variables on which R depends. The composition variables x_n, characterize the amount of solute in the process stream at each stage. Hence it is appropriate to refer to the x-variables as <u>state variables</u>.

The variables u_n, on the other hand, represent those quantities which physically would be manipulated to exercise control over the extraction. Hence these variables are called <u>decision or control variables</u>, [3].

This is a discrete-time continuous-state control problem.

d) Let us finally discuss a simple inventory problem. A company wants to plan its production and inventory schedule for the coming month. It is assumed that the demand on the i^{th} day, $0 \leq i \leq N$, is $d(i)$ for its product. To meet unexpected demand it is necessary that the inventory stock $s(i)$ (state variable) should not fall below $\hat{s} > 0$. If $m(i)$ denotes the amount manufactured (control variable) on the i^{th} day, then the system will be described by the difference equation

$$s(i+1) = s(i) + m(i) - d(i) .$$

Suppose that the initial stock is s_0, and the cost of storing inventory s for one day is $c(s)$ whereas the cost of manufacturing amount n is $b(m)$. Then the cost minimization control problem can be formalized as:

$$\min_{\{m(i)\}} \quad J = \sum_{i=0}^{N} [c(s(i)) + b(m(i))]$$

subject to: $\quad s(i+1) = s(i) + m(i) - d(i) \quad , \quad \forall i$

$$s(0) = s_0$$

$$s(i) \geq \hat{s} \ , \ m(i) \geq 0 \ , \ s(i) \text{ and } m(i) \text{ entier}, \quad \forall i .$$

This is a discrete-time, discrete-state control problem. If $d(t)$ is a stochastic variable with known probability distribution then our criterion will be to minimize the expected value of J, that is expected costs. This is then a stochastic control problem, [4].

3. PROBLEM FORMULATION

From the examples of the last section we can conclude that the formulation of an optimal control problem requires:

(i) specification of the state and control variables,

(ii) a mathematical description of the process to be controlled, e.g. specification of the equations of motion or transformation functions, in state variable form, of the dynamical system to be controlled,

(iii) a set of constraints on the control or/and state variables, and

(iv) specification of the performance criterion to be maximized.

An optimal control problem assumes then that we have found explicitly expression for the equation of motion, but how is this done? This, so called identification, problem could be solved by constructing a model of the process using physical, energetical, chemical and/or economical laws, but the main disadvantage of this approach resides in its complexity and lack of generality.

4. IDENTIFICATION AND CONTROL

The problem of identifying the characteristics of a system may be considered as dual to that of controlling the system. One

cannot control a system unless it has been identified, either a priori or while control is applied. The knowledge of the differential equations of a process is one possible identification, though not the only one. We may alternatively tabulate the possible controls and their respective responses at a given future time at which we are interested. There are various methods of identification, which employ a number of different concepts concerning the form of the identification model (say differential equations, difference equations, transfer functions, gradient expressions, etc.). None of the different identification techniques can be employed to identify systems of all kinds. Each of the techniques has its own ranges of applicability. It is possible to consider a theory of identification, which deals with the estimation of parameters from input and output data, i.e., from a history of measurements, and in which identification is improved with an increase in the number of measurements. Subsequently, errors in identification lead to errors in control and these errors are employed to improve further identification. Hence, identification theory is similar, or in fact dual, to control theory, where errors in control (assuming that the system has been identified) are employed to improve the next control.

Broadly speaking, one distinguishes between different identification situations as follows:

a) linear and nonlinear systems,

b) stationary and nonstationary systems, the latter being systems whose parameters vary with time,

c) discrete and continuous systems,

d) single-input and multi-input techniques,

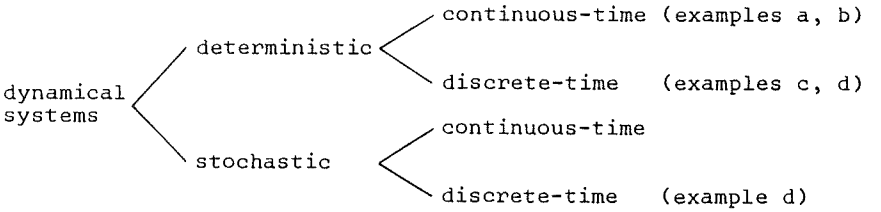
e) deterministic and stochastic processes, and

f) degree of a priori knowledge regarding the system.

In what follows we will assume that the identification problem has been solved and that we are faced with a control problem [5].

In some cases, as for instance on-line process control, this two-step approach cannot be employed and simultaneous identification and control is needed. We will not be dealing with this kind og problems that is usually discussed in the field of Adaptive Control Processes [6].

An identified dynamical system can be classified for control purposes in the following way:



In what follows we will be mainly interested in both continuous-time and discrete-time <u>deterministic</u> dynamical systems, that is in systems where all parameters and functions are assumed being deterministics. Some of the results obtained for these cases can be easily generalized to some special type of stochastic systems, this will be shown in chapter 9.

As we have seen in sec. 2 some problems are more realistic formulated as continuous-time while others are better represented as discrete-time, but sometimes due to numerical needs a continuous-time problem may be discretized and, viceversa, sometimes due to analytical needs a discrete-time system may be formulated as a continuous-time one.

5. THE MATHEMATICAL MODEL OF A CONTINUOUS-TIME OPTIMAL CONTROL PROBLEM

Let be $x_1(t), x_2(t), \cdots, x_n(t)$ the <u>state variables</u> (or simple the <u>states</u>) of the process at time t, and $u_1(t), u_2(t), \cdots, u_m(t)$ the <u>control inputs</u> to the process at time t then the system may be described by the so called <u>equations of motion</u> (transformation functions, state equations)

$$\begin{aligned}
\dot{x}_1 &= f_1(x_1(t),\cdots,x_n(t),u_1(t),\cdots,u_m(t),t) \\
\dot{x}_2 &= f_2(x_1(t),\cdots,x_n(t),u_1(t),\cdots,u_m(t),t) \\
&\vdots \\
\dot{x}_n &= f_n(x_1(t),\cdots,x_n(t),u_1(t),\cdots,u_m(t),t)
\end{aligned}$$

or, using a vector notation

$$\dot{x}(t) = f(x(t),u(t),t)$$

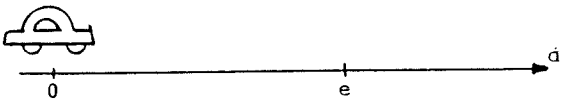
where

$x(t) = [x_1(t),\cdots,x_n(t)]' \in R^n$ is the state vector, and

$u(t) = [u_1(t),\cdots,u_m(t)]' \in R^m$ is the control vector, and

$f: R^n \times R^m \times R \to R^n$.

Example 1



A car is to be driven from 0 in a straight line. $d(t)$ is the distance of the car from 0 at time t. Suppose that the car can be approximated by a unit point mass that can be accelerated by using the throttle or decelerated by using the brake, so that

$$\ddot{d}(t) = \alpha(t) + \beta(t)$$

where the control α is throttle acceleration and β is braking deceleration. Let us choose position and velocity as state variables, that is

$$x_1(t) = d(t) \quad \text{and} \quad x_2(t) = \dot{d}(t)$$

and letting

$$u_1(t) = \alpha(t) \quad \text{and} \quad u_2(t) = \beta(t)$$

The equation of motion becomes

$$\dot{x}_1(t) = x_2(t)$$
$$\dot{x}_2(t) = u_1(t) + u_2(t)$$

or, using matrix notation

$$\dot{x}(t) = A\, x(t) + B\, u(t)$$

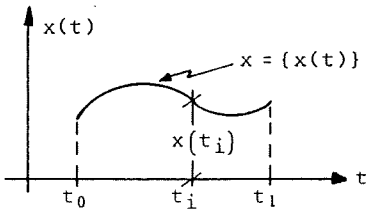
where

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}$$

This is the model of the process in state form.

A history of control input values during the interval $[t_0, t_1]$ is denoted by $u = \{u(t)\}$ and is called a control history, or simply a control. A history of state values in the interval $[t_0, t_1]$ is called a state trajectory and is denoted by $x = \{x(t)\}$. In the literature the terms "history", "curve", "function" and "trajectory" is usually used interchangeably.

It is important to keep in mind the difference between a function and the value of a function as illustrated below



The next step is to define the constraints on the state and the control vectors.

Example 2

Consider the last example of driving the car from 0 to e. Assume that the car starts from rest and stops upon reaching point e. The state constraints are

$$\left.\begin{aligned} x_1(t_0) &= 0 \\ x_2(t_0) &= 0 \end{aligned}\right\} \Rightarrow x(t_0) = 0$$

and

$$\left.\begin{aligned} x_1(t_1) &= e \\ x_2(t_1) &= 0 \end{aligned}\right\} \Rightarrow x(t_1) = \begin{bmatrix} e \\ 0 \end{bmatrix}$$

and if we assume that the car does not back up, then

$$\begin{aligned} 0 \leq x_1(t) \leq e \\ 0 \leq x_2(t) \end{aligned}$$

Moreover, if the maximum acceleration is $M_1 > 0$, and if the maximum deceleration is $M_2 > 0$, then the controls must satisfy

$$\begin{aligned} 0 \leq u_1(t) \leq M_1 \\ -M_2 \leq u_2(t) \leq 0 \end{aligned}$$

Finally, if the car starts with G liters of gas and there are no service stations on the way another constraint is

$$\int_{t_0}^{t_1} \left[k_1 u_1(t) + k_2 x_2(t)\right] dt \leq G$$

which assumes that the rate of gas consumption is proportional to both acceleration and speed.

Let us now make these concepts more precise. A control history which satisfies the control constraints during the entire time interval $[t_0,t_1]$ is called an admissible control. We shall denote the set of admissible controls by U, then $\{u(t)\}$ is an admissible control if and only if $\{u(t)\} \in U$.

A state trajectory which satisfies the state variable constraints during the entire time interval $[t_0,t_1]$ is called an admissible (state)trajectory. The set of admissible state trajectories will be denoted by X, then $\{x(t)\}$ is admissible if and only if $\{x(t)\} \in X$.

In general the final state of a system will be required to lie in a specified region S of the (n+1)-dimensional state-time space. We shall call S the target set. If the final state and the final time are fixed then S is a point

An optimal control is defined as one admissible control that maximizes (or minimizes) the performance measure of the system.

Example 3

Suppose that in our last example the objective is to make the car reach point e as quickly as possible; then the performance measure J is

$$J = -(t_1-t_0)$$

and it has to be maximized.

We will assume that the performance of the system, to be maximized, is evaluated by

$$J = h(x(t_1),t_1) + \int_{t_0}^{t_1} g(x(t),u(t),t)dt$$

where t_0 and t_1 are the initial and final time and h and g are scalar functions. t_1 may be specified or 'free'.

We can now formulate the optimal control problem.

Find an admissible control u^* which causes the system

$$\dot{x} = f(x(t),u(t),t)$$

to follow an admissible trajectory x^* that maximizes

$$J = h(x(t_1),t_1) + \int_{t_0}^{t_1} g(x(t),u(t),t)dt$$

u^* is called an optimal control and x^* an optimal trajectory.

This problem can be considered as a mathematical programming problem (static optimization) in infinite dimensional space, the space being that of all continuous real valued functions $\{u(t)\} \in U$. This can easily be seen by discretizing our problem in N intervals of length Δ, then

$$J = \lim_{\substack{N \to \infty \\ \Delta \to 0 \\ N\Delta = t_1 - t_0}} J^N$$

Most of the theorems that we have seen while discussing static optimization can be shown to be also valid for more general spaces than R^n [7].

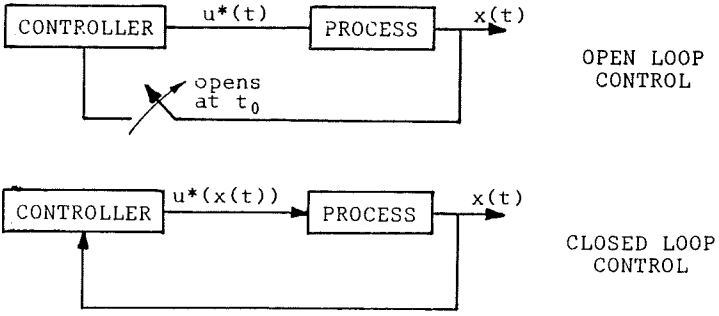
Our problem can also be considered as a generalization of the problems usually discussed in calculus of variations. Here one works with the concept of functionals and it can be shown that many of the results of the static theory have their analogous in calculus of variations [8].

Both approaches will lead us to necessary conditions that due to the special form of the optimal control problem have some very important characteristics that facilitate the evaluation of $u^* = \{u^*(t)\}$. These necessary conditions are known as Pontryagin's maximum principle [9], to be seen in chapter 6.

5.1 Types of optimal control

There are two types of control. One is open loop control or optimal control function in which the optimal trajectory $u^* = \{u^*(t)\}$, is determined as a function of time for a specified initial state value. This open loop control is completely specified at the initial time t_0, and the state trajectory x^* is determined by integrating the equations of motion.

The other type of control is closed loop control (also optimal control law, optimal feedback control, optimal control strategy, optimal control policy), in which the optimal control trajectory is determined as a function of the current state variables and time, e.g. $u^* = \{u^*(x(t),t)\}$. Here we write $x(t)$ instead of $x^*(t)$ to emphasize that the control law is optimal for all admissible $x(t)$, not just for some special state value at time t [10].



Example 4

Most clothes dryers are regulated by open loop control, by a timer which must be set in advance. A home heating system, by contrast, is typically regulated by a thermostat which turns the furnace on if the room temperature is too low and turns it off if the room temperature is too high. Thus the control of furnace depends on the current state variable, the room temperature.

The problem of obtaining closed loop control is called that of synthesis.

If the open loop control is used, in general, because of errors or disturbances, the system will not stay on the nominal open loop trajectory, $x^*(t)$, but will be in some neighbourhood about it. Thus it is necessary to determine the control perturbation, $\delta u^*(t)$, to be added to the open loop control so as to reach the desired terminal state in an optimal manner starting from the state $x^* + \delta(x)$ existing at time t. This area is termed optimal feedback control.

We will center our discussion around deterministic systems. In such cases open and closed loop control yield identical results, so the emphasis will be on open loop control, which is easier to determine. Closed loop control is generally superior to open loop control in yielding a higher maximum for the criterion in the case of stochastic control in which stochastic variables with given distributions appear in the problem, and in the case of adaptive control, in which identification and control are performed simultaneously (see further chapter 9).

Some of the most typical performance measures are:

NAME	CRITERION	COMMENTS
Minimum time or time optimal control problem	$J = -(t_1-t_0) = -\int_{t_0}^{t_1} dt$	with t_1 the first instant of time when $x(t)$ and S intersect
Terminal control problem	$J = -[x(t_1)-r(t_1)]' H[x(t_1)-r(t_1)]$ where H is a real symmetric positive semi-definite matrix	to minimize the deviation of the final state of a system from its desired value $r(t_1)$
Minimum control effort problem	$J = -\int_{t_0}^{t_1} [u'(t)R(t)u(t)]dt$ where $R(t)$ is a real symmetric positive definite weighting matrix	to transfer the system from an arbitrary initial state x_0 to a specified target S with 'a minimum control effort'
Tracking or servomechanism problem	$J = -\int_{t_0}^{t_1} \{[x(t)-r(t)]'Q(t)[x(t)-r(t)] + u(t)'R(t)u(t)\}dt$ where $Q(t)$ is a real symmetric matrix that is positive semi-definite $\forall t \in [t_0,t_1]$ and $R(t)$ is a real symmetric positive definite matrix $\forall t \in [t_0,t_1]$	to maintain the system state $x(t)$ as close as possible to the desired state $r(t)$, $\forall t \in [t_0,t_1]$ with 'a minimum control effort'
Regulator problem		This is a special case of a tracking problem for which $r(t) = 0$, $\forall t \in [t_0,t_1]$

Some of the most typical equations of motion are the following:

NAME	EQUATION OF MOTION
Non-linear and time-varying	$\dot{x}(t) = f(x(t),u(t),t)$
Non-linear and time-invariant (stationary, autonomous)	$\dot{x}(t) = f(x(t),u(t))$
Linear and time-varying	$\dot{x}(t) = A(t)x(t) + B(t)u(t)$ where A(t) and B(t) are n×n and n×m matrices
Linear and time-invariant	$\dot{x}(t) = Ax(t) + Bu(t)$

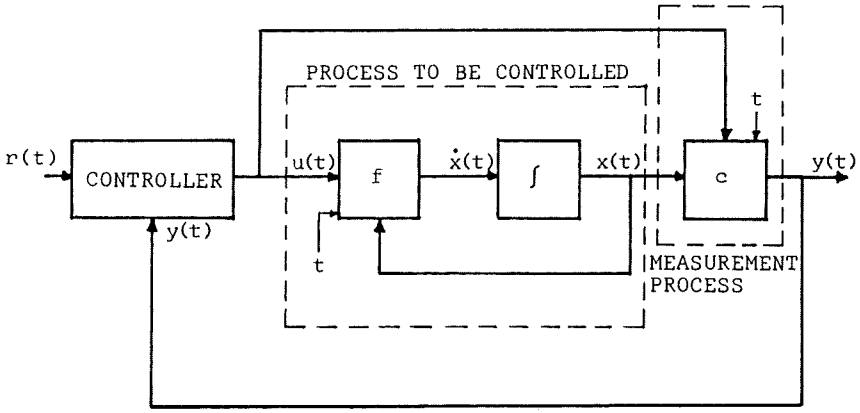
The physical quantities that can be measured are called the outputs and are denoted by

$$y(t) = [y_1(t),\cdots,y_q(t)]' . \quad \text{In general}$$

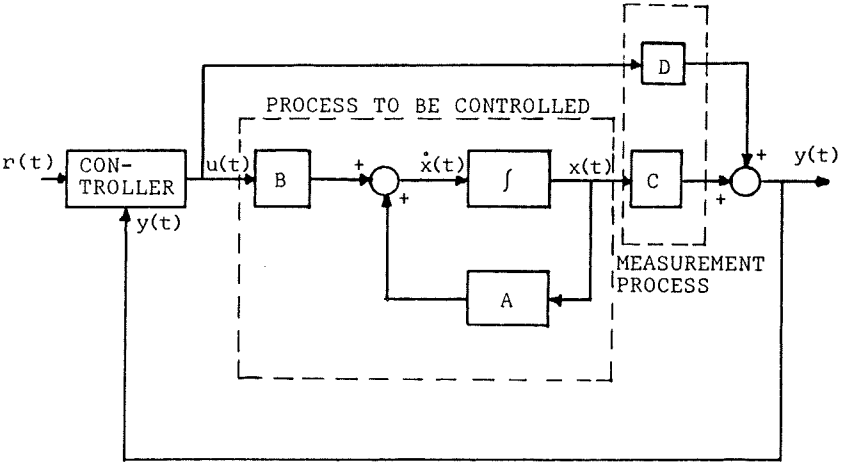
$$y(t) = c(x(t),u(t),t)$$

A graphical illustration of a control problem is shown below. $r(t)$, which has not been included in the state equations, represents any inputs that are not controlled. They are called reference or command input. In what follows we will assume that $y(t) = x(t)$ [11].

A nonlinear system representation



A linear system representation



5.3 Suboptimal control (heuristic methods)

Solutions to optimal control problems are not always easily obtained. In addition, even when it is possible to obtain the desired solution it may be impractical to implement it, that is, to construct a controller capable of duplicating the exact mathematical result. One way out of the dilemma is to solve a restricted problem in which the form of the controller that will be allowed is postulated. The problem then remaining is to choose the values of the controller parameters so as to achieve optimality within these constraints. This is the so-called specific optimal control problem.

One may ask how the specific form for the suboptimal controller is to be chosen? Here knowledge of the solution of the unrestricted original optimal control problem becomes valuable. It would seem that the suboptimal controller should be built to permit it to approximate the response that one would obtain if the optimal controller were built. Knowledge of the exact solution has one further benefit. A comparison can be made between the optimum value and the value of the performance measure obtained using the suboptimal controller. In this manner a judgment can be made on the acceptability of the suboptimal controller designed. Should the deviation be too large, further design refinements can be attempted to enable the suboptimal solution to more closely approximate the optimal one.

It is many times not possible to find an analytic representation of the optimal control law for a problem. If an analytic representation of the control law is required, a satisfactory suboptimal one may sometimes be found in the following manner. It is assumed that the physical process can be approximated by a simple one, for which an analytic representation of the control law can be found exactly. The suboptimal control law for the actual process is then obtained as a modification of the optimal control law for the simplified process.

It is important to emphasize that the value of the performance function obtained using a suboptimal controller should always be tested against that obtained using the optimal control for the actual process to check the degree of optimality of the approximate solution [12].

5.4 Controllability and observability

The concepts of controllability and observability provide necessary and in some cases sufficient conditions for a given problem to posses a solution. The concepts have been referred to as "duals" because controllability involves the relation between the state variables of the system and the system inputs (the control vector), while observability involves the relation between the state variables and the system outputs. Controllability and observability may be defined for a general system; however, at the present time it is only for linear systems that conditions have been found to ensure that a system be controllable and observable.

Intuitively, state controllability is a property of a system which guarantees that any initial state can be transferred to any desired terminal state in finite time. A system that is output-controllable is one where any desired terminal output can be attained in finite time starting from an arbitrary initial state.

Observability is an expression of our ability to determine at t_i the state variables $x(t_i)$ based upon measurement of the unforced system output over some interval $[t_i, t_1]$ where $t_1 \geq t_i$ and finite. It is obvious that if each state variable is available for measurement, the plant is observable. That this sufficient condition is not also necessary is what makes the concept interesting.

Example 5

As a simple illustration, consider a plant defined by the two state variables x_1, x_2 and governed by the system equations:

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) + u(t) \\ \dot{x}_2(t) &= x_2(t) \end{aligned} \tag{1}$$

The measurable output of the plant is assumed to be:

$$y(t) = x_1(t) \tag{2}$$

First note from (1) that starting from a specified initial state $[x_1(t_i), x_2(t_i)]$ it is not possible to reach any desired terminal state $x_2(t_1)$. This follows, because the control does not influence the x_2 equation in (1). Thus the system is not state-controllable.

If the terminal conditions for the problems are such that only $x_1(t_1)$ is specified, then by (1) it is apparent that the control can be made to drive the system to the desired terminal value. In general, when the system control vector is such that the terminal-state manifold can be reached, even if this does not include all states, the system is said to be controllable in reduced sense. Note also, that measuring the unforced system output $y(t) = x_1(t)$ over some interval $[t_i, t_1]$ is sufficient to also determine $x_2(t_i)$, since by (1): $x_2(t_i) = \dot{x}_1(t_i)$ for the unforced system. Thus the system is observable. If on the other hand, the measurable system output is not given by (2) but rather by

$$y(t) = x_2(t) \tag{3}$$

the system is not observable. This follows, because measurement of x_2 over the interval $[t_i, t_1]$ is not sufficient to determine $x_1(t_i)$ but only the change from it over the interval. Thus with (3) as the output, the system is not observable.

It can be shown that a linear, stationary system

$$\dot{x}(t) = Ax(t) + Bu(t)$$

is controllable $(x(t_1) = 0)$ if and only if the $n \times mn$ matrix

$$E = [B|AB|A^2B|\cdots|A^{n-1}B]$$

has rank n. If there is only one control input ($m=1$), a necessary and sufficient condition for controllability is that the $n \times n$ matrix E be nonsingular.

Analogously, it can be shown that the linear, stationary system

$$\dot{x}(t) = Ax(t) + Bu(t)$$
$$y(t) = Cx(t)$$

is observable if and only if the $n \times qn$ matrix

$$G = [C'|A'C'|(A')^2C'|\cdots|(A')^{n-1}C']$$

has rank n.

Since we have made the simplifying assumption that $y(t) = x(t)$ (all of the states can be measured), the question of observability will not arise in what follows. Moreover, the systems to be studied will be assumed to possess the property of controllability to the degree required by the solution [13].

6. THE MATHEMATICAL MODEL OF A DISCRETE-TIME OPTIMAL CONTROL PROBLEM

In discrete-time systems, the time-evolution of the state vector is described by a set of vector difference equations. Discrete systems can arise:

a) due to sampling of an inherently continuous-time system, or

b) there are discrete-time systems which need not be related to any continuous-time system.

A concrete problem is usually modelled as a continuous-time problem to obtain analytical results and as a discrete-time problem for numerical purposes. Many of the concepts and definitions we have discussed for continuous-time systems can be analogously defined for the discrete case.

The general discrete-time control problem has the following structure:

$$\max_{\{u(i)\}} J = \sum_{i=0}^{N-1} [r(x(i),u(i),i)]$$

$$\left(\text{or } J = \sum_{i=0}^{N-1} [r(x(i),u(i),i)] + \varphi(x(N)) \right)$$

subject to:

equations of motion: $x(i+1) - x(i) = f(x(i),u(i),i)$, $i = 0,1,\cdots,N-1$

$$\left(\text{or } x(i+1) = f(x(i),u(i),i) \right)$$

initial condition: $x(0) \in X_0$

final condition: $x(N) \in X_N$

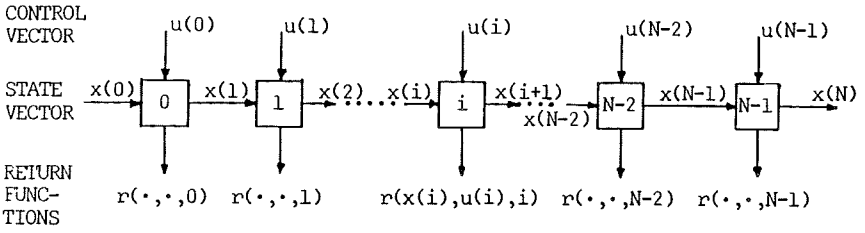
state-space constraint: $x(i) \in X_i$, $i = 1,2,\cdots,N-1$

control constraint: $u(i) \in U_i$, $i = 0,1,\cdots,N-1$

where $x(i) \in R^n$, $u(i) \in R^m$ and $r(\cdot,\cdot,i) : R^{n+m} \to R$, $f(\cdot,\cdot,i) : R^{n+m} \to R^n$ are given real valued functions. This discrete-time problem is continuous in space. Our problem can be considered as one of solving a static optimization problem with $N \times p$ decision variables. Moreover the results we have obtained regarding necessary and sufficient conditions can be applied once X_i and U_i are specified, the same is true in what concerns numerical methods. Now, due to the special structure of this dynamical model, the necessary conditions will take a special form known as the discrete maximum principle. Moreover, stronger sufficient conditions may also be obtained [14].

From a numerical point of view Bellman's dynamic programming is a computer oriented procedure that is usually employed to find optimal control laws. The importance of this approach resides in the fact that it is also applicable to systems that are discrete in space and/or stochastic [15].

The following graphical representation of discrete-time control problems is usually helpful:



The next chapter will also discuss theoretical results for discrete-time control problems.

7. REFERENCES

[1] Intriligator, M.D.: Mathematical Optimization and Economic Theory, Prentice-Hall 1971.

[2] Schutz, D.G., and Melsa, J.L.: State functions and Linear control systems, McGraw Hill 1968.

[3] Wilde, D., and Beightler: Fundations of optimization, Prentice-Hall 1967.

[4] Valqui Vidal, R.V.: OR in production planning, IMSOR 1970.

[5] Lee, T.H., et al.: Computer process control: Modeling and optimization, John Wiley 1968.

[6] Nahorski, Z., and Valqui Vidal, R.V.: Stabilization of a linear discrete-time system using simultaneous identification and control, International Journal of Control 1974, vol. 19, N^{o} 2, 353-364.

[7] Luenberger, D.G.: Optimization by vector space methods, John Wiley 1969.

[8] Citron, S.J.: Elements of optimal control, Holt, Rinehat, and Winston, 1969.

[9] Pontryagin, L.S., et al.: The mathematical theory of optimal processes, John Wiley 1962.

[10] Kliger, I.: "On closed-loop optimal control", IEEE Trans. Aut. Control (1965), 207.

[11] Ogata, K.: State Space Analysis of Control Systems, Prentice-Hall 1967.

[12] Rekasius, Z.V.: "Suboptimal Design of Intentionally Nonlinear Controller", IEEE Trans. Automatic Control (1964), 380-386.

[13] Kalman, R.E.: "On the general theory of control systems", Proc. First IFAC Congress (1960), 481-493.

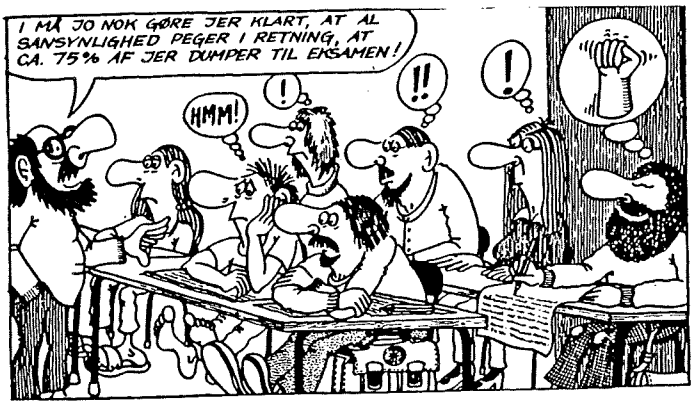
[14] Fan, L., and Wang, Ch.: The discrete maximums principle a study of multistage systems optimization, John Wiley 1964.

[15] Bellman, R., and Dreyfus, S.E.: Applied Dynamic Programming Princeton University Press 1962.

CHAPTER 6

DYNAMIC OPTIMIZATION:

Theory


I MÅ JO NOK GØRE JER KLART, AT AL SANSYNLIGHED PEGER I RETNING, AT CA. 75% AF JER DUMPER TIL EKSAMEN!
HMM!
!
!!
!

1. INTRODUCTION

In the last chapter the optimal (deterministic) control problem has been formulated. In this chapter we will be concerned with the theoretical aspects of both discrete-time and continuous-time control problems.

It is easily seen that the discrete-time control problem is an especial case of our general mathematical programming problem, therefore all our results of chapter 2 are also applicable to this kind of problems. Moreover the results obtained in chapter 2 can be generalized to be valid for linear topological spaces. Once this is done the results obtained in the static theory could be also applied to the continuous-time case. This will not be done here because it will demand a high level of mathematical knowledge [1].

Here we will be concerned with the main theoretical results of optimal control theory. This theory will provide necessary and sufficient conditions for optimal control. Moreover it will also give us an understanding of the geometrical properties of the control problem. This will be done in a compact and informal, but not elementary way. For the sake of compactness most proofs will not be given, some of them are quite elementary while others can be rather cumbersome. Anyway all proofs can be found in some of the references of sec. 4.

Emphasis will be placed in the understanding of two of the main approaches to control problems, that is dynamic programming and the maximum principle. This will be done in sec. 2 for the discrete-time control problem while the continuous-time case will be discussed in sec. 3.

2. DISCRETE-TIME OPTIMAL CONTROL PROBLEMS

In chapter 5 the optimal control problem has been formulated. In this section we will be concerned with the theoretical as-

pects of discrete-time control problems. This kind of problems are also denominated as multistage or sequential decision processes due to the fact that they can be conceived as problems of dynamic resource allocation.

In this section we will develop optimality conditions for the discrete-time control problem. Sec. 2.1 is devoted to sufficient conditions that will give origin to a numerical method known as dynamic programming. In the next section necessary conditions are stipulated and they will take the form of the well-known discrete maximum principle. This principle will take a weak form if it is derived from Kuhn-Tucker necessary conditions while a strong form can be obtained by applying geometric-type proofs from a control theoretic viewpoint, that will be shown to be closely related to Everett's theory.

It is important to keep in mind that dynamic programming can be applied to solve some optimization problems that cannot be formulated as control problems, that is for instance the case of the socalled combinatoric problems.

In what concerns the discrete maximum principle, a great deal of misunderstandings is present in most of the well-known literature. Therefore the necessity to stipulate both a weak and a strong form of this principle, this will not happen when we will discuss the continuous-time control problem.

2.1 Sufficient conditions for optimality

A decision tree and the principle of optimality

Let us consider a discrete-time control problem and assume, for the sake of simplicity, that the state and control vectors have only one component, and that $x(0)$ is given. That is we have to

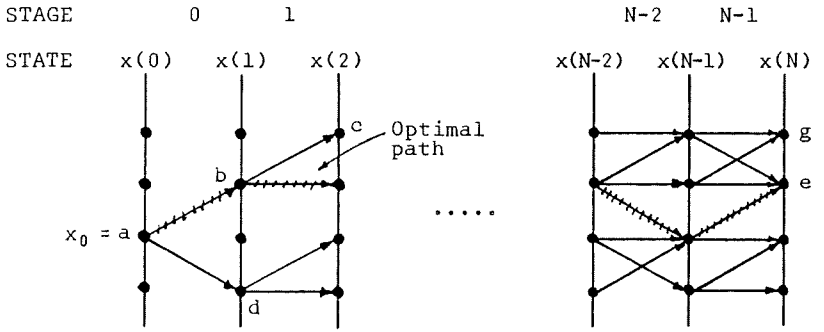
$$\max_{\{u(i)\}} J = \sum_{i=0}^{N-1} r(x(i),u(i),i)$$

subject to:

$$x(i+1) = f(x(i),u(i),i) \quad , \quad i = 0,1,\cdots,N-1$$

and $$x(0) = x_0$$

Let us further assume that each decision and state variable can take on only a finite number of values, we can then represent our problem graphically by a 'decision tree' as shown below.



The circles, called nodes, correspond to the states, and the lines between circles, called arcs, correspond to decisions. Thus at the first stage we start from the initial state x_0 and there are two possible decisions represented by the two arcs emanating from the node. Associated with each arc is a return $r(x(i),u(i),i)$ and an output $x(i+1)$. Admissible controls correspond to paths (a set of adjoining arcs) between x_0 and some of the values of $x(N)$. The return from a path is the sum of the returns from the arcs included in the path. Our problem is to find a maximum path, that is a path yielding maximum return. The optimal path a-b-e is also shown in the figure. Let us denote $F(k,x(k))$ as the total maximum return from $x(k)$, a given value, to the last stage (N-1). The maximum cost, J, is therefore

$$J_{ae} = F(0,a) = r(a,u^*(0),0) + J_{be}$$

ASSERTION: If a-b-e is the optimal path for stages 0 to (N-1) starting from a, then b-e is the optimal path for stages 1 to N-1 starting from b.

Proof by contradiction [2]: Suppose b-c-g is the optimal path for stages 1 to (N-1), then

$$J_{bcg} > J_{be}$$

and

$$r(a,u^*(0),0) + J_{bcg} > r(a,u^*(0),0) + J_{be} = F(0,a)$$

but this last relation can be satisfied only by violating the condition that a-b-e is the optimal path from a to e. Thus the assertion is proved, and $J_{be} = F(1,b)$.

This proof simply states that if the remaining decisions were not optimal then the whole policy could not be optimal. Bellman [3] has called the above property of an optimal control the principle of optimality:

> An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.

Let us now assume that we know the optimal paths for stages 1 to (N-1) both for $x(1) = b$ and $x(1) = d$, then the optimal path for stages 0 to (N-1) can be found by comparing

$$J_{ab} + F(1,b)$$
$$J_{ad} + F(1,d)$$

The maximum of these costs must be one associated with the optimal control at point a. That is

$$F(0,a) = \max_{u(0)} \{r(a,u(0),0) + F(1,f(a,u(0),0))\}$$

or, if the functions $F(1,x(1))$ are given for all admissible values of $x(1)$ then we can find $u^*(0)$ by solving an optimization problem in one variable. Thus our original N-stages discrete-time control problem has been decomposed in a family of (N-1) stages discrete control problems with the same structure than the original one and an optimization problem in one variable. Moreover, the further application of the optimality principle will permit us the complete decomposition of our problem. In other words, we are solving an optimization problem in N-variables by decomposing it in a series of optimization problems in one variable. Dynamic programming (or DP for short) is a computational technique to be presented in the next chapter, which is based on the above discussed concepts.

Functional equations

We will now show that the rather intuitive results obtained in the last section have a more general validity. However, for notational convenience we neglect final conditions and state-space constraints. Let us consider the following discrete-time control problem

$$F(0,x_0) = \max_{\{u(i)\}} \left[\sum_{i=0}^{N-1} r(x(i),u(i),i) + \varphi(x(N)) \right]$$

subject to:

$$x(i+1) = f(x(i),u(i),i) \quad , \quad i = 0,1,\cdots,N-1$$

$$x(0) = x_0$$

$$u(i) \in U_i \quad , \quad i = 0,1,\cdots,N-1$$

where the state $x(i)$ and the control $u(i)$ belong to arbitrary sets X and U respectively. X and U may be finite sets, or finite-dimensional vector spaces, or even infinite-dimensional spaces. $x_0 \in X$ is given. The U_i are given subsets of U, and, finally, $r(\cdot,\cdot,i) : X \times U \to R$, $\varphi : X \to R$, $f(\cdot,\cdot,i) : X \times U \to X$ are given functions.

The main concept underlying DP involves embedding the last control problem $[0,x_0]$, in which the system starts in state x_0 at time 0, into a family of optimal control problems with the same dynamics, objective function, and control constraints as in problem $[0,x_0]$ but with different initial states and initial stages. More precisely, for each $x \in X$ and k between 0 and N-1, consider the problem $[k,x]$:

$$F(k,x) = \max_{\{u(i)\}} \left[\sum_{i=k}^{N-1} r(x(i),u(i),i) + \varphi(x(N)) \right]$$

subject to:

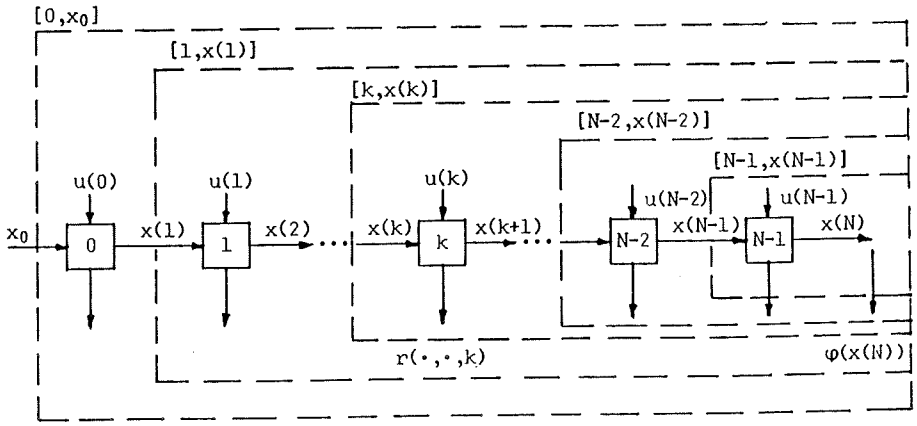
$$x(i+1) = f(x(i),u(i),i) \quad , \quad i = k,k+1,\cdots,N-1$$

$$x(k) = x$$

$$u(i) \in U_i \quad , \quad i = k,k+1,\cdots,N-1$$

and assume that an optimal solution exists for all $0 \leq k \leq N-1$, and all $x \in X$.

This family of problems is illustrated below.



Lemma 1 Suppose $u^*(k),\cdots,u^*(N-1)$ is an optimal control for problem $[k,x]$ and let $x^*(k) = x$, $x^*(k+1),\cdots,x^*(N)$ be the corresponding optimal trajectory. Then for any ℓ, $k \leq \ell \leq N-1$, $u^*(\ell),\cdots,u^*(N-1)$ is an optimal control for problem $[\ell,x^*(\ell)]$.

This is nothing else than a precise statement of the principle of optimality and can be as well easily proved by contradiction. This lemma can be used to prove the decomposition principle under more general conditions [4].

Theorem 1

Define $F(N,x) = \varphi(x)$, then $F(k,x)$ satisfies the backward recursion equation (functional equation)

$$F(k,x) = \max_{u(k)\in U_k} \{r(x,u(k),k) + F(k+1,f(x,u(k),k))\}$$

$$\text{for} \quad 0 \leq k \leq N-1$$

This theorem can also be proved by noting that

$$F(k,x) = \max_{\{u(k)\}} \left\{r(x,u(k),k) + \sum_{i=k+1}^{N-1} r(x(i),u(i),i) + \varphi(x(N))\right.$$

$$= \max_{u(k)\in U_k} \max_{\{u(k+1)\}} \left\{r(x,u(k),k) + \sum_{i=k+1}^{N-1} r(x(i),u(i),i) + \varphi(x(N))\right.$$

$$= \max_{u(k)\in U_k} \left\{r(x,u(k),k) + \left(\max_{\{u(k+1)\}} \sum_{i=k+1}^{N-1} r(x(i),u(i),i) + \varphi(x(N))\right.\right.$$

$$= \max_{u(k)\in U_k} \{r(x,u(k),k) + F(k+1,f(x,u(k),k))$$

We have then simplified (decomposed) problem $[k,x]$ into two smaller optimization problems

1. A problem $[k+1,x(k+1)]$, and
2. A one-stage (static) optimization problem

Let us further specify two results.

Corollary 1

Let $u(k),\cdots,u(N-1)$ be any control for the problem $[k,x]$ and let $x(k) = x,\cdots,x(N)$ be the corresponding trajectory, then

$$F(\ell,x(\ell)) \geq r(x(\ell),u(\ell),\ell) + F(\ell+1,f(\ell,x(\ell),u(\ell))$$

$$\text{for } k \leq \ell \leq N-1$$

and equality holds for all $k \leq \ell \leq N-1$ <u>if and only if</u> the control is optimal for problem $[k,x]$.

Corollary 2

For $k = 0,1,\cdots,N-1$, let $\psi(k,\cdot) : X \to U_k$ be such that

$$r(x,\psi(k,x),k) + F(k+1,f(x,\psi(k,x),k)) =$$

$$\max_{u \in U_k} [r(x,u,k) + F(k+1,f(x,u,k))]$$

then $\psi(k,\cdot)$, $k = 0,\cdots,N-1$ is an <u>optimal closed loop control</u>, i.e., for any k,x the control $u^*(k),\cdots,u^*(N-1)$ defined by

$$u^*(\ell) = \psi(\ell,x^*(\ell)) \text{ , } k \leq \ell \leq N-1 \text{ , where}$$

$$x^*(\ell+1) = f(x^*(\ell), \psi(\ell,x^*(\ell),\ell)) \text{ , } k \leq \ell \leq N-1$$

$$x^*(k) = x$$

is optimal for problem $[k,x]$.

Example 1

Consider the following control problem

$$J = \max - \sum_{i=0}^{1} \{x^2(i) - \tfrac{1}{2}(-2)^i[(u(i) + 2)^2 - 1]\}$$

subject to: $x(i+1) = x(i) + u(i)$, $i = 0,1$

$x(0) = 1$

$-3 \leq u(i) \leq 1$, $i = 0,1$

where $x(i) \in R$ and $u(i) \in R$.

The functional equations are

$$F(1,x) = \max_{-3 \leq u(1) \leq 1} [-x^2 + \tfrac{1}{2}(-2)[(u(1) + 2)^2 - 1]] \quad \text{and}$$

$$F(0,x(0)) = \max_{-3 \leq u(0) \leq 1} [-x^2(0) + \tfrac{1}{2}(-2)^0[(u(0) + 2)^2 - 1] + F(1,x(0) + u(0))]$$

The last stage is optimized at $u^*(1) = -2$, a constant independent of x , and we obtain

$$F(1,x) = -x^2 + 1$$

then

$$F(0,x(0)) = \max_{-3 \leq u(0) \leq 1} [-x^2(0) + \tfrac{1}{2}[(u(0) + 2)^2 - 1] - (x(0) + u(0))^2 + 1]$$

This is maximized at $\underline{u^*(0) = 0}$, for $x(0) = 1$ and we obtain

$$J = F(0,1) = 0.5$$

and the optimal control is: $\{u^*(0),u^*(1)\} = \{0,-2\}$.

Note that the optimal solution is obtained even if our problem is neither concave nor convex.

Complementary remarks

1.- We have obtained the above mentioned results by comparing the optimal decision with all the other decisions. This global comparison, therefore, leads to optimality conditions which are sufficient. Moreover, the results obtained are quite general in the sense that they do not require differentiability, concave functions, or convex sets, or even the restriction to a finite-dimensional state-space.

2.- The obtained results say nothing at all about how the one-stage optimization problems are to be solved, each of these could conceivably be an extremely difficult problem. Thus only in some special cases it is possible to find a "closed-form" analytical solution to the recursion equation [5], otherwise we have to recur to numerical methods, this will be discussed in the next chapter.

3.- In general one is able to use either a forward solution or a backward solution, and one has the option of selecting one or the other. A backward solution, the one we have emphasized, works in time so that the first step of the recursion equation corresponds to the last decision (in time) to be made. The forward solution works forward in time, with the first step in the recursion equations being that corresponding to the first decision to be made.

Mitten's conditions

We have shown some decomposition principles for discrete-time control problems with additive returns. These results can be generalized to other types of performance function as shown by Mitten [6]. Let us assume that the objective function of our discrete-time control problem is

$$R(r(x(0),u(0),0),r(x(1),u(1),1),\cdots,r(x(N-1),u(N-1),N-1))$$

such that it satisfies the following two sufficient conditions.

(a) Separability

$$R(r(\cdot,\cdot,0),\cdots,r(\cdot,\cdot,N-1)) = R_1[r(\cdot,\cdot,0),R_2[r(\cdot,\cdot,1),\cdots,r(\cdot,\cdot,N-1)]]$$

where R_1 and R_2 are real-valued functions, and

(b) Monotonicity

R_1 is monotonically non-decreasing function of R_2 for every $r(\cdot,\cdot,0)$.

Then the problem can be decomposed, i.e.

$$\max_{u(0),\cdots,u(N-1)} R(\quad) = \max_{u(0)} R_1[r(\cdot,\cdot,0), \max_{u(1),\cdots,u(N-1)} R_2[r(\cdot,\cdot,1),\cdots,r(\cdot,\cdot,N-1)]]$$

Then if we assume complete decomposition Theorem 1 can be easily generalized. Mitten's proof of this result is rather cumbersome because he is dealing with more complex forms of multistage processes, in [7] an elementary proof can be found.

Classical examples where these sufficient conditions can be applied are:

multiplication of positive stage returns

$$R = r(\cdot,\cdot,0)\cdot r(\cdot,\cdot,1) \cdots r(\cdot,\cdot,N-1)$$

minimum of the individual stage returns

$$R = \min\{r(\cdot,\cdot,0),r(\cdot,\cdot,1),\cdots,r(\cdot,\cdot,N-1)\}$$

Recently, Cambini [8] has obtained weaker sufficient conditions for the validity of the recurrence equations and sufficient conditions under which a general mathematical programming problem can be formulated as a discrete-time control problem.

2.2 Necessary conditions for optimality derived from the static theory

Let us begin by considering the following (elementary) discrete-time control problem

$$\max_{\{u(i)\}} \sum_{i=0}^{N-1} r(x(i),u(i),i)$$

subject to

$$x(i+1) = f(x(i),u(i),i) \quad , \quad i = 0,1,\cdots,N-1$$

$$x(0) = a$$

where $x(i) \in R^n$, $u(i) \in R^m$ and $r(\cdot,\cdot,i) : R^{n+m} \to R$, $f(\cdot,\cdot,i) : R^{n+m} \to R^n$ are given differentiable functions. Applying Kuhn-Tucker theorem we can stipulate necessary conditions for optimality, see chapter 2. Suppose that the constraint qualification is satisfied, and $\{x^*(i)\}$, $i = 0,\cdots,N$; $\{u^*(i)\}$, $i = 0,\cdots,N-1$, is an optimal solution, then there exist $p^*(i) \in R^n$, $1 \leq i \leq N$, a Lagrange multiplier vector so that at the optimal solution

$$\frac{\partial L}{\partial u(i)} = \frac{\partial r}{\partial u(i)} + p(i+1) \frac{\partial f}{\partial u(i)} = 0 \quad , \quad i = 0,1,\cdots,N-1$$

$$\frac{\partial L}{\partial x(i)} = \frac{\partial r}{\partial x(i)} + p(i+1) \frac{\partial f}{\partial x(i)} - p(i) = 0 \quad , \quad i = 1,\cdots,N-1$$

$$x(0) = a$$

$$\frac{\partial L}{\partial x(N)} = -p(N) = 0$$

$$\frac{\partial L}{\partial p(i+1)} = f(x(i),u(i),i) - x(i+1) = 0 \quad , \quad i = 0,\cdots,N-1$$

for $x(i) = x^*(i)$, $u(i) = u^*(i)$ and $p(i) = p^*(i)$, where L is the Lagrangian form:

$$L = \sum_{i=0}^{N-1} r(x(i),u(i),i) - \sum_{i=0}^{N-1} p(i+1)(x(i+1)-f(x(i),u(i),i))$$

Considerable elegance and mnemonic simplification is achieved if we define the Hamiltonian function H(i) by

$$H(x(i),u(i),p(i+1)) = r(x(i),u(i),i) + p(i+1)f(x(i),u(i),i) \qquad i = 0,1,\cdots,N-1$$

then Kuhn-Tucker conditions can be written as

(1)	$\frac{\partial H(i)}{\partial u(i)} = 0$	$i = 0,1,\cdots,N-1$	(stationarity of H(i))
(2)	$\frac{\partial H(i)}{\partial x(i)} = p(i)$	$i = 1,\cdots,N-1$	(adjoint or co-state equations)
(3)	$\frac{\partial H(i)}{\partial p(i+1)} = x(i+1)$	$i = 0,1,\cdots,N-1$	(equations of motion)
(4)	$x(0) = a$		boundary conditions
(5)	$p(N) = 0$		

at $x(i) = x^*(i)$, $u(i) = u^*(i)$ and $p(i) = p^*(i)$

From the foregoing we see that our discrete maximum principle (in weak form) requires that:

> the set of control variables $\{u^*(i)\}$ that causes our discrete-time control problem to take on a stationary value be obtained as follows: introduce an additional set of Lagrange multipliers (adjoint or co-state variables) $p(i)$, $i = 1,\cdots,N$, which satisfy (2) and (5); form a stagewise Hamiltonian function that satisfies (2) and (3); and solve for the values of the control variables that cause the Hamiltonian function to take on a stationary value at each stage [9].

The above stipulated necessary conditions are also sufficient if for instance the $r(\cdot,\cdot,i)$ are concave and the $f(\cdot,\cdot,i)$ are linear. Furthermore, in this case (1) becomes

$$\max_{u(i)} H(x^*(i),u(i),p^*(i+1))$$

That is maximization of the Hamiltonian with respect to the control variables is necessary in order to maximize the objective function. This is the strong form of the discrete maximum principle that is also valid under less restrictive assumptions as it will be seen in the next section.

In the available literature there exist different ways to formulate the discrete-time control problem and therefore the necessary conditions will also be different. Thus, sometimes the objective function is written as

$$\max_{\{u(i)\}} \varphi(x(N))$$

In such situation the necessary conditions remain the same except for (5) that becomes $p(N) = \frac{\partial \varphi(x(N))}{\partial x(N)}$.

In other cases the equation of motion is expressed as

$$x(i+1) - x(i) = f(x(i),u(i),i) \quad , \quad i = 0,1,\cdots,N-1$$

and (2) and (3) become

$$\frac{\partial H(i)}{\partial x(i)} = p(i) - p(i+1) \quad , \quad i = 1,\cdots,N-1$$

$$\frac{\partial H(i)}{\partial p(i+1)} = x(i+1) - x(i) \quad , \quad i = 0,1,\cdots,N-1$$

respectively.

These results are summarized in the following tableau

for $\{x^*(i),u^*(i)\}$, $\exists\ p^*(i)$ such that	equations of motion	
	$x(i+1) = f(x(i),u(i),i)$	$x(i+1)-x(i) = f(x(i),u(i),i)$
$\max \left(\sum_{i=0}^{N-1} r(\cdot,\cdot,i) \right)$	$H(i) = r(x(i),u(i),i) + p(i+1)f(x(i),u(i),i)$	$H(i) = r(x(i),u(i),i) + p(i+1)f(x(i),u(i),i)$
	$\frac{\partial H(i)}{\partial u(i)} = 0 \quad i = 0,1,\cdots,N-1$	$\frac{\partial H(i)}{\partial u(i)} = 0 \quad i = 0,1,\cdots,N-1$
	$\frac{\partial H(i)}{\partial x(i)} = p(i) \quad i = 1,\cdots,N-1$	$\frac{\partial H(i)}{\partial x(i)} = p(i) - p(i+1) \quad i = 1,2,\cdots,N-1$
	$\frac{\partial H(i)}{\partial p(i+1)} = x(i+1) \quad i = 0,1,\cdots,N-1$	$\frac{\partial H(i)}{\partial p(i+1)} = x(i+1) - x(i) \quad i = 0,1,\cdots,N-1$
Objective function	$x(0) = a$	$x(0) = a$
	$p(N) = 0$	$p(N) = 0$
	$H(i) = p(i+1)f(x(i),u(i),i)$	$H(i) = p(i+1)f(x(i),u(i),i)$
	$\frac{\partial H(i)}{\partial u(i)} = 0 \quad i = 0,1,\cdots,N-1$	$\frac{\partial H(i)}{\partial u(i)} = 0 \quad i = 0,1,\cdots,N-1$
$\max [\varphi(x(N))]$	$\frac{\partial H(i)}{\partial x(i)} = p(i) \quad i = 1,\cdots,N-1$	$\frac{\partial H(i)}{\partial x(i)} = p(i)-p(i+1) \quad i = 1,\cdots,N-1$
	$\frac{\partial H(i)}{\partial p(i+1)} = x(i+1) \quad i = 0,1,\cdots,N-1$	$\frac{\partial H(i)}{\partial p(i+1)} = x(i+1)-x(i) \quad i = 0,1,\cdots,N-1$
	$x(0) = a$	$x(0) = a$
	$p(N) = \frac{\partial \varphi(x(N))}{\partial x(N)}$	$p(N) = \frac{\partial \varphi(x(N))}{\partial x(N)}$

These results can easily be generalized, thus the next tableau stipulates the necessary conditions for a general discrete-time control problem. Here $x(i) \in R^n$, $u(i) \in R^m$, and the functions $r(\cdot,\cdot,i) : R^{n+m} \to R$, $f(\cdot,\cdot,i) : R^{n+m} \to R^n$, $q_i : R^n \to R^{m_i}$, $g_i : R^n \to R^{\ell_i}$ and $h_i : R^m \to R^{s_i}$ are once-differentiable.

Assume $\{x^*(i), u^*(i)\}$ maximizes

$$\sum_{i=0}^{N-1} r(x(i),u(i),i)$$

subject to the constraints below

Suppose that the constraint qualification is satisfied, then there exist Lagrange multipliers $p^*(i) \in R^n$, $i = 0,1,\cdots,N$; $\lambda^{i*} \in R^{m_i}$, $i = 0,1,\cdots,N$; $\alpha^{0*}, \alpha^{N*} \in R^{\ell_i}$, and $\gamma^{i*} \in R^{s_i}$, $i = 0,1,\cdots,N-1$, such that

Equations of motion:

$$x(i+1)-x(i) = f(x(i),u(i),i) \qquad i = 0,\cdots,N-1$$

Adjoint equations:

$$p(i)-p(i+1) = \frac{\partial H(i)}{\partial x(i)} - \left[\frac{\partial q_i(x(i))}{\partial x(i)}\right]' \left[\lambda^i\right]' \qquad i = 0,\cdots,N-1$$

Initial conditions:

$$q_0(x(0)) \leq 0\,, \quad g_0(x(0)) = 0$$

Transversality or boundary conditions:

$$p(0) = \left[\frac{\partial g_0(x(0))}{\partial x(0)}\right]' \left[\alpha^0\right]'$$

$$\lambda^0 \geq 0$$
$$\lambda^0\, q_0(x(0)) = 0$$

Final conditions:

$$q_N(x(N)) \leq 0\,, \quad g_N(x(N)) = 0$$

$$p(N) = \left[\frac{\partial g_N(x(N))}{\partial x(N)}\right]' \left[\alpha^N\right]' - \left[\frac{\partial q_N(x(N))}{\partial x(N)}\right]' \left[\lambda^N\right]'$$

$$\lambda^N \geq 0$$
$$\lambda^N\, q_N(x(N)) = 0$$

State-space constraint:

$$q_i(x(i)) \leq 0 \qquad i = 1,\cdots,N-1$$

$$\lambda^i \geq 0$$
$$\lambda^i\, q_i(x(i)) = 0$$

Control constraint:

$$h_i(u(i)) \leq 0 \qquad i = 0,\cdots,N-1$$

Stationarity conditions:

$$\frac{\partial H(i)}{\partial u(i)} - \left[\frac{\partial h_i(u(i))}{\partial u(i)}\right]' \left[\gamma^i\right]' = 0$$

$$\gamma^i \geq 0$$
$$\gamma^i\, h_i(u(i)) = 0$$

The last tableau stipulates the discrete maximum principle in its weak form. If the $r(\cdot,\cdot,i)$ are concave and the remaining functions are linear then the necessary conditions are also sufficient. Moreover, in this case the stationary conditions can be written as

$$\max_{u(i)} H(x^*(i),u(i),p^*(i+1))$$

$$h_i(u(i)) \leq 0$$

Example 2

Consider the problem

$$\max - \sum_{i=0}^{N-1} [x(i)-u(i)]^2$$

subject to: $x(i+1) = u(i)$, $i=0,\ldots,N-1$

$x(0) = a$

where $x(i) \in R$ and $u(i) \in R$.

In this problem we can stipulate the discrete maximum principle in its strong form.

The Hamiltonian becomes

$$H(i) = -[x(i)-u(i)]^2 + p(i+1)u(i)$$

The necessary (and sufficient) conditions become

$$\max_{u(i)} H(i) \Rightarrow 2(x(i)-u(i)+p(i+1) = 0 \Rightarrow u^*(i) = \frac{2x^*(i)+p^*(i+1)}{2}$$

$$\frac{\partial H(i)}{\partial x(i)} = p(i) \Rightarrow -2(x(i)-u(i)) = p(i) \Rightarrow u^*(i) = \frac{2x^*(i)+p^*(i)}{2}$$

$$\frac{\partial H(i)}{\partial p(i+1)} = x(i+1) \Rightarrow x^*(i+1) = u^*(i)$$

and the boundary conditions: $x(0)=a$, $p^*(N)=0$.

From these conditions we find that

$$p^*(i+1) = p^*(i) = p^*(N) = 0$$

and the optimal control trajectory

$$u^*(0) = u^*(1) = \ldots = u^*(N-1) = a$$

Example 3

Let us consider the problem of example 1, the Hamiltonian functions are

$$H(1) = -x^2(1)-[(u(1)+2)^2-1] + p(2)\ (x(1)+u(1))$$
$$H(0) = -1 + \tfrac{1}{2}[(u(0)+2)^2-1] + p(1)\ (1+u(0))$$

Disregard by the moment the restrictions: $-3 \leq u(i) \leq 1$, then the necessary conditions are

$$\frac{\partial H(1)}{\partial u(1)} = 0 \Rightarrow -2(u(1)+2) + p(2) = 0$$

$$\frac{\partial H(0)}{\partial u(0)} = 0 \Rightarrow u(0) + 2 + p(1) = 0$$

$$\frac{\partial H(1)}{\partial x(1)} = p(1) \Rightarrow -2x(1) + p(2) = p(1)$$

$$\frac{\partial H(1)}{\partial p(2)} = x(2) \Rightarrow x(2) = x(1) + u(1)$$

$$\frac{\partial H(0)}{\partial p(1)} = x(1) \Rightarrow x(1) = 1 + u(0)$$

$$x(0) = 1$$
$$p(2) = 0$$

That has only one solution

$$p^*(1) = -2\ ,\ p^*(2) = 0$$
$$x^*(1) = 1\ \ ,\ x^*(2) = -1$$

and the control trajectory: $\{u^*(0),u^*(1)\}=\{0,-2\}$ that is also admissible. That is the discrete maximum principle in its weak form gives a global optimum.

Note that while $u^*(1)$ maximizes $H(1)$, $u^*(0)$ minimizes $H(0)$, that is the discrete maximum principle in its strong form is not applicable here.

Example 4

Consider the following problem

$$\max - \sum_{i=0}^{N-1} |u(i)|$$

subject to:
$$x(i+1) = x(i) + Ax(i)+bu(i) \ , \ i=0,\ldots,N-1$$
$$-1 \leq u(i) \leq 1 \ , \quad i=0,\ldots,N-1$$
$$x(0) = x_o$$
$$x(N) = x_N$$

where A is $[n\times n]$, b is $[n\times 1]$, $x(i) \in R^n$ and $u(i) \in R$. The maximum principle cannot be applied directly here although the problem is concave, that is due to the fact that $|u(i)|$ is not a continuously differentiable function.

This inconvenience can be avoided by making the following substitution:

$$u(i) = v(i) - w(i)$$
$$|u(i)| = v(i) + w(i)$$
$$0 \leq v(i) \leq 1 \ , \ 0 \leq w(i) \leq 1$$

for $i=0,\ldots,N-1$.

Note that this transformation is valid only if either or both $v(i)$ or/and $w(i)$ are zero for each $i=0,\ldots,N-1$, but it is easily shown that this is the case at the optimal solution.

A discrete maximum principle derived from a control theoretic approach

In the last section we approached the discrete-time problem from a mathematical programming point of view. The strong form of the discrete maximum principle can be obtained under less restrictive assumptions by applying a control theoretic viewpoint, i.e. by explicit consideration of trajectories. The strong form states, essentially, that a necessary condition for maximization of the objective function is that the stagewise Hamiltonian be maximized with respect to the control variables.

The subtleties that distinguish the weak and strong forms of the discrete maximum principle had not been recognized by many earlier investigators, that is for example the case of a well-known book of Fan and Wang [10]. Halkin [11] has derived a strong form of the discrete maximum principle based upon assumptions that guaranteed that the sets of reachable states at each stage were convex. It is easy to construct counter-examples that show that, if this convexity assumption is violated, one obtains incorrect results. However, there are discrete optimal control problems for which this convexity requirement can be relaxed, as shown by Holtzman [12] by the requirement of "directional convexity" (to be defined below). These two last papers give a lucid derivation of the strong form of the discrete maximum principle using simple topological and geometric arguments, and they are excellent introduction to the important geometric approach to optimal control theory.

Let us consider the following discrete-time control problem

$$\max \sum_{i=0}^{N-1} r(x(i),u(i),i)$$

subject to

$$x(i+1) - x(i) = f(x(i),u(i),i) \quad , \quad i = 0,1,\cdots,N-1$$

$$x(0) = a$$

$x(N) \in S$ (a smooth[1] target in the state space)

$u(i) \in U_i$, $i = 0,1,\cdots,N-1$

Here it is assumed that the following conditions are satisfied.

(1) The $f(\cdot,\cdot,i) : R^{n+m} \to R^n$ are continuous and well-defined for all $x(i)$ and $u(i)$ such that the Jacobian matrix, $\frac{\partial f(\cdot,\cdot,i)}{\partial x(i)}$, is continuous and bounded for all $x(i)$.

(2) If I is the identity matrix, then it is assumed that the matrix $I + \frac{\partial f(\cdot,\cdot,i)}{\partial x(i)}$ is nonsingular for all $x(i),u(i)$.

(3) The $r(\cdot,\cdot,i) : R^{n+m} \to R$ are well-defined and continuous for all $x(i)$ and $u(i)$, and that the vector $\frac{\partial r(\cdot,\cdot,i)}{\partial x(i)}$ is continuous and bounded for all $x(i)$ and for every fixed $u(i)$[2].

(4) The requirement of <u>directional convexity</u> is as follows. Let u and v be any two given vectors such that $u \in U_i$ and $v \in U_i$ Define a vector $[(n+1) \times 1]$

$$y_i(x,u) = \begin{bmatrix} f(x(i),u(i),i) \\ r(x(i),u(i),i) \end{bmatrix}$$

Then the directional convexity assumption is:

For all x for any given $u \in U_i$ $v \in U_i$ and $\alpha \in [0,1]$, there is a $\beta \geq 0$ and a $w \in U_i$ such that

---

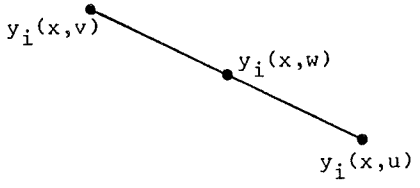
1) S is 'smooth' if the tangents exist for all boundary points of S.

2) Note that we no longer require the functions f and r to be differentiable in u; in fact, we do not even require these functions to be continuous in u.

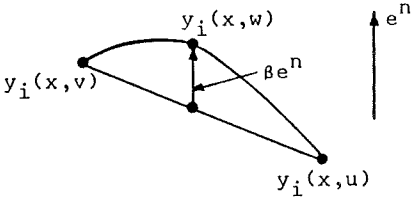
$$y_i(x,w) = \alpha y_i(x,v) + (1-\alpha)y_i(x,u) + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \beta \end{bmatrix}$$

$$= \alpha y_i(x,v) + (1-\alpha)y_i(x,u) + \beta e^n$$

Note that if r is concave, if f is linear and U_i is convex, this last assumption is satisfied. The diagrams below illustate these concepts:



Convexity



Directional convexity

The directional convexity assumption loosely implies that the set of reachable states at stage i is convex in the 'right' direction. Thus, the convexity assumption is not necessary in the proof; rather, only certain portions of the set need be convex. This requirement of directional convexity represents one of the assumptions used in the derivation of the following theorem. As such, it does not represent a necessary condition for optimality. In other words, a given problem may violate the assumption of directional convexity and it may still satisfy the strong form of the discrete maximum principle. On the other hand, one cannot be sure whether or not the principle holds whenever the assumption of directional convexity is violated.

The necessary conditions are now stated here as follows [14].

Theorem 2

Suppose that $\{u^*(i)\}$ is the optimal control sequence. Let $\{x^*(i)\}$ denote the generated optimal state sequence. Then, there is corresponding co-state sequence $\{p^*(i)\}$, $i = 1,\cdots,N$ such that the following relations hold

(1) Canonical difference equations

$$x^*(i+1) - x^*(i) = \left.\frac{\partial H(i)}{\partial p(i+1)}\right|_* \quad , \quad i = 0,\cdots,N-1$$

$$-p^*(i+1) + p^*(i) = \left.\frac{\partial H(i)}{\partial x(i)}\right|_* \quad , \quad i = 1,\cdots,N-1$$

(2) Boundary conditions

$$x(0) = a$$

$$x^*(N) \in S$$

$p^*(N)$ normal to S

(3) Maximization of the Hamiltonian

$$\max_{u(i)\in U_i} H(x^*(i),u(i),p^*(i+1),i) \qquad i = 0,1,\cdots,N-1$$

where H ( ) is defined as in page 227 ---

Example 5

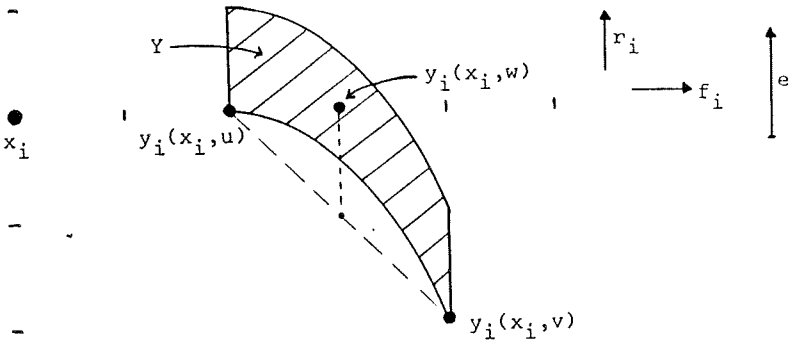
Consider a problem with

$$r(x(i),u(i),i) = -\tfrac{1}{2}u_1^2 + u_2$$

$$x(i+1) - x(i) = u_1 + 2$$

$$(u_1,u_2)' \in U = \{(u_1,u_2)' \mid 0 \le u_1 \le 2 \wedge 0 \le u_2 \le 1\}$$

We see that the set Y of reachable states $y_i(x_i,u)$ is as illustrated below:



Y is not convex. By introducing the unit vector e parallel with r_i and the non-negative scalar β we see that Y is directionally convex: Let for instance $u = (0,0)' \in U$, $v = (2,0)' \in U$ and $\alpha = \frac{1}{2}$. Then we can find that for instance $\beta = 1$ and $w = (1,\frac{1}{2})' \in U$ satisfy

$$
\begin{aligned}
& y_i(x,w) \\
&= (3,0)' \\
&= \tfrac{1}{2}(4,-2)' + \tfrac{1}{2}(2,0)' + (0,1)' \\
&= \alpha y_i(x,v) + (1-\alpha) y_i(x,u) + \beta(0,1)'
\end{aligned}
$$

Example 6

Let us consider the problem of example 3 and now we do not need to make the above mentioned substitution, because Theorem 2 can be directly applied. Thus, the Hamiltonian becomes

$$H(i) = -|u(i)| + p(i+1)\left[x(i) + Ax(i) + bu_i\right]$$

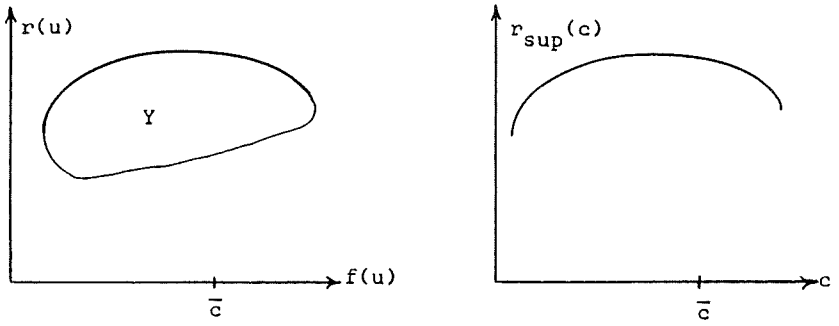
and the discrete maximum principle in its strong form is applicable.

2.3 Interpretation of the discrete-time control problem in terms of Everett's theory

These results can be interpreted in the light of the results developed in chapter two [29]. Here we considered a problem which can be formulated as

$$\max\ r(u)$$
$$f(u) = \bar{c}$$
$$u \in U$$

Let $(r(u),f(u))'$ be called a state, and let Y be called the set of possible states for $u \in U$. This is illustrated on the next page, left.

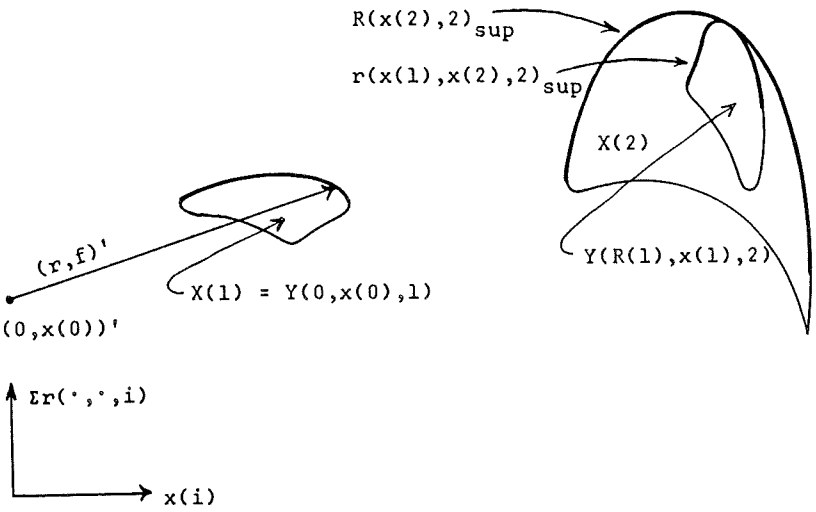


Now obviously $(r(u^*),f(u^*))'$ will never be placed in the interior of Y, since we could then always obtain a higher value of r(u) by moving upwards. Therefore $(r(u^*),f(u^*))'$ is situated at the upper boundary of Y. This boundary, conceived as a function of c, is called the perturbation function $r_{sup}(c)$, and was studied in chapter two.

In analogy with this we define for our dynamic problem

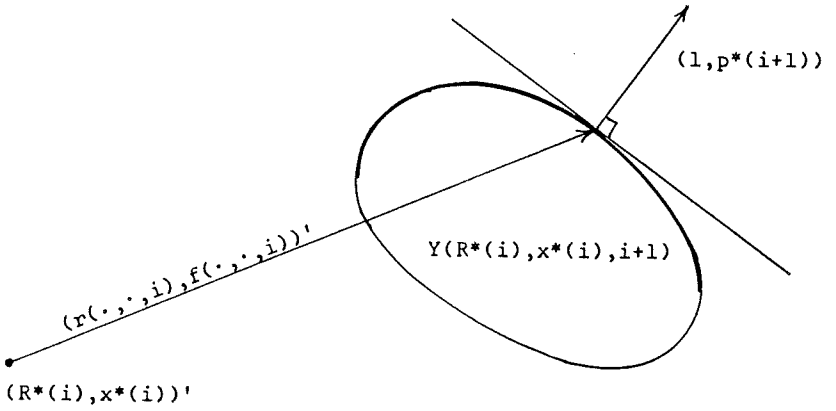
$$R(i) = \sum_{j=0}^{i-1} r(x(j),u(j),j)$$

that is, $R(i)$ is the accumulated criterion function; let $(R(i),x(i))'$ be called a cumulated state, and let $X(i)$ be the set of possible accumulated states at stage i. Further let $Y(R(i-1),x(i-1),i)$ be the set of possible states for any given $(R(i-1),x(i-1))'$. Finally let $R(x(i),i)_{sup}$ denote the overall perturbation function and $r(x(i-1),x(i),i)_{sup}$ denote the perturbation function for any given $x(i-1)$. These definitions are illustrated on the next page:



Since x is independent of r, an optimal $(R(i),x(i))'$ will always be situated at $R(x(i),i)_{sup}$, the upper boundary of $X(i)$, and $r(x^*(i-1),x(i),i)_{sup}$, the upper boundary of $Y(R^*(i-1),x^*(i-1),i)$. Assuming a linear support to $r(x^*(i),x(i+1),i+1)_{sup}$ exists at $x^*(i+1)$ it is easy to see that the

scalar product between $(1,p^*(i+1))$ and $(r(x^*(i),u(i),i), f(x^*(i),u(i),i))'$ attains its maximal value at $x^*(i+1)$. Or, since this scalar product is equal to $H(x^*(i),u(i),p^*(i+1))$, the strong form of the maximum principle is verified.



Since we will always place an optimal $(R(i),x(i))'$ on $r(\cdot,\cdot,i)_{sup}$ that is, on the upper boundary of $Y(\cdot,\cdot,i)$, it is not necessary for the existence of a linear support that $Y(\cdot,\cdot,i)$ be convex only that it is directional convex. This, then, is equivalent to saying that $r(\cdot,\cdot,i)_{sup}$ is concave.

We see that the strong form of the maximum principle is nothing but repeated application of the homogeneous strong Lagrangean principle. Likewise, the weak form of the maximum principle is a repeated application of the homogeneous weak Lagrangean principle. Like in the static case it is not always possible to have the first component in the vector $(w,p(i))$ equal 1. Unlike the static case it is not usual to apply nonlinear supports in the discrete maximum principle.

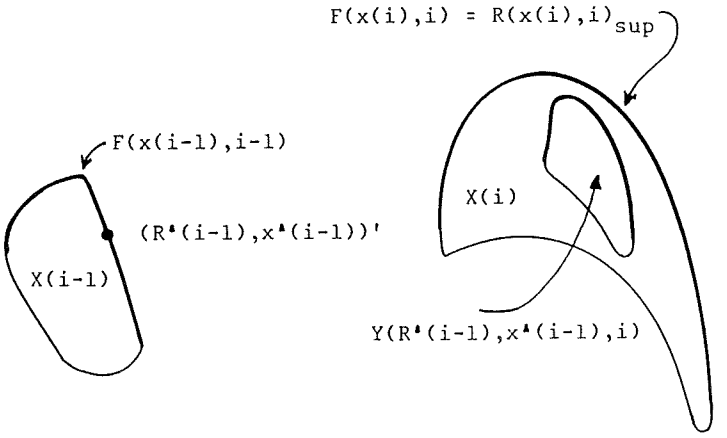
The relationship between the discrete maximum principle and dynamic programming can also conveniently be discussed within the framework of perturbation functions.

Let $$F(x(i),i) = \max_{\{u(j)\}} \sum_{j=0}^{i=1} r(x(j),u(j),j)$$

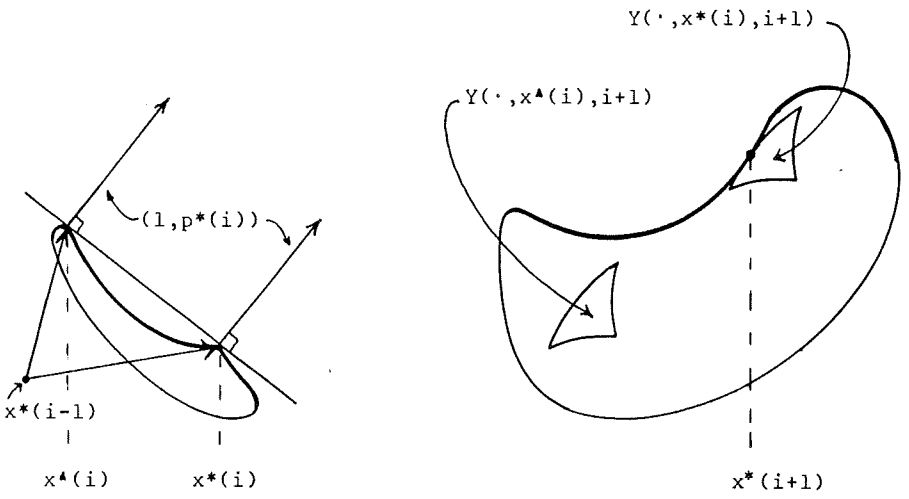
subject to $$x(i) = x(i-1) + f(x(i-1),u(i-1),i-1)$$

that is, the optimal performance function in a forward DP solution. Then it is seen, that $F(x(i),i) = R(x(i),i)_{sup}$.

In other words, in dynamic programming the overall perturbation function $R(x(i),i)_{sup}$ is calculated at each stage. Starting at stage i at $(R^*(i-1),x^*(i-1))'$ situated at $R(x(i-1),i-1)_{sup}$ the optimal performance with respect to this $x^*(i-1)$ is found. Repeating for all other $(R(i-1),x(i-1))'$ situated at $R(x(i-1),i-1)_{sup}$ and comparing these $r(x(i-1),x(i),i)_{sup}$, the overall optimal performance $R(x(i),i)_{sup}$ is found. The principle that only $(R(i-1),x(i-1))'$ situated at $R(x(i-1),i-1)_{sup}$ are considered, is the principle of optimality (in the forward version).



The relationship between the necessary and sufficient conditions is illustrated below:

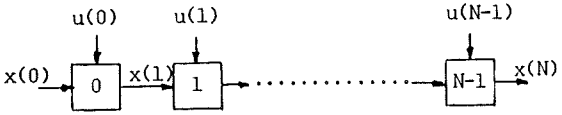


Whether we choose the optimal $x^*(i)$ or the non-optimal $x^A(1)$ we get the same, maximal, Hamiltonian. Thus it is impossible through the maximum principle to distinguish between $x^A(i)$ and $x^*(i)$. The maximum principle thus provide necessary conditions. The dynamic programming approach, however, does distinguish between $x^A(i)$ and $x^*(i)$ and thus provide sufficient conditions.

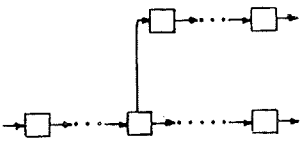
Complementary remarks

1.- There are several variants of the basic problem to which the discrete maximum principle applies. Thus by altering the boundary conditions or by introducing additional variables we can consider discrete, stagewise problems for which the end values of the state variables are specified, the initial values of the state variables are unspecified, or there appear additional terms in the constraining equations, see further [10].

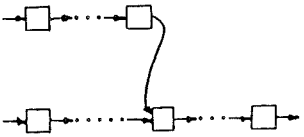
2.- We have developed sufficient and necessary conditions for socalled serial systems, i.e.



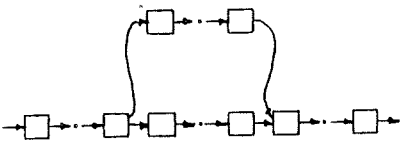
These results have been extended to deal with complex systems as those shown below



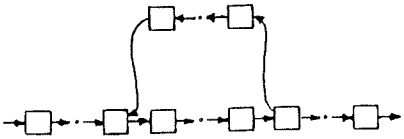
Diverging branch system



Converging branch system



Feedforward loop system



Feedback loop system

or even more complex systems with several loops and branches at the same time. Systems of this nature are particularly common in the chemical-process industries, see further [10], [7].

3.- Recently, it became apparent that a unified approach to develop necessary conditions to discrete-time control problems is not only feasible, but also highly desirable from a theoretical viewpoint. This is done by utilizing the geometrical properties of the discrete-time control to obtain constraint qualifications that are weaker than those obtained in the static optimization problem, for a lucid derivation, see [14], [15].

3. CONTINUOUS-TIME OPTIMAL CONTROL PROBLEMS

In this section we will be dealing with continuous-time optimal control problems and it will be seen that the two approaches, dynamic programming and the maximum principle, utilized to deal with discrete-time optimal control can also be applied to study theoretically the continuous-time case.

The study of continuous-time optimal control problems is a wide and difficult branch of control theory. Most good books, as for instance [16], demand a high level of mathematical background. Moreover, most of these books are especially oriented to the control of technical and aerospace systems, what makes difficult their access to operations research workers. Here emphasis will be placed to the main results and their geometrical conception. The geometrical approach, although its lack of mathematical elegance as opposite to the pure mathematical, provides deeper insight on the optimal properties of systems that are evoluating in time that are, as it will be seen later on, highly related to the classical mechanics of moving bodies.

Sec. 3.1 deals with the dynamic programming approach, while the next section, sec. 3.2, stipulates the necessary conditions as

given by the maximum principle. Furthermore, some variants that can be transformed to the standard formulation of sec. 3.2 are described in sec. 3.3. The bang-bang principle and the calculus of variations problem are some very important special cases that will be discussed in sec. 3.4.

Finally, the last section deals with the relationship between the two approaches we have presented in sec. 3.1 and sec. 3.2.

3.1 Sufficient conditions for optimality

Hamilton-Jacobi equation

For the sake of simplicity let us first consider the following elementary (fixed-end-point) control problem

$$\max_{\{u(t)\}} J = \int_{t_0}^{t_1} g(x(t),u(t),t)dt + h(x(t_1),t_1)$$

$$\dot{x} = f(x(t),u(t),t)$$

t_1 and $x(t_0) = x_0$ given values .

Assuming a solution exists let

$$F(t,x)$$

be the optimal performance function for the problem $[t,x]$, i.e. the problem

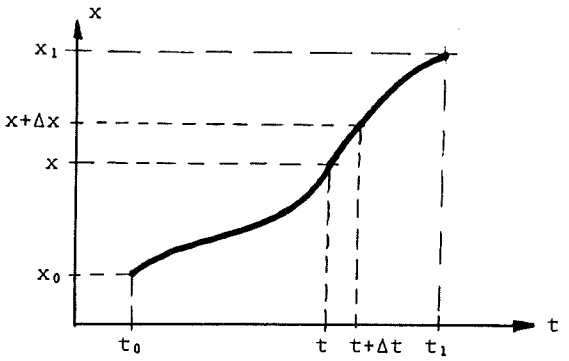
$$\max_{\{u(t)\}} \left[\int_{t}^{t_1} g(x(t),u(t),t)dt + h(x(t_1),t_1)\right]$$

$$\dot{x} = f(x(t),u(t),t)$$

Our problem is thereby embedded in a wider class of problems characterized by their (n+1) initial parameters. According to the Principle of Optimality [3]

$$F(t,x) = \max_{\{u(t)\}} [g(x,u(t),t)\Delta t + F(t+\Delta t, x+\Delta x)]$$

This is illustrated below (remark the similarity to the discrete-time case).



Assuming that $F(t,x)$ is single-valued and <u>continuous differentiable</u> function of the (n+1) variables (this is a critical assumption), we can expand $F(t+\Delta t, x+\Delta x)$ in a Taylor series about the point (t,x) [1] to obtain

$$F(t,x) = \max_{\{u(t)\}} \left[g(x,u(t),t)\Delta t + F(t,x) + \frac{\partial F(t,x)}{\partial x}\Delta x + \frac{\partial F(t,x)}{\partial t}\Delta t + \text{terms of higher order}\right]$$

where $\frac{\partial F(t,x)}{\partial x} = \left(\frac{\partial F(t,x)}{\partial x_1}, \dots, \frac{\partial F(t,x)}{\partial x_n}\right)$ is a row vector.

This yields

$$0 = \max_{\{u(t)\}} \left[g(x,u(t),t) + \frac{\partial F(t,x)}{\partial x} \cdot \frac{\Delta x}{\Delta t} + \frac{\partial F(t,x)}{\partial t} + o(\Delta t)\right]$$

---

1) The couple (t,x) is usually referred as a <u>phase</u>.

Taking the limit as $\Delta t \to 0$ gives

$$\boxed{-\frac{\partial F(t,x)}{\partial t} = \max_{\{u(t)\}} \left[g(x,u(t),t) + \frac{\partial F(t,x)}{\partial x} f(x,u(t),t) \right]}$$

This nonlinear partial differential equation, called Hamilton-Jacobi equation, has as boundary condition

$$F(t_1,x(t_1)) = h(x(t_1),t_1)$$

Once $F(t,x)$ is found, the optimal value of the objective function is then

$$F(t_0,x_0)$$

Let us now consider the following, more general, continuous-time problem

$$\max J = \int_{t_0}^{t_1} g(x(t),u(t),t)dt + h(x(t_1),t_1)$$

$$\dot{x} = f(x(t),u(t),t)$$

$$x(t_0) = x_0 \quad \text{(a given value)}$$

$$u : [t_0,t_1] \to U \quad \text{and} \quad \{u(t)\} \text{ piecewise continuous}$$

where $x \in R^n$, $u \in R^m$, $U \subseteq R^m$. $h : R^{n+1} \to R$ is assumed differentiable and f and g, are assumed to satisfy the following conditions

i) they are continuous differentiable in the variables (x,u) for each fixed $t \in [t_0,t_1]$

ii) except for a finite subset $D \in [t_0,t_1]$, the function f, g, $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial u}$, $\frac{\partial g}{\partial x}$ and $\frac{\partial g}{\partial u}$ are continuous, and

iii) all the last derivatives are bounded for all $t \in [t_0,t_1]$.

These assumptions will guarantee the existence of a solution as shown in every standard treatise on differential equations. It is easily shown that under the same differentiability (smoothness) assumption on $F(t,x)$, the Hamilton-Jacobi equation becomes [4]

$$-\frac{\partial F(t,x)}{\partial t} = \max_{\{u(t)\in U\}}\left[g(x,u(t),t) + \frac{\partial F(t,x)}{\partial x} f(x,u(t),t)\right]$$

with boundary condition: $F(t_1,x(t_1)) = h(x(t_1),t_1)$.

The next theorem expresses the fact that the last equation stipulates sufficient conditions for an optimal control, i.e. if $F(t,x)$ is found that satisfies Hamilton-Jacobi equation then it is the maximum cost function.

Theorem 3

Suppose there exists a differentiable function $F : [t_0,t_1] \times R^n \to R$ which satisfies the Hamilton-Jacobi equation and its respective boundary condition. Suppose there exists a function $\psi : [t_0,t_1] \times R^n \to U$ piecewise continuous in t and Lipschitz [1] in x (this assures a unique solution to the equation of motion), satisfying

$$g(x,\psi(t,x),t) + \frac{\partial F(t,x)}{\partial x} f(x,\psi(t,x),t)$$

$$= \max_{\{u(t)\in U\}}\left[g(x,u(t),t) + \frac{\partial F(t,x)}{\partial x} f(x,u(t),t)\right]$$

Then ψ is an optimal feedback control and $F(t_0,x_0)$ is the value of the maximum of the objective function.

---

1) For any two vectors x^1 and x^2, there exists a finite positive constant K, such that:

$$\|\psi_j(x^1,t)-\psi_j(x^2,t)\| \leq K\|x_j^1-x_j^2\| \quad , \; j = 1,2,\cdots,n \; .$$

If ψ is differentiable and all derivatives are bounded then this condition is satisfied.

Example 7

Let us consider the following continuous-time control problem

$$\max_{\{u(t)\}} \; cx(T)$$

subject to

$$\dot{x} = u(t)x(t) \quad , \quad 0 \leq t \leq T$$

$$x(0) = x_0$$

$$x(t) \geq 0 \quad , \quad 0 \leq t \leq T$$

$$0 \leq u(t) \leq 1 \quad , \quad 0 \leq t \leq T$$

where $c > 0$, $x_0 > 0$, $x(t) \in R$ and $u(t) \in R$.

The Hamilton-Jacobi equation becomes

$$-\frac{\partial F(t,x)}{\partial t} = \max_{\{u(t)\}} \left[\frac{\partial F(t,x)}{\partial x} u(t)x(t)\right]$$

and $\quad F(T,x(T)) = cx(T)$

As $x(t) \geq 0$, the optimal solution will become

if $\frac{\partial F(t,x)}{\partial x} > 0$, $u^*(t) = 1$ and $-\frac{\partial F(t,x)}{\partial t} = x \frac{\partial F(t,x)}{\partial x}$

if $\frac{\partial F(t,x)}{\partial x} < 0$, $u^*(t) = 0$ and $-\frac{\partial F(t,x)}{\partial t} = 0$

Since $\frac{\partial F(T,x)}{\partial x} > 0 \Rightarrow u^*(T) = 1$. The question is now to find if there is a change of the value of the control variable. This simple case can be solved by guessing a form of the solution and see if it can be made to satisfy the differential equation and the boundary conditions. Assume a solution of the form

$$F(t,x) = cxe^{T-t}$$

and it is easily verified that it satisfies these conditions.

Now: $\frac{\partial F}{\partial x} = ce^{T-t} > 0$ then $u^*(t) = 1$, $0 \leq t \leq T$

and $x^* = x_0 e^t$, $0 \leq t \leq T$

and according to Theorem 1 this is an optimal control. Note that the assumption of differentiability in $F(t,x)$ is satisfied in this example.

Complementary remarks

1.- The smoothness assumption is not satisfied for many problems [16] and it is generally not known in advance whether they hold for any particular problem, in such situations it is sometimes possible fo find disjoint regions where solutions to the Hamilton-Jacobi equation can be found. This process of "piecing together" suitable regions often leads to a global proof of optimality. However, we must usually solve our problem in order to determine the regions, and so this procedure is essentially a check on the optimality of a control derived by applying another procedure [17].

2.- We note that the Hamilton-Jacobi equation, being a partial differential equation, is often quite difficult (if not impossible) to solve. This is another reason of why this equation is most often used to check optimality.

3.- The Hamiltonian function for the continuous problem is defined as

$$H(x(t),u(t),p(t),t) = g(x(t),u(t),t) + p(t)f(x(t),u(t),t)$$

where $p(t)$ is a row vector $[1\times n]$. Then the Hamilton-Jacobi equation becomes

$$0 = \frac{\partial F(t,x)}{\partial t} + H\left(x,u^*\left(x, \frac{\partial F(t,x)}{\partial x}, t\right), \frac{\partial F(t,x)}{\partial x}, t\right)$$

$$= \frac{\partial F(t,x)}{\partial t} + \max_{\{u(t)\in U\}} H\left(x,u(t), \frac{\partial F(t,x)}{\partial x}, t\right)$$

3.2 Necessary conditions for optimality

The Pontryagin's maximum principle

In the discrete-time case we obtained necessary conditions for optimality by perturbing, in the small, our problem and performing a linear approximation around the optimal solution, this is the idea behind Kuhn-Tucker theorem. A similar perturbation approach can be applied to our continuous control problem, but this requires a level of mathematical sophistication beyond the scope of these notes. In what follows we will utilize a rather intuitive approach.

Consider the optimal control problem

$$\max_{\{u(t)\}} J = \int_{t_0}^{t_1} g(x(t),u(t),t)dt$$

subject to

$$\dot{x}(t) = f(x(t),u(t),t) \quad , \quad t_0 \leq t \leq t_1$$

$$x(t_0) = x_0 \quad \text{(a given value)}$$

$$x(t_1) \in R^n \quad \text{(free)}$$

$$\{u(t)\} \in U \text{, i.e., } u: [t_0,t_1] \to U$$

where $x(t) \in R^n$, $u(t) \in R^m$ and the functions g and f are differentiable. Moreover t_1 is fixed. Proceeding by analogy to the static case, we define a Lagrangian function which equals the expression to be maximized plus the inner product of the Lagrange multiplier (co-state) vector and the constraints. Since the constraints and the co-states variables are defined

over the entire time interval, however, the inner product is properly treated under the integral sign.

The Lagrangian functional becomes

$$L(u) = \int_{t_0}^{t_1} \{g(x(t),u(t),t) + p(t)[f(x(t),u(t),t) - \dot{x}]\}dt$$

where the co-state vector is

$$p(t) = [p_1(t),p_2(t),\cdots,p_n(t)]$$

Let us now integrate the term $-p(t)\dot{x}$ by parts, then

$$L(u) = \int_{t_0}^{t_1} \{g(x(t),u(t),t) + p(t)f(x(t),u(t),t) + \dot{p}(t)x(t)\}dt - [p(t_1)x(t_1) - p(t_0)x(t_0)]$$

or

$$L(u) = \int_{t_0}^{t_1} \{H(x(t),u(t),p(t),t) + \dot{p}(t)x\}dt - [p(t_1)x(t_1) - p(t_0)x(t_0)]$$

Let us consider the effect of a variation in the control trajectory from $\{u^*(t)\}$ to $\{u^*(t) + \Delta u^*(t)\}$ with a corresponding variation in the state trajectory of $\{x^*(t)\}$ to $\{x^*(t) + \Delta x^*(t)\}$ and a variation in the co-state vector trajectory of $\{p^*(t)\}$ to $\{p^*(t) + \Delta p^*(t)\}$. The variation in the Lagrange functional is

$$\Delta L(u^*) = \int_{t_0}^{t_1} \left\{\frac{\partial H(x^*,u^*,p^*,t)}{\partial u}\Delta u(t) + \left[\dot{p}^*(t) + \frac{\partial H(x^*,u^*,p^*,t)}{\partial x}\right]\Delta x(t) + \left[\frac{\partial H(x^*,u^*,p^*,t)}{\partial p} - \dot{x}^*\right]\Delta p(t)\right\}dt - p(t_1)\Delta x(t_1) + \text{higher-order terms}$$

The co-state vectors are arbitrary, so let us select them to make the coefficient of Δx equal to zero, that is

$$\dot{p}^*(t) = -\frac{\partial H(x^*,u^*,p^*,t)}{\partial x}$$

the so-called co-state equations, with boundary condition

$$p^*(t_1) = 0$$

Let us further observe that the coefficient of Δp is zero, since for admissible control trajectories

$$\dot{x}^* = \frac{\partial H(x^*,u^*,p^*,t)}{\partial p} = f(x^*(t),u^*(t),t)$$

that is

$$\Delta L(u^*) = \int_{t_0}^{t_1}\left[\frac{\partial H(x^*,u^*,p^*,t)}{\partial u}\right]\Delta u(t)dt + \text{higher-order terms}$$

As $J(u^*) = L(u^*)$, a necessary condition for u^* to be a local maximum is $J(u) - J(u^*) = J \leq 0$,

$$\frac{\partial H(x^*,u^*,p^*,t)}{\partial u}\,\Delta u(t) = H(x^*,u^* + \Delta u^*(t),p^*(t),t) - H(x^*,u^*(t),p^*(t),t)$$

and $u^* + \Delta u^*(t)$ is a sufficiently small neighbourhood of u^*, then the higher-order terms are small, then for u^* to be a maximizing control it is necessary that

$$\Delta L(u^*) = \int_{t_0}^{t_1}[H(x^*,u^* + \Delta u^*(t),p^*,t) - H(x^*,u^*,p^*,t)]dt \leq 0$$

for all admissible $\Delta u^*(t)$. From here it is possible to show [18] that a necessary condition for u^* to maximize J is

$$H(x^*,u^*,p^*,t) \geq H(x^*,u(t),p^*(t),t) \quad \text{for} \quad \forall t \in [t_0,t_1]$$

at all admissible controls, otherwise it is possible to construct admissible controls such that $\Delta L(u^*) > 0$. The last inequality indicates that an optimal control must necessarily maximize the Hamiltonian function, this is the so-called Pontryagin's maximum principle [16].

Let us now summarize the necessary conditions for $u^*(t)$ to be an optimal control.

$$\left.\begin{aligned} \dot{x}^*(t) &= \frac{\partial H(x^*,u^*,p^*,t)}{\partial p} \\ \dot{p}^*(t) &= -\frac{\partial H(x^*,u^*,p^*,t)}{\partial x} \\ &\max_{u(t)\in U} H(x^*,u(t),p^*,t) \end{aligned}\right\} \quad \forall t \in [t_0,t_1]$$

and boundary conditions

$$x(t_0) = x_0$$

$$p(t_1) = 0$$

For the sake of completeness, we will now stipulate the maximum principle as a theorem [4], [16]. Let us consider the general control problem

$$\max_{\{u(t)\}} J = \int_{t_0}^{t_1} g(x(t),u(t),t)dt + h(x(t_1),t_1)$$

subject to

equations of motion:	$\dot{x} = f(x(t),u(t),t)$, $t_0 \leq t \leq t_1$
initial condition:	$x(t_0) = x_0$
final condition:	some no yet specific boundary condition where t_1 might be free or fixed
control constraint:	$\{u(t)\} \in U$, i.e. $u: [t_0,t_1] \to U$ (constant compact set) and $\{u(t)\}$ piecewise continuous

where $x(t) \in R^n$, $u(t) \in R^m$, $U \subseteq R^m$, $f : R^n \times R^m \times [t_0,t_1] \to R^n$, $g : R^n \times R^m \times [t_0,t_1] \to R$ and $h : R^n \times [t_0,t_1] \to R$.

We now make the following regularity assumptions on the differential equations to assure the existence of solutions:

i) the functions

$$f \ , \ \frac{\partial f}{\partial x} \ , \ \frac{\partial f}{\partial t} \ , \ g \ , \ \frac{\partial g}{\partial x} \ , \ \frac{\partial g}{\partial t}$$

are continuous (observe that we do not require that f and g have continuous partial derivatives with respect to $u(t)$[1]).

We will further make the following assumptions on the terminal cost function:

ii) the functions

$$h \ , \ \frac{\partial h}{\partial x} \ , \ \frac{\partial h}{\partial t} \ , \ \frac{\partial^2 h}{\partial x^2} \ , \ \frac{\partial^2 h}{\partial x \, \partial t} \ \text{and} \ \frac{\partial^2 h}{\partial t^2}$$

are continuous.

Theorem 4

Let $\{u^*(t)\} \in U$ be an optimal control and let $\{x^*(t)\}$ be the corresponding trajectory. In order that $\{u^*(t)\}$ be optimal, it is necessary that there exists a function $p^*(t)$ such that:

a) $p^*(t)$ corresponds to $u^*(t)$ and $x^*(t)$, so that they are a solution of the canonical system

$$\dot{x}^*(t) = \frac{\partial H(x^*,u^*,p^*,t)}{\partial p}$$

$$\dot{p}^*(t) = -\frac{\partial H(x^*,u^*,p^*,t)}{\partial x} \quad \begin{pmatrix}\text{adjoint}\\ \text{equation}\end{pmatrix}$$

---

1) The requirement of continuous partial derivatives is somewhat stronger than is necessary. In fact this theorem is usually stated by assuming that the Lipschitz condition is satisfied.

and $x(t_0) = x_0$ and some extra boundary conditions, where

$$H(x,u,p,t) = g(x(t),u(t),t) + p(t)f(x(t),u(t),t)$$

b) $u^*(t)$ satisfies the maximum principle

$$H(x^*,u^*,p^*,t) = \max_{u \in U} H(x^*,u(t),p^*,t)$$

for all $t \in [t_0,t_1]$ except possibly for a finite set.

This theorem represents the necessary conditions for optimality. We observe that the first equation of the canonical system is the equation of motion and is actually independent of $p(t)$. Remark also that the Hamiltonian has to be maximized at the points of continuity of $\{u^*(t)\}$, the points at which the Hamiltonian may not be maximized must be discontinuities of $\{u^*(t)\}$.

For minimum time control problems $g = 1$ and $h = cx(t_1)$, then H is the scalar product of the vector $p(u)$ and the velocity vector $\dot{x}(t)$, therefore the geometrical sense of the theorem is that the optimal control tries to "drive away" the system optimally in some direction determined by the vector $p^*(u)$, at each moment.

Observe also that these necessary conditions do not depend upon the final conditions or upon the time being free or fixed, to these necessary conditions we have just to add some boundary conditions (exactly n because x_0 is given) for the canonical equation. Let us suppose that we wished to minimize rather than to maximize. What changes in our necessary conditions would this lead to? We can see immediately that the form of the canonical system will not change. On the other hand, in condition b), the maximization of H will be replaced by the minimization of H, while the boundary conditions to be seen later remain the same [17]. In effect, then, the only necessary condition which distinguishes between maximization and minimization is condition b), the other conditions may thus be

viewed as necessary conditions for an "extremal" (notice the analogy to stationary solutions to static problems).

Boundary conditions

If in our "proof" of the maximum principle we have not specified the final conditions for $x(t_1)$ and t_1, it is easily shown that the following condition has to be satisfied [18]

$$\left[\frac{\partial h(x^*(t_1),t_1)}{\partial x} - p^*(t_1)\right]\Delta x(t_1) + \left[\frac{\partial H(x^*(t_1),u^*(t_1),p^*(t_1),t_1)}{\partial t} + \frac{\partial h(x^*(t_1),t_1)}{\partial t}\right]\Delta t_1 = 0$$

Thus for our first case t_1 is fixed, that is $\Delta t_1 = 0$, $h = 0$ and since $x(t_1)$ is free, then $p^*(t_1) = 0$. On the other hand, if $x(t_1)$ and t_1 were given values, then $x(t_1) = x_1$, the required 2n conditions. The tableau below specifies the boundary conditions under different assumptions. Sometimes the target set S is defined by a set of k equations, $T(x(t)) = 0$ or $T(x(t),t) = 0$ and the following "smoothing" assumptions

i) the functions T, $\frac{\partial T}{\partial x}$, and $\frac{\partial T}{\partial t}$ are continuous, and

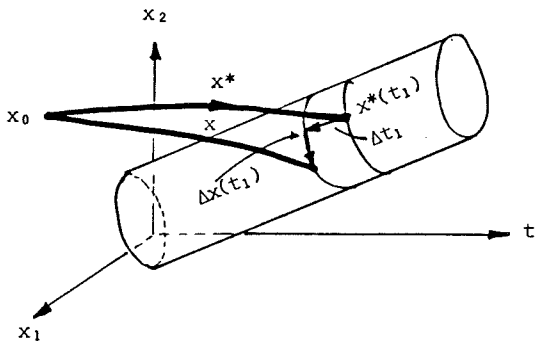
ii) the Jacobian matrix $\frac{\partial T}{\partial x}$ is of rank k.

At the optimal solution (x^*,u^*,p^*):

Problem	Description	Boundary condition equations	Remarks
t_1 fixed	1. $x(t_1)=x_1$ specified final state	$x(t_0)=x_0$ $x(t_1)=x_1$	2n equations to determine 2n constants of integration
	2. $x(t_1)$=free	$x(t_0)=x_0$ $\frac{\partial h}{\partial x}\Big/_{t=t_1} -p(t_1)=0$	2n equations to determine 2n constants of integration
	3. $x(t_1)$ on the surface $T(x(t))=0$ where $T(k\times 1)$	$x(t_0)=x_0$ $\frac{\partial h}{\partial x}\Big/_{t=t_1} -p(t_1)= \sum_{i=1}^{k} d_i \cdot \frac{\partial T_i}{\partial x}\Big/_{t=t_1}$ $T(x(t_1))=0$	(2n+k) equations to determine the 2n constants of integration and the variables $d_1,\cdots,d_k$

Problem	Description	Boundary-condition equation	Remarks
t_1 free	4. $x(t_1)=x_1$ specified final state	$x(t_0)=x_0$ $x(t_1)=x_1$ $H_{t_1}+\left.\frac{\partial h}{\partial t}\right\|_{t=t_1}=0$	(2n+1) equations to determine the 2n constants of integration and t_1
	5. $x(t_1)$ free	$x(t_0)=x_0$ $\left.\frac{\partial h}{\partial x}\right\|_{t=t_1}-p(t_1)=0$ $H_{t_1}+\left.\frac{\partial h}{\partial t}\right\|_{t=t_1}=0$	(2n+1) equations to determine the 2n constants of integration and t_1
	6. $x(t_1)$ on the moving point $\theta(t)$	$x(t_0)=x_0$ $x(t_1)=\theta(t_1)$ $H_{t_1}+\left.\frac{\partial h}{\partial t}\right\|_{t=t_1}$ $+\left[\left.\frac{\partial h}{\partial x}\right\|_{t=t_1}-p(t_1)\right]\left[\frac{d\theta}{dt}\right]_{t=t_1}=0$	(2n+1) equations to determine the 2n constants of integration and t_1
	7. $x(t_1)$ on the surface $T(x(t))=0$	$x(t_0)=x_0$ $\left.\frac{\partial h}{\partial x}\right\|_{t=t_1}-p(t_1)=\sum_{i=1}^{k} d_i\cdot\left.\frac{\partial T_i}{\partial x}\right\|_{t=t_1}$ $T(x(t_1))=0$ $H_{t_1}+\left.\frac{\partial h}{\partial t}\right\|_{t=t_1}=0$	(2n+k+1) equations to determine the 2n constants of integration, the variables $d_1,\cdots,d_k$ and t_1
	8. $x(t_1)$ on the moving surface $T(x(t),t)=0$	$x(t_0)=x_0$ $\left.\frac{\partial h}{\partial x}\right\|_{t=t_1}-p(t_1)=\sum_{i=1}^{k} d_i\cdot\left.\frac{\partial T_i}{\partial x}\right\|_{t=t_1}$ $T(x(t_1),t_1)=0$ $H_{t_1}+\left.\frac{\partial h}{\partial t}\right\|_{t=t_1}$ $=\sum_{i=1}^{k} d_i\cdot\left.\frac{\partial T_i}{\partial t}\right\|_{t=t_1}$	(2n+k+1) equations to determine the 2n constants of integration, the variables $d_1,\cdots,d_k$ and t_1

As an example take case 8. The final state lies on the intersection of the k hypersurfaces defined by T , and it is moving in time.



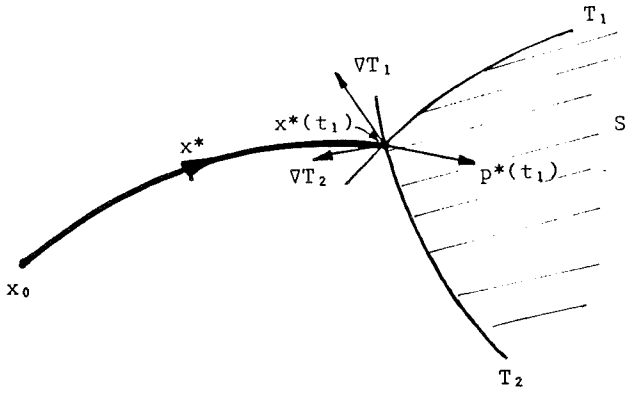
Then it is easily shown that the admissible values of the (n+1) vector

$$\begin{bmatrix} \Delta x(t_1) \\ \cdots\cdots \\ \Delta t_1 \end{bmatrix}$$

are normal to each of the gradient vectors

$$\begin{bmatrix} \frac{\partial T_1(x^*(t_1),t_1)}{\partial x} \\ \cdots\cdots\cdots\cdots \\ \frac{\partial T_1(x^*(t_1),t_1)}{\partial t} \end{bmatrix}, \ \cdots\cdots, \ \begin{bmatrix} \frac{\partial T_k(x^*(t_1),t_1)}{\partial x} \\ \cdots\cdots\cdots\cdots \\ \frac{\partial T_k(x^*(t_1),t_1)}{\partial t} \end{bmatrix}$$

Then the two vectors coefficients of $\Delta x(t_1)$ and Δt_1 must be linear combinations of the above mentioned gradient vectors as shown in the tableau. If $h = 0$, then it is easily seen that the vector $p^*(t_1)$ has to be orthogonal to the surface S at $x^*(t_1)$, this is why the final boundary conditions are sometimes known as transversality conditions.



Example 8

Let us consider the continuous-time control problem of example 7.

The Hamiltonian function is

$$H = p(t)u(t)x(t)$$

The necessary conditions become

$$\max H \qquad \begin{cases} 1 & p(t) > 0 \\ 0 & p(t) > 0 \\ ? & p(t) = 0 \end{cases}$$

$$0 \leq u(t) \leq 1 \Rightarrow u^*(t) =$$

$$x(t) > 0$$

$$\dot{p} = -\frac{\partial H}{\partial x} \Rightarrow \dot{p} = -u(t)p(t)$$

$$\dot{x} = \frac{\partial H}{\partial p} \Rightarrow \dot{x} = u(t)x(t)$$

$$x(0) = x_0 > 0$$

$$p(T) = c > 0$$

Since $p(T) > 0 \Rightarrow u^*(T) = 1$. Let us now find out if $p(t)$ changes of sign. Going backwards

$$\dot{p} = -p(t) \quad \text{and} \quad p(T) = c$$

This has as a solution: $p(t) = ce^{T-t}$ that is bigger than zero for all, $0 \leq t \leq T$. That is the optimal control becomes: $u^*(t) = 1$, $0 \leq t \leq T$ and

$$x^*(t) = x_0 e^t \ , \ 0 \leq t \leq T$$

$$p^*(t) = ce^{T-t} \ , \ 0 \leq t \leq T$$

Further comments

1.- To prove Theorem 4 is not an easy task, thus under the weak differentiability assumptions our variational approach is not longer suitable. A complete, based on topological arguments, proof was first published in English in full detail - in 1962 in the book of Pontryagin, et al. [16], an "easy to understand" heuristic proof following the same arguments can be found in [17]. However, the derivation given by Lee and Markus [19] is more satisfactory. A geometrical approach can be found in [20] and [21]. A generalization of the maximum principle to include extensions to infinite-dimensional spaces can be found in [22].

2.- Pontryagin and his co-workers [16] have also derived another useful necessary conditions, these are:

i) If t_1 is fixed, and H does not depend explicitly on time, then

$$H(x^*(t),u^*(t),p^*(t)) = \text{constant} \ , \ \forall t \in [t_0,t_1]$$

ii) If t_1 is free, and the H does not depend explicitly on time, then

$$H(x^*(t),u^*(t),p^*(t)) = 0 \ , \ \forall t \in [t_0,t_1]$$

The constancy of function H, analogous to the Hamiltonian function of analytical mechanics, corresponds to the law of the conservation of energy in conservative mechanical systems (the H, as is well known expresses the systems total energy).

3.- We have assumed that the initial condition x_0 was given, Theorem 4, is still valid for the case that x_0 has to be included in a 'smooth' surface I. The only modification has to be done on the boundary condition $x(t_0) = x_0$, where similar conditions to the transversality conditions for $x(t_1)$ have to be introduced. That is $p^*(t_0)$ is normal to I at $x^*(t_0)$ [4].

4.- In order to be able to handle pathological cases Theorem 4 is usually presented including an additional constant $p_0^* \geq 0$ so that the Hamiltonian is now

$$H(x,u,p,t) = p_0\, g(x,u,t) + pf(x,u,t) .$$

Theorem 4 remains the same except for the change of the Hamiltonian. The pathological cases occur when $p_0^* = 0$; when $p_0^* \neq 0$, we may choose $p^* = 1$. As far as I know useful constraint qualifications as it was the case for the static problem are not known [19].

5.- Although the (strong) necessary conditions for the discrete-time problem parallels those of the continuous-time control problem, we should notice that some additional assumptions were necessary on the discrete-time case to achieve such results. In general, it is <u>impossible</u> to obtain a maximum principle for the discrete case that is as strong as that for the corresponding continuous. The difference can be traced to the fact that in continuous-time a control history perturbation of large magnitude but arbitrarily short duration can be considered which introduces only small changes in the resulting trajectory; while in discrete-time

a large magnitude perturbation of control at any instant introduces a large change in the trajectory. Thus, only in continuous-time it is possible to introduce control perturbations that are simultaneously large in magnitude (at a given instant) but small in their effect on the objective and constraint functionals.

6.- Applying the maximum principle, one starts from the relation:

$$\max_{u(t)\in U} H(x^*,u(t),p^*,t)$$

in order to deduce a relation between the optimal control $u^*(t)$ and the corresponding $x^*(t),p^*(t)$. In general, this relation takes the form $u^*(t) = k(x^*(t),p^*(t))$. If $x^*(t),p^*(t)$ uniquely specify $u^*(t)$, then one deals with the so-called normal problems. However, there are cases where this necessary condition does not provide us with a well-defined expression for the optimal control, such problems are called singular, and they arise quite often whenever the Hamiltonian function H is linear in the control vector. Singularity does not necessarily mean that either the optimal control does not exist or it cannot be defined; it simply states that the above mentioned necessary condition does not lead to a well-defined relation between $u^*(t)$, $x^*(t)$ and $p^*(t)$ and we must manipulate the other necessary conditions in an effort to determine such a well-defined relationship. Unfortunately, at this time, general results regarding the existence, optimality and implementation of singular controls are very limited [17], [23].

7.- The maximum principle is valid, provided that an optimal control does, indeed, exist. Question regarding the existence of optimal controls are complicated; it is doubtful that it will ever be possible to state general existence results, that at the same time are easy to verify, without relatively strong assumptions. Most of the results which

are available are very mathematical in nature, see for instance [19].

The problem of deducing whether or not an optimal control exists is two-fold:

i) Is it possible to find at least one control which satisfies the imposed control constraints $u(t) \in U$, and which transfers the system from x_0 to S?

ii) If the answer to i) is yes; under which conditions can one guarantee that an optimal control will also exist?

The problem of existence is usually related to show that the set U is compact and that the functional $J(u(t))$ is continuous. For the special cases of linearity in $u(t)$, many results are available [17]. In practice, one usually attempts to find an optimal control rather than to try to prove that one exists [13].

8.- The necessary conditions are not, in general, sufficient for optimality. By analogy with the static case, it is natural to expect that suitable concavity/convexity assumptions have to be imposed on the functions and sets involved. Thus, one can relate the search for sufficient conditions to the <u>convexity</u> of the set of reachable states[1)]. It is not difficult to show that if $(x^*(t),u^*(t))$ solves the given problem it is <u>sufficient</u> that for all admissible $(x(t),u(t))$ and for all t for which $p(t)$ is differentiable (and $h(x(t_1),t_1) = 0$)

$$H(x^*(t),u^*(t),p(t),t) - H(x(t),u(t),p(t),t) \geq \dot{p}(t)(x(t) - x^*(t))$$

---

1) The set of reachable states at time t, A_t, we shall mean the subset A_t of the state space such that: $A_t = \{x : \exists\{u(t)_{[t_0,t]}\} \in U_t$ (set of admissible controls from t_0 to t) such that the corresponding solution $x(t)$ to the equations of motion, starting at x_0, satisfies the property $x(t) = x\}$.

This simple sufficiency condition is suitable as a reference but not very useful as it stands since nothing is said on how to find $p(t)$. Some conditions under which this inequality is satisfied for problems without state-space constraints (and $p_0 = 1$) are the following

i) U convex, and H is jointly concave in x and u for fixed p and t.

ii) $H(x(t),u^*(t),p^*(t),t)$ is concave in $x(t)$ for any value of $t \in [t_0,t_1]$.

iii) $H(x(t),u(t),p(t),t)$ is linear in $x(t)$ [24].

It is easily seen that i) implies ii) but the reverse is obviously not correct, further results along these ideas and correct proofs can be found in [26].

Other type of sufficient conditions can be obtained by combining the necessary conditions with the sufficient conditions provided by the Hamilton-Jacobian equation, see further results in [17].

3.3 Variants of the problem

The maximum principle is valid for a quite general family of continuous-time control problems. Nevertheless by suitable transformations many problems can be suitable modified so that Theorem 4 is still valid. Let us see some examples.

Higher-order equations of motion

Assume that the evolution of the system is described by

$$\ddot{x} = f(\dot{x},x,u)$$

with $$x(t_0) = x_0 \quad , \quad \dot{x}(t_0) = a$$

The problem can be reduced to the same form as the basic problem as follows. Define

$$\frac{dx}{dt} = x_1 \quad , \quad \text{with} \quad x_1(t_0) = a$$

Then the second-order differential equation becomes

$$\frac{dx_1}{dt} = \dot{x}_1 = f(x, x_1, u)$$

Our equations of motion are then

$$\dot{x} = x_1 \qquad , \quad x(t_0) = x_0$$
$$\dot{x}_1 = f(x, x_1, u) \quad , \quad x_1(t_0) = a$$

and Theorem 4 is applicable.

It is worth noting that equations of motion described by an $n^{\underline{th}}$ order differential equations can also be handled in the same manner.

Choice of extra parameters [20]

Sometimes the problem depends on an additional parameter α, that has to be found such that the performance function is maximized, for example

$$\dot{x} = f(x(t), u(t), \alpha, t)$$

Now let us introduce an additional state variable

$$x_{n+1}(t) = \alpha \quad , \quad \forall t \in [t_0, t_1]$$

and

$$\frac{d\, x_{n+1}(t)}{dt} = 0 \quad , \quad \forall t \in [t_0, t_1]$$

This new equation is introduced in the last system of equations. Using this transformation nonautonomous systems can be transformed to autonomous.

Simultaneous control and state constraints

Thus far we have admitted controls which belong to a prescribed constant subset $U \subseteq R^m$ and the states were not restricted other that at the initial and final time. A similar approach used to handle restrictions in the static case could be applied. Thus, under some regularity conditions, if we have some constraints as

$$q(x(t),u(t)) = 0$$

then introduce a Lagrange vector $y(t)$, and the H becomes

$$H(x(t),u(t),p(t),y(t),t) = g(x(t),u(t),t) + p(t)f(x(t),u(t),t) - y(t)q(x(t),u(t))$$

and the conditions of Theorem 4 are still valid. For the case of inequalities of the type ($\leq$), we have further to demand $y(t) \geq 0$ and $y(t)q(x(t),u(t)) = 0$ (complementary slackness) at the optimal solution.

Obviously this modification is needed for the case that one or more of the restrictions are active, that is for intervals of time for which $y(t) > 0$, otherwise the unrestricted version of Theorem 4 is applied ($y(t) = 0$). Some additional necessary conditions upon the vector $p(t)$ have to be added at the point where an interval is entered, these are known as jump conditions, see further [16]. There exist other approaches to handle constraints, see for instance [18] and [20].

3.4 Some special continuous-time problems

Minimum-time control of linear systems

Let us consider the following problem.

$$\max_{\{u(t)\}} J = -\int_{t_0}^{t_1} dt = -(t_1 - t_0)$$

subject to

$$\dot{x}(t) = a(x(t),t) + B(x(t),t)u(t)$$

$$M_i^- \leq u_i(t) \leq M_i^+ \ , \ i = 1,\cdots,m \ , \ \forall t \in [t_0,t_1]$$

given $t_0, x(t_0) = x_0 \ , \ x(t_1) = 0$

where $a(x(t),t)$ is a $(n\times 1)$ vector and B is a $(n\times m)$ matrix whose components may be explicitly dependent on the states and t. M_i^+ and M_i^- are known upper and lower bounds for the $i^{\underline{th}}$ control component. The Hamiltonian function is

$$H = -1 + p(t)[a(x(t),t) + B(x(t),t)u(t)]$$

Let us express B as

$$B(x(t),t) = [b_1(x(t),t) \mid b_2(x(t),t) \mid \cdots\cdots \mid b_m(x(t),t)]$$

where b_i are column vectors of B. To max H is equivalent to

$$\max_{M_i^- \leq u_i \leq M_i^+} \left[\sum_{i=1}^{m} p^*(t)[b_i(x^*(t),t)]u_i(t) \right]$$

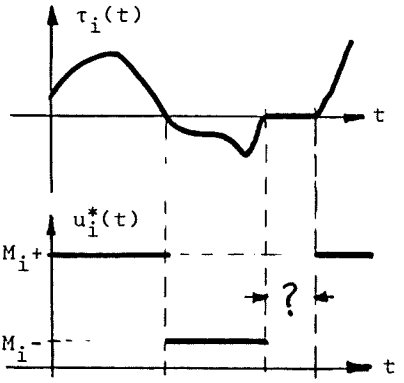
By properties of Linear Programming, the form of the optimal control is

$$u_i^*(t) = \begin{cases} M_i^+ & p^*(t)b_i(x^*(t),t) > 0 \\ M_i^- & p^*(t)b_i(x^*(t),t) < 0 \\ \text{undetermined} & p^*(t)b_i(x^*(t),t) = 0 \end{cases}$$

This is the mathematical statement of the well-known bang-bang principle, that is the optimal control to obtain minimum-time

response is maximum effort throughout the interval of operation.

The function $\tau_i(t) = p^*(t) \cdot b_i(x^*(t),t)$ is known as the switching function and is illustrated below



Notice that if $\tau_i(t)$ passes through zero, a switching of the control $u_i^*(t)$ is indicated. If for some finite interval $\tau_i(t) = 0$, the maximation of H provides no information about how to select $u_i^*(t)$; then the problem is not normal, because we have some singular conditions.

If the equation of motion is linear and stationary (autonomous), that is

$$\dot{x} = Ax(t) + Bu(t)$$

where A, B are constant matrices, and

$$-1 \leq u_i(t) \leq 1 \quad , \quad i = 1,2,\cdots,m$$

Then, obviously, if an optimal control exists, it is bang-bang. Moreover the following results can be shown [16], [17]:

i) A singular interval cannot exist. That is for a singular interval to exist, it is necessary that the system be uncontrollable. Conversely, if the system is completely controllable, a singular interval cannot exist.

Thus existence of optimal control can be found by testing the system for controllability [18]. Further results are:

ii) Existence theorem. If all of the eigenvalues of A have nonpositive real parts, then an optimal control exists that transfers x_0 to the origin.

iii) Uniqueness theorem. If an extremal control exists it is unique. Since an optimal control, if one exists, must be an extremal control, e.g. must satisfy the necessary conditions, these theorems indicate that a control which satisfies the maximum principle and the required boundary conditions must be the optimal control. Thus if an optimal control exists, satisfaction of the maximum principle is both necessary and sufficient for time-optimal control of antonomous, linear systems.

iv) Number of switchings theorem. If the eigenvalues of A are all real, and a (unique) time-optimal control exists, then each control component can switch at most (n-1) times.

Further discussion about the design of linear time-optimal systems can be found in [17].

Calculus of variations

The calculus of variations is a branch of mathematics that deals with a special form of our continuous-time problem. This is the case where constraints in the state and control trajectory are not present and the equation of motion is: $\dot{x}(t) = u(t)$. Historically this is the first type of control problems that have been solved. The different conditions an optimal solution

must satisfy are quite analogue to the static optimization problem, and they can be found by applying the classic concept of functional variation [27], by making use of the Hamilton-Jacobi equation [3] or by utilizing the maximum principle [16].

The classical calculus of variations problem is

$$\max_{\{x(t)\}} J = \int_{t_0}^{t_1} g(x(t),\dot{x}(t),t)dt$$

given that: $x(t_0) = x_0$

$x(t_1) = x_1$

i) Euler equations

Assume that $g(x,\dot{x},t)$ and $x(t)$ are twice continously differentiable then a necessary condition for an extremal solution, if one exists, is

$$\frac{\partial g}{\partial x} - \frac{d}{dt}\left(\frac{\partial g}{\partial \dot{x}}\right) = 0$$

This Euler equation is a system of second-order differential equations in x. Notice the analogy to the stationarity condition in the static theory. The Euler equations take some special forms under some assumptions as shown below.

$g(x,\dot{x},t)$ does not depend explicitly on	Euler equation
$\dot{x}(t)$	$\frac{\partial g}{\partial x} = 0$ (stationarity condition)
$x(t)$	$\frac{\partial g}{\partial \dot{x}} = \text{constant}$
t	$g - \frac{\partial g}{\partial \dot{x}_i}\dot{x}_i = \text{constant}$ $i = 1,2,\cdots,n$

ii) Legendre condition

This condition is analogous to the second order necessary condition in the static case. Thus at the optimal solution $\{x^*(t)\}$ the Hessian matrix

$\frac{\partial^2 g}{\partial \dot{x}^2}$ must be negative definite or negative semidefinite.

This is a necessary condition for an extremal and a sufficient condition for a local maximum.

iii) Weierstrass condition

This condition is analogous to the one in the static case that the objective function be concave. If $\{x(t)\}$ is an extremal solution then for any other admissible trajectory

$$E(x,\dot{x},t,\dot{z}) \leq 0$$

is a sufficient condition for a global maximum, where E is Weierstrass excess function, defined as

$$E(x,\dot{x},t,\dot{z}) = g(x,\dot{z},t) - g(x,\dot{x},t) - \left(\frac{\partial g}{\partial \dot{x}}\right)(\dot{z}-\dot{x})$$

This condition is always met if g is a concave function on x.

iv) Weirstrass-Erdman corner conditions

While the trajectory $\{x(t)\}$ is continuous, $\{\dot{x}(t)\}$ need be only piecewise condinuous and, hence, may actually consist of segments of curves joined at points called corners at which $\dot{x}(t)$ is discontinuous. The Weierstrass-Erdman corner conditions require that $\frac{\partial g}{\partial \dot{x}}$ and $g - \frac{\partial g}{\partial \dot{x}}\dot{x}$ be continuous across the corner. Thus if a corner occurs at time τ

$$\left[\frac{\partial g}{\partial \dot{x}}\right]_{\tau^-} = \left[\frac{\partial g}{\partial \dot{x}}\right]_{\tau^+}$$

$$\left[g - \frac{\partial g}{\partial \dot{x}}\,\dot{x}\right]_{\tau^-} = \left[g - \frac{\partial g}{\partial \dot{x}}\,\dot{x}\right]_{\tau^+}$$

Further remarks

1. The Euler conditions can be extended to take account of constraints obtaining then conditions quite similar to Kuhn-Tucker conditions for the static case [16].

2. More complex boundary conditions can be handled as it was shown above while discussing the maximum principle.

3.5 Relationship of the maximum principle to dynamic programming

The study of this relationship will provide us with a very simple geometrical interpretation of the maximum principle. For the sake of simplicity, let us consider the following continuous-time problem

$$\max_{\{u(t)\}} J = \sum_{i=1}^{n} c_i\, x_i(t_1)$$

subject to

$$\dot{x} = f(x,u,t) \quad , \quad t_0 \leq t \leq t_1$$

so that

$$x(t_0) = x_0$$

$$u(t) \in U$$

and subject to some boundary conditions, for instance $x(t_1) = x_1$. Assume that the maximum principle is applicable to this problem, then the similarity between the maximum principle and the dynamic programming approach is not a casual one as stipulated below.

Theorem 5

Let $F(t,x)$, value of J on the optimal control if at time t the system is found at x, be a continuous and continuously differentiable function of x and t, then

i) for all t, $\{x^*,u^*\}$ satisfies the maximum principle condition with respect to $p^*(t)$, where

$$p^*(t) = \frac{\partial F(t,x^*)}{\partial x}$$

and

$$-\frac{\partial F(t,x^*)}{\partial t} = H(x^*,u^*,p^*,t)$$

ii) $F(t,x)$ satisfies the Hamilton-Jacobi equation.

Example 9

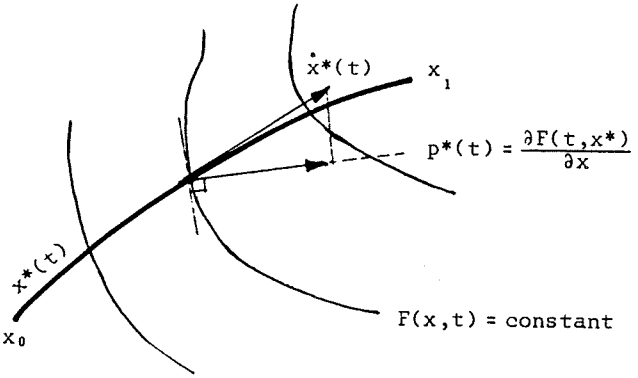
If we take the problem of example 7, where

$$p^*(t) = ce^{T-t}, \quad F(t,x) = cxe^{T-t} \quad \text{and}$$

$$H = cx^*e^{T-t}u^*(t)$$

it is easily verified that Theorem 5 is satisfied.

This theorem has a geometric interpretation [28]. Recall that the vector $p^*(t)$ defines the direction of "maximum acceleration" of the system at each moment. As follows from the last theorem, this direction, at each fixed moment of time, is determined in its turn by the gradient of the function $F(t,x)$ with respect to $x(t)$. It thus turns out that, at each fixed moment of time, whatever the position of the system, the optimal control tends to "accelerate" the latter in the direction defined by the gradient of the function $F(t,x)$. This is illustrated below:



The maximum principle requires that $\{u(t)\}$ be so selected that the projection of the velocity $\dot{x}(t)$ of the system in the phase space upon the direction $p(t)$ normal to the iso-surface, $F(t,x)$ = constant, shall be maximal. This relationship of the function $p(t)$ to the gradient of the function $F(t,x)$ is retained even for more complicated boundary conditions.

The last theorem provides also a useful interpretation of the co-state variables, thus the extremal co-states are the sensitivity of the maximum value of the performance measure to changes in the state value.

Another important conclusion of our last theorem is the fact that Hamilton-Jacobi equation implies the maximum principle. The maximum principle does not, however, imply the Hamilton-Jacobi equation since the maximum principle does not require the basic assumption that the $F(t,x)$ be continuously differentiable. Thus the maximum principle permits discontinuities in $p(t)$ and the "jump" condition has to be added. We note further that the existence of singular controls is also related to the presence of discontinuities of the derivatives $\frac{\partial F(t,x)}{\partial x}$, a singular trajectory occurs precisely along a surface of discontinuity.

Finally, the maximum principle has another important advantage upon dynamic programming, thus it, in essence, breaks up the solution of the Hamilton-Jacobi equation into two steps. The first step, max H, can generally be easily taken, and, it often yields insight into the nature of the optimal control (see for example the bang-bang principle).

4. REFERENCES

[1] Luenberger, D.: Optimization by vector space methods, John Wiley 1969.

[2] Kirk, D.E.: An introduction to dynamic programming, IEEE Trans. Education (1967), 212-219.

[3] Bellman, R.E., and Dreyfus, S.E.: Applied Dynamic Programming, Princeton University Press 1962.

[4] Varaiya, P.P.: Notes on Optimization, Van Nostrand 1972.

[5] Wagner, H.M.: Principles of Operations Research, Prentice-Hall 1969.

[6] Mitten, L.C.: Composition principles for synthesis of optimal multistage processes, Operations Research, 12, 1964.

[7] Nemhauser, C.L.: Introduction to dynamic programming, Wiley & Sons 1966.

[8] Cambini, A.: On the validity of the recurrence equation of dynamic programming, Universita di Pisa, Dipartimento de ricerca operativa e scienze statistiche, 1974.

[9] Gottfried, B.S., and Weisman, J.: Introduction to optimization theory, Prentice-Hall 1973.

[10] Fan, L., and Wang, Ch.: The discrete maximum principle, Wiley & Sons 1964.

[11] Halkin, H.: Optimal control for systems described by difference equations in Advances in Control Systems, Ed. Leondes, Academic Press 1964.

[12] Holtzman, J.M.: On the maximum principle for non-linear discrete-time systems, IEEE Transactions on Automatic Control, vol. 11, 1966.

[13] Athans, M.: The status of optimal control theory and applications for deterministic systems, IEEE Transactions on Automatic Control, vol. 11, 1966.

[14] Canon, M., et al.: Theory of optimal control and mathematical programming, McGraw-Hill 1970.

[15] Luenberger, D.: "Mathematical Programming and Control Theory: trends of interplay" in Perspectives on Optimization, Ed. Geoffrion, A.M. Addison-Wesley, 1972.

[16] Pontryagin, et al.: The mathematical theory of optimal processes, J. Wiley 1962.

[17] Athans, M., and Falb, P.: Optimal control, McGraw-Hill 1966.

[18] Kirk, D.: Optimal control theory, an introduction, Prentice-Hall 1970.

[19] Lee, E.B., and Markus, L.: Foundations of optimal control theory, John Wiley 1967.

[20] Leitman, G.: An introduction to optimal control, McGraw-Hill 1966.

[21] Halkin, H.: On the necessary condition for optimal control systems, J. d'Analyse Mathematique, vol. 12, 1963.

[22] Neustadt, L.W.: A general Theory of Extremals, J. Computer and System Sciences, 3, No. 1, 1969.

[23] Bryson, A., and Yo-chi Ho: Applied optimal control, Ginn and Company 1969.

[24] Rozonoér, L.I.: L.S. Pontryagin's maximum principle in the theory of optimum systems, I, II and III in [25].

[25] Oldenburger, R., ed.: Optimal and self-optimizing control, The MIT Press 1966.

[26] Seierstand, A., and Sydsæter, K.: Sufficient Conditions in Optimal Control Theory, Institute of Economics, University of Oslo 1975.

[27] Gelfand, I.M., and Fomin, S.V.: Calculus of variations, Englewood Cliffs, Prentice-Hall 1963.

[28] Feldbaum, A.A.: Optimal control systems, Academic Press 1965.

[29] Ravn, H.F.: The discrete Maximum Principle with Nonlinear Hamiltonian, IMSOR 1976.

CHAPTER 7

DYNAMIC OPTIMIZATION:

Computational methods

1. INTRODUCTION

Numerical methods to solve optimal control problems are still in the research stage. Few of them have been implemented. Lacking valuable comparative computing experimentations, it is even difficult to judge the relative merits of the methods intended to solve the same problem. Therefore, it is difficult to give a precise description of the different approaches, many strategies are available. Moreover, like all fields in expansion, inevitable confusions arise: similar words are used for different meanings, some attack computing difficulties by using analytical expressions; others attack the analytical part of the optimization procedure for solving indirectly the computing difficulties, etc.

The purpose of this chapter is to give an overview on the global strategies, based on our theoretical discussion, that are available to solve numerically optimal control problems. Numerical and computer programming aspects will not be discussed in detail, see further [1] and the references in [2] and [3].

The two main approaches are dynamic programming and the (continuous and discrete) maximum principle. They will be described in their most elementary versions, and their advantages and disadvantages will be pointed out.

2. DISCRETE-TIME OPTIMAL CONTROL

As we have seen, this problem can be regarded as a finite-dimensional mathematical programming problem and the full range of algorithms, we have discussed for the static optimization problems, becomes available. There has been a good deal of computational experience accumulated on problems of this type which, on the whole, has essentially verified that general nonlinear programming algorithms can effectively solve control

problems. Actually most of the case studies solved by static optimization methods are control problems. Indeed, control problems often serve as convenient sources of large dimensional test problems for new algorithms. Generally, however, these straight forward applications of known algorithms to control problems carry little significance toward the fundamental development of either the static or dynamic optimization. Special, but very applicable, cases are linear programming problems with special structure and problems with quadratic performance function and linear stage-transformations and constraints, see further [4].

Here we will be concerned with those algorithms that can be developed based on our theoretical discussion on discrete-time optimal control problems. We have discussed two approaches: dynamic programming and the discrete maximum principle, the table below shows the range of applicability of these two methods.

		Assumption	
	DISCRETE-TIME CONTROL PROBLEM	Discrete in space	Continuous in space
Approach	Dynamic Programming	Applicable	Applicable
	Discrete Maximum Principle		Applicable

Thus dynamic programming is a more general approach, but for problems that are continuous in space it is in general impossible to find out which approach is more suitable. As it will be seen below both methods have their advantages and disadvantages.

The standard computational algorithm of dynamic programming

The functional equation of dynamic programming gives us a computational procedure to find an optimal control law, thus under the assumptions stipulated in our theoretical discussion

$$F(k,x) = \max_{u(k) \in U} \{r(x,u(k),k) + F(k+1,f(x,u(k),k))\}$$

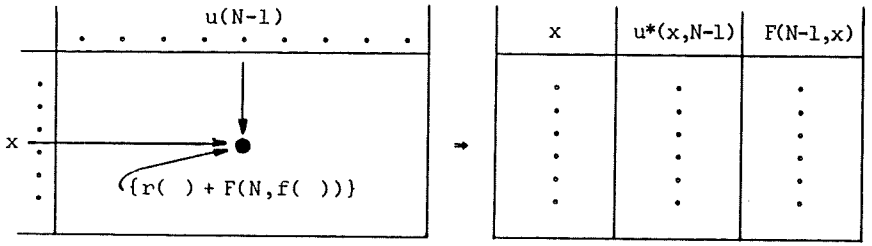
for $0 \leq k \leq N-1$

and

$$F(N,x) = \varphi(x) .$$

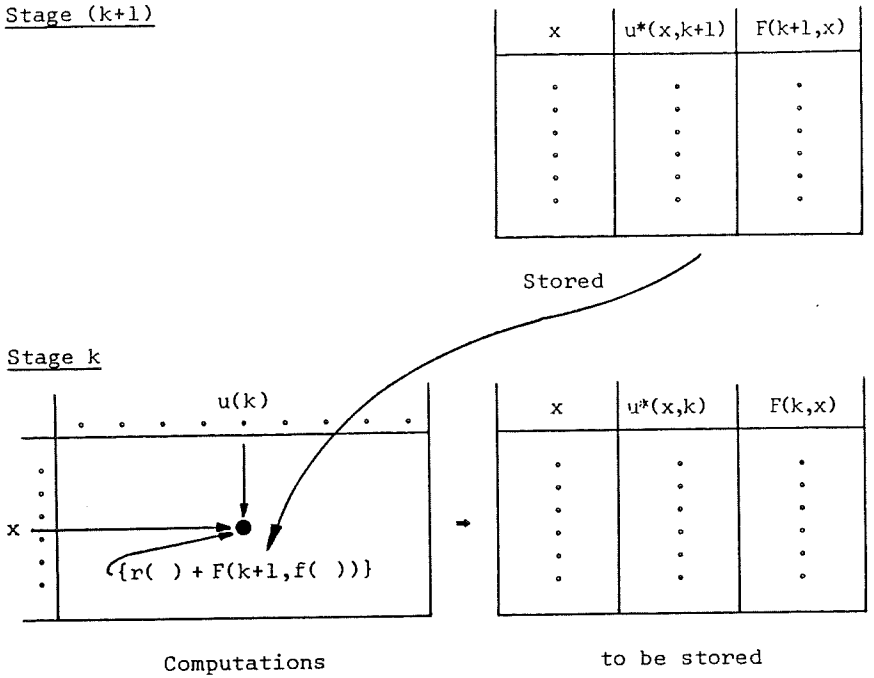
Assume now that each state variable $x_i(k)$, $i = 1,2,\cdots,n$, can take V_i values, and each control variable $u_j(k)$, $j = 1,2,\cdots,m$, can take C_j values.

Initially, $F(N,x)$ is found for all admissible x , this is done directly. Next, at $k = N-1$, for each admissible state x , each admissible control $u(N-1)$ is applied, and the next (admissible) state $f(x,u(N-1),N-1)$ is computed. Now $F(N,f(x,u(N-1),N-1))$ is found from the last tabulated values and $r(x,u(N-1),N-1)$ is computed directly. The sums of these quantities for each admissible state are then compared, and the maximum value is stored at $F(N-1,x)$. The optimal control $u^*(x,N-1)$, is stored as the value of u for which the maximum is attained. This is illustrated below



Computations to be stored

The procedure continues in this manner, with $F(k,x)$ and $u^*(x,k)$ being computed in terms of $F(k+1,x)$, this is schematically shown below.



The procedure stops when $k = 0$ is reached, at this point $x(0)$ is given, $u^*(x(0),0)$ can then be computed and the optimal performance value is $F(0,x(0))$. The other optimal values $u^*(x,k)$ can be found backwards by looking at the stored tables containing the optimal control laws, until the entire sequence control is obtained.

Remarks

1. This computational procedure is very appealing for a number of reasons. In the first place, thorny questions of existence and uniqueness are avoided; as long as there is at least one feasible control sequence, then this procedure guarantees that the global maximum is found. Furthermore, extremely general types of system equations, performance criteria, and constraints can be handled. Constraints ac-

tually reduce the computational burden by decreasing the admissible sets. Finally, the optimal control is obtained as a closed loop control.

2. Dynamic programming uses the principle of optimality to reduce dramatically the number of calculations required to determine the optimal control law. Let us compare the dynamic programming algorithm with direct enumeration of all possible control sequences. Consider a control problem with $n = m = 1$, and $V_1 = 10$, $C_1 = 4$. The table below shows a comparison of the number of calculations required by dynamic programming and by direct enumeration.

Number of stages	Dynamic programming	Direct enumeration
1	40	40
2	80	200
3	120	840
4	160	3400
5	200	13640
6	240	54600
L	40 L	$\sum_{k=1}^{L} [10 \cdot 4^k]$

Note: In this example it is assumed that $\varphi(\ ) = 0$, and that $x(0)$ is not given.

The important point is that the number of calculations required by direct enumeration increases exponentially with the number of stages, while the computational requirements of dynamic programming increase linearly.

3. There is one serious drawback with dynamic programming: for high-dimensional systems the number of high-speed storage

locations becomes prohibitive. This difficulty is usually known as the curse of dimensionality [5]. Recall that to evaluate $F(k,x)$ we need access to the values of $F(k+1,x)$. Thus, for $n = 3$ with $V_i = 100$, $\forall i$, this means that $10^2 \times 10^2 \times 10^2 = 10^6$ storage locations are required; this number approaches the limit of rapid-access storage available with current computers. There is nothing to prevent us from using low-speed storage; however, this will drastically increase computation time. Several techniques that have been developed to alleviate the curse of dimensionality will be discussed below.

4. The standard algorithm assumes that the state and control variables are discrete variables, otherwise the admissible range of state and control values has to be quantized, in this case the goodness of the optimal solution will depend on the number of grid points, moreover interpolation will be required to evaluate $F(k+1,x)$ for the case that the output state at a given stage does not fall exactly on a grid value. If the interpolation formulas are of the proper form, it is sometimes possible to use efficient search procedures to solve the maximization problem at each stage, rather than quantization of control and direct comparison.

5. In certain well-behaved problems it is possible to obtain an accurate approximation to $F(k+1,x)$ over all x by expressing $F(\ )$ as a low-order polynomial in x. In such cases a saving in high-speed memory requirement can be made by storing only the coefficients of the polynomial, rather than values of $F(\ )$ for all quantized x.

6. Continuous-time control problems can be solved by quantizing in space and time to find then a discrete-time control problem. A substantial saving in high-speed memory requirements can be obtained by using the so-called "state increment dynamic programming". This method and other techniques that have been developed to alleviate the curse of dimen-

sionality can be found in [3], [5] and [1]. In these references and those suggested in our theoretical discussion many practical applications within different areas can be found.

7. A main disadvantage of dynamic programming is the fact that a feasible solution is first found when the first stage is reached, that is when the optimal solution is found. This can be avoided by applying the next approach.

Computational algorithms based on the discrete maximum principle

Having established a set of conditions that defines the discrete form of the maximum principle, we seek a computational algorithm that will allow us to implement the theoretical development. This is not a particularly easy task, since we are required to solve a set of mixed-boundary difference equations.

We begin by restating the canonical difference equations

$$x^*(i+1) - x^*(i) = f(x^*(i),u^*(i),i)$$

and

$$p^*(i) = \frac{\partial r(x^*(i),u^*(i),i)}{\partial x} + p^*(i+1)\left[I + \frac{\partial f(x^*(i),u^*(i),i)}{\partial x}\right] .$$

With these forward recursive equations that transform the state variables and backward recursive adjoint equations, the following algorithm suggests itself:

a. Choose an initial set of decision variables.

b. Using these estimates for the decision variables and the boundary conditions, generate a set of state variables.

c. The estimated values for the decision variables and the state variables are now used, together with the boundary conditions, to generate a set of adjoint variables.

d. A new set of decision variables is obtained by rendering the stagewise Hamiltonian stationary with the state variables and the adjoint variables retaining their most recent values. We then return to step b and obtain a new set of state variables and a new set of adjoint variables.

e. This iterative procedure is continued until successive sets of the state variables have converged to within some preassigned tolerance.

Steps b and c are easy to perform if the initial state is given and the final state is free, otherwise iterative procedures have to be developed to carry on these steps in order to satisfy the transversality conditions.

In what concerns step d, several of the numerical methods to solve static optimization problems could be applied.

Remarks

1. The above algorithm was first suggested by Katz [6], but it is not the only means of implementation of the discrete maximum principle as we will see later on. It is probably the best-known computational algorithm, however, because of its simplicity and its straightforward nature. A general difficulty with algorithms that attempt to solve problems of this type is connected with convergence problems. The method has been shown to be convergent for systems having certain structural features, the most notable being linear, straight-chain systems. Unfortunately, the method may not converge for more complicated problems [7]. Applications of this procedure to the aerospace and chemical-process areas are shown by Fan and Wang [8].

2. Another method is the following: Choose an initial set of state and adjoint variables, then maximize the stagewise

Hamiltonian to find the control variables that are then substituted in the canonical equations to obtain new values of the state and adjoint variables; iterations continue till the boundary conditions are satisfied. Additional discussion on computational methods has been discussed in [9] and [10].

3. This method is not so heavily subject to storage and dimensionality restrictions. However, the introduction of the adjoint variables does increase the dimensionality of the solution space and it requires solution of a mixed-boundary-type problem that gives convergence problems. Even if a solution is obtained using the maximum principle, we have no assurance that the desired optimum has been located, since the maximum principle represents only a necessary condition for an optimum. Finally, use of the discrete maximum principle is restricted to problems in which the stage-transformations equations and stage return functions have continuous partial derivatives with respect to the state and decision variables.

4. Few attempts have been made to derive an algorithm applicable to processes with bounded state variables. However, the resulting formulations become, in general, very complicated when they are applied in solving an optimization problem numerically. The problems with bounded state variables do not give any trouble to the method of dynamic programming since, in this method, the optimal decisions are determined for the whole allowable domain of the state variables, and hence the optimal policy thus obtained automatically satisfies the constraints on the state variables.

3. CONTINUOUS-TIME OPTIMAL CONTROL

The evaluation of optimal controls for continuous-time problems is, in general, a difficult and time-consuming task. It is only

for very few special cases that an analytical expression for the optimal control can be found. Thus, the evaluation of the optimal controls must be carried out using an iterative procedure and a digital computer. When employing a digital computer to solve continuous-time optimal control problems, operations such as the solution of differential equations cannot be executed exactly. This means that either the original problem itself or its solution (or both) must be approximated in some appropriate fashion. There are two primary strategies to this approximation problem:

a) To discretize the original problem at the outset, converting it to a discrete-time optimal control problem to be solved by the methods described in the last section. There are a number of important questions regarding the appropriateness of such approximations in terms of their convergence to the true solutions, see for instance [11]. Unfortunately, this approach can easily get out of hand yielding a problem of enormously large dimension with poor accuracy properties. Moreover, if the maximum principle is to be applied, we should remember that the results obtained for continuous-time systems are more general than those for the discrete-time case. And

b) The strategy largely employed by control theorists, is based on analysis of the continuous-time problem. The problem is constantly regarded as one in continuous-time and one attempts to carry out, at least approximately, the calculations dictated by the continuous-time analysis. The primary components of these calculations is integration of the equations of motion and the adjoint equations. Efficient integration procedures are the predictor-corrector method, or the Runge-Kutta procedure, see [12], [13].

It is impossible to muster a definitive argument that favors either of these strategies over the other. Which is better is dependent on the actual problem to be solved. For the majority

of control problems, however, the second strategy has a great advantage because of the relatively short time required for accurate integration. It is only for problems with many constraints on the state and control vectors, or which have a particularly simple structure, such as being linear or quadratic programs that the first strategy is suitable.

If the second strategy is chosen, one usually has to solve the so-called "two-point boundary value problem" arising from the application of the maximum principle of Pontryagin to the optimal control problem.

Two-point boundary-value problems

Assuming that the state and control variables are not constrained by any boundaries, that the final time t_1 is fixed, and that $x(t_1)$ is free, we can summarize the two-point boundary-value problem that results from the maximum principle:

$$\dot{x}^*(t) = \frac{\partial H}{\partial p} = f(x^*,u^*,t) \tag{1}$$

$$\dot{p}^*(t) = -\frac{\partial H}{\partial x} = -\frac{\partial g(x^*,u^*,t)}{\partial x} - \frac{\partial f(x^*,u^*,t)}{\partial x} p^*(t) \tag{2}$$

$$\frac{\partial H}{\partial u} = 0 = \frac{\partial f(x^*,u^*,t)}{\partial u} p^*(t) + \frac{\partial g(x^*,u^*,t)}{\partial u} \tag{3}$$

$$\begin{aligned} x^*(t_0) &= x_0 \\ p^*(t_1) &= \frac{\partial h(x^*(t_1),t_1)}{\partial x} . \end{aligned} \tag{4}$$

Let us assume that (3) can be solved to obtain

$$u^*(t) = k(x^*,p^*,t) .$$

If this expression is substituted in (1) and (2), we have a set of 2n first-order ordinary differential equations, the so-called

reduced differential equations, involving $x^*(t)$, $p^*(t)$, and t. The boundary conditions for these differential equations are given by (4). Unfortunately, the boundary values are split, so that direct integration of (1) and (2) cannot be found. Thus the difficulty of solving optimal control problems by using the maximum principle is caused by the combination of split boundary values and nonlinear differential equations. We have then to recur to iterative procedures.

One of the best known procedures is the so-called variation of extremals method. Here one tries to solve the reduced differential equations. One can proceed in two ways:

a) Guess an initial value for $p(t_0)$. Integrate, forward in time, (1) and (2) using $x(t_0)$ and $p(t_0)$. Check if the boundary conditions at time t_1 are satisfied, if not, change $p(t_0)$ using say, a gradient method or Newton's method [13].

b) Guess $x(t_1)$ and $p(t_1)$ that satisfy (4). Integrate backward in time. Adjust $x(t_1)$ until the state trajectory passes through the given initial state $x(t_0)$.

A more detailled discussion of these methods can be found in [14].

The second popular method is the so-called gradient method. The basic idea is to guess a control function, such that the integration of the canonical equations yields a solution which satisfies (4). One then adjusts the initial guess until (3) is satisfied, or until the performance measure is maximized (notice the similarity of the gradient method for the static case). This method is easier to generalize to take account of control constraints [16].

The third popular method of computing optimal controls is the often called quasi-linearization method. In this method one tries to solve the reduced differential equations, thus one guesses trajectories $x(t)$ and $p(t)$ which meet the boundary

conditions. These values will not, in general, satisfy the reduced differential equations. One then linearizes the differential equations about the initial guesses. The linearized equations are solved for a new set of trajectories which still meet the boundary conditions. The procedure is repeated until the generated trajectories approach the solutions of (1) and (2).

These three approaches have obviously to be modified if the states and control trajectories are constrained, and if the final time is not fixed, see further [15]. The table below gives a comparison of the features of these three iterative methods for solving nonlinear two-point boundary-value problems [14].

Feature	Variation of extremals	Gradient method	Quasi-linearization
Initial guess	$p(t_0)$ (or $x(t_1)$)	$u(t)$	$x(t)$ and $p(t)$
Stop rule	$p(t_1) = \frac{\partial h(x(t_1),t_1)}{\partial x}$	$\frac{\partial H}{\partial u} \simeq 0$	Canonical equations
Importance of initial guess	Divergence may result from poor guess	Not usually crucial to convergence	Divergence may result from poor guess
Convergence	Once convergence begins (if it does) it is rapid	Approaches a maximum rapidly, then slows down drastically	Converges quadratically in the vicinity of the optimum

Remarks

1. Since there is always some arbitrary initial guess involved, the convergent iterative process will converge to a solution which is "close" (in some sense) to the initial guess. If the extremal controls are not unique, then it is easy to see that two different initial guesses may converge to two different extremal solutions (neither of which need be optimal). In the absence of any sufficiency conditions or of

any uniqueness properties of the extremal controls, one must compute all of the extremal controls and then compare the values of the corresponding cost functionals in order to isolate the optimal control. This problem is compounded by the fact that the most commonly used iterative computational algorithms are local in nature.

2. The general feeling remains that there does not exist one method or one procedure suitable for treating all nonlinear dynamical optimization problems. This will never be the case for the simple reason that these problems cannot be grouped within one single class. As a consequence, the previously described techniques will only give the whole of their potential power if they are applied to each particular case, tested, compared, reorganized, etc. This is the practical and the unique approach for filling the gap between theory and application.

4. REFERENCES

[1] Pun, L.: Introduction to optimization practice, John Wiley 1969.

[2] Athans, M.: The status of optimal control theory and applications, IEEE Transactions on Automatic Control, vol. 11, 1966.

[3] Larson, R.E.: A survey of Dynamic Programming Computational Procedures, IEEE Transactions on Automatic Control, 767-774, 1967.

[4] Luenberger, D.G.: Mathematical Programming and Control Theory in "PERSPECTIVES ON OPTIMIZATION", ed. A.M. Geoffrion, Addison-Wesley Publishing Company 1972.

[5] Nemhauser, G.L.: Introduction to Dynamic Programming, John Wiley 1966.

[6] Katz, S.: Best operating points for staged systems, Ind. Eng. Chem. Fundamentals, 4, 226, 1962.

[7] Denn, M.M.: Convergence of a method of successive approximations in the theory of optimal processes, Ind. Eng. Chem. Fundamentals, 4, 231, 1965.

[8] Fan, L., and Wang, Ch.: The discrete maximum principle, John Wiley 1964.

[9] Lapidus, L., and Luus, R.: Optimal control of engineering processes, Ginn-Blaisdell 1967.

[10] Gurel, O., and Lapidus, L.: The maximum principle and discrete systems, Ind. Eng. Chem. Fundamentals, 7, 617, 1968.

[11] Cullum, J.: Perturbations of optimal control problems, J. SIAM Control, vol. 4, No. 3, 1966.

[12] Henrici, P.: Discrete Variable Methods in Ordinary Differential Equations, John Wiley 1968.

[13] Hamming, R.: Numerical Methods for Scientists and Engineers, McGraw-Hill 1962.

[14] Kirk, D.E.: Optimal Control Theory, Prentice-Hall 1970.

[15] Sage, A.P.: Optimum Systems Control, Englewood Cliffs: Prentice-Hall 1968.

[16] Mufti, I.H.: Computational methods in optimal control problems, Springer-Verlag 1970.

CHAPTER 8

DYNAMIC OPTIMIZATION:

Case studies

Matajura understood. Without asking for any promises in terms of time, he became Banzo's servant. He cleaned, he cooked, he washed, he gardened. He was ordered never to speak of fencing or to touch a sword. He was very sad at this; but he had given his promise to the master, and resolved to keep his word. Three years passed for Matajura as a servant.

One day while he was gardening, Banzo came up quietly behind him and gave him a terrible whack with a wooden sword. The next day in the kitchen the same blow fell again. Thereafter, day in, day out, from every corner and at any moment, he was attacked by Banzo's wooden sword. He learned to live on the balls of his feet, ready to dodge at any movement. He became a body with no desires, no thought — only eternal readiness and quickness.

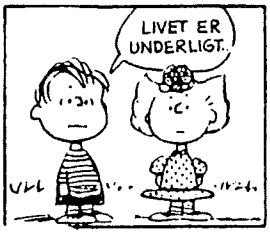
Banzo smiled, and started lessons. Soon Matajura was the greatest swordsman in Japan.

1. INTRODUCTION

Most of the case studies we have discussed and solved applying static optimization methods in chapter 4 can also be solved employing dynamic programming methods or the discrete maximum principle. Further applications to control problems have been reviewed in [1] and [2].

Here we will discuss some economic and technical applications of the continuous maximum principle. Pontryagin's maximum principle has been employed with great success in solving technical control problems within aerospace and chemical systems. That is not the case in what concerns operations research and economic problems, where very few dynamical models have been implemented. This poor rate of implementation is probably related to the lack of sounded numerical methods to deal with constrained control problems. The different problems will be presented in their most elementary version, further study can be followed by consulting the reference's list at the end of this chapter.

Dynamic programming applications will be seen in part THREE.



2. CASE 1: A simple inventory problem

We want to determine the optimal lot size under the following assumptions:

a) The demand rate, r, is constant.

b) Demand is instantaneously satisfied.

c) A constant amount, Q (the lot size) is ordered at fixed time intervals, $T = Q/r$ (the period).

d) The costs involved are constant,

- h the cost of storing one inventory unit for one time unit,
- A the set up cost.

The problem is formulated as follows:

$$\min J = cx(T)$$

where

$$\dot{x} = h\left(1 - \frac{rt}{Q}\right) + \frac{Ar}{Q^2} \quad , \quad x(0) = 0 \quad , \quad c = r$$

Indeed J is found to be

$$J = \frac{Ar}{Q} + \frac{hQ}{2}$$

the average total cost and x defines the cumulated costs.

The Hamiltonian becomes

$$H = p\left[h\left(1 - \frac{rt}{Q}\right) + \frac{Ar}{Q^2}\right]$$

and the adjoint equation

$$\dot{p} = 0 \quad , \quad p(T) = -r \Rightarrow p(t) = -r \quad , \quad t \in [0,T]$$

and the stationarity condition

$$\frac{\partial H}{\partial Q} = 0 \Rightarrow p\left[\frac{hrt}{Q^2} - \frac{2Ar}{Q^3}\right] = 0$$

Since $p \neq 0$, the only bounded solution is then

$$Q^* = \frac{2A}{ht} .$$

Now for $t = T \Rightarrow T = \frac{Q}{r}$, we obtain the well-known form

$$Q^* = \sqrt{\frac{2Ar}{h}} .$$

3. CASE 2: A production planning problem

Let us consider a production system governed by the equations

$$\dot{I}(t) = P(t) - S(t) \quad , \quad I(0) = I_0$$
$$\dot{S}(t) = -\lambda P(t) \quad , \quad S(0) = S_0$$

where

$I(t)$ is the inventory level (state)

$S(t)$ is the rate of sales (state)

$P(t)$ is the rate of production (control)

and $\underline{P} \leq P(t) \leq \bar{P}$ (bounds in control), and

λ is a constant.

We shall minimize costs, that is

$$Z(T) = \int_0^T [cP(t) + hI(t)]dt$$

where c is the cost per unit time of producing at rate $P(t)$, and h is the cost per unit time of holding inventory level $I(t)$.

By defining

$$Z(t) = \int_0^t [cP(t) + hI(t)]dt \ , \ Z(0) = 0$$

The model becomes

$$\min_{\underline{P} \le P(t) \le \bar{P}} Z(T)$$

subject to

$$\begin{bmatrix} \dot{I} \\ \dot{S} \\ \dot{Z} \end{bmatrix} = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & 0 \\ h & 0 & 0 \end{bmatrix} \begin{bmatrix} I \\ S \\ Z \end{bmatrix} + \begin{bmatrix} 1 \\ -\lambda \\ c \end{bmatrix} P$$

The adjoint equations are

$$\dot{p} = -A' p = \begin{bmatrix} 0 & 0 & -h \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} \quad \text{and} \quad \begin{aligned} p_1(T) &= 0 \\ p_2(T) &= 0 \\ p_3(T) &= -1 \end{aligned}$$

This is a system of homogeneous linear differential equations that solved gives

$$p = \begin{bmatrix} -hT + ht \\ -hTt + \frac{ht^2}{2} + \frac{hT^2}{2} \\ -1 \end{bmatrix}$$

The Hamiltonian is then

$$H = p_1 \dot{I} + p_2 \dot{S} + p_3 \dot{Z}$$

$$= P\left[-\frac{\lambda h}{2} t^2 + (h+\lambda hT)t - \frac{\lambda hT^2}{2} - hT - c \right] + S(hT-ht) - hI$$

The optimal control is then bang-bang

$$P^* = \begin{cases} \bar{P} & \text{if } \tau(t) > 0 \\ \underline{P} & \text{if } \tau(t) < 0 \\ \text{unspecified} & \text{if } \tau(t) = 0 \end{cases}$$

with the switching function

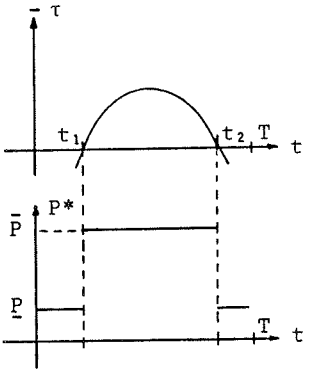
$$\tau(t) = -\frac{\lambda h}{2} t^2 + (h+\lambda hT)t - \frac{\lambda hT^2}{2} - hT - c$$

It is easy to see that singular control does not exist.

To find the form of the optimal policy, we must examine the zeroes of $\tau(t)$

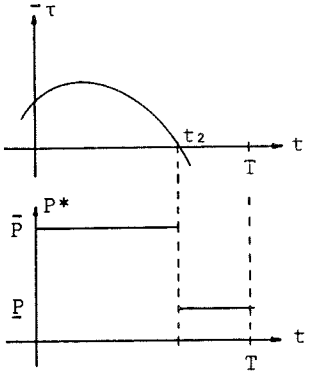
$$t = \frac{1+\lambda T}{\lambda} \pm \frac{1}{\lambda} \sqrt{1 - \frac{2c\lambda}{h}}$$

Letting t_1 be the smaller root and t_2 the larger root, if real, we sketch below the optimal policy for the four possible cases.



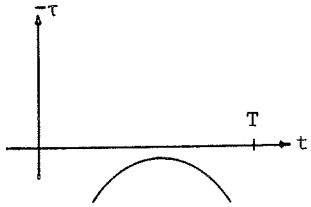
CASE 1:

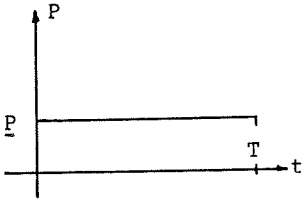
t_1 , t_2 , real positive



CASE 2:

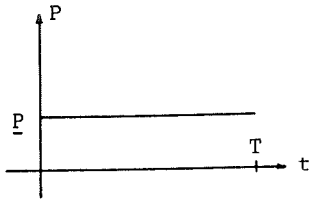
t_1 , t_2 , real; t_2 positive





CASE 3:

t_1 , t_2 , real, negative



CASE 4:

t_1 , t_2 , imaginary

Now it is easy to see that the form of the optimal solution is highly dependent on the value of the constant λ.

Thus if λ h/2c, the optimal solution is CASE 4.

Try yourself the other possibilities!

4. CASE 3: Control of a chemical plant

Consider the problem of maximizing the final amount of the product y during a two-stage chemical reaction in which $x \to y \to z$. Assuming first-order kinetics for the x and y rates, the reaction rates are taken to be

$$\dot{x} = -ax$$

$$\dot{y} = ax - by$$

$$\dot{z} = g(x,y,z)$$

where the rate coefficients a,b are related so that

$$b = \rho a^k$$

with ρ and k positive constants.

The amount of the waste product z formed is seen not to influence the x and y reactions, thus we drop the last differential equation.

Given that $y(t_1)$ is to be maximized by proper choice of the control a(t) with the following conditions

$$\text{at } t = t_0 = 0: \quad x(t_0) = x_0 \ , \ y(t_0) = y_0$$

$$t = t_1 \quad \text{(specified)} \ .$$

The adjoint equations are

$$\dot{p}_1 = a(p_1 - p_2)$$

$$\dot{p}_2 = bp_2 .$$

The Hamiltonian is

$$H = p_1(-ax) + p_2(ax-by)$$

that for $k \geq 1$ is maximized at

$$a = \left[\frac{1}{\rho k} \frac{x}{y} \left(\frac{p_2 - p_1}{p_2} \right) \right]^{\frac{1}{k-1}}$$

while this values minimize H for $k < 1$. Here it is obviously assumed that $p_2(t) > p_1(t)$, so that an extremal control exists.

We have then left a two-point boundary-value problem with transversality conditions $p_1(t_1) = 0$, $p_2(t_1) = 1$ that can be solved by a numerical method.

Let us now assume that the admissible control must lie in the range:

$$a_\ell \leq a(t) \leq a_u .$$

The adjoint equations remain the same, but at points where the control goes on and off the control boundary during the process, the corner conditions are:

$$p_1/_{t^+} = p_1/_{t^-} \qquad p_2/_{t^+} = p_2/_{t^-}$$

$$H/_{t^+} = H/_{t^-}$$

at points of entrance and departure from the control boundary.

Another modification is that while maximizing H we have to take account of the bounds on $a(t)$. Then optimal solutions can be found for both $k < 1$ and $k \geq 1$. For the last case $a^*(t)$ is either an interior point or an extreme point, while for $k < 1$ the $a^*(t)$ is always an extreme point.

5. CASE 4: A road building problem

A road is to be built over uneven terrain. The road must be constructed in such a manner that the slope at no point exceeds a maximum value, θ_m, in magnitude.

Let $y(x)$ be the equation describing the road to be built and $t(x)$ the equation of the terrain. Assume that the costs to be minimized are given by

$$J = \int_0^L [t(x) - y(x)]^2 \, dx$$

subject to:

$$\left| \frac{dy(x)}{dx} \right| \leq \theta_m \, .$$

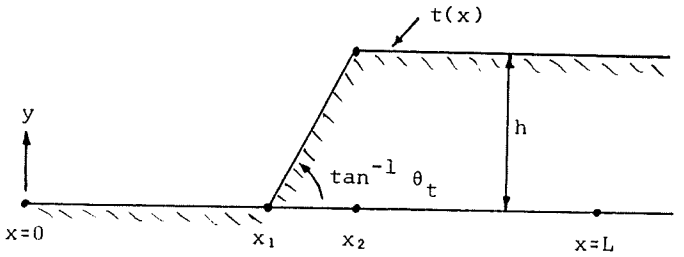
As a simple example assume that

$$t(x) = 0 \qquad , \quad 0 \leq x \leq x_1$$

$$t(x) = \theta_t (x - x_1) \qquad , \quad x_1 \leq x \leq x_2$$

$$t(x) = h \qquad , \quad x_2 \leq x \leq L$$

as shown below.



At its ends the construction must join with roadways already built, requiring that the end values of $y(x)$ be specified: $y(0) = 0$, $y(L) = h$.

This problem can be easily formulated as a standard control problem

$$\max z(L)$$

subject to

$$\frac{dy}{dx} = \theta(x)$$

$$\frac{dz}{dx} = -[t(x) - y(x)]^2$$

$$|\theta(x)| \leq \theta_m$$

where $\theta(x)$ is the control variable. The end conditions are

$$\text{at } x = 0 \ , \ y(0) = 0 \ , \ z(0) = 0$$

$$\text{at } x = L \ , \ y(L) = h \ .$$

The adjoint equations are

$$\dot{p}_1 = 2p_2[t(x) - y(x)]$$

$$\dot{p}_2 = 0$$

with transversality conditions: $p_2(L) = -1$.

The Hamiltonian becomes

$$H = p_1 \theta - p_2[t(x) - y(x)]^2$$

that is linear in θ . The optimal control is then bang-bang with switching function $\tau(t) = p_1$. There are then 2 control possibilities

$$p_1 \neq 0 \Rightarrow \theta^*(x) = \theta_m \quad \text{sign} \quad \tau(t)$$

$$p_1 = 0 \Rightarrow \theta^*(x) = \frac{dt(x)}{dx} \quad \text{for some interval (singular control)}$$

That is either the road takes on its maximum positive or negative slope or the ground terrain is identically followed.

6. CASE 5: A one-sector economy

Let us consider a modified model of the one-sector economy described in chapter 5, sec. 2. The control model in question is

$$\max_{\{c(t)\}} J = \int_0^T e^{-rt} \frac{1}{1-\eta} c(t)^{1-\eta} \, dt$$

subject to

$$\dot{k}(t) = f(k(t) - c(t) - \delta k(t)$$

$$k(0) = \bar{k}$$

$$y_T = \bar{y}$$

where

J	is the performance function
r	is a time rate of welfare discount
$c(t)$	is instantaneous consumption at time t
T	is the final time period
η	is the elasticity of marginal utility with respect to consumption
$\dot{k}(t)$	is the net investment
$f(k(t)$	is the production function
$k(t)$	is the capital stock

δ is the rate of depreciation

$\bar{k}$ is the initial stock

$y_T = \bar{y}$ is the terminal output uniquely determined by k_T

For computational purposes we can write the discrete-time version of this problem, e.g.

$$J = \sum_{i=0}^{T-1} \frac{1}{(1+\varphi)^i} \frac{1}{1-\eta} c_i^{1-\eta}$$

$$k_{i+1} = f(k_i) - c_i + (1-\delta)k_i$$

$$k_0 = \bar{k}$$

$$y_T = f(k_T) = \bar{y}$$

The welfare function

$$u(c) = \frac{1}{1-\eta} c_i^{1-\eta} \quad , \quad \eta \geq 0 \ , \ \eta \neq 1$$

has the following properties

$$u'(c_i) = c_i^{-\eta} \geq 0$$

$$\frac{u'(c_i)}{u(c_i)/c_i} = 1 - \eta = \text{constant elasticity of utility}$$

$$u''(c_i) = -\eta\, c_i^{-\eta-1} \leq 0 \qquad \text{disminishing marginal welfare}$$

$$\lim_{\eta \to 0} u(c_i) = c_i$$

In what concerns the production function, we assume that:

$$f(k(t)) = e^{zt}\ \gamma\ k(t)^{\beta}\left(\ell_0\ e^{rt}\right)^{1-\beta}$$

$$= e^{[r(1-\beta)+z]t}\ \gamma\ \ell_0^{1-\beta}\ k(t)^{\beta}$$

$$= e^{qt}\ a[k(t)]^{\beta}$$

where

$$a = \gamma\ \ell_0^{1-\beta}$$

which can be approximated in discrete time with a one-period investment lag by

$$y_i = f\left[k_i\right] = (\ell+g)^i\ a\ k_i^{\beta}$$

where

- z = the rate of neutral technical progress
- γ = efficiency parameter
- β = elasticity of output with respect to capital
- ℓ_0 = initial labor force
- r = rate of growth of labor force
- g = $r(1-\beta) + z$

Applying the discrete-time maximum principle we obtain the following necessary conditions

$$k_{i+1} = (1+g)^i\ a\ k_i^{\beta} - c_i + (1-\delta)k_i \tag{1}$$

$$p_{i+1} = \left[(1+g)^i\ \beta\ a\ k_i^{\beta-1} + 1-\delta\right]^{-1}\ p_i \tag{2}$$

$$c_i = \left[(1+\varphi)^i\ p_{i+1}\right]^{-1/\eta} \tag{3}$$

$$p_T = \upsilon \tag{4}$$

$$k_0 = \bar{k} \tag{5}$$

where

p_i = the adjoint variables, that is the "shadow price" of capital

and

υ = a "shadow price" on the terminal constraint.

This system of equations can be solved by the following routine:

(a) Choose a p_0

(b) Use k_0 and p_0 in (2) to obtain p_1

(c) Use p_1 in (3) to obtain c_0

(d) Use c_0 and k_0 in (1) to obtain k_1

(e) Repeat steps (b) through (d), increasing the index by one on each repetition until k_T is obtained

(f) Compare y_T to $\bar{y}$; stop if they are sufficient "close", otherwise choose a new p_0 and return to (b).

7. REFERENCES

[1] Athans, M.: The status of optimal control theory and applications for deterministic systems, IEEE Transactions on Automatic Control, vol. 11, 1966.

[2] Larson, R.: A survey of dynamic programming computational procedures, IEEE Transactions on Automatic Control, Dec. 1967.

CASE 1 and 2

[3] Connors, M., and Teichroew, D.: Optimal control of dynamic operations research models, International textbook company, 1967.

[4] Adiri, I., and Ben-Israel, A.: An extension and solution of Arrow-Karlin type production models by the Pontryagin maximum principle, Cahier du CEDRO, vol. 8, N° 3, 1966.

[5] Lansdowne, Z.: The theory and applications of generalized linear control processes, Technical report, N° 10, Stanford University, Department of OR, 1970.

[6] Sprzeuzhouski, A.: A problem in optimal stock management, Journal of optimization theory and applications, vol. 1, N° 3, 1967.

[7] Bensoussan, A., et al.: Management Applications of Modern Control Theory, North Holland, 1974.

CASE 3 and 4

[8] Citron, S.: Elements of optimal control, Holt, Rinehart and Winston, Inc., 1969.

[9] Fan, L.T.: The continuous maximum principle, J. Wiley and Sons, 1966.

[10] Zahradnik, R: Theory and techniques of optimization for practicing engineers, Barnes and Noble, 1971.

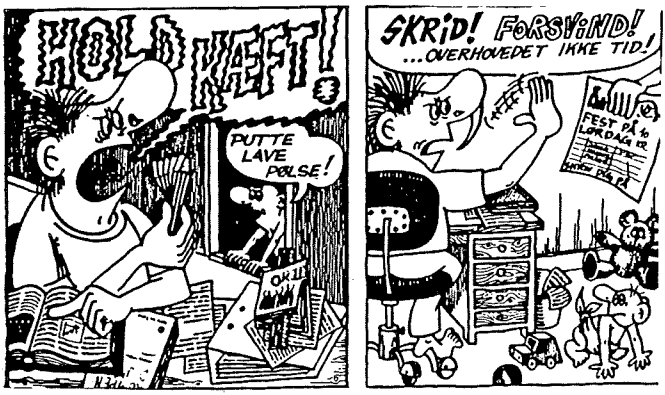
CASE 5

[11] Tintner, G., and Sengupta, J.: Stochastic Economics, Academic Press, 1972.

[12] Intriligator, M.: Mathematical optimization and Economic theory, Prentice-Hall, 1971.

[13] Mirakhor, A.: Application of the Optimal Control Theory to Economic Analysis, Eco. Comp. and Eco. Cybernetics Studies and Research, N^{O} 3, 1973.

[14] Kendrich, D., and Taylor, D.: Numerical methods and nonlinear optimizing models for Economic Planning, in Studies in Development Planning, Cambridge Massachusetts, 1971, Harvard University Press.



PART THREE

STOCHASTIC OPTIMIZATION

CHAPTER 9

STOCHASTIC OPTIMIZATION:

An introduction


-VI MÅ STRYGE DEN TUR TIL RUC!
JEG HAR SIMPELTHEN IKKE TID!
-ÆRGERLIGT, DET KUNNE NU HÅVE VÆRET INTERESSANT, AT SE DEN POLITISKE DICIPLINERING!

1. INTRODUCTION

If some or all of the parameters of our deterministic model (both static and dynamic) are assumed to be stochastic, then our decision problem becomes much more complicated from both a theoretical and practical viewpoint. Depending on the degree of "stochasticity" of our problem, on the information available, and so on, one has to select a strategy to be able to handle the problem by some of the methods we have developed to solve deterministic problems. This methodological aspect of the problem will be briefly discussed in the next section.

Stochastic static optimization is usually known as Stochastic Programming or more precisely as Stochastic Linear Programming to emphasize the fact that most of the models and methods discuss the case of a LP-model where some or all the parameters are stochastic. Sec. 3 will be a rough introduction to the subject to emphasize the need to define new criteria for optimality and feasibility. An up to date discussion on this subject can be found in [1].

The last section will be a very short introduction to the, from the mathematical viewpoint, rather cumbersome subject of stochastic control. Emphasis is placed on the different practical ways to find a solution to the stochastic control problem and some important results are briefly mentioned.

This chapter is an introduction to the subject, where most of the results are presented in an unformal way with the purpose to show that some of the methods we have developed for the deterministic case can also, under suitable assumptions, be employed to deal with such complicated problems. Further study can be followed by consulting the references.

2. STRATEGIES TO DEAL WITH STOCHASTIC OPTIMIZATION PROBLEMS

Stochastic optimization problems are those decision problems where some of the parameters in the objective function or/and the restrictions are stochastic. In stochastic optimization we assume that we know the probability distribution or the frequency function of the stochastic parameters. Employing a terminology from Decision Theory we could say that stochastic optimization deals with decision problems under risk. Depending on the "importance" of the stochastic parameters upon the results of the decision problem, we can utilize several different strategies to deal with stochastic optimization problems. Let us discuss them:

a) Replace all the stochastic variables by their mean value, we have then a deterministic optimization problem. This approach can be utilized when the variations of the stochastic parameters are so small that they will not have a major influence upon the optimal decision or control variables. In the case studies we have seen in the previous chapters this strategy has been implicitly employed.

b) Utilizing a) a first estimate of the decision variables is performed. Later on these values are suitable modified by analyzing the stochastic properties of the problem. This approach can be utilized when we expect that only minor changes on the first estimate values will occur.

c) Due to the stochastic parameters the objective function becomes stochastic and the feasible set becomes also stochastic. Here a new (deterministic) payoff function is defined, for example minimization of the expected value of the objective function, and a new definition of feasibility is required, for instance the probability that a constraint is not satisfied is a small number. Thus in a single model we take account of the structural and stochastic properties of the system in study, these models can often be solved

utilizing some of the methods discussed in earlier chapters. This approach, in the static case, is usually known as STOCHASTIC PROGRAMMING and, in the dynamic case, is denominated STOCHASTIC CONTROL. This strategy is suitable for those cases where the maximal losses due to unsounded decisions are not too large, otherwise we have to employ the next strategy.

d) In this strategy the study is centered around the stochastic properties of the problem. Thus if the variations of the stochastic parameters are such that great losses can occur it is usually impossible to find a single objective function. The search for "optimal" decisions under such situations is a creative, not an analytical, effort.

Further discussion can be found in [2] and [3].

3. STOCHASTIC PROGRAMMING

Let us consider the following Linear Programming (LP) model

$$\max_{x} z = cx$$

subject to:

$$Ax \leq b$$

$$x \geq 0 \ .$$

In Stochastic Programming we consider this LP model, where some or all the parameters (that is c,A,b) are stochastic. Under such conditions z is also stochastic, and a new objective function has to be defined. Moreover the feasible set is no longer constant and a new criterion for feasibility has to be defined.

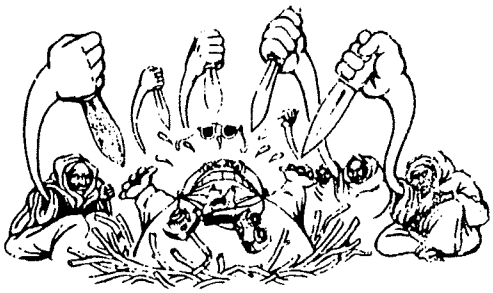
Stochastic Programming models are usually classified in two categories:

a) "Wait and see" approach, where the decision is first made after the stochastic parameters have been observed. This is not a decision problem, here one tries to study the effects of the stochastic parameters upon z assuming that an optimal policy will later on be selected.

b) "Here and now" approach, where the decision is made before the stochastic parameters have been observed. Here the stochastic model is transformed to an"equivalent deterministic model", with the same optimal decision.

We will center our discussion on the last type of models, for a discussion of "wait and see" models see [1].

"Wait and see"



"Here and now"!

3.1 The objective function

In the table below is shown some of the best known objective functions employed in Stochastic Programming.

Criterion	Definition	Deterministic equivalent model if only c is stochastic
E-model (Expected value)	$\max E[z]$	$\max E[z] = \sum_{j=1}^{n} E[c_j]x_j$ $Ax \leq b$ $x \geq 0$ (LP-problem)
V-model (Variance)	$\min V[z]$ $E[z] \geq E_0$ E_0 given	$\min V[z] = x'Cx$ $\sum_{j=1}^{n} E[c_j]x_j \geq E_0$ $Ax \leq b$ $x \geq 0$ (Convex quadratic problem)
EV-model	$\max\{E[z]-\lambda V[z]\}$ $\lambda \geq 0$ (risk-aversion coefficient)	$\max\{E[c] \cdot x - \lambda x' Cx\}$ $Ax \leq b$ $x \geq 0$ (Concave quadratic problem)
F_α-model (α-fractil)	$\max f$ $P\{z \leq f\} = \alpha$ $0 < \alpha < 0{,}5$	If: $z \in N(E[z],\tau[z])$ $\max\{E[c]x - \tau_\alpha\sqrt{x'Cx}\}$ $Ax \leq b$ $x \geq 0$ where: $\tau_\alpha = -\Phi^{-1}(\alpha)$ being Φ the cumulated distribution function for the standard N(0,1). (Concave programming problem)
P_k-model (Safety-first principle)	$\max p = P\{z>k\}$ $= \frac{E[z]-k}{\tau(z)}$ k given	If: $z \in N(E[c],\tau[z])$ $\max\left[\frac{E[c]x-k}{\sqrt{x'Cx}}\right]$ $Ax \leq b$ $x \geq 0$ (Pseudo-concave programming problem when $k < E[c]x$)

The last table also contains the deterministic equivalent model for the case of a constant feasible set and the stochastic parameters c are independent of x. The E-, V- and EV-models are easy to stipulate, while for the F_α- and P_k-models a further normality assumption on the objective function is required to be able to find an analytical expression of the deterministic equivalent model. While in the E-model we are indifferent to risk in the V- and EV-models we measure risk by calculating the variance of the objective function. The F_α-model compares the different α-fractiles and chooses the largest, while the P_k-model, also controls the payoff at the left tail of the cumulate distribution function, compares the probability for the payoff being greater than a given value k and chooses the largest.

Most of above mentioned criteria have been developed in connection with real-life problems and they have been used mainly due to the fact that the deterministic equivalent can be solved employing already known methods. A very important theoretical question is to what extent these criteria are "rational" in the sense that they satisfy the conditions stipulated by Utility Theory for rational behaviour. Actually it is easy to find simple examples that show that in general most of these criteria are not "rational". It is possible to develop other criteria that are rational in this sense, but another problem arises, that is: the deterministic equivalent cannot be analytically found or it is so complex that it is extremely difficult to solve it. Due to this fact, in practice, it is usually recommended to utilize several of these criteria to find several solutions that can be compared before a final choice is found. Here simulation plays a central role to evaluate the behaviour of the different models.

3.2 The feasible set

If the parameters of some of the constraints of the standard LP-model become stochastic we have to redefine our problem. In general there are two ways to take account of the possibility

of not being able to satisfy a constraint due to stochasticity. The first one is to assure that the probability that a restriction is satisfied is near one. That is to demand that

$$P\left\{ \sum_{j=1}^{n} a_{ij}\, x_j \leq b_j \right\} \geq \alpha_i \quad , \quad \alpha_i > 0{,}5 \ .$$

In other words infeasibility is permitted to happen very seldom.

The second way to take account of stochastic feasible sets is to introduce a penalty that has to be added to the objective function and it represents the expected costs incurred due to the fact that unfeasible solutions might be chosen. Let us further concretize these ideas by describing briefly some of the best known models.

3.3 Chance constrained programming

The simplest version of this model is the case when only b is stochastic and the b_i's are stochastic independent, the problem is then

$$\max\ cx$$

$$P\left\{ \sum_{j=1}^{n} a_{ij}\, x_j \leq b_i \right\} \geq \alpha_i \quad , \quad i = 1,\cdots,m$$

the deterministic equivalent is easily found as being the following LP-problem

$$\max\ cx$$

$$\sum_{j=1}^{n} a_{ij}\, x_j \leq F_{b_i}^{-1}(1-\alpha_i) \quad , \quad i = 1,\cdots,m$$

where $F_{b_i}(\ )$ is the cumulate distribution function of b_i. Assume now that the a_{ij} are stochastic and independent to each other, to find a deterministic equivalent we have to as-

sume that $a_{ij} \in N(m_{ij}, s_{ij}^2)$ and if the b_i are deterministic, it is easily shown that the deterministic equivalent is the following nonlinear program:

$$\max cx$$

$$\sum_{j=1}^{n} m_{ij} x_j + \tau_N(\alpha_i) \sqrt{\sum_{j=1}^{n} s_{ij}^2 x_j^2} \leq b_i \quad , \quad i = 1, \cdots, n$$

where $\tau_N(\alpha) = \Phi^{-1}(\alpha) > 0$ for $\alpha > 0{,}5$. This is a more complex problem and numerical methods can be employed to find a solution.

In this model the case of b_i being stochastic and independent of the a_{ij} can be easily added, thus

$$\sum_{j=1}^{n} a_{ij} x_j \leq b_i \Leftrightarrow \sum_{j=1}^{n+1} a_{ij} x_j \leq 0$$

where

$$a_{n+1} = -b \quad \text{and} \quad x_{n+1} = 1 \ .$$

For the case of c being also stochastic, we have then to modify our objective function and rather complex deterministic equivalent can be found under normality assumptions.

Chance Constrained Programming is a rather popular method due to its simplicity, but recently it has been shown that this approach has serious theoretical drawbacks in what concerns "rationality". The problem resides in the fact that the model says nothing about the consequences of infeasibility that occurs with probability $(1-\alpha)$. The next two approaches avoid this disadvantage.

3.4 Two-stage programming

Let us consider the following LP-problem.

$$\min z = cx$$

$$Ax = b$$
$$x \geq 0$$

Assume that c is deterministic, and b and/or A are stochastic, then we will have some variations in the vector $(b-Ax)$ that will be described by

$$My = b - Ax \quad , \quad y \geq 0$$

where y is a vector $(n_1 \times 1)$ and M a matrix $(m \times n_1)$ with deterministic elements.

The model in this approach becomes

$$\min_x \left[cx + E\left\{ \min_y qy \right\} \right]$$

$$Ax + My = b$$
$$x, y \geq 0$$

where q is a penalty that has to pay when infeasibility occurs (here a linear penalty function is assumed, other functions could be employed). Here for the objective function we are utilizing an E-model, other models could be used but the solution procedures become more complicated. This model is called Two-stage Programming because first we have to find x "here and now", then observe b and/or A and at a second stage we have to find y "wait and see".

If we define

$$Q(x,b,A) = \min qy$$

$$My = b - Ax$$
$$y \geq 0$$

and $Q(x) = E_{b,A}[Q(x,b,A)]$.

The deterministic equivalent is then

$$\min_{x \geq 0}[cx + Q(x)] \ .$$

It is only for some simple cases that an analytical expression to $Q(x)$ can be found, but in most computational methods one has to estimate $\nabla Q(x)$ at a given point, this can be easily done for the case that the "wait and see" problem is a LP model.

3.5 Fixed-charge penalty

Here a constant penalty T_i is added to the expected cost if restriction "i" is not satisfied. The model is for c and b stochastic

$$\min\left[E[c]x + \sum_{i=1}^{n} T_i(1 - F_i(t_i))\right]$$

$$Ax - t = 0$$
$$x \geq 0$$

where

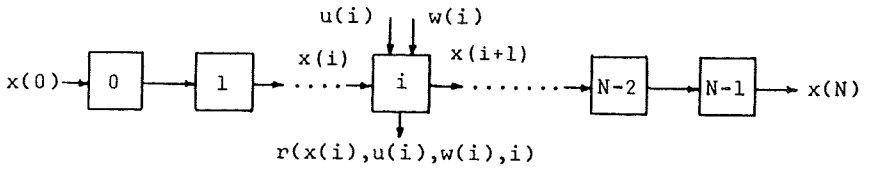
$$1 - F_i(t_i) = P\{t_i < b_i\}$$

is the probability of having $(\Sigma\, a_{ij}\, x_i < b_i)$, that is the demand condition $Ax \geq b$ is not satisfied (notice that our problem here is $\min\{cx \mid Ax \geq b \ , \ x \geq 0\}$). This is a non-linear problem and in general $F_i(t_i)$ is not concave. Concavity for the case of the b following a normal distribution can be assured if we demand that $t_i \geq Eb_i$. It has been shown that this model is theoretically better founded than the Chance Constrained Programming in what concerns rationality, moreover this model is less sensitive to errors in the estimates of T_i

than Chance Constrained Programming is to errors in the estimates of α_i, see further [1].

4. STOCHASTIC CONTROL

Let us consider the graphical representation of a discrete-time control problem where at each stage a stochastic (k×1) vector $w(i)$, $i = 0,1,\cdots,N-1$, is considered as illustrated below



Our discrete-time stochastic control problem becomes then

$$\max_{u(i)} J = E_{w(i)}\left[\sum_{i=0}^{N-1} \{r(x(i),u(i),w(i),i)\}\right]$$

subject to

$$x(i+1) = f(x(i),u(i),w(i),i) \quad , \quad i = 0,1,\cdots,N-1$$

$$x(0) = x_0$$

where $E[\ ]$ is the expected value operator and $w(i)$ are either continuous or discrete stochastic variables with known probability distributions.

Under varying assumptions concerning the information available to the controller, different optimal control policies result. The definition of admissible controls in a stochastic system

relies heavily on the amount of available information. Closed-loop controls were defined as those where the information available up to the present time is used in order to take immediate decisions, while open-loop controls are independent of the available information. Obviously open-loop controls are on average worse than closed loop controls, with exception of the deterministic case where both modes of control are equivalent.

The following have been suggested as alternative modes of control and use of information [4]:

a) Closed Loop Optimal (J_a), controls are computed as feedback laws via dynamic programming, or equivalents. In computing them, all information available of the time of exercising the control is assumed to be employed as a conditioning argument for the control laws. Here the optimal control is a stochastic function, with exception of the first period that is a constant.

b) Open Loop - No updating (J_b), controls are conditioned only on initial information. Even if further information becomes available, these initially computed controls are enacted up to the end of the time horizon. All the optimal controls are constants.

c) Open Loop-Updating (J_c), same as (b) except that controls are recomputed as new information becomes available. Only the first-period decision of b) are used.

d) Mean value approximations (J_d), in this approach all stochastic variables are set equal to their expected value. Then the closed-loop optimal control law is computed for the resulting deterministic. This procedure may be used in either an up-dating or no-updating version.

e) Best within a class, here one uses qualitative information about the form of the optimal law to deduce an interesting

and easily parameterized class of control laws. One then solves for the best control law within the specified class.

It is obvious that the correct way to solve stochastic control problems is by employing a closed loop optimal law, but in practice it is usually extremely difficult or impossible to find this optimal control and in those cases where it can be found it is not sure that such control can be implemented. This justifies the existence of the other modes of control. A relation between the performance functions may be shown to be

$$J_a \geq J_c \geq J_b \geq J_d$$

assuming no up-dating in d).

In this section we assume that maximization of the expected value of the objective function is a relevant approach to the concrete problem otherwise alternative criteria can be suggested as it was the case for the static problem. Finally, it is important to emphasize that most of the results in this section are also valid for continuous-time stochastic control problems.

4.1 Functional equations

The principle of optimality will also permit us to establish a decomposition principle for the discrete-time stochastic control problem that will permit us to evaluate analytically or numerically closed-loop controls. Let us assume that the vectors $w(i)$, $i = 0,1,\cdots,N-1$, are stochastic independent. If $F(k,x)$ denotes now the total expected maximum return from x , a given value, to the last stage (N-1), then

$$F[k,x] = \max_{\{u(k)\}} \; E_{\{w(k)\}} \left[\sum_{i=k}^{N-1} r(x(i,u(i),w(i),i) \right]$$

$$= \max_{u(k)} \max_{\{u(k+1)\}} E_{w(k)} \left[E_{\{w(k+1)\}} \left[r(x,u(k),w(k),k) + \sum_{i=k+1}^{N-1} r(x(i),u(i),w(i),i) \right] \right]$$

$$= \max_{u(k)} E_{w(k)} \left[r(x,u(k),w(k),k) + \max_{\{u(k+1)\}} E_{\{w(k+1)\}} \left[\sum_{i=k+1}^{N-1} r(\ ) \right] \right]$$

$$= \max_{u(k)} E_{w(k)} \left[r(x,u(k),w(k),k) + F(k+1,f(x,u(k),w(k),k)) \right] .$$

We have then decomposed our $[k,x]$ problem into two problems as it was the case for the deterministic problem, and a terminal boundary condition is

$$F(N-1,x) = \min_{u(N-1)} E_{w(N-1)} [r(x,u(N-1),w(N-1),N-1)]$$

and the optimal performance function is

$$J_a = F(0,x_0)$$

and the optimal control is found in a closed loop form $u^*(x,k)$ for each stage. These are stochastic functions, because $x(k)$ is stochastic except for $k = 0$, where $x(0)$ is given.

A computational procedure analogous to the one described for the deterministic case can be applied, that is at each stage the minimization can be performed by quantization or by employing gradient methods [5]. That is the existence of the stochastic vectors gives in principle no additional difficulties as far as the $w(k)$ at different stages are uncorrelated. This assumption can always be relaxed at the expense of defining additional state variables to account for the correlation, thus the recursion equation becomes

$$F(k,x,w(0),\cdots,w(k-1)) = \max_{u(k)} \; E_{w(k)/w(0),\cdots,w(k-1)} [r(x,u(k),w(k)) + F(k^1,f(x,u(k),w(k),k),w(0),\cdots,w(k))]$$

where $F(\ )$ is the expected value of the objective function for problem $[k,x]$ given the available information on $w(k)$, $k = 0,\cdots,k-1$, and E is a conditional expectation operator. This approach is highly restricted by the course of dimensionality problem. Under some special conditions an analytical solution can be found to this rather complicated expression, this will be seen in the next section.

4.2 Certainty equivalence or separation principle

Let us consider an unrestricted discrete-time stochastic control problem with a concave quadratic objective function whose expected value has to be maximized and with a linear equation of motion.

It is not difficult to show that under these conditions $u^*(0)$, the first-period closed loop optimal control, is identical to $u^{**}(0)$ the corresponding solution to the deterministic model that is obtained by replacing all the stochastic variables by their mean value. In general this principle asserts that the optimal value $u^*(k)$ is determined by

$$\max\left(\sum_{i=k}^{N-1} r(x(i),u(i),\bar{w}(i),i)\right)$$

subject to:

$$x(i+1) = f(x(i),u(i),\bar{w}(i),i) \quad , \quad i = k,k+1,\cdots,N-1$$

given: $x(k) = x_k$

where $\bar{w}(i)$ is the conditional expectation of $\bar{w}(i)$, $i = k,\cdots,N-1$, on $w(0),\cdots,w(k-1)$, the information available at stage k. Of course, the values of $u(0),\cdots,u(k-1)$ have

already been determined by time k. The values obtained for $u^{**}(k+1),\cdots,u^{**}(N-1)$ are to be regarded as estimated future decisions at stage k, which will be revised as more information is gathered, and the time for actual decision approaches.

In other words, to get the optimal decision rule at stage k, we can split the procedure into two successive parts: the first part, estimation (or identification), consists of estimating the mean value of the stochastic variables given the available information; the second one, control, consists of solving the deterministic control problem starting from the best estimate at the present time. Equivalent results can be found for the continuous-time stochastic control problem [6].

This principle is known under two different names: the certainty equivalence principle for operations research workers [7] and the separation principle for control theorists [8]

Applications of this important principle to production planning, macro-economic planning and control problems can be found in [7], [9].

4.3 Stochastic maximum principle

If the stochastic dynamic system can be described by stochastic differential equations, then under suitable assumptions a maximum principle for this case can also be formulated, although the mathematical background to deal with such problem becomes much more sophisticated [6]. Intuitively speaking, the necessary conditions state that the optimal stochastic control vector $u^*(t)$ must maximize the conditional expectation of the Hamiltonian given the information available at t over the class of admissible controls. This is so because $p(t)$ is a vector of stochastic variables and the canonical equations are also stochastic differential equations.

5. REFERENCES

[1] Jensson, P.: Stokastisk Programmering. Licentiatafhandling ved IMSOR, 1975.

[2] Tintner, G., and Sengupta, J.K.: Stochastic Economics, Academic Press, 1972.

[3] Hanssmann, F.: Operations Research Techniques for Capital Investment, J. Wiley, 1968.

[4] Kleindorfer, P.: Stochastic Optimization of Dynamic Planning Models, TIMS XX International Meeting in Tel Aviv, Israel, 1973.

[5] Nemhauser, G.: Introduction to Dynamic Programming, John Wiley, 1966.

[6] Bensoussan, A., et al.: Management Applications of Modern Control Theory, North-Holland, 1974.

[7] Valqui Vidal, R.V.: Operations Research in Production Planning, IMSOR, 1970.

[8] Tou, J.: Modern Control Theory, McGraw-Hill, 1964.

[9] Theil, H.: Optimal decision rules for government and industry, North-Holland, 1964.

EPILOGUE

In the last chapters we have been dealing with a mathematical discipline known as optimization in what concerns theory, computational aspects and applications.

It is not possible for me to conclude about this subject in the same way as G.H. Hardy does about number theory and his own work in this field:

> "I have never done anything 'useful'. No discovery of mine has made, or is likely to make, directly or indirectly, for good or ill, the least difference to the amenity of the world. I have helped to train other mathematicians, but mathematicians of the same kind as myself, and their work has been, so far at any rate as I have helped them to it, as useless as my own. Judged by all practical standards, the value of my mathematical life is nul; and outside mathematics it is trivial anyhow. I have just one chance of escaping a verdict of complete triviality, that I may be judged to have created something worth creating. And that I have created something is undeniable: the question is about its value".
>
> (A mathematicians's apology, 1969)

Until about 1940, three main methods of optimization existed, that is:

(i) Use of algebraic, geometric and trigonometric inequalities

(ii) Use of differential calculus and Lagrange's method of undetermined multipliers, and

(iii) Use of the calculus of variations

with applications to physics and engineering at that time.

1. Use of geometric and algebraic inequalities: Duality

Problems of optimization arose very early in the history of mathematics. Some typical problems were:

(i) What is the shortest path between any two points in a plane or in space?

(ii) What is the shortest path between any two points on the surface of a sphere (say on the Earth).

(iii) Of all rectangles with a given perimeter, which has the maximum area?

(iv) Of all closed curves with a given perimeter, which encloses the maximum area?

The problems were essentially geometrical in nature and their solutions depended on the use of geometrical and algebraic inequalities. Thus the solution of the first problem could depend on the geometrical theorem that the sum of two sides of a triangle is greater than the third.

The solution of the third problem could be made to depend on the well-known algebraic inequality that for any n given positive numbers,

$$\text{Arithmetic Mean} \geq \text{Geometric Mean} \geq \text{Harmonic Mean}$$

For n=2, it gives

$$\frac{x+y}{2} \geq \sqrt{(xy)} \tag{1}$$

If x and y are the lengths of the sides of a rectangle with perimeter 1, then using (1),

$$\sqrt{(xy)} \leq \frac{1}{4} \text{ or } xy \leq \frac{1^2}{16} \tag{2}$$

but xy is the area of the rectangle. Thus the area of any rectangle with perimeter l, is less than or equal to $l^2/16$ but the area actually becomes equal to this value when x=y=l/4. Thus of all rectangles with given perimeter l, the square of side l/4 has the maximum area.

If on the other hand, we consider rectangles with given area A, the above inequality gives

$$2(x+y) \geq 4\sqrt{A} \qquad (3)$$

so that the perimeter is minimum when $x=y=\sqrt{A}$. Thus of all rectangles with given area, the square has the least perimeter.

The above two problems are related to one another. In the first perimeter was fixed and the area had to be maximized; in the second, the area was fixed and the perimeter had to be minimized. Such optimization problems very often occur in pairs each of which is called the dual of the other.

In fact the above problems are constrained optimization problems. We have to maximize or minimize an objective function subject to some restrictions on the values by one or more constraint functions. In the first case the objective function corresponds to area and the constraint function corresponds to perimeter and the objective function has to be maximized. In the second case, the objective function corresponds to perimeter and the constraint function corresponds to area and the objective function has to be minimized. Each problem is the dual of the other.

The fourth problem has a story. A king was pleased by a subject and ordered that he could run round a closed path and whatever land was enclosed by the path would be his. If the maximum distance he could run in a day was l, the problem was to find the closed curve of length l which enclosed the maximum area. It was proved that the curve was a circle of radius $l/2\pi$. The dual problem would be to find the curve of minimum perimeter enclosing a given area A. The curve would be a circle of radius $\sqrt{(A/\pi)}$.

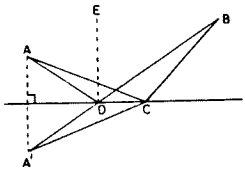


Figure 1

Another interesting problem was posed early in the history of mathematics. A man has to go from A to B after collecting water from a river (see figure 1). At what point should he collect the water so that the distance travelled by him is the shortest? If he collects the water at point C and A' is the mirror image of A in the straight river bank, the distance travelled by him is

$$\begin{aligned} AC + CB &= A'C + CB \\ &\geq AB \\ &= AD + DB \end{aligned} \tag{4}$$

Thus the distance travelled will be least when the water is collected at D where D is the point of intersection of the straight river bank with the line BA'. It is easily seen that if DE is the normal to the river bank at D, then

$$\angle ADE = \angle BDE \tag{5}$$

When later Fermat enunciated his principle in optics that light travels from one point to another in such a way that the time taken by it is least, the above solution gave the proof of the fact that angle of incidence is equal to angle of reflection.

The above methods of optimization are based on inequalities. To prove that the maximum value of $f(x_1,x_2,\ldots,x_n)$ is A, we prove (i) $f(x_1,x_2,\ldots,x_n) \leq A$, and (ii), for some values of $x_1,x_2,\ldots,x_n$,

the function actually becomes equal to A. Similarly to prove that the minimum value of the function is B, we prove that (i) $f(x_1,x_2,\ldots,x_n) \geq B$, and (ii), for some values of x, the function actually becomes equal to B.

2. Need for calculus: Lagrage's method of undetermined multiples

From Fermat's principle of least time and the fact that in a homogeneous medium, the velocity of light is constant, it follows that in a homogeneous medium, light travels in a straight line. If light travels from a point A in one medium (a vacuum) to a point B (see figure 2) in another medium (with refractive index μ), then the time taken in travelling form A to B is

$$\frac{AC}{c} + \frac{CB}{c/\mu} = \sqrt{(h_1^2+x^2)}/c+\mu\sqrt{[h_2^2+(k-x)^2]}/c \qquad (6)$$

where c is the velocity of light in a vacuum. We have to choose x so that this time is the minimum.

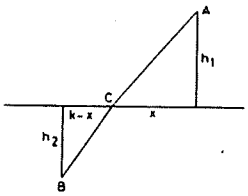


Figure 2

Problems like the above require a new technique of optimization, viz. differential calculus invented by Newton and Liebnitz. This can provide the maximum or minimum of any function f(x) in an interval provided it is continuous and differentiable a sufficient number of times. We put $f'(x)=0$ and solve for x. At each

of the points obtained, we find f"(x). If f"(x) is negative, we get a maximum and if f"(x) is positive, we get a minimum. We find the value of f(x) at all points where we get a maximum and also the value of f(x) at the ends of the interval; one of the values will be the maximum of the function in the interval.

The techniques can easily be extended to find the maximum or minimum of a function of several variables, viz. $f(x_1,x_2,\ldots,x_n)$. It was extended by Lagrange to optimization of $f(x_1,x_2,\ldots,x_n)$ subject to constraints

$$g_i(x_1,x_2,\ldots,x_n)=0, \quad (i=1,2,\ldots,m) \tag{7}$$

He essentially optimizes

$$f(x_1,x_2,\ldots,x_n)+\sum_{i=1}^{m}\lambda_i g_i(x_1,x_2,\ldots,x_n) \tag{8}$$

where λ's are called multipliers. He sets

$$\frac{\partial f}{\partial x_j}+\sum_{i=1}^{m}\lambda_i\frac{\partial g_i}{\partial x_j}=0, \quad (j=1,2,\ldots,n) \tag{9}$$

and uses (7) and (9) to solve for $x_1,x_2,\ldots,x_n$; $\lambda_1,\lambda_2,\ldots,\lambda_m$.

Thus, suppose we have to find the maximum volume of a rectangular parallelepiped, the sum of whose edges has a vixed value 1. We maximize

$$xyz+\lambda[x+y+z-1/4]$$

Differentiating we get

$$yz+\lambda=0, \quad zx+\lambda=0, \quad xy+\lambda=0$$

giving

$$x=y=z=1/12$$

and the maximum volume is $(1/12)^3$ and is obtained when the rectangular parallelepiped is a cube. We have of course also to establish that the optimum value here gives in fact the maximum volume. (Chapter 2 is a complete discussion of this subject).

3. Development of the calculus of variations

The object here is to optimize an integral

$$\int_a^b F(x,y,\frac{dy}{dx})dx \tag{10}$$

where F is a known function. If we know y as a function of x, we can find the value of the integral. Our object is to find that function which optimizes the integral. We consider some typical examples:

(i) To find the shortest distance between two points in a plane, we have to minimize

$$\int_{x1}^{x2} \sqrt{[1+(\frac{dy}{dx})^2]}dx \tag{11}$$

(ii) To find the shortest distance between two points on the sphere $x^2+y^2+z^2=a^2$, we take

$$x=a \sin\theta \cos\phi,\ y=a \sin\theta \sin\phi,\ z=a \cos\theta \tag{12}$$

and minimize

$$= \int_{\theta_1}^{\theta_2} \sqrt{[a^2+a^2\sin^2\theta(\frac{d\phi}{d\theta})^2]}d\theta \tag{13}$$

(iii) To find the shortest distance between two points in three-dimensional space, we have to find functions $x(t)$, $y(t)$, $z(t)$ such that

$$\int_{t1}^{t2} \sqrt{[(\frac{dx}{dt})^2 + (\frac{dy}{dt})^2 + (\frac{dz}{dt})^2}]dt$$

is the minimum.

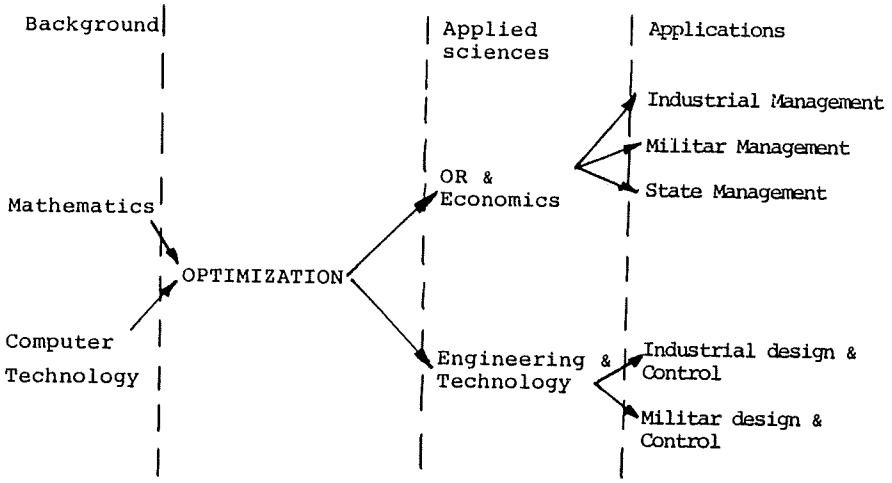
Chapter 6 discusses these and some more general problems of this type.

4. Development after World War II

After World War II a new class of optimization techniques became available and paper and books have been published at a great pace. What made this possible? Gottfried and Weisman write:

> "Two factors made this possible - the development of high-speed electronic digital computers, and the application of mathematical analysis to the development of numerical techniques for obtaining maxima or minima. These numerical procedures (which are largely the basic tools in the field of operations research) bypass many of the difficulties associated with the calculus. Moreover, the modern digital computer, with its large memory and its extremely rapid calculational ability, enables the practical utilization of these numerical techniques in a reasonable amount of time and at a tolerable expense".
>
> (Introduction to Optimization Theory, 1973)

This is not the whole truth. To understand this development we have to see also at the applications. The diagram below shows that optimization is usually a tool used by applied sciences (OR, Economics, Engineering, etc.). It is the huge development of these sciences together with mathematics and computer technology that explains the growth of optimization.



What made possible the huge development of these applied sciences? This is a long story that is closely related to the industrial, militar and state development of the high industrialized societes both at East and West.

Are you interested in this story? then read

(i) Systemvidenskaberne og den industrielle produktion, R.V.V. Vidal, IMSOR, 1980,

(ii) En introduktion til offentlig planlægning, Lyngvig & Vidal, IMSOR, 1980.

APPENDIX

a) Functions

A function of n variables is symbolized by $f(x)$, where $x = (x_1, x_2, \cdots, x_n)'$.

A real valued function $f(x)$ is said to be upper semicontinuous at x_0 if, given $\varepsilon > 0$, there is a $\delta > 0$ such that

$$f(x) - f(x_0) < \varepsilon \quad \text{for} \quad |x - x_0| < \delta$$

A function $f(x)$ is said to be lower semicontinuous at x_0 if $-f(x)$ is upper semicontinuous af x_0. If $f(x)$ is both upper and lower semicontinuous, it is continuous.

A function $f(x)$ is differentiable at the point x_0 if there exists n numbers $a = (a_1, a_2, \cdots, a_n)$ for which

$$\lim_{|h| \to 0} \frac{\left| f(x_0 + h) - f(x_0) - \sum_{j=1}^{n} a_j h_j \right|}{|h|} = 0$$

where $h = (h_1, \cdots, h_n)'$ is any point in R^n and $|h|$ is the norm of h. The a_j's are partial derivatives, where

$$a_j = \frac{\partial f(x_0)}{\partial x_j} \quad , \quad j = 1, \cdots, n$$

A function of $f(x)$ is unimodal with a minimum at x^*, if and only if

$$x_3 < x_4 < x^* \Rightarrow f(x_3) > f(x_4) > f(x^*)$$

$$x^* < x_4 < x_3 \Rightarrow f(x^*) < f(x_4) < f(x_3)$$

Note that a unimodal function needs not be continuous.

The first partial derivatives of a function $f(x)$ at a certain point, if they exist, define the slopes of the tangent to the function with respect to the n-coordinate axes. A useful notation for summarizing these first partial derivatives is to use the so-called gradient vector which is given by

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \cdots, \frac{\partial f(x)}{\partial x_n} \right)$$

The n^2 second partial derivatives of $f(x)$ can be considered to be the elements of the nxn matrix

$$H(x) = \left\| \frac{\partial^2 f(x)}{\partial x_i \, \partial x_j} \right\|$$

$H(x)$ is denominated the Hessian matrix of $f(x)$. $H(x)$ is a symmetric matrix.

Assume that the second partial derivatives of $f(x)$ exist and are continuous. Let $h = (h_1, h_2, \cdots, h_n)'$, Taylor's expansion around x_0 is given by

$$f(x_0 + h) = f(x_0) + \nabla f(x_0)h + \tfrac{1}{2}h'H(x)h + \cdots$$

b) Quadratic forms

To determine whether a function has a local optimum at a stationary point we must examine the behaviour of the term containing the Hessian matrix in a Taylor series expansion, in a neighbourhood of its stationary point. This term $(h'H(x)h)$ is an example of a quadratic form.

A quadratic form in n variables $x_1, x_2, \cdots, x_n$ is a numerical function of these variables which can be written:

$$Q(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} \, x_i \, x_j = x'Qx$$

where

$$x = (x_1, x_2, \cdots, x_n)' \quad \text{and}$$

$$Q = \| b_{ij} \|$$

It will be noted that when $i \neq j$, the coefficient of $x_i x_j$ is $(b_{ij} + b_{ji})$, thus Q can always be assumed to be a symmetric matrix, because, if it is not, we can uniquely define new coefficients

$$a_{ij} = a_{ji} = \frac{b_{ij} + b_{ji}}{2}, \quad \forall(i,j)$$

then we can always assume that a matrix associated with a quadratic form is symmetric.

Eigenvalues and eigenvectors

Given a (nxn) matrix A, any non-zero vector x satisfying

$$Ax = \lambda x$$

is called an eigenvector (proper or characteristic vector). The scalar λ is called the eigenvalue (proper or characteristic value). Since $x \neq 0$, λ is given as roots of the characteristic equation:

$$\det(A - \lambda I) = 0$$

We have the well-known results:

- Eigenvalues and eigenvectors of a real symmetric matrix are all real.

- Eigenvectors of a real symmetric matrix corresponding to distinct eigenvalues are orthogonal to each other.

- Given an arbitrary real quadratic form $Q(x)$, where Q is symmetric, there exists an orthogonal matrix T, that is $T' = T^{-1}$, such that

$$T'QT = \text{diagonal matrix } \Lambda$$

$$= \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$$

We say then that Q is diagonalized by the orthogonal matrix T. Some of the λ's may be zero.

Definiteness

A quadratic form $Q(x)$ (or the matrix Q) is called positive definite if $Q(x) > 0$ for all $x \neq 0$. $Q(x)$ (or Q) is called positive semidefinite if $Q(x) \geq 0$ for every x and there exists $x \neq 0$ for which $Q(x) = 0$. A form $Q(x)$ is called negative definite (semidefinite) if $-Q(x)$ is positive definite (semidefinite). $Q(x)$ is said to be indefinite if it is positive for some points x and negative for others.

We have the well-known results:

- $Q(x)$ is positive (negative) definite if and only if all the eigen values of Q are positive (negative). Now, if each λ_j is non-negative (non-positive) and at least one is zero, then $Q(x)$ is positive (negative) semidefinite.

- Given a positive definite matrix Q, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be its eigenvalues. Then

$$\lambda_1 = \max_{x \neq 0} \frac{x'Qx}{x'x} \quad \text{and} \quad \lambda_n = \min_{x \neq 0} \frac{x'Qx}{x'x}$$

This states therefore that the maximization (or minimization) of a positive definite quadratic form $Q(x)$ is equivalent to finding the largest (or smallest) eigenvalue of the matrix Q.

- A symmetric matrix Q is positive definite if and only if every one of the quantities

$$b_{11}\ ,\quad \begin{vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{vmatrix}\ ,\quad \begin{vmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{vmatrix}\ ,\ \cdots\ ,\ \det(Q)$$

is positive (Sylvester's criterion).

Let us consider the quadratic form

$$Q(x) = bx + \tfrac{1}{2}x'Qx$$

where $b = (b_1,\ldots,b_n)'$ and Q is symmetric and positive definite. The n directions given by the vectors $(s^1,\ldots,s^n)$, all of them different of 0, are Q-conjugate, if

$$i \neq j \Rightarrow (s^i)'Qs^j = 0\ ,\ \text{where}\quad i = 1,\cdots,n \quad j = 1,\cdots,n$$

c) Convex sets

A subset $X \subseteq R^n$, is convex if and only if given two points x and y in X, then:

$$\alpha x + (1-\alpha)\, y \in X\ ,\ \forall\alpha\ ,\ 0 \leq \alpha \leq 1$$

Geometrically, a set is convex if and only if two points in the set, all points on the line segment connecting these points also lie in the set. Examples of convex sets are:

A hyperplane in R^n, defined as $\left\{ x \in R^n \mid \sum_{j=1}^{n} a_j\, x_j = b \right\}$.

A (closed) half space in R^n, defined as $\left\{ x \in R^n \mid \sum_{j=1}^{n} a_j\, x_j \leq b \right\}$.

A convex cone, i.e. C is a convex cone if and only if whenever x and y belongs to C then $x + y$ and αx belongs to C, where $\alpha \geq 0$.

A point x is a convex combination of the points $x_1, x_2, \cdots, x_p$ if it can be expressed as:

$$x = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p$$

where

$$\alpha_1, \alpha_2, \cdots, \alpha_p \geq 0 , \quad \sum_{i=1}^{p} \alpha_i = 1$$

and a set X is convex if and only if every convex combination of points in X belongs to X.

If sets A and B are convex, then $A \cap B$ and

$$A + B = \{x \mid x = a + b , \ a \in A , \ b \in B\}$$

are also convex. However $A \cup B$ is not necessarily convex. The intersection of a finite number of closed half spaces is convex and is called a polyhedral convex set.

An extreme point of a convex set is an element of the set which cannot be expressed as a convex combination of two other points in the set. A set is strictly convex if and only if it is convex and all its boundary points are extreme points. A convex set need not, however, have any extreme point. An example is any open convex set.

The convex hull of a set A is the "smallest" convex set containing A, i.e. the intersection of all convex sets containing A. The set A equals its convex hull if it is convex.

The convex hull of a finite number of points in R^n is a convex polyhedron - a bounded polyhedral convex set which is the set of all convex combinations of the given points.

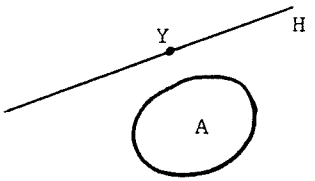
A closed bounded convex set is the convex hull of its extreme points. Given any convex closed set A in R^n and a point y in R^n, if $y \notin A$ then there exists a bounding hyperplane

$$H = \left\{ x \in R^n \mid \sum_{j=1}^{n} a_j x_j = b \right\}$$

containing y for which all points in A lie in one of the closed half spaces determined by H; i.e.

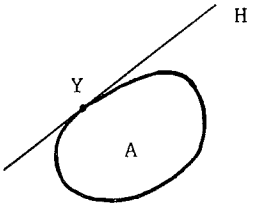
$$\sum_{j=1}^{n} a_j y_j = b \quad \text{and either}$$

$$\sum_{j=1}^{n} a_j z_j \leq b \quad \text{or} \quad \sum_{j=1}^{n} a_j z_j \geq b \quad \forall z \in A$$

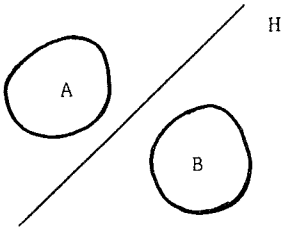


H is a bounding hyperplane for A

If y is a boundary point of A then there exists a supporting hyperplane H which contains y and for which all points in A lie in one of the closed half spaces determined by H. Given two non-empty convex sets A and B in R^n which are disjoint or have only boundary points in common there is a separating hyperplane H for which all points in A lie in one of the closed half spaces determined by H and all points in B lie in the other closed half space determined by H.

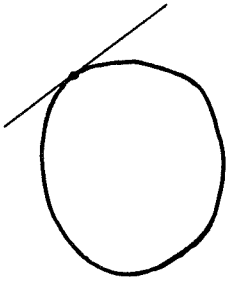


Supporting hyperplane

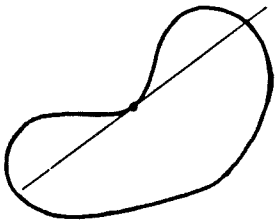


Separating hyperplane

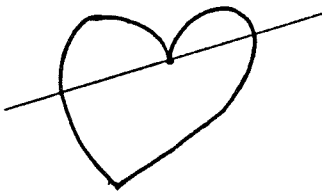
As supporting and tangent hyperplanes play a central role in mathematical programming, the diagram below shows several cases. A boundary point of a set may possess a hyperplane which is both supporting and tangent (case (i)), tangent without being supporting (case (ii)), supporting without being tangent (case iv)), or no hyperplane of any character may exist (case (iii)).



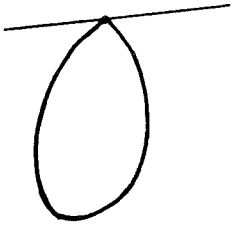
(i)



(ii)



(iii)



(iv)

Moreover case (iv) admits infinitely many supporting planes.

d) Special type of functions

A real valued function $f(x)$ defined on a convex set X is convex if and only if given any two distinct points x and y in X:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$$

for all α, where $0 \leq \alpha \leq 1$

An equivalent definition of convex function, if $f(x)$ is differentiable, is: $f(x)$ is convex if and only if

$$f(x_2) \geq f(x_1) + \nabla f(x_1)(x_2 - x_1)$$

for any x_1, x_2.

$f(x)$ is strictly or strongly convex if and only if the strict inequality holds. The function $f(x)$ is concave if and only if $-f(x)$ is convex. Some of the properties of convex functions are:

- If $f(x)$ and $g(x)$ are convex functions defined on X, then $f(x) + g(x)$, $\max[f(x),g(x)]$, and $cf(x)$, $c > 0$, are all convex.

- If $f(x)$ is convex and has continuous second partial derivatives on an open convex set X, then the Hessian matrix is positive semidefinite and conversely. If $H(x)$ is positive definite then $f(x)$ is strictly convex and conversely.

A differentiable function $f(x)$ is pseudoconvex if and only if

$$\nabla f(x_1)(x_2 - x_1) \geq 0$$

implies

$$f(x_2) \geq f(x_1) , \quad \text{for} \quad x_2, x_1 \in X$$

$f(x)$ is pseudoconcave if and only if $-f(x)$ is pseudoconvex. Obviously a convex differentiable function is also pseudoconvex but the converse is not true.

The real valued function $f(x)$ defined on a convex set X is quasi-convex if and only if given any two distinct points $x, y \in X$:

$$f(\alpha x + (1 - \alpha)y) \leq \max[f(x), f(y)]$$

$$\forall \alpha , \quad 0 \leq \alpha \leq 1$$

$f(x)$ is quasi-concave if and only if $-f(x)$ is quasi-convex.

The function $f(x)$ is quasi-convex if and only if the sets

$$\{x \in X \mid f(x) \leq b\}$$

for any real number b, are convex.

If $f(x)$ is differentiable, then it is quasi-convex if and only if

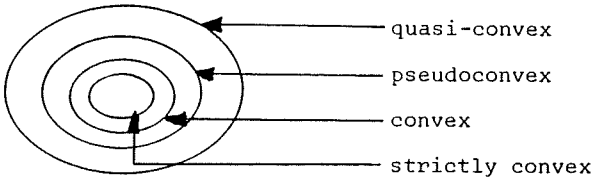
$$f(x_2) \leq f(x_1)$$

implies

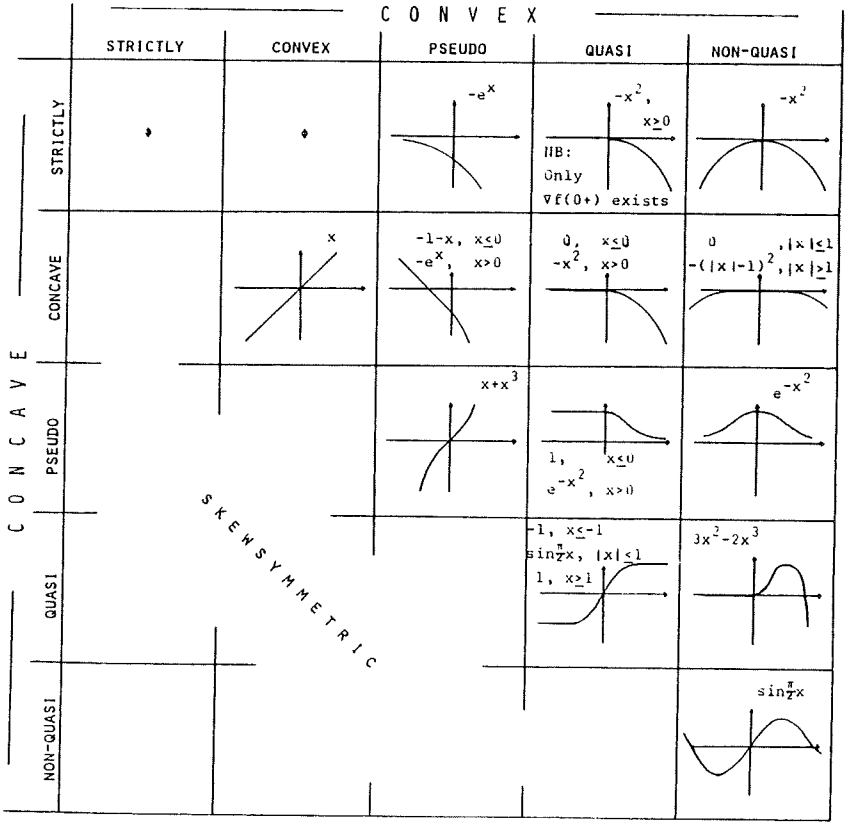
$$\nabla f(x_1)(x_2 - x_1) \leq 0$$

$$\text{for} \quad x_2, x_1 \in X$$

The relation between these types of functions are illustrated in the next page for $f \in C^1$



Obviously a convex differentiable function is also pseudoconvex and quasi-convex but the converse is not necessarily true. The table below illustrates with some examples the relations between concavity and convexity properties of functions.



Some of the most interesting properties of these functions are:

- If $f(x)$ is convex in R^n, and $g(\ )$ is a non-decreasing convex function in R, then $h(x) = g(f(x))$ is convex in R^n.

- If in the last property $g(\ )$ is only a non-decreasing function in R, then $h(x) = g(f(x))$ is quasi-convex in R.

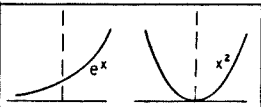
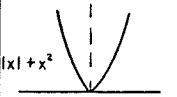
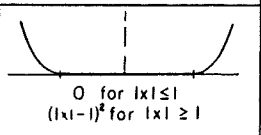
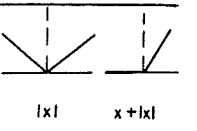
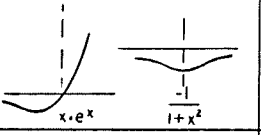
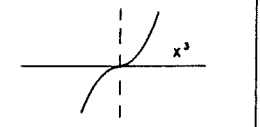
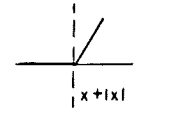
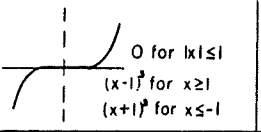
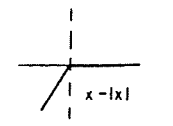
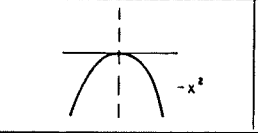
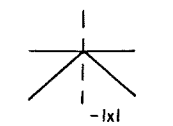
- If $h(x)$ is concave, then $f(x) = \frac{1}{h(x)}$ is convex in $X = \{x \mid h(x) > 0\}$.

- The function $f(x) = \frac{cx + \alpha}{dx + \beta}$ is both pseudoconvex and pseudoconcave for all convex sets X, where $dx + \beta \neq 0$.

- a^x and $-\log(x)$ are convex and x is a scalar.

- If $f(x) > 0$ within X, an open convex set of R^n. Then if $\log f(x)$ is convex, $f(x)$ is also convex in X.

- If $g(x)$ is convex in R^n, then $f(x) = e^{g(x)}$ is also convex in R^n.

Type	Examples of continuously differentiable functions $f(x)$	Not everywhere differentiable functions $f(x)$
strongly convex		
convex, not strongly convex		
pseudoconvex, not convex		contradicts the definition
strongly quasi-convex, not pseudoconvex		
quasiconvex, not strongly quasi-convex		
not quasiconvex		

Examples of various types of functions

e) Supergradient and subgradient

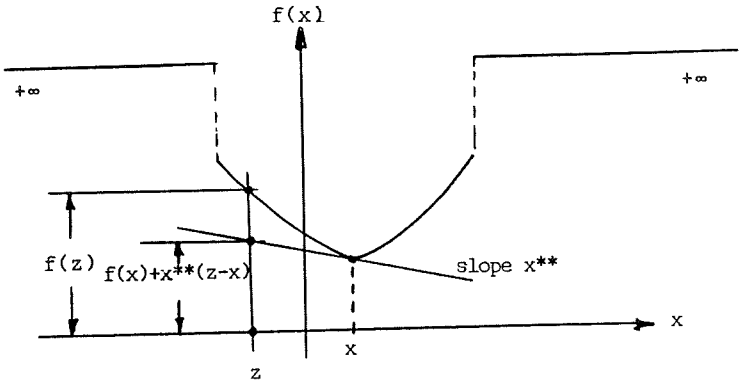
A row vector x^* is said to be a supergradient of $f(x)$ at a point x if

$$f(z) \leq f(x) + x^*(z - x) \quad , \quad \forall z \in R^n$$

A (row) vector x^{**} is said to be a subgradient of $f(x)$ at a point x if

$$f(z) \geq f(x) + x^{**}(z - x) \quad , \quad \forall z \in R^n$$

The figure below illustrate this definition.



f) Farkas' Lemma

Let $\{P_0, P_1, \cdots, P_r\}$ be an arbitrary set of vectors. There exist $\beta_i \geq 0$ such that

$$P = \sum_{i=1}^{r} \beta_i P_i$$

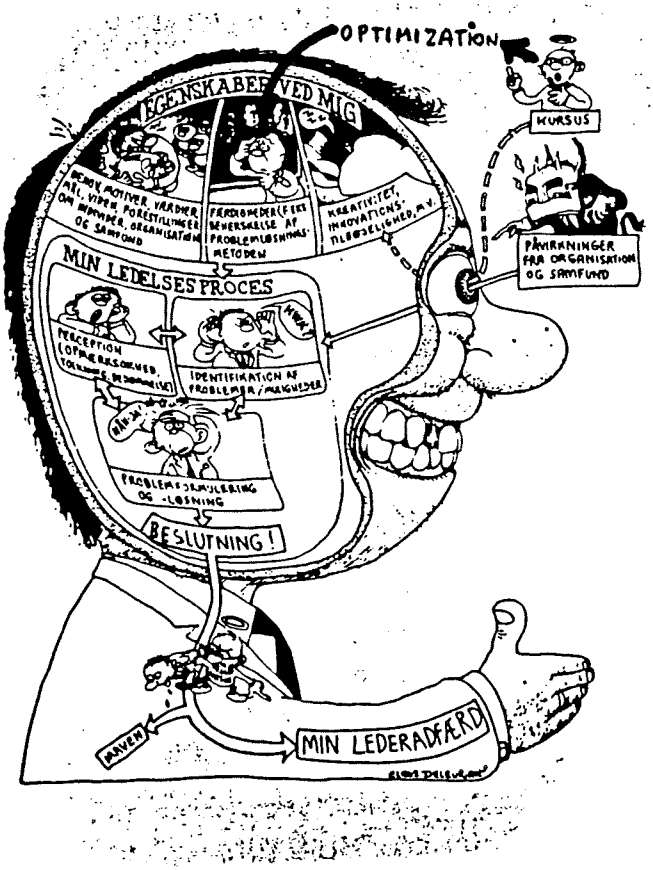
if and only if

$$P' y \geq 0$$

for all y satisfying

$$P_i' y \geq 0 \quad , \quad i = 1, \cdots, r$$

----- 0 -----


OPTIMIZATION
EGENSKABER VED MIG
MOTIVER, VÆRDIER
FERDIGHEDER (FRA
KREATIVITET,
KURSUS
PÅVIRKNINGER
FRA ORGANISATION
OG SAMFUND
MIN LEDELSES PROCES
PERCEPTION
IDENTIFIKATION AF
PROBLEMER / MULIGHEDER
BESLUTNING!
MAVEN
MIN LEDERADFÆRD